Multiscale modelling of DNA point mutations: the effect of the environment and replication enzymes

Max Winokan

A dissertation submitted for the degree of

Doctor of Philosophy

December 2023

Quantum Biology Doctoral Training Centre, University of Surrey,

Guildford, Surrey GU2 7XH, United Kingdom

Abstract

DNA mutation occurring in the absence of external factors remains an open problem. Watson and Crick's pioneering work on DNA structure suggested that rare protonation states of DNA bases could be a source of such *spontaneous* mutations. Löwdin proposed quantum tunnelling as a source of tautomeric A^*T^* and G^*C^* nucleotide pairs. During replication, such tautomeric pairs can form mismatches (A^*C, GT^*, AC^*, and G^*T) that evade error correction mechanisms by mimicking the structure of canonical DNA. Existing computational models of DNA tautomerism have often oversimplified the dynamical aspects of replication enzymes and their biological microenvironment. This thesis extends established methods to provide a comprehensive description of the quantum phenomenon of proton transfer in DNA. The effect of mechanical separation forces on the behaviour of AT and GC base pairs serves as a model of strand separation conditions induced by a helicase enzyme. The results from complementary quantum mechanical (QM) and molecular mechanics (MM) models reveal that tautomeric populations can be trapped despite their (sub-picosecond) lifetimes, challenging previous assumptions which dismiss their relevance to mutations. The effect of a more realistic replication environment on proton transfer within DNA is explored. Multiscale QM/MM models allow for the determination of the free energy surface experienced by protons in both aqueous DNA and in complex with an explicit helicase enzyme. The results highlight that PcrA helicase reduces the stability of G^*C^*, potentially due to a natural selection pressure to reduce spontaneous mutations. Equilibrium descriptions of the proton transfer obtained from umbrella sampling were complemented with an instantaneous approach where the proton can be isolated from other atomic motions, leading to a novel time-dependent potential for the proton transfer. An alternative mechanism for spontaneous mutations is considered, involving proton transfer within a polymerase enzyme during DNA synthesis. Dynamical QM/MM simulations suggest that the polymerase's dynamics could prime a GT *wobble* mismatch into a *tunnelling-ready* state resulting in a heightened rate of mismatches. Finally, the role of tautomers in genetic disorders is contextualised, and the limitations of existing theoretical descriptions are acknowledged. The potential for future research, including investigating quantum effects in DNA methylation, gene expression, and epigenetics is outlined.

Declaration

This thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images or text resulting from the work of others (whether published or unpublished, and including any content generated by a deep learning/artificial intelligence tool) are fully identified as such within the work and attributed to their originator in the text, bibliography or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification. I agree that the University has the right to submit my work to the plagiarism detection service TurnitinUK for originality checks. Whether or not drafts have been so-assessed, the University reserves the right to require an electronic version of the final document (as submitted) for assessment as above.

Signature:

Date: December 4, 2023

Acknowledgements

First and foremost, I would like to express my utmost gratitude to my PhD supervisors Marco Sacchi and Jim Al-Khalili for the opportunity to pursue my research in this exciting and novel field.

Marco has been an exemplary supervisor since I started my PhD, available seemingly night and day for advice and thorough critique of computational chemistry methods and all aspects of research. Marco 's passion for quantum biology and computational chemistry is admirable, and Marco has been exceptionally engaged with my project. Our Friday meetings were a weekly highlight that included electric discussions on topics ranging from quantum chemistry to travel photography. Without Marco's attention to simulation details and high standard for analytical rigour the PhD would not have been possible.

From small group tutorials in my first week as a Physics undergraduate at Surrey to a continuous guiding light throughout my PhD, Jim has been exceptionally supportive and a great source of stimulating dialogue. I will always be grateful for the opportunities Jim has given me at Surrey, from funding the QB DTC and PhD, to outreach possibilities such as animations for his TV documentary and an all-access pass to the Cheltenham Science Festival. Despite his busy schedule, Jim was always reliably available for the most thorough proofreading I could have hoped for.

In addition to my supervisors, Louie Slocombe has been a highlight of my time at Surrey. At my dining table in Guildford, I remember Louie passionately pitching the topic of his PhD and thinking how much I would have liked to work on proton transfer in DNA myself. Little did I know that Louie and I would work together for four years, starting as PhD students wrestling with QM/MM, and publishing four (and counting) papers together. Aside from being a close friend, Louie has been an excellent collaborator and is one of the most ambitious, professional, and fun people in academia.

I also thank Cedric Vallee for many great memories in the QB DTC office and numerous European cocktail bars. Cedric has been an excellent friend, academic collaborator, and table tennis opponent. I can't wait to continue our adventures at Diamond Light Source and beyond.

While the above individuals were all a notably positive part of my research, many more people supported me throughout my time at Surrey: my collaborators Ben King and Paul Stevenson; Johnjoe McFadden and Youngchan Kim for leadership of the QB DTC; all my friends and peers in the QB office; Justin Read, Alessia Gualandris, Alexis Diaz Torres, Richard Sear, Paul Stevenson and many other members of the Physics department; and all administrative and support staff at the university.

I am also grateful for the financial support from the University of Surrey, Leverhulme Trust, John Templeton Foundation, HecBioSim consortium, and UK Carr-Parinello Consortium, who provided funding for my research and resources for many hours of calculations on the HPCs EUREKA, EUREKA2, Archer, and Archer2.

Beyond academics, I am very grateful for the support of my partner Jen, parents Richard and Viola, sister Una, and wider family for their support and understanding for the countless hours spent behind my laptop at home and on vacations. I thank my housemates Iain, Sean, Billy and Katherine for the great memories like the Super Dinner League and Loyd's Gross Men fixtures. I also thank all my friends at Team Surrey Ultimate for much-needed relief from academia across many training sessions, board game socials, and tournaments away. I am also very grateful to Paolo Rigiroli, Matt Thorman, and Electronic Arts for supporting me during my mid-PhD internship where I learned many skills that accelerated my academic career. I also thank Frank von Delft, Warren Thompson, Kate Fieseler and the rest of XChem at Diamond Light Source for their support and understanding of my busy transition from PhD to post-doctoral researcher.

Finally, I would like to thank my confirmation examiners Professors Guoping Lian and Brendan Howlin, and PhD examiners Professors Brendan Howlin and Adrian Mulholland.

Publications

1. Louie Slocombe, Max Winokan, Jim Al-Khalili, and Marco Sacchi. "Proton transfer during DNA strand separation as a source of mutagenic guanine-cytosine tautomers". In: *Communications Chemistry* 5.1 (Nov. 2022), p. 144. DOI: https://doi.org/10.1038/s42004-022-00760-x [1]

2. Louie Slocombe, Max Winokan, Jim Al-Khalili, and Marco Sacchi. "Quantum Tunnelling Effects in the Guanine-Thymine Wobble Misincorporation via Tautomerism". In: *The Journal of Physical Chemistry Letters* 14 (Jan. 2023), p. 9-15. DOI: https://doi.org/10.1021/acs.jpclett.2c03171 [2]

3. Benjamin King, Max Winokan, Paul Stevenson, Jim Al-Khalili, Louie Slocombe, and Marco Sacchi. "Tautomerisation Mechanisms in the Adenine-Thymine Nucleobase Pair during DNA Strand Separation". In: *The Journal of Physical Chemistry B* 127 (Mar. 2023), p. 4220-4228. DOI: https://doi.org/10.1021/acs.jpcb.2c08631 [3]

4. Max Winokan, Louie Slocombe, Jim Al-Khalili, and Marco Sacchi. "Multiscale simulations reveal the role of PcrA helicase in protecting against spontaneous point mutations in DNA". In: *Scientific Reports* (2023). DOI: https://doi.org/10.1038/s41598-023-48119-z [4]

Contents

List of Figures

List of Tables

Glossary & Acronyms

A Adenine. 3, 7

AA Amino Acid. xx, 153

AA Amino Acids. 42, 48

ACM Adiabatic Connection Method. 25

AM Adiabatic Mapping. x, xi, xix, 31, 32, 34, 39, 40, 41, 42

ASCII American Standard Code for Information Interchange. 151

AT Adenine-Thymine. xi, 1, 2, 9, 11, 38, 41, 42, 119, 124, 126, 128

B3LYP Becke, 3-parameter, Lee-Yang-Parr. xi, 9, 38, 49, 51

BLYP Becke-Lee-Yang-Parr. 25

C Cytosine. 3, 7

CASTEP Cambridge Serial Total Energy Package. xi, 38, 40, 42

DCP Dispersion Correcting Potentials. 25

DFT Density Functional Theory. xi, 1, 9, 10, 24, 25, 26, 28, 38, 39, 41, 47, 49, 51, 52, 69, 104, 123, 124, 125, 126, 129

DNA Deoxyribonucleic Acid. x, xi, xii, xix, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 20, 21, 26, 28, 33, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 51, 52, 69, 115, 117, 119, 120, 121, 123, 124, 125, 127, 128, 129, 130

DoF's Degrees of Freedom. 128

DPT Double Proton Transfer. x, xii, xix, 7, 8, 9, 10, 11, 12, 37, 49, 51, 115, 117, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130

dsDNA double-stranded DNA. 3, 4, 5, 8, 120, 121, 122, 125, 127, 128, 130

EM Energy Minimisation. xix, 28, 29, 31, 41

FF Force Field. x, xi, 21, 22, 26, 27, 28, 37, 38, 39, 41, 42

FF Force Fields. 49

FWHM Full Width at Half Maximum. 47

G Guanine. 3, 7, 129

GAFF General Amber Force Field. xi, 38

GC Guanine-Cytosine. xii, xix, xx, 1, 2, 7, 9, 10, 11, 26, 41, 46, 49, 50, 51, 117, 118, 119, 121, 124, 125, 126, 127, 128

GGA Generalised Gradient Approximation. 25

Gromacs Groningen Machine for Chemical Simulations. xi, 40

HF Hartree-Fock. 9, 25

HPC High-Performance Computing. 52, 69, 104

IBT Intra-Base Transfer. 122

iPES instantaneous Potential Energy Surface. 128

KIE Kinetic Isotope Effect. 129

LDA Local Density Approximation. 25

MD Molecular Dynamics. xi, xii, 1, 26, 29, 30, 41, 42, 43, 44, 45, 46, 47, 52, 69, 104, 125, 127, 128, 129

MEP Minimum Energy Path. 19, 31, 33, 125, 128, 129

MGGA Meta Generalised Gradient Approximation. 25

MM Molecular Mechanics. 1, 21, 22, 26, 27, 28, 30, 34, 37, 39, 41, 48, 49, 123, 124

MP-2 second-order Møller-Plesset perturbation theory. 9, 10

NA Nucleic Acid. xx, 42, 47, 48, 153

NA Nucleic Acids. 42, 43, 48

NEB Nudged Elastic Band. xi, xii, 18, 31, 32, 33, 34, 47, 48, 125, 126, 128, 129

NMR Nuclear Magnetic Resonance. 8, 11

NPT Number Pressure Temperature. xi, 44

NTP Nucleoside Triphosphate. 5, 128, 129

NVT Number Volume Temperature. xi, 44, 45

ONIOM *our own n-layered integrated molecular orbital molecular mechanics*. 10, 27

OQS Open Quantum Systems. 20, 123, 130

PBE Perdew-Burke-Ernzerhof. xi, 40

PcrA Plasmid copy reduced. 2, 5, 37, 124, 127, 128, 130

PES Potential Energy Surface. x, xi, xii, 18, 31, 32, 34, 48, 128

PMF Potential of Mean Force. 35, 36, 49, 127, 128

PMF Potentials of Mean Force. 51

PT Proton Transfer. 20, 31, 37, 123, 124, 125, 126, 129

QM Quantum Mechanics. 1, 19, 22, 26, 27, 28, 33, 37, 38, 39, 47, 48, 124

QM/MM Quantum Mechanics/Molecular Mechanics. x, xii, 1, 10, 11, 12, 22, 26, 27, 28, 36, 37, 41, 47, 48, 49, 51, 104, 124, 127, 128, 129

RC Reaction Coordinate. x, xi, 31, 32, 35, 49

RC Reaction Coordinates. 35

RMS Root Mean Square. xi, xxiii, 44, 129

RMSD Root Mean Square Displacement. xi, 44

RNA Ribonucleic Acid. 5

SC Side Chain. xx, 153

SCF Self-Consistent Field. 9

SD Steepest Descent. 28

SE Schrödinger Equation. 23, 24

SE Semi-Empirical. 37, 123, 124

SMD Steered Molecular Dynamics. 22, 35, 49, 52, 69, 125, 126

SPE Single-Point Energy. 36, 128

SPE's Single-Point Energies. 34

ssDNA single-stranded DNA. xi, xii, 3, 4, 5, 6, 10, 41, 43, 49, 117, 120, 122, 124, 127, 128, 130

T Thymine. 3, 7, 129

tRNA transfer Ribonucleic Acid. 8

TRS Tunnelling Ready State. 2, 104, 129

TS Transition State. x, 17, 18, 19, 31, 32, 33, 35, 128

TST Transition State Theory. 9, 17, 18, 19, 29

US Umbrella Sampling. xii, 18, 22, 31, 32, 34, 35, 36, 37, 49, 50, 124, 127, 128

UV Ultra-Violet. 8

vdW van der Waal. x, xi, 3, 25, 43

w.r.t. with respect to. 19

WC Watson-Crick. 3, 5, 7, 8, 41, 120, 124, 128, 129

WHAM Weighted Histogram Analysis Method. 35

WKB Wentzel-Kramers-Brillouin. 19

XCF Exchange Correlation Functional. 25, 49

XCF's Exchange Correlation Functionals. 25

XRD X-Ray Diffraction. 43

ZPE Zero-Point Energy. 19

1. Introduction

DNA encodes the genetic information to synthesise all the proteins necessary for life. In order to facilitate a stable hereditary system nature has evolved complex machinery to ensure that the genetic code is unravelled, expressed, and duplicated with astounding accuracy. Despite this, certain levels of mutation introduce diversity into populations. Favoured phenotypes are more likely to pass on their genetic material, and in this way natural selection optimises a genome over many generations. Within our cells there is an intrinsic competition between sources of DNA damage and sophisticated error-correcting mechanisms. DNA encodes genetic information through a sequence composed of four nucleobases which reliably hydrogen bond to form the pairs Adenine-Thymine (AT) and Guanine-Cytosine (GC).

Mutations which arise in the absence of external factors are known as spontaneous, and their source remains an open problem. Spontaneous mutation has been postulated to occur due to altered protonation states known as tautomers. These tautomers can dimerise with the canonically incompatible nucleobases while maintaining the helical secondary structure of DNA. Without tautomerisation, base pair mismatches disrupt the structure of DNA, by forming wobble conformations which can be detected by error-correcting mechanisms. Thus, tautomeric DNA bases may form mispairs that evade detection and can lead to mutation.

Determining the feasibility of spontaneous mutation via proton transfer experimentally has proved difficult, and many authors have turned to theoretical models of proton transfer in DNA. In this thesis, I aim to develop our computational models of proton transfer in DNA by introducing biologically realistic environmental interactions. Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) methods can be used to combine a highly accurate quantum mechanical region within a large-scale atomistic Molecular Mechanics (MM) system. Density Functional Theory (DFT) provides an *ab initio* description of the QM region, and applying Molecular Dynamics (MD) algorithms allows us to describe the biological ensemble.

The protons in hydrogen bonds are known to be delocalised, prompting investigations into the nuclear quantum effects involved in their transfer. For the first time, explicit modelling of replication enzymes with quantum chemical descriptions of DNA reveals how different biological microenvironments can interact with mutagenic tautomeric states. In this thesis, the effect of two key stages of DNA replication on tautomerism is explored. Models of strand separation by helicase enzymes, and the synthesis of new duplex DNA by polymerase enzymes reveal that quantum effects

can have non-trivial consequences for our genome.

1.1. Thesis Structure

This thesis is organised into two parts. The first introduces the reader to the biological context, chemical theory, and computational models in this work, before moving on to the presentation of novel results and discussion thereof.

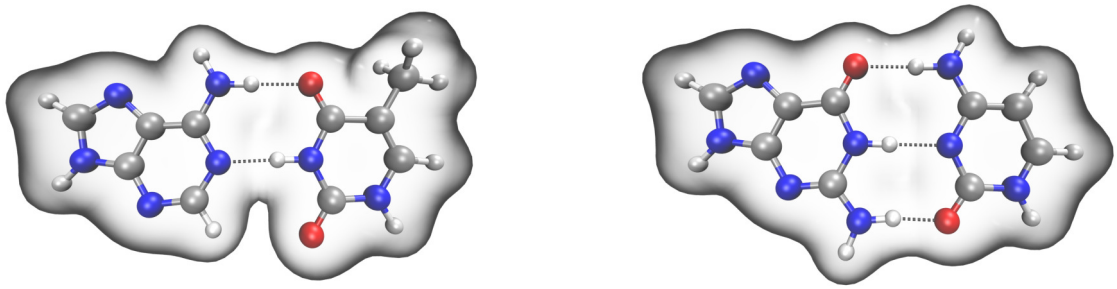
Chapter 2 introduces the reader to the context of the work presented in this thesis, including relevant biological mechanisms and a review of previous works. Chapter 3 provides some background into the behaviour of chemical reactions in biology, and theories that describe their kinetics. Chapter 4 details the models used to approximate the molecular systems and the simulation algorithms that are used to manipulate and investigate such systems. Chapter 5 describes how computational chemistry models and algorithms can be applied to map reactions of interest.

Chapter 6 details precursory investigations into the behaviour of hydrogen-bonding in DNA and results on its stability. Additionally, the PcrA Helicase-DNA complex is introduced. Chapter 7 investigates the mechanical action of strand separation on tautomerism in GC. Chapter 8 similarly models strand separation on tautomerism in AT, and restores its viability as a candidate for point mutations. Chapter 9 Introduces the first explicit model of tautomerism within a helicase enzyme and indicates that protein-DNA interactions may decrease the population of tautomers. Chapter 10 details how proton transfer may lead to the misincorporation of a GT wobble mismatch, including explicit models of the polymerase enzyme which aids the population of a Tunnelling Ready State (TRS). Chapter 11 contains exploratory results which suggest the need for theoretical advancement to determine the true population of tautomers existing in inherently dynamical biology. Chapter 12 discusses the key findings from the research and comments on our current understanding of spontaneous mutations.

2. Biological Context and Literature Review

In the 19th century, Charles Darwin postulated that evolution and inheritance of phenotypes were linked. He observed that offspring possess variants different from their parents. From his famous expeditions to the Galapagos islands, he reported that better traits are *naturally selected* due to their higher survival rate, and thus the genetic material encoding them is passed on at a greater rate[5]. Elsewhere, Gregor Mendel's investigations of pea plants and their cross-breeding revealed that some traits are dominant and some are recessive. This led to the development of the mathematical model of inheritance where genetic material is stored as alleles[6].

It would take almost a century to reveal the structure of the molecule encoding our genetic material. In 1953 Franklin and Gosling published x-ray crystallographic data of DNA[7]. From these findings Watson and Crick determined DNA's now famous double helix structure[8]. The reliable pairing of Adenine (A) with Thymine (T) (see Figure 2-1a) and Guanine (G) with Cytosine (C) (see Figure 2-1b) is now referred to as the canonical or Watson-Crick (WC) pairing. The purine (A and G) and pyrimidine (T and C) nucleobases are combined with a ribose sugar and a negatively charged phosphate group to form nucleotides. Any number of nucleotides can be chained together to form single-stranded DNA (ssDNA). Two complementary ssDNA strands can form a double-stranded DNA (dsDNA) duplex molecule.



(a) Adenine-Thymine (b) Guanine-Cytosine

Figure 2-1 The two canonical Watson-Crick nucleobase dimers. White spheres are hydrogen, carbon is grey, nitrogen is blue, and oxygen is red. Black dotted lines show the hydrogen bonds. The black surface around the dimers is the vdW surface.

The chemistry of DNA involves several types of interactions. While covalent bonds join the nucleotides together into ssDNA, non-covalent interactions govern the formation of DNA's helical secondary structure. Adjacent bases in the same strand of DNA exert a stabilising effect on each other

via π-π stacking, and the two strands of DNA are bound together via a hydrophobic interior, where hydrogen bonds readily form between compatible nucleobases. Stacking enhances bond-accepting capacity but reduces donating strength of DNA bases[9]. Under aqueous conditions (the solubility of DNA is facilitated by its negative charge), DNA takes on a right-handed helical form known as B-DNA. While the solvent effects stabilise the B-DNA molecule[10], external factors such as pH and salt concentration can lead to less common forms of DNA: A-DNA, and Z-DNA[11, 12]. In fact, environmental effects cause the same sequence of DNA to take on different conformational geometries[13]. This dynamical nature of DNA requires careful treatment for experimental and theoretical investigators alike. DNA and in fact, most macromolecules should be modelled as a statistical ensemble of many competing conformations.

2.1. DNA Replication

The human genome contains approximately 2.5×10^8 base pairs[14]. DNA replication plays an important role in life, generally making flawless copies of chromosomal DNA in a process involving a number of highly specialised enzymes which collectively are termed the "replisome". Broadly speaking the process can be summarised as follows.

1. The two chains of double-stranded DNA (dsDNA) are separated by a class of enzymes named "helicases".

2. Both of the resulting single-stranded DNA (ssDNA) chains serve as a template from which DNA "polymerases" form a new duplex.

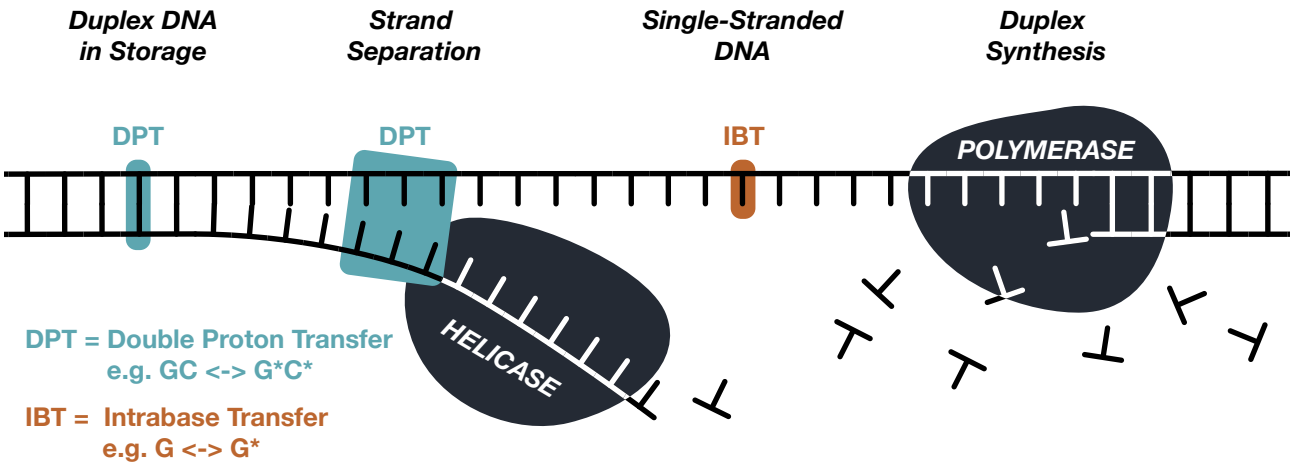


Figure 2-2 Illustration of the DNA replication process and sites for proton transfer therein.

Numerous types of replisomes exist across different species, often with a unique helicase performing strand separation. Broadly speaking, helicases perform an active *translocation* of ssDNA through a binding site, taking energy to perform this action from the hydrolysis of an ATP

molecule. In this thesis, we discuss an explicit model of a bacterial helicase named Plasmid copy reduced (PcrA) which is involved in generating plasmids. The dynamics of PcrA includes a *stepping motor/inchworm* mechanism suggested by the determination of its crystal structure[15], experiments[16, 17] and theoretical considerations[18, 19]. Additionally, key amino acid residues in the ssDNA binding site of PcrA have been identified by Dillingham, *et al.* in [20], by performing an alanine scan where the side chain of a single residue is removed and observing the effect on the helicase's function.

Polymerase enzymes also differ across replisomes, and even within a single cell there are several polymerases with varying roles and fidelities. DNA polymerases (as opposed to RNA polymerases) synthesize double-stranded DNA from a single-stranded template. The polymerase proceeds along the DNA in a stepwise fashion and allows for a Nucleoside Triphosphate (NTP) to diffuse into a "palm" domain. Once a NTP has entered the binding pocket of the polymerase the *thumb* domain of the enzyme will close down, attempting to incorporate the new nucleotide. Depending on the pairing of this nucleotide with the ssDNA template, and the ability of the polymerase to detect mismatches (fidelity), these may be rejected at this point. While the free energy difference between WC and mismatch paired nucleotides is small (< 0.15 eV), high-fidelity polymerases can recognise wobble pairings such as (G-T) by the disruption of the canonical helical secondary structure[21]. This along with exonuclease activity - which is a proof-reading mechanism that serves to cleave mismatches off the end of DNA chains in higher fidelity polymerases - ensures that the rate of errors produced by polymerases is exceptionally low at around one in ten million[22, 23]. Following a misincorporation, further errors can be removed during the mismatch repair phase of the cell cycle[24]. In Chapter 10, an explicit model of a λ-polymerase was implemented. This human polymerase is a monomer and lacks exonuclease activity which is a proof-reading mechanism that serves to cleave mismatches off the end of DNA chains in higher fidelity polymerases[25, 26, 27, 28, 29].

2.2. Mutations

Genetic information is expressed via the transcription of DNA into RNA and then translation into a sequence of amino acids which form proteins. The transcription of DNA into RNA is 1:1 with every DNA nucleotide coding for its RNA counterpart which, with the exception of uracil replacing thymine, have the same nucleotide components. Translation of RNA into amino acids, however, involves triplets of nucleotides known as codons. With 20 amino acids to code for, neither singlet (4 permutations) nor doublets of nucleic acids ($4 \times 4 = 16$ permutations) would be sufficient to code for any protein. With 64 available codons, 20 amino acids, and three stop codons (which truncate the polypeptide chain synthesis), there is significant redundancy in codons. This redundancy provides the first layer of defence against mutations as a single letter change only has a fractional chance of changing the amino acid that will be synthesised. In the case where the DNA point mutations do not

result in changes to the phenotype, it is deemed *silent/synonymous*. Conversely, non-synonymous mutations can change the amino acid encoded by the codon (missense), or prematurely terminate the protein (nonsense). Missense can result in a conservative mutation where the amino acid type remains unchanged or a non-conservative mutation if the amino acid type is also altered.

Despite the error-correcting mechanisms described in the previous section, mutations still arise during DNA replication. The cellular environment protects genetic material from external influence, but radiation, metallics, electric fields, carcinogens, and ultraviolet factors may tamper with the DNA structure or modify base pairing chemistry[30, 31, 32, 33, 34]. When mutations occur in the absence of external factors, the mutation is *spontaneous*. Spontaneous mutations can arise through a rotation of a nucleobase (Hoogsteen base pairing), removal of a nucleobase (depurination), hydrolysis of an amine group (cytosine $\rightarrow$ uracil via deamination), slipped strand mispairing (a polymerase skips over a ssDNA hairpin), ionisation, and non-canonical protonation states (tautomerism)[8, 22, 35, 36, 37, 38].

While the total rate of mutations has been experimentally determined in various replisomes, there is little clarity on the precise contributions of different routes to mutation[39]. It is known that polymerases attempt to pair a mismatch approximately one in ten million base pairs, with exonuclease activity providing around 100-fold improvement to fidelity[23]. Subsequent post-polymerase mismatch repair mechanisms provide a further 1000-fold improvement[40, 41, 42]. This leads to a total fidelity of around 1 error in 1×10^{10} base pairs[43, 44].

Mismatch repair in the human replisome primarily involves two enzymes. MutS which passively scans for mismatches, and MutL which actively begins the repair process[45, 46, 47, 48]. Mismatch repair has been shown to require both these enzymes and energy from hydrolysing ATP[49]. The physical mechanisms of mismatch recognition are intriguing, and MutS is able to probe the non-covalent interactions between complementary strands to detect issues, going beyond the recognition of non-canonical backbone conformations employed by the majority of DNA repair enzymes[50, 51].

2.3. Tautomerism in DNA

Watson and Crick were the first to suggest that spontaneous mutation may be due to altered protonation states they called “tautomeric shifts”[8]. Protonation states can be impacted by acid-base chemistry and the pK_a values for nucleobases in solution have been experimentally determined[52]. Cellular pH (6-8) rules out nucleobase ionisation leading to altered protonation states[53, 54]. In 1963, an alternative mechanism for these tautomeric shifts was suggested by Löwdin via proton transfer[36, 55]. Specifically, Löwdin suggested that quantum tunnelling of the protons in the hydrogen bonds of DNA can cause mutations. The delocalisation of protons due to their quantum wave can lead to a finite probability of the proton ending up bound to a non-canonical site. Löwdin suggested that tunnelling may drive mutations in all life.

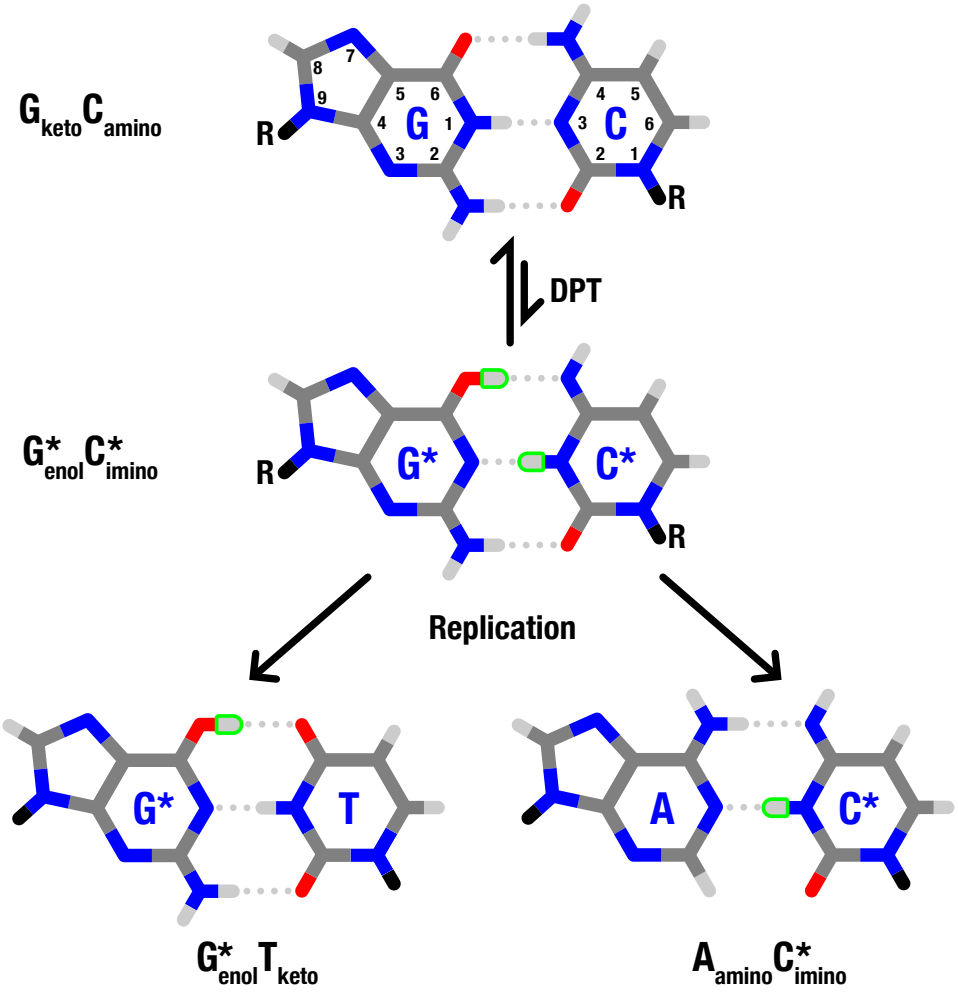


Figure 2-3 Mechanism for spontaneous mutation of GC via a Double Proton Transfer (DPT).

The mechanism for spontaneous mutations proposed by Löwdin is illustrated in Figure 2-3. Canonically the WC pair Guanine-Cytosine (GC) forms three hydrogen bonds, two of which are between the keto group of Guanine and the amino group of Cytosine. Löwdin identified that a Double Proton Transfer (DPT) across this pair of hydrogen bonds leads to a tautomeric enol-imino form that conserves the WC geometry. This tautomeric dimer is referred to as G^*C^*. Notably, both the enol form of Guanine and the imino form of Cytosine fit the hydrogen bonding profile of Thymine (keto) and Adenine (amino), respectively. This allows for the formation of non-canonical base pairs G^*T and AC^* . These nucleotide pairs form geometries resembling WC pairs despite being mismatches and thus are postulated to evade error correction mechanisms that rely on the detection of disrupted DNA's helical secondary structure.

Topal and Fresco showed that these non-WC these purine-pyrimidine mismatches still fit in the geometric constraints of the double helix[56]. Such mismatches have been observed in aqueous duplex DNA and within the active site of DNA-polymerase[37, 57, 58, 59, 60, 61, 62]. The rates of spontaneous mutation have also been experimentally determined up to a maximum of one error in 1×10^8[63, 64, 65, 66]. For tautomerism to be relevant the probability of equilibrium population of tautomers should exceed 1×10^{-8} to allow for a certain headroom to evade error correction[24, 67].

2.4. Existing evidence for tautomerism

Unfortunately, direct evidence of spontaneous mutation via tautomerism is rare. Genome analysis has consistently revealed a GC:AT mutation bias, meaning GC base pairs are more commonly mutated[68, 69, 70, 71]. External factors such as reactive oxygen species, and UV do not explain this bias[71]. Simulations suggest that DPT occurs more frequently in GC compared to AT[72, 73]. Further simulations suggest the tautomers lack the required stability for biological relevance[74, 75]. Nuclear Magnetic Resonance (NMR) spectroscopy of GC in tRNA found tautomerisation rates on the order of 100/s and a G^*C^* population fraction of 0.1[76]. Such NMR relaxation dispersion of GT showed a $GT^* \leftrightarrow G^*T$ tautomerisation reaction to be at least 1×10^5/s; Enol tautomeric base pairs exist within DNA sequences, and GT^* mimics WC GC. Observation of GT^* DPT is the best experimental representation of the DPT in canonical GC[62, 77].

2.5. Strand separation and base pair opening

Strand separation involves breaking the hydrogen bonds in double-stranded DNA, and is the first step of replication/transcription and mismatch repair (see Figure 2-2). Imino protons (see Figure 2-3) in the hydrogen bonding reveal the opening-closing kinetics as closed base pairs do not exchange imino protons. Open base pairs can exchange these protons with water (acid-base reaction), which can be revealed by NMR. Base pair opening activation energies range between 0.5 and 1.1 eV[78, 79, 80, 81]. Strand separation in live cells is thought to occur over milliseconds to seconds[82, 83].

2.6. Previous computational modelling of DNA

Quantum mechanical (electronic structure) methods are more accurate for modelling bond-breaking and bond-forming investigations in proton transfer. However, classical molecular mechanics can still reveal details about DNA interactions. This includes binding affinities of intercalating drugs[84], electronic properties[85], hydration effects[86], and melting points[87]. Structural stability of double-stranded DNA comes from electrostatic interactions, VdW interactions (base pair hydrogen bonding) and intrastrand π-π stacking. The accuracy with which these interactions are modelled is determined by the choice of the force field. Inaccuracies in force-fields can build up over long simulations, leading to inaccurate flexibility patterns for example[88, 89]. Many different force fields are available for DNA, and none are perfect. Extensive benchmarking is available in [90]. Polarisable force fields may further improve accuracy[91, 92]. Molecular dynamics simulations are often started from, benchmarked to and compared against experimental crystal structures leading to many force fields being extremely accurate at reproducing equilibrium conformations but struggling to capture non-equilibrium dynamics.

2.7. Previous computational modelling of proton transfer

Initial *in silico* proton transfer investigations are often performed on the formic acid dimer. Here, Density Functional Theory (DFT) provides excellent agreement with higher levels of theory such as second-order Møller-Plesset perturbation theory (MP-2) and improves accuracy over Self-Consistent Field (SCF) models[93].

For tautomerism in Guanine-Cytosine (GC) and in Adenine-Thymine (AT), a rich literature of computational investigations exists. For example, Florian, *et al.* [72] modelled proton transfer in GC with Hartree-Fock (HF) and MP-2. Of the six possible tautomeric sites only G^-C^+ and G^*C^* were stable. Furthermore, from Transition State Theory (TST) several requirements for biological relevance were outlined:

1. the reaction asymmetry (ΔE_{rxn}) between canonical GC and G^*C^* must be greater than 0.56 eV;
2. the canonical lifetime must be less than the replication period (1×10^{-8}s);
3. the tautomeric lifetime should be less than the timescale of strand separation (1×10^{-10}s);
4. the forward barrier to tautomerisation (ΔE_{fwd}) should be less than 1.21 eV;
5. the reverse barrier to tautomerisation (ΔE_{rev}) should exceed 0.13 eV.

Florian, *et al.* also predicted two mechanisms for the DPT leading to G^*C^* : a concerted process where both protons transfer synchronously and a stepwise process via a zwitterionic intermediate G^-C^+. Furthermore, MP-2 calculations revealed that the concerted transfer is preferred in the gas-phase dimer with reaction energetics satisfying all requirements except reaction asymmetry, which was approximately 0.4 eV. Reaction kinetics can be modelled with Transition State Theory (TST)[94].

Villani, *et al.* continued these efforts into DPT in gas-phase GC with a DFT study using B3LYP/cc-pvdz level of theory, but predicted lower barriers to tautomerisation (ΔE_{fwd}= 0.55 eV and ΔE_{rev}= 0.24 eV), reporting cyclic dynamical behaviour with a time period of 0.8ps (0.8×10^{-12}) (smaller than strand separation)[95]. Averaged over these periodic dynamics G^*C^* and G^-C^+ were populated 0.05-0.15% of the time, each, with a more significant population for G^-C^+ 0.15%.

Newer literature (circa 2006 onwards) has suggested that the asynchronous transfer via G^-C^+ is preferred[95, 96, 97, 98]. This may be due to the recent trend to improve models of the environment within which the GC is modelled. Gonzalez, *et al.* [99] and Hayashi, *et al.* [100] found that higher polarisability in continuum fields resulted in the DPT becoming stepwise.

Brovarets, *et al.* have provided alternative requirements for biological relevance in [74]:

1. G^*C^* must be dynamically stable and $\Delta G_{\text{rev}} > 0$;
2. the equilibrium population of G^*C^* should be between 1×10^{-8} and 1×10^{-11};
3. Electronic interaction of G^*C^* must not exceed dissociation energies of canonical GC;

4. The lifetime of G*C* must exceed strand separation (1×10^{-9}s) [1];

5. Individual ssDNA tautomers should have lifetimes exceeding 1×10^{-4} to allow them to reach the polymerase;

6. equilibration time for G*C* must be less than milliseconds (time between helicase's ATP stepping motor action).

Brovarets, *et al.* used B3LYP/6-311++G** and MP2 with dielectric $\epsilon = 4$ to model the hydrophobic DNA-protein environment of the replisome[74]. They state that G*C* will not lead to mismatches in the transition from vacuum to $\epsilon = 4$ due to the short lifetime, and lack of free energy barrier for the decay of G*C* to GC ($\Delta G_{\text{rev}} < 0$)[74].

Single base tautomerism is reduced by solvent interactions (by approximately one order of magnitude), nevertheless, the equilibrium tautomer population was found to exceed the rate of observed spontaneous mutations (1×10^{-1} to 1×10^{-4})[101, 102, 103]. The population of G*C* was found to be much smaller than the ssDNA tautomers[104]. The introduction of explicit water molecules (microhydration) in DFT models of the proton transfer in GC stabilised the tautomer[96, 97]. Additionally, the inclusion of stacking effects altered the DPT suggesting the need for more realistic environmental models[98].

Quantum Mechanics/Molecular Mechanics (QM/MM) can improve the environmental models for DFT via various embedding schemes. Using QM/MM, several works suggest higher transition state energy than isolated G*C* for the concerted transfer and observe a contraction of the base pair during the transfer, which is resisted by the backbone of the DNA duplex[105, 106]. For the stepwise process, ONIOM (see Section 4.4.1) QM/MM of a microhydrated codon revealed a barrierless DPT, but solvent effects stabilised G*C* [96]. Nevertheless, the tautomer is regarded as a secondary effect due to the low predicted thermal equilibrium population of 1.0×10^{-11}. Recent QM/MM approaches including a base pair and water molecules suggest that the secondary structure of B-DNA destabilises the G*C* tautomer, as the reverse barrier is decreased due to the higher asymmetry from free energy corrections relative to the gas-phase[107]. Roßbach, *et al.* rigorously benchmarked the convergence of QM/MM parameters finding that electrostatic embedding provided the most reliable convergence of single point energies, and that a very large quantum mechanical region is favourable[108]. Due to a discrepancy in their optimisation scheme (the system was never optimised in QM/MM only at the MM level of theory), this may be an overestimate[109]. Later investigation of the single proton transfer in QM/MM GC reveals that conformational variations account for larger changes between replicas than QM region size[110]. The level of theory can affect the structure and energy of DNA models. B3LYP is desirable where possible for its comparable results to MP-2[111, 112, 113, 114].

[1] N.B. this is an order of magnitude slower than suggested by [72], our work will show that both of these values greatly underestimate the speed of relevant atomic motions (see Chapters 7 and 8). Additionally, the more current understanding of GC tautomerism suggests that equilibrium tautomer populations, and not lifetimes, are most important in determining biological relevance.

Several relevant articles have appeared during the preliminary stages of this PhD research from Gheorghiu, *et al.* in [115, 116] and Slocombe, *et al.* in [103, 117]. Gheorghiu applied a QM/MM model of aqueous duplex DNA and observed a statistical ensemble of tautomerism pathways including synchronous and asynchronous DPT. For GC, Gheorghiu reported a stepwise process to occur in 84% of replica simulations, and reported that the small reverse barrier (< 0.1 eV) vanished when considering Gibbs free energies. Gheorghiu, *et al.* claim that previous purely quantum mechanical models have "oversimplified" the proton transfer processes and report an equilibrium population for G*C* of 1×10^{-9} and conclude that biological relevance is unlikely due to the short tautomeric lifetime (2.4×10^{-14}). In [116], Gheorghiu further investigated external electric fields on the secondary structure of DNA and GC tautomerism and found negligible effects on both, which is reassuring as electric fields are used in gene therapy. On the other hand, Slocombe *et al.* 's work in [103] reports increased stability for the G*C* tautomer ($\Delta E_{\text{rev}} = 0.266$ eV, $\tau_r = 5.5 \times 10^{-10}$). Crucially, Slocombe reports single base tautomers to be stable for hours, far exceeding the requirement outlined by Brovarets in [74]. Using an open quantum systems approach in [117] Slocombe, *et al.* determined that tautomerism in GC equilibrates extremely quickly (1.0×10^{-9}) and that thus a comparison of lifetimes is not applicable. Instead, tautomeric populations in chemical equilibrium must be considered. The open quantum systems treatment suggests that nuclear quantum effects play a considerable role in proton transfer and that quantum tunnelling dominates the transfer process. Path integral molecular dynamics and nuclear quantum effects are gathering more interest in the computational descriptions of DNA[118, 119].

There is a comparable body of literature on the tautomerism in AT, however, it is consistently reported that AT is a less favourable candidate than GC. [120, 121, 103] report negligible reverse barriers ($\Delta E_{\text{rev}} < 0.01$ eV). While some suggest that nuclear quantum effects and a water-like dielectric enhance the dynamical stability of A*T* (see [122, 100, 123]), others report that free energy and hydrophobic continuum fields are not sufficient to stabilise the tautomer[104, 99, 75, 124, 125, 73, 126, 127, 128, 129]. Additionally Gheorghiu, *et al.* do not observe any stable DPT products in QM/MM [115], and neither do Slocombe, *et al.* in [103].

Following the publication of [117] in early 2022, the state of tautomerism as a candidate for causing spontaneous point mutations can be summarised as follows: It is generally agreed that base pair tautomerism would only cause a very small number of spontaneous mutations in line with observed spontaneous mutation rates *in vivo*. Experimental evidence of the proton transfer in DNA is scarce, but NMR experiments have been able to measure proton transfer rates in DNA. Recent simulations have revealed higher equilibrium populations for GC (1×10^{-7} to 1×10^{-9}) vs AT (1×10^{-9} to 1×10^{-12}), which may explain the observed GC vs AT mutation bias. GC tautomerism is metastable with small reverse barriers leading to a fast equilibration and short lifetimes. The general trend is that an asynchronous DPT pathway is favourable. Water-assisted pathways are not energetically favourable as the thermodynamic probability of water-assisted pathways is three

orders of magnitude smaller than direct DPT. The aqueous environment and DNA structure are important for determining the DPT thus a variety of QM/MM replicas is crucial to model a realistic ensemble of conformations.

3. Biochemical Reactions

"In biology, life is the avoidance of equilibrium, and the attainment of equilibrium is death."

- Peter Atkins [130]

Chemistry is the study of change in matter. Physical chemistry underpins all biological life and governs the chaotic mess of a myriad of chemical reactions in dynamic equilibria contained within. In the simplest case, a chemical reaction describes transitions back and forth between two states - the reactant and the product.

In order to describe reactions - however delicate or violent - several useful quantities are used in physical chemistry. This section will focus on those most relevant to this project.

3.1. Reaction Quotients, Coordinates, and Rates

In physical chemistry, changes in product/reactant mixtures are mapped along a reaction quotient Q:

$$Q = \frac{\text{product activity}}{\text{reactant activity}} \approx \frac{\text{product molality}}{\text{reactant molality}} \approx \frac{\text{product concentration}}{\text{reactant concentration}} \tag{3.1}$$

To illustrate this, consider a reaction $A \rightarrow B$. In the pure reactant state the number of product molecules (n_B) is zero, and the reaction quotient is zero ($Q = n_B/n_A = 0$). Conversely in the pure product state the number of reactant molecules $n_A = 0$, the reaction quotient is infinite ($Q = n_B/n_A = \infty$). Mixed systems will have a finite and positive reaction quotient [131]. To describe change in these reactions a quantity ξ or "extent of reaction" is defined, for the case $A \rightarrow B$, $d\xi = -dn_A = +dn_B$. And for a general reaction:

$$\sum_r \nu_r X_r \rightleftharpoons \sum_p \nu_p X_p, \tag{3.2}$$

the change in a given reaction component can be defined in terms of ξ:

$$dn_i = d[X_i] = \pm \nu_i d\xi, \tag{3.3}$$

Where the the index i can be any of the reactant or product indices (r or p, respectively), and $[X_i]$ is the amount of reaction component/species X_i, with stoichiometric number ν_i.

In atomistic simulations, the number and atomic composition in the system is constant (except in the rarely employed Grand canonical ensemble[1]) [132]. Thus the difference between the two states is essentially a question of atomic position and bonding only. When studying reactions computationally, it is futile to attempt to keep track of every single position, bond length, and/or angle. As a result, reactions are described according to a handful of carefully chosen generalised reaction coordinates. See Chapter 5 for more details.

To introduce time dependence, reaction rate constants k are employed. For our model reaction ($A \rightarrow B$) the forward and reverse rate constants can be expressed as a change in state population:

$$k_{\text{forward}} = \frac{dn_B}{dt} = -\frac{dn_A}{dt} \tag{3.4}$$

$$k_{\text{reverse}} = -\frac{dn_B}{dt} = \frac{dn_A}{dt} \tag{3.5}$$

If the rate constants are known and given an initial configuration, the states populations can be calculated for any future time value. Microscopically, the rate constants depend on the quantum numbers of the involved states. Macroscopically, the rate constants are a function of temperature as the probability of finding a system in a certain state depends on its energy and the Boltzmann distribution [132].

3.2. Gibbs Energy

The Gibbs energy (G) is the "available energy" in a system [133]. Combining the internal energy (U), the work required to construct the physical dimensions (pV), and capacity for entropy (S) of the system gives the Gibbs energy:

$$G = U + pV - TS, \tag{3.6}$$

where p and V are the familiar pressure and volume units, respectively [133, 131]. The change in Gibbs energy (along the reaction) can also be related to the chemical potential (a measure of a system's potential to undergo change), μ [130],

$$\Delta_r G = \mu_P - \mu_R = \frac{dG}{dn}. \tag{3.7}$$

The change in Gibbs energy during a reaction, allows for its categorisation. For negative Gibbs energy change, the reaction is spontaneous (exergonic) ($\Delta_r G < 0$), requiring no external force/energy. In the case of a positive change in Gibbs energy the reaction must be driven by an external free energy input and is non-spontaneous (endergonic) [134, 135]. Additionally, in the case where

[1]The grand canonical ensemble fixes chemical potential, volume, and temperature, but not the number of particles. See Section 4.7

a forward reaction (reactant to product) is endergonic, its reverse reaction must be spontaneous. The result of this is that any chemical mixture will tend towards minimising the Gibbs energy [131]. Using the Gibbs energy as a thermodynamic potential, the equilibrium of any chemical system can be evaluated.

3.3. Chemical Equilibrium

The equilibrium state of any system is determined by minima in the Gibbs energy landscape plotted against the reaction coordinate(s) [133, 131]. At these minima the local derivative of the Gibbs energy will be zero ($\Delta_r G = 0$). See Figure 3-1.

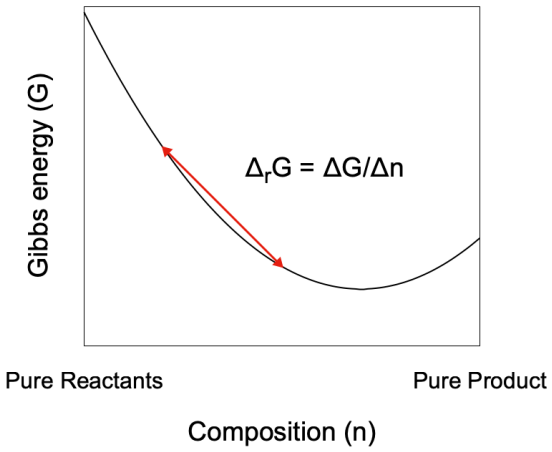


Figure 3-1 Change in Gibbs energy along a reaction. Given the free energy surface shown here, there will be more products in the equilibrium composition than reactants, this is due to the minimum being nearer the product state.

Given enough time to equilibrate, chemical reactions can generally be found in dynamic equilibrium where a mix of states is present. Often, significant concentrations of both products and reactants are present in the equilibrium composition. This means that any change from product to reactant will be compensated for by its reverse reaction. These net changes are encapsulated in the equilibrium constant $K = (Q)_{\text{equilibrium}}$ [131].

Like the reaction quotient Q, the equilibrium constant can be expressed in many ways. Generally it is the ratio of the population of the product over the population of the reactant. These populations can also be substituted with activities, partial pressures, concentrations - or most commonly in this work - with rate constants:

$$K = \frac{k_f}{k_r} \tag{3.8}$$

Systems in chemical equilibrium will do their best to oppose any changes from the environment. To compensate for an increase in pressure (due to an external force) the system will reduce the number of molecules. To oppose increases in temperature, a system in equilibrium will absorb

energy as heat, favouring the products in an endothermic reaction (and vice-versa in an exothermic reaction) [131].

The compartmentalisation of the individual molecules into energy levels is determined by the Boltzmann distribution, and is of course temperature dependent. From the Boltzmann distribution we assume that lower temperatures result in greater populations of less energetic energy levels, whereas higher temperatures populate energy levels more evenly. In the case where the energy levels are of similar density in both product and reactant state, endothermic reactions have equilibria favouring the reactant [132, 131]. See Figure 3-2.

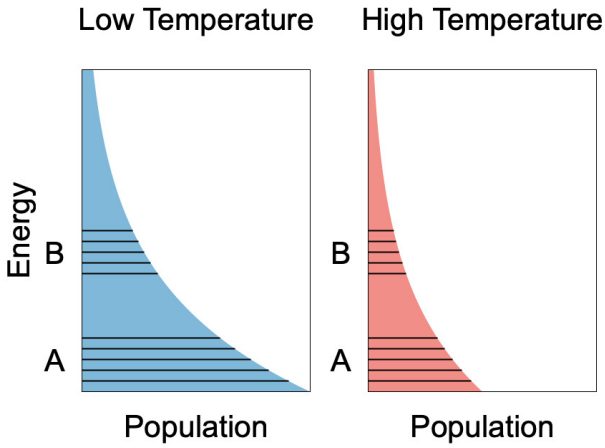


Figure 3-2 Boltzmann distributions of a system with sets of energy levels A and B corresponding to different reaction states with fixed density of states, at low and high temperatures.

If the density of states in the product is greater than that of the reactant's, it is possible for the equilibrium to favour the product as there are more states to populate [132, 131]. See Figure 3-3.

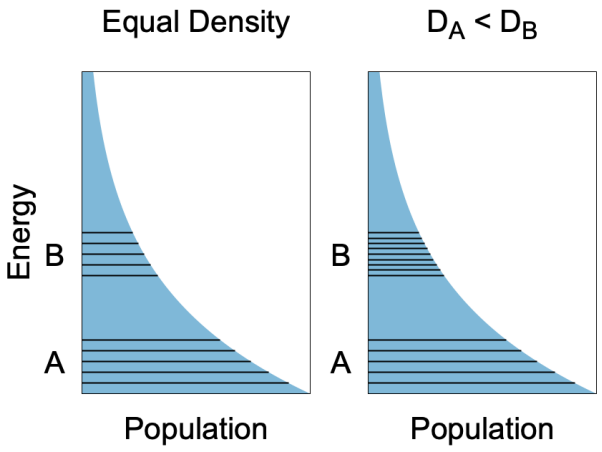


Figure 3-3 Boltzmann distributions of a system with sets of energy levels A and B corresponding to different reaction states at constant temperature, where the density of states of B is greater than that of A.

Of course many free energy surfaces also contain saddle points and maxima, in addition to the stable minima. The equilibrium composition is determined by the energy difference between the two minima, and not the path between them [136]. This is encapsulated in the relation between the

standard reaction Gibbs energy ($\Delta_r G^{\ominus}$, to the equilibrium constant:

$$\Delta_r G^{\ominus} = -RT \ln K, \tag{3.9}$$

where R is the gas constant [131]. In biological systems, the participation of hydrogen ions in the reaction $\mathrm{A} + \nu \mathrm{H}^+(\mathrm{aq}) \rightarrow P$, and assuming a pH of 7 [131], is accounted for via

$$\Delta_r G^{\oplus} = \Delta_r G^{\ominus} + 7\nu RT \ln 10, \tag{3.10}$$

where ν is the stoichiometric number of H^+ ions consumed in the reaction [131].

The barriers encountered along the reaction path contribute to the kinetic resistance to the equilibration of the system. Given enough time, any system will reach equilibrium, irrespective of the size of intermediate barriers. The rates of reactions between energetic minima, and across energy landscape saddles or ridges, are explained further by Transition State Theory (TST).

Catalysts, including enzymes, provide alternative routes from reactants to products [130]. The standard difference in Gibbs energy ($\Delta_r G$), remains unchanged and therefore the equilibrium composition is unaffected [130].

3.4. Transition State Theory

The Transition State (TS) is an intermediate stationary point between the reactant and product states. Along the reaction coordinate, the TS will be a maximum. See Figure 3-4.

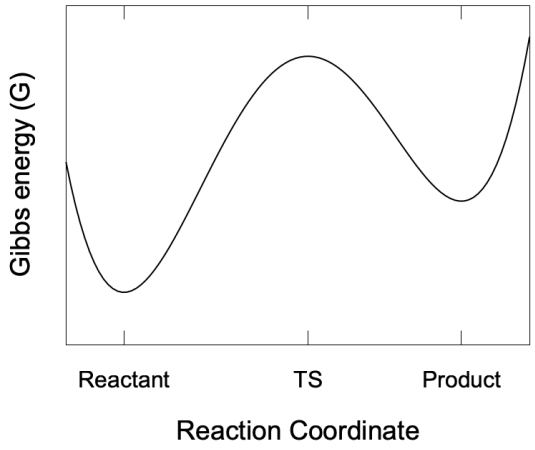


Figure 3-4 Change in Gibbs energy along a reaction. TS denotes the Transition State

Transition State Theory (TST) is a theory describing reactions across free energy surfaces. To formalise TST several assumptions are made [132]:

- An equilibrium distribution across all quantum states at all points along the reaction coordinate.

- The probability of state occupancy is Boltzmann-like ($e^{-\Delta E/kT}$).
- Every molecule that passes the TS will continue to the product state.

Given the above assumptions, TST prescribes that the rate of a reaction, k_{rate} can be obtained from the activation free energy, $\Delta G^{\ddagger}$, using Equation (3.11) [132, 137].

$$k_{rate} = \frac{k_b T}{h} \exp\left(-\frac{\Delta G^{\ddagger}}{RT}\right) \tag{3.11}$$

Where $\Delta G^{\ddagger}$ refers to the activation free energy, or the barrier height of the reaction. For a forward reaction the activation free energy is $\Delta G_{\text{TS}} - \Delta G_{\text{reactant}}$ [132]. The rate constant calculated with Equation (3.11) will always be an upper limit as the theory assumes that no molecule that has crossed the barrier will change direction and re-cross to the reactant state.

To compensate for re-crossing, a transmission constant κ can be introduced [132]. This coefficient can also describe quantum tunnelling contributions to the reaction rate. κ is usually around unity, with low (tunnelling dominated) temperatures giving $\kappa > 1$, and high temperatures leading to high rates of re-crossing $\kappa < 1$. When $\kappa = 1$ there is no compensation/presence of tunnelling or re-crossing.

The familiar equilibrium constant K can be expressed as a function of the free energy difference between the reactant and product states ΔG_0 [132],

$$K = \exp\left(-\frac{\Delta G_0}{RT}\right). \tag{3.12}$$

3.4.1. Comparison to Experimental Rate Constants

Transition State Theory (TST) prescribes that the rate of a reaction, k_r can be obtained from the activation free energy, $\Delta G^{\ddagger}$, using Equation (3.13) [138, 137],

$$\Delta G^{\ddagger} = -RT \ln \frac{k_r}{k_B T/h}. \tag{3.13}$$

Such free energies of activation can be obtained from Umbrella Sampling (see Section 5.3. When free energy is not available, an estimated activation entropy $\Delta S^{\ddagger}$ must be added to the Potential Energy Surface obtained for example from a converged NEB calculation.

3.5. Minimum Energy Paths

Two minima on a Gibbs energy surface can be joined by infinitely many paths. The path that is taken by the system will vary, depending not only on thermodynamic parameters but also on random chance. In most cases however, reaction kinetics can be extracted from a single path through the Gibbs energy landscape, provided this path minimises the energy barrier [139]. Given

this minimised Transition State, the path of steepest descent to both minima is known as the Minimum Energy Path (MEP) or path of least action. Many computational techniques exist to find and optimise the MEP for a reaction, and these will be discussed in Chapter 5.

In the low-temperature limit, the reaction will always proceed along the Minimum Energy Path, as energy is scant. In the high-temperature limit, the reaction will occur along the shortest path w.r.t. the internal coordinates, as energy is in abundance [132].

3.6. Tunnelling Corrections

Classical reaction rate theories have been remarkably successful. Protons, due to their low mass, not only exhibit quantum mechanical properties such as delocalisation and tunnelling, but in some cases their behaviour is dominated by QM[117]. This is especially important in proton transfer reactions where classical theories fail to accurately predict empirically observed rates[140, 125, 117]. Over the last century, many corrections have been proposed that incorporate QM effects such as the Zero-Point Energy (ZPE) and tunnelling. These corrections range in accuracy and computational/analytical expense, and additionally are often only accurate in a small subset of situations.

In this thesis potential energy profiles for proton transfer are presented, and properties such as the reaction free energy (asymmetry) and activation free energy (forward barrier) determine the classical equilibrium behaviour of the reaction. Transition State Theory (TST) can be used to obtain semi-classical reaction rate constants from free energy differences between reaction states and the transition surface between them. Several approximations to extend TST to include tunnelling have been developed. A reaction rate from TST k_{TST} can be adjusted to include tunnelling and re-crossing

$$k_{\mathrm{QM}} = \Gamma\kappa\, k_{\mathrm{TST}}, \tag{3.14}$$

where Γ is the recrossing factor taken to be unity in TST, and κ the tunnelling factor. Non-dissipative tunnelling corrections, which do not include an environmental coupling factor, provide values for κ taking additional properties of the reaction energy surface into account. The Wigner correction coefficient defines κ as,

$$\kappa = 1 + (\hbar\beta)^2 \frac{\omega_b^2}{24}, \tag{3.15}$$

where ω_b is the first imaginary mode of vibration at the Transition State (TS)[141, 142, 143]. More sophisticated tunnelling corrections also become more sensitive to the shape of the reaction barrier, especially at the TS. Theories such as the Bell correction using the Wentzel-Kramers-Brillouin (WKB) approximation depend strongly on the analytical form of the potential and provides a temperature-dependent κ:

$$\kappa(T) = \beta \exp(E_b \beta) \int_0^\infty \exp(-E\beta) P(E) dE, \tag{3.16}$$

based on the barrier height E_b, permeability of the barrier $P(E)$. The formulation of $P(E)$ varies greatly in complexity based on the shape of the barrier[144, 143]. Both the Wigner and Bell corrections are most accurate with symmetrical potentials and barrier energies not greatly exceeding the thermal energy[117].

Dissipative or "open" quantum treatments of reaction theory extend complexity and biological realism by coupling the system to a thermostat model of the local microenvironment. While out of the scope of this thesis, theoretical development of quantum descriptions of proton transfer are required to fully answer the prevalence of tunnelling in DNA tautomerism. An outlook into the possibilities of utilising a time-dependent Open Quantum Systems (OQS) theory to describe PT in DNA is discussed in Chapter 11.

4. Computational Chemistry

The descriptions of quantum chemical simulation methods utilised in this work are broadly separated into two categories. The first of these contains the techniques by which the energy and dynamics of molecular systems may be approximated with numerical methods, described in this chapter. Following this, we describe methods by which chemical reactions can be sampled and mapped in Chapter 5.

4.1. Modelling Molecules

Computational chemistry relies on models of atomic interactions, with varying degrees of complexity and approximation. These "levels of theory" determine not only the accuracy and realism of such investigations, but also their performance and computational cost. A thorough understanding facilitates a careful selection of the correct tool for the problem at hand, as limited resources result in a compromise between computational expense and theoretical accuracy. Empirical and semi-empirical computational chemistry theories can be tuned to approach experimental results, but where such results are not available it is prudent to use the highest available level of *ab initio* theory.

4.2. Force Field Methods

Molecular Mechanics (MM) is a molecular modelling method using classical mechanics to describe forces and energies. In essence, each atom is a point charge and mass, with fixed springs representing covalent bonding, and non-bonded interactions are modelled with one-dimensional potential functions.

An important concept in MM is the subdivision of large biological molecular systems into residues: small molecular subunits that are repeated in large molecules. These include for example the amino acids that make up proteins and the nucleic acids that form DNA. Through residues, Force Field (FF) methods are able to describe almost all proteins (to a certain degree of accuracy), without bespoke parameters for each one [145]. The Force Field is a database of parameters that determines the topology of each residue sub-unit, spring constants and equilibrium distances for all bonded interactions, and parameters for non-bonded interactions. Each residue template in the FF will

have descriptors for atomic species (atom types), partial charges, and bonding information rigidly defined. Each atom type gives a further distinction between atoms of the same atomic number (element), that includes parameters specific to their chemically bonded neighbours [132].

The total energy of the Force Field is given by a sum of interaction terms given in Equation (4.1) [132].

$$E_{\mathrm{FF}} = E_{\mathrm{stretch}} + E_{\mathrm{bend}} + E_{\mathrm{torsional}} + E_{\mathrm{vdW}} + E_{\mathrm{Coulomb}} + E_{\mathrm{cross}} \tag{4.1}$$

Generally these interactions can be categorised as bonded ($E_{\mathrm{stretch}} + E_{\mathrm{bend}} + E_{\mathrm{torsional}}$), non-bonded ($E_{\mathrm{vdW}} + E_{\mathrm{Coulomb}}$), and cross-term E_{cross} interactions. These interactions are illustrated in Figure 4-1. Additional terms may be added to Equation (4.1) to *steer* the simulation. This can take the form of additional potential energy contributions that bias the system away from equilibrium. Such restrained simulations are named Steered Molecular Dynamics (SMD) and form the basis of Umbrella Sampling (US) (see Section 5.3). Further details of the algebraic formulation of these interactions can be found in computational chemistry textbooks such as [132].

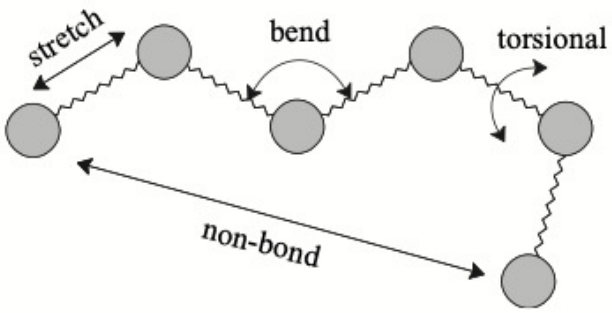


Figure 4-1 Illustration of the energy contributions with a Force Field method. Adapted from [132]

The parameters in a Force Field will have been validated against experiment and high-level QM theory calculations, but they will always be an approximation, as the molecular subunits will not always behave identically across different macromolecules. Where modelling of electronic effects or uncommon molecules is required, hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) methods can be employed which embed a quantum mechanical subregion of interest into a FF model.

4.3. Quantum Mechanical Methods

While Molecular Mechanics (MM) provides an expedient way of modelling molecular systems, it has the major drawback that electron dynamics are almost entirely neglected. As a century of modern physics has proven, the behaviour of electrons should be described by quantum mechanics. The classical 'rigid balls connected by springs' model of MM is inadequate on the atomic scale.

Quantum Mechanical (QM) modelling of chemical systems centres on approximating solutions to

the many-body Schrödinger Equation (SE):

$$\hat{H}\Psi = E\Psi. \tag{4.2}$$

Equation (4.2) is an eigenvalue equation of the many-body wave function Ψ for the molecular system. The Hamiltonian operator $\hat{H}$ and its associated total energy eigenvalue E, is the fundamental description of all the electronic, and nuclear interactions. Solving the SE for molecular systems poses considerable challenges. Many-body wave functions are not analytically solvable. To resolve these issues, approximations are used to simplify the complexity of calculations, and corrections are used to reach acceptable accuracy with respect to empirical and higher level *ab initio* results. The Hamiltonian describing the energy of a molecular system can be written as:

$$\hat{H} = -\hat{T}_e - \hat{T}_n + \hat{V}_{nn} + \hat{V}_{ee} - \hat{V}_{ne}, \tag{4.3}$$

where the energy operators take the following forms:

$$\text{Electronic kinetic energy } \hat{T}_e = \sum_i \frac{1}{2}\nabla_i^2 \tag{4.4a}$$

$$\text{Nuclear kinetic Energy } \hat{T}_n = \sum_A \frac{1}{2M_A}\nabla_A^2 \tag{4.4b}$$

$$\text{Electon-nucleus interaction potential } \hat{V}_{nn} = \sum_{i,A} \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} \tag{4.4c}$$

$$\text{Electron-electron interaction potential } \hat{V}_{ee} = \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{4.4d}$$

$$\text{Nucleus-nucleus interaction potential } \hat{V}_{ne} = \sum_{A<B} \frac{Z_A Z_B}{|\mathbf{R}_A - \mathbf{R}_B|} \tag{4.4e}$$

The indices i and j refer to electrons, while A and B correspond to nuclei. To determine the value each of the energy operators in Equation (4.4) for a given state, the energies are summed over all the relevant indices to account for all components of the many-body system. $\mathbf{r}$ and $\mathbf{R}$ refer to electronic and nuclear positions[1], while M is the nuclear mass, and Z is the nuclear charge. ∇^2 is the cartesian Laplacian operator, applied to the corresponding electronic or nuclear coordinates.

In order to solve the SE, the Born-Oppenheimer approximation is made such that the molecular wave-function can be fully separated into nuclear (Ψ_n) and electronic (Ψ_e) wave functions, each describing the nuclear and electronic many-body state:

$$\Psi(\mathbf{r}, \mathbf{R}) = \Psi_e(\mathbf{r}, \mathbf{R})\Psi_n(\mathbf{R}) \tag{4.5}$$

[1] relativistic effects are ignored, and thus the postions can be freely defined relative to a point of reference such as the centre of mass of the system.

This approximation is motivated by orders of magnitude mass difference between nuclei and electrons. Removing the nuclear degrees of freedom, we can write a solution to the many-body SE as:

$$\hat{H}_e\Psi_e(\mathbf{r},\mathbf{R}) = \left(\hat{T}_e + \hat{V}_{ee} + \hat{V}_{ne}\right)\Psi_e(\mathbf{r},\mathbf{R}) = E_e(\mathbf{R})\Psi_e(\mathbf{r},\mathbf{R}). \tag{4.6}$$

While the wave function $\Psi_e(\mathbf{r},\mathbf{R})$ does include electronic degrees of freedom, $E_e(\mathbf{R})$ is now a function of the nuclear coordinates only. This is motivated by the assumption that due to the mass difference between electrons and nuclei, the response time of the electrons to changes in nuclear coordinates is effectively immediate. This implies that for a given set of nuclear coordinates there exists only one ground state configuration of electrons, that is assumed to be instantly achieved following nuclear motion. Despite this simplification, solving for the ground state energy is not computationally trivial.

4.3.1. Density Functional Theory

Density Functional Theory (DFT) is an *ab initio* Quantum Mechanical (QM) model for molecular systems, based on the Hohenberg-Kohn theorem [132]. The Hohenberg-Kohn theorem postulates that the electron density ρ of a system and its ground state energy have a one-to-one correspondence [146] whose exact mathematical relation is not known. Crucially, for a set of nuclear coordinates, Hohenberg and Kohn postulate that there exists a single ground-state electron density ρ_0 that defines the ground-state energy of the total system,

$$E_0 = E[\rho_0(\mathbf{r})]. \tag{4.7}$$

The relationship between the electronic density and system energy is approximated via a functional (function of a function), in this case $E[\rho_0(\mathbf{r})]$. [146, 147]. The electronic density ρ can also be used to retrieve the number of electrons via an integral, and the position and heights of *cusps* within the density will describe the position and charge of the system's nuclei [132, 147]. Thus DFT reduces the many body problem of molecular energy to a single functional of electronic density. The first of which is the energy functional relating electron density to energy. The electronic energy functional comprises three parts,

$$E[\rho] = \hat{T}[\rho] + \hat{V}_{ee}[\rho] + \hat{V}_{ne}[\rho]. \tag{4.8}$$

The electron-nucleus term is trivially calculated, conversely the repulsion due to electron-electron interaction requires further analysis. Separating out classical Coulomb interactions $\hat{J}[\rho]$ and using a non-interacting reference system to describe the kinetic energy $\hat{T}_s[\rho]$, Kohn and Sham ([147]) showed that the $\hat{T}[\rho] + \hat{V}_{ee}[\rho]$ can be written as:

$$\hat{T}[\rho] + \hat{V}_{ee}[\rho] = \hat{T}_s[\rho] + \hat{J}[\rho] + \hat{E}_{\text{xc}}[\rho], \tag{4.9}$$

where the exchange and correlation interactions can be combined into a single unknown functional, the Exchange Correlation Functional (XCF) denoted by $\hat{E}_{\text{xc}}[\rho]$.

The use of an XCF simplifies the computation of electron-electron interactions $\hat{V}_{ee}$ by treating the electronic degrees of freedom separately. Each electron interacts only with the system's nuclei, with electron-electron interactions treated via the XCF. Exact mathematical formulations of the XCF are unknown, and there exist a wide array of approximate approaches with varying accuracy and computational expense.

Iterations on the general formulation of the Exchange Correlation Functional have resulted in several types of XCF's. Beginning with the Local Density Approximation (LDA) which assumes that the uniform electron gas (UEG) gives a comparable energy density to the system at hand, using Dirac's expression for a homogeneous electron. Generalised Gradient Approximation (GGA) functionals introduce the gradient of the UEG. Meta Generalised Gradient Approximation (MGGA) functionals introduce a second-order (Laplacian) term. Hybrid or Adiabatic Connection Method (ACM) functionals add the exact Hartree-Fock (HF) exchange energy. One such hybrid functional, deployed frequently in the work presented in this thesis takes the following form:

$$E_{\text{xc}}^{\text{B3LYP}} = (1-a)E_{\text{x}}^{\text{LDA}} + aE_{\text{x}}^{\text{HF}} + b\Delta E_{\text{x}}^{\text{B}} + (1-c)E_{\text{c}}^{\text{LDA}} + cE_{\text{c}}^{\text{LYP}}, \tag{4.10}$$

where $a = 0.1161$, $b = 0.9262$, $c = 0.8133$, and the subscripts B and LYP denote the Becke exchange and Lee-Yang-Parr correlation contributions from BLYP.

We have seen how density functional theory formulates the ground state properties of the molecular system based on the electron density, but describing this electron density poses further challenges in obtaining molecular orbitals. To this end, molecular orbitals are described as a linear combination of atomic orbitals, which, in turn, are described as a linear interpolation of basis functions describing primitive orbital functions which when combined approximate the true orbital. Orbital basis sets can take the form of Gaussian, Green, or plane-wave functions. Due to their efficiency contracted Gaussian functions are often chosen. The functions are contracted to truncate the exponential decay ($\exp(-r)$) of electron orbitals to reduce the space that must be integrated over. As a greater number of Gaussian functions are used in the basis set the closer the resulting molecular orbital approaches the exact answer. Basis sets are further extended with polarization functions which account for angular momentum beyond the valence orbits of the atoms (improving descriptions of chemical bonding), and diffuse functions that extend orbitals to account for extended delocalisation seen for example with hydrogen bonds.

One of the shortcomings of Density Functional Theory (DFT) is the inaccuracy of long-range non-covalent intermolecular van der Waal (vdW) interactions. To improve accuracy additional Disper-

sion Correcting Potentials (DCP) are introduced [148]. Dispersion corrections account for hydrogen bonding that governs the secondary structure of proteins[149], while diffuse basis sets capture some of these interactions, extending the basis sets radially from the nuclei adds considerable computational expense. Models such as Grimme's D3 dispersion model[150, 151, 152] improve the description of hydrogen bonding with limited additional computational cost by adding an energy term proportional to R^{-6}, where R is the inter-atomic distance.

4.4. Quantum Mechanics / Molecular Mechanics

Now that we have introduced methods to approximate the energy of atomic systems through Force Field (FF) methods, and seen how to model electronic structure via Density Functional Theory (DFT) methods, it is natural to consider their combination. The motivation is obvious: classical FF methods do not account for the motion of electrons, and hence cannot be used to describe bond breaking or bond formation [132]. DFT and other high-level electronic structure theories are computationally expensive and (without exceptionally large research grants) system size can rarely exceed dozens of atoms [132].

However, the implementation of this elusive hybrid of Quantum Mechanics (QM) and Molecular Mechanics (MM) or QM/MM is difficult. Most bespoke Molecular Dynamics (MD) packages do not support electronic structure calculations, and vice versa. Several implementations of interfacing codes exist, serving to mediate the communication between two separate executables.

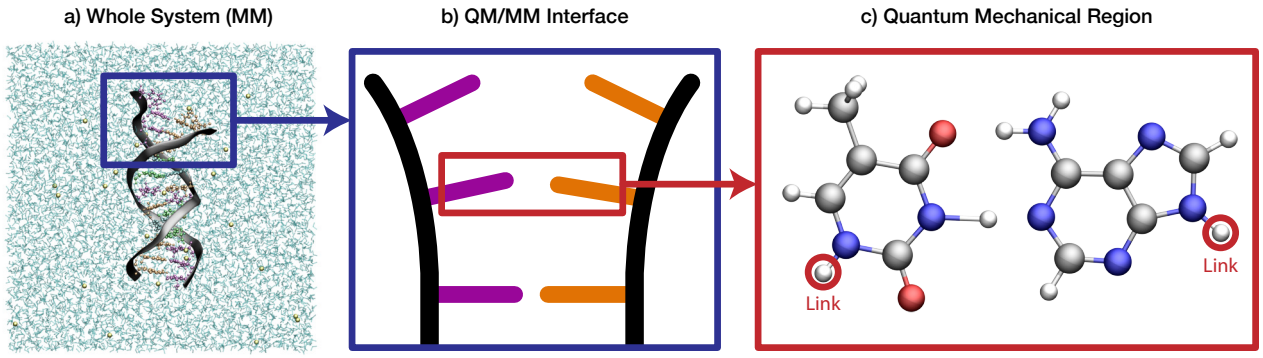


Figure 4-2 Multiscale Quantum Mechanics/Molecular Mechanics (QM/MM) schematic of an aqueous DNA double helix. a) shows the entire system including the DNA duplex represented by a cartoon ladder structure of the helical backbone, ball and stick representation of the DNA nucleobases (adenine in orange, thymine in purple, guanine in green, and cytosine in cyan), solvent molecules (teal liquorice), and counter ions (yellow spheres). b) shows a schematic of the interface between the two levels of theory. c) shows the atoms included in the quantum mechanical region. The encircled hydrogen atoms are link atoms which transfer forces to the DNA backbone.

QM/MM allows for the study of quantum effects in many-body systems, as long as they are within a small QM subregion. Take Figure 4-2 for example, the entire system might comprise a solvated DNA duplex, while the QM region is just a single GC pair. This is especially useful in considering the quantum effects in large biological molecules that would take a tremendous amount

of computational resources to study purely quantum mechanically [153]. The choice of QM region is important. When mapping reactions, the molecules involved in bond breaking/formation must be within the QM region [153]. Additionally the size of the QM region may have an effect on the energetics, with larger regions reducing error [108].

4.4.1. Additive versus Subtractive

The total energy of a hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) system can be expressed as a sum of contributions. Two main flavours of Hamiltonian are employed: additive and subtractive.

In the additive case:

- The energy of the QM region is evaluated by the QM Hamiltonian
- The rest of the system is evaluated by the MM Hamiltonian
- The interaction between the two regions is quantified via an embedding scheme. See Section 4.4.2

The total energy for an additive QM/MM method is thus given by:

$$E_{\text{total}} = E_{\text{QM}} + E_{\text{MM}} + E_{\text{QM/MM}}. \tag{4.11}$$

The subtractive case, also known as *our own n-layered integrated molecular orbital molecular mechanics* (ONIOM) does not explicitly treat the interaction between the two regions. The energy is given by:

$$E_{\text{ONIOM}} = E_{\text{QM}} + E_{\text{MM1}} - E_{\text{MM2}}, \tag{4.12}$$

where

- The energy of the QM region is evaluated by the QM calculator
- The entire system is evaluated by the MM calculator (E_{MM1})
- The energy of the QM region is evaluated by the MM calculator and subtracted (E_{MM2})

The subtractive case does not require an explicit embedding scheme to describe the interaction between the two regions. It does require a Force Field description of the QM region, which may lead to problems if the molecules in that region are difficult to parametrise accurately [132].

4.4.2. Embedding Schemes

Several methods of embedding a high-level QM theory within a QM/MM Hamiltonian exist. The simplest of which is mechanical embedding where the electronic descriptions (charges) do not interact

with each other. This leaves the steric interactions, which are modelled by typical FF non-bonded potentials such as Lennard-Jones [132].

The next level of complexity is introduced where MM partial charges can polarise the QM region. This embedding scheme is called electronic embedding.

An intermediate is to include charge-charge interactions between partial charges calculated from a population analysis[2] of the QM region. Furthermore, through polarizable embedding these charges can be allowed to modify the electric moments of the atoms in the MM region. This requires the use of a polarizable Force Field [132].

4.4.3. Link Atoms

Often the border of the QM region lies between two covalently bonded atoms, and thus must be treated with care. When the bond is cut, electrons are left unpaired in the QM region [132]. In such a case hydrogen link atoms or orbitals are introduced. These are not felt by the MM calculator.

Additionally rotational and torsional terms may be introduced across the QM/MM boundary, which will often need to be manually implemented if the QM region does not have a FF description.

4.4.4. Convergence

The choice of energy description, embedding schemes, link atoms, and QM region all have an effect on the total system energy. Therefore careful benchmarking is often required to validate the often unique multiscale approaches used by different groups. For the proton transfer in DNA chains a good reference for the convergence of QM/MM parameters is Rossbach et al [108] who found that increasing QM region size up to around 1000 atoms improves accuracy.

4.5. Finding stationary states

We have described how Force Field (FF) methods can describe the total energy of a molecular system, in a particular static state. We will see in due course the concept of chemical equilibrium (Section 3.3), whereby systems will tend to follow paths of steepest energetic descent towards local energy minima. To accurately perform reaction mapping, algorithms of geometry optimisation are employed. Optimisations are performed to find stationary points in the multi-dimensional energy surface. In the case where a minimum is desired, the optimisation is also known as Energy Minimisation (EM). All quantum chemical software packages will contain one or more optimisation algorithms. The simplest of which is Steepest Descent (SD), where the forces are computed and the positions are iterated to decrease the forces [132]. This is repeated over many iterations until the

[2] Density Functional Theory (DFT) codes produce electronic densities, which can be used to estimate partial atomic charges for a system using population analyses. These partial charges account for the fact that varying electronegativity causes polar bonds and (effectively) fractional charges for each atom. Electrons are more likely to be found near electronegative atoms.

force and/or energy difference is within certain tolerances. At this point, the optimisation is said to have converged.

4.6. Time integration

Now that we have introduced Energy Minimisation - the methods by which the local minima of a system can be found - we can proceed to the models of dynamics. Biological systems are in constant flux, and often far from equilibrium. To model this, time-dependency must be introduced. Molecular Dynamics (MD) is a method by which model systems are propagated through time. In order to introduce this time-dependence, Newton's laws of motion are discretised and iterated with small timesteps. The most common algorithm is the Verlet time integrator:

$$\mathbf{r}(t+\Delta t) = (2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \mathbf{a}(t)^2 + ..., \tag{4.13}$$

$$\mathbf{a} = \frac{\mathbf{F}}{m} = -\frac{1}{m}\frac{\partial \mathbf{V}}{\partial \mathbf{r}}. \tag{4.14}$$

$$\tag{4.15}$$

As with any discretised integral, the size of the infinitesimal integration step is of great importance to both the accuracy and precision of the result. In Molecular Dynamics, the time step chosen must be significantly (an order of magnitude) shorter than the fastest motion present in the system. Usually, the fastest molecular vibrations are of the order 10^{14} Hz, and as a result, time steps of 1 femtosecond (10^{-15} s) are used [132].[3] This is one of the limiting factors of MD, because even processes that take nanoseconds, will require millions of iterations to model. Considering that many biological processes take many microseconds or more, the dynamics that can be modelled with MD is limited. The entire DNA replication cycle for example takes thousands of seconds [74], and includes dozens of enzymes comprising hundreds of thousands of atoms each. Therefore one must be increasingly selective in the processes of study, especially as the level of theory is increased. Alternatively, reactions may be mapped with techniques discussed in Chapter 5.

4.7. Thermodynamic ensembles

Statistical mechanics forms the foundation upon which Transition State Theory and Molecular Dynamics is built. While it has not been explicitly described in this report, statistical mechanics has made an appearance through the Boltzmann distributions in Chapter 3. Supposing that the time integrator in MD conserves total energy E, then the system is a microcanonical ensemble. Such an ensemble is also known as NVE, after the quantities that are conserved/constant despite

[3] Larger time steps may be used in combination with a technique known as SHAKE, which fixes all bond vibrations involving hydrogen as these are often not relevant to the dynamics.

the dynamics. The number of particles, N, is fixed because most MD programs do not allow for particles to disappear and appear during the simulation. The volume V is perhaps a less obvious conserved quantity to those unfamiliar with MM, but it is practically always fixed.

One might then wonder what other thermodynamic properties are involved in Molecular Dynamics. Analysis of NVE Molecular Dynamics runs allows for the calculation of temperature (from the average kinetic energy) and pressure (from interactions with the simulation box's boundaries and the density). Both temperature and pressure will fluctuate in the microcanonical ensemble, and where this is not desirable, may be scaled by modifying the positions and velocities of the particles [132].

Simply scaling the velocities to achieve a desired temperature does not in fact lead to an NVT (canonical) ensemble, as the dynamics are affected. An improvement over this is the use of a coupled thermal bath:

$$dT/dt = \frac{1}{\tau}\left(T_0 - T\right), \tag{4.16}$$

which can be implemented by scaling the velocities by:

$$\sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_0}{T} - 1\right)}, \tag{4.17}$$

where T_0 is the target temperature, and is set to the value that the system has been initialised with (through a Boltzmann distribution of velocities).

However, even this more sophisticated method of coupling to a thermal bath is only an approximation. In most cases, a coupled thermal bath is acceptable as the average behaviour is correct, despite the incorrect fluctuations (with a time constant τ). Nosé-Hoover coupling provides a true canonical ensemble through a bath that is integral to the system and evolves together with other variables.

All of the above descriptions for temperature coupling are true for pressure as well, allowing for simulations of isobaric (NpT) nature as well.

5. Mapping Reactions

Chapter 4 introduced computational chemistry methods that can be used to approximate the internal energies of molecular systems and their dynamics. In order to investigate chemical reaction kinetics, the reaction must first be mapped (sampled). The equilibrium state of a chemical system depends only on the reaction free energy:

$$\Delta G_{\text{reaction}} = \Delta G_{\text{product}} - \Delta G_{\text{reactant}} \tag{5.1}$$

For simple systems, this can be obtained by performing a geometry optimisation (Energy Minimisation (EM)) of the reactant and product states separately and applying free energy corrections. To calculate reaction kinetics, additionally the maximum of the Minimum Energy Path (MEP) connecting the two states (i.e. the Transition State (TS)) must be determined.[1] This can be achieved via an optimisation scheme or a reaction mapping method. This chapter details three common reaction mapping techniques: Adiabatic Mapping (AM), Nudged Elastic Band (NEB), and Umbrella Sampling (US).

In order to describe the reaction path, a set of Reaction Coordinates (RCs) must be chosen. Reaction coordinates can vary in formulation but are always a set of degrees of freedom that allow for continuous interpolation between reaction states. I.e. for a Proton Transfer (PT) a suitable pair of Reaction Coordinates may be the distance of the proton to its donor atom, and to its acceptor.

5.1. Adiabatic Mapping

Adiabatic Mapping (AM) is a scheme by which the Potential Energy Surface (PES) along a known reaction path can be estimated. The system is initiated in a known Transition State (TS), and then through a series of restrained minimisations the potential energy profile is obtained. Figure 5-1 illustrates a three-step mapping from the Transition State to the product state.

Adiabatic Mapping provides an inexpensive way to quickly estimate the PES, as only a series of minimisations is required. However, it is very sensitive to the details of the transition state, and especially in large systems it may not be accurate even when averaged over a large number of initial structures [132]. For example, if one is interested in the breaking and forming of a specific bond

[1] While in a one-dimensional reaction profile, the TS is a maximum it represents a "saddle-point" in multi-dimensional energy surface, where the point is a minimum along some dimensions, and a maximum in others.

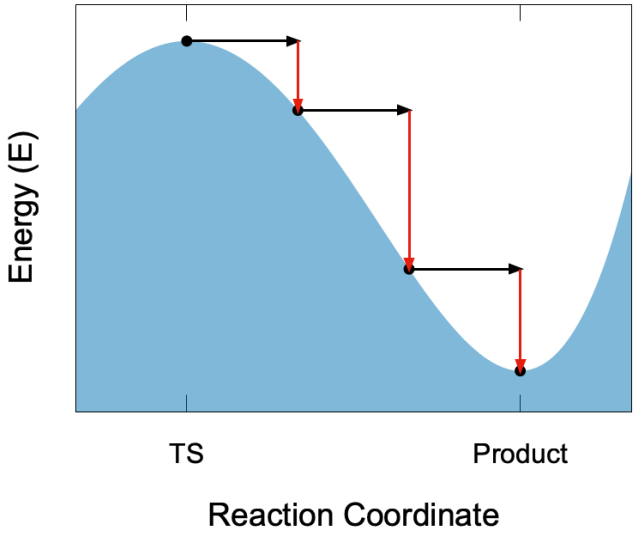


Figure 5-1 Illustration of Adiabatic Mapping (AM) with very large Reaction Coordinate (RC) steps. The system is started in the Transition State (TS) state, and the Reaction Coordinate is changed either instantaneously or via a strong restraining force (black arrows). Then the system is minimised while keeping the RC constant (red arrows). This process is repeated to map the entire Potential Energy Surface.

one might set an RC to be:

$$\text{RC} = l_{AB} - l_{BC}, \tag{5.2}$$

where if the RC increases, the length of the bond joining A to B (l_{AB}) increases, while the length of the bond joining B to C (l_{BC}) decreases. The difficulty is that AM does not discriminate between motions not involved in the calculation of the RC. Suppose this reaction happens in a small ligand surrounded by a large enzyme, if an Adiabatic Mapping method is used, there will be infinitely many conformations of the protein that still result in the TS value of the RC [137]. These can be averaged out by statistical repetitions of AM. This does not remove the fundamental issue that the enzyme's varying conformations affect the energy, and thus make it very difficult to completely minimise to extract the Potential Energy Surface of the reaction of interest [132]. For these reasons, when a reaction must be sampled in a noisy/dynamic environment, the method of Umbrella Sampling (US) is preferred. In US here the sampled conformations can be statistically deconvoluted to clean up the reaction profile.

5.2. Nudged Elastic Band

The Nudged Elastic Band (NEB) method presents another reaction mapping algorithm, that connects a set of images (system replicas) or beads between two endpoints in conformational space [154, 132]. Instead of mapping along a fixed reaction path such as Adiabatic Mapping (AM) and Umbrella Sampling (US), NEB will optimise the reaction path to be the path of least action [155].

This is achieved by defining a target function consisting of the sum of energies of all the images and a penalty to maintain an even distribution of images along the path [155]. This penalty is parametrised as a spring constant k, giving rise to the method's name. The magnitude of the spring constant can affect the resulting reaction path, overly large spring constants can result in the MEP not passing through local minima (corner cutting), weak spring constants result in paths that undersample near higher energy states [132]. The series of images is 'nudged' by taking only the component of the forces acting perpendicular to the path [154]. An example of a NEB path through a conformational landscape is provided in Figure 5-2.

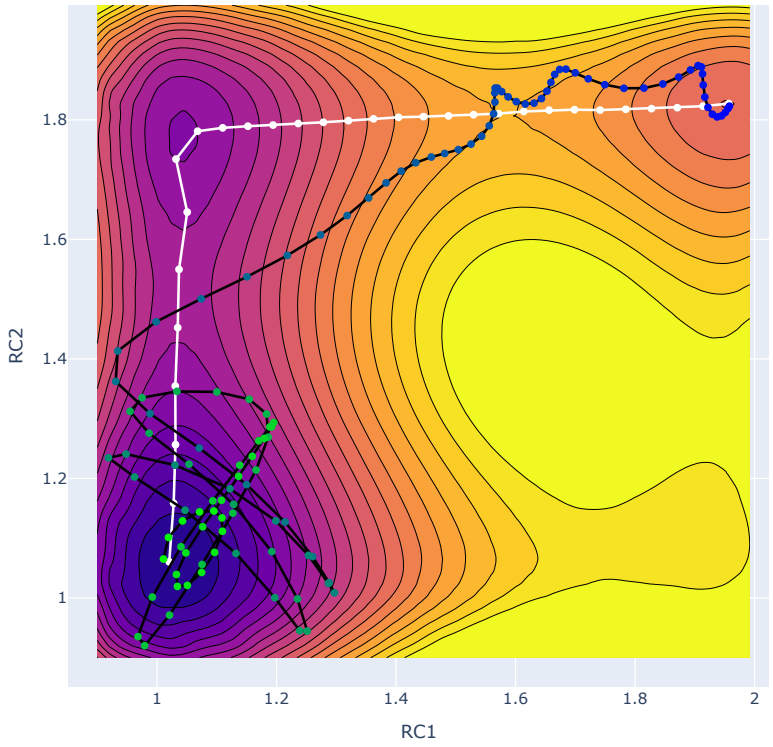


Figure 5-2 An example of an optimised NEB path (in white) through a potential energy landscape defined by two reaction coordinates (heat map with cold colours representing low energy, and cold representing higher energy). Here the path of least action through a two-dimensional energy landscape has been obtained. Additionally, the stochastic decay path from the top-right local minimum to the global minimum (bottom-left) is shown with black lines. Taken from Chapter 9.

The NEB method has been used extensively for the study of proton transfers, including those in DNA hydrogen bonds [103, 156]. While popular, the NEB method can prove difficult to converge, leading to expensive QM calculations, and it does not sample the free energy. As a result, normal mode/vibrational calculations are required to obtain the imaginary frequency of the TS for tunnelling corrections (see Section 3.6). To improve computational efficiency, a machine-learned approach to NEB "ML-NEB"[157, 158] has been implemented and used in [103] and Chapters 7, 8 and 10

In addition to this optimisation method, there exists a dynamical approach to populating the images through simulated annealing [155, 156]. This is less popular with QM calculations due

to the high cost of evaluating Single-Point Energies, but serves as a useful tool with Molecular Mechanics. With simulated annealing, the images are started at both reaction endpoints, and gradually thermally distributed to a high temperature(> 400 K) and then gradually cooled to settle the images. Figure 5-3 shows such an annealing profile, and Figure 5-4 shows the progression of a barrier as a result of the annealing.

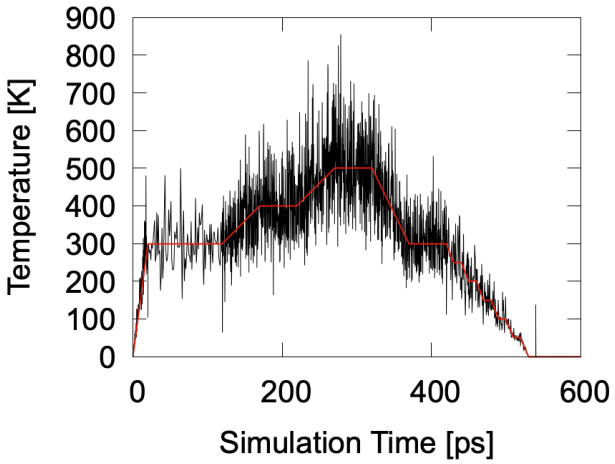


Figure 5-3 Temperature profile for simulated annealing.

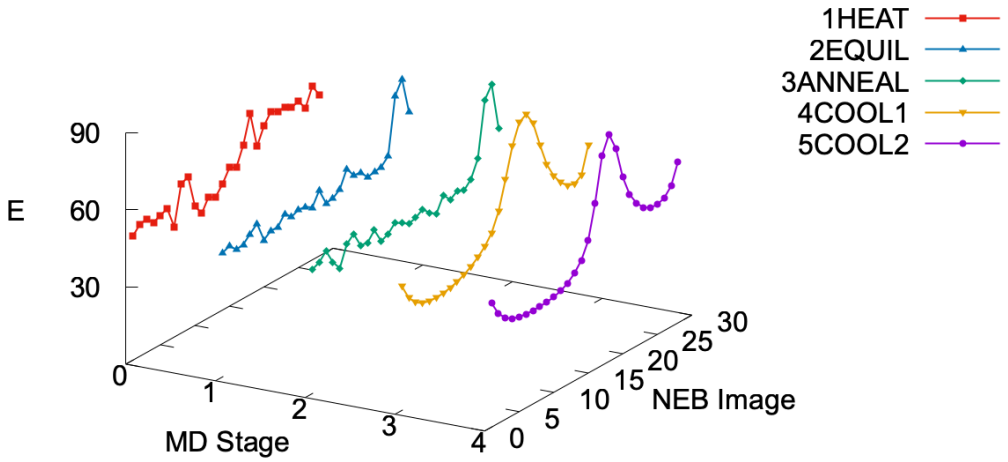


Figure 5-4 Progression of NEB PES during simulation annealing. We can see that the estimated path of least action obtained by NEB improves as images are thermally populated and then cooled to absolute zero.

5.3. Umbrella Sampling

Similarly to Adiabatic Mapping discussed previously, Umbrella Sampling (US) steps along a determined reaction path with a series of biasing restraint potentials [159, 132]. The biased/modified potential function of the Umbrella Sampling is shown in Equation (5.3). As with Adiabatic Map-

ping knowledge of at least the Transition State, and ideally more of the reaction path, is required [137]. This path is split into a series of 'windows' at which Steered Molecular Dynamics (SMD) is performed. This allows for the system to explore the local conformational space, and the free energy can be determined (also referred to as Potential of Mean Force (PMF)) [132].

$$V'(\mathbf{r}) = V(\mathbf{r}) + k_U(\mathbf{r} - \mathbf{r_0})^2 \tag{5.3}$$

Sequentially stepping along the Reaction Coordinates will provide statistical distributions of the system along the Reaction Coordinate [160, 137]. These will be difficult to interpret at first, but using the Weighted Histogram Analysis Method (WHAM) will reveal the energetics of the reaction [137, 161, 162]. Figure 5-5 shows exemplar sampling statistics for three Umbrella Sampling windows. The harmonic restraint potentials restrict the system to a certain Reaction Coordinate range. However, the actual distribution will always deviate somewhat. This can be interpreted by WHAM to determine the relative free energies of the windows.

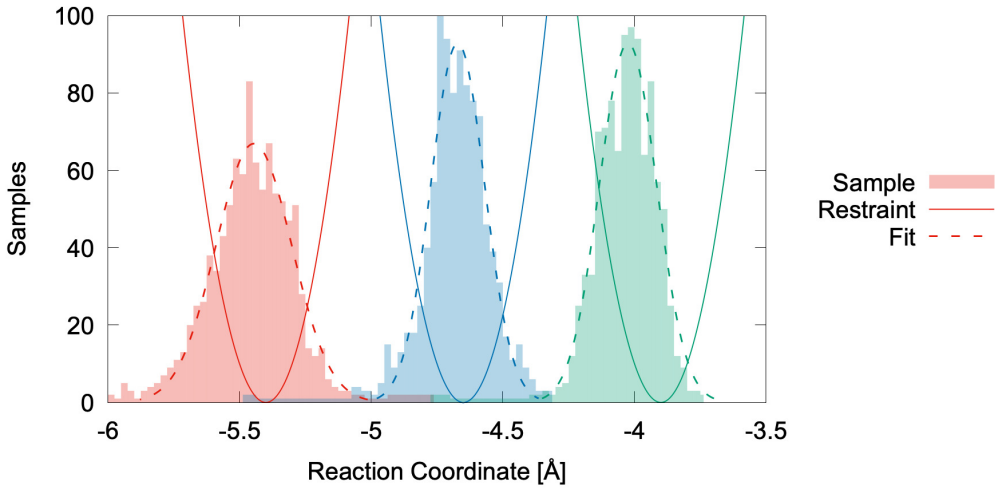


Figure 5-5 Example of umbrella sampling Reaction Coordinate statistics.

As with any Steered Molecular Dynamics, the ensemble distribution will differ from the standard Boltzmann one [159, 132]. The expression in Equation (5.4) can be used to unbias the results of Umbrella Sampling simulations.

$$\langle A \rangle = \frac{\langle A(\mathbf{r}) e^{U(\mathbf{r})/kT} \rangle_{V'}}{\langle e^{U(\mathbf{r})/kT} \rangle_{V'}}, \tag{5.4}$$

where the biased distribution, averaged over the modified potential (denoted by $\langle\rangle_{V'}$), is deconvoluted to give the unbiased average distribution $\langle A(\mathbf{r}) \rangle$ [132, 161].

Because Umbrella Sampling samples a small region around the reaction path, the Potential of Mean Force obtained may deviate slightly from the sampled path, suggesting that it is more energetically favourable than the prescribed path. Adaptive Umbrella Sampling makes use of this path as the initial condition for a new US run. Repeating this should lead to convergence with the true

PMF [132].

Umbrella Sampling can be computationally expensive when performed at high-level QM/MM due to the large number of SPEs and gradient calculations required (10^4 or more) [137]. Umbrella Sampling (US) is a mainstay of free energy sampling in computational chemistry, and has been applied extensively to enzyme catalysed reactions as well as proton tunnelling in DNA[107]. The proposal of this work is to combine these two to determine the effects of replisome enzymes on the rate of proton tunnelling in DNA.

6. Hydrogen bonding and the stability of DNA

This chapter includes preliminary investigations into hydrogen bonding and the stability of DNA. Several levels of theory, with increasing computational expense, are benchmarked and used to characterise the hydrogn bonds in DNA. Attempts are made to extend Molecular Mechanics (MM) models to non-canonical tautomers in DNA. Following characterisation of the DNA base pairing via hydrogen bonds, the ensemble stability of aqueous DNA and DNA in complex with PcrA helicase will be investigated. PcrA Helicase has been well-studied to reveal its the kinetics and dynamics of its stepping-motor mechanism, and the key amino acid residues required for the translocation of DNA (see Chapter 2). Finally, preliminary attempts at mapping intrabase Proton Transfer (PT) and inter-base Double Proton Transfer (DPT) in Quantum Mechanics/Molecular Mechanics (QM/MM) models of aqueous DNA and the PcrA Helicase-DNA complex are reported. These initial results show the shortcomings of Semi-Empirical Quantum Mechanics (QM) models and highlight the challenge of obtaining sufficient sampling for Umbrella Sampling (US) at high levels of theory.

6.1. Hydrogen Bonds in Nucleobase Dimers

This section includes the preliminary investigations into the accuracy and limits of Molecular Mechanics (MM) in describing the hydrogen-bonds in DNA.

6.1.1. Optimised Nucleobases and Dimers

The equilibrium gas phase and solvated structures of the canonical and tautomeric DNA nucleobase dimers were obtained through geometry optimisation through both *ab initio* QM and Force Field MM methods, respectively. Additionally, the isolated nucleobases structures were optimised.

By comparing the equilibrium energy of the dimer to the sum of the two constituent nucleobases the binding energy can be estimated. See Table 6-1.

$$\begin{aligned} D_{e,\mathrm{AT}} &= E_{\mathrm{AT}} - E_{\mathrm{A}} - E_{\mathrm{T}} \\ D_{e,\mathrm{GC}} &= E_{\mathrm{GC}} - E_{\mathrm{G}} - E_{\mathrm{C}} \end{aligned} \tag{6.1}$$

6.1.2. Parametrising the Tautomers

In order to use the tautomeric nucleobase structures with Force Field methods, the latter must be parametrised. The tautomeric forms of Adenine and Thymine have been detailed in literature [163, 164, 165]. This subsection briefly details the internal parametrisation attempts of the tautomeric Adenine-Thymine dimer.

AmberTools provides the **`antechamber`** program which uses semi-empirical QM methods to parametrise the bonds in small molecules. **`antechamber`** will assign partial charges and atomtypes within the General Amber Force Field (GAFF) that best reproduce the QM results. These methods were benchmarked against DFT and semi-empirical QM methods.

Figure 6-1 compares the charges obtained for the Adenine tautomer from **`antechamber`** to population analyses of DFT minimisations. Another informative comparison is that of the minimised structure A*T* using GAFF, DFT, and semi-empirical PM3. These structures are compared in Figure 6-2.

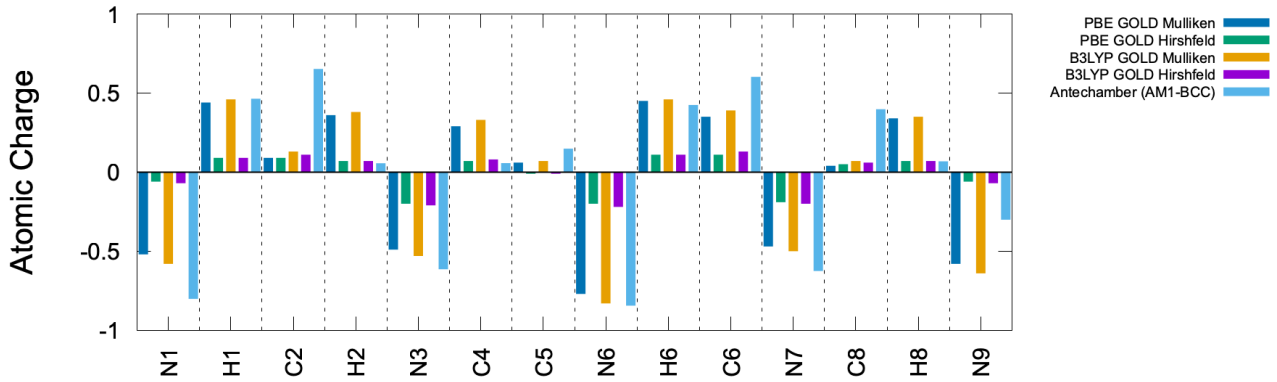


Figure 6-1 Adenine tautomer charges from CASTEP (DFT) and Amber's **`antechamber`**. We can see that **`antechamber`** reproduces charges similar to CASTEP's Mulliken populations, with the improvement of much less polar C-H bonds.

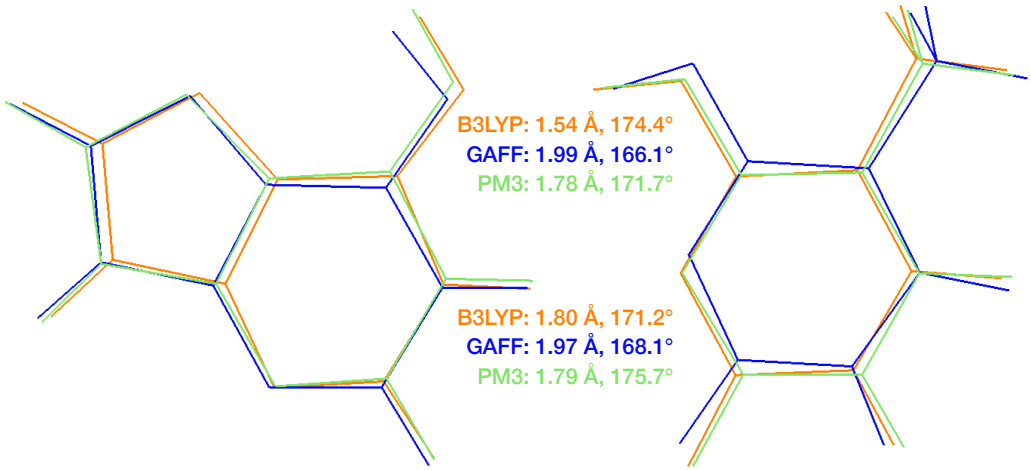


Figure 6-2 Minimised structures of the Adenine-Thymine tautomeric dimer using CASTEP & B3LYP (orange), Amber's GAFF with **`antechamber`** parameters (blue), and Amber's semi-empirical PM3 (green). For each level of theory the hydrogen-acceptor distance and donor-hydrogen-acceptor angle is given.

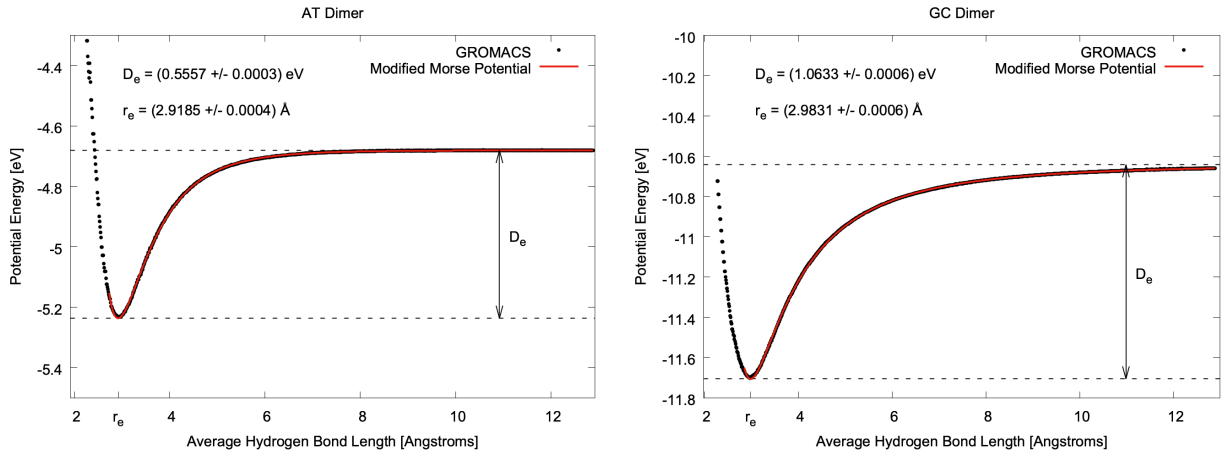
We can see that while **antechamber** provides acceptable charges and bond parameters, even a semi-empirical QM method produces results more in line with high-level DFT theory. Where possible, it appears favourable to use an embedded semi-empirical QM region or better to model custom residues of interest.

6.1.3. Characterising the Hydrogen Bonds

The hydrogen bond strength of the canonical DNA dimers was obtained through an Adiabatic Mapping (AM) process (see Section 5.1) with Force Field MM and DFT QM methods. This can be compared to the subtractive method described in Section 6.1.1, as well as values from the literature (see Table 6-1).

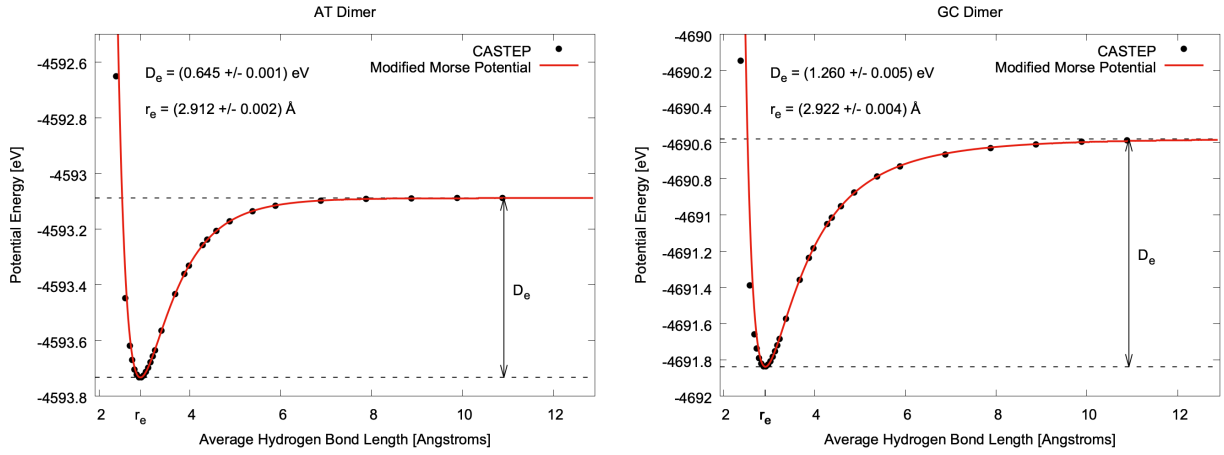
The potential energy of DNA dimers at a series of fixed separations was relaxed and their potential energy was calculated, as in an Adiabatic Mapping procedure. During relaxation position restraints were applied to keep the system planar and the hydrogen bonds at 180 degrees. By plotting this potential energy against the average donor-acceptor distance, the potential function of the hydrogen bonds can be determined. By fitting a modified Morse potential [166], as shown in (Equation (6.2)), the dissociation energy, equilibrium separation, and energy of the monomeric system can be obtained. See Figure 6-3.

$$\text{Morse}(x) = D_e \exp\left(2\rho\frac{1-x^\phi}{\phi}\right) - 2\exp\left(\rho\frac{1-x^\phi}{\phi}\right) - A. \tag{6.2}$$



(a) Adenine-Thymine

(b) Guanine-Cytosine



(c) Adenine-Thymine

(d) Guanine-Cytosine

Figure 6-3 DNA nucleobase dimer hydrogen bond potential curves obtained using Gromacs (Figures 6-3a and 6-3b) and CASTEP PBE (Figures 6-3c and 6-3d) Adiabatic Mapping.

Table 6-1 provides a summary of the results found in this work, with a comparison to experimental and computational literature values.

Table 6-1 Comparison of equilibrium donor-acceptor distance (r_e) and dissociation energies (D_e) for hydrogen bonds in DNA. D_e values obtained from modified morse potential fit of Adiabatic Mapping (AM) potentials and energy subtraction from Energy Minimisation (EM) ($D_e = E_{\text{dimer}} - E_{\text{base1}} - E_{\text{base2}}$). r_e is either the mean of the hydrogen bonds equilibrium lengths from EM or the minima of the modified morse potential from AM. Rows with values taken from the literature are denoted by a reference in their method column.

Dimer	Method	r_e [Å]	D_e [eV]
AT	MM AM	2.9185 ± 0.0004	0.5557 ± 0.0003
AT	DFT AM	2.912 ± 0.002	0.645 ± 0.001
AT	MM EM	2.88	0.549
AT	DFT EM	2.918	-
AT	DFT EM [167]	2.83	0.564 ($\Delta H = 0.512$)
AT	Mass Spectrometry [168]	-	$\Delta H = 0.525$
GC	MM AM	2.9831 ± 0.0006	1.0633 ± 0.0006
GC	DFT AM	2.922 ± 0.004	1.260 ± 0.005
GC	MM EM	2.87	1.001
GC	DFT EM	2.841	-
GC	DFT EM [167]	2.83	1.13 ($\Delta H = 1.03$)
GC	Mass Spectrometry [168]	-	$\Delta H = 0.911$

It is clear that both DFT methods and well-parametrised MM generally reproduce the correct hydrogen-bond lengths, even though the experimental values [169, 170] show a great deal of fluctuation due to thermal effects. As expected, the three hydrogen-bonds in Guanine-Cytosine produce a higher dissociation energy than Adenine-Thymine, both in our work and in other literature [170, 171, 172].

6.1.4. Non-Canonical Nucleobase Dimers

While the characterisation of the hydrogen bonds in the Watson-Crick (WC) canonical dimers compares favourable to expected values, repeating this process for the tautomers is difficult due to the quality of parametrisation. Our results in Figure 6-4 show that the dissociation energy is extremely sensitive to the chosen partial charge set.

Due to the sensitivity to parametrisation, the novel hydrogen bond potentials for the A*T* and A*C (seen in Figure 6-5) are not deemed sufficiently validated for publication. In the rest of this work, a multiscale Quantum Mechanics/Molecular Mechanics (QM/MM) approach is favoured over custom Molecular Mechanics (MM) parameters as this reduces the need for careful validation of the Force Field (FF) parameters.

6.2. Ensemble Stability of DNA & Helicase

The bacterial PcrA helicase (shown in Figure 6-6) has been a popular candidate for both experimental and computational investigations. Previous Molecular Dynamics (MD) of PcrA helicase has revealed the stepping motor action of the enzyme in translocating single-stranded DNA (ssDNA)

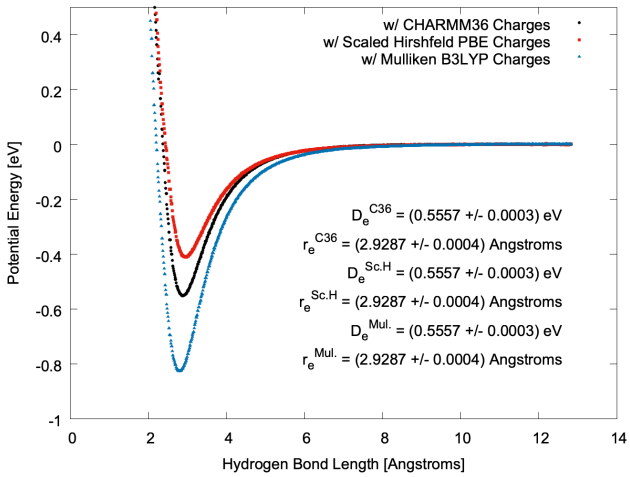
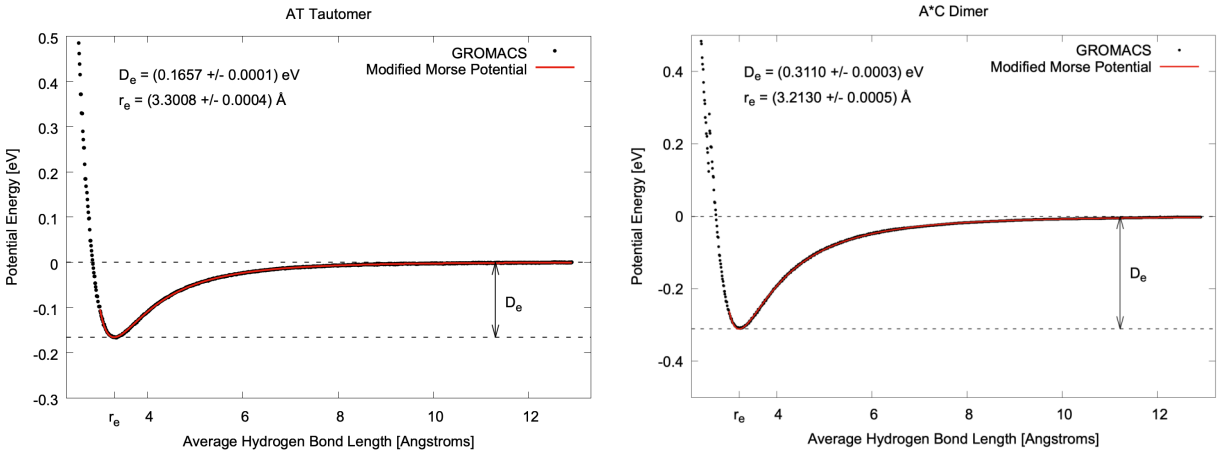


Figure 6-4 Adenine-Thymine Dimer hydrogen bond potential energy surface with charges from CASTEP population analysis instead of from the CHARMM36 Force Field.



(a) Adenine*-Thymine*

(b) Adenine*-Cytosine

Figure 6-5 Tautomeric DNA nucleobase dimer hydrogen bond potential curves obtained using Gromacs Adiabatic Mapping

[173, 18]. The PcrA helicase has also been extensively studied experimentally, providing insight into the translocation rate [16, 174, 17], and the roles of individual residues in the DNA binding site [20].

The ensemble stability of the DNA duplex and DNA-Helicase complex were determined through nanosecond duration Molecular Dynamics (MD) simulations. Of particular interest is the behaviour of the hydrogen bond-lengths (donor-acceptor distance) during the simulations, and whether the presence of helicase can be correlated to the breaking of hydrogen bonds, especially at the end of the duplex.

As shown in Figure 6-6, the DNA-Helicase complex includes two strands of DNA, one of which terminates after 14 Nucleic Acids (NA). Table 6-2 shows the sequence of the DNA in the complex. The NA base pairs in the duplex most likely to separate due to thermal effects are the first and fourteenth in the sequence. There may also be a difference in the behaviour of the thirteenth and fourteenth base pair due to the helicase enzyme's Amino Acids.

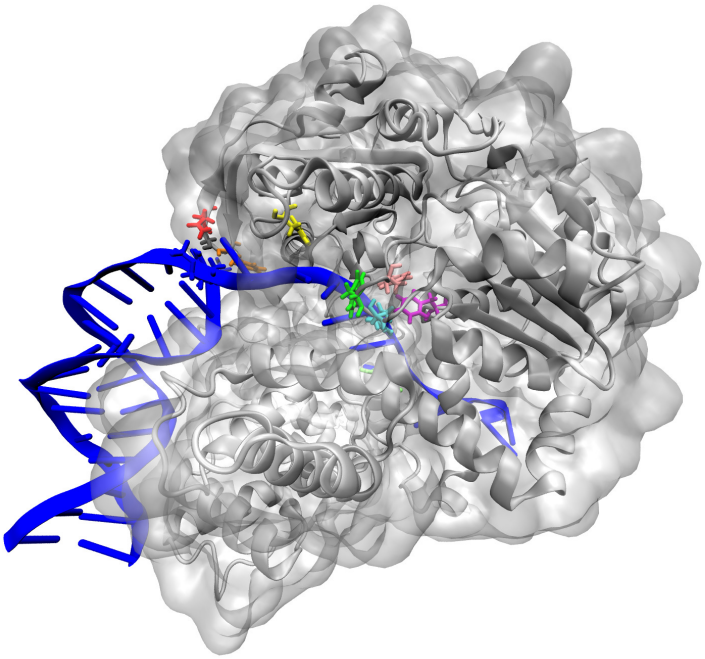


Figure 6-6 Visualisation of the PcrA Helicase-DNA product complex system. DNA (blue) and protein (grey) are shown with cartoon representations. For the protein, key amino acids for the translocation of ssDNA are coloured, and the vdW surface is shown in translucent grey.

Table 6-2 Sequence of the DNA in the helicase complex.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
C	A	G	T	G	C	A	G	T	G	C	T	C	G	T	T	T	T	T	T	T	T	T	T
G	T	C	A	C	G	T	C	A	C	G	A	G	C										

6.2.1. Preparation of the simulation system

The PDB entries, 3PJR[1] and 2PJR[2], contain the substrate and product states of the PcrA Helicase complexed with the same sequence of DNA. These structures originate from X-Ray Diffraction (XRD) documented in [15] and have relatively poor resolution of around 3Å. Due to this poor resolution, missing residues, and fragmented DNA, the unmodified structures are unsuitable for simulation. A previous MD study [18] of this system, details a rigorous procedure by which the structures were made comparable and generally cleaned up. These structures were generously provided by the principal author Dr Jin Yu.

In our work, the Nucleic Acids in the DNA duplex were shifted and rotated to conform to the regular B-DNA helical macrostructure. Minimisations and equilibration fixed the backbone rotation irregularities introduced by the manual adjustment of atomic positions. In all of the MD simulations, the system was solvated and neutralised through the addition of sodium cations. Additionally, the fourteenth base pair (terminal of the duplex) was nearly completely dissociated ('frayed') and a separate 'fixed' structure for both of the complex states was created.

[1] https://www.rcsb.org/structure/2PJR
[2] https://www.rcsb.org/structure/3PJR

6.2.2. Ensemble Molecular Dynamics

The two variants of the DNA-Helicase product complex, frayed (2CR) and fixed (2CX); as well as extracted DNA molecules, frayed (2DR) and fixed (2DX); were simulated in both NVT and NPT ensembles at 300K and 1 atm (where appropriate).

The Root Mean Square Displacement (RMSD) of the DNA-Helicase product complex - seen in Figure 6-7 - indicates that over the short (nanosecond) timescale, the Helicase reduces the dynamics of the DNA relative to its starting conformation, indicating that the enzyme stabilises the dynamics of the DNA. The stepping motor action of PcrA helicase is known to be on the order of 50 bases per second [16], so the nanosecond time scale is essentially stationary.

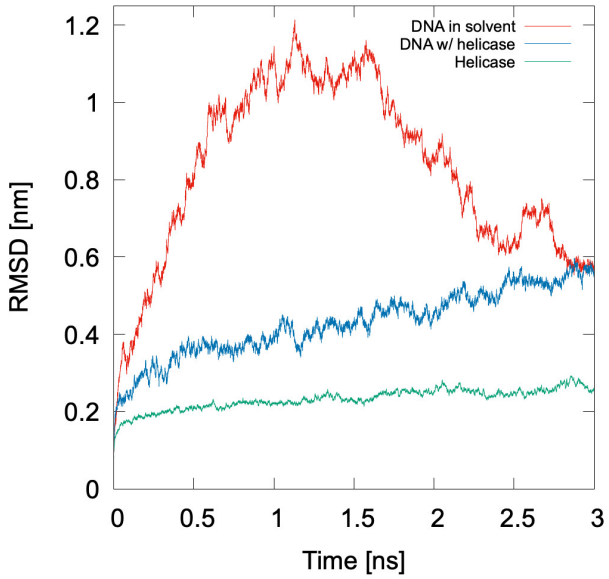
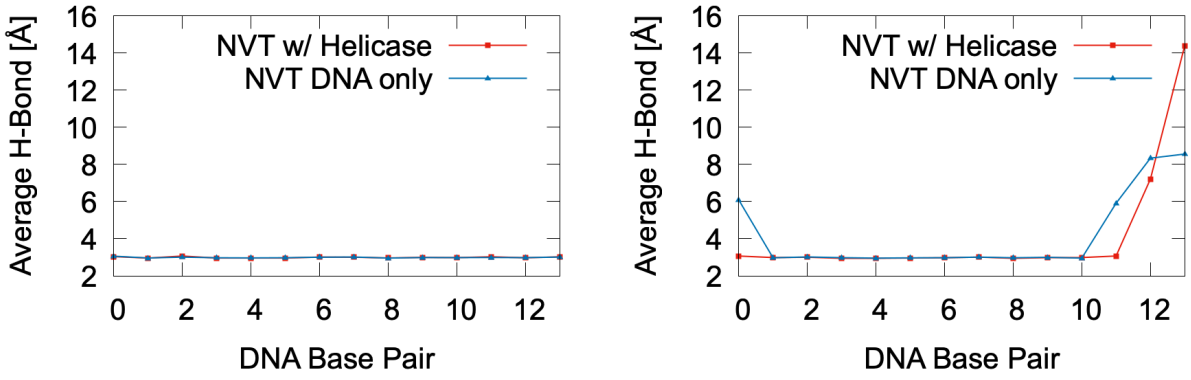


Figure 6-7 Root Mean Square Displacement (RMSD) of the DNA (blue) and protein (green) of the DNA-Helicase product complex during NPT MD. The RMSD of DNA in solvent (without Helicase, shown in red) suggests the Helicase enzyme stabilises the motion of the DNA.

The hydrogen-bond length dynamics during these ensemble simulations was analysed to determine if any strand-separation events induced by the helicase could be observed. Figure 6-8 shows that the passive fluctuations of the helicase do not separate the duplex when the DNA begins in the 'fixed' configuration. Neither is there a clear distinction with and without helicase if the DNA begins in the 'frayed' configuration. This is not wholly unexpected as the helicase's stepping motor action is known to be active and consume ATP [16].



(a) 'Fixed' Duplex

(b) 'Frayed' Duplex

Figure 6-8 Average hydrogen bond lengths in the duplex portion of the DNA during 3 ns of NVT MD.

One aspect of the hydrogen-bond dynamics that is concealed in the average behaviour is the occurrence of temporary dissociations of individual hydrogen bonds. Consider the three hydrogen bonds shown in Figure 6-9. Between about $1.0 - 1.7$ ns of the simulation, two of the three hydrogen-bonds between the GC dimer dissociate for several hundred picoseconds, before returning to their equilibrium length. This data shows that the separation of the base pair is asymmetric, in that the hydrogen bonds are not broken synchronously. It is suggested that such large fluctuations may kickstart the translocation of the DNA by helicase, as the partially dissociated dimer will be far easier to separate.

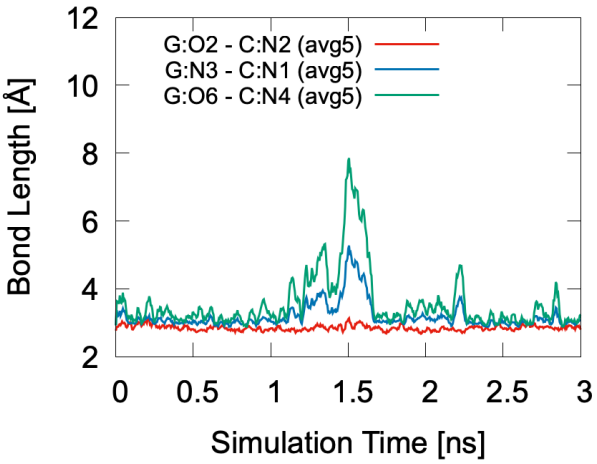
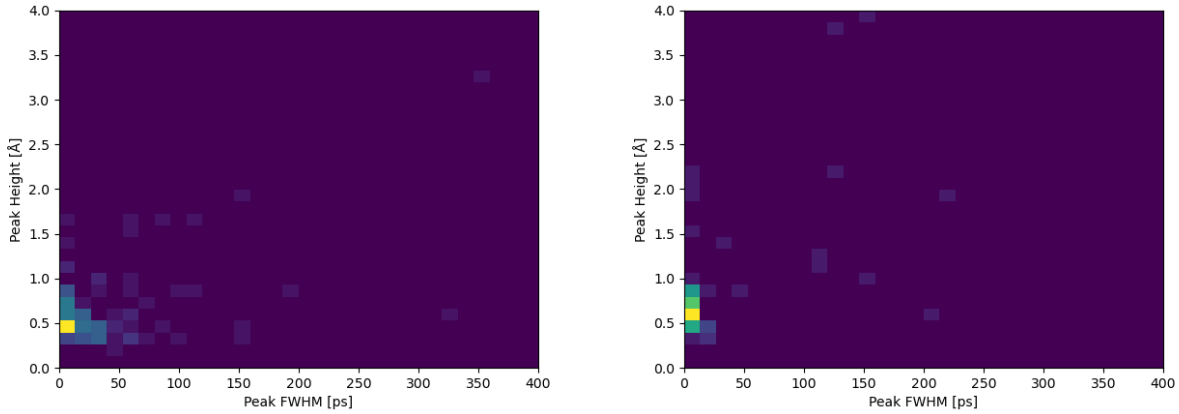


Figure 6-9 Example of temporary hydrogen bond dissociation in DNA base pair during ensemble MD in helicase complex. In the figure a large temporary dissociation of the base pair is seen between 1 and 1.7 ns, as well as several smaller fluctuations. Each of these opening events may be the seeding point from which irreversible strand separation occurs, and may trap tautomeric populations.



(a) with Helicase

(b) Helicase Removed

Figure 6-10 Histograms showing hydrogen bond fluctuation statistics of the penultimate base pair with and without helicase during NVT ensemble MD.

A closer inspection of Figure 6-9, shows that there are many more temporary dissociations of the dimer. In order to sample this behaviour statistically, an automated peak fitting algorithm was used. For both the complex and isolated DNA. Gaussian functions were fitted to the hydrogen bond dynamics from a combined 18 ns of ensemble MD. Histograms summarising the statistics are shown in Figure 6-10. Despite the limited statistics, the presence of the enzyme is correlated with longer (higher peak FWHM) dissociations, and a more consistent dissociation (less variance in peak height).

6.3. Proton Transfer along Hydrogen-Bonds in DNA

Now we come to the focal point of this PhD project's remit: proton transfer in DNA during encounters with the replisome. Requiring more than simply conformational changes, the chemical reaction of proton transfer inherently requires a quantum mechanical (QM) description. Following tests and benchmarks of Quantum Mechanics/Molecular Mechanics (QM/MM) in several atomistic simulation packages, Amber was initially found to be the most practical. Amber's Molecular Dynamics (MD) workflow can easily be extended to include both semi-empirical and ab initio QM (DFT through NWChem) with additive QM/MM.

6.3.1. Adenine Intrabase Transfer

The first trial of reaction mapping in Amber was performed on the adenine intrabase transfer. Figure 6-11 shows the Adenine Nucleic Acid (NA) and the exaggerated path of the proton during transfer. Practically, due to the transfer happening within the same residue, the two reaction endpoints can be described with the same `.prmtop` topology file. This makes such a transfer relatively simple to study in Amber.

A 28-image dynamical Nudged Elastic Band (NEB) routine with a simulated annealing over 600

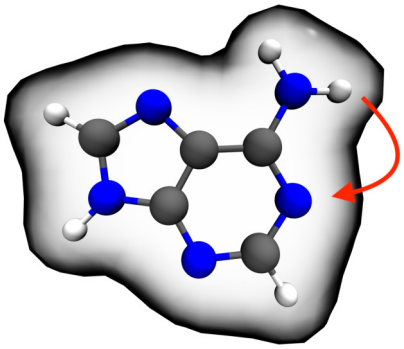
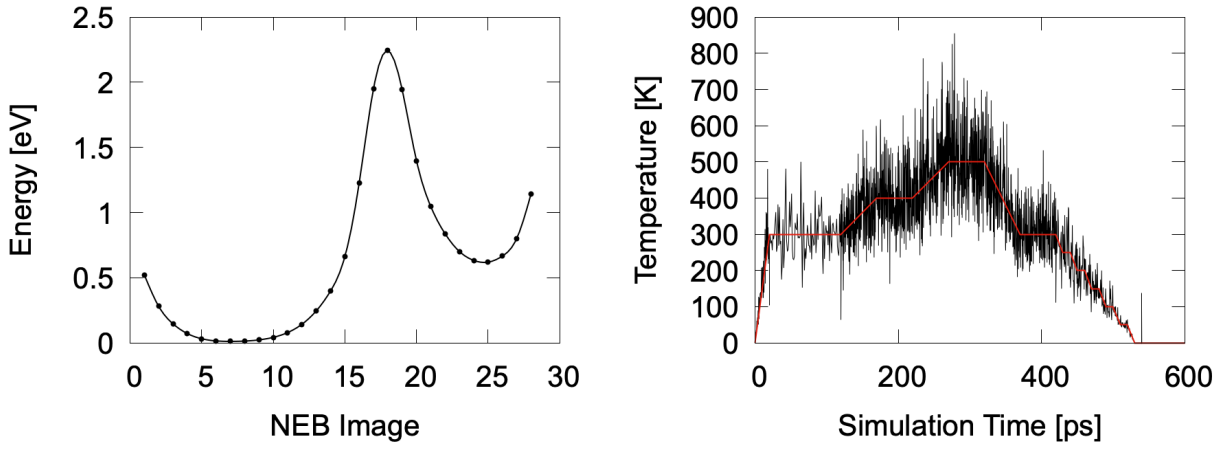


Figure 6-11 The intrabase transfer in Adenine.

ps was used. The energy calculator was of the additive QM/MM type, with the Nucleic Acid in semi-empirical (PM6) QM, and the solvent in classical MM (TIP3P). Section 6.3.1 shows the resulting barrier obtained. Compared to [103] both the asymmetry (0.63 eV in [103] vs 0.5 eV in this work) and barrier energy (2.25 eV in [103] vs 2.1 eV in this work) are underestimated. However, the discrepancies can be attributed to the differences in the level of theory.



(a) Barrier

(b) Annealing Profile

Figure 6-12 Amber/PM6 QM/MM NEB barrier of intra-base transfer in Adenine. Figure 6-12a shows the PES along the path of least action obtained from a simulated annealing NEB method with an annealing profile seen in Figure 6-12b

.

6.3.2. Helicase Complex

We now progress from studying the transfer in the small-scale NA simulations to the large-scale DNA-Helicase complex. As shown in Figure 6-6 of the previous section, the full molecular system is very large, comprising 724 AA in the enzyme, and 25 NA. Counting the solvent molecules, the entire system contains over one hundred thousand atoms. This is far too large for a full quantum mechanical calculation, thus a QM/MM theory is once again implemented. This time only a sphere of 20 Å around the dimer of interest (the 14th base pair in the sequence depicted in Table 6-2) is allowed to move, and all other atoms are frozen.

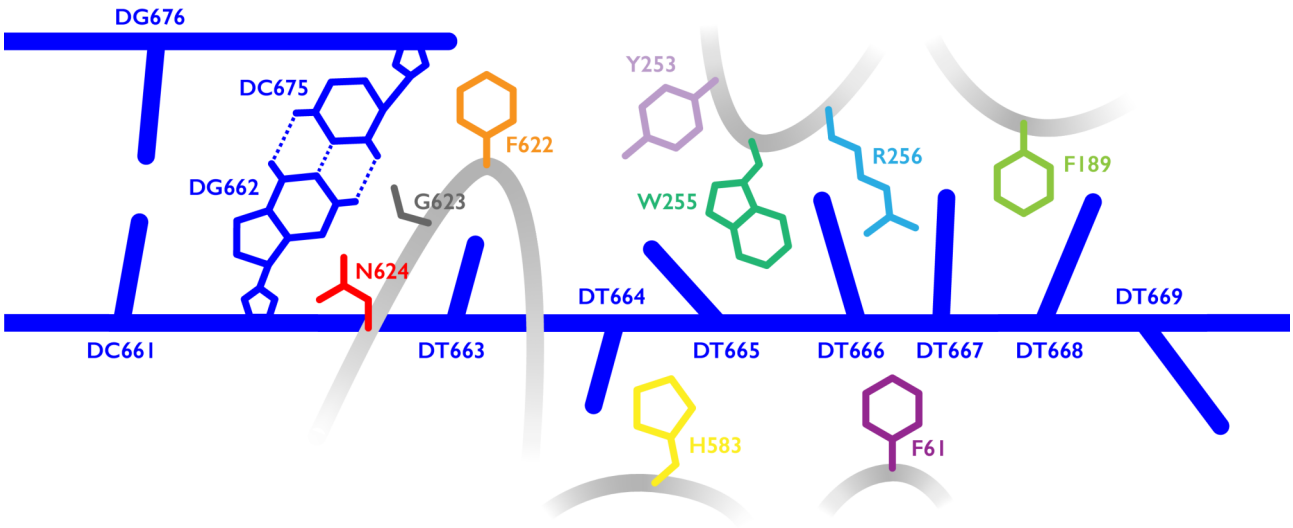


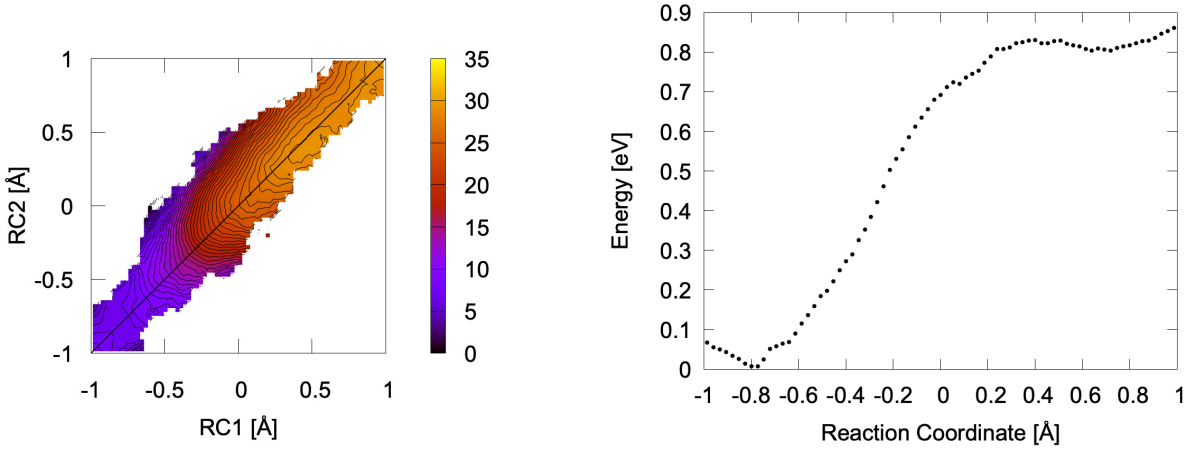
Figure 6-13 Illustration of the PcrA Helicase DNA-binding site. DNA in blue showing the end of the duplex with the Guanine-Cytosine dimer where DPT is of interest, and single-stranded thymine tail. The three amino acids near the dimer of interest (N624, G623, & F622), as well as those integral to ssDNA translocation, are drawn attached to their grey protein backbone. For a clarification on the shorthand used in this diagram see Section A.

The ssDNA binding site of PcrA helicase comprises many key residues, as identified in [20]. Figure 6-13 illustrates the local structure of the PcrA Helicase ssDNA binding site. The residue F622[3] is instrumental in the separation of the DNA by PcrA helicase [20]. In addition to F622, the residues N624 and G623 may contribute to the DPT in GC.

An Umbrella Sampling (US) (see Section 5.3) procedure was used to map the Double Proton Transfer (DPT) reaction with two Reaction Coordinate (RC) dimensions (one for each hydrogen transfer). For the Molecular Mechanics (MM) the ff14sb and OL15 Force Fields were used for the protein and DNA, respectively. The solvent model was TIP3P. Initially, this reaction was sampled to PM6/MM level of theory. In this case, 75 sampling windows were utilised and 1 picosecond of Steered Molecular Dynamics (SMD) was performed for each window (1000 simulation steps). The free energy surface with energy contour lines is shown in Figure 6-14a, while Figure 6-14b shows the Potential of Mean Force (PMF) through this surface.

This calculation was repeated at a higher QM/MM level of theory comprising DFT in NWChem with the B3LYP XCF and the 6-311++g** basis set. Due to the increased computational cost, the sampling scheme was reduced to 15 windows of 250 femtoseconds (250 simulation steps). Section 6.3.2 similarly shows the two-dimensional reaction surface and PMF for the DPT.

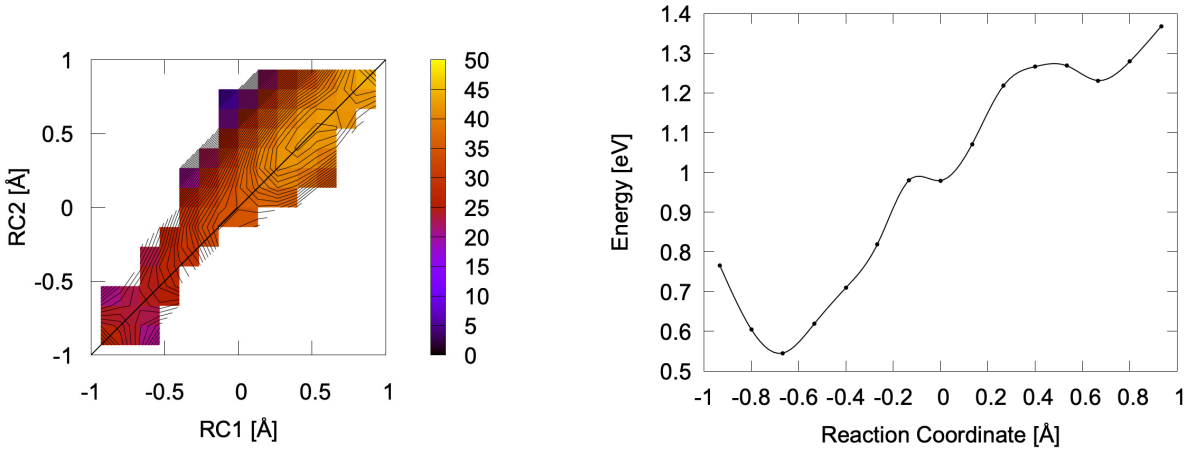
[3]For a clarifcation on the shorthand used in this work see Section A.



(a) Free Energy Surface

(b) Reaction Free Energy

Figure 6-14 Amber/PM6 QM/MM Umbrella Sampling results for concerted double proton transfer in Guanine-Cytosine base pair embedded in PcrA Helicase.



(a) Free Energy Surface

(b) Reaction Free Energy

Figure 6-15 Amber/NWChem (B3LYP) QM/MM Umbrella Sampling results for concerted double proton transfer in Guanine-Cytosine base pair embedded in PcrA Helicase.

System	Theory	ΔG [eV]	$\Delta\Delta G_r$ [eV]
GC in solution [103]	DFT to B3LYP	$0.427^{\dagger}$	$0.266^{\dagger}$
GC in DNA [115]	QM/MM w/ DFT to B3LYP	0.496	0.044
GC in Helicase complex	QM/MM w/ semi-empirical PM6	0.801	0.022
GC in Helicase complex	QM/MM w/ DFT to B3LYP	0.686	0.037

Table 6-3 The asymmetry ΔG and reverse barrier $\Delta\Delta G_r$ for the concerted DPT in GC with various systems but comparable theory. A † denotes potential energy instead of free energy.

Considering the Potentials of Mean Force (PMF) shown in Section 6.3.2 we can see that there is a large energy asymmetry between the canonical and tautomeric states, as well as a very shallow tautomeric minimum, in both the semi-empirical and DFT results. The asymmetry ΔG and reverse barrier $\Delta\Delta G_r$ from this workm as well as two select references, are shown in Table 6-3.

The preliminary results of this work show a further extension of the trend that increasing local complexity of the environment surrounding the GC base pair in which DPT is to occur, increases the asymmetry as well as instability (lower reverse barrier) of the tautomeric metastable state.

7. Proton transfer during DNA strand separation as a source of mutagenic guanine-cytosine tautomers

This chapter is based on the article: Louie Slocombe, Max Winokan, Jim Al-Khalili, and Marco Sacchi. "Proton transfer during DNA strand separation as a source of mutagenic guanine-cytosine tautomers". In: *Communications Chemistry* 5.1 (Nov. 2022), p. 144. DOI: https://doi.org/10.1038/s42004-022-00760-x [1] Supplementary information is included in Appendix Section B.

My contributions to this project include the design and execution of the Steered Molecular Dynamics (SMD) methodology for the determination of an intrinsic atomistic separation speed of aqueous duplex DNA under replisome conditions. This included: preparation of aqueous DNA structures for simulation; determination and testing of MD and steering force parameters; preparing and overseeing many replica SMD simulations on High-Performance Computing resources; development of novel analysis techniques to determine dynamical characteristics of base pair opening; and discussion of the implication of the SMD results in comparison to literature and the Density Functional Theory results obtained by co-author Louie Slocombe. Additional contributions were made to the writing and proofreading of the article body, preparation of the publications figures and proposed table of contents image (Figure 7-1).

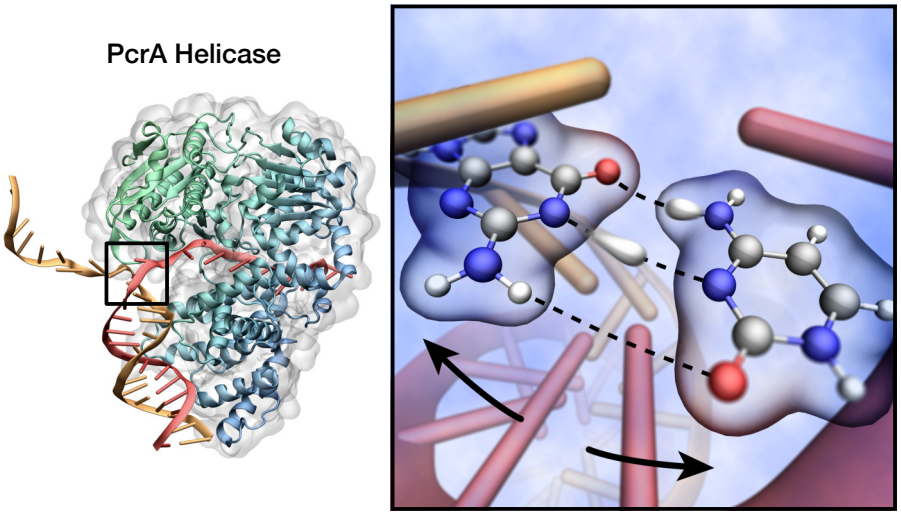


Figure 7-1 Double Proton Transfer during DNA Strand Separation.

7.1. Abstract

Proton transfer between the DNA bases can lead to mutagenic Guanine-Cytosine tautomers. Over the past several decades, a heated debate has emerged over the biological impact of tautomeric forms. Here, we determine that the energy required for generating tautomers radically changes during the separation of double-stranded DNA. Density Functional Theory calculations indicate that the double proton transfer in Guanine-Cytosine follows a sequential, step-like mechanism where the reaction barrier increases quasi-linearly with strand separation. These results point to increased stability of the tautomer when the DNA strands unzip as they enter the helicase, effectively trapping the tautomer population. In addition, molecular dynamics simulations indicate that the relevant strand separation time is two orders of magnitude quicker than previously thought. Our results demonstrate that the unwinding of DNA by the helicase could simultaneously slow the formation but significantly enhance the stability of tautomeric base pairs and provide a feasible pathway for spontaneous DNA mutations.

7.2. Introduction

In biology, the separation of a DNA duplex occurs via fraying at its terminal base pairs due to random thermal effects or the action of a helicase enzyme during the DNA replication cycle. Once the separation of DNA has started, it will likely propagate down the duplex due to the force cascading down the ribose-phosphate backbone. In the case of helicase enzymes, an active, stepping-motor action pulls on one of the strands of DNA through a narrow opening in the enzyme, thereby forcing apart the nucleobase pairs[18].

The interaction of the non-canonical, tautomeric state of a nucleobase pair in DNA within the helicase has thus far been overlooked in the literature. Specifically, the process of DNA strand separation and its impact on proton transfer has been assumed to be simply a matter of comparing the proton transfer timescale (via quantum and classical effects) to the timescales of the biological process.

If a tautomer passes through the replication machinery, it will form a mismatch with the wrong corresponding base on the copy strand. For instance, the tautomer of guanine will pair with thymine instead of cytosine (G-C $\leftrightarrow$ G*-C* $\rightarrow$ G*-T, where the star denotes the tautomeric non-standard form) [175, 176]. Furthermore, the mismatched base pair can evade fidelity check-points of the replisome by adopting a structure similar to a Watson and Crick base pair [175, 176] resulting in an error in the genetic code and hence a point mutation.

Florian and Leszczynski [177] first proposed that for the tautomeric mechanism to be biologically relevant, the tautomers must remain stable during the long process of DNA unwinding and strand separation, which are the prerequisite steps for the synthesis of the new DNA strand by the polymerase. Consequently, the lifetimes of the tautomers should exceed this characteristic time for the

base-pair opening ($\sim 10^{-10}s$) [177].

In the last decade, numerous authors [33, 178, 111, 179, 115, 180, 176] argued that the tautomers' lifetime is much shorter than the helicase separation time. Therefore no tautomeric population would successfully survive the DNA strand separation by the enzyme. If the G-C tautomer has a short lifetime and reverts to the standard canonical form, the potentially mutagenic point defect is rendered ineffective during the uncoiling process. Subsequently, the tautomer is not propagated into the two single-stranded DNAs. On the contrary, if the tautomeric lifetime is longer than the double-strand separation time, the tautomeric form will survive the biological process. Under closer inspection, the timescale reasoning requires further justification and refinement. Here, we will unpick some of the core assumptions and provide evidence for the need for a more careful investigation of enzyme effects on the DNA tautomers.

In the following sections, we first use quantum chemical models to determine the effect of an induced separation of the two strands of DNA on the structures of the G-C and G*-C* dimers and on the characteristics of the minimum energy pathway linking the two endpoints between the bases. We find that the features of the proton transfer are quasi-linearly correlated with the separation distance. To accompany our quantum chemistry calculations of the G-C dimer, we also evaluate the occurrence of separation events in classically simulated aqueous DNA subjected to a small separation force. We find a wide variety of opening events but reveal a characteristic separation speed unaffected by choice of steering force.

7.3. Results

We model the separation of the DNA bases using Density Functional Theory (DFT) at the B3LYP + XDM / 6-311++G** [181] level of theory (NWChem [182]) with an implicit solvent. In the DFT calculations, we truncate the model to the G-C dimer, constrain the R-group atom (where the base would join the rest of the DNA), and separate the bases. Fig. 7-2 provides a summary of the scheme. See the Methods section for further information on the separation methods.

We systematically vary the separation distance between the bases and study the effect of this splitting on the hydrogen bond lengths and energies. Fig. 7-3 shows the structural changes of the G-C base /as a result of the induced separation distance. For the canonical form Fig. 7-3a and Fig. 7-3c, initially, there are no visible changes to the structure other than the elongation of the hydrogen bonds holding the bases together. However, as the separation distance increases, the two bases undergo an internal rotation relative to each other (measured by the angle θ in Fig. 7-2). The rotation helps to minimise the length of one of the O–H–N hydrogen bonds (B1 or B3) while the other two bonds are stretched. In the DFT calculations, there is a clear preference for the O–H–N hydrogen bond (B1) to maintain its equilibrium length while the base rotates. The non-uniformity of the separation implies that the bases do not synchronously split apart, but instead separate

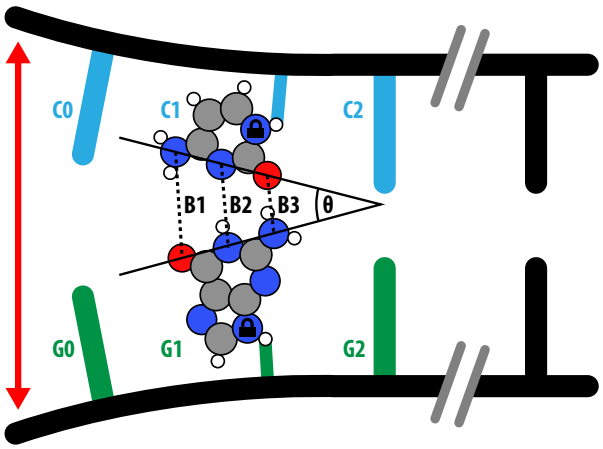


Figure 7-2 **The separation scheme used to investigate how the canonical and tautomeric G-C base pairs separate.** Four G-C base pairs of the 14 base pair DNA duplex used in the molecular dynamics simulations are shown, with a separating force (red arrow) applied to the first base pair's (G0-C0) backbone. DFT calculations were performed only on base-pair G1-C1, where atoms marked with the lock icon were fixed. We define the three hydrogen bonds; Bond 1 (B1) as the distance measured from DG:O^6-DC:N^4, Bond 2 (B2) from DG:N^1-DC:N^1, and Bond 3 (B3) DG:N^2-DC:O^2. The opening angle θ measures the asymmetry with which the hydrogen bonds stretch.

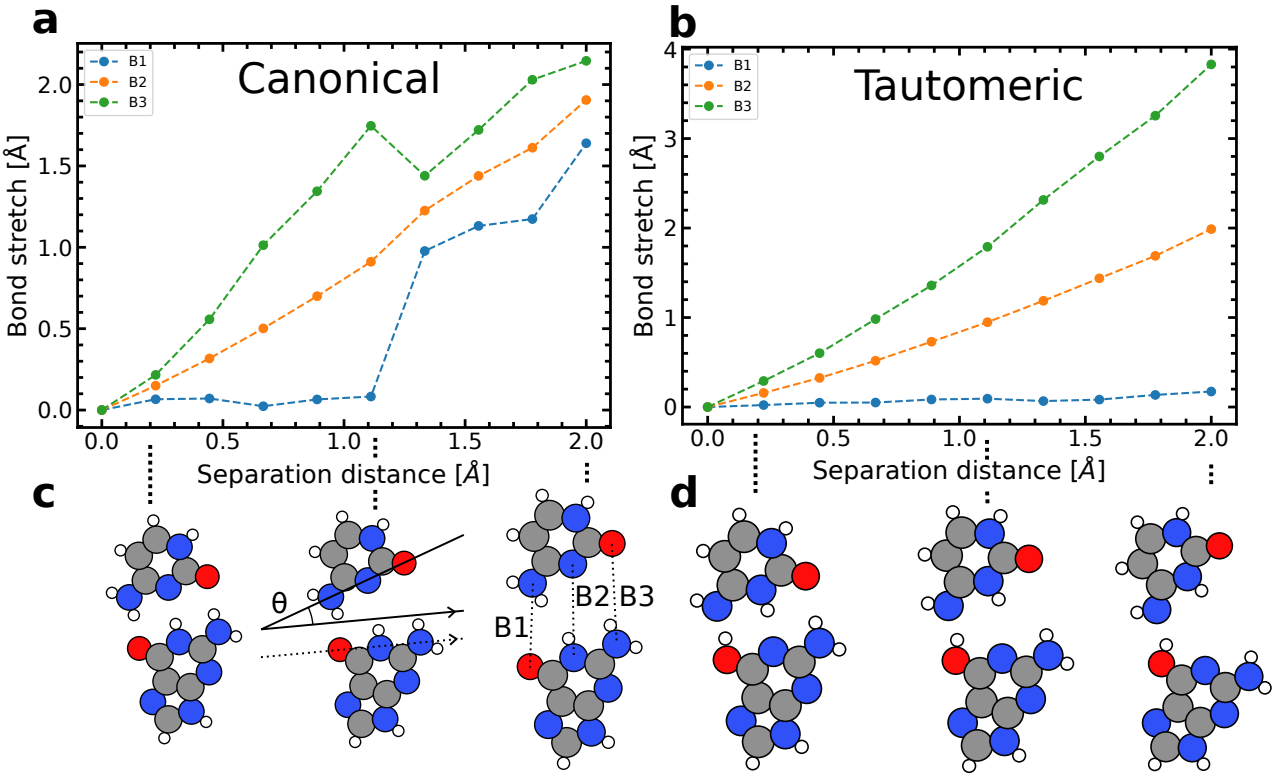


Figure 7-3 **Bond length dependency on the separation distance of the G-C dimer.** Here the separation distance is defined as the distance between the non-hydrogen bonded atoms participating in the hydrogen bonds) calculated by DFT. (a) The stretching of the canonical form of G-C from their unconstrained equilibrium lengths. The equilibrium lengths for the canonical base are 2.89 Å, 2.96 Å, 2.89 Å for B1, B2, and B3 respectively. (b) The stretching of the bonds of the tautomeric form of G-C, where two hydrogens have transferred. The equilibrium lengths are 2.61 Å, 2.89 Å, 3.01 Å for the tautomeric base. Whereas (c) and (d) show the structural changes of the G-C base's canonical and tautomeric forms, respectively. During separation, the bond lengths and angle significantly change.

asymmetrically. The rotation about the fixed R-group is physically consistent since it is the only covalently bonded link between the base and the rest of the DNA.

A non-linear change in the bonding angle suggests that the proton transfer mechanism is fundamentally different from the idealised equilibrium picture previously modelled [177, 33, 178, 111, 179, 115, 180, 176] (see references within), where the bases are assumed to be largely unaffected by the pulling of the helicase enzyme. However, these calculations indicate that the helicase-DNA interaction cannot be ignored and requires further investigation.

A comparison between the canonical vs the tautomeric form is shown in Fig. 7-3a and Fig. 7-3b, demonstrating that there is some significant difference between the rotation of the base, depending on where two of the three hydrogen bond protons are located. Here, the O-H bond of the tautomeric G offers a much more comprehensive hydrogen bonding range due to being on the outer edge of the molecule - in comparison to the standard form of C. As a result, the O-H bond of the tautomeric G remains in a hydrogen bond for much longer as the separation distance increases.

The bond length stretching is further highlighted in the bottom panels of Fig. 7-3. The panels show the change in the length of each hydrogen bond in the canonical G-C and tautomeric form. For the canonical case, the O-H-N bond is shown to stay relatively constant during the separation until $1.2\,\text{Å}$. After this point, it begins to stretch in line with the other hydrogen bonds. Whereas for the tautomeric form, the top bond is essentially not involved in the breaking until a separation distance of $2.0\,\text{Å}$. Further details can be found in Supplementary Note 1.

7.3.1. Dynamics of the Separation Process

Building upon our DFT calculations, we explore the separation dynamics of a G-C base pair within a more extensive model system comprising aqueous double-stranded DNA, with 14 base-pairs in total. For these calculations, we apply a steering force during a Molecular Dynamics (MD) simulation to model the external action of a helicase enzyme. Computational details are available in the Methods section. The pulling force was applied between the backbone atoms of the first G-C base pair to increase the likelihood of separation during the simulations. The three hydrogen bond lengths (B1, B2, and B3 in Fig. 7-2) of the base pair in question were analysed over a range of MD replicas to gather statistics on the separation dynamics. A large number of distinct but short lived fluctuations are observed, mimicking the breathing of DNA. Should these fluctuations possess properties independent of the steering force, we can argue that they are characteristic of DNA strand separation, and thus also transferable to enzymatic action.

Fig. 7-4 provides an example fit of a separation time series, as well as the resulting statistics across all our dynamics simulations. While separation speed varies in a complicated manner with the pulling force, there is a significant overlap of the standard error of adjacent base pairs and forces. Thus, we conclude that the separation dynamics occur with a separation speed of approximately $1.2\,\text{Å}\,\text{ps}^{-1}$, without a significant correlation to the force or base pairs. Now satisfied that

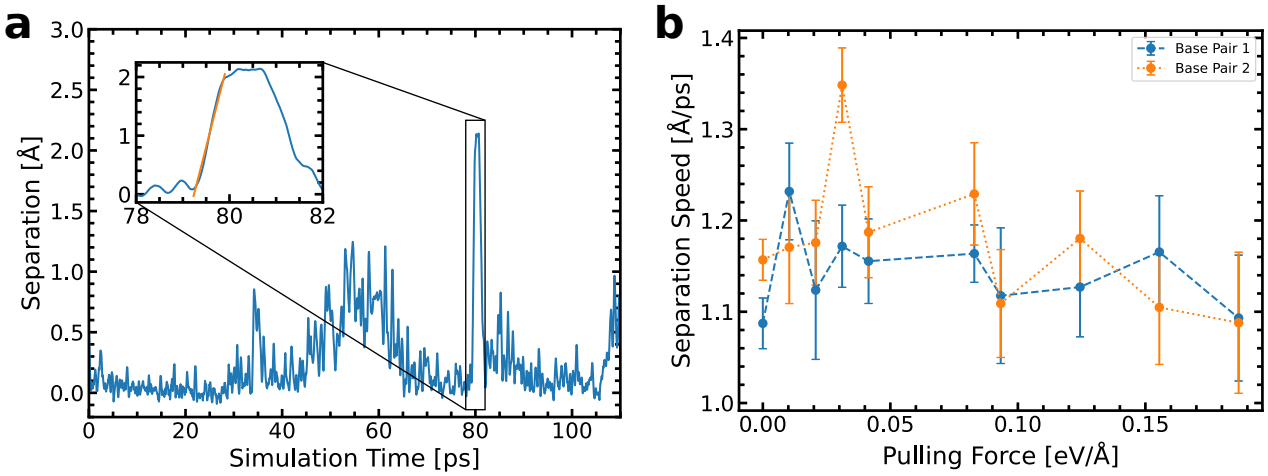


Figure 7-4 **The procedure for estimating the base pair separation speed.** (a) Demonstrating a separation event during a $200\,\mathrm{ps}$ molecular dynamics simulation on the time series of the base pair separation (blue solid line). The inset figure includes a linear line of best fit (orange solid line) of the separation event to determine the speed. (b) The arithmetic mean and standard error of separation speeds for a range of forces taken from a sample of 210 molecular dynamics simulations containing n=1442 separation events for base pair G1-C1 (blue error bars joined by blue dashed line) and for base pair G2-C2 (orange error bars joined by orange dashed line).

we are not introducing significant bias, we turn our attention to the atomistic mechanism of the separation events.

Following the previous timescale hypothesis [177], and assuming that, after $2.0\,\text{Å}$ of separation, no reverse proton transfer occurs (see Supplementary Note 2), the tautomer's lifetime must exceed $\sim 1.7\,\mathrm{ps}$. This requirement is two orders of magnitude shorter than the quoted characteristic time for the base-pair opening $\sim 100\,\mathrm{ps}$ [177]. Descriptions of the mechanism of DNA strand separation by helicase enzymes are informed by the rate at which the enzyme translocates DNA by measuring the number of base-pairs processed by the enzyme in a short period[183, 16, 173, 18, 19, 17]. This ignores the possibility that individual stages of the helicase's dynamics occur considerably faster. Our atomistic view of the separation shows that splitting individual base pairs is much quicker than the overall speed of the helicase action since it does not include translocation activity. While the tautomer certainly does not outlive a complete cycle of helicase's stepping motor action, it needs to outlive the time taken to separate the base pair up to $2.0\,\text{Å}$. Our results demonstrate that the separation of a base pair due to external action occurs within a timescale relevant to that of the proton transfer. Further on in this letter, we explore the implications by considering the effect of base pair separation on the energetics of the tautomerisation via double proton transfer and reconsider the relevancy of the timescale.

7.3.2. Opening Angles

To further clarify the complicated way in which an external force separates DNA, an opening angle (θ) was defined for each separation event (see Fig. 7-2 and Methods section). Fig. 7-5 demonstrates

a bimodal distribution, peaked at $\theta = (18.9 \pm 0.1)^\circ$ and $(-17.4 \pm 0.1)^\circ$, with only a few events showing synchronous, symmetric stretching of the three hydrogen bonds (zero opening angle). Consequently, the separation dynamics does not occur in a perfectly symmetric fashion as generally assumed[177, 33] and instead, an asymmetric breaking mechanism between the two DNA strands is much more probable (97.8% of events have $|\theta| > 3^\circ$), with a clear preference for specific opening angles which have the minimum energy requirements. In comparison to our DFT calculations, the introduction of the backbone fluctuations and thermal ensemble induces a bias towards one or the other of the opening directions. 39% of the trajectories follow a path that closely follows the direction described by the DFT result, where the system is restrained at the R-group. For the negative opening angle distribution (the one consistent with the DFT picture), and given the structure of the G-C base pair, an opening angle magnitude of $(-17.4 \pm 0.1)^\circ$ suggests that the length B1 stays fixed while B3 is stretched by approximately $\sim (1.4 \pm 0.4)\,\text{Å}$. This value compares well to the DFT geometry optimisations (see Fig. 7-3) which predicts B3 to stretch up to $\sim 1.7\,\text{Å}$ (relative to B1) before B1 is dilated. This suggests that the bond angles of the DFT calculations on the single base pairs are consistent with the MD picture, which includes the larger structure. However, it is unclear how the larger structure influences the proton transfer.

The bimodal distribution in the DNA separation events demonstrates a diverse and rich environment of energetic scenarios that are radically different from the idealised assumption made by previous authors, who either reduce the problem to a comparison of lifetimes disregarding the mechanisms of strand separation [177, 95, 184, 111, 115, 180] or perform their calculations only in the static aqueous dimer [177, 104, 95, 96, 122, 33]. Although several authors have pointed to the fact that the complex external environments may strongly determine the influence of tautomers on mutation [178, 179, 117, 176], our MD results show just how diverse the biological environment experienced by DNA is.

7.3.3. Proton Transfer

For each separation distance, we perform an analysis of the double proton transfer scheme using a machine learning approach to the nudged elastic band algorithm [157, 158], which yields the minimum energy path and determines the transition state of the reaction. We connect the canonical to the tautomeric form producing an energy landscape for the double proton transfer, see Fig. 7-6a and Supplementary Note 1 and 2.

We define the reaction energy asymmetry as the energy difference between the canonical G-C and the double proton transfer, tautomeric G*-C* product. In Fig. 7-6b, the asymmetry is displayed as a function of the separation reaction coordinate. Initially, the reaction asymmetry corresponds to the unconstrained calculation ($0.51\,\text{eV}$) at the equilibrium distance. As the separation distance increases, the asymmetry drops to a minimum of $0.37\,\text{eV}$ at $1.5\,\text{Å}$. From 0 to $1.0\,\text{Å}$ the reaction asymmetry briefly dips and then rises; this is due to a complex interplay between the local rearrangement of

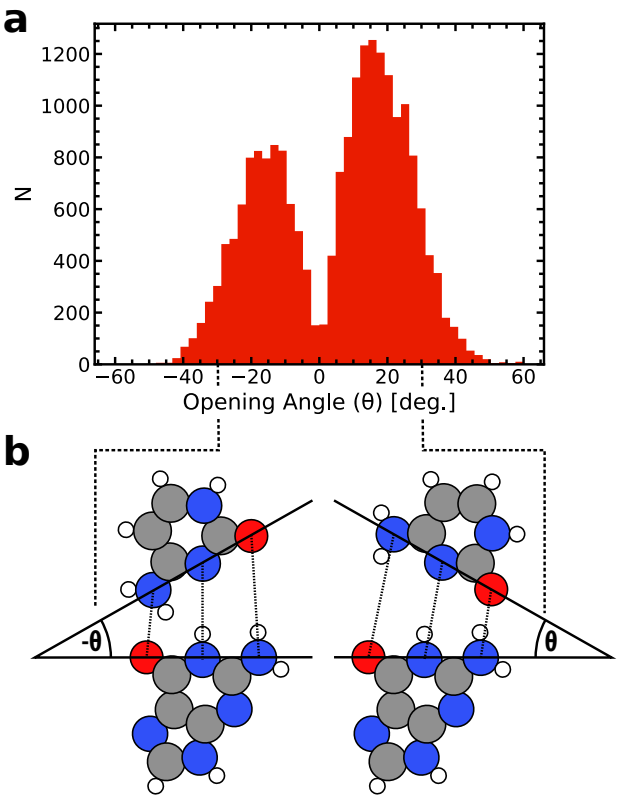


Figure 7-5 **The statistical distribution of opening angles.** (a) The statistical distribution of opening angles during steered molecular dynamics in base pairs 1 and 2 (red filled histogram bars). Negative angles suggest that B1 stays fixed while B3 opens. The static analysis with DFT suggests an opening angle of -22 degrees is energetically favourable without thermal effects. (b) Example of two snapshot geometries from MD runs, highlighting the direction of the opening angle (θ).

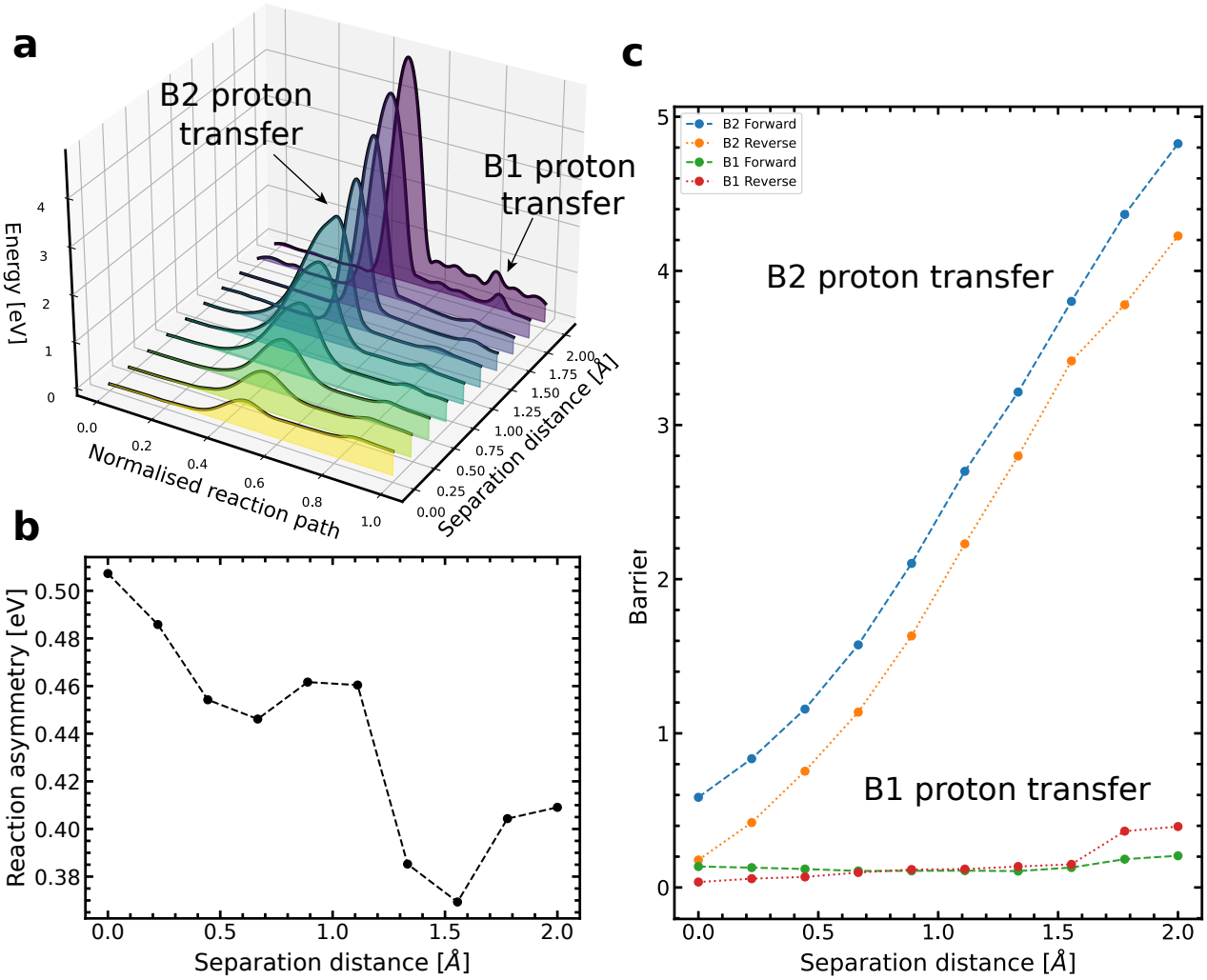


Figure 7-6 **The double proton transfer tautomerisation reaction pathway** (a) Double proton transfer tautomerisation as a function of separation distance and the reaction path image. (b) Demonstrating the changes in the reaction asymmetry as a function of the base separation distance. (c) The changes in the forward and reverse reaction barriers of the canonical to tautomeric double proton transfer scheme in G-C as a function of the imposed separation distance. The plotted barrier energy is the transition state energy referenced to either the canonical, tautomeric or intermediate (single proton transfer) stable state.

atoms, the bonding configuration, and the rotation of the base about the aforementioned R-groups. At separation distances greater than $1.5\,$Å, the asymmetry begins to increase.

The proton transfer pathway can be described as follows. Initially, we observe the O-H bonds stretch with little to no rotation or buckling of the overall structure of either base; instead, the bases remain essentially facing each other ($\theta = 0.0°$). We note that the B2(N-H-N) proton moves first, followed by the B1(O-H-N) proton. However, as the separation distance increases, both bases rotate, in the opposite sense, during the reaction pathway to minimise the bond lengths (see figures in Supplementary Note 1). The general reaction pathway changes so that the bases rotate after the first proton transfer, preceded by another transfer. For more information, see Supplementary Note 1.

We observe a two-step transfer process where the middle hydrogen initially moves (B2), followed by

the B1 hydrogen. Thus, the reaction path comprises two energy barriers, indicating the presence of a stable single proton transfer intermediate between two transition states. The intermediate corresponds to a structure in which only the B2 hydrogen atom has moved from G to C. Conversely, there is no stable intermediate corresponding to a single proton transfer for the B1 hydrogen bond to move. Fig. 7-6c summarises how the two barriers change with separation distance. The B2 hydrogen reaction barrier is very sensitive to the separation distance, and it rapidly grows with increasing separation (from $0.57\,\mathrm{eV}$ to $4.80\,\mathrm{eV}$). On the other hand, the B1 hydrogen reaction barrier stays approximately constant for distances below $0.9\,\text{Å}$ due to the bases rotating, keeping the top bond at the equilibrium length. While the B1 hydrogen reaction barrier is constant during the initial separation, then as the base distance increases, it increases rapidly as the bond lengthens.

Gheorghiu *et al.* [115] have observed both the concerted and stepwise proton transfer mechanism, while others, including Brovarets *et al.* [111] and Slocombe *et al.* [180], only observe a concerted mechanism. Gheorghiu *et al.* [109] found that the proton transfer mechanisms varied during the ensemble quantum mechanics/molecular mechanics simulations and that the concerted DPT mechanism for G-C is only a small subsample of a more extensive collection of viable mechanisms: double proton transfer, concerted vs stepwise, single proton transfer, concerted vs rearrangement. Instead, for G-C, the stepwise process dominates with a probability of 0.84 vs 0.12 for the concerted mechanism due to the interaction with the larger DNA structure and local solvent environment. Gheorghiu reports that the first reaction barrier is $(0.61 \pm 0.05)\,\mathrm{eV}$ and the second $(0.07 \pm 0.03)\,\mathrm{eV}$ and a reaction asymmetry of $(0.59 \pm 0.05)\,\mathrm{eV}$.

In this study, we found that at the global minimum (no separation distance), there is a reaction asymmetry of $0.507\,\mathrm{eV}$, which is slightly larger than in our previous work [180] due to the interactions with the solvent and the incorporation of dispersion corrections. The first barrier has energy $0.574\,\mathrm{eV}$, and the second barrier has energy $0.516\,\mathrm{eV}$. Thus, there is a $0.058\,\mathrm{eV}$ reverse barrier from the double to single proton transfer product. The single proton transfer minimum has an energy of $0.399\,\mathrm{eV}$ relative to the canonical form.

Consequently, during the cleavage process, the energetic landscape of the reaction will change as a function of time. As a result, the reaction barrier and the energy difference between the reactant and products would drastically change. The change in the energetic landscape could also depend on the timescale of the separation rate compared to the period of the vibrations in G-C. Provided that the vibrational modes of the bases are similar or quicker than the timescale of the separation, the bases will have time to rearrange during the separation as calculated here. The rearrangement during the separation must then be incorporated into the model determining the rate. Conversely, if the separation is quicker than the system's dynamics, one can assume that all the atoms are stuck in place while the bases dissociate (frozen approximation).

On the other hand, Slocombe *et al.* [117] demonstrated there is a continuous exchange of the canonical and tautomeric forms due to the fast reaction rates, in turn, due to a significant quantum

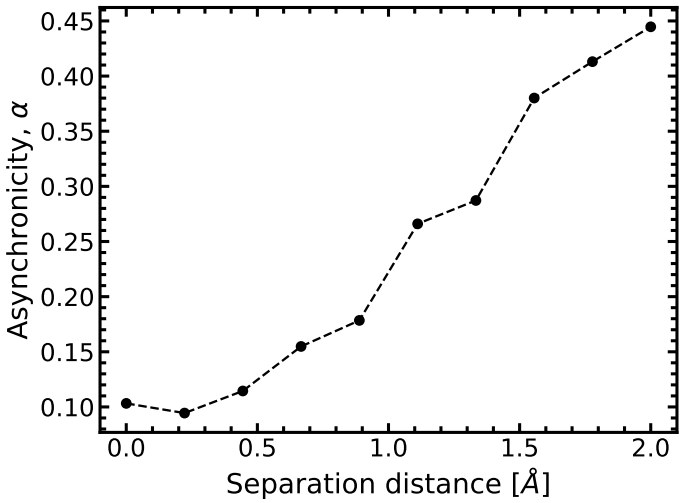


Figure 7-7 **Measuring the asynchronicity as the DNA bases disassociate.** Here, the asynchronicity (α, black circles joined by black dotted line) is calculated for each double proton reaction and displayed as a function of induced separation distance.

component. While the lifetime of a single tautomer might be short, the formation rate (forward reaction rate) might be high, such that over an ensemble of bases passing through the helicase, a proportion of them pass. During the separation process, the canonical reactant is continuously forming the product. Consequently, it is a combined process that competes with the separation timescale.

Including a more comprehensive description of the cellular environment (stacking and solvent effects) has recently been suggested to alter the reaction asymmetry [178] such that the tautomeric state vanishes on the free energy surface. Such interactions could avert the tautomer's formation and prevent it from potentially leading to a mutation. Gheorghiu *et al.* [115] suggest coupling to the larger environment offers a diverse ensemble of reaction pathways. To investigate this, we took a snapshot from an MD run presented above and re-determined the reaction path when including the rung below the separating base (pair G2-C2 in Fig 7-2). In the calculation we assume that the base above is fully separated and thus, we can omit it from our simulation system as it no longer introduces stacking interactions with the base in question. Furthermore, we assume a separation of timescales whereby the base which begins to separate would relax, while the base below would remain largely unchanged by the base separation. This assumption is further justified in Supplementary Note 3. Using the same methods as described before, at $0.39\,\text{Å}$ separation distance the first energy barrier for proton transfer has an energy of $1.08\,\text{eV}$. This can be compared to $1.06\,\text{eV}$ at the same distance but with two isolated bases. The second barrier has an energy of $0.098\,\text{eV}$ (vs $0.120\,\text{eV}$ with the isolated base pair), with a reaction asymmetry of $0.467\,\text{eV}$ (vs $0.462\,\text{eV}$ with the isolated base pair). The second reaction barrier shows the largest difference to the PES of the single base pair proton transfer, with a change of $0.02\,\text{eV}$. This finding is consistent with Das *et al.* [110] who conducted MD simulations and showed that adjacent base pair stacking modifies the proton transfer profile on the order of $0.04\,\text{eV}$.

Fig. 7-7 displays the asynchronicity as a function of the separation distance induced between the DNA bases. Asynchronicity is a measure of the separation between the two proton transfer events; further detail can be found in the methods section and Supplementary Note 4. Here we use asynchronicity to quantify how the proton transfer mechanism changes during the base dissociation process. Fig. 7-7 demonstrates that the double proton transfer initially has some concurrence indicated by a low asynchronicity value. However, as the bases are further apart, the first and second proton transfer becomes an increasingly separate event. The disconnection of the two proton transfer events along the reaction path could lead to a distribution of outcomes of product states. Furthermore, the increased asynchronicity indicates an increased localisation of the single proton transfer. Consequently, the single proton transfer could also occur along with the double proton pathway. However, due to the prohibitively significant initial forward reaction barrier, the population of either product becomes increasingly unlikely as the DNA base is further separated by the replication machinery.

On the other hand, proton-coupled electron and hole transfer (PCET) is a prominent feature of radiation-induced excited state dynamics and subsequent DNA damage [185]. The current theoretical framework of PCET has successfully treated these types of problems [186, 187]. For example, applying PCET to the excited state dynamics of the one-electron oxidation of G-C decouples the proton and electron transfer from the middle hydrogen bond N of G (B2 in our labelling) to the N on C [185]. Similarly, a PCET excited-state deactivation mechanism for G-C has been proposed from experiment and theory (see review Kumar *et al.* [185]). Furthermore, Femtosecond transient absorption spectroscopy suggests that the excited state PCET between the DNA stands has a pronounced deuterium isotope effect [188]. These findings match the transition state calculations presented in this paper, as we determine the same B2 proton transfer pathway.

7.4. Discussion

In this work, we analysed the double proton transfer rate during DNA strand separation and proved that a simple comparison of the tautomeric lifetime is insufficient to determine the survival probability of the tautomer during this process. We propose that the proton transfer potential is not static; instead, the synchronicity of the transfer process and the activation barriers for each proton transfer drastically change as the DNA strands are pulled apart. For the G-C base, we observe both rotation and internal rearrangement of the bond lengths to minimise the energy requirement during the breaking of the hydrogen bonds as the bases split; this has a profound effect on the proton transfer energy landscape. As a result, the double proton transfer mechanism becomes an asynchronous and stepwise process with two different and well-defined reaction barriers. In particular, we observe a linear dependence of the energy of the first barrier on the separation distance. Consequently, the G*-C* tautomeric state becomes more stable as the hydrogen bonds break. The

overall reverse barrier of the proton transfer (G*-C* $\rightarrow$ G-C) increases rapidly as a function of the separation distance between G and C; this yields a drastic increase in the lifetime of the potentially mutagenic tautomer (G*-C*). On the other hand, the forward barrier (G-C $\rightarrow$ G*-C*) also increases as a function of separation distance. Thus, although the survival lifetime of G*-C* increases dramatically during the process of DNA strand separation, the overall probability of trapping a G*-C* tautomer is probably extremely low. At the equilibrium distance, there is debate over a metastable G*-C* state [115], while there is a substantial energy barrier for larger separation distances.

Consequently, we determine that a direct comparison between the biological timescale and the lifetime of the tautomeric state is misleading when making assertions about tautomeric populations becoming mutations. Furthermore, we suggest that the method of determining the proton transfer rate kinetics needs to be revised since the system is out of equilibrium. Instead, a well-parameterised time-dependent kinetic model is required to describe the low initial population and its subsequent trapping.

In summary, the work here only scratches the surface of describing the biology involved in producing the mutations from the proton transfer mechanism. However, we have gone further than the status quo and laid the path forward to accurately determining the mutation mechanism. Finally, to fully answer whether G-C tautomers lead to the point mutation during DNA replication, we underscore the requirement to combine a time-dependent kinetic model to resolve the competing biological splitting timescale with calculations of the interaction between helicase and the tautomer.

7.5. Methods

7.5.1. Modelling the Separation Process using Density Functional Theory

We model the separation process using DFT methods. We use NWChem 7.0.2 [182] at the B3LYP + XDM / 6-311++G** level of theory. We use the B3LYP exchange-correlation functional [181] with XDM, a non-empirical dispersion scheme [189, 190] to account for long scale dispersion relations we expect to play a more dominant role as the bases are further apart. We pick XDM over other models since it offers greater accuracy and flexibility at reasonable computing cost [191]. Recently, Gheorghiu *et al.* [109] have benchmarked the optimum combined exchange-correlation functional, basis, and dispersion correction and determined that the combination provides fair agreement with higher levels of theory at a reasonable computational expense.

For the DFT calculations, we embed the DNA bases in an implicit continuum solvation model [192, 193, 194] with a low dielectric factor. We use a dielectric factor of $\epsilon = 8.0$ [195, 196], describing the combined influence of the surrounding water molecules and protein interface, which we expect to see when the DNA interacts with the helicase.

We performed an unconstrained geometry optimisation of the canonical and tautomeric forms of G-C using the L-BFGS algorithm [197] implemented in the Atomic Simulation Environment (ASE)

[198, 199]. All the structures were optimised using a force tolerance of $0.01\,\mathrm{eV\,\AA^{-1}}$.

We define the separation reaction coordinate as the distance between the constrained R-groups of the bases (where the base would join onto the rest of the DNA). With this reaction coordinate defined, we can move the bases apart by shifting the bases some distance along the separation coordinate. In reality, the separation reaction coordinate is likely not a straight line due to the interactions with the DNA backbone and enzyme, restricting the movement of the bases. We first limit our system size and focus on getting the QM calculation accurate as a first approximation. Molecular dynamics investigations of DNA duplex separation provide insight into the random nature of these fluctuations and their timescales.

If we allow enough time for the bases to relax during the separation process, we can optimise the geometry of the base at each separation distance. During the optimisation, we apply a constraint to the R-group where the base joins the backbone. The geometry is allowed to relax at each separation distance while the coordinate of the R group atoms is fixed. The constraints prevent the system from drifting back together and simulate the strain imposed on the base from the rest of the DNA separation. The rationale for fixing the R-group is that the separation forces originate from the backbone and propagate to the base via the R-group. We can determine the reaction asymmetry by repeating the calculation for the canonical and tautomeric states.

7.5.2. Obtaining the Reaction Pathway

We obtained the potential energy landscapes describing the proton transfer reactions using a machine learning approach to the classical all-nudged elastic band algorithm (ML-NEB) [157, 158]. The ML-NEB approach minimises the number of DFT single-point energy calculations required to accurately depict the minimum energy path. In our treatment, we collect the movement of the protons transferring (and other atoms moving to facilitate the transfer) into a single axis. The reaction pathway contains a general description of the transfer process; the energetic landscape of this pathway is then explored using ML-NEB. The ML-NEB algorithm incorporates a Gaussian regression model to produce a surrogate description of the accurate minimum energy path. Thus, the uncertainty in the energy points on surrogate minimum energy path becomes the convergence criteria.

ASE [198, 199] was used throughout this work to connect NWChem to Python3 and the ML-NEB algorithm. All pathway calculations are optimised to a force tolerance of $0.01\,\mathrm{eV\,\AA^{-1}}$, with a maximum uncertainty on each image to be $0.02\,\mathrm{eV}$. To increase the resolution of the reaction path while keeping the computational time down, we perform the ML-NEB calculations in two steps. After relaxing the pathway using the 15 images, we interpolate between every image and insert a new image, bringing the total number of images to 29. We then relax the extended pathway, providing a higher resolution of the reaction path.

7.5.3. Molecular Dynamics

Molecular dynamics were performed in GROMACS 2018 [200]. The 14 base pair B-DNA duplex system was constructed from two identical chains of single-stranded DNA with the sequence: $C^3CCACGTACGTGGG^5$. Surrounded in a box of explicit SPCE solvent extending $2\,\mathrm{nm}$ in each cartesian direction, and sodium ions to neutralise the system. The force field used for the DNA was CHARMM36[201]. Several replica systems were minimised, equilibrated, and simulated with a pulling force acting on the backbones of the first base-pair. For each replica the system was first minimised to a maximum force of $12\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$, the equilibration took place over $500\,\mathrm{ps}$ of NVT ensemble with $1\,\mathrm{fs}$ timestep, and a temperature of $310\,\mathrm{K}$ maintained via a Nose-Hoover thermostat with coupling constant of $0.2\,\mathrm{ps}$. In excess of $50\,\mathrm{ns}$ of simulation data was collected and analysed, distributed across 66 replicas with 10 different forces.

To gather statistics on the separation dynamics, the three hydrogen bond length time series of the base pair in question were analysed. The bond length time series were initially passed through a Savitsky-Golay filter with window size 63 and polynomial order 2. A separation metric was defined as the arithmetic mean of the hydrogen bond extensions relative to their equilibrium value. This separation metric was studied across each MD run, in the many instances where the separation peaked above the noise floor, a least squares regression was performed with a linear function whose slope is used to estimate the separation speed of the separation event in question. The fit was limited to the first two Angstroms of separation, as beyond this the hydrogen bonds are deemed to be broken. The uncertainty in the slope of the linear function provides a metric for the quality of the fit. Linear best fits with a negative slope, and those with relative uncertainty above 5 percent were discarded.

To classify the asymmetry behaviour of each separation event, opening angle was defined from the dot product between the vectors connecting the donor/acceptor atoms of each nucleobase. I.e. for guanine: $\mathbf{G} = \overrightarrow{\mathrm{DG{:}N^2\ DG{:}O^6}}$, and for cytosine: $\mathbf{C} = \overrightarrow{\mathrm{DC{:}O^2\ DC{:}N^4}}$. To distinguish between opening with the top bond opening first and bottom bond staying fixed, and vice versa the cross product of the two vectors was calculated and a negative value was applied if it was anti-aligned with the direction of the double helix.

7.5.4. Proton Transfer Asynchronicity

To further analyse how the proton transfer mechanism changes during the base dissociation process we determine the asynchronicity (α) of the double proton transfer. As a concept, asynchronicity is defined by a slight separation of the two proton transfers, i.e. one proton transfers, other heavy ions rearrange and then the second proton transfers. We formally define asynchronicity as

$$\alpha = \frac{|\alpha_{\mathrm{B1}} - \alpha_{\mathrm{B2}}|}{||q_{\mathrm{IRC}}||}. \tag{7.1}$$

Where,

$$\alpha_i = \text{argmax}\left(\frac{\partial x_i}{\partial q_{\text{IRC}}} \cdot \frac{\partial x_i}{\partial q_{\text{IRC}}}\right). \tag{7.2}$$

Here x_i is the Cartesian vector of atom i and q_{IRC} is the reaction coordinate. The partial derivative of the Cartesian vector tracks the motion of, say, B1 or B2 along the reaction coordinate. The dot product normalises the B1 or B2 motion relative to the collective rearrangement of all atoms. If there is no motion of atom i, it does not contribute to the reaction path, then $\alpha_i \to 0$. If $\alpha_{\text{B1}} \sim \alpha_{\text{B2}}$ the protons transfer at the same point on the reaction coordinate, thus the process is synchronous. While when $\alpha > 0$, one proton moves before another, and larger values of α indicate a large separation of the transfer events. In the extreme case when the protons transfer at each opposing end of the reaction coordinate, α tends to unity.

7.6. Acknowledgements

This work was made possible through the support of the Leverhulme Trust doctoral training centre grant number DS-2017-079 and from the John Templeton Foundation grant number 62210. We acknowledge helpful discussions with the members of the Leverhulme Quantum Biology Doctoral Training Centre; particular thanks goes to Johnjoe McFadden. Further thanks go to Antonio Pantelias, who both offered many productive conversations. The authors thank the University of Surrey for access to the Eureka HPC. This work used the ARCHER2 UK National Supercomputing Service. We are grateful for computational support from the UK Materials and Molecular Modelling Hub, partially funded by EPSRC EP/R029431. This work was supported by HECBioSim, the UK High End Computing Consortium for Biomolecular Simulation, which is supported by the EPSRC (EP/L000253/1).

7.7. Data availability

The data presented in the figures of this article are available from the corresponding authors upon reasonable request. The reaction pathways and structures are available on Github.

7.8. Code availability

The analysis source codes are available on Github.

7.9. Author contributions

M.S. and J.A-K. conceived and designed this research, L.S. performed the density functional theory calculations and M.W. the molecular dynamics calculations. All the authors contributed to the

preparation of the manuscript and have approved the final version of the manuscript.

8. Tautomerisation Mechanisms in the Adenine-Thymine Nucleobase Pair during DNA Strand Separation

This chapter is based on the article: Benjamin King, Max Winokan, Paul Stevenson, Jim Al-Khalili, Louie Slocombe, and Marco Sacchi. "Tautomerisation Mechanisms in the Adenine-Thymine Nucleobase Pair during DNA Strand Separation". In: *The Journal of Physical Chemistry B* 127 (Mar. 2023), p. 4220-4228. DOI: https://doi.org/10.1021/acs.jpcb.2c08631 [3] Supplementary information is included in Appendix Section C. This chapter differs from the published work as grammatical errors have been corrected.

My contributions to this project include the design and execution of the Steered Molecular Dynamics (SMD) methodology for the determination of an intrinsic atomistic separation speed of aqueous duplex DNA under replisome conditions. This included: preparation of aqueous DNA structures for simulation; determination and testing of MD and steering force parameters; preparing and overseeing many replica SMD simulations on High-Performance Computing resources; development of novel analysis techniques to determine dynamical characteristics of base pair opening; and discussion of the implication of the SMD results in comparison to literature and the Density Functional Theory results obtained by co-authors Ben King and Louie Slocombe. Additional contributions were made to the writing and proofreading of the article body, preparation of the publications figures, journal cover (Figure 8-1), and table of contents image (Figure 8-2).

Figure 8-1 Cover of The Journal of Physical Chemistry B, Issue 127, March 2023.

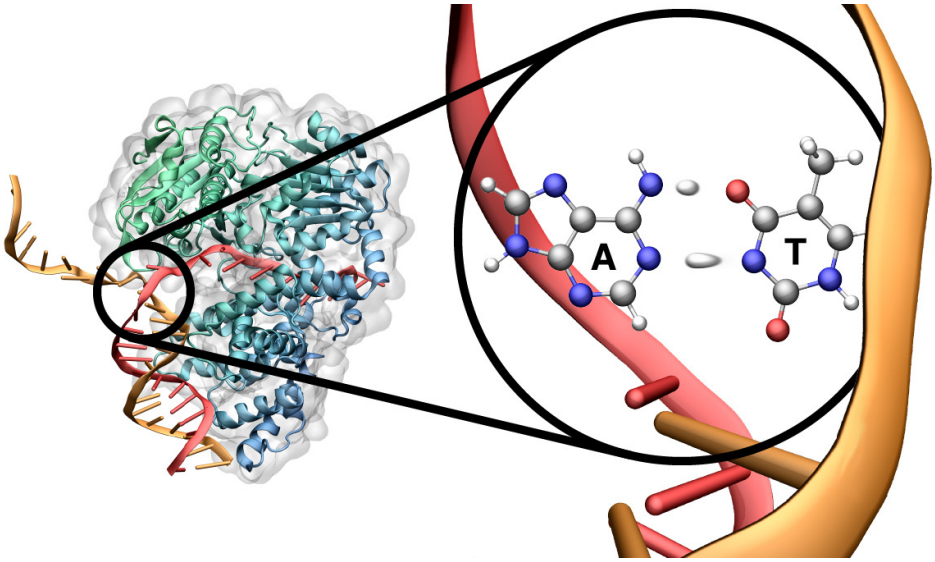


Figure 8-2 Table of contents figure for: Tautomerisation Mechanisms in the Adenine-Thymine Nucleobase Pair during DNA Strand Separation.

8.1. Abstract

The adenine-thymine tautomer (A*-T*) has previously been discounted as a spontaneous mutagenesis mechanism due to the energetic instability of the tautomeric configuration. We study the stability of A*-T* while the nucleobases undergo DNA strand separation. Our calculations indicate an increase in the stability of A*-T* as the DNA strands unzip and the hydrogen bonds between the bases stretch. Molecular Dynamics simulations reveal the timescales and dynamics of DNA strand separation and the statistical ensemble of opening angles present in a biological environment. Our results demonstrate that the unwinding of DNA, an inherently out-of-equilibrium process facilitated by helicase, will change the energy landscape of the adenine-thymine tautomerisation reaction. We propose that DNA strand separation allows the stable tautomerisation of adenine-thymine, providing a feasible pathway for genetic point mutations via proton transfer between the A-T bases.

8.2. Introduction

Spontaneous mutagenesis describes how the genetic code of DNA can incorporate errors without the influence of external factors. Generally, these errors disrupt the canonical (or Watson-Crick) base pairings, adenine-thymine (A-T) and guanine-cytosine (G-C). The root idea that the tautomerisation of DNA may be a mechanism promoting genetic mutation dates back to a short, tentative hypothesis from Watson and Crick's second paper of 1953, where the process of mitosis (the self-replication of DNA) was first theorised[202]. Since then, many studies have investigated the validity of Watson and Crick's claim. Tautomerisation is a phenomenon of structural isomerism not just exclusive to spontaneous mutagenesis but observed in many compounds across organic chemistry (for example, in biological enzymes[203]) that occurs via proton transfer mechanisms. For concision, we take tautomerisation to mean DNA base pair tautomerisation in this paper. The process of tautomerisation proceeds as follows. Each hydrogen bond between the canonical base pairs of DNA depends on the bonding strength of a hydrogen atom to the more electronegative atom on the opposite base (either nitrogen or oxygen). This hydrogen atom is preferentially associated with its donor base, but, via tunnelling or a classical over-the-barrier hopping mechanism, the proton (of the hydrogen atom) may travel along a minimum energy pathway to associate with the acceptor of the opposite base, leaving its electron with the base it was initially covalently bonded to. In this case, each base becomes charged, and the product is a zwitterionic base pair. This mechanism can be described as: A-T $\leftrightarrow$ A^+-T^- and G-C $\leftrightarrow$ G^--C^+. It is also possible that this initial proton transfer will prompt the transfer of a second proton belonging to the other base, with the final product being two neutral bases, each with misplaced hydrogen atoms (see Figure 8-3). This double proton transfer is known as tautomerisation, and each tautomeric base is known as a tautomer. The overall tautomerisation mechanism can be written: A-T $\leftrightarrow$ A*-T*, G-C $\leftrightarrow$ G*-C* (where '*' denotes a tautomer). If the tautomeric form is established when mitosis begins, the tautomerised base pair

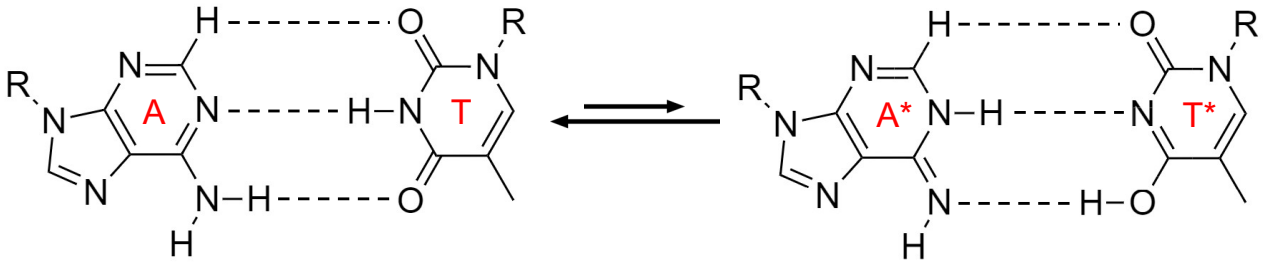


Figure 8-3 The tautomerisation mechanism for the A-T base pair. A = adenine, T = thymine, A* = the tautomer of adenine, and T* = the tautomer of thymine. Here, 'R' indicates where the base binds to the sugar-phosphate backbone of the DNA strand. The dashed lines represent the hydrogen bonds between the base pairs. The forwards-backwards arrow indicates the reversibility of the process, with the larger backwards arrow indicating that the canonical configuration is more stable and preferred energetically.

will be cleaved, preventing each base pair to revert to its respective canonical form. See Figure 8-3 for an illustration of the chemical reaction of tautomerisation for the A-T base pair.

A tautomer can bond in the following pairs (defined by their molecular geometries): A*-C, A-C*, G*-T and G-T*[115]. Due to the geometric similarity between purine-pyrimidine bonding within both canonical and tautomeric pairings, the tautomer will not disrupt the replication process. The tautomeric mismatch can evade correction by the replisome fidelity checks, and a genetic error can be established in two DNA double helices, propagating through subsequent generations of DNA replication. It is also possible for the zwitterionic products of single proton transfer to bond in an *anti-syn* mismatch (where a base bonds in a flipped orientation) to form non-standard base pair configurations[204]. However, these anti-syn products are energetically unfavourable and cannot be incorporated into a full DNA sequence due to sterical repulsion[77, 205]. Therefore, a single proton transfer cannot propagate as a genetic mutation since it will interrupt the process of mitosis and be corrected.

To be of any relevance for genetic mutation, the products of tautomerisation must survive the mitotic replication of DNA. The replication includes the cleaving of the DNA duplex, caused by the enzyme helicase forcing itself between the two strands of DNA and splitting the base pairs that bind the double helix together. The absence of the A-T tautomer in the equilibrium state of DNA, where the double helix is unperturbed by external forces pulling the structure apart, has been well-studied[177, 178, 111, 180], but the existence of tautomers on single, separated DNA strands have not been confirmed. Florian and Leszczynski[177] state that the tautomers must outlive the strand separation timescale of $\sim 100\,\mathrm{ps}$. However, multiple studies have argued that the tautomeric lifetime is much shorter than this timescale[33, 179]. The timescale for which the tautomeric forms of base pairs can exist depends on the stability of the energetic minimum that the tautomeric forms exhibit.

Several computational studies have attempted to measure the minimum energy pathway of the tautomerisation reaction in both A-T and G-C base pairs. These studies have been detailed in the reviews by Kim *et al.*[176] and Srivastava[179], but vary in methodology and, consequently, have produced different conclusions about the stability of the A*-T* tautomeric base pair. Some

studies suggest the configuration is metastable[104, 96], some suggest that there is a shallow energy minimum[126] and others suggest that there exists a substantial energy minimum which is sufficiently stable to be considered a candidate for the spontaneous mutagenesis mechanism[73].

In response to the disagreement classifying the nature of A*-T* stability, Brovarets *et al.*[111] employed Møller-Plesset second-order perturbation theory (MP2) with the 6-311++G(d,p) basis set to model the double proton transfer tautomerisation in both A-T and G-C base pairs. The author reported the lifetimes of the A*-T* and G*-C* base pairs as $6.5\,\mathrm{fs}$ and $160\,\mathrm{fs}$, respectively, revealing that the A*-T* configuration could not exist for long enough in equilibrium to be relevant to spontaneous mutagenesis.

To reduce the computational cost of the simulations whilst retaining much accuracy, Soler-Polo *et al.*[178] implemented a quantum mechanical/molecular mechanical (QM/MM) approach where the main system (the base pair) was computationally modelled with quantum mechanical calculations and a surrounding environment of the DNA double helix, and the aqueous solution was modelled with classical molecular mechanical calculations. This study found that the transition state of the G-C tautomerisation is asymmetric; the state G*-C* is less stable than the state G-C.

A study of the tautomerisation process in both A-T and G-C base pairs was conducted by Slocombe *et al.* and confirmed the transition state calculations in the G-C base pair of Soler-Polo. Using density functional theory (DFT), Slocombe *et al.* were able to discern the energy barriers of the G-C and A-T base pairs. It was discovered that the G-C base pair allows a stable minimum for the canonical and tautomerisation configurations with the same asymmetry of the transition state as Soler-Polo. However, the A-T base pair, while having a stable canonical configuration, has only a metastable tautomeric configuration, in agreement with multiple prior studies.

A limitation in each of the studies above is that calculations were performed on the DNA structure in equilibrium, where the strands are not undergoing separation as they would during replication. However, as proposed by Florian and Leszczynski[177], the tautomeric state must be stable enough to survive the DNA cleavage within the mitosis scheme, hence, the separation of DNA strands must be considered in order to validate proton transfer between bases as a genetic mutation mechanism.

In a recent letter by Winokan *et al.* [1], tautomerisation of the G-C base pair while undergoing DNA strand separation was investigated using DFT and molecular dynamics (MD) approaches. The conclusions of the DFT investigations determined that the tautomerisation energy barrier increases as the strands move away from one another, as would be expected. The motion of the bases within this process was also given rotation degrees of freedom around a fixed atom of the base bonded to the sugar-phosphate backbone. It was determined that the hydrogen bonds between the bases extended quasi-linearly during DNA strand separation and that the bases rotate to preferentially limit the extension of the upper and lower, but not central, hydrogen bonds between the G-C base pair. Additionally, the MD investigation provided estimations for the timescale of DNA stand separation. The authors concluded that the lifetime of the tautomeric state must outlive $1.7\,\mathrm{ps}$ - a speed two

orders of magnitude smaller than quoted by Florian and Leszczynski[177].

In this paper, we model the tautomerisation of A-T during DNA strand separation. To do this, we use DFT to calculate the stretching of the hydrogen bonds B1 and B2 and between A-T and the opening angle, θ, of the base separation (see Figure 8-4) as a function of strand separation distance and compute the transition states of the tautomerisation reaction across a range of strand separations. MD was employed to investigate the dynamics of A-T tautomerisation and estimate the timescale of the reaction. A detailed description of the methods used in this investigation can be found in the Supplementary Information document. The MD trajectories reveal a significant variation in the separation dynamics with profound physical implications for the tautomerisation of A-T.

8.3. Method

The DNA strand separation process is modelled by density functional theory (DFT) at the B3LYP + XDM / 6-311++G** [181, 189] level of theory. The software conducting this method is NWChem 7.0.2 [182]. We select the combination of B3LYP [181] (exchange-correlation functional) with XDM [206, 207] (non-empirical dispersion scheme) and 6-311++G** (basis set incorporating dispersion and polarisation corrections) to satisfy accuracy requirements while allowing for a reasonable computational cost. The precedent for this combination of factors comes from a recent study by Gheorghiu [109], who has optimised the combination of methodological approaches to attain the level of accuracy required in the multi-scale modelling of DNA processes. The DNA system that the calculations pertain to is a single A-T base pair (depicted by A1-T1 in the scheme of Figure 8-4) which exists within an implicit continuum solvation model [192, 193, 194] with a low dielectric factor ($\epsilon = 8.0$) [195, 196]. The low dielectric factor is in concordance with the proximity of helicase and the aqueous solution of water molecules in which the system is embedded.

We impose non-equilibrium DNA strand separation across thirteen systematic increments. The separation of the strands is measured by the summed distance that the R-groups are displaced from equilibrium. The optimisation of the atomic geometries within the base pairs was conducted using the L-BFGS method [197]. We use the Atomic Simulation Environment (ASE)[198, 199] to implement the L-BFGS algorithm in order to optimise the atomic geometries and allow the system to 'relax' with a force tolerance of $0.01\,\mathrm{eV\,\AA^{-1}}$. During the relaxation, we fix the location of the R-group of the base to have control over the definition of the DNA strand separation. The bases are also allowed to rotate about the fixed nitrogen atoms.

We calculate the transition state of the proton transfer reactions from the canonical configuration to the tautomeric configuration of the A-T base pair across fifteen snapshot images for the first eight increments of the DNA strand separation (from $0.0\,$Å to $1.56\,$Å separation) using a machine learning nudged elastic band (ML-NEB) [157, 158] approach. ML-NEB seeks to minimise the energy pathway

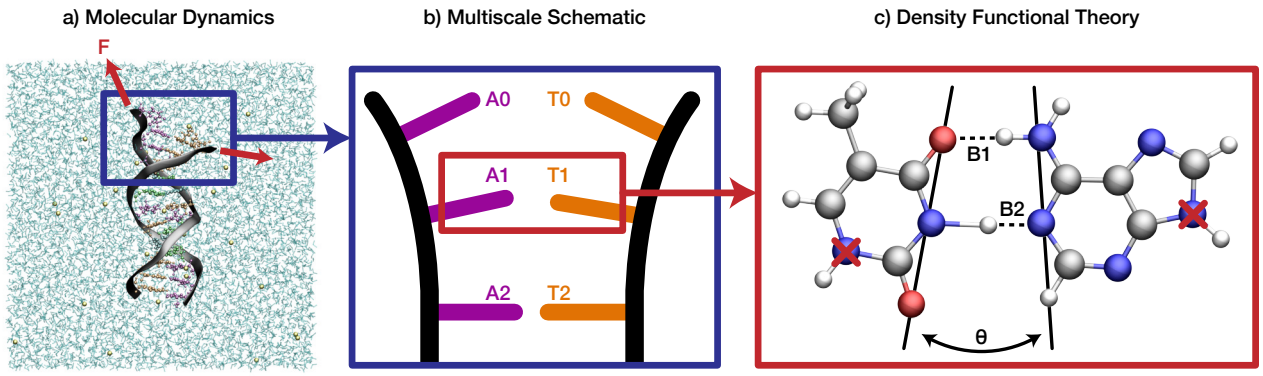


Figure 8-4 The scheme modelling DNA strand separation for the canonical form of the adenine-thymine base pair. In the MD simulation (a), the two DNA strands are forced apart by a constant steering force mimicking the action of helicase[18]. The base pairs A0-T0 and A2-T2 are superfluous to the DFT calculations, as were the sugar-phosphate backbone strands, which are both contained within the molecular mechanical region. DFT calculations were applied to the quantum mechanical region of the A1-T1 base pair (c) and modelled the extension of the two hydrogen bonds B1 and B2 and the transition state of an asynchronous double proton transfer reaction along the bonds B1 and B2 with an intermediate state of $A1^{+}$-$T1^{-}$. A1 and T1 comprise the chemical formula $C_{10}H_{11}N_7O_2$. The A1-T1 bond opens under this action at an angle of θ. The nitrogen atoms with lock icons were fixed to the sugar-phosphate DNA backbone and served as the point around which the bases could rotate.

of the reaction to obtain the transition state for each separation. The bases are permitted to rotate around the fixed nitrogen atoms. In this way, the bonds B1 and B2 can extend with the DNA strand separation, but not necessarily at equal rates since the bases rotate individually, limiting the extension of one bond and accelerating the extension of the other. This feature is measured by obtaining the respective bond stretching for B1 and B2 as a function of DNA strand separation. We compute the reaction asymmetry to demonstrate the asynchronicity of the reaction pathway. The ML-NEB algorithm's advantage is minimising the number of single-point DFT calculations to determine the energy and forces to obtain a minimum reaction energy path. An implicit assumption in transition state theory that we implement is that the reaction pathway obtained is a one-dimensional potential energy surface; therefore, the proton transfer occurs across a single dimension (the reaction coordinate along the minimum energy path) as a first approximation. ML-NEB utilises a Gaussian regression model to reconstitute the full minimum energy path. This renewed minimum energy path is described as a surrogate minimum energy path. This surrogate minimum energy path's convergence criteria are derived from the data points' uncertainty along its energy pathway. The reaction path is calculated and optimised with a force tolerance of $0.01\,\mathrm{eV\,\AA^{-1}}$ and a maximum energy uncertainty of $0.02\,\mathrm{eV}$. Mathematically, we define the reaction path as the magnitude of two vectors: the difference between the vectors between the hydrogen donor and hydrogen acceptor in bonds B1 and B2. Throughout this investigation, the NWChem software was connected to Python3 *via* ASE.

We use molecular dynamics (MD) to investigate the details of the DNA strand separation process. These calculations were performed in Gromacs version 2018[208] with the CHARMM36 (March 2019) force field [201] and SPCE water model [209]. The MD system is constructed of 14 base pairs within a double strand of DNA, initially in equilibrium and an aqueous environment. For

each replica simulation, the system is minimised to $12\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$, before being placed in an NVT ensemble at $310\,\mathrm{K}$ where a $500\,\mathrm{ps}$ equilibration is conducted with a $1\,\mathrm{fs}$ timestep. Following successful equilibration, a production $200\,\mathrm{ps}$ MD trajectory is simulated with an optional constant steering force between the backbones of the terminal A-T base pair of the DNA duplex. The steering force, if present, was applied along the vector connecting the centres of mass of the nucleotide's backbone atoms with the force encouraging their separation. In addition to replicas where no force was applied, five other steering forces were investigated ranging from $25\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$ to $125\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$. The strength of the biasing force was chosen to be of the same order of magnitude as the hydrogen bonds binding the A-T pair. In total, over 34 nanoseconds of MD trajectories across six force constants and 49 replicas were analysed.

We calculate the occurrences of the opening angles θ (see Figure 8-4) across the range of DNA strand separation $0.0\,\text{Å}$ – $2.0\,\text{Å}$ and estimate the speed at which the strand separation occurs. In the quantum mechanical investigations, the opening angle smoothly increases during strand separation. In the MD calculations, we examine whether this observation remains true in the biological ensemble. We study the opening angle in two scenarios: where the B1 bond is the first to open and where the B2 bond is the first to open. We collect a histogram of opening angles and their occurrences across a 75° range for each scenario of B1 and B2 opening first. We treat a positive angle as opening from the B2 end of the base pair and a negative angle as opening from the B1 end of the base pair.

In both DFT and MD calculations, we simulate an environment contained within a dielectric solvent shell. Therefore, the environmental interaction with the base pair system is consistent between both methodologies.

8.4. Results

The results of the DFT investigation are comprised of data that are each a function of DNA strand separation: hydrogen bond stretching (Figure 8-5), the A-T tautomerisation transition state (Figure 8-6a), the reaction asymmetry (Figure 8-6b), and the forwards and reverse energy barriers of each concerted proton transfer (Figure 8-6c).

The results of the MD investigation consist of a histogram of the statistical ensemble of angle occurrences (Figure 8-7) and the separation speeds of the DNA strands for each simulated helicase pulling force (Figure 8-8).

8.5. Discussion

In both the canonical and tautomeric configurations of A-T, the bond extension under strand separation (see Figure 8-5a and 8-5b) is comparable for both individual cases of the extension of the

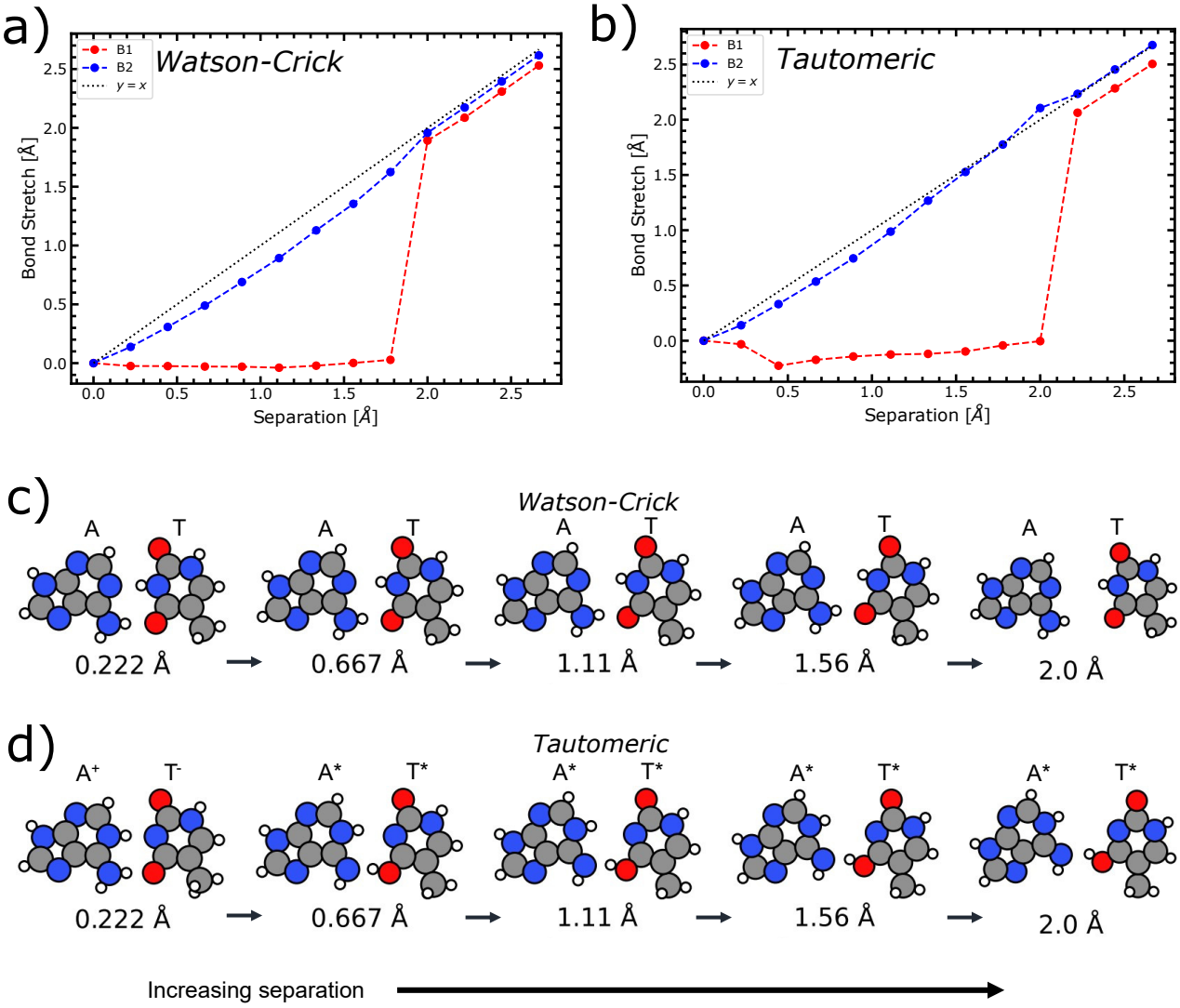


Figure 8-5 The results of the DFT model compute the scheme by which the base pair A1-T1 (see Figure 8-4) separates in relation to the base rotation and the stretching of the hydrogen bonds. (a and b) The stretching of the bonds B1 and B2 in the A1-T1 base pair as a function of the DNA strand separation distance for both the a) Watson-Crick canonical configuration of the A1-T1 base pair and the b) tautomeric A1-T1 base pair. (c and d) A five-step incremental molecular depiction of the A-T base pair under DNA strand separation for c) the Watson-Crick canonical configuration and d) the tautomeric configuration. The numerical labels, in angstroms, represent the distance separation of the DNA strands. In this regime, the distance 0.0 Å represents the DNA strands in equilibrium separation. The right-hand arrows define the progression of the strand separation. In the first image of panel d), there is no stable double proton transfer mechanism for tautomerisation, and the product of single proton transfer is zwitterionic, hence the labels 'A^+' and 'T^-'.

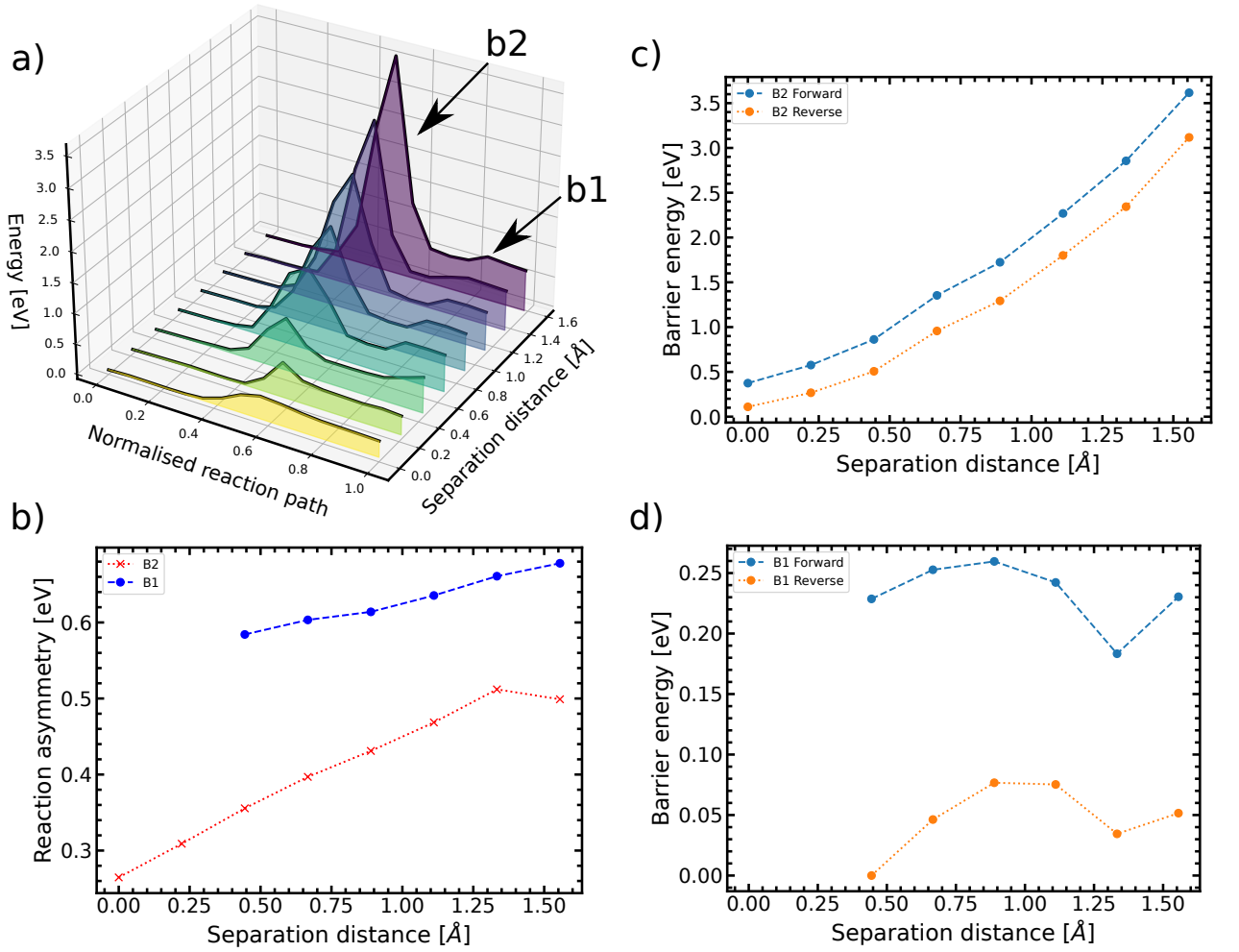


Figure 8-6 (a) The minimum energy pathway of the adenine-thymine tautomerisation reaction as a function of normalised reaction path and DNA strand separation distance. Double proton transfer becomes stable at the third DNA strand separation increment, although this is difficult to see on the plot. (b) The reaction asymmetry defines the difference between the energy of the product (tautomeric state) and the energy of the reactant (canonical state). There is a single proton transfer product for all data points and a double proton transfer product after the third data separation point. (c) and (d) The reaction energy barrier between the forward and backward tautomerisation reaction is plotted as a function of separation distance for both single and double proton transfer tautomerisation. The two proton transfer events along bonds B1 and B2 are asynchronous.

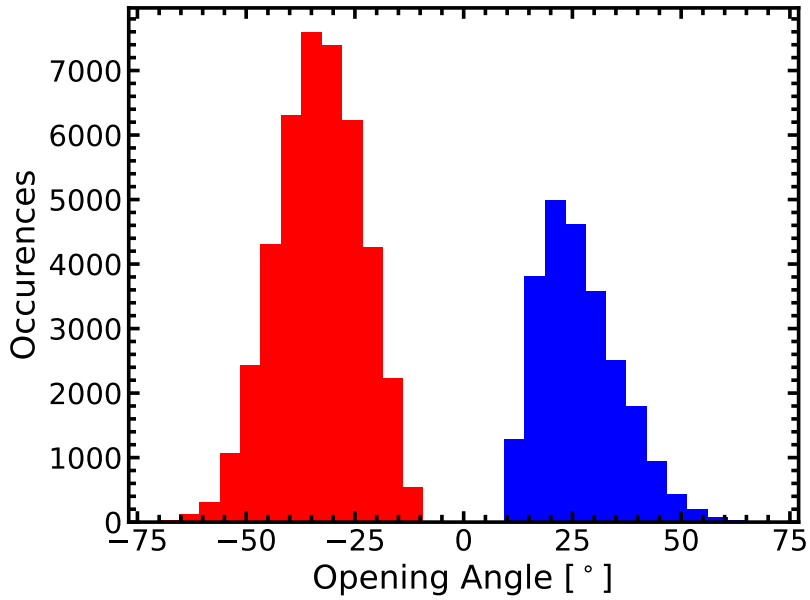


Figure 8-7 The histogram of opening angles against the occurrences corresponding to them across the range of DNA strand separation represents the energetic preferences of each opening angle summed across the range of DNA separation. The positive angles represent the opening of the base pair starting at bond B2, and the negative angles represent the opening of the base pair starting at bond B1.

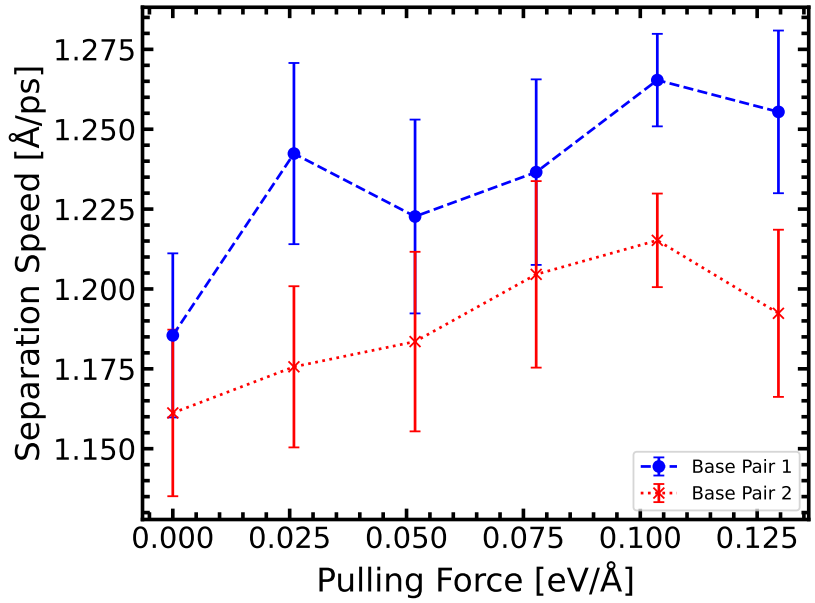


Figure 8-8 The DNA strand separation speed instigated by the simulated helicase pulling force. Base Pair 1 refers to the base pair A1-T1, and Base Pair 2 refers to the base pair A2-T2 (see Figure 8-4). The separation speed data points provided are the arithmetic mean of 210 MD simulations (with the standard error providing the error bars) produced for six simulated helicase pulling forces.

two hydrogen bonds holding the base pair together, defined (see Figure 8-4) as B1 and B2. Bond B1 extends at a much more gradual rate than B2, showing that the bases are rotating with respect to one another. In both canonical and tautomeric cases, the rotation of the bases preferentially limits the extension of the B1 bond. We observe that bond B2 effectively stretches at the same rate at which the DNA strands are separated. The implication of this is that the bond B1 is the stronger bond of the two. Another similarity between the canonical and tautomeric configurations is that the opening angle between the base pair grows steadily until a cut-off point (which is different for each configuration) where the angular interaction of the bases reverts to approximately the original respective opening angle, where $\theta = 0$. This reversion occurs at a lower separation distance for the canonical configuration (~ 1.8 Å) than for the tautomeric configuration (~ 2.0 Å). By establishing that the bond B1 appears to be stronger than B2, one can go further to say that the tautomeric configuration of the bond B1 is stronger. Figure 8-5b confirms that the tautomeric version of B1 is stronger since, initially, the bond shortens, momentarily forcing an increase in the rate of angular opening. Once the angular reversions visible in Figure 8-5a and 8-5b have occurred, both bonds extend in both the canonical and tautomeric configurations at a very similar rate. This rate is similar to the rate of strand separation.

In Figure 8-5c and Figure 8-5d, the molecular structures of a subset of the strand separation mechanism are presented. One can observe the gradual increase in θ complementing the bond stretching patterns visible in Figure 8-5a and 8-5b. Two key features in Figure 8-5c and Figure 8-5d are the last image of the Watson-Crick sequence and the first image of the tautomeric sequence. In the last Watson-Crick image (at 2.0 Å separation), the angle of base separation has reverted to $\theta = 0$ - the bond stretching is no longer asymmetric. This reversion to roughly the equilibrium angle is observed in Figure 8-5a. The first tautomeric image represents only a single proton transfer; hence the product is a zwitterionic base pair. In the first two data points (at a strand separation of 0.0 Å and 0.22 Å), there is no stable energy minimum for the double proton transfer tautomeric form. Therefore, for the A*-T* configuration to be biologically relevant as a mutation mechanism, the DNA strands must first be separated by some distance as a criterion of spontaneous mutagenesis. Beyond this paradigm-shifting distance, the tautomeric configuration becomes more and more stable due to the potential energy surface having an increasing energy barrier. However, this energy barrier increase also requires greater energy to overcome. Therefore, we can conclude that the tautomeric form A*-T* becomes more stable under DNA separation but at the cost of the double proton transfer reaction becoming more unlikely.

Figure 8-6a represents the previous conclusion - the energy barrier between the canonical and tautomeric state grows as the distance between the DNA strands increases. However, while there is no energy minimum at the maximum coordinate on the normalised reaction path axis for the first two separation increments, an energy minimum becomes visible for larger separation distances, representing a stable state for the A*-T* tautomer. Also, in Figure 8-6a, one can observe the de-

velopment of a central energy minimum for the third separation increment onward. This energy minimum is created by the intermediate zwitterionic state. We observe that the energetically preferable double proton transfer mechanism is asynchronous (a step-wise transfer mechanism) and consists of a single proton transfer from the thymine base to create the zwitterionic intermediate A^+-T^- followed by an induced, second single proton transfer from the adenine base. The height of the second energy barrier is much lower than the first (across all strand separations and increasingly so as strand separation increases), showing that it is much more energetically favourable for the zwitterionic product to undergo single proton transfer again to produce the A*-T* tautomeric state, rather than a single proton transfer which reverts to the canonical state.

Figure 8-6b shows the quasi-linear reaction asymmetry of both the single and double proton transfer reactions. The reaction asymmetry depicts the variation in the stability of both the single and double proton transfer products in relation to the canonical state. One observes that the canonical state is the most stable. Figure 8-6b shows that the base pair structure at each increment has access to a single proton transfer minimum, but this is not the case for the double proton transfer, which can only begin to obtain stability of the double proton transfer tautomer at $\geq 0.444\,\text{Å}$ strand separation (the third imposed strand separation increment).

Figure 8-6c and d represent the change in barrier heights for transfer as a function of separation distance. The double proton transfer occurs in two single proton transfer stages, initiated by a proton donation from thymine to adenine. This first stage concerns the transfer of the proton in bond B2 from the nitrogen of thymine in B2 to the nitrogen of adenine in B2 and has a higher energy barrier across all increments of the strand separation. Initially, there is no response of a second proton transfer for the first two increments of strand separation, and the single proton transfer is all that occurs. Since there is no second proton transfer up until $0.444\,\text{Å}$, the first two data points for both 'B1 Forward' and 'B1 Reverse' are $0\,\text{eV}$. However, the energy barriers of the second proton transfer grow in response to the gradual creation of a stable minimum for the double proton transfer tautomeric state. Since this second proton transfer is instigated by the first, it has a lower energy barrier and grows at a much slower rate across the strand separation compared to the steep linear increase of the initial single proton transfer.

There have been indications in previous literature that the tautomerisation transition state is significantly affected by the environment the base pair exists within conducted by, for example, Li *et al.*[210]. In their study of the particular case of DNA base pair tautomerisation occurring in the wobble mismatch geometry wG-T, Li *et al.* report that the presence of an aqueous environment or the presence of a DNA duplex makes the transition state of a wG-T→G-T^* reaction (where 'wG-T' indicates a 'wobble' mismatch bonding) slightly more endoergic. In contrast, the presence of a DNA polymerase enzyme makes the transition state slightly more exoergic. Therefore, exact and realistic environmental modelling can either marginally stabilise or destabilise the tautomer.

For the molecular dynamics investigation, we find that the DNA strand separation occurs at a

speed of $\sim 1.25\,\text{Å}\,\text{ps}^{-1}$. If we assume that, at a separation distance of $2.0\,\text{Å}$, the tautomerisation reaction does not reverse back to the canonical state, the lifetime of the tautomer need only exceed $\sim 1.6\,\text{ps}$. This assumption is very reasonable since, at separations $> 1.56\,\text{Å}$, the reverse energy barrier would require temperatures of $> 3.5 \times 10^4\,\text{K}$ ($E = k_B T$, where k_B is the Boltzmann constant). Such high temperatures are not biological. Therefore, the tautomers can be trapped on separated DNA strands, surviving the mitotic division, ready to incorporate errors in the genetic code through further generations of mitosis.

Figure 8-7 shows the bimodal statistical distribution of the opening angles, θ, for the MD simulations across the DNA strand separation range. For bond B2 opening first, the process favours an opening angle of $\sim 20°$, showing that the most energetically stable mode by which this takes place is when this angle is observed. The second scenario shows the process favours an opening angle of $\sim -35°$, corresponding to the initial opening bond being B1. Across both scenarios, the most energetically favourable configuration of the DNA strand separation overall is at an angle of $\sim 35°$ with the DNA unzipping from the B1 bond (i.e. the strand in Figure 8-4 unzipping from the bottom upwards). Figure 8-7 shows that favouring $\sim -35°$ as an opening angle is universal and cumulative across the DNA separation process.

8.6. Conclusions

Using DFT calculations, we propose that the double proton transfer tautomer of the adenine-thymine base pair (A*-T*) can exist in an energetically stable state on the condition that the DNA duplex is separated by more than $0.444\,\text{Å}$. As the DNA strands separate further from one another, the A*-T* state becomes more stabilised due to the increase in the energy barriers of the tautomerisation reaction. However, this increasing energy barrier also diminishes the probability of the tautomerisation reaction occurring. Before the separation that stabilises the A*-T* state is reached, only single proton transfer products are stable, but these reactions are not biologically relevant to the spontaneous mutagenesis process. However, the double proton transfer product is biologically relevant to the spontaneous mutagenesis process as its tautomeric nucleobases can bond in non-canonical pairings that evade replisome fidelity checks. The conclusion that the A*-T* base pair can be a genetic mutation instigator is contrary to several previous studies[177, 104, 96] which have only examined the tautomerisation of A-T while the DNA duplex is in an equilibrium state, finding that the state A*-T* at equilibrium is metastable and therefore biologically irrelevant.

Using MD calculations, we find that the mode of DNA strand separation that is energetically favoured is where the helicase enzyme forces open the B1 bond first at an angle of $\sim 35°$. We also find that the minimum lifetime of the tautomer to be mechanistically feasible for genetic mutation is $\sim 1.6\,\text{ps}$. This minimum lifetime of the A*-T* base pair is two orders of magnitude shorter than previously suggested by Florian *et al.*[177] and in agreement with the minimum lifetime of the G*-C*

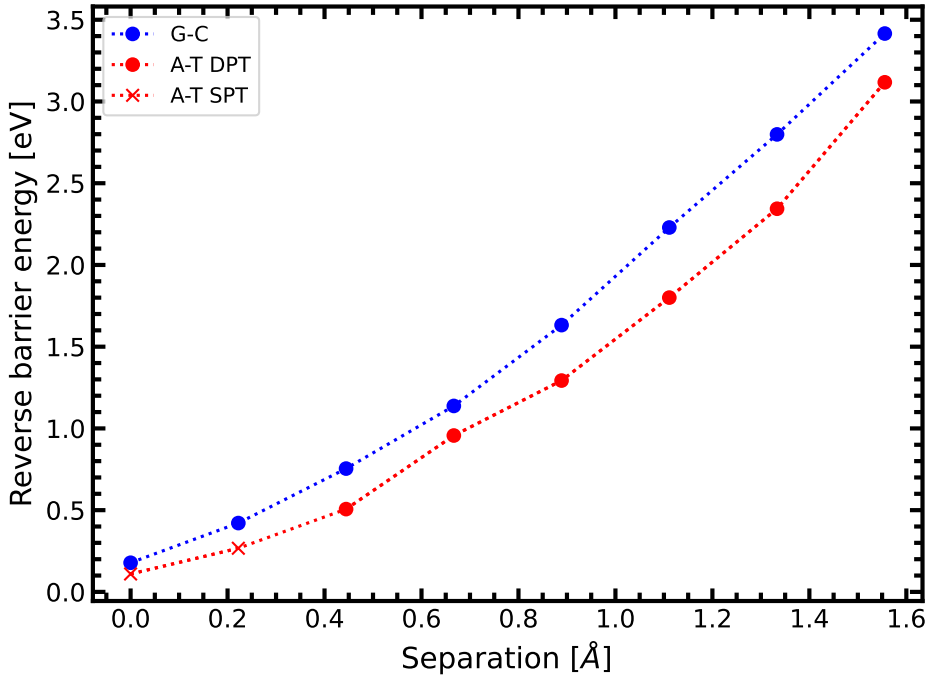


Figure 8-9 The reverse energy barrier of the first proton transfer as a function of the DNA strand separation. The A-T reverse energy barrier data is split into two portions: SPT (single proton transfer), of which the product is the zwitterionic A^{+}-T^{-} state, and DPT (double proton transfer), of which the product is the tautomeric base pair state A^{*}-T^{*}.

base pair suggested by Slocombe *et al.*[1] of $1.7\,\mathrm{ps}$.

A representation of the stability of the tautomers (reverse energy barrier of the tautomerisation reaction) G*-C* and A*-T* is shown in Figure 8-9. We observe that the reverse energy barrier of both tautomers increases during strand separation at a similar rate. The relation between separation distance increase and energy barrier increase is close to quadratic ($y \propto x^{1.894}$ for A-T tautomerisation and $y \propto x^{1.783}$ for G-C tautomerisation). By asserting that the stability of the tautomeric product is associated with the height of the tautomerisation energy barrier, we can say that the states G*-C* and A*-T* are very similar in energetic stability from $0.44\,\text{Å}$ where A*-T* begins to be formed.

In this paper, we examine the mechanical effect of the separation process on A-T tautomerisation facilitated by the helicase enzyme. It is possible that the helicase micro-environment might affect the reaction free-energy to stabilise or destabilise the A-T tautomer. However, it still needs to be determined to what extent the helicase is simply a mechanical device that forces the DNA strands apart or whether the helicase provides a significant electrostatic interaction to alter the energy landscape of A-T tautomerisation. Two points speak in favour of this paper. Firstly, the base pair A1-T1 is enclosed in a molecular environment between A0-T0 and A2-T2, which one would expect to shield the quantum mechanical region from the electrostatic potential of the helicase enzyme. Secondly, the quantum mechanical calculations at equilibrium with no induced separation in Figure 8-6 are in excellent agreement with those of Gheorghiu *et al.*, who consider the environmental effect of water complexes in DNA proton transfer.

To the best of our knowledge, our theoretical study provides the first strong argument against

the widely-held belief that adenine-thymine tautomers are not relevant to spontaneous mutagenesis. By incorporating the dynamical process of DNA strand separation - a key step in the mitotic replication process and much more realistic than a static DNA picture - we show that adenine-thymine tautomerisation is just as relevant in genetic mutation mechanisms as guanine-cytosine tautomerisation.

8.7. Data availability

The data presented in the figures of this article are available from the corresponding authors upon reasonable request. The reaction pathways and structures are available on Github.

Author Contributions

L.S., M.S., and J.A-K. conceived and designed this research, L.S. and B.K. performed the density functional theory calculations and M.W. the molecular dynamics calculations. P.S. assisted in the preparation of the manuscript, specifically critical review, commentary and revision. All the authors contributed to the preparation of the manuscript and have approved the final version of the manuscript.

9. Multiscale simulations reveal the role of PcrA helicase in protecting against spontaneous point mutations in DNA

This chapter is based on the article Max Winokan, Louie Slocombe, Jim Al-Khalili, and Marco Sacchi. "Multiscale simulations reveal the role of PcrA helicase in protecting against spontaneous point mutations in DNA". In: *Scientific Reports* (2023). DOI: https://doi.org/10.26434/chemrxiv-2023-x4rl5 [4]. The article has been passed peer review and has been accepted for publication in Scientific Reports. Supplementary information is included in Appendix Section D.

My co-authors Louie Slocombe, Jim Al-Khalili, and Marco Sacchi contributed to the conception and design of the research, as well as writing and proofreading of the manuscript text.

9.1. Significance Statement

Proton transfer between the DNA nucleobases results in rare tautomeric forms that can cause mutations. Previous works have addressed the energy requirements of GC and AT tautomerisation but have failed to capture the effect of the enzymatic microenvironment in the model of DNA replication. In this study, we calculated the free energy profiles for the double proton transfer in Guanine-Cytosine in the presence of a bacterial PcrA Helicase enzyme, which plays a crucial role in unwinding DNA. We conclude that the presence of the Helicase destabilises the G^*C^* tautomer, potentially offering evolutionary protection from the proton transfer events.

9.2. Abstract

Proton transfer across hydrogen bonds in DNA can produce non-canonical nucleobase dimers and is a possible source of single-point mutations when these forms mismatch under replication. Previous computational studies have revealed this process to be energetically feasible for the guanine-cytosine (GC) base pair, but the tautomeric product (G^*C^*) is short-lived. In this work we reveal, for the first time, the direct effect of the replisome enzymes on proton transfer, rectifying the shortcomings of existing models. Multi-scale quantum mechanical/molecular dynamics (QM/MM) simula-

tions reveal the effect of the bacterial PcrA Helicase on the double proton transfer in the GC base pair. It is shown that the local protein environment drastically increases the activation and reaction energies for the double proton transfer, modifying the tautomeric equilibrium. We propose a regime in which the proton transfer is dominated by tunnelling, taking place instantaneously and without atomic rearrangement of the local environment. In this paradigm, we can reconcile the metastable nature of the tautomer and show that ensemble averaging methods obscure detail in the reaction profile. Our results highlight the importance of explicit environmental models and suggest that asparagine N624 serves a secondary function of reducing spontaneous mutations in PcrA Helicase.

9.3. Introduction

The reliable nature of the DNA base pairing rules discovered by Franklin, Watson, and Crick [7, 211] is crucial to maintain accurate replication of genetic material. A low but finite rate of DNA mutations allows for the gradual evolution of the genome, as well as causing genetic diseases. Spontaneous mutations can occur when the canonical Watson and Crick (WC) pairing is disturbed, causing a mismatch to be formed. Such a mistake can be made permanent through replication, thus altering the genome.

During replication, duplex DNA is unwound and has its strands separated to form two single-stranded templates by a helicase enzyme. Each template is assembled into a new double-stranded molecule via a polymerase enzyme. Polymerase enzymes have remarkable reading fidelity and error-correction mechanisms and are crucial in ensuring genome stability. The reliable nature with which guanine binds to cytosine, and adenine to thymine, is at the core of guaranteeing this stability. Mismatches such as guanine-thymine (G-T) do not form a Watson and Crick-like dimer and are thus rejected by the polymerase [56, 212]. However, proton transfers can create rare tautomers that may evade these cellular error-detecting facilities by mimicking the WC geometry.

One pathway for creating point mutations is proton transfer (PT) across the nucleotides' hydrogen bonds. For example, in its canonical (standard) configuration, the guanine-cytosine (GC) dimer is in amino-keto form. A double proton transfer (DPT) across two of its three hydrogen bonds results in a metastable state known as a tautomer and is written as G^*C^*. The reaction from GC (amino-keto) to G^*C^* (imino-enol) is shown in Figure 9-1.

The tautomer G^*C^* causes both nucleotide bases to have an altered hydrogen bonding profile, which no longer matches their canonical counterparts. This modified steric profile causes problems during DNA replication as a tautomer in the template strand will not match using the WC pairing, instead leading to mismatches. The tautomeric form, G^*C^* is 'Watson-Crick'-like because it leaves the secondary structure of the DNA duplex unaffected and can thus evade error-correction mechanisms. While a guanine nucleobase in the template strand would canonically be matched with a cytosine base, the imino form of guanine (G^*) pairs instead with thymine, causing a point mutation

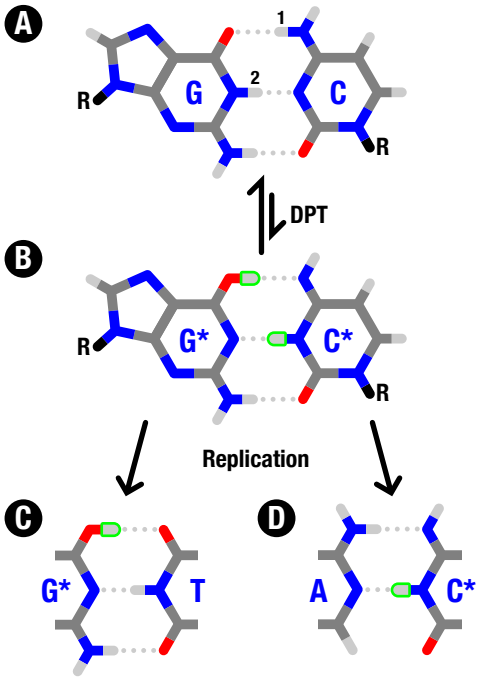


Figure 9-1 A mechanism for creating spontaneous point mutations in which a canonical GC base pair (A) undergoes a double proton transfer (DPT) to an intermediate G*C* tautomer (B) before a helicase enzyme separates the dimer allowing for the creation of the G*-T (C) and A-C* (D) mismatches by a polymerase enzyme. Highlighted in green are the transferred hydrogen atoms/protons forming the nonstandard structures. In the canonical GC base pair (A) the two protons involved in the DPT are labelled "1" and "2".

in the genome as the letter C is replaced with T. Similarly, for the cytosine (enol) tautomer, the genetic letter G is replaced by A.

Pivotal to the feasibility of the WC-like point mutation hypothesis is that the tautomeric dimer can survive in the noisy cellular environment at biological temperatures - and crucially - in the presence of the replisome enzymes. Furthermore, significant nuclear quantum effects have been demonstrated for this reaction, indicating that quantum tunnelling facilitates the DPT in DNA [213, 117] and thus may contribute to genomic variation [176].

Recent Density Functional Theory (DFT) calculations and hybrid quantum mechanics / molecular mechanics (QM/MM) calculations have revealed the energy landscape governing the DPT in canonical DNA base pairs. Several computational studies of DPT in DNA base pairs show that the contribution of DPT to the production of rare tautomeric forms of DNA base pairs is negligible in AT but not in GC due to the differences in their reaction energy profile. The shallow local minimum corresponding to the tautomeric state of AT present in the potential energy surface [73, 126, 104, 125, 178, 205] vanishes when accounting for free energy corrections [104]. Consequently, AT has been dismissed as a candidate for spontaneous point mutations [214, 122]. In GC dimers, the depth of the tautomeric well has been studied with [178] and without [215, 178] free energy considerations. However, in the free energy surface, the tautomeric well is, crucially, still present [104]. Therefore the GC base pair is the more likely candidate for single-point mutations via the tautomerisation process [213]. Typically, other authors have quoted activation energies in the range

0.42-0.63 eV and reaction asymmetries of 0.29-0.54 eV for the DPT in GC, and both synchronous and asynchronous transfers have been observed[184, 1, 109, 216]. The preference for the two protons to transfer in a stepwise fashion appears with more realistic models of the environment that include larger duplex DNA structures such as Angiolari, *et al.*[216] and Gheorghiu, *et al.*[115].

Whether the tautomers produced by DPT can have a lasting biological impact by forming WC-like mismatches depends on the role of the replisome in their formation and stability [184, 213, 111, 115, 176]. But, to date, no computational investigation has characterised the DPT in the presence of replisome enzymes. This work aims to determine the plausibility of DPT in a helicase-DNA complex.

For replication, the DNA has to unwind and the two strands need to separate. These tasks are performed in cooperation by the topoisomerase and helicase enzymes. One of the most studied and thus best-understood helicase enzymes is that of the gram-positive bacterium named the plasmid copy-reduced (PcrA) helicase. The PcrA helicase belongs to the helicase superfamily 1 and contains a single protein chain, with 649 amino acids forming a monomer with four domains of interest. There have been several experimental and computational studies of the PcrA helicase, revealing stepping motor dynamics [15, 16, 173, 18, 174, 17], roles of individual residues in the ssDNA binding site [20], and the details of the ATP cleavage cycle [19].

The topology of the PcrA resembles a torus (see panel B of Figure 9-2), with a single strand of DNA being translocated through, pushing apart the duplex. Figure 9-2 shows the single-stranded DNA (ssDNA) binding site of PcrA helicase and the end of the DNA duplex. Slocombe *et al.* [1, 3] proposed that DNA bases about to be split by the helicase enzyme encounter a unique environment that modifies the energetics of the tautomerism but did not include an explicit enzyme model. The last base pair of the duplex (base pair N with residues DG662 and DC701) is the last chance for the DPT to occur in the PcrA Helicase complex before the strands are separated. In this work, we model the DPT within a GC base pair embedded in two contrasting biological scenarios: (1) in aqueous duplex DNA and (2) at the entrance to the ssDNA binding site of PcrA helicase, in order to reveal the effect an explicit replisome environment has on the DPT. This work aims to untangle the mechanical and chemical interactions between the helicase and the DNA bases. Additionally we will investigate the effect of the hydrogen bond stretching on the proton transfer.

9.4. Results

We employ QM/MM to model the proton transfer reaction profile between the hydrogen bonds of DNA. Within the QM/MM framework, we use umbrella sampling (US) to obtain the reaction free energy from an ensemble of steered dynamical trajectories where the system is harmonically restrained to a given reaction coordinate value. Details on the restraints are given in Sections 1.1-1.4 and Figure S1 of the supplementary information (SI). In Sections 2.1-2.2 and Figures S4-S5 of the SI we validate our level of theory. Initially, we use US on two distinct environments with

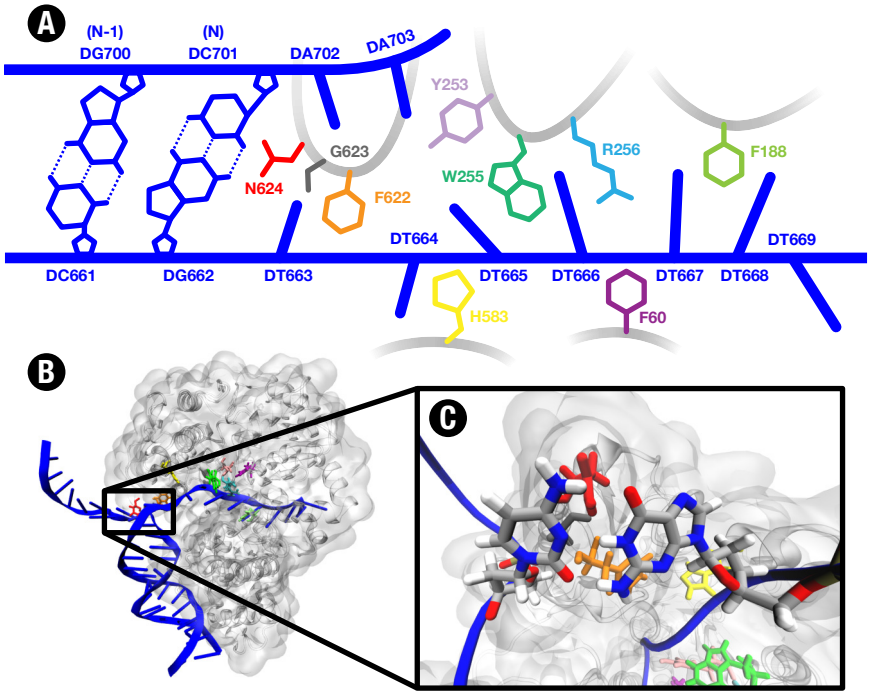


Figure 9-2 The simulation system for studying proton transfer within the replisome environment comprises the PcrA helicase (grey) in a complex with two DNA strands (blue). (A) shows a schematic of the interface between DNA and the single-stranded DNA binding site of the protein. The final two base pairs of the duplex DNA, pair N (DG662-DC701) and pair N-1 (DC661-DG700), and note-worthy amino acids of the enzyme are highlighted. (B) shows a 3D rendering of the enzyme-DNA complex. (C) is a 3D render zoomed in on the DC701-DG662 QM region within the enzyme-DNA complex where proton transfer is investigated.

contrasting environmental complexity. Firstly, we model the aqueous DNA system where the bases of interest are embedded within a stack of bases above and below and surrounded by water and counter ions. Secondly, we model the double proton transfer in GC within an explicit replisome environment (see Figure 9-2). For both cases the DNA sequence is given in Table S1 of the SI.

We perform calculations for the duplex's base pair closest to the helicase (N), and for the rung before (N-1), and at two distinctive timescale limits to home in on the role of the distance between DNA and helicase during DNA strand separation. In the US simulations, it is assumed that all the vibrational degrees of freedom, with the exception of the reaction coordinate, are allowed to relax during the proton transfer. The US trajectories provide a fully-equilibrated model to the proton transfer dynamics because every point on the sampled path is able to equilibrate along the degrees of freedom perpendicular to the reaction coordinate. Conversely, we investigate an approximation in which the proton transfer occurs near-instantaneously (i.e. via fast quantum tunnelling), and thus, the surrounding environment is effectively frozen. In the GC dimer the proton transfer is dominated by tunnelling[117], and atomic rearrangement can play a role in priming a GT mismatch into a 'tunnelling-ready state'[2]. The DPT is likely to occur over a timescale in between the two models, there is a competition between the strand separation and the proton transfer. For both the aqueous dsDNA and DNA-Helicase complex, we have obtained potential energy surfaces for

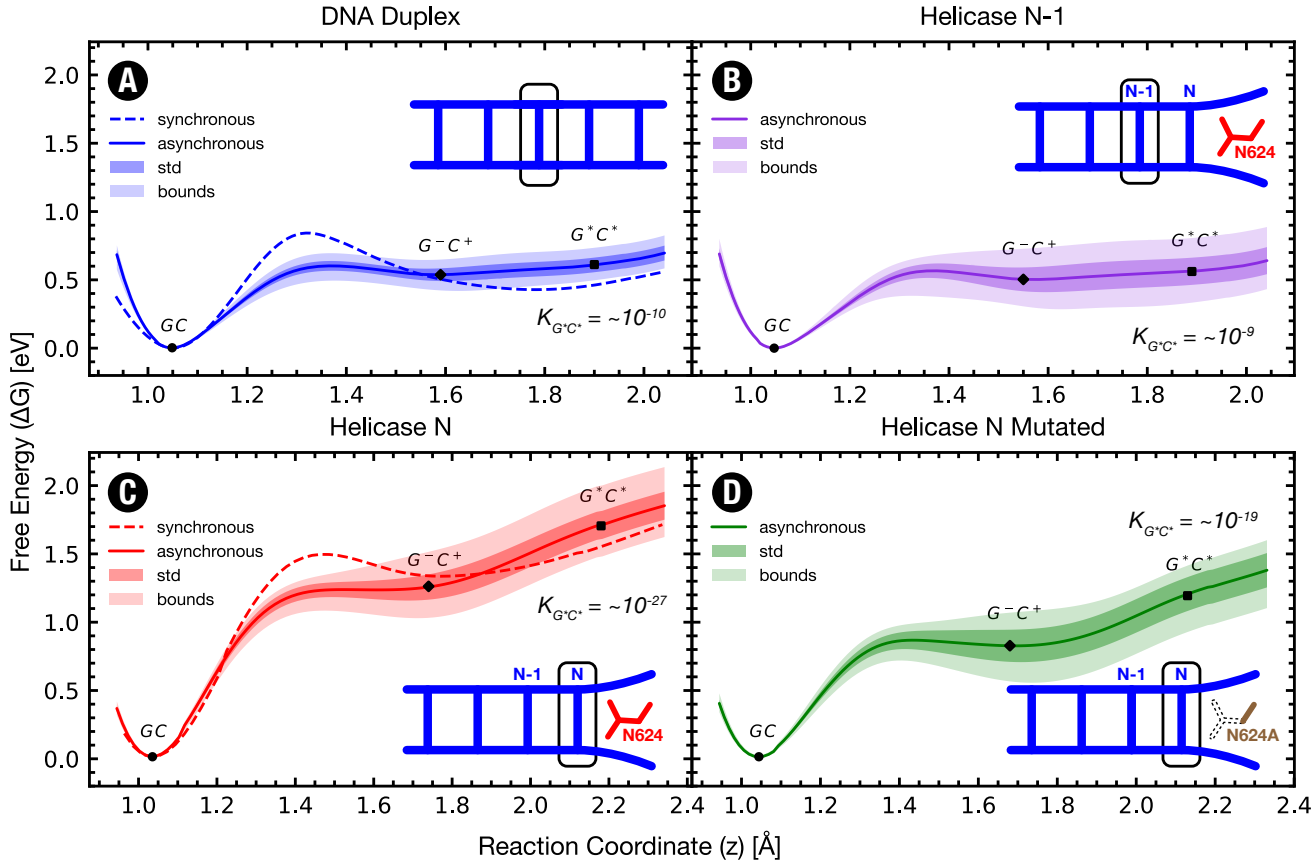


Figure 9-3 Potentials of mean force (PMF) for the asynchronous double proton transfer in guanine-cytosine obtained from QM/MM Umbrella Sampling (US) simulations for Aqueous Duplex DNA (A) and two base pairs in the wild type PcrA Helicase-DNA Complex, pair N-1 (B) and pair N (C). Panel (D) additionally shows the umbrella sampled DPT PMF for the PcrA Helicase with the N624A mutation. For aqueous DNA panel (A) and Helicase base pair N panel (C) the associated potential of mean force for the concerted DPT is shown as a dashed curve. The concerted barriers with full bootstrapping statistics can be found in Figure S7 of the supplementary information (SI). Table S2 of the SI gives the sampling times for each of these PMFs. For each panel the equilibrium constant of the tautomeric state $K_{G^*C^*}$ is superimposed. For all US profiles presented here the motion along four distance reaction coordinates (donor-hydrogen and hydrogen-acceptor for each of the two hydrogen bonds) was projected onto a single reaction coordinate using the average relative change from the canonical minimum see Figure S1 and Section 1.4 of the SI.

the instantaneous double proton transfer and the minimum energy pathways. Further details are provided in Section 1.5 and Figures S9-S14 of the SI.

We start with our control system, which is a simulated double-stranded DNA in an aqueous solution within QM/MM. We use this as a baseline for understanding the effect of the helicase on proton transfer and as it allows us to compare to previous studies [110, 115, 109].

We determine two categories of free energy reaction paths: asynchronous and synchronous. A synchronous path is one in which both protons transfer simultaneously, whereas an asynchronous path corresponds to the situation in which the middle proton between the two nitrogen atoms transfers first (see Figure S2 of the SI). The properties of the free energy reaction paths obtained from US are shown in Table 9-1. These two mechanisms align with those proposed by Gheorghiu

Table 9-1 Statistical properties of umbrella sampling (US) potentials of mean force (PMF) for the double proton transfer in guanine-cytosine for two pathways: 'synchronous' where both protons transfer simultaneously, and 'asynchronous' where the middle proton between the two nitrogen atoms transfers first. Reaction energies (ΔG_{rxn}), forward and reverse energy barriers (ΔG_{fwd} and ΔG_{rev}, respectively), and equilibrium constants (K) are reported for each reaction transition and the different replication scenarios studied in this work. DNA duplex refers to the DPT in aqueous linear double-stranded DNA. Helicase N-1 refers to the DG700-DC661 base pair of DNA in complex with PcrA Helicase, and Helicase N to base pair DG662-DC701 of the same complex. N624A refers to the mutation of asparagine N624 into alanine, which does not form hydrogen bonds.

Synchronous US	Property	DNA Duplex	Helicase N-1	Helicase N	Helicase N624A
GC	ΔG_{rxn}(eV)	0.47 ± 0.06	0.48 ± 0.22	1.35 ± 0.13	-
$\downarrow$	ΔG_{fwd}(eV)	0.85 ± 0.20	0.56 ± 0.20	1.53 ± 0.09	-
G^*C^*	ΔG_{rev}(eV)	0.41 ± 0.21	0.08 ± 0.24	0.18 ± 0.15	-
	Reverse barrier (%)	100	100	100	-
	K	$2.50 \cdot 10^{-8}$	$1.72 \cdot 10^{-8}$	$1.46 \cdot 10^{-22}$	-

Asynchronous US	Property	DNA Duplex	Helicase N-1	Helicase N	Helicase N624A
GC $\rightarrow$ G*C*	ΔG_{rxn}(eV)	0.60 ± 0.05	0.55 ± 0.16	1.68 ± 0.10	1.18 ± 0.13
	K	$1.68 \cdot 10^{-10}$	$1.50 \cdot 10^{-9}$	$6.70 \cdot 10^{-28}$	$9.53 \cdot 10^{-20}$
GC	ΔG_{rxn}(eV)	0.55 ± 0.05	0.49 ± 0.13	1.22 ± 0.09	0.82 ± 0.10
$\downarrow$	ΔG_{fwd}(eV)	0.60 ± 0.04	0.57 ± 0.09	1.24 ± 0.07	0.86 ± 0.07
G^-C^+	ΔG_{rev}(eV)	0.06 ± 0.03	0.09 ± 0.06	0.05 ± 0.05	0.05 ± 0.04
	Reverse barrier (%)	100	91	42	68
G^-C^+	ΔG_{rxn}(eV)	0.06 ± 0.04	0.05 ± 0.06	0.46 ± 0.03	0.35 ± 0.04
$\downarrow$	ΔG_{fwd}(eV)	0.06 ± 0.04	0.05 ± 0.06	0.46 ± 0.03	0.35 ± 0.04
G^*C^*	ΔG_{rev}	-	0.04 ± 0.06	-	-
	Reverse barrier (%)	0	25	0	0

Table 9-2 Statistical properties of minimum energy path profiles for the instantaneous double proton transfer in guanine-cytosine for two asynchronous DPT pathways: via the two zwitterions G^-C^+ and G^+C^-. Reaction energies (ΔE_{rxn}), forward and reverse energy barriers (ΔE_{fwd} and ΔE_{rev}, respectively) are reported for each reaction transition and duplex DNA and the PcrA Helicase-DNA complex. The minimum energy path included the G^+C^- Zwitterion in two out of seven replicas with the Helicase and zero out of seven replicas for DNA. Properties of the energetic minima for each replica instantaneous surface are provided in Table S3 of the supplementary information. † are derived from a single replica

Asynchronous MEP	Property	DNA Duplex	Helicase N	Helicase N624A
GC $\rightarrow$ G*C*	ΔE_{rxn} (eV)	1.13 ± 0.25	1.84 ± 0.27	1.69 ± 0.23
GC	ΔE_{rxn} (eV)	0.61 ± 0.17	1.35 ± 0.15	1.19 ± 0.11
$\downarrow$	ΔE_{fwd} (eV)	0.77 ± 0.23	1.38 ± 0.15	1.23 ± 0.11
G^-C^+	ΔE_{rev} (eV)	0.15 ± 0.11	0.05 ± 0.05	0.04 ± 0.01
	Reverse barrier (%)	100	80	100
	Product Lifetime (fs)	200 ± 270	30 ± 10	-
G^-C^+	ΔE_{rxn} (eV)	0.52 ± 0.13	0.43 ± 0.17	0.50 ± 0.12
$\downarrow$	ΔE_{fwd} (eV)	0.70 ± 0.17	1.41 ± 0.18	1.23 ± 0.35
G^*C^*	ΔE_{rev}	0.18 ± 0.10	0.98 ± 0.09	0.73 ± 0.23
	Reverse barrier (%)	100	100	-
	Product Lifetime (fs)	120 ± 150	260 ± 150	-
GC	ΔE_{rxn} (eV)	-	1.93 ± 0.34	2.04†
$\downarrow$	ΔE_{fwd} (eV)	-	2.24 ± 0.17	2.41†
G^+C^-	ΔE_{rev} (eV)	-	0.32 ± 0.17	0.38†
	Reverse barrier (%)	-	100	100
	Product Lifetime (fs)	-	-	-
G^+C^-	ΔE_{rxn} (eV)	-	0.04 ± 0.03	-0.15†
$\downarrow$	ΔE_{fwd} (eV)	-	0.60 ± 0.02	0.39†
G^*C^*	ΔE_{rev}	-	0.56 ± 0.00	0.54†
	Reverse barrier (%)	-	100	100
	Product Lifetime (fs)	-	-	-

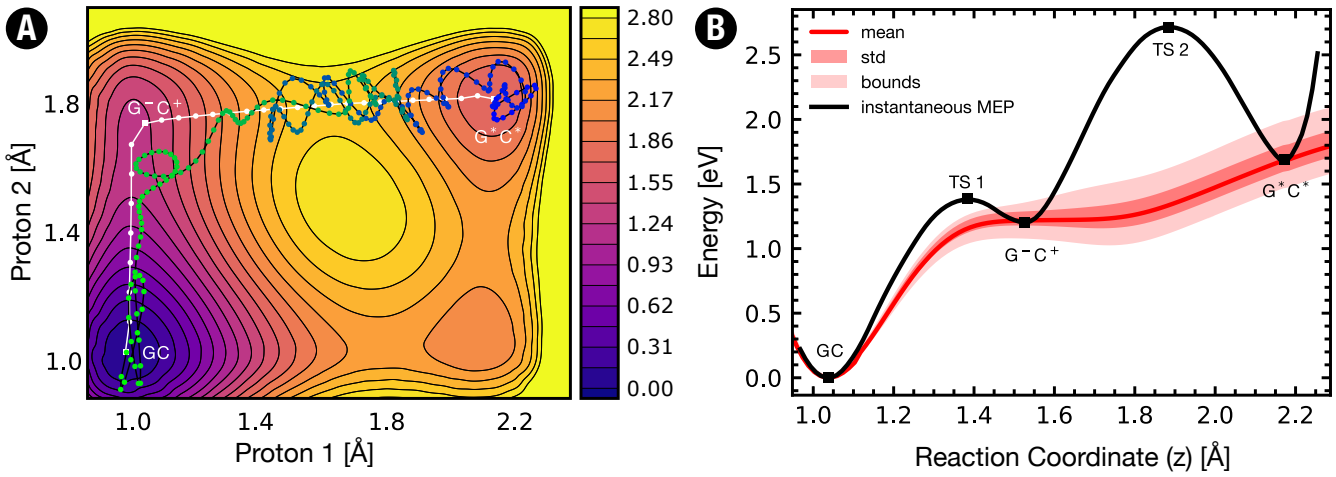


Figure 9-4 Instantaneous DPT in a snapshot of the PcrA Helicase-DNA complex. (A) Contoured heatmap illustrating the DPT's instantaneous energy surface, the minimum energy pathway within this landscape (white line with circular dots), and the decay path from G^*C^* to GC (black line with multicoloured circles). For the decay path, the dots are coloured according to the simulation time, with blue corresponding to the start of the simulation and green at the end of the decay. The color scale provides the potential energy in eV. (B) The free energy of the DPT in an ensemble from US (red curve and red shaded regions) compared to the minimum energy path for DPT within the instantaneous energy landscape (black curve). The instantaneous MEP in panel B corresponds to the energy profile taken along the MEP in panel A.

et al. [115, 109] who suggested that the reaction proceeds via a statistical mixture of the two paths. The asynchronous pathway is also in agreement with the Path Integral Molecular Dynamics (PIMD) US results of Angiolari, *et al*[216]. Other studies suggest that during the dissociation of the bases, the asynchronicity of the proton transfer increases [1, 3].

Here, the reaction asymmetry (ΔG_{rxn}) for the GC $\rightarrow$ G^*C^* process in aqueous DNA is 0.54 ± 0.08 eV, in good agreement with other aqueous GC dimer studies [104, 217, 122, 1]. Results from Angiolari, *et al*'s PIMD US[216] include a reaction asymmetry of 0.61 ± 0.02 eV for an isolated GC dimer, and 0.83 ± 0.01 eV for an explicitly solvated three base pair DNA duplex for the DPT. The higher asymmetry in the larger hydrated system is attributed by the authors to the presence of a water molecule near the O6 acceptor of the guanine, which in our work is observed in the PcrA-helicase complex (see Discussion). Furthermore, the lower activation energy (ΔG_{fwd}) for the asynchronous mechanism (0.60 ± 0.04 eV) compared to the synchronous mechanism (0.85 ± 0.20 eV) suggests that the former is kinetically preferred. For the cases of aqueous duplex DNA and base pair N in complex with PcrA Helicase the potentials of mean force for the synchronous transfer are shown in Figure 9-3. Synchronous proton transfer PMFs are shown with statistical uncertainty in Figure S7 of the supplementary information, including the base pair N-1 case. Further determination of the most favourable DPT pathways are explored using minimum energy paths later in this article. For the asynchronous path, we report a small but clear free energy barrier between the canonical (GC) and zwitterionic (G^-C^+) states and a barrier-less transition between G^-C^+ and G^*C^* see Table 9-1 and Figure 9-3. Here, the error bars arise from the statistical ensemble of replicas.

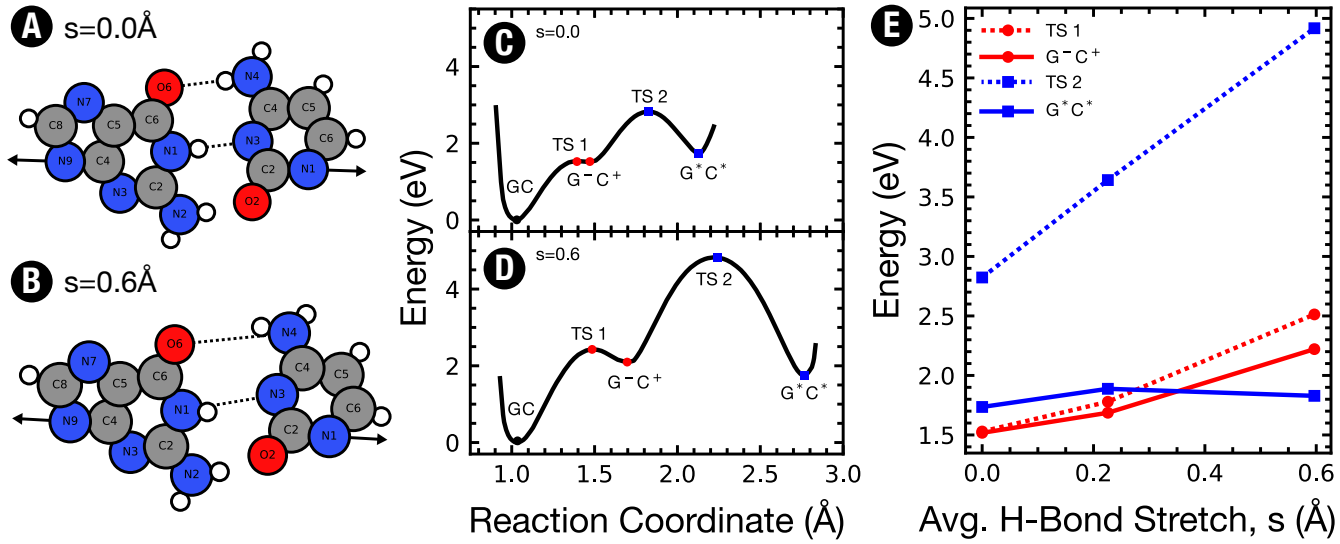


Figure 9-5 Scheme for determining the effect of mechanical separation on the instantaneous double proton transfer (DPT) energy landscape. Panel A shows the equilibrium distance GC dimer taken as the starting point for the steered molecular dynamics (SMD). This GC dimer forms the QM region within the base pair N PcrA helicase-DNA complex system as used elsewhere in this work. Two black arrows denote the atoms the constant non-equilibrium pulling force was applied to during the SMD and the direction. The DPT was sampled across the two hydrogen bonds denoted by dotted lines. Panel B contains the same elements as Panel A, but for a later snapshot in the SMD corresponding to a separation of $0.60\,$Å. Panel C shows the minimum energy path (MEP) through the instantaneous potential energy landscape (PES) for the equilibrium separation snapshot. Panel D is the equivalent of Panel C for $0.60\,$Å separation. Panel E shows the energies of the zwitterion (G^-C^+), tautomer (G^*C^*), and transition states at separation values relative to the canonical GC of $0.23\,$Å and $0.60\,$Å.

We will now turn our attention to the case when the DNA is in contact with the larger replisome environment. Initially, we focus on the end of the double-stranded duplex, that is in close contact with the enzyme's binding site, since this is the last base pair to potentially undergo proton transfer before entering the helicase. The two base pairs of interest are the duplex's second nearest base pair (N-1) and the final base pair in the stack (denoted N). These two base pairs are shown in panel A of Figure 9-2. Under the initial inspection of Figure 9-3, it appears that when DNA forms a complex with PcrA helicase, the free energy surfaces for both the synchronous and asynchronous DPT are significantly different from that of the DNA duplex. Conversely, Figure 9-3 and Table 9-1 show that the single GC base pair between N-1 and the helicase ssDNA binding site sufficiently mimics the duplex DNA environment, resulting in the aqueous DNA and helicase N-1 cases being nearly indistinguishable. Here the aqueous DNA and helicase N-1 cases have a forward reaction barrier and reaction free energy within each other's uncertainty for both the synchronous and asynchronous proton transfer mechanisms. This result is consistent with other works [108] demonstrating that three base pairs of DNA are sufficient to model stacking effects on DPT in double-stranded DNA when considering proton transfer in the middle base pair. Consequently, in the N-1 case, the base pair above appears to "shield" the interaction of the helicase from the proton transfer mechanism by providing a local environment closely resembling duplex DNA. Whether by mechanical constraints,

induced immobility and conformational deformation induced by electrostatic interactions with the helicase, the G^*C^* tautomer in base pair N-1 is protected from the destabilising effect observed in base pair N.

Base pair N, situated at the ssDNA binding site of PcrA helicase, is affected by π-π interactions from only one side (from the base pair below) and is near the side chain of an asparagine amino acid (N624) of the helicase enzyme, shown in red in Figure 9-2. The GC base pair N has also been partially opened as the hydrogen bonds have been stretched by 0.3 Å relative to the equilibrium in both the DNA and helicase N-1 cases. Thus, for base pair N, the energetics of the proton transfer is radically changed. The overall endothermicity of the reaction has increased, causing the G^*C^* and G^-C^+ states to be less likely to be populated compared to the aqueous DNA case. The resulting proton transfer profiles indicate that the G*C* state is thermodynamically unstable since it sits on a steep free energy slope which suggests the system is most likely to return to the canonical reactant. In addition, the G^-C^+ intermediate becomes metastable with only 42% of bootstrapping profiles exhibiting a small reverse barrier back to the canonical GC. Previous work has attributed the greatest effect of a helicase enzyme on the DPT in GC to be due to separation of the base pair which reduces the rate of proton transfer. In our US of the DPT in base pair N of the PcrA complex, the purely mechanical effect of the helicase is incorporated in the QM/MM model by taking snapshots from MD. For base pair N we find the base pair to be separated by an additional 0.15 Å with respect to the aqueous DNA trajectories. According to purely mechanical models[1, 3] a 22% increase to the forward reaction barrier and 1% decrease of the reaction asymmetry is expected at this distance. Our US indicates corresponding changes of 280% to the forward barrier and 180% for the reaction energy. We thus conclude that mechanical separation alone is not enough to describe the changes to the proton transfer energy profile in the presence of the helicase, as neither the radical increase in forward barrier or asymmetry between duplex DNA and base pair N in the helicase-DNA complex can be accounted for by existing separation models.

We further investigate the interaction of helicase and the proton transfer mechanism by homing in on the closest residue of helicase. The role of the asparagine in destabilising G^-C^+ and G*C* is revealed by the US results for the N624A mutated case. Here, asparagine (N624) has been substituted with an alanine (N624A), which has a much smaller (hydrophobic) side chain and cannot form hydrogen bonds with the DNA. As asparagine N624 is not part of the ssDNA binding site of PcrA helicase, which plays a key role in the inchworm mechanism of PcrA[218, 20], its substitution with alanine should only negligibly affect the enzyme's function. This is further corroborated by the presence of various amino acid motifs in the sites 622-626 (see Section 1.8 and Table S4 of the SI), indicating that these residues are not key to PcrA's primary biological function in strand separation through ssDNA translocation. The precise impact of the N624A mutation on the efficacy of the helicase would need to be experimentally determined. When the alanine mutation is present, some of the destabilising effects of the helicase on G^-C^+ and G*C* are mitigated. The reaction asymmetry

of the G^*C^* product relative to GC is reduced by 30%, and a slight reverse barrier between GC $\rightarrow$ G^-C^+ appears 62% more frequently. In addition to any electrostatic differences due to removing the asparagine's sidechain, the canonical GC state was also found to be more structurally compressed with the N624A mutation relative to the wild type. During the QM/MM steered MD sampling of the canonical GC within the N624A mutated helicase, the hydrogen bonds were stretched by $0.16\,\text{Å}$ relative to equilibrium duplex DNA.

In summary, within our model, we can reproduce the energetic landscape for the aqueous duplex DNA and match it with prior literature [115, 109]. We significantly expanded the system's complexity and fidelity with respect to the biological environment by including the helicase enzyme and determine that the helicase radically alters the free energy landscape suppressing the probability of proton transfer in GC. Furthermore, we found that modifying a single local residue reduces the electrostatic interaction between the helicase and the DNA bases about to dissociate, thus leading to a significant reduction in reaction energy. Overall, this indicates that the helicase environment suppresses the formation of the tautomer and increases the probability of the canonical form of DNA as the strands separate.

We now consider a scenario in which there is not a sufficient amount of time for the system to relax due to very fast DNA strand separation; instead, the system's vibrational degrees of freedom perpendicular to the reaction coordinate remain "frozen" during the proton transfer. We propose a method to explore the minimum energy pathway that instantaneously transferring protons would take between canonical and tautomeric GC. This allows for the determination of the quantum tunnelling pathway as suggested by previous studies of the DPT in the GC base pair and aqueous DNA[213, 117, 115, 216], and provides an explanation for the picosecond stability of the tautomer despite the lack of a reverse energy barrier in the US PMF. Here we assume that the helicase is poised to dissociate the base pair N DNA rung and consider the instantaneous proton transfer energy landscape. A similar non-equilibrium model has been applied to the aqueous GC and the GT wobble mismatch in the thumb domain of DNA Polymerase where we showed that nuclear quantum effects play a critical role in the proton transfer reaction of isolated GC bases [117, 2], leading to a fast proton exchange. Accordingly, we expect both the quantum and classical rate of proton transfer to be significantly modified by the increased reaction energies in the DNA-Helicase complex. In this scenario, it is pertinent to consider a GC $\rightarrow$ G^*C^* reaction whose product state is solely populated via a fast proton transfer mechanism. In this case, the system will not be subject to a thermalised ensemble energy barrier, but to an instantaneously "frozen" energy landscape with all other degrees of freedom constrained. We obtained multiple frozen potential energy surfaces (PES) for a sample of (canonical GC) QM/MM US snapshots and determined QM/MM single point energies (SPEs) for a two-dimensional grid of proton positions (see Section 1.5 of the SI). Within these landscapes, the locations of local minima were determined using the BFGS optimiser, and the minimum energy path (MEP) between the canonical GC and tautomeric G^*C^* was determined using a nudged elastic

band (NEB) algorithm. These methods are described in secton 1E of the supplementary information. The MEP is shown as the white line on panel A of Figure 9-4, and the energy profile along this MEP is overlaid on the corresponding US result in panel B of Figure 9-4. Table 9-2 displays the properties of the reaction profile along the MEP through the PES for aqueous duplex DNA, base pair N in the Helicase-DNA complex with the asparagine N624 wild type, and alanine N624A mutation. The instantaneous PES model produces comparable reaction asymmetries for each of the studied systems, but find higher transition state energies resulting in increased stability of the rare zwitterionic and tautomeric states. We propose that these raised transition states are due to conformational changes which accommodate the non-canonical protonation states which are populated in US but are frozen in the instantaneous PES. While the MEPs for aqueous DNA always pass via an intermediate G^-C^+ zwitterion, in both helicase scenarios, a small fraction of PES replicas exhibited MEPs passing through the alternate zwitterion G^+C^-. All seven replica MEPs proceeded via G^-C^+ for aqueous DNA while for the wild-type PcrA Helicase, only five of the seven replicas chose this path, and the N624A mutation showed this behaviour in three of the four replicas.

From the 2D-PES, we can determine the position of the proton corresponding to the G^*C^* local minimum. To investigate the behaviour of the G^*C^* state, we take the tunnelling product predicted by the MEP and place the protons back in the QM/MM ensemble snapshot from which the PES was generated (see Section 1.6 of the SI). All QM/MM trajectories resulted in the two protons transferring back to the canonical GC protonation state within $0.6\,\mathrm{ps}$ during these unbiased ensemble simulations (see Figure S3 of the SI). The decay from the G^*C^* tautomer to canonical GC was observed to take several pathways, one of which is illustrated on panel A of Figure 9-4. While for some replicas, the MEP was found to go via the alternate zwitterion G^+C^-, no decay trajectories were found to take this path. However, in three out of seven trajectories within the wild-type PcrA helicase, the protons passed over or near the energetic maximum corresponding to a synchronous DPT. This behaviour was not observed in aqueous DNA.

Slocombe *et al.* [1, 3] have previously shown that imposing a separation on the DNA dimer stabilises the tautomeric product by increasing the reaction barrier of the minimum energy path (MEP). The primary action of the Helicase is to separate the strands of DNA, which will inevitably alter the DPT reaction profiles. During the umbrella sampling the GC dimer is compressed at the transition states, and both the canonical and tautomeric GC separations are higher for base pair N of the helicase than base pair N-1 and duplex DNA (see Section 2.3 and Figure S6 of the SI). We have repeated our instantaneous MEP methodology for three steered MD snapshots at increasing separations (see Section 1.7 of the SI). Akin to the work by Slocombe *et al.* [1, 3], a steering force was applied to the backbone atoms of the GC dimer in the PcrA Helicase-DNA complex. Panels A and B of Figure 9-5 show two example GC dimer configurations. We have verified that in the DNA-Helicase complex, the GC dimer opens asymmetrically (with the DG:N2-DC:O2 hydrogen-bond staying at near equilibrium) as reported in [1]. The instantaneous PES of the DPT was obtained, and

the MEP was again determined with NEB. While the overall reaction asymmetry between canonical GC and the G^*C^* tautomer was uncorrelated with the separation in this study, the two transition states and intermediate zwitterionic minimum in the triple well potential increased linearly with separation. Broadly these results indicate an agreement with previous mechanical studies which indicate a rapid reduction of proton transfer between canonical and tautomeric states as separation increases, which supports a "trapping" phenomenon where separation of a G^*C^* base pair leads to greater stability of the tautomer. Our instantaneous PES study at increasing separation indicate that trapping is initially possible, but our US profiles do not exhibit a reverse reaction barrier, suggesting that such trapping would be short-lived as the system thermalises within 1 picosecond.

9.5. Discussion

The US trajectories for the GC base pair at the opening of the ssDNA binding site of PcrA helicase show a considerable increase in the reaction free energy of both the G^-C^+ and G^*C^* states. Furthermore, the presence of the helicase reduces the reverse barrier of the DPT reaction leading to a tautomer with lower stability than in isolated DNA. It is crucial to determine the stability of the G^*C^* tautomeric base pair to clarify if they could survive the strand separation and potentially lead to point mutations. While previous models have anticipated the reaction barrier between canonical GC and G^-C^+ to increase with the mechanical action of a helicase enzyme, our results demonstrate for the first time the effect of realistic model of the helicase-DNA complex on the stability of the tautomeric states. Additionally, we observe that asparagine N624 forms hydrogen bonds with a water molecule in the first solvation shell (see Figure S8 of the SI), holding it in place close to the dissociating DNA and modifying the free energy landscape for proton transfer. This is further demonstrated by the increased density of water atoms surrounding the GC donor and acceptor atoms in base pair N of the helicase complex during classical MD, as seen in the Figure S19 of the SI. Most notably, in base pair N of the PcrA complex the hydrogen acceptor O6 of the guanine has a 50% increased density around 1.7 Å when compared to both aqueous DNA and the N624A point mutated complex. This is consistent with the formation of a hydrogen bond in close proximity to a water molecule as we indeed observed in our QM/MM US trajectories, in good agreement with the results of Angiolari, et al[216]. Angiolari, et al. also suggest the presence of a water molecule in this configuration plays a role in destabilising the rare tautomeric and zwitterionic intermediate protonation states in GC[216]. Substituting the asparagine with a hydrophobic alanine reduces this destabilising effect (see Figure S8 of the SI). The remaining differences between duplex DNA and the alanine mutated helicase profile are attributed to the increase in base pair separation and altered chemical environment modelled by the electrostatic embedding of the QM/MM.

In duplex DNA, the GC $\rightarrow$ G^*C^* tautomerisation takes place with a significant tautomeric population maintained at dynamic equilibrium [213, 117, 1]. However, the decreased stability of the G^*C^*

base pair at the active site of PcrA helicase suggests that one of the roles of the helicase is to protect against the formation of excess mutations. The stepping motor action of PcrA helicase, includes fast translocation dynamics ($1\,\text{Å}/\text{ps}$) which are suspended for extended periods (10's milliseconds) while a new ATP molecule diffuses into the binding pocket and is hydrolysed [16, 18, 1]. The most likely route for a point mutation to successfully pass the helicase enzyme is for the G^*C^* base pairs population to be formed before base pair N in the DNA-helicase complex due to the significant reaction barrier and barrier-less reverse reaction in the helicase active site. The millisecond lag time between helicase's translocation steps and the fractional picosecond lifetime of the tautomers suggests a further reduction in the tautomer population, and chemical equilibrium will have been reached.

Our novel instantaneous model for the proton transfer PES indicates that any such transfer reaction requires a significantly increased activation energy when compared to the equilibrium description (from US). It is conceivable that a rare tunnelling-ready state (TRS) may be populated via thermal fluctuations of the canonical configuration, in which the system is primed for transfer, akin to the mechanism reported in the Wobble-GT dimer in a polymerase palm domain[2]. A TRS would place the canonical system closer to the transition state and reduce the effective free energy barrier towards the zwitterion and tautomeric states. To promote tunnelling the protons would need to approach a configuration whose energy is degenerate with respect to the target well. In the PcrA helicase complex our PESs indicate a GC system would need to gain $1.35 \pm 0.13\,\text{eV}$ of thermal energy to be degenerate with the zwitterion G^-C^+. The intermediate G^-C^+ requires $0.43 \pm 0.17\,\text{eV}$ to account for the reaction free energy to the G^*C^* tautomeric well. Accounting for the low probability of thermally occupying a TRS in the canonical well (on the order 10^{-22}), and then the zwitterionic well (10^{-7}), we expect the gain from thermally occupying states closer to the TS to be negligible.

However, while the tautomeric population is significantly diminished by the presence of the helicase, a small fraction of non-canonical GC dimers becomes trapped behind the rapidly increasing energy barrier when the hydrogen bonds begin dissociating. Consequently, the exact population of tautomers passing this process is determined by solving the microkinetics of the DPT process in a time-dependent PES, which we plan to study in closer detail in future studies. Here, we provide a semi-quantitative assessment of the probability of point mutations surviving the DNA splitting by assuming that the population (1×10^{-8}, see Table 9-1) in the rung down equilibrates in the helicase active site to a much lower concentration of 1×10^{-22}. Determination of the tunnelling rate in a noisy environment and variable energy landscape poses a significant theoretical challenge. Simple corrections such as those proposed by Wigner[141] and the WKB approximation[219] are not strictly valid in this 'deep tunnelling' regime and, as a result, would return a negligible tunnelling correction [213]. On the other hand, an accurate open quantum system description is required to capture the dynamics of the helicase environment. However, the magnitude of the free energy barriers reported here expects to diminish the tunnelling [117].

Controlling the rate of mutations is one of the functions of a successful biological replication mechanism [20], and whether or not the suppression of G^*C^* mutagens by PcrA helicase is the result of a specific evolutionary pressure is an important question. Furthermore, the roles of well-conserved residues in the ssDNA binding site of PcrA helicase are known, and the asparagine N624 site is not shown to be vital to the stepping motor action of the enzyme as it in not part of the ssDNA binding domain[20].

A survey of PcrA helicase sequences across 59 species was performed, exhibiting N624 in the majority (51%) of cases. Other amino acids present in this position are arginine (denoted R624 at 24%), glutamine (Q624 at 14%), and tyrosine (Y624 at 12%). R624 and Q624 strongly correlate with changes to other vicinal residues and occur within a multitude of different motifs between sites 622 and 626. These alternate sequences, as well as consideration of other helicase enzymes such as the 31 DNA helicases in the human genome[220] and viral helicases[221, 222, 223], might indeed be a fruitful route for future research, although we expect a similar increase in hydration around the hydrogen bonding site in most helicases leading to a destabilisation of rare protonation states.

Since the specific role of asparagine has been demonstrated in this work by the reduction in tautomer destabilisations following an alanine substitution, and N624 is not crucial for the successful function of PcrA helicase, it can be hypothesised that this specific amino acid was naturally selected to reduce mutations. In addition to demonstrating the need for further theoretical work on this area, we also provide an experimentally testable hypothesis that the N624A alanine substitution would reduce the rate of point mutation formation through the DPT.

While we have demonstrated the importance of modelling the explicit biological environments for understanding the role of proton transfer in DNA and the need for multiscale modelling to be included in future studies of GC tautomerism, to determine the precise interplay between a dynamically changing free energy landscape and the quantum delocalisation of GC's protons. We have shown that the dynamics of DNA separation occur on the picosecond timescale and are sufficient to change the DPT reaction in duplex DNA radically [1]. The QM/MM MD decay trajectories give us critical insight into the short-lived nature of the metastable states, going beyond simple reaction kinetics. We observe G^*C^* lifetimes on the order of tenths of picoseconds, suggesting that these processes proceed within the same timescale as strand separation and would need novel theoretical rate calculations before any definitive conclusions may be drawn.

9.6. Methods

The duplex B-DNA structure was generated with Avogadro[224]. The atom and residue names were modified and suitable 3' and 5' termini were added using MolParse[225]. A modified and optimised version of the PcrA Helicase-DNA substrate complex PDB entry 3PJR,[15] was obtained from Yu et al. [18]. The systems were solvated and neutralised in a cubic box of SPCE water with $1\,\mathrm{nm}$

clearance and sodium counter-ions.

Energy minimisation, molecular mechanics, and molecular dynamics were performed using Gromacs 2018[200], with the March 2019 version of the CHARMM36 force field[226, 227, 201]. All calculations in Gromacs used the Verlet cutoff scheme applying 1 nm Coloumb and vdW radii, utilising fourth order Particle Mesh Ewald long-range electrostatics, and cubic periodic boundary conditions.

Energy minimisation was performed on both these systems using a steepest descent algorithm until the maximum force did not exceed 12 kJ/mol/nm. Equilibration was performed in an NVT ensemble using a leap-frog integrator, timestep of 1 fs, and Nose-Hoover temperature coupling with time constant 0.2 ps and reference temperature 310 K, and three-dimensional periodic boundary conditions for a total of 3 ns.

Hybrid QM/MM calculations were performed with the Gromacs-CP2K distribution[200, 228]. The GC nucleobase pair QM region was modelled with density functional theory, solved with CP2K[228] with the BLYP exchange correlation functional, DZVP-MOLOPT-GTH basis set, GTH-BLYP potential, electrostatic embedding, and the VDW3 dispersion correction. For QM/MM dynamics, Gromacs applied the leap-frog algorithm, with forces obtained from DFT in CP2K for the QM region and CHARMM36 elsewhere.

Umbrella Sampling was performed with Gromacs-CP2K at a CHARMM36/DFT/BLYP level of theory using four distance reaction coordinates describing the double proton transfer. Each umbrella sampling window applied a 20000 kJ/mol/nm^2 harmonic potential along the four reaction coordinates, and was simulated for at least 8 ps per replica per window. Depending on the level of equilibration, up to the first 2 ps of each simulation were discarded. The potentials of mean force (PMF) along the sampled reaction were obtained by the Weighted Histogram Analysis Method (WHAM) implemented in Gromacs with a tolerance of 1.0×10^{-6} and statistical variance was estimated with 100 bootstrapping samples. A logistic function was fit to the RMS difference between WHAM PMFs for the same system but varying sampling times to ensure even convergence between simulation systems (see Figure S5 and Table S2 of the SI).

To determine the instantaneous double proton transfer potential energy surface, snapshots were taken from the QM/MM umbrella sampling simulations corresponding to the canonical GC. Each proton had its position interpolated between 0.9 and 2.5 Angstrom distance from its covalently bonded donor atom. Single point energies (SPEs) were calculated using the previously described electrostatically embedded QM/MM in CP2K with two levels of theory: identical to the umbrella sampling methodology with BLYP/DZVP-MOLOPT-GTH and B3LYP/DZVP-MOLOPT-GTH which also utilised the auxillary density matrix method using the cFIT3 basis. The grid of SPEs was interpolated using the Clough-Tocher piecewise cubic, C1 smooth, curvature-minimizing interpolant in two-dimensions inplemented in SciPy.[229] Within this surface local minima were determined using the BFGS optimiser implemented in ASE,[198] and the minimum energy path connecting the

canonical and tautomeric states was obtained using the nudged elastic band (NEB). To investigate the metastable nature of the G^*C^* tautomer, post DPT product structures were generated from each instantaneous PES, and these were placed in unbiased 310 K NVT QM/MM MD simulations.

To obtain a strand separation trajectory in the DNA-Helicase complex a short (5ps) steered molecular dynamics simulation was performed on an equilibrated structure. A constant separating force (500 kJ/mol/nm^2) was applied to the DC:N1 and DG:N9 backbone atoms forcing them directly apart.

To determine the hydration of GC base pairs in various systems, additional molecular dynamics simulations (33 ns) were performed for three systems: aqueous DNA, the wild type PcrA helicase-DNA complex, and the N624A point mutation of the same enzyme complex. From these trajectories the hydrogen bonding geometries were analysed with MolParse[225], and the rdf function was used to calculate the radial density of water around the hydrogen bonding sites of the GC dimer.

Further details regarding the computational simulation methodology can be found in Section 1 of the supplementary information.

9.7. Acknowledgements

We are grateful for financial support from the Leverhulme Trust doctoral training centre grant number DS-2017-079. This work was made possible through the support of Grant 62210 from the John Templeton Foundation. M.S. is grateful for support from the Royal Society (URF/R/191029). We acknowledge helpful discussions with the members of the Leverhulme Quantum Biology Doctoral Training Centre; particular thanks go to Johnjoe McFadden and Cedric Vallee. This work used the ARCHER2 UK National Supercomputing Service. We are grateful for computational support from the UK Materials and Molecular Modelling Hub, partially funded by EPSRC EP/R029431. This work was supported by HECBioSim, the UK High-End Computing Consortium for Biomolecular Simulation, supported by the EPSRC EP/L000253/1).

9.8. Author contributions statement

L.S., M.S., and J.A-K. conceived and designed this research. M.W. performed the computational chemistry calculations. All the authors contributed to the preparation of the manuscript and have approved the final version of the manuscript.

9.9. Data Availability

The raw data used and analysed during the study are available from the corresponding author on request. In addition to further details provided in the Supplementary Information, input, parameter,

and analysis files sufficient to reproduce the results published in this work are available on Github at:

`https://github.com/mwinokan/GC-tautomerism-in-PcrA-Helicase`

9.10. Author Declaration

There are no conflicts of interest to declare.

10. Quantum Tunnelling Effects in the Guanine-Thymine Wobble Misincorporation via Tautomerisation

This chapter is based on the article: Louie Slocombe, Max Winokan, Jim Al-Khalili, and Marco Sacchi. "Quantum Tunnelling Effects in the Guanine-Thymine Wobble Misincorporation via Tautomerism". In: *The Journal of Physical Chemistry Letters* 14 (Jan. 2023), p. 9-15. DOI: https://doi.org/10.1021/acs.jpclett.2c03171 [2]. Supplementary information is included in Appendix Section E.

My contributions to this project include the design and execution of the Quantum Mechanics/-Molecular Mechanics (QM/MM) methodology for the determination of the role of the polymerase enzyme in priming the system for proton tunnelling. This included: preparation of simulation inputs; preparing and overseeing many replica QM/MM MD simulations on High-Performance Computing resources; development of novel analysis techniques to determine and visualise the behaviour of the Tunnelling Ready State; and discussion of the implication of the QM/MM work in comparison to literature and the Density Functional Theory results obtained by co-author Louie Slocombe. Additional contributions were made to the writing and proofreading of the article body, and preparation of the publications figures.

10.1. Abstract

The misincorporation of a non-complimentary DNA base in the polymerase active site is a critical source of replication errors that can lead to genetic mutations. In this work, we model the mechanism of wobble mispairing and the subsequent rate of misincorporation errors by coupling first-principles quantum chemistry calculations to an open quantum systems master equation. This methodology allows us to accurately calculate the proton transfer between bases, allowing the misincorporation and formation of mutagenic tautomeric forms of DNA bases. Our calculated rates of genetic error formation are in excellent agreement with experimental observations in DNA. Furthermore, our quantum mechanics/molecular mechanics model predicts the existence of a short-lived "tunnelling-ready" configuration along the wobble reaction pathway in the polymerase active site,

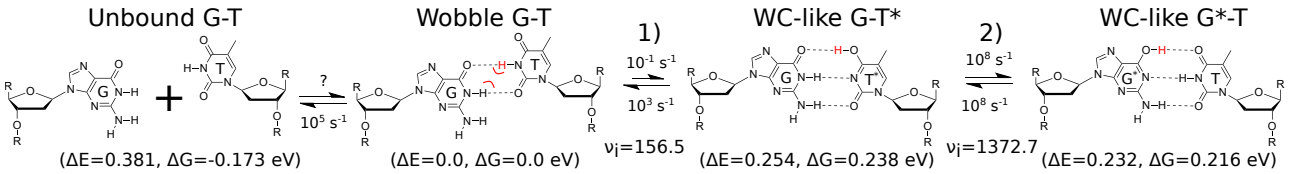


Figure 10-1 Schematic representation of the G-T wobble mispair and the conversion to a Watson-Crick-like configuration via a proton transfer process. For reaction 1 to occur and for the wobble confirmation to adopt a Watson-Crick-like configuration, a proton must rearrange (shown in red). The reaction rates are shown above the arrows. Reaction 1 competes with the unbinding rate of the wobble mispair shown by the first set of arrows on the left. Reaction 2 denotes the further proton transfer reaction in the Watson-Crick-like configuration. Here, the * denotes the tautomeric enol form of the base.

dramatically increasing the rate of proton transfer by a hundredfold, demonstrating that quantum tunnelling plays a critical role in determining the transcription error frequency of the polymerase.

10.2. Article Body

DNA polymerase is an enzyme that catalyses the synthesis of DNA molecules by matching complimentary deoxyribonucleoside triphosphates (dNTP) to the template DNA strand using the standard Watson-Crick (WC) base pair rules. However, when a non-complimentary dNTP diffuses into the active site during the polymerase dNTP sampling, the polymerase domain will transition from an open to an ajar conformation. Thus, forming a different non-standard hydrogen-bonded base-pairing arrangement called a wobble mispair. While there are other sources of replication errors, the fidelity of replication primarily depends on the ability of polymerases to select and incorporate the correct complementary base (see Fig. 10-1) and reject these wobble mispairs. However, it is proposed [59, 230, 62, 77, 210, 231, 232, 233] that the wobble mismatch can form alternative tautomeric configurations that can mimic the WC geometry and lead to erroneous DNA base matches, such as wobble(G-T)→G*-T, where G* is the tautomeric (enol) form of the G base [176]. Watson-Crick-like mispairs have been observed in the active sites of DNA polymerases [59, 29, 234], and ribosomes in enzymatically competent conformations [235, 230, 236]. Both NMR relaxation dispersion experiments and simulations [62, 77, 210] indicate that the concentration of tautomeric mismatches in the cellular environment is significant and has a considerable impact on the replication fidelity of the polymerase. They have all demonstrated that the population of Watson-Crick-like G-T mispairs depends on the local environment, such as the base sequence and the local solvation environment.

Recent theoretical work on the Watson-Crick bonded bases entering the helicase enzyme has shown that quantum effects lead to the formation of meta-stable tautomeric forms of DNA [180, 117]. Quantum chemical models of G-C and A-T base pairs [180] describe the double proton transfer's potential energy surface (PES) in both canonical base pairs. The main difference between the A-T and G-C PES is that A-T has a considerable forward barrier for the tautomer formation but a small reverse barrier that causes its tautomeric form to be highly unstable [111, 115, 178]. On

the other hand, G-C has a sizeable reverse barrier, giving a tautomeric lifetime comparable to the replication process. Moreover, quantum tunnelling leads to a fast proton exchange between the bases [237], such that the timescale of the helicase cleavage is much slower than the proton transfer dynamics [176]. Consequently, using a semi-classical interpretation [180, 117], the potentially mutagenic tautomeric form is continuously formed and destroyed over timescales several orders of magnitude quicker than the helicase cleavage, after which the bases are split into their monomeric forms. However, using a quantum interpretation, the tunnelling proton's wave functions evolve on a shorter timescale, so two clear probability distributions (in the canonical and tautomeric configuration) emerge. As previously demonstrated, once the tautomer is formed and the DNA is opened, it is stabilised and is unlikely to revert to its canonical form due to a prohibitively large reaction barrier [180, 1]. However, it still needs to be determined to what degree environmental effects play a role in destabilising the tautomer. Some initial evidence suggests that the DNA [115] environment reduces the reverse barrier, but it is unclear for the DNA and helicase complex.

Recent NMR experiments using isotopic substitution suggest that when the DNA enters the polymerase active site, the wobble tautomerisation reaction might be facilitated by tunnelling [238]. Rangadurai *et al.* investigated the dynamics of the transition between a wobble and Watson-Crick-like G-T in duplex DNA by performing NMR relaxation dispersion in both H_2O and D_2O. The authors reported that the kinetic isotope effect (KIE) in the exchange rate between the two conformations of the mismatch was threefold slower in heavy water. This result provides the first experimental evidence supporting the hypothesis that quantum effects are involved in wobble tautomerisation.

In the replication machinery, during the polymerase dNTP sampling, the sample is rejected if G is mismatched against T. We investigate a pathway that connects the wobble mismatch to a Watson-Crick-like pairing (shown as pathway 1 in Fig. 10-1), leading to base misincorporation through a phosphodiester bond formation. In this scheme, proton transfer must occur for the bases to slide to a Watson-Crick-like pairing, either classically, via an "over the barrier" mechanism, or via quantum tunnelling. To avoid replication errors, the polymerase should reject such mismatches; otherwise, the wrong base pairing can undergo further proton transfer, connecting two Watson-Crick-like tautomeric forms (shown as pathway 2). Additional pathways are explored in supplementary note 1.

We investigate the reactions: wobble(G-T) $\rightleftharpoons$ G-T* (reaction 1) and wobble(G-T) $\rightleftharpoons$ G*-T, whereby the reactants start as a wobble mismatch and, via proton transfer, result in a Watson-Crick-like conformation. We determine that the reaction wobble(G-T) to G*-T proceeds through a G-T* intermediate state in a step-wise mechanism. The minimum energy pathway can therefore be described by two steps; in the first step, the wobble(G-T) passes through the transition state of wobble(G-T) $\rightleftharpoons$ G-T*. In the second step, through an intermediate local minimum, the G-T* intermediate converts to G*-T [239, 231]. In comparison, the G-T* reaction (reaction 1) contains one transition state with no intermediate minimum.

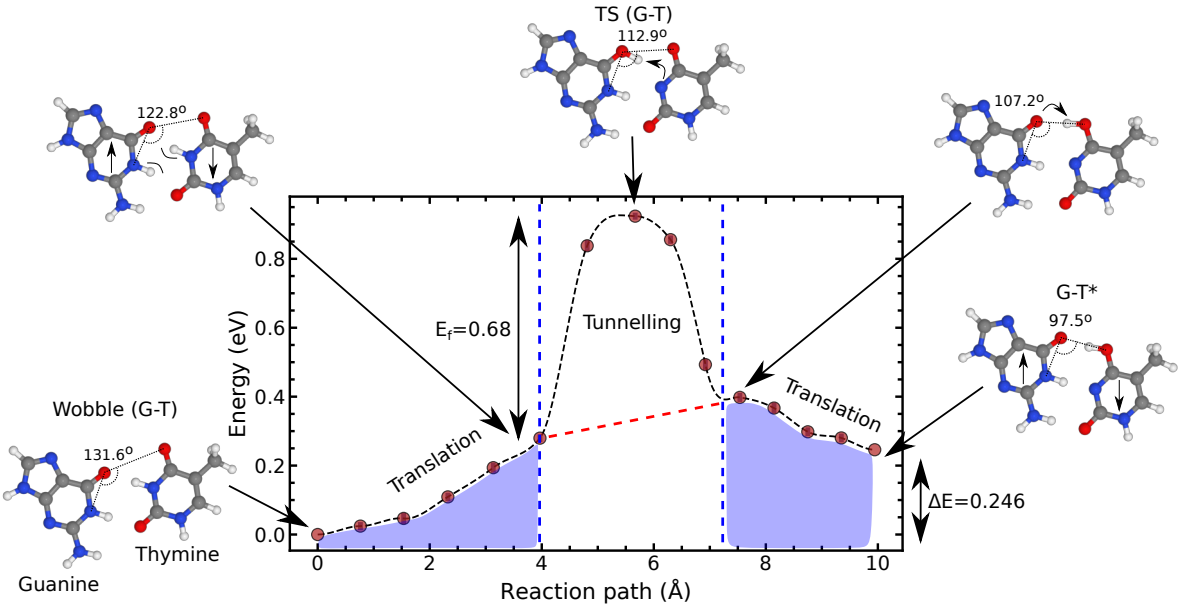


Figure 10-2 Minimum energy path of the wobble(G-T)⇌G-T* proton transfer reaction pathway. The reaction 1 paths are obtained using a machine learning approach to the nudged elastic band method. The reaction path contains classical rearrangement of the bases and a high reaction barrier where we suppose the proton can tunnel through.

Fig. 10-2 shows the minimum energy path of this reaction, for which the forward barrier is $0.926\,\mathrm{eV}$ and the reverse barrier is $0.680\,\mathrm{eV}$. We perform a normal mode analysis to calculate the free energy values of the reactant, transition state, and product. We determine that the free energy values are smaller than the electronic energy barriers. The free energy contributions reduce the forward barrier by 20% and the reverse barrier by 30%, resulting in a free energy profile consistent with Li *et al.* [210]. A summary can be found in supplementary note 1, and a detailed comparison of the reaction barrier parameters to the literature in note 3.

On the wobble(G-T) ⇌ G-T* (reaction 1) reaction path, we observe three regions (see Fig. 10-2). The first region (0-$4\,$Å) largely corresponds to the collective movement of the bases relative to each other as they drift to a Watson-Crick-like bonding angle. In this region, the ΔE is essentially constant as the molecules move over a flat PES in which weak Van der Waals interactions dominate. The fast and activated proton transfer occurs between 4-$7.5\,$Å. In this region, the proton of the Thimine N-H bond first transfers to the oxygen of the carboxylic group of G (as described by the arrow in the transition state of Fig. 10-2). The same proton subsequently hops back to the nearest oxygen of T. Finally, the region of the reaction path closest to the product ($> 7.5\,$Å) corresponds to a further collective translation of the bases toward a Watson-Crick-like configuration, with little rearrangement in the bond of the transferred proton.

Despite several previous attempts to model the creation of G-T wobble mismatches [231, 239, 210], the presence and role of quantum effects in this reaction have not been addressed, with previously reported semi-classical models severely underestimating the experimental reaction rates. In the following, we introduce a first-principles based quantum dynamic approach for modelling

proton tunnelling in a realistic cellular environment, which accounts for the noise and thermal fluctuations of the biological system. We then employ this method to calculate the G-T wobble mismatch reaction pathway to the Watson-Crick-like configuration and the double proton transfer scheme in the Watson-Crick-like configuration (see Fig. 10-1). Quantum and classical contributions to the reaction rate are determined, and we discuss the contribution of proton tunnelling in forming Watson-Crick-like tautomers within the polymerase active site.

The open quantum systems approach employed in this study is based on Caldeira and Leggett's quantum Brownian Motion model [240] in which the protons in the hydrogen bonds are embedded in an Ohmic bath of quantum oscillators, which represent the cellular environment. The interactions between the DNA and the environment are integrated over time using the path integral formalism introduced by Feynman and Vernon [241]. The equivalent phase-space version is given by the Wigner-Moyal Caldeira and Leggett equation (WM-CL),

$$\frac{\partial W}{\partial t} = \underbrace{-\frac{p}{\mu}\frac{\partial W}{\partial q} + \frac{\partial V}{\partial q}\frac{\partial W}{\partial p} - \frac{\hbar^2}{24}\frac{\partial^3 V}{\partial q^3}\frac{\partial^3 W}{\partial p^3} + \mathcal{O}\left(\hbar^4\right)}_{\text{Schrödinger dynamics}} \underbrace{+\gamma\frac{\partial pW}{\partial p}}_{\text{Dissipation}} \underbrace{+\gamma\mu k_{\mathrm{B}}\tilde{T}\frac{\partial^2 W}{\partial p^2}}_{\text{Decoherence}} . \tag{10.1}$$

Here, W is a quasi-probability density encapsulating the proton's quantum state as a function of both position (q), and momentum (p) [141, 242]. The first set of terms in Eq. 10.1 corresponds to the Schrödinger dynamics of a particle with effective mass μ. The subsequent two terms correspond to the dissipation and decoherence arising from the coupling to the quantum bath. Here, γ is the phenomenological friction constant that describes the strength of the coupling to the bath [240], k_{B} is Boltzmann's constant and $\tilde{T}$ represents the effective bath temperature.

The advantage of employing an open quantum system model is that it incorporates the interactions with the local environment in the quantum dynamics. These interactions significantly affect the system's dynamics and can either impede or encourage the system's evolution, known as a quantum Zeno or anti-Zeno effect [243]. Furthermore, the coupling to the environment results in quantum dissipation, such that the information in the system is lost to its environment and decoherence, where a quantum system loses its wave-like properties. As a consequence, classical behaviour emerges.

Assuming that the system-to-environment coupling constant is dominated by the thermal fluctuations of the surrounding water molecules, we can estimate the value of γ. Water has a vibrational spectrum in the range $3300 - 3900\,\mathrm{cm}^{-1}$ [244]. So, assuming that the fastest oscillators in this range dominate, we use an Ohmic spectral density for our environment oscillators [240] having a coupling parameter $\gamma = 3900\,\mathrm{cm}^{-1}$. We determine the quantum contribution to the reaction rate by monitoring the flux of the density passing through the transition state; see supplementary note 2 for further details. The forward and reverse reaction rate constants, k_{f} and k_{r}, are obtained from,

$$k_{\mathrm{f,r}} = \frac{\kappa}{h\beta}e^{-G_{\mathrm{f,r}}\,\beta} , \tag{10.2}$$

where $\beta = 1/(k_{\mathrm{B}}T)$ and $G_{\mathrm{f,r}}$ corresponds to the Gibbs free energy barrier of the forward and reverse reaction, respectively. The tunnelling factor, κ, encapsulates the quantum-to-classical contribution to the rate, incorporating quantum effects such as tunnelling and non-classical reflections.

Firstly, we determine the quantum and classical rates for reaction 1 using our open quantum systems approach. Reaction 1 has a prohibitively high and wide reaction barrier (see Fig. 2), resulting in a low classical and quantum reaction rate. We evaluate that the quantum-to-classical ratio is small, $\kappa = 1.02$, suggesting that tunnelling is negligible; here, in this case, dissipative and decoherent effects from the biological environment suppress the tunnelling. We find that the overall reaction rate is dominated by an over-the-barrier classical mechanism, with a value of $5.244 \times 10^{-1}\,\mathrm{s}^{-1}$ - which is consistent with both the experimental value (0.6-$68\,\mathrm{s}^{-1}$ [77, 238]) of the G-T wobble system in DNA. The reaction rate is several orders of magnitude smaller than the dNTP unbinding rate, which is of the order $70\,000\,\mathrm{s}^{-1}$ [77]. Furthermore, we determine the effect of isotopic substitutions on the reaction rate and found that the reaction rate is essentially unaffected by deuterium substitution (KIE=1.1). Consequently, due to the slow reaction rate, the dNTPs unbinding rate and subsequent base rejection compete with the proton transfer mechanism. As a result, statistically, some of the wobble mismatches will eventually diffuse from the polymerase's active site before proton transfer occurs. Since the diffusion timescale competes with the proton transfer timescale, the final population of tautomers incorporated will be reduced as accounted for by the kinetic network in Ref. [77].

To compare how the change of environment and subsequent change to the reaction profile impacts the tunnelling, we extract the free energy pathway data from Li *et al.*[210] and apply our tunnelling approach. A detailed description can be found in supplementary note 3. In summary, we determined that regardless if the G-T wobble is exclusively in an aqueous solution or a more complex DNA environment, the tunnelling is primarily suppressed to the degree that it is insignificant.

However, we note that the PES in Fig. 10-2 describes three fundamentally different molecular motions, and only the inner barrier (section 2 in Fig. 10-2) corresponds to the proton transfer between the bases. In contrast, regions 1 and 3 correspond to overall translations of the bases without significant changes in the hydrogen bond length. This PES topology is compatible with a tunnelling-ready state composed of the reactant's activated structure seen at the end of region 1. The activation process concerns the reorganisation of the heavy atoms where thermal energy is required for the reactants to reach an activated tunnelling-ready state, whereby the reactant and product states become similar in energy.

Here we explore the minimum energy pathway of proton transfer in the tunnelling-ready state. Further details of the methods can be found in supplementary note 1. The subsequent minimum energy pathway is shown in Fig. 10-3. Here, the reaction pathway shows three pseudo-minima corresponding to the bases already part-way slid into a Watson-Crick-like shape, the second, where the proton has transferred to the other base, and the third, the return of the proton back to the same

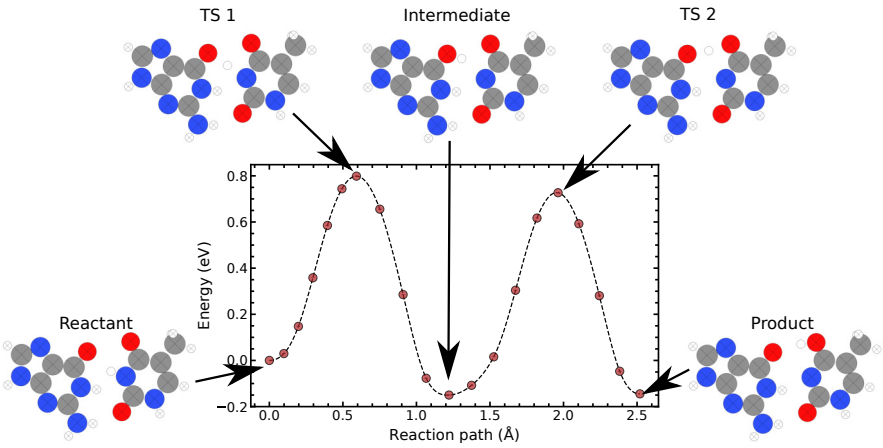


Figure 10-3 Tunnelling-ready Minimum energy path of the wobble(G-T)⇌G-T* reaction. The proton transfer reaction pathway, reaction 1, assumes that the bases have already partly slid into a Watson-Crick-like shape. Each minima and maxima along the path are labelled. Crossed-out atoms indicate that they have been constrained.

base. The last two minima indicate that if we assume that the proton transfer is much faster than the rest of the atomic motion during the reaction, a bifurcation of the reaction pathway is possible. In fact, after the first initial proton transfer, the rest of the atoms could rearrange, trapping the population in the middle well.

To calculate the contribution of quantum tunnelling in this activated tunnelling-ready state, we re-evaluate the PES considering only where the proton is transferring. We then calculate the rate of proton tunnelling from the tunnelling-ready state through region 2. We calculate the inner barrier section (between 4-$7.5\,\text{Å}$) by taking the image of the start of the barrier from reaction 1 and assuming that the local polymerase environment has thermally induced this conformation change. Using this approach, we calculate the quantum contribution to the reaction rate and find that the overall rate is much larger. The rate is now $1.279 \times 10^{-1}\,\text{s}^{-1}$, with a $\kappa = 99.0$ indicating a large contribution from tunnelling. By substituting hydrogen with deuterium, we find that the tunnelling-corrected reaction rate exhibits a KIE of 10.15, which is compatible with experimental results [77] that predict a three-fold reduction in rate.

We now focus on explicitly how the polymerase active site interacts with the G-T wobble mismatch and the tunnelling-ready state. Free energy pathway calculations from Li *et al.*[210] suggest that the polymerase introduces a 46% increase in the proton transfer reaction barrier. However, as the DFT calculations show, for the wobble(G-T)⇌G-T* reaction to occur, the nucleotide dimer must first be compressed into a "tunnelling-ready" state. Therefore, it is desirable to know whether this state is populated in a biologically relevant thermal ensemble. To this end, hybrid quantum-classical Quantum-Mechanical/Molecular-Mechanical Molecular-Dynamics (QM/MM MD) simulations were performed, wherein the entire polymerase enzyme and solvent are included explicitly in the simulation system. Many short replica simulations were computed from the wobble configuration obtained through a crystal structure, totalling over $2800\,\text{ps}$ of QM/MM MD. This investigation was repeated

without the enzyme to highlight the compressing effect of the enzyme's "thumb" region. Both simulation systems are described and illustrated in supplementary note 4.

A metric for the overlap distance between the simulation snapshot and the "tunnelling-ready" state from the ML-NEB calculations is defined as Δ. The metric allows the frequency of how often the state is populated in a biologically relevant thermal ensemble to be determined. Using these QM/MM MD simulations, a cumulative histogram of this data is shown in Fig. 10-4 for both aqueous DNA and the polymerase DNA complex. Full computational details are provided in supplementary note 4.

The MD simulations demonstrate that the G-T dimer remains in either a wobble configuration consistent with the reactant configuration or exists in a transitory unbound state (see the schematic representation in Fig. 10-1). Among all the QM/MM MD replica simulations, 0.003% of the trajectory was within $0.096\,\text{Å}$ root-mean-squared distance from the "tunnelling-ready" state for the enzyme-DNA complex. Without the enzyme, no Δ values below 0.096 were observed. The lower delta regime corresponding to the TRS was informed by considering the Δ difference between the 5th and 6th ML-NEB data points from Fig. 10-2. Assuming a uniform distribution of events, this is equivalent to the dimer compressing once every $35.2\,\text{ps}$ in the polymerase active site. Despite the approximately $0.275\,\text{eV}$ energetic penalty to compression shown in Fig. 10-2, our dynamical simulations show that this state is reachable within a realistic biological environment. Crucially, without the presence of the enzyme, no population was found below $\Delta = 0.97\,\text{Å}$, suggesting that the enzyme facilitates the compression. These results justify performing proton transfer calculations for the wobble(G-T)⇌G-T* reaction from such a compressed "tunnelling-ready" state as shown in Fig. 10-3. While the proton transfer mechanism starts from a more compressed G-T wobble conformation, reaction rate calculations must now also consider the sparsity with which this compressed state is observed, as the barrier shown in Fig. 10-3 is only accessible less than 0.003% of the time.

On the other hand, for the G*-T ⇌ G-T* (reaction 2 in Fig. 10-1), the barrier is considerably smaller than for the wobble transfer reaction (reaction 1 in Fig. 10-1), $0.356\,\text{eV}$ vs $0.926\,\text{eV}$. As shown in Fig. 10-5, the region of the PES comprised between 0.0 and $0.5\,\text{Å}$ corresponds to the small translation of the two bases towards each other to facilitate the transfer. First, the middle proton in the N-H-N bond transfers (denoted by the arrow in Fig. 10-5). Then, the O-H-O proton transfers, as observed in the secondary hump after the transition state. We determine that the reduced mass of the double proton transfer is $1.49\,m_p$ at the transition state (see supplementary note 1).

Here, the forward reaction barrier, $E_f = 0.356\,\text{eV}$, and asymmetry $\Delta E = 0.017\,\text{eV}$ are small compared to the wobble-to-enol transfer and compares well with previous calculations ($E_f = 0.34\,\text{eV}$ [210]). The low reaction barrier leads to a fast proton transfer with a large forward and reverse reaction rate on the order of $10^8\,\text{s}^{-1}$. We use the open quantum systems model to determine a quantum-to-classical rate ratio of $\kappa = 18.1$ and a KIE of 4.25. The high κ and KIE for this reaction suggest that quantum effects play a significant role in reaction 2 and that, due to the fast pro-

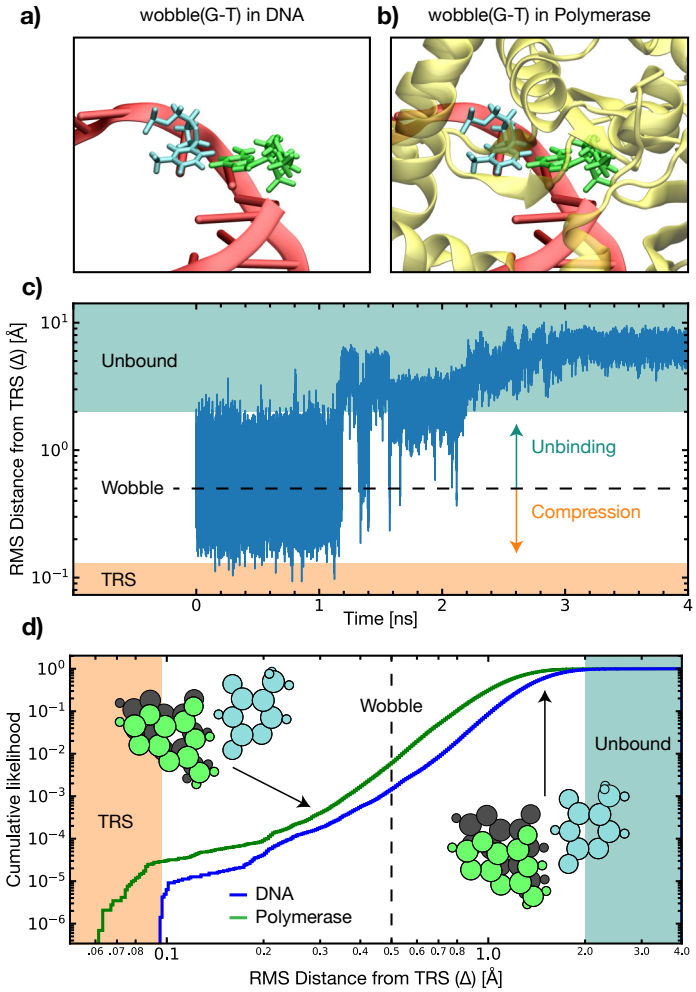


Figure 10-4 Dynamical investigation into the biological relevance of the compressed "tunnelling-ready" state (TRS) of the wobble(G-T) mismatch. The compression of the wobble(G-T) mismatch is considered in a DNA insertion site with the polymerase enzyme (panel b)) and without the enzyme (panel a)). An RMS distance is defined to the TRS and plotted c) during a single long molecular dynamics trajectory and d) aggregated from over $2800\,\mathrm{ps}$ of QM/MM MD simulations. In panel c), the RMS distance to the wobble(G-T) configuration is shown as a black dashed line, and two additional regimes are illustrated. Firstly, an unbound regime is defined with $\Delta > 2.0\,\text{Å}$, and a set of "tunnelling-ready"/compressed states with $\Delta < 0.096\,\text{Å}$. In panel d), the cumulative likelihood across a range of Δ values is graphed for the Polymerase-DNA complex (green line) and aqueous DNA (blue line). In this context, the cumulative likelihood determines the probability of finding the dimer at an RMS distance below the given value. Two example conformations are shown relative to the TRS (grey circles) position.

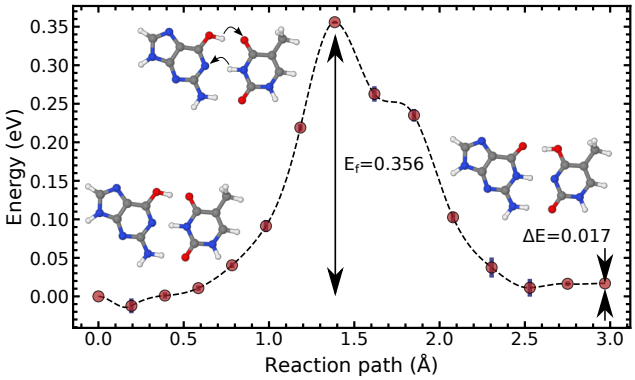


Figure 10-5 Minimum energy path of the G*-T⇌G-T* reaction. The double proton transfer reaction pathway, reaction 2, assuming the conversion to a Watson-Crick-like state has already occurred.

ton transfer timescale, the system can quickly reach equilibrium. After a single proton transfer successfully forms the tautomeric Watson-Crick-like form, the protons can continue to transfer between the bases (reaction 2) via a fast double proton transfer. Consequently, following another step of replication in the polymerase, an error will likely be induced on both daughter strands, as the enol forms will readily mismatch with the wrong base [111].

In summary, we have employed quantum chemical calculations to determine the reaction pathway of several reactions for generating tautomers of the G-T wobble mispair. We applied an open quantum systems approach to account for the decoherent and dissipative local environment [240] and identified quantum and classical contributions to the reaction rates. For the wobble(G-T) ⇌ G*-T mechanism, we find that the reaction proceeds via a step-wise process involving G-T*. Consequently, we focused on wobble(G-T) ⇌ G-T*. The proton transfer reaction from the wobble to the Watson-Crick pathway has a significantly high and broad reaction barrier, which implies an insignificant contribution from quantum tunnelling and a slow classical rate. We noted that for the wobble(G-T)⇌G-T* reaction to occur, the nucleotide dimer must first be compressed into a "tunnelling-ready" state, we probed this state using QM/MM MD to determine how likely it is populated in a biologically relevant thermal ensemble. We evaluated that this state is more likely to be populated in the polymerase environment and leads to an increase in quantum tunnelling.

As highlighted by previous computational studies, the role of proton transfer in spontaneous mutation is a complex affair [115, 233, 210]. However, the proton transfer mechanism in the polymerase is a prominent candidate as a source of mutations as it is later in the replication cycle and could play a more significant role than other equilibria competing during mutation. Furthermore, for mechanisms involving double-stranded Watson-Crick DNA, it needs to be clarified if the helicase or another mechanism reduces the proton transfer populations via electrostatic destabilisation or exonuclease proofreading mechanisms.

To conclude, our model predicts tunnelling rates that match the experimental NMR observed rates to a high degree of accuracy—opening the possibility that quantum mechanics is required to

explain the biologically relevant functionality of polymerase.

Acknowledgements

This work was made possible through the support of the Leverhulme Trust doctoral training centre grant number DS-2017-079 and from the John Templeton Foundation grant number 62210. M.S. is grateful for support from the Royal Society (URF/R/191029). We acknowledge helpful discussions with the members of the Leverhulme Quantum Biology Doctoral Training Centre; particular thanks go to Johnjoe McFadden. Further thanks go to Antonio Pantelias, who offered many productive conversations. In addition, the authors thank the University of Surrey for access to Eureka. Via our membership of the UK's HEC Materials Chemistry Consortium, funded by EPSRC (EP/R029431) and the UKCP Consortium, funded by EPSRC grant ref EP/P022561/1, this work used the ARCHER2 UK National Supercomputing Service.

Data availability

The data for the reaction pathway is available on Github. Additionally, data presented in this article are available from the corresponding authors upon reasonable request.

Author contributions

L.S., M.S., and J.A-K. conceived and designed this research. L.S. built the computational apparatus. M.W. provided the QM/MM calculations. All the authors contributed to the preparation of the manuscript.

11. The biological relevance of tautomerism

In previous chapters, it has been demonstrated that the Double Proton Transfer (DPT) in base pairs of DNA can lead to metastable tautomeric nucleobases, which may lead to mutations. For this pathway to be biologically relevant however, it is not sufficient to indicate a possibility of populating these states, instead, it is prudent to obtain a description of the probability of misincorporation of nucleotide mismatches and the resulting changes to protein synthesis.

Our cells are constantly undergoing a cycle of multiplication, providing countless opportunity for tautomers to encounter strand separation and lead to mismatches. However, a point mutation in a random part of DNA in the human body is exceptionally unlikely to cause significant effects. This may seem counter-intuitive, but the majority of cells in our body are hyper-specialised with only a small part of their DNA coding for genes, and a smaller fraction still that will be translated into proteins. Diseases such as cancer are often proposed as a dangerous consequence of mutations. Changing even a single letter of DNA can modify the structure of a protein via a 'missense' amino acid mutation, or insert a stop codon creating a completely non-viable protein. Despite this, even nonsense mutations such as an incorrectly terminated protein will not cause lasting issues in the average (somatic) cell, as if its viability is compromised, it will simply not continue to produce descendants.

There are two remaining instances of replication where spontaneous mutation can have drastic effects, in pluripotent cells such as stem cells, and in the production of gametes. Pluripotent cells, found for example in our bone marrow maintain the ability of expressing any part of their genome, and during development, a single stem cell may be the seed from which entire organs with 10^8 or more cells form[245]. Similarly, unwanted mutations in gametes produced via meiosis pose significant threat, if a gamete is produced with a genotype causing disease, this mistake will appear in every single cell of the offspring.

11.1. Point mutations during the production of gametes

While a point mutation in a random part of the genome of somatic cells in the human body is unlikely to have any significance, during the production of gametes even a single point mutation can be devastating (if its in the coding portion of DNA). In this section a model is proposed to estimate the fraction of gametes produces with mutated proteins due to the misincorporation of

double proton transfer products of GC.

A simplified reaction network is used to determine the equilibrium population of tautomeric G and C in single stranded DNA. The constant exchange of canonical GC into tautomeric G*C* interacts with the irreversible mechanism of strand separation, carried out by the helicase enzyme.

$$\begin{array}{ccc} G^*C^* & \xrightarrow{k_h} & G^* + C^* \\ k_f \uparrow\downarrow k_r & & \\ GC & \xrightarrow{k_h} & G + C \end{array} \tag{11.1}$$

This gives us a way to calculate the equilibrium population as a function of three rate constants $K_{\mathrm{G^*+C^*}}(k_f, k_r, k_h)$. The atomic separation speed induced by a helicase is taken to be 1 Å/ps, as reported in [1, 3]. Taking the forward rate from [213] as $k_f = 1500/\mathrm{s}$, and the decay rate from G*C* to canonical GC from ensemble lifetime calculations in Chapter 9 ($k_r = 10^{13}/\mathrm{s}$), the dynamics this network describes are shown in Figures 11-1 and 11-2.

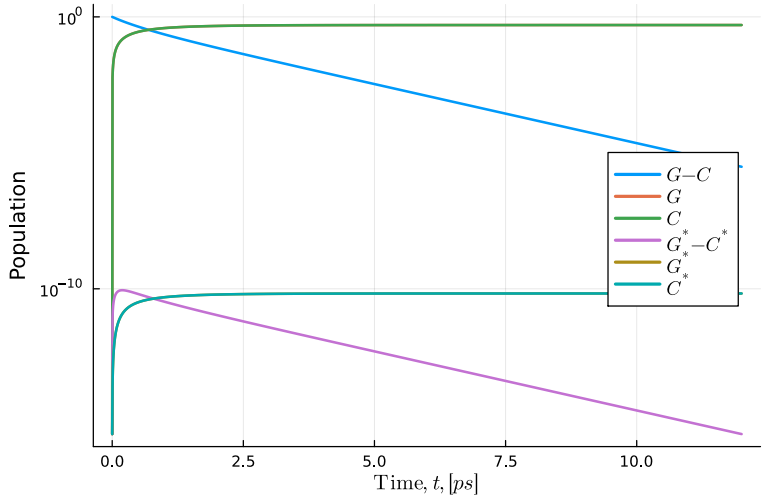


Figure 11-1 Reaction dynamics of GC tautomerism in strand separation.

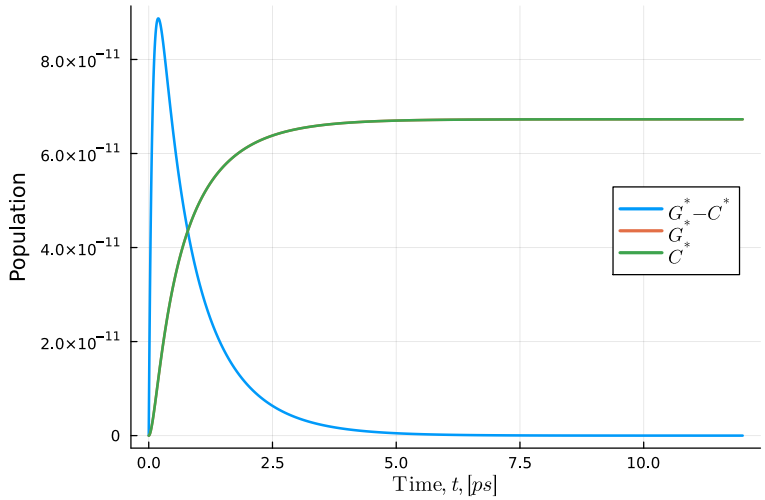


Figure 11-2 Reaction dynamics of the GC tautomer G*C* , and single-stranded tautomers G* and C* in a simple model of strand separation.

Since the helicase acts unidirectionally the population of GC and G*C* are continuously separated into G + C, and G* + C* . Despite the fast rate of G*C* decay, the helicase traps a finite population of single-stranded tautomers. Assuming, each G* and C* will eventually form a permanent mismatch, each successful tautomerisation produces two mismatches: G*T and AC*.

Meiosis is the mechanism by which gamete cells are created. Starting from a cell with the regular amount of genetic material (diploid), a replication cycle occurs, doubling the DNA. Following this, two meiosis cycles split this cell into four granddaughter cells. The granddaughter cells are haploid, each containing only half the genetic material. Thus we can estimate the expected number of mutations in a gamete via the following equation:

$$N_{\text{mutations}} = \frac{1}{2} \times N_{\text{coding GC}} \times K_{\text{G}^*+\text{C}^*}(k_f, k_r, k_h) \tag{11.2}$$

The factor $1/2$ arises from meiosis: the genetic material is doubled, before it is halved twice.

Data from the human genome sequencing[14] provides a GC fraction of 0.41. Using the approximation that 5% of DNA is coding, and 2.5×10^8 base pairs of the human genome[14], we obtain approximately 5 million coding GC base pairs.

DNA is transcibed into amino-acids by reading triplet codons of nucleobases in the ssDNA. The codons corresponding to amino acids are non-unique meaning that there are several combinations encoding each peptide. To determine the impact of tautomerism a Python code was used to investigate how the substitutions C→T and G→A would affect amino acid codons. The reverse substitutions were ignored as AT does not form a stable DPT product[213].

Table 11-1 shows all possible amino acid substitutions arising from the DNA point mutations C→T and G→A. Considering that one such mutation would occur in a given codon, the likelihoods reveal whether the mutation is likely to be dangerous, i.e. non-conservative or a stop codon.

Table 11-1 can give us approximate probabilities of different consequences of point mutations. Given that a point mutation has arisen due to the DPT in GCthe chance of a non-synonymous point mutation within our model is 65.6%. Table 11-2 provides further detail on the types of point mutations that are predicted by our model.

Equation (11.2) was applied to a range of values of k_f and is plotted in Figure 11-3. The decay rate from G*C* to canonical GC was taken from ensemble lifetime calculations in Chapter 9 ($k_r = 10^{13}$/s). The model defined in Equation (11.2) is directly proportional to the equilibrium population of G*C* tautomers, and it appears the stable state of the reaction network shown in Equation (11.1) is linearly dependent on k_f. The forward rate from [213] is shown as a black vertical line, this $k_f = 1500$/s would suggest that approximately 1 in 40,000 gametes/children would have a premature stop codon in a part of their coding DNA (a near-worst-case scenario). For any amino acid substitution to occur this rate would be approximately 1 in 3000.

Is this value reasonable? A staggering 1 in 33 babies are born with defects[246], accounting for up

Table 11-1 Likelihood and classification of non-synonymous amino acid point mutations given a successful pair of mismatches due to GC tautomerism. Synonymous/silent mutations are omitted.

Mutation	Likelihood	Classification
A → T	40.0%	non-conservative
A → V	40.0%	conservative
C → Y	66.7%	non-conservative
D → N	66.7%	non-conservative
E → K	66.7%	non-conservative
G → D	20.0%	non-conservative
G → E	20.0%	non-conservative
G → R	20.0%	non-conservative
G → S	20.0%	non-conservative
H → Y	66.7%	non-conservative
L → F	28.6%	conservative
M → I	100.0%	conservative
P → L	40.0%	non-conservative
P → S	40.0%	non-conservative
Q → *	66.7%	nonsense / STOP
R → *	7.7%	nonsense / STOP
R → C	15.4%	non-conservative
R → H	15.4%	conservative
R → K	15.4%	conservative
R → Q	15.4%	non-conservative
R → W	7.7%	non-conservative
S → F	22.2%	non-conservative
S → L	22.2%	non-conservative
S → N	22.2%	conservative
T → I	50.0%	non-conservative
T → M	16.7%	non-conservative
V → I	50.0%	conservative
V → M	16.7%	conservative
W → *	100.0%	nonsense / STOP

Table 11-2 Conditional probabilities of amino acid point mutation classifications within the model of GC tautomerism.

Synonymous			34.4%
Non-Synonymous	Any		65.6%
	Nonsense	STOP	7.3%
	Missense	Conservative	17.7%
	Missense	Non-Conservative	40.6%

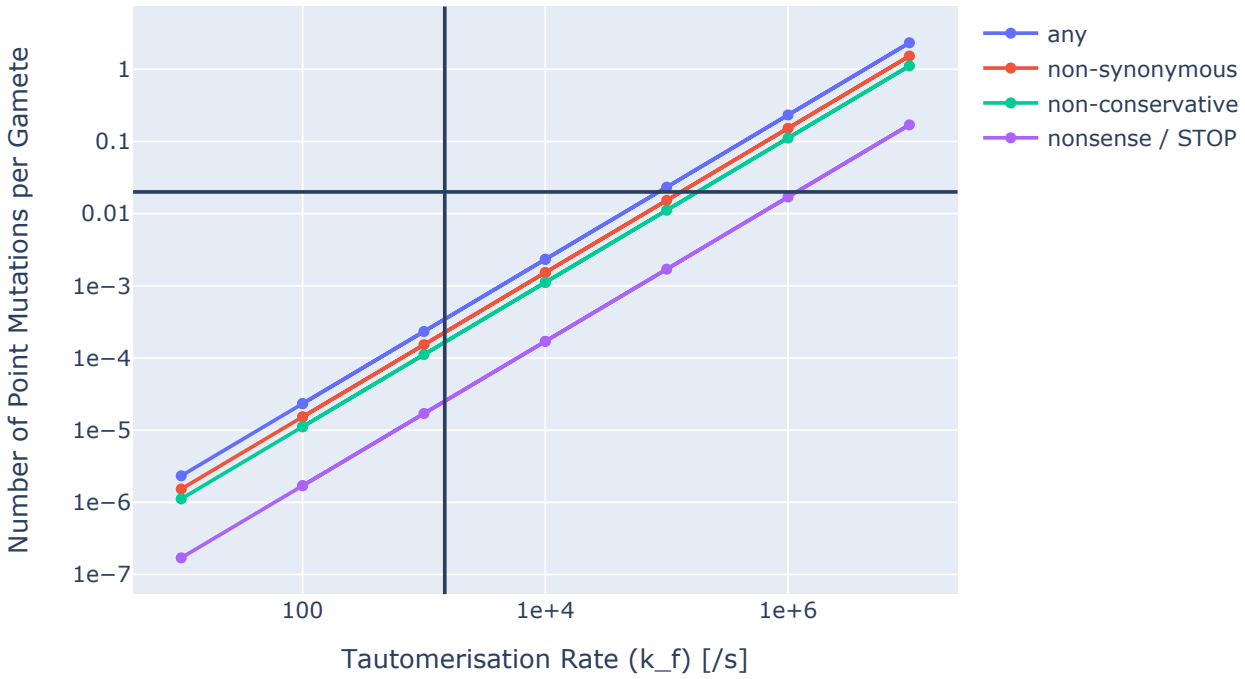


Figure 11-3 Number of point mutations per gamete based on the rate of tautomerisation in GC

to a fifth of infant mortality[247]. Premature stop codons in mRNA are the cause of a large number of diseases including cystic fibrosis, Duchenne muscular dystrophy, β-thalassemia, and several cancers[248]. A survey of the human genome found that 12% of reported single point mutations led to premature stop codons[249]. Once other mutation mechanism such as deletions, splicing, and insertions are included, nonsense mutations account for up to a third of all inherited disease[250]. Kumar *et al.* report that 1 in 50 people are affected by single-gene disorder[251]. While spontaneous point mutations due to GC tautomerism are certainly not the only cause of single-gene disorders, a rate of nonsense mutations of 1 in 40,000 gametes is certainly not ruled out by the measured prevalence of single-gene disorders.

In this model several simplifications and assumptions have been made, nonetheless we provide a simple preliminary connection between tautomerism and genetic disease. For the sake of clarity, the approximations in the model are detailed here.

Firstly, we assumed that AT is not a viable candidate for producing point mutations via the DPT. In equilibrium DNA, A*T* is not stable [213], however Chapter 8 and [3] have shown the stability of A*T* to be restored by non-equilibrium separations, it may be that helicase enzymes do in fact trap some AT tautomers. It is still expected that tautomerism is more feasible via GC as there will be more equilibrium population of tautomers to trap, compared to AT. Including AT tautomerism in the model would increase the predicted rate of mutations.

Additionally, the portion of DNA that a point mutation would be relevant in is estimated by the fraction of coding DNA. In reality, non-coding DNA thought to originate from ancient viral DNA informs gene expression and thus can have an indirect effect on disease[252, 253]. A more sophisticated model would consider the locus of mutations more carefully.

While the two assumptions presented above lead to an underestimate of mutation rate, there are several approximations that lead our model to overestimate the mutation rate. Firstly, considerations of the complex processes of meiosis and DNA replication itself has necessarily been simplified. For one, within our model we do not consider that mutations arising during meiosis and the production of gametes could affect the viability of the gamete, *i.e.* a point-mutation may never lead to a genetic disorder as the gamete is sufficiently altered and may not achieve successful fertilisation. Furthermore, it is not entirely clear whether a premature stop codon will act as a dominant or recessive gene. Some studies report that up to 90% of mutations are recessive to the wild type[254].

The assumption that each tautomeric G^*C^* that has successfully passed strand separation results in two point mutations is also likely an overestimate. This assumption is three fold; tautomeric nucleobases in ssDNA do not decay back to their canonical form, meaning that each separated G^*C^* results in two mismatches G^*T and AC^*; due to resembling the Watson-Crick structure G^*T and AC^* are not rejected by the polymerase or removed by other error-correcting mechanisms; and finally, the helicase strand separation dynamics are sufficiently fast, effectively trapping tautomeric populations.

This final assumption is the natural starting place for further theoretical consideration as predicted population of ssDNA tautomers is highly dependent on results presented in this thesis. The next section will further discuss the shortcomings of current reaction rate models in inherently out-of-equilibrium situations such as DNA replication.

11.2. Tautomerism in the dynamical environment of strand separation

Building upon Equation (11.1), one can consider the interaction of a GC base pair with a helicase enzyme in the following way:

$$\begin{array}{ccccccc} G^*C^*_{\text{dsDNA}} & \xrightarrow{k_h} & G^*C^*_{\text{HELICASE}} & \xrightarrow{k_h} & G^*_{\text{ssDNA}} & + & C^*_{\text{ssDNA}} \\ \text{DPT} \uparrow\downarrow & & \text{DPT}' \uparrow\downarrow & & \text{IBT} \uparrow\downarrow & & \text{IBT} \uparrow\downarrow \\ GC_{\text{dsDNA}} & \xrightarrow{k_h} & GC_{\text{HELICASE}} & \xrightarrow{k_h} & G_{\text{ssDNA}} & + & C_{\text{ssDNA}} \end{array} \tag{11.3}$$

The GC begins in complex with the helicase in a state equivalent to double-stranded DNA (bottom left of Equation (11.3)), see also base pair N-1 in the PcrA helicase work in Chapter 9. Double Proton Transfer (DPT) can lead to the formation of the $G^*C^*_{\text{dsDNA}}$ tautomer with one set of dynamics. Both forms of GC are translocated by the helicase into the base pair N position via a rate constant k_h. Following this translocation, the base pair is now situated with a nearby asparagine residue

which greatly discourages the formation of the tautomer. We have shown however that even the small separations at the start of strand separation from $G^*C^*_{\text{HELICASE}}$ to $G^*_{\text{ssDNA}} + C^*_{\text{ssDNA}}$ can trap any tautomeric population. This is further illustrated in Figure 11-4:

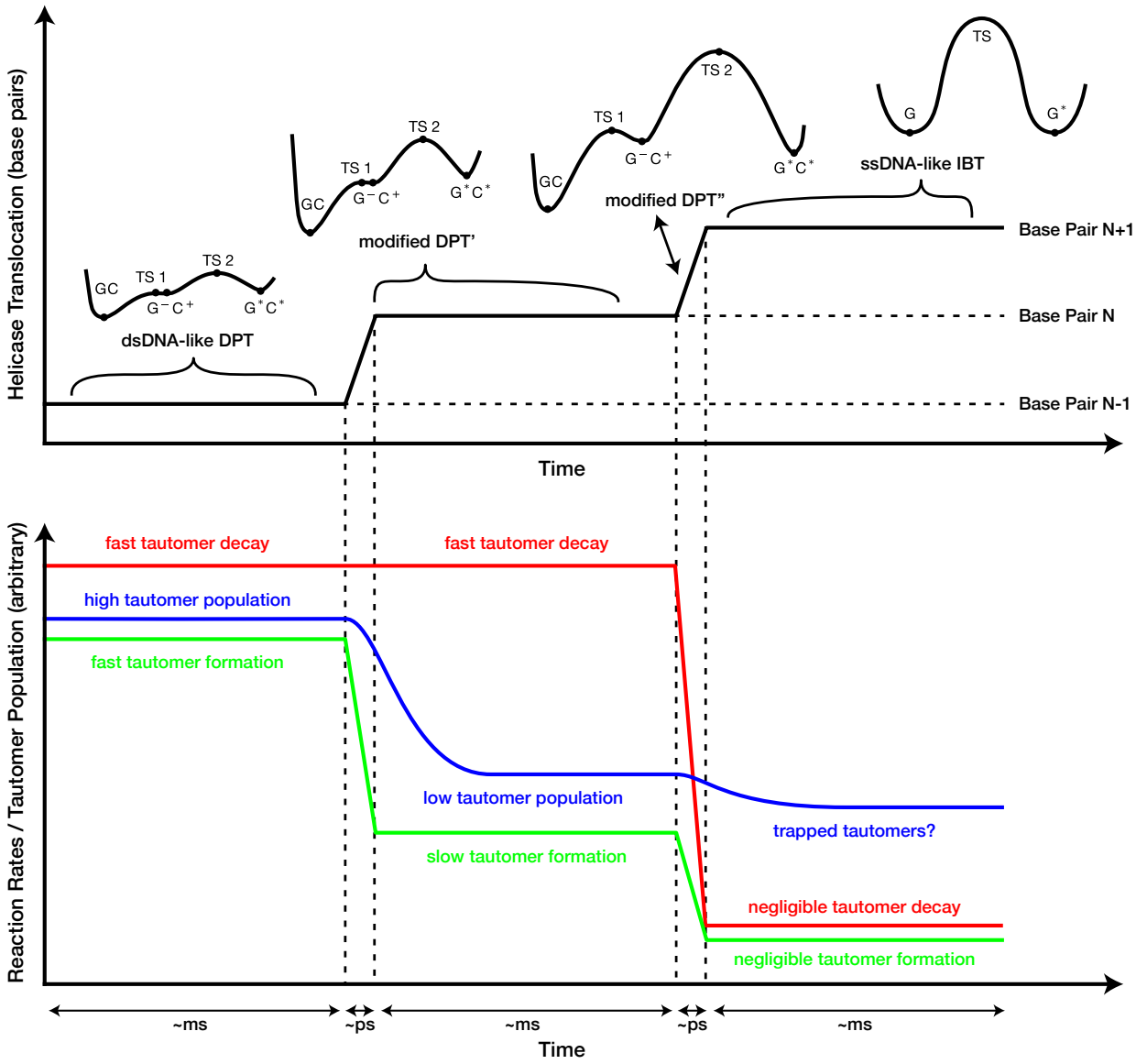


Figure 11-4 Illustration of different proton transfer regimes during strand separation.

Figure 11-4 illustrates the different proton transfer energetics experienced by a GC dimer undergoing translocation by a helicase during strand separation. Initially the GC is in a dsDNA scenario in the PcrA helicase complex (base pair N-1) as described in Chapter 9. We have shown that this resembles the GC DPT regime observed by other authors such as in [213, 117, 115]. In this regime we expect a DPT with a fast formation rate, and slightly faster decay rate. Additionally, Slocombe, *et al.* used open quantum systems models in [117] to determine a very high tautomeric population of 1.73×10^{-4}, it was found that the proton transfer is dominated by proton tunnelling. Due to shortcomings of current deep tunnelling corrections to semi-classical transition state theory, an equivalent population for the base pair N scenario of DNA in complex with PcrA Helicase could not

be obtained. In base pair N the DPT is modified and denoted as DPT′. This value is expected to lie between the rate in dsDNA and the equilibrium value expected due to the asymmetry of the reaction (approximately 10^{-22}). A further modification occurs due to the mechanical separation induced in the dsDNA → ssDNA transition as shown in Chapters 7 to 9, we denote this as DPT″. Once in ssDNA it is presumed that the energy landscape determining the Intra-Base Transfer (IBT) is near-symmetric with a large barrier and reaction path[213]. As described earlier, this is a pivotal assumption in the spontaneous mutations hypothesis and urgently needs further investigation (see Chapter 12).

Determining the population of tautomers passing through the dynamics of strand separation is far from trivial. Difficulties with modelling deep tunnelling are compounded by the need for the development of open quantum systems theory to include a time-dependent potential landscape as described at the end of Chapter 9. Nonetheless, knowledge of the behaviour of tautomerism during various stages of strand separation allows us to qualitatively describe expected results. The stepping motor action of helicases such as PcrA includes fast (picosecond) motion as described in Chapters 7 and 8 and longer (millisecond) lag times as an ATP molecule diffuses in, undergoes hydrolysis by the enzyme to release the energy necessary for translocation. During these lag times, we assume the population of tautomers to fully equilibrate to their steady-state. Despite this, in the transition from base pair N to ssDNA, the energies of the transition states are found to sharply increase, trapping the population of tautomers above their equilibrium value for both the modified DPT′ and the IBT.

At the core of the problem is the competition between the decay of the double-stranded G*C* and the atomistic speed of the helicase separation, even with the simple model proposed in Equation (11.1) the dependence on k_h is non-linear.

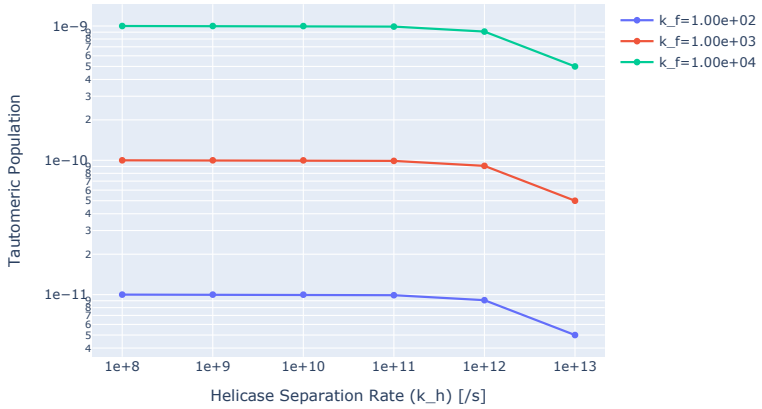


Figure 11-5 Effect of helicase separation speed within a simplistic model of strand separation.

12. Discussion and Conclusions

The source of DNA mutations in the absence of external factors remains an open problem. Watson and Crick's seminal work to reveal the helical secondary structure of DNA also suggested rare protonation configurations of DNA bases could lead to such spontaneous mutations[8]. Their theory was developed by Löwdin who proposed that a Double Proton Transfer (DPT) via quantum tunnelling could generate tautomeric pairs of nucleotides that might lead to mismatches during replication (see Section 2.3)[55]. Challenges in experimental observations of spontaneous mutation due to the rarity of their occurrence and fast atomistic mechanisms, have spawned a rich literature of theoretical investigations using ever-improving computational chemistry models to bridge this gap in understanding. In Chapter 2 we reviewed the biological context and previous literature on which this PhD project was built. Other authors combined state-of-the-art computational chemistry with quantum tunnelling corrections ranging from Wigner corrections to an Open Quantum Systems (OQS) approach[115, 116, 103, 117]. Nevertheless, a considerable oversight remains in the detail of environmental models. Shortcomings in both the number of explicit protein degrees of freedom and their dynamics oversimplified the relevant biology. This thesis details several projects where substantial efforts were made to include the non-equilibrium dynamics induced by replication enzymes and/or explicit models of the biological micro-environment surrounding the Proton Transfer (PT), both of which have been largely overlooked in the literature. Previous computational models of the DPT in DNA had been limited by small system size (often only two nucleotides) with simplistic environmental embedding (no explicit solvent and local protein environment), and disregarded biological dynamics by mapping reactions at minimum energy (zero-kelvin) or equilibrium conformations and applying free energy corrections after the fact. Even when considering ensemble based sampling, all previous research did this in unbiased equilibrium simulations, despite many replication mechanisms being active, and thus injecting energy/force into the DNA.

Chapters 3 to 5 detailed the chemical theory governing reactions in biology, computational chemistry methods to approximate the energy and dynamics of molecular systems, and techniques for mapping reactions within simulations, respectively.

Chapter 6 reports early investigations into hydrogen bonding and the stability of DNA. We report difficulties in extending Molecular Mechanics (MM) models of tautomers in DNA due to the lack of optimised force field parameters for non-canonical DNA. It was shown that in some cases Semi-Empirical (SE) models can approximate higher levels of theory such as Density Functional Theory

(DFT). The energetics and equilibrium geometry of the Watson-Crick (WC) and select non-canonical nucleotide base pairs were detailed for DFT and MM. Following the characterisation of the DNA base pairing via hydrogen bonds, the ensemble stability of aqueous DNA and DNA in complex with PcrA helicase was determined. Other authors have studied the PcrA helicase to reveal the kinetics and dynamics of its stepping-motor mechanism, and the key amino acid residues required for the translocation of DNA (see Chapter 2). The explicit role of strand separation enzymes on the DPT has been completely neglected thus far, both in terms of altering the local chemical environment, and the effect of strand separation dynamics on the PT. While translocation rates have been experimentally determined for many helicases, there remains a need to understand the separation dynamics on an atomistic scale. We postulate that such dynamics would vary the hydrogen bonding distance and thus the energetics of PT. While the average length of the hydrogen bonds in DNA appears stable, fluctuations in the donor-acceptor distance were observed which exhibited a wide variety of duration, opening angles, and opening speeds. It was found that even with no active translocation, the helicase contributed to extending the duration of these fluctuations. Finally, preliminary attempts at mapping intrabase PT and inter-base Double Proton Transfer (DPT) in Quantum Mechanics/-Molecular Mechanics (QM/MM) models of aqueous DNA and the PcrA Helicase-DNA complex were reported. These initial results revealed shortcomings of SE Quantum Mechanics (QM) models and highlight the challenge of obtaining sufficient sampling for Umbrella Sampling (US) at high levels of theory. These difficulties would not be resolved until much later in the thesis (see Chapter 9).

The relevance of tautomers in causing spontaneous mutations in DNA depends on their rate of formation and their likelihood of surviving strand separation by a helicase enzyme. Tautomerisation rates have been calculated for the DPT in both Guanine-Cytosine (GC) and Adenine-Thymine (AT) base pairs, but the effect of strand separation on the energy landscapes remained unreported. Seeking to alleviate this, we published two papers considering the effect of non-equilibrium separation on the DPT in the aqueous WC nucleobase dimers, and the dynamics of a simplified model for strand separation in [1, 3] the results of which are also included in this thesis in Chapters 7 and 8, respectively.

Starting with Guanine-Cytosine, we reported optimised geometries for the nucleobase dimer with an imposed constraint applied to the nitrogen atoms connecting to the ribose sugar of the backbone. As the constrained distance was varied, this would allow us to model the strand separation mechanism of a helicase enzyme, which forcefully separated the duplex into two single-stranded DNA (ssDNA) chains. This simple model of strand separation immediately revealed results that extended our understanding of the DPT under biological conditions. Firstly, the optimised geometries at non-equilibrium separations revealed that the three hydrogen bonds between GC did not stretch symmetrically. As the imposed separation was increased the dimer opened preferentially around one of the two hydrogen bonds involved in DPT in equilibrium GC (denoted B1). This bond B1 remained at equilibrium for up to 1 Å of separation and acted as a pivot for the base pair's opening.

This behaviour was observed for both canonical and tautomeric GC, and in fact, tautomeric GC maintained B1 at equilibrium beyond 2 Å of separation. This stretching asymmetry had not been previously reported or considered in the literature and indicated that the effect of separation on the DPT may impact the synchronicity of the two transferring protons, as well as the reaction energy landscape.

Secondly, to determine whether base pair opening in a biological ensemble would follow a similar pattern, aqueous double-stranded DNA (dsDNA) was simulated in Steered Molecular Dynamics (SMD) with a steering force pulling apart the backbone of the first base pair of the duplex. As reported previously in Section 6.2.2 (Figures 6-8 and 6-10) aqueous DNA is undergoing constant fluctuations around the minimum energy conformation. A separation metric was used which calculated the average donor-acceptor distance relative to the equilibrium distance. It was postulated that nature would evolve to employ a minimal pulling force to reduce the energetic cost of the cellular cycle, while a force would still be necessary to ensure the process happened irreversibly. Observing the separation during SMD revealed that many such fluctuations were taking place and the slope of their rising edge was found to be consistently on the order of 1 Å per picosecond. This separation speed was not correlated to the imposed pulling force, as a similar speed was also observed in unbiased MD simulations. Crucially, this speed placed the atomistic timescale of strand separation on the order of picoseconds, which is two magnitudes faster than what is generally reported in the literature (~ 100 ps) when arguing the irrelevance of tautomers due to their short lifetimes[72]. Additionally, the SMD supported the view that the base pair opening of GC is asymmetrical. An asymmetric breaking of the hydrogen bonds was much more prevalent (97.8% of events have $|\theta| > 3°$), with a clear preference for specific opening angles that have the minimum energy requirements. This results in a bimodal distribution with 39% of opening trajectories following the minimum energy opening suggested by DFT. The negative peak of the distribution agreed well with the optimised geometries from DFT. This dynamical variety suggests that ensemble modelling techniques should be employed in future work in order to sufficiently describe diverse dynamics of strand separation. So far, this paper has highlighted the asymmetrical and wide-ranging nature of base pair opening, even suggesting that fluctuations of amplitudes of the order of Angstroms occur on a two orders of magnitude shorter timescale than previously thought, the effect of these dynamics on the DPT had not yet been determined.

Between each pair of optimised GC and G^*C^* geometries for a given separation, a Nudged Elastic Band (NEB) calculation was performed to obtain the Minimum Energy Path (MEP) between the two states. For the equilibrium separation, an asynchronous DPT was determined with the PT across B2 being rate limiting with a larger forward barrier of $\sim 0.6\mathring{A}$ and $< 0.1\mathring{A}$ for B1. The reaction asymmetry was found to be $\sim 0.5\mathring{A}$. These DPT profiles at equilibrium distance were in good agreement with those reported by other authors[177, 33, 178, 111, 179, 115, 180, 176] and validated our computational methodology (see Section 2.7). Increasing the dimer separation showed a drastic effect

on the DPT energetics. Specifically, the forward barrier of the rate-limiting Proton Transfer across B2 was shown to increase quasi-linearly with separation. At an imposed separation of 1 Å increased the forward rate by around 300% up to ~ 2.4 eV, meanwhile the reaction asymmetry was reduced by $< 10\%$ to ~ 0.46 eV. At the same time a reverse barrier to the B1 transfer appears with larger separations < 0.1eV. These NEB profiles indicate that even the fractional Å separation events frequently observed in the dynamical simulations and a key part of the strand separation mechanism would cause a significant difference to the DPT. This increased forward reaction barrier for the B2 transfer, together with the emergence of a reverse reaction barrier to the B1 transfer, suggests that the non-equilibrium dynamics of strand separation stabilises the mutagenic G*C* conformation.

Following the discovery that G*C* is stabilised by strand separation, it was pertinent to reassess the possibility of the tautomerisation mechanism in AT. The DPT product A*T* was found to be short-lived by all but a few authors, and thus dismissed as a candidate for spontaneous mutations (see Section 2.7)[120, 121, 122, 100, 123, 104, 99, 75, 124, 125, 73, 126, 127, 128, 129, 115, 103]. As discussed for the case of GC, the focus on lifetime is misleading. If tautomerisation is a constant dynamical process, the tautomer population at any given instance is more meaningful than the lifetime of individual tautomeric states.

The emergence of additional stability of G*C* under strand separation prompted further investigation into a complementary work for the DPT in DPT the details of which can be found in Chapter 8 and published in [3]. In this work, the same methodology developed for GC was replicated for AT. Again a multiscale workflow was implemented where SMD simulations complemented constrained DFT. Similar to GC, non-equilibrium separations of AT revealed an asymmetric opening around one of the bonds, in this case B1, which refers to the nitrogen-oxygen hydrogen bond in AT. Again, the tautomeric dimer showed the same pivoted opening around B1 which breaks down at larger separations (> 1.6Å). At equilibrium separation, no stable DPT product (A*T*) was observed, however, at separations above 0.5Å a stable product emerged. A separation of 1 Å caused the rate-limiting transfer across B2 to increase by 400% to ~ 2eV. Additionally, a > 0.25 eV reverse barrier for B1 suggests that the A*T* tautomeric dimer is exceptionally stable during strand separation. Despite this new-found stability, the contribution of AT to spontaneous mutations is likely to be a secondary effect due to the inability of A*T* to form under equilibrium conditions. This suggests that for any trapping of tautomers, the tautomerisation must occur during strand separation, to which we have shown the barriers are prohibitively high. Complementary to the SMD performed for GC, the strand separation speed and opening angle distribution were determined for AT. It was found that a similar bimodal opening angle distribution was observed for AT, and the separation speed was again characteristic at ~ 1 Å/picosecond and uncorrelated to the biasing force.

With [1, 3] we significantly improved our understanding of the role of strand separation in spontaneous mutation via DPT, as we showed that trapping of tautomeric populations of G*C* and A*T* was possible even with lifetimes on the order of fractional picoseconds. Nevertheless, there remained

a need for a significant improvement in the modelling of the replisome environment in which the DPT was mapped. While ensemble QM/MM works had considered the DPT in aqueous duplex DNA, enzyme interactions in strand separation had not been modelled. To this end, the same PcrA-DNA complex introduced in Chapter 6 was taken as an environment for the DPT, in contrast to aqueous duplex DNA. These results are presented in Chapter 9. In this research, in which we presented the comparison between the GC DPT in aqueous dsDNA, to that of GC in the dsDNA of the complex. Namely, the last (N) and second to last (N-1) GC base pair of the duplex was considered. An Umbrella Sampling (US) reaction mapping scheme was employed within a QM/MM model of the system. Following validation of the asynchronous DPT in aqueous dsDNA, it was found that the base pair N-1 did not have an altered DPT profile, and the two scenarios were statistically indistinguishable. In contrast to base pair N-1, the last base pair of the duplex (N), had a radically different Potential of Mean Force (PMF) profile. The reaction asymmetry of the DPT was almost tripled from 0.60 eV to 1.68 eV, meanwhile, there was no observed reaction barrier to both the zwitterionic intermediate (G^-C^+) and the tautomeric DPT product (G^*C^*). This indicated that PcrA creates a unique chemical environment around base pair N, that completely discourages the formation of G^*C^* tautomer. It was found that even if G^*C^* is formed by this mechanism, its lifetime was expected to be negligible ($\sim$fs).

Observing the conformations populated during the US, a key amino acid residue was identified in asparagine (N624). The protonated α-amino group of the asparagine is in close proximity to the hydrogen bonds of GC, and encourages the presence of a water molecule forming a hydrogen bond bridge between the GC and the asparagine. In order to quantify the effect of the asparagine specifically, the US was repeated for a mutated helicase with an alanine residue in place of the asparagine (N624A). With this mutated PcrA complex, the reaction asymmetry was reduced to 1.18 eV which should increase the equilibrium population of G^*C^* tautomers by around eight orders of magnitude. This in turn would increase the number of ssDNA tautomers that pass through strand separation and provides an experimentally testable hypothesis for future work. The level at which asparagine N624 specifically is conserved in the wild-type PcrA helicase is also of note. N624 exists in the majority ($> 50\%$) of wild-type PcrA, but not in all the sequences. Minor populations of PcrA sequences include tyrosine, arginine, or glutamic acid in site 624 and these are still able to perform the stepping motor action for ssDNA translocation. We suggest that the preference for asparagine N624 goes beyond the primary function of strand separation, and may be due to a natural selection pressure to reduce spontaneous mutations.

Despite the lack of significant reverse reaction barriers and large asymmetry of the G^-C^+ and G^*C^* states in the US PMFs - which should indicate highly unstable states - in unbiased ensemble QM/MM MD simulations a system prepared in the tautomeric G^*C^* state does not decay back to the canonical well until several hundred simulation steps had passed. This discrepancy suggests that the ensemble-averaged PMF obtained via US does not capture the full picture of proton transfer.

We explain that US equilibrates all Degrees of Freedom (DoF's) that are orthogonal to the reaction coordinate, which is desirable when elucidating chemical equilibria. However, proton transfer happens on a timescale much faster than the equilibration of molecular DoF's and is ill-described by equilibrium theories, instead, we postulate that the protons experience an "instantaneous" Potential Energy Surface (iPES). To determine the iPES a scheme was designed where snapshots from QM/MM MD were taken as starting points and the transferring protons were scanned across their hydrogen bonds, keeping all other DoF's frozen. By taking a QM/MM Single-Point Energy (SPE) at each of these partially-transferred conformations the two-dimensional transfer surface could be interpolated. The MEP through this iPES was used as a fast-transfer approximation for the proton transfer. Remarkably these MEPs showed a similar reaction asymmetry to the US PMFs, but the Transition State (TS) energies between the minima were raised, indicating greater stability for the rare $G^{-}C^{+}$ and $G^{*}C^{*}$ states. Placing the $G^{*}C^{*}$ system back into ensemble QM/MM MD allowed us to track the decay of $G^{*}C^{*}$ back to canonical GC, via the $G^{-}C^{+}$ intermediate. Again, we observed a metastability where the $G^{*}C^{*}$ is maintained for ~ 0.1 picoseconds, despite the lack of an energetic minimum in the US PMF. We conclude that while the proton transfer occurs in a regime well-approximated by the iPES, the molecular DoF's begin to thermalise following a transfer, and adjust to the new protonation state. Following this equilibration, the TSs between the canonical, intermediate, and tautomeric states are once again lowered, hence reducing the stability of the zwitterion and tautomer. With this dual-timescale description of the proton transfer landscape, we have reconciled the metastability of $G^{*}C^{*}$ and $G^{-}C^{+}$, as the MEP through the iPES indicates a stability that decays with equilibration.

Finally, using the iPES technique we present an analogue to the NEB at non-equilibrium separation employed in Chapters 7 and 8. Here, further snapshots from QM/MM are taken as starting points for the iPES, with the addition of a weak steering force, opening up base pair N. The subsequently obtained MEPs indicated that the TSs energies increased linearly with separation, but the overall reaction asymmetry remained unchanged. This is consistent with our findings for aqueous GC and AT. While trapping of rare tautomeric and potentially mutagenic states is still possible within the PcrA micro-environment, the initial tautomeric populations are reduced to negligible quantities due to the large reaction asymmetry. On the whole, strand separation presents a challenging environment for the generation of ssDNA tautomers for spontaneous mutagenesis.

Furthermore, in the replication cycle, a polymerase enzyme synthesises dsDNA from the ssDNA templates that result from strand separation. In Chapter 10, which includes work published in [2], we investigate an alternative route to spontaneous mutations. Contrary to the mechanism proposed by Löwdin, where a DPT occurs during strand separation and ssDNA tautomers are paired into WC-like mismatches in the polymerase, this mechanism takes place solely in the DNA polymerase. Polymerases work by allowing for Nucleoside Triphosphate (NTP) molecules to diffuse into the palm domain and an attempt is made to pair them to the template strand. Normally, mismatches are

rejected at this stage if they do not form WC-like pairs. But it is possible for canonical Guanine and Thymine to rearrange into a WC-like G*Tvia a series of PTs. While other theoretical works preceded ours, we were the first to determine the role of quantum tunnelling in this process, and additionally, elucidate the role of the enzyme's dynamics in facilitating this mechanism. First, using DFT and NEB, the MEP between the wobble mismatched GT and G*T were determined. The NEB profile that emerged first showed a relatively linear potential energy cost to rearrange the molecules to facilitate the PT, before a high energy barrier is encountered ($E_f = 0.68$ eV), and another molecular rearrangement. The initial molecular rearrangement raises the potential energy by ~ 0.3eV over 4 Å of the reaction path, and on the other side of the barrier the potential energy drops by ~ 0.25eV over a further 2.5 Å. Coupled to a thermostat, the system will perform the initial translation via thermal Brownian motion. Should this occur, the reduced barrier width and height will amplify the rate of quantum tunnelling, and hence we have named this scenario a Tunnelling Ready State (TRS). A further NEB profile was produced linking the TRSs on either side of the transfer barrier. Through this barrier, the rate was determined to be 0.128/s with a tunnelling contribution of $\kappa = 99$ indicating the transfer occurred classically only 1% of the time. Additionally, the substitution of the hydrogen atoms with deuterium yielded a Kinetic Isotope Effect (KIE) of ~ 10 which is comparable to the experiment.

There was now a clear way in which the polymerase could interact with the GT wobble mismatch in order to promote misincorporation via Proton Transfer. The thumb domain of a polymerase compresses the potential base pairing to attempt to bind the two nucleobases. We postulated that this compression may prime the mismatch for misincorporation if the dimer is brought closer to the TRS. QM/MM MD simulations were performed on the DNA-polymerase complex, and a complementary system with the enzyme removed. In both cases, replicas were started in the wobble conformation and a metric was designed to measure the Root Mean Square (RMS) distance to the TRS determined from DFT(Δ). During initial classical MD simulations, a replica started in the wobble configuration would exhibit compression and relaxation around the initial Δ-value during ~ 1 ns before the thymine NTP was rejected and unbound. During this initial binding period, the random fluctuations cause compressions approaching the TRS. To gather statistics on the frequency of such compression towards the TRS, over a nanosecond of QM/MM MD replica simulations were performed for each of the DNA and DNA-polymerase complex systems. It was determined that in complex with the polymerase the wobble GT approached the TRS closer and more frequently than without the enzyme. We conclude that the polymerase's primary function, which compresses trial base pairs to attempt duplex synthesis, primes the wobble GT into a misincorporation via proton tunnelling.

Finally, Chapter 11 reflects on the biological meaning of the tautomerisation rates obtained in previous chapters, and how theoretical developments may shape the future of spontaneous mutation. In order to cause a mutation, tautomeric states need not only to be populated via a DPT,

but a mismatch must also be formed in the duplex DNA. The worst-case scenario is represented by a nonsense mutation (see Section 2.2), where a point mutation in DNA results in a change in genotype from an amino acid codon to a stop codon, prematurely terminating the protein during synthesis. In Chapter 11, we argued that even nonsense mutations in somatic cells are unlikely to have lasting effects, and that mutations in gametes and stem cells would pose a greater threat. A reaction network is used to estimate the fraction of G^*C^* that survives strand separation. By considering the process of meiosis as well as estimating the fraction of coding GC, the probability of single-stranded G^* and C^* causing point mutations was evaluated. Considering that point mutations are more likely to occur from GC than AT - as verified by experimental observations, and the computational work described in this thesis - we report that 65.5% of point mutations resulting from a DPT would cause a non-synonymous change to the genome, and a 7.3% chance of nonsense. Even high rates of DPT of 1500/s would suggest that approximately 1 in 40,000 gametes/children would have a premature stop codon in a part of their coding DNA (a near-worst-case scenario). For any non-synonymous mutation, this rate would be 1 in 3000. While this number may seem high, estimates of single-gene genetic disorders in humans lie around 1 in 50. Even such an optimistic DPT rate would suggest that in excess of two and a half million people on Earth have a genetic disorder as a result of a non-synonymous amino acid substitution in their genome, this is a secondary effect when considering the overall prevalence of genetic disorders in humans.

Having placed the role of tautomerism in context with genetic disorders, in Chapter 11 we point out the shortcomings of our theoretical descriptions of the frequency of tautomerism to date. Traditional descriptions of proton transfer within static energy landscapes break down during inherently out-of-equilibrium scenarios such as those encountered during strand separation, dsDNA synthesis, and even DNA in storage. Despite the difficulties in developing a theory to assess the evolution of a $GC \rightarrow G^*C^*$ reaction in a dynamic environment, the work presented in this thesis allows us to gain a qualitative picture of the feasibility of trapping DPT products during strand separation. Further theoretical work will allow for a more precise description of how a dynamically increasing transition state, due to the increased separation induced by a helicase, may trap an existing tautomer population. This may be achieved with Open Quantum Systems (OQS) models of proton transfer, where the time-evolution of the nuclear degrees of freedom can be numerically solved in a time-dependent potential function. Experimental validation of the efficacy of the N624A mutation in reducing error-rates in PcrA helicase would underline our hypothesis that evolutionary pressures exist to suppress quantum tunnelling in replication enzymes. Other routes to mutation via quantum effects may emerge in the methylation of DNA, gene expression, and/or epigenetics or the non-canonical motifs present in the telomeres of DNA. The ability of ssDNA tautomers to survive a potentially solvent-assisted decay back to their canonical counterparts and their nanopore sequencing traces beg for further investigation, as any theory of spontaneous mutations via a DPT hinges on this assumption.

Bibliography

[1] Louie Slocombe et al. “Proton transfer during DNA strand separation as a source of mutagenic guanine-cytosine tautomers”. In: *Communications Chemistry* 5.1 (Nov. 2022), p. 144. ISSN: 2399-3669. DOI: 10.1038/s42004-022-00760-x. URL: https://doi.org/10.1038/s42004-022-00760-x.

[2] Louie Slocombe et al. “Quantum Tunnelling Effects in the Guanine-Thymine Wobble Misincorporation via Tautomerism”. In: *The Journal of Physical Chemistry Letters* 14 (2022), pp. 9–15.

[3] Benjamin King et al. “Tautomerisation Mechanisms in the Adenine-Thymine Nucleobase Pair during DNA Strand Separation”. In: *The Journal of Physical Chemistry B* 0.0 (0). PMID: 36939840, null. DOI: 10.1021/acs.jpcb.2c08631.

[4] Max Winokan et al. “Multiscale simulations reveal the role of PcrA helicase in protecting against proton transfer in DNA”. In: (2023). Scientific Reports.

[5] Charles Darwin. *The Works of Charles Darwin, Volume 27: The Power of Movement in Plants*. NYU Press, 2010.

[6] Gregor Mendel. “Versuche über Pflanzen-Hybriden”. In: *Verhandlungen des Naturforschenden Vereines in Brünn* 4 (1866), pp. 3–47.

[7] Rosalind E Franklin and Raymond G Gosling. “Molecular configuration in sodium thymonucleate”. In: *Nature* 171 (1953), pp. 740–741.

[8] James D Watson and Francis HC Crick. “Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid”. In: *Nature* 171.4356 (1953), pp. 737–738.

[9] Giovanni Villani. “Theoretical investigation of the coupling between hydrogen atoms transfer and stacking interaction in guanine–cytosine dimers”. In: *Physical Chemistry Chemical Physics* 15.44 (2013), pp. 19242–19252.

[10] Bobo Feng et al. “Hydrophobic catalysis and a potential biological role of DNA unstacking induced by environment effects”. In: *Proceedings of the National Academy of Sciences* 116.35 (2019), pp. 17169–17174. ISSN: 0027-8424. DOI: 10.1073/pnas.1909122116. eprint: https://www.pnas.org/content/116/35/17169.full.pdf. URL: https://www.pnas.org/content/116/35/17169.

[11] Daniel Svozil et al. “DNA conformations and their sequence preferences”. In: *Nucleic Acids Research* 36.11 (2008), pp. 3690–3706.

[12] Stephen Neidle and Gary N Parkinson. “Quadruplex DNA crystal structures and drug design”. In: *Biochimie* 90.8 (2008), pp. 1184–1196.

[13] Nicholas V Hud. “Nucleic acid-metal ion interactions”. In: (2009).

[14] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4 et al. “Initial sequencing and analysis of the human genome”. In: *nature* 409.6822 (2001), pp. 860–921.

[15] Sameer S Velankar et al. “Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism”. In: *Cell* 97.1 (1999), pp. 75–84.

[16] Mark S Dillingham, Dale B Wigley, and Martin R Webb. “Demonstration of unidirectional single-stranded DNA translocation by PcrA helicase: measurement of step size and translocation speed”. In: *Biochemistry* 39.1 (2000), pp. 205–212.

[17] Jeehae Park et al. “PcrA helicase dismantles RecA filaments by reeling in DNA in uniform steps”. In: *Cell* 142.4 (2010), pp. 544–555.

[18] Jin Yu, Taekjip Ha, and Klaus Schulten. “Structure-based model of the stepping motor of PcrA helicase”. In: *Biophysical journal* 91.6 (2006), pp. 2097–2114.

[19] Christopher P Toseland et al. “The ATPase cycle of PcrA helicase and its coupling to translocation on DNA”. In: *Journal of molecular biology* 392.4 (2009), pp. 1020–1032.

[20] Mark S Dillingham et al. “Defining the roles of individual residues in the single-stranded DNA binding site of PcrA helicase”. In: *Proceedings of the National Academy of Sciences* 98.15 (2001), pp. 8381–8387.

[21] Lawrence A Loeb and Keith C Cheng. “Errors in DNA synthesis: a source of spontaneous mutations”. In: *Mutation Research/Reviews in Genetic Toxicology* 238.3 (1990), pp. 297–304.

[22] Leslie Pray. “Discovery of DNA structure and function: Watson and Crick”. In: *Nature Education* 1.1 (2008), p. 100.

[23] Scott D McCulloch and Thomas A Kunkel. “The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases”. In: *Cell research* 18.1 (2008), pp. 148–161.

[24] Dana Branzei and Marco Foiani. “Regulation of DNA repair throughout the cell cycle”. In: *Nature reviews Molecular cell biology* 9.4 (2008), pp. 297–308.

[25] Miguel Garcia-Diaz et al. “A structural solution for the DNA polymerase λ-dependent repair of DNA gaps with minimal homology”. In: *Molecular cell* 13.4 (2004), pp. 561–572.

[26] Miguel Garcia-Diaz et al. “A closed conformation for the Pol λ catalytic cycle”. In: *Nature structural & molecular biology* 12.1 (2005), pp. 97–98.

[27] Miguel Garcia-Diaz et al. “Role of the catalytic metal during polymerization by DNA polymerase lambda”. In: *DNA repair* 6.9 (2007), pp. 1333–1340.

[28] Andrea F Moon et al. “The X family portrait: structural insights into biological functions of X family polymerases”. In: *DNA repair* 6.12 (2007), pp. 1709–1725.

[29] Katarzyna Bebenek, Lars C Pedersen, and Thomas A Kunkel. “Replication infidelity via a mismatch with Watson–Crick geometry”. In: *Proceedings of the National Academy of Sciences* 108.5 (2011), pp. 1862–1867.

[30] Alya A Arabi and Chérif F Matta. “Effects of intense electric fields on the double proton transfer in the Watson–Crick Guanine–Cytosine base pair”. In: *The Journal of Physical Chemistry B* 122.37 (2018), pp. 8631–8641.

[31] Xifeng Li, Zhongli Cai, and Michael D Sevilla. “Investigation of proton transfer within DNA base pair anion and cation radicals by density functional theory (DFT)”. In: *The Journal of Physical Chemistry B* 105.41 (2001), pp. 10115–10123.

[32] Jason M Walsh, Penny J Beuning, et al. “Synthetic nucleotides as probes of DNA polymerase specificity”. In: *Journal of Nucleic Acids* 2012 (2012).

[33] Denis Jacquemin et al. “Assessing the importance of proton transfer reactions in DNA”. In: *Accounts of chemical research* 47.8 (2014), pp. 2467–2474.

[34] Greg C Randall and Patrick S Doyle. “DNA deformation in electric fields: DNA driven past a cylindrical obstruction”. In: *Macromolecules* 38.6 (2005), pp. 2410–2418.

[35] Francis HC Crick. “The origin of the genetic code”. In: *Journal of molecular biology* 38.3 (1968), pp. 367–379.

[36] Per-Olov Löwdin. “Proton tunneling in DNA and its biological implications”. In: *Reviews of Modern Physics* 35.3 (1963), p. 724.

[37] Sean J Johnson and Lorena S Beese. “Structures of mismatch replication errors observed in a DNA polymerase”. In: *Cell* 116.6 (2004), pp. 803–816.

[38] Isaac Joseph Kimsey. “Visualizing Rare Watson-Crick-Like Tautomeric and Anionic Mismatches in DNA and RNA”. PhD thesis. Duke University, 2016. URL: https://hdl.handle.net/10161/12885.

[39] AR Morgan. “Base mismatches and mutagenesis: how important is tautomerism?” In: *Trends in biochemical sciences* 18.5 (1993), pp. 160–163.

[40] Thomas A Kunkel and Dorothy A Erie. “DNA mismatch repair”. In: *Annu. Rev. Biochem.* 74 (2005), pp. 681–710.

[41] Paul Modrich. “Mechanisms in E. coli and human mismatch repair (Nobel Lecture)”. In: *Angewandte Chemie International Edition* 55.30 (2016), pp. 8490–8501.

[42] Mark J Schofield and Peggy Hsieh. “DNA mismatch repair: molecular mechanisms and biological function”. In: *Annual Reviews in Microbiology* 57.1 (2003), pp. 579–608.

[43] Iwona J Fijalkowska, Roel M Schaaper, and Piotr Jonczyk. “DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair”. In: *FEMS microbiology reviews* 36.6 (2012), pp. 1105–1121.

[44] Heewook Lee et al. “Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing”. In: *Proceedings of the National Academy of Sciences* 109.41 (2012), E2774–E2783.

[45] Jonathan A Eisen. “A phylogenomic study of the MutS family of proteins”. In: *Nucleic acids research* 26.18 (1998), pp. 4291–4300.

[46] Zhenguo Lin, Masatoshi Nei, and Hong Ma. “The origins and early evolution of DNA mismatch repair genes—multiple horizontal gene transfers and co-evolution”. In: *Nucleic acids research* 35.22 (2007), pp. 7591–7603.

[47] Josef Jiricny. “The multifaceted mismatch-repair system”. In: *Nature reviews Molecular cell biology* 7.5 (2006), pp. 335–346.

[48] Melissa L Fishel et al. “Apurinic/apyrimidinic endonuclease/redox factor-1 (APE1/Ref-1) redox function negatively regulates NRF2”. In: *Journal of Biological Chemistry* 290.5 (2015), pp. 3057–3068.

[49] Pengyu Hao et al. “Recurrent mismatch binding by MutS mobile clamps on DNA localizes repair complexes nearby”. In: *Proceedings of the National Academy of Sciences* 117.30 (2020), pp. 17775–17784.

[50] Ingrid Tessmer et al. “Mechanism of MutS searching for DNA mismatches and signaling repair”. In: *Journal of Biological Chemistry* 283.52 (2008), pp. 36646–36654.

[51] Josef Jiricny. “Mismatch repair: the praying hands of fidelity”. In: *Current Biology* 10.21 (2000), R788–R790.

[52] Manuela Haunschmidt, Wolfgang Buchberger, and Christian W Klampfl. “Investigations on the migration behaviour of purines and pyrimidines in capillary electromigration techniques with UV detection and mass spectrometric detection”. In: *Journal of Chromatography A* 1213.1 (2008), pp. 88–92.

[53] Christian FRELIN et al. “The regulation of the intracellular pH in cells from vertebrates”. In: *European journal of biochemistry* 174.1 (1988), pp. 3–14.

[54] Pallavi Thaplyal and Philip C Bevilacqua. “Experimental approaches for measuring pKa’s in RNA and DNA”. In: *Methods in enzymology*. Vol. 549. Elsevier, 2014, pp. 189–219.

[55] Per-Olov Löwdin. “Quantum genetics and the aperiodic solid: Some aspects on the biological problems of heredity, mutations, aging, and tumors in view of the quantum theory of the DNA molecule”. In: *Advances in quantum chemistry*. Vol. 2. Elsevier, 1966, pp. 213–360.

[56] Michael D Topal and Jacques R Fresco. “Complementary base pairing and the origin of substitution mutations”. In: *Nature* 263.5575 (1976), pp. 285–289.

[57] Tracey E Barrett et al. “Crystal structure of a G: T/U mismatch-specific DNA glycosylase: mismatch recognition by complementary-strand interactions”. In: *Cell* 92.1 (1998), pp. 117–129.

[58] Gordon A Leonard, Ewan D Booth, and Tom Brown. “Structural and thermodynamic studies on the adenine. guanine mismatch in B-DNA”. In: *Nucleic acids research* 18.19 (1990), pp. 5617–5623.

[59] Weina Wang, Homme W Hellinga, and Lorena S Beese. “Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis”. In: *Proceedings of the National Academy of Sciences* 108.43 (2011), pp. 17644–17648.

[60] Jie Pan and Sarah A Woodson. “Folding intermediates of a self-splicing RNA: mispairing of the catalytic core”. In: *Journal of molecular biology* 280.4 (1998), pp. 597–609.

[61] Apurba K Sau et al. “Evidence for A+ (anti)-G (syn) mismatched base-pairing in d-GGTAAGCGTACC”. In: *FEBS letters* 377.3 (1995), pp. 301–305.

[62] Isaac J Kimsey et al. “Visualizing transient Watson–Crick-like mispairs in DNA and RNA duplexes”. In: *Nature* 519.7543 (2015), pp. 315–320.

[63] John W Drake. “Spontaneous mutation: comparative rates of spontaneous mutation”. In: *Nature* 221 (1969), pp. 1132–1132.

[64] John W Drake et al. “Rates of spontaneous mutation”. In: *Genetics* 148.4 (1998), pp. 1667–1686.

[65] Michael W Nachman and Susan L Crowell. “Estimate of the mutation rate per nucleotide in humans”. In: *Genetics* 156.1 (2000), pp. 297–304.

[66] Augustine Kong et al. “Rate of de novo mutations and the importance of father’s age to disease risk”. In: *Nature* 488.7412 (2012), pp. 471–475.

[67] Alexander K Showalter and Ming-Daw Tsai. “A reexamination of the nucleotide incorporation fidelity of DNA polymerases”. In: *Biochemistry* 41.34 (2002), pp. 10571–10576.

[68] Penelope R Haddrill and Brian Charlesworth. “Non-neutral processes drive the nucleotide composition of non-coding sequences in Drosophila”. In: *Biology letters* 4.4 (2008), pp. 438–441.

[69] Guenter Albrecht-Buehler. “The spectra of point mutations in vertebrate genomes”. In: *Bioessays* 31.1 (2009), pp. 98–106.

[70] Stephan Ossowski et al. “The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana”. In: *science* 327.5961 (2010), pp. 92–94.

[71] Dawei Fu et al. “Effects of carboxymethyl chitosan on the blood system of rats”. In: *Biochemical and Biophysical Research Communications* 408.1 (2011), pp. 110–114.

[72] Jan Florian and Jerzy Leszczyński. “Spontaneous DNA Mutations Induced by Proton Transfer in the Guanine-Cytosine Base Pairs: An Energetic Perspective”. In: *Journal of the American Chemical Society* 118.12 (1996), pp. 3010–3017.

[73] Giovanni Villani. “Theoretical investigation of hydrogen atom transfer in the adenine–thymine base pair and its coupling with the electronic rearrangement. Concerted vs. stepwise mechanism”. In: *Physical Chemistry Chemical Physics* 12.11 (2010), pp. 2664–2669.

[74] Ol’ha O Brovarets’ and Dmytro M Hovorun. “Why the tautomerization of the G· C Watson–Crick base pair via the DPT does not cause point mutations during DNA replication? QM and QTAIM comprehensive analysis”. In: *Journal of Biomolecular Structure and Dynamics* 32.9 (2014), pp. 1474–1499.

[75] Ol’ha O Brovarets and Dmytro M Hovorun. “Can tautomerization of the A· T Watson–Crick base pair via double proton transfer provoke point mutations during DNA replication? A comprehensive QM and QTAIM analysis”. In: *Journal of Biomolecular Structure and Dynamics* 32.1 (2014), pp. 127–154.

[76] H Rüterjans et al. “Evidence for tautornerisrn in nucleic acid base pairs. 1 H NMR study of 15 N labeled tRNA”. In: *Nucleic acids research* 10.21 (1982), pp. 7027–7039.

[77] Isaac J Kimsey et al. “Dynamic basis for dG• dT misincorporation via tautomerization and ionization”. In: *Nature* 554.7691 (2018), pp. 195–201.

[78] Paul D Johnston and Alfred G Redfield. “An NMR study of the exchange rates for protons involved in the secondary and tertiary structure of yeast tRNA Phe”. In: *Nucleic acids research* 4.10 (1977), pp. 3599–3615.

[79] K Snoussi and J-L Leroy. “Imino proton exchange and base-pair kinetics in RNA duplexes”. In: *Biochemistry* 40.30 (2001), pp. 8898–8904.

[80] Sebastian Wärmländer, Anjana Sen, and Mikael Leijon. “Imino proton exchange in DNA catalyzed by ammonia and trimethylamine: evidence for a secondary long-lived open state of the base pair”. In: *Biochemistry* 39.3 (2000), pp. 607–615.

[81] Irene A Chen and Jack W Szostak. “A kinetic study of the growth of fatty acid vesicles”. In: *Biophysical journal* 87.2 (2004), pp. 988–998.

[82] Daniel Coman and Irina M Russu. “A nuclear magnetic resonance investigation of the energetics of basepair opening pathways in DNA”. In: *Biophysical journal* 89.5 (2005), pp. 3285–3292.

[83] Arthur Kornberg and Tania A. Baker. *DNA Replication*. W.H. Freeman and Company, 1998.

[84] Zhanhang Shen et al. "Binding of anticancer drug daunomycin to a TGGGGT G-quadruplex DNA probed by all-atom molecular dynamics simulations: Additional pure groove binding mode and implications on designing more selective G-quadruplex ligands". In: *Journal of Molecular Modeling* 23 (2017), pp. 1–11.

[85] Cees Dekker and Mark Ratner. "Electronic properties of DNA". In: *Physics World* 14.8 (2001), p. 29.

[86] Samir Kumar Pal and Ahmed H Zewail. "Dynamics of water in biological recognition". In: *Chemical Reviews* 104.4 (2004), pp. 2099–2124.

[87] Viet Hoang Man et al. "Comparative melting and healing of B-DNA and Z-DNA by an infrared laser pulse". In: *The Journal of Chemical Physics* 144.14 (2016), p. 145101.

[88] Péter Várnai and Krystyna Zakrzewska. "DNA and its counterions: a molecular dynamics study". In: *Nucleic acids research* 32.14 (2004), pp. 4269–4280.

[89] Alberto Pérez et al. "Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers". In: *Biophysical journal* 92.11 (2007), pp. 3817–3829.

[90] Rodrigo Galindo-Murillo et al. "Assessing the Current State of Amber Force Field Modifications for DNA". In: *Journal of Chemical Theory and Computation* 12.8 (2016). PMID: 27300587, pp. 4114–4127. DOI: 10.1021/acs.jctc.6b00186.

[91] Justin A. Lemkul and Alexander D. Jr. MacKerell. "Polarizable Force Field for DNA Based on the Classical Drude Oscillator: II. Microsecond Molecular Dynamics Simulations of Duplex DNA". In: *Journal of Chemical Theory and Computation* 13.5 (2017). PMID: 28398748, pp. 2072–2085. DOI: 10.1021/acs.jctc.7b00068.

[92] Alexey Savelyev and Alexander D. MacKerell Jr. "All-atom polarizable force field for DNA based on the classical drude oscillator model". In: *Journal of Computational Chemistry* 35.16 (2014), pp. 1219–1239. DOI: https://doi.org/10.1002/jcc.23611. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23611.

[93] Henryk Chojnacki et al. "Preliminary density functional calculations on the formic acid dimer". In: *Computers & chemistry* 19.3 (1995), pp. 181–187.

[94] Donald G Truhlar, Bruce C Garrett, and Stephen J Klippenstein. "Current status of transition-state theory". In: *The Journal of physical chemistry* 100.31 (1996), pp. 12771–12800.

[95] Giovanni Villani. "Theoretical investigation of hydrogen transfer mechanism in the guanine–cytosine base pair". In: *Chemical physics* 324.2-3 (2006), pp. 438–446.

[96] JP Cerón-Carrasco et al. "Intermolecular proton transfer in microhydrated guanine- cytosine base pairs: A new mechanism for spontaneous mutation in DNA". In: *The Journal of Physical Chemistry A* 113.39 (2009), pp. 10549–10556.

[97] S Tolosa, JA Sansón, and A Hidalgo. “Mechanisms for guanine–cytosine tautomeric equilibrium in solution via steered molecular dynamic simulations”. In: *Journal of Molecular Liquids* 251 (2018), pp. 308–316.

[98] Toru Matsui et al. “Sequence-dependent proton-transfer reaction in stacked GC pair II: The origin of stabilities of proton-transfer products”. In: *Chemical Physics Letters* 478.4-6 (2009), pp. 238–242.

[99] Robert L González-Romero et al. “Ultralow and anisotropic thermal conductivity in semiconductor As 2 Se 3”. In: *Physical Chemistry Chemical Physics* 20.3 (2018), pp. 1809–1816.

[100] Tomoyuki Hayashi and Shaul Mukamel. “Infrared signatures of proton transfer in guanine· cytosine and adenine· thymine base pairs: Dft study”. In: *Israel journal of chemistry* 44.1-3 (2004), pp. 185–191.

[101] Leonid Gorb et al. “A quantum-dynamics study of the prototropic tautomerism of guanine and its contribution to spontaneous point mutations in Escherichia coli”. In: *Biopolymers: Original Research on Biomolecules* 61.1 (2001), pp. 77–83.

[102] Yevgeniy Podolyan, Leonid Gorb, and Jerzy Leszczynski. *Ab initio study of the prototropic tautomerism of cytosine and guanine and their contribution to spontaneous point mutations*. 2003.

[103] Louie Slocombe, Jim S Al-Khalili, and Marco Sacchi. “Quantum and Classical effects in DNA point mutations”. In: *Physical Chemistry and Chemical Physics* (2021).

[104] Leonid Gorb et al. “Double-proton transfer in adenine- thymine and guanine- cytosine base pairs. A post-hartree- fock ab initio study”. In: *Journal of the American Chemical Society* 126.32 (2004), pp. 10119–10129.

[105] Vincent Zoete and Markus Meuwly. “Double proton transfer in the isolated and DNA-embedded guanine-cytosine base pair”. In: *The Journal of chemical physics* 121.9 (2004), pp. 4377–4388.

[106] Francesco Luigi Gervasio, Mauro Boero, and Michele Parrinello. “Double proton coupled charge transfer in DNA”. In: *Angewandte Chemie International Edition* 45.34 (2006), pp. 5606–5609.

[107] Diego Soler-Polo et al. “Proton Transfer in Guanine-Cytosine Base Pairs in B-DNA”. In: *Journal of chemical theory and computation* 15.12 (2019), pp. 6984–6991.

[108] Sven Roßbach and Christian Ochsenfeld. “Influence of coupling and embedding schemes on QM size convergence in QM/MM approaches for the example of a proton transfer in DNA”. In: *Journal of chemical theory and computation* 13.3 (2017), pp. 1102–1107.

[109] Alexander Gheorghiu. “Ensemble-based multiscale modelling of DNA base pair tautomerism in the absence and presence of external electric fields”. PhD thesis. University College London, 2020.

[110] Susanta Das, Kwangho Nam, and Dan Thomas Major. “Rapid convergence of energy and free energy profiles with quantum mechanical size in quantum mechanical–molecular mechanical simulations of proton transfer in DNA”. In: *Journal of chemical theory and computation* 14.3 (2018), pp. 1695–1705.

[111] Ol’ha O Brovarets and Dmytro M Hovorun. “Atomistic mechanisms of the double proton transfer in the H-bonded nucleobase pairs: QM/QTAIM computational lessons”. In: *Journal of Biomolecular Structure and Dynamics* 37.7 (2019), pp. 1880–1907.

[112] James Kermode et al. *Preconditioned Geometry Optimisers for the CASTEP and ONETEP codes*. Report. Archer, 2018. URL: https://www.archer.ac.uk/community/eCSE/eCSE11-07/eCSE11-07.php.

[113] Tushar van der Wijst et al. “Performance of various density functionals for the hydrogen bonds in DNA base pairs”. In: *Chemical Physics Letters* 426.4-6 (2006), pp. 415–421.

[114] Attila Bende. “Hydrogen bonding in the urea dimers and adenine–thymine DNA base pair: anharmonic effects in the intermolecular H-bond and intramolecular H-stretching vibrations”. In: *Theoretical Chemistry Accounts* 125 (2010), pp. 253–268.

[115] A Gheorghiu, PV Coveney, and AA Arabi. “The influence of base pair tautomerism on single point mutations in aqueous DNA”. In: *Interface focus* 10.6 (2020), p. 20190120.

[116] Alexander Gheorghiu, Peter V Coveney, and Alya A Arabi. “The influence of external electric fields on proton transfer tautomerism in the guanine–cytosine base pair”. In: *Physical Chemistry Chemical Physics* 23.10 (2021), pp. 6252–6265.

[117] Louie Slocombe, Marco Sacchi, and Jim Al-Khalili. “An Open Quantum Systems approach to proton tunnelling in DNA”. In: *Communications Physics* 5.1 (2022), pp. 1–9.

[118] Radek Pohl et al. “Proton transfer in guanine–cytosine base pair analogues studied by NMR spectroscopy and PIMD simulations”. In: *Faraday discussions* 212 (2018), pp. 331–344.

[119] Wei Fang et al. “Inverse temperature dependence of nuclear quantum effects in DNA base pairs”. In: *The journal of physical chemistry letters* 7.11 (2016), pp. 2125–2131.

[120] Vojtech Hrouda, Jan Florian, and Pavel Hobza. “Structure, energetics, and harmonic vibrational spectra of the adenine-thymine and adenine*-thymine* base pairs: gradient nonempirical and semiempirical study”. In: *The Journal of Physical Chemistry* 97.8 (1993), pp. 1542–1557.

[121] Jan Florian, Vojtech Hrouda, and Pavel Hobza. “Proton transfer in the adenine-thymine base pair”. In: *Journal of the American Chemical Society* 116.4 (1994), pp. 1457–1460.

[122] Alejandro Pérez et al. “Enol tautomers of Watson- Crick base pair models are metastable because of nuclear quantum effects”. In: *Journal of the American Chemical Society* 132.33 (2010), pp. 11510–11515.

[123] Santiago Tolosa, Jorge Antonio Sansón, and Antonio Hidalgo. “Theoretical thermodynamic study of the adenine–thymine tautomeric equilibrium: electronic structure calculations and steered molecular dynamic simulations”. In: *International Journal of Quantum Chemistry* 117.20 (2017), e25429.

[124] Adam D. Godbeer. “Quantum Tunnelling Effect in DNA Base Pair Mutation”. PhD thesis. 2014.

[125] AD Godbeer, JS Al-Khalili, and PD Stevenson. “Modelling proton tunnelling in the adenine–thymine base pair”. In: *Physical Chemistry Chemical Physics* 17.19 (2015), pp. 13034–13044.

[126] Giovanni Villani. “Theoretical investigation of hydrogen transfer mechanism in the adenine–thymine base pair”. In: *Chemical physics* 316.1-3 (2005), pp. 1–8.

[127] Ol’ha O Brovarets’, Kostiantyn S Tsiupa, and Dmytro M Hovorun. “Novel pathway for mutagenic tautomerization of classical A-T DNA base pairs via sequential proton transfer through quasi-orthogonal transition states: A QM/QTAIM investigation”. In: *PLoS One* 13.6 (2018), e0199044.

[128] Ol’ha O Brovarets’, Kostiantyn S Tsiupa, and Dmytro M Hovorun. “Unexpected A· T (WC)¡-¿A· T (rWC)/A· T (rH) and A· T (H)¡-¿ A· T (rH)/A· T (rWC) conformational transitions between the classical A· T DNA base pairs: A QM/QTAIM comprehensive study”. In: *International Journal of Quantum Chemistry* 118.18 (2018), e25674.

[129] Ol’ha O. Brovarets and Dmytro M. Hovorun. “Can tautomerization of the A·T Watson–Crick base pair via double proton transfer provoke point mutations during DNA replication? A comprehensive QM and QTAIM analysis”. In: *Journal of Biomolecular Structure and Dynamics* 32.1 (2014). PMID: 23383960, pp. 127–154. DOI: 10.1080/07391102.2012.755795.

[130] Peter Atkins and Julio De Paula. *Physical chemistry for the life sciences*. Oxford University Press, USA, 2011.

[131] Peter W Atkins and Julio De Paula. *Physical Chemistry*. 8th ed. Oxford university press, Oxford UK, 2006.

[132] Frank Jensen. *Introduction to computational chemistry*. John wiley & sons, 2017.

[133] Josiah Willard Gibbs. “A method of geometrical representation of the thermodynamic properties by means of surfaces”. In: *Transactions of Connecticut Academy of Arts and Sciences* (1873), pp. 382–404.

[134] Lubert Stryer. *Biochemistry*. 3rd ed. Freeman and Company, New York, 1988.

[135] Raymond Chang. *Physical chemistry for the biosciences*. University Science Books, 2005.

[136] Eldon Emberly. *Introduction to Biological Physics: Boltzmann Distribution*. Lecture Notes, Simon Fraser University.

[137] Richard Lonsdale, Jeremy N Harvey, and Adrian J Mulholland. "A practical guide to modelling enzyme–catalysed reactions". In: *Chemical Society Reviews* 41.8 (2012), pp. 3025–3038.

[138] Donald G Truhlar et al. "Ensemble-averaged variational transition state theory with optimized multidimensional tunneling for enzyme kinetics and other condensed-phase reactions". In: *International journal of quantum chemistry* 100.6 (2004), pp. 1136–1152.

[139] Alain C Vaucher and Markus Reiher. "Minimum energy paths and transition states by curve optimization". In: *Journal of chemical theory and computation* 14.6 (2018), pp. 3091–3099.

[140] Yanying Liu et al. "Understanding the Large Kinetic Isotope Effect of Hydrogen Tunneling in Condensed Phases by Using Double-Well Model Systems". In: *The Journal of Physical Chemistry B* 125.22 (2021). PMID: 34033714, pp. 5959–5970. DOI: 10.1021/acs.jpcb.1c02851. eprint: https://doi.org/10.1021/acs.jpcb.1c02851. URL: https://doi.org/10.1021/acs.jpcb.1c02851.

[141] Eugene Wigner. "On the Quantum Correction For Thermodynamic Equilibrium". In: *Physical Review* 40 (1932), p. 749.

[142] R P_ Bell. "The tunnel effect correction for parabolic potential barriers". In: *Transactions of the Faraday Society* 55 (1959), pp. 1–4.

[143] Louie Slocombe. "Tunnelling Corrections". 2023.

[144] M. Razavy. *Quantum Theory Of Tunneling (2nd Edition)*. World Scientific Publishing Company, 2013. ISBN: 9789814525039. URL: https://books.google.co.uk/books?id=%5C_j67CgAAQBAJ.

[145] Jiri Sponer and Filip Lankas. *Computational Studies of RNA and DNA*. Vol. 2. Springer Science & Business Media, 2006.

[146] Pierre Hohenberg and Walter Kohn. "Inhomogeneous electron gas". In: *Physical review* 136.3B (1964), B864.

[147] Walter Kohn and Lu Jeu Sham. "Self-consistent equations including exchange and correlation effects". In: *Physical review* 140.4A (1965), A1133.

[148] Gino A DiLabio, Mohammad Koleini, and Edmanuel Torres. "Extension of the B3LYP–dispersion-correcting potential approach to the accurate treatment of both inter-and intra-molecular interactions". In: *Theoretical Chemistry Accounts* 132.10 (2013), pp. 1–13.

[149] George A Jeffrey and Wolfram Saenger. *Hydrogen bonding in biological structures*. Springer Science & Business Media, 2012.

[150] Stefan Grimme et al. “A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu”. In: *The Journal of chemical physics* 132.15 (2010), p. 154104.

[151] Stefan Grimme. “Density functional theory with London dispersion corrections”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (2011), pp. 211–228.

[152] Stefan Grimme et al. “Consistent structures and interactions by density functional theory with small atomic orbital basis sets”. In: *The Journal of chemical physics* 143.5 (2015), p. 054107.

[153] Jacques P Bothma, Joel B Gilmore, and Ross H McKenzie. “The role of quantum effects in proton transfer reactions in enzymes: quantum tunneling in a noisy environment?” In: *New Journal of Physics* 12.5 (2010), p. 055002.

[154] Hannes Jónsson, Greg Mills, and Karsten W Jacobsen. “Nudged elastic band method for finding minimum energy paths of transitions”. In: (1998).

[155] Delaram Ghoreishi et al. “Fast implementation of the nudged elastic band method in AMBER”. In: *Journal of chemical theory and computation* 15.8 (2019), pp. 4699–4707.

[156] David H Mathews and David A Case. “Nudged elastic band calculation of minimal energy paths for the conformational change of a GG non-canonical pair”. In: *Journal of molecular biology* 357.5 (2006), pp. 1683–1693.

[157] Martin Hangaard Hansen, José A Garrido Torres, Paul C Jennings, et al. “An Atomistic Machine Learning Package for Surface Science and Catalysis”. In: *arXiv preprint arXiv:1904.00904* (2019).

[158] José A Garrido Torres, Paul C Jennings, Martin H Hansen, et al. “Low-scaling algorithm for nudged elastic band calculations using a surrogate machine learning model”. In: *Phys. Rev. Lett.* 122.15 (2019), p. 156001.

[159] Glenn M Torrie and John P Valleau. “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling”. In: *Journal of Computational Physics* 23.2 (1977), pp. 187–199.

[160] Johannes Kästner. “Umbrella sampling”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.6 (2011), pp. 932–942.

[161] Shankar Kumar et al. “The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method”. In: *Journal of computational chemistry* 13.8 (1992), pp. 1011–1021.

[162] Marc Souaille and Benoıt Roux. “Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations”. In: *Computer physics communications* 135.1 (2001), pp. 40–57.

[163] Yuko Tsuchiya et al. “Keto-enol tautomer of uracil and thymine”. In: *The Journal of Physical Chemistry* 92.7 (1988), pp. 1760–1765.

[164] Raymond J Abraham and Paul E Smith. “Charge calculations in molecular mechanics 6: the calculation of partial atomic charges in nucleic acid bases and the electrostatic contribution to DNA base pairing”. In: *Nucleic acids research* 16.6 (1988), pp. 2639–2657.

[165] C Fonseca Guerra et al. “Adenine tautomers: relative stabilities, ionization energies, and mismatch with cytosine”. In: *The Journal of Physical Chemistry A* 110.11 (2006), pp. 4012–4020.

[166] Longjiu Cheng and Jinlong Yang. “Modified Morse potential for unification of the pair interactions”. In: *The Journal of chemical physics* 127.12 (2007), p. 124104.

[167] Célia Fonseca Guerra et al. “The nature of the hydrogen bond in DNA base pairs: the role of charge transfer and resonance assistance”. In: *Chemistry–A European Journal* 5.12 (1999), pp. 3581–3594.

[168] IK Yanson, AB Teplitsky, and LF Sukhodub. “Experimental studies of molecular interactions between nitrogen bases of nucleic acids”. In: *Biopolymers: Original Research on Biomolecules* 18.5 (1979), pp. 1149–1170.

[169] Zhengrong Wu et al. “H...N hydrogen bond lengths in double stranded DNA from internucleotide dipolar couplings”. In: *Journal of biomolecular NMR* 19.4 (2001), pp. 361–365.

[170] Célia Fonseca Guerra et al. “Hydrogen bonding in DNA base pairs: reconciliation of theory and experiment”. In: *Journal of the American Chemical Society* 122.17 (2000), pp. 4117–4128.

[171] Halina Szatylowicz and Nina Sadlej-Sosnowska. “Characterizing the strength of individual hydrogen bonds in DNA base pairs”. In: *Journal of chemical information and modeling* 50.12 (2010), pp. 2151–2161.

[172] Martin J Scanlan and Ian H Hillier. “An ab initio study of tautomerism of uracil, thymine, 5-fluorouracil, and cytosine”. In: *Journal of the American Chemical Society* 106.13 (1984), pp. 3737–3745.

[173] Katherine Cox et al. “Molecular dynamics simulations of a helicase”. In: *Proteins: Structure, Function, and Bioinformatics* 52.2 (2003), pp. 254–262.

[174] Wenke Zhang et al. “Directional loading and stimulation of PcrA helicase by the replication initiator protein RepD”. In: *Journal of molecular biology* 371.2 (2007), pp. 336–348.

[175] J. D. Watson and F. H. C. Crick. “THE STRUCTURE OF DNA”. In: *Cold Spring Harbor Symp. Quant. Biol.* 18 (1953), pp. 123–131. DOI: 10.1101/SQB.1953.018.01.020. eprint: http://symposium.cshlp.org/content/18/123.full.pdf+html. URL: http://symposium.cshlp.org/content/18/123.short.

[176] Youngchan Kim et al. “Quantum Biology: An Update and Perspective”. In: *Quantum Reports* 3.1 (2021), pp. 80–126. ISSN: 2624-960X. DOI: 10.3390/quantum3010006. URL: https://www.mdpi.com/2624-960X/3/1/6.

[177] Jan Florián and Jerzy Leszczyński. “Spontaneous DNA Mutations Induced by Proton Transfer in the Guanine·Cytosine Base Pairs: An Energetic Perspective”. In: *J. Am. Chem. Soc.* 118.12 (1996), pp. 3010–3017. DOI: 10.1021/ja951983g. eprint: https://doi.org/10.1021/ja951983g. URL: https://doi.org/10.1021/ja951983g.

[178] Diego Soler-Polo et al. “Proton transfer in guanine-cytosine base pairs in B-DNA”. In: *Journal of chemical theory and computation* 15.12 (2019), pp. 6984–6991.

[179] Ruby Srivastava. “The Role of Proton Transfer on Mutations”. In: *Frontiers in Chemistry* 7 (2019). ISSN: 2296-2646. DOI: 10.3389/fchem.2019.00536. URL: https://www.frontiersin.org/article/10.3389/fchem.2019.00536.

[180] Louie Slocombe, JS Al-Khalili, and Marco Sacchi. “Quantum and classical effects in DNA point mutations: Watson–Crick tautomerism in AT and GC base pairs”. In: *Physical Chemistry Chemical Physics* 23.7 (2021), pp. 4141–4150.

[181] Axel D. Becke. “Density-functional thermochemistry. III. The role of exact exchange”. In: *The Journal of Chemical Physics* 98.7 (1993), pp. 5648–5652. DOI: 10.1063/1.464913.

[182] Edoardo Apra et al. “NWChem: Past, present, and future”. In: *The Journal of chemical physics* 152.18 (2020), p. 184102.

[183] Timothy M Lohman and Keith P Bjornson. “Mechanisms of helicase-catalyzed DNA unwinding”. In: *Annual review of biochemistry* 65.1 (1996), pp. 169–214.

[184] Ol’ha O. Brovarets’ and Dmytro M. Hovorun. “Why the tautomerization of the G·C Watson–Crick base pair via the DPT does not cause point mutations during DNA replication? QM and QTAIM comprehensive analysis”. In: *Journal of Biomolecular Structure and Dynamics* 32.9 (2014). PMID: 23909623, pp. 1474–1499. DOI: 10.1080/07391102.2013.822829.

[185] Anil Kumar and Michael D Sevilla. “Proton-coupled electron transfer in DNA on formation of radiation-produced ion radicals”. In: *Chemical reviews* 110.12 (2010), pp. 7002–7023.

[186] David R Weinberg et al. “Proton-coupled electron transfer”. In: *Chemical Reviews* 112.7 (2012), pp. 4016–4093.

[187] Robin Tyburski et al. “Proton-coupled electron transfer guidelines, fair and square”. In: *Journal of the American Chemical Society* 143.2 (2021), pp. 560–576.

[188] Kimberly de La Harpe, Carlos E Crespo-Hernández, and Bern Kohler. “Deuterium isotope effect on excited-state dynamics in an alternating GC oligonucleotide”. In: *Journal of the American Chemical Society* 131.48 (2009), pp. 17557–17559.

[189] Erin R Johnson and Axel D Becke. “Van der Waals interactions from the exchange hole dipole moment: application to bio-organic benchmark systems”. In: *Chemical Physics Letters* 432.4-6 (2006), pp. 600–603.

[190] Axel D Becke, Alya A Arabi, and Felix O Kannemann. “Nonempirical density-functional theory for van der Waals interactions”. In: *Canadian Journal of Chemistry* 88.11 (2010), pp. 1057–1062.

[191] A Otero-De-La-Roza and Erin R Johnson. “Non-covalent interactions and thermochemistry using XDM-corrected hybrid and range-separated hybrid density functionals”. In: *The Journal of chemical physics* 138.20 (2013), p. 204109.

[192] A. Klamt and G. Schüürmann. “COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient”. In: *J. Chem. Soc., Perkin Trans. 2* (5 1993), pp. 799–805. DOI: 10.1039/P29930000799. URL: http://dx.doi.org/10.1039/P29930000799.

[193] Darrin M. York and Martin Karplus. “A Smooth Solvation Potential Based on the Conductor-Like Screening Model”. In: *The Journal of Physical Chemistry A* 103.50 (1999), pp. 11060–11079. DOI: 10.1021/jp9920971.

[194] Aleksandr V. Marenich, Christopher J. Cramer, and Donald G. Truhlar. “Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions”. In: *The Journal of Physical Chemistry B* 113.18 (2009). PMID: 19366259, pp. 6378–6396. DOI: 10.1021/jp810292n.

[195] Jed W Pitera, Michael Falta, and Wilfred F van Gunsteren. “Dielectric properties of proteins from simulation: the effects of solvent, ligands, pH, and temperature”. In: *Biophysical journal* 80.6 (2001), pp. 2546–2555.

[196] Lin Li et al. “On the dielectric “constant” of proteins: smooth dielectric function for macromolecular modeling and its implementation in DelPhi”. In: *Journal of chemical theory and computation* 9.4 (2013), pp. 2126–2136.

[197] M. C. Payne et al. “Iterative minimization techniques for ab initio total-energy calculations - molecular-dynamics and conjugate gradients”. In: *Rev. Mod. Phys.* 64 (1992), pp. 1045–1097.

[198] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, et al. “The atomic simulation environment—a Python library for working with atoms”. In: *J. Phys.: Condens. Matter* 29.27 (2017), p. 273002. URL: http://stacks.iop.org/0953-8984/29/i=27/a=273002.

[199] S. R. Bahn and K. W. Jacobsen. “An object-oriented scripting interface to a legacy electronic structure code”. In: *Comput. Sci. Eng.* 4.3 (2002), pp. 56–66. DOI: 10.1109/5992.998641.

[200] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. “GROMACS: A message-passing parallel molecular dynamics implementation”. In: *Computer Physics Communications* 91.1 (1995), pp. 43–56. ISSN: 0010-4655. DOI: https://doi.org/10.1016/0010-4655(95)00042-E. URL: https://www.sciencedirect.com/science/article/pii/001046559500042E.

[201] Katarina Hart et al. “Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium”. In: *Journal of chemical theory and computation* 8.1 (2012), pp. 348–362.

[202] J. Watson and F. Crick. “Genetical Implications of the Structure of Deoxyribonucleic Acid”. In: *Nature* 171 (1953), pp. 964–967. DOI: https://doi.org/10.1038/171964b0.

[203] Natalia S Nemeria et al. “Reaction mechanisms of thiamin diphosphate enzymes: defining states of ionization and tautomerization of the cofactor at individual steps”. In: *The FEBS journal* 276.9 (2009), pp. 2432–3446.

[204] Alexey Rozov et al. “Novel base-pairing interactions at the tRNA wobble position crucial for accurate reading of the genetic code”. In: *Nature communications* 7.1 (2016), pp. 1–10.

[205] Eugene S Kryachko and John R Sabin. “Quantum chemical study of the hydrogen-bonded patterns in A· T base pair of DNA: Origins of tautomeric mispairs, base flipping, and Watson–Crick to Hoogsteen conversion”. In: *International journal of quantum chemistry* 91.6 (2003), pp. 695–710.

[206] Axel D Becke and Erin R Johnson. “A density-functional model of the dispersion interaction”. In: *The Journal of chemical physics* 123.15 (2005), p. 154101.

[207] Erin R Johnson and Axel D Becke. “A post-Hartree–Fock model of intermolecular interactions”. In: *The Journal of chemical physics* 123.2 (2005), p. 024101.

[208] Henk Bekker et al. “Gromacs-a parallel computer for molecular-dynamics simulations”. In: *4th International Conference on Computational Physics (PC 92)*. World Scientific Publishing. 1993, pp. 252–256.

[209] HJC Berendsen, JR Grigera, and TP Straatsma. “The missing term in effective pair potentials”. In: *Journal of Physical Chemistry* 91.24 (1987), pp. 6269–6271.

[210] Pengfei Li et al. “Environmental Effects on Guanine-Thymine Mispair Tautomerization Explored with Quantum Mechanical/Molecular Mechanical Free Energy Simulations”. In: *Journal of the American Chemical Society* 142.25 (2020). PMID: 32459476, pp. 11183–11191. DOI: 10.1021/jacs.0c03774.

[211] James D Watson and Francis HC Crick. “The structure of DNA”. In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 18. Cold Spring Harbor Laboratory Press. 1953, pp. 123–131.

[212] Alexey Rozov, Natalia Demeshkina, Eric Westhof, et al. “New structural insights into translational miscoding”. In: *Trends Biochem. Sci.* 41.9 (2016), pp. 798–814.

[213] L. Slocombe, J. S. Al-Khalili, and M. Sacchi. “Quantum and classical effects in DNA point mutations: Watson–Crick tautomerism in AT and GC base pairs”. In: *Phys. Chem. Chem. Phys.* 23 (7 2021), pp. 4141–4150. DOI: 10.1039/D0CP05781A. URL: http://dx.doi.org/10.1039/D0CP05781A.

[214] Ol’ha O Brovarets’ and Dmytro M Hovorun. “Proton tunneling in the A-T Watson-Crick DNA base pair: myth or reality?” In: *Journal of Biomolecular Structure and Dynamics* 33.12 (2015), pp. 2716–2720.

[215] Robert Rein and Frank E Harris. “Studies of Hydrogen-Bonded Systems. II. Tunneling and Tautomeric Equilibria in the N–H··· N Hydrogen Bond of the Guanine—Cytosine Base Pair”. In: *The Journal of Chemical Physics* 42.6 (1965), pp. 2177–2180.

[216] Angiolari, Federica and Huppert, Simon and Pietrucci, Fabio and Spezia, Riccardo. “Environmental and Nuclear Quantum Effects on Double Proton Transfer in the Guanine–Cytosine Base Pair”. In: *The Journal of Physical Chemistry Letters* 14 (2023), 5102–5108.

[217] M Noguera, M Sodupe, and J Bertrán. “Effects of protonation on proton-transfer processes in guanine–cytosine Watson–Crick base pairs”. In: *Theoretical Chemistry Accounts* 112 (2004), pp. 318–326.

[218] Mark S Dillingham, Panos Soultanas, and Dale B Wigley. “Site-directed mutagenesis of motif III in PcrA helicase reveals a role in coupling ATP hydrolysis to strand separation”. In: *Nucleic acids research* 27.16 (1999), pp. 3310–3317.

[219] Gregor Wentzel. “Eine verallgemeinerung der quantenbedingungen für die zwecke der wellenmechanik”. In: *Zeitschrift für Physik* 38.6-7 (1926), pp. 518–529.

[220] Umate, Pavan and Tuteja, Narendra and Tuteja, Renu. “Genome-wide comprehensive analysis of human helicases”. In: *Communicative & integrative biology* 4.1 (2011), 118–137.

[221] Spratt, Austin N and Gallazzi, Fabio and Quinn, Thomas P and Lorson, Christian L and Sönnerborg, Anders and Singh, Kamal. “Coronavirus helicases: Attractive and unique targets of antiviral drug-development and therapeutic patents”. In: *Expert opinion on therapeutic patents* 31.4 (2021), 339–350.

[222] Xu, Ting and Sampath, Aruna and Chao, Alex and Wen, Daying and Nanao, Max and Chene, Patrick and Vasudevan, Subhash G and Lescar, Julien. “Structure of the Dengue virus helicase/nucleoside triphosphatase catalytic domain at a resolution of 2.4 A”. In: *Journal of virology* 79.16 (2005), 10278–10288.

[223] Tian, Hongliang and Ji, Xiaoyun and Yang, Xiaoyun and Zhang, Zhongxin and Lu, Zuokun and Yang, Kailin and Chen, Cheng and Zhao, Qi and Chi, Heng and Mu, Zhongyu and others. "Structural basis of Zika virus helicase in recognizing its substrates". In: *Protein & cell* 7.8 (2016), 562–570.

[224] Marcus D Hanwell et al. "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform". In: *Journal of cheminformatics* 4.1 (2012), pp. 1–17.

[225] Max Winokan and Cedric Vallee. *MolParse: A python package for parsing, modifying, and analysis of molecular structure files.* `https://github.com/mwinokan/MolParse`. 2023.

[226] Robert B Best et al. "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles". In: *Journal of chemical theory and computation* 8.9 (2012), pp. 3257–3273.

[227] Jing Huang et al. "CHARMM36m: an improved force field for folded and intrinsically disordered proteins". In: *Nature methods* 14.1 (2017), pp. 71–73.

[228] Jürg Hutter et al. "cp2k: atomistic simulations of condensed matter systems". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.1 (2014), pp. 15–25.

[229] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature methods* 17.3 (2020), pp. 261–272.

[230] Alexey Rozov et al. "Structural insights into the translational infidelity mechanism". In: *Nature communications* 6.1 (2015), pp. 1–9.

[231] Ol'ha O. Brovarets and Dmytro M. Hovorun. "Quantum dancing of the wobble G-T(U/5BrU) nucleobase pairs and its biological roles". In: *Chemical Physics Impact* 1 (2020), p. 100006. ISSN: 2667-0224. DOI: `https://doi.org/10.1016/j.chphi.2020.100006`. URL: `https://www.sciencedirect.com/science/article/pii/S2667022420300062`.

[232] Kei Odai and Keisho Umesaki. "Kinetic Study of Transition Mutations from G–C to A–T Base Pairs in Watson–Crick DNA Base Pairs: Double Proton Transfers". In: *The Journal of Physical Chemistry A* 125.37 (2021), pp. 8196–8204.

[233] Shreya Chandorkar et al. "Multiscale Modeling of Wobble to Watson–Crick-Like Guanine–Uracil Tautomerization Pathways in RNA". In: *International journal of molecular sciences* 22.11 (2021), p. 5411.

[234] Myong-Chul Koag, Kwangho Nam, and Seongmin Lee. "The spontaneous replication error and the mismatch discrimination mechanisms of human DNA polymerase β". In: *Nucleic acids research* 42.17 (2014), pp. 11233–11245.

[235] Natalia Demeshkina et al. "A new understanding of the decoding principle on the ribosome". In: *Nature* 484.7393 (2012), pp. 256–259.

[236] Alexey Rozov et al. “Tautomeric GU pairs within the molecular ribosomal grip and fidelity of decoding in bacteria”. In: *Nucleic Acids Research* 46.14 (2018), pp. 7425–7435.

[237] Gizem Çelebi et al. “Time delay during the proton tunneling in the base pairs of the DNA double helix”. In: *Progress in Biophysics and Molecular Biology* (2021).

[238] Atul Rangadurai et al. “Probing conformational transitions towards mutagenic Watson–Crick-like G· T mismatches using off-resonance sugar carbon R 1ρ relaxation dispersion”. In: *Journal of Biomolecular NMR* (2020), pp. 1–15.

[239] Ol’ha O. Brovarets’ and Dmytro M. Hovorun. “The nature of the transition mismatches with Watson–Crick architecture: the G*·T or G·T* DNA base mispair or both? A QM/QTAIM perspective for the biological problem”. In: *Journal of Biomolecular Structure and Dynamics* 33.5 (2015). PMID: 24842163, pp. 925–945. DOI: 10.1080/07391102.2014.924879.

[240] Amir O Caldeira and Anthony J Leggett. “Path integral approach to quantum Brownian motion”. In: *Physica A: Statistical mechanics and its Applications* 121.3 (1983), pp. 587–616.

[241] Richard Phillips Feynman and FL Vernon Jr. “The theory of a general quantum system interacting with a linear dissipative system”. In: *Annals of physics* 281.1-2 (2000), pp. 547–607.

[242] J. Weinbub and D. K. Ferry. “Recent advances in Wigner function approaches”. In: *Applied Physics Reviews* 5.4 (2018), p. 041104. DOI: 10.1063/1.5046663. eprint: https://doi.org/10.1063/1.5046663. URL: https://doi.org/10.1063/1.5046663.

[243] Zixian Zhou et al. “Quantum Zeno and anti-Zeno effects in open quantum systems”. In: *Physical Review A* 96.3 (2017). ISSN: 2469-9926 2469-9934.

[244] Fabian Gottwald, Sergei D. Ivanov, and Oliver Kühn. “Applicability of the Caldeira–Leggett Model to Vibrational Spectroscopy in Solution”. In: *The Journal of Physical Chemistry Letters* 6.14 (2015). PMID: 26266853, pp. 2722–2727. DOI: 10.1021/acs.jpclett.5b00718.

[245] Steven A Frank. *Dynamics of cancer: incidence, inheritance, and evolution*. princeton university press, 2007.

[246] Centers for Disease Control, Prevention (CDC, et al. “Update on overall prevalence of major birth defects–Atlanta, Georgia, 1978-2005”. In: *MMWR. Morbidity and mortality weekly report* 57.1 (2008), pp. 1–5.

[247] Danielle M Ely and Anne K Driscoll. “Infant mortality in the United States, 2020: data from the period linked birth/infant death file.” In: *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* 71.5 (2022).

[248] Kim M Keeling, Ming Du, and David M Bedwell. “Therapies of nonsense-associated diseases”. In: *Nonsense-mediated mRNA Decay—Landes Bioscience Editor: Lynne E. Maquat* (2006), pp. 121–136.

[249] Michael Krawczak et al. “Human gene mutation database—a biomedical information and research resource”. In: *Human mutation* 15.1 (2000), pp. 45–51.

[250] Pamela A Frischmeyer and Harry C Dietz. “Nonsense-mediated mRNA decay in health and disease”. In: *Human molecular genetics* 8.10 (1999), pp. 1893–1900.

[251] Pankaj Kumar et al. “Prevalence and patterns of presentation of genetic disorders in a pediatric emergency department”. In: *Mayo Clinic Proceedings*. Vol. 76. 8. Elsevier. 2001, pp. 777–783.

[252] Frank MJ Jacobs et al. “An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons”. In: *Nature* 516.7530 (2014), pp. 242–245.

[253] Luis P Villarreal et al. *Viruses and the evolution of life*. ASM press, 2005.

[254] AO Wilkie. “The molecular basis of genetic dominance.” In: *Journal of medical genetics* 31.2 (1994), pp. 89–98.

Appendix

A. Naming Conventions

Atom Names of Nucleic Acids

For molecular dynamics all the atoms in a residue need a unique name. The ASCII sketches below show the naming conventions used in this work for the canonical (backbone-less) nucleotides.

Listing 1 *Adenine*

```
H61  H62
  \  /
   N6
   |
   C6
  // \
 N1  C5--N7\\
 |   ||     C8-H8
 C2  C4--N9/
 /\\ /    \
H2 N3      H9
```

Listing 2 *Thymine*

```
     H51    O4...
      |     ||
H52-C5M    C4   H3...
     |  \  /  \  /
  H53   C5    N3
        ||    |
      H6-C6   C2
          \   / \\
           N1    O2
            \
             H1
```

Listing 3 *Guanine*

```
        O6
        ||
        C6
       /  \
   H1-N1   C5--N7\\
      |    ||     C8-H8
      C2   C4--N9/
     / \\ /     \
H21-N2  N3       H9
    |
    H22
```

Listing 4 *Cytosine*

```
   H42  H41
     \  /
      N4
      |
      C4
     / \\
 H5-C5  N3
    ||   |
 H6-C6   C2
     \  / \\
      N1   O2
      |
      H1
```

Table 12-1 Table of Nucleic Acid (NA) and Amino Acid (AA) residue shorthands sorted by type. SC denotes Side Chain.

Molecule	Shorthand	Residue Key	Type
Adenine	A	DA	Purine NA
Thymine	T	DT	Pyrimidine NA
Guanine	G	DG	Purine NA
Cytosine	C	DC	Pyrimidine NA
Arginine	R	ARG	Positively Charged SC AA
Histidine	H	HIS/HID/HIE	Positively Charged SC AA
Lysine	K	LYS	Positively Charged SC AA
Aspartic Acid	D	ASP	Negatively Charged SC AA
Glutamic Acid	E	GLU	Negatively Charged SC AA
Serine	S	SER	Polar Uncharged SC AA
Threonine	T	THR	Polar Uncharged SC AA
Asparagine	N	ASN	Polar Uncharged SC AA
Glutamine	Q	GLN	Polar Uncharged SC AA
Cysteine	C	CYS	Special Case SC AA
Selenocysteine	U	SEC	Special Case SC AA
Glycine	G	GLY	Special Case SC AA
Proline	P	PRO	Special Case SC AA
Alanine	A	ALA	Hydrophobic SC AA
Valine	V	VAL	Hydrophobic SC AA
Isoleucine	I	ILE	Hydrophobic SC AA
Leucine	L	LEU	Hydrophobic SC AA
Methionine	M	MET	Hydrophobic SC AA
Phenylalanine	F	PHY	Hydrophobic SC AA
Tyrosine	Y	TYR	Hydrophobic SC AA
Tryptophan	W	TRP	Hydrophobic SC AA

B. Supporting Information: Proton Transfer During DNA Strand Separation as a Source of Mutagenic Guanine-Cytosine Tautomers

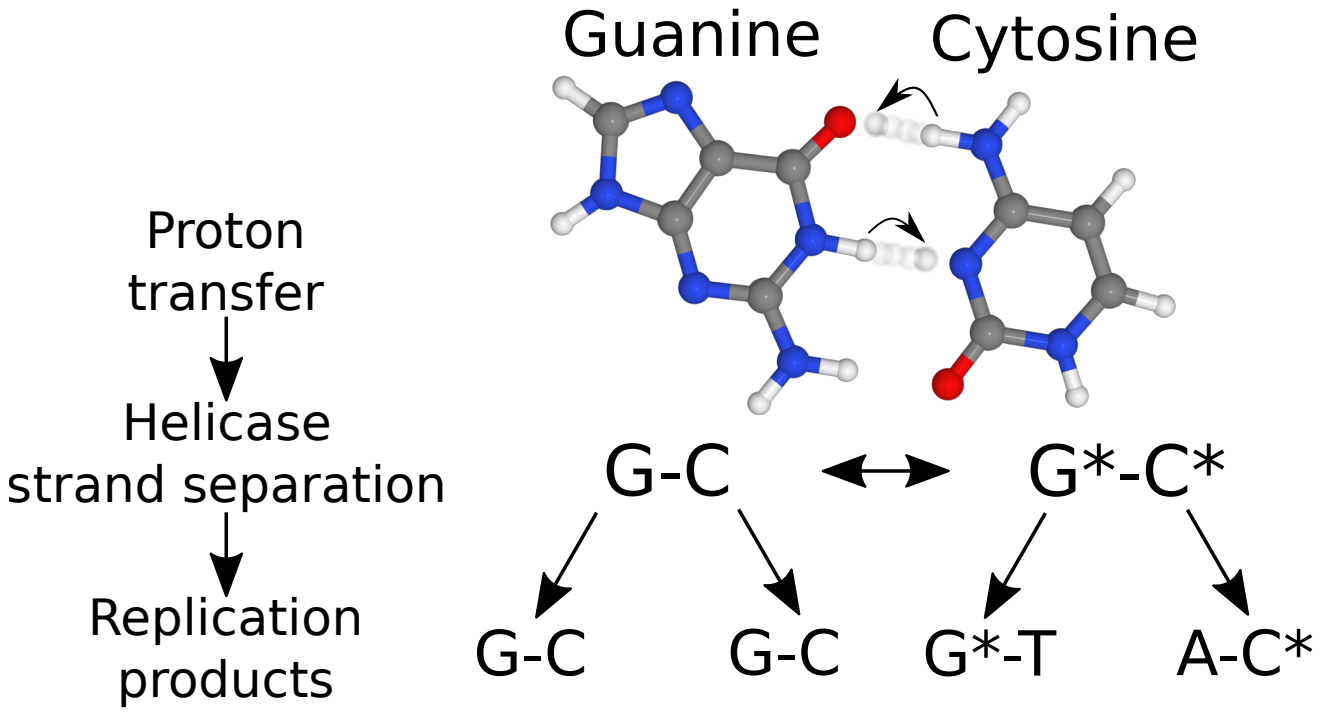
Supporting Information: Proton Transfer During DNA Strand Separation as a Source of Mutagenic Guanine-Cytosine Tautomers

Louie Slocombe,[1,2,*] Max Winokan,[1,†] Jim Al-Khalili,[3,‡] and Marco Sacchi[2,§]

[1] *Leverhulme Quantum Biology Doctoral Training Centre, University of Surrey, Guildford, GU2 7XH, UK.*
[2] *Department of Chemistry, University of Surrey, Guildford, GU2 7XH, UK.*
[3] *Department of Physics, University of Surrey, Guildford, GU2 7XH, UK.*

(Dated: October 5, 2022)

This material contains supplemental information for the results presented in the manuscript *Proton Transfer During DNA Strand Separation as a Source of Mutagenic Guanine-Cytosine Tautomers*. The data presented in the figures of the article are available from the corresponding authors upon reasonable request. The reaction pathways and structures are available on Github. Furthermore, the analysis source codes are also available on Github.



Supplementary Figure 1: The double proton transfer reaction along the hydrogen bonds between the G-C base pair. The scheme depicts the resulting products of this transfer. If the tautomers pass through the strand separation their daughter strands mismatch with the wrong base pair.

* louie.slocombe@surrey.ac.uk
† m.winokan@surrey.ac.uk
‡ J.Al-Khalili@surrey.ac.uk
§ m.sacchi@surrey.ac.uk

CONTENTS

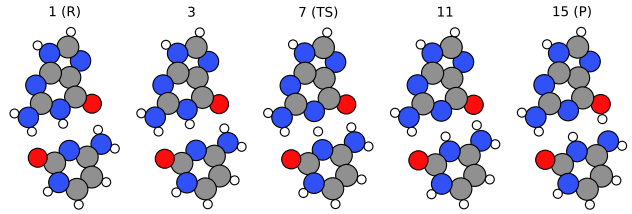
SUPPLEMENTARY NOTE 1: FURTHER REACTION PATHWAY ANALYSIS

Reading from left to right, each plot in a row of Fig. 2 shows the image number of the reaction path. We use 15 images along the reaction pathway, the first corresponding to the canonical form and the last to the tautomeric form. In each row, the seventh image is the transition state. Each row corresponds to a reaction path for a given separation distance. Thus reading down, we observe how the proton transfer landscape changes as the bases separate.

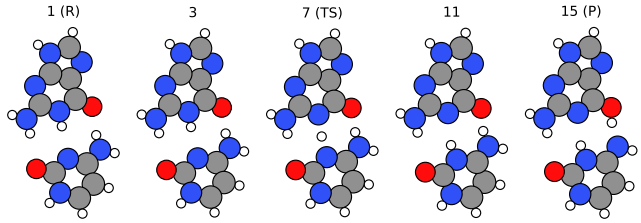
Fig. 2 shows the splitting of the canonical and tautomeric form of G-C while being constrained at the R-group mentioned before. Initially, there are visible changes other than the elongation of the hydrogen bonds holding the bases together. As the separation distance increases, there is a slight internal rotation of the bases relative to each other. The rotation minimises the length of the hydrogen bonds. There is a clear preference for the top hydrogen bond to maintain its equilibrium length while the base rotates. The rotation only happens since we are fixing the R-group, where the base joins onto the sugar and the rest of the DNA backbone. The R-group is where the base will be pulled from since this is the only covalently bonded link between the base and the rest of the DNA.

In comparison, there is some difference between the rotation of the base for the canonical vs the tautomeric form (see the bottom panel of Fig. 2). Here, the O-H bond of the tautomeric G offers a much more comprehensive hydrogen bonding range due to being on the outer edge of the molecule - in comparison to the standard form of C. As a result, the O-H bond of the tautomeric G remains in a hydrogen bond for much longer as the separation distance increases.

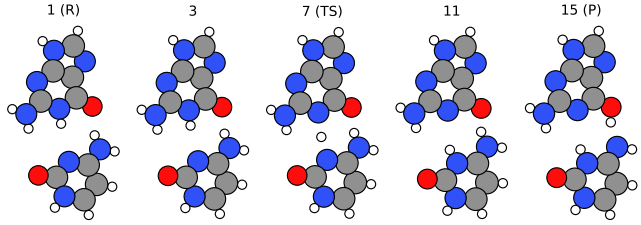
In reality, the splitting reaction coordinate is likely not a straight line due to the interactions with the DNA backbone and local water environment, restricting the movement of the bases. Consequently, we expect too see a range of opening angles as observed by molecular dynamics calculations.



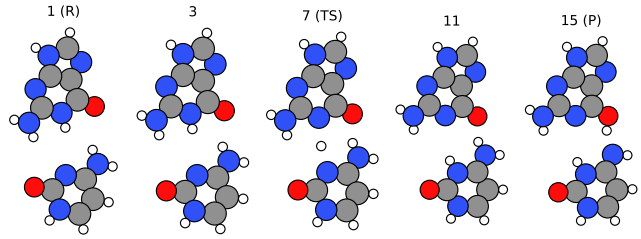
(a) 0.0 Å



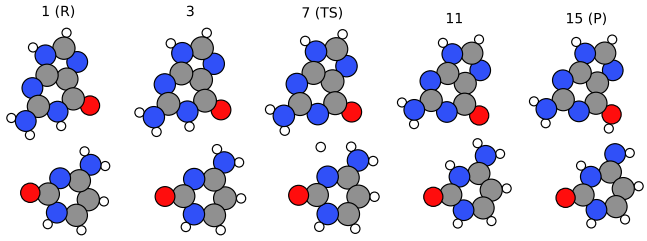
(b) 0.444 Å



(c) 0.889 Å

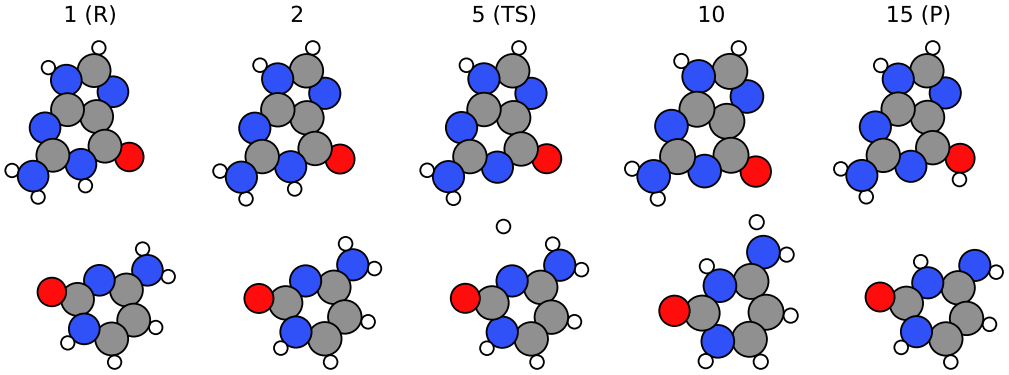


(d) 1.33 Å

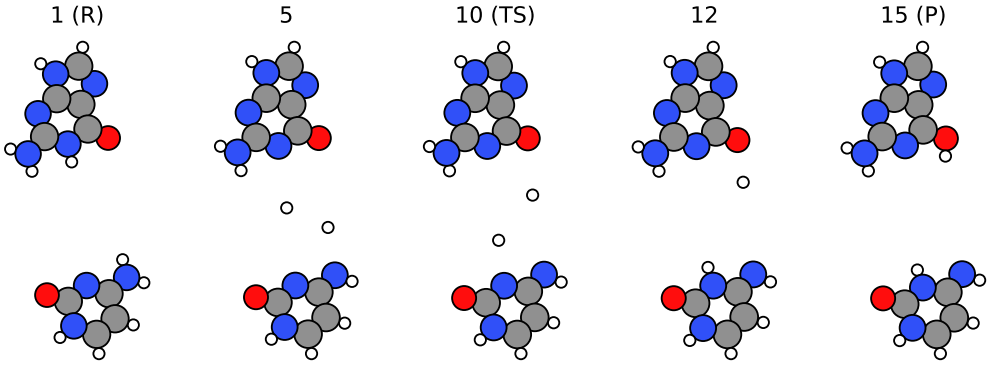


(e) 1.78 Å

Supplementary Figure 2: The double proton transfer reaction pathway. Each row is the reaction path, and each column is the pathway over increasing separation distances. The R label corresponds to the reactant, TS, the transition state, and P the product.



(a) 2.5 Å



(b) 5.0 Å

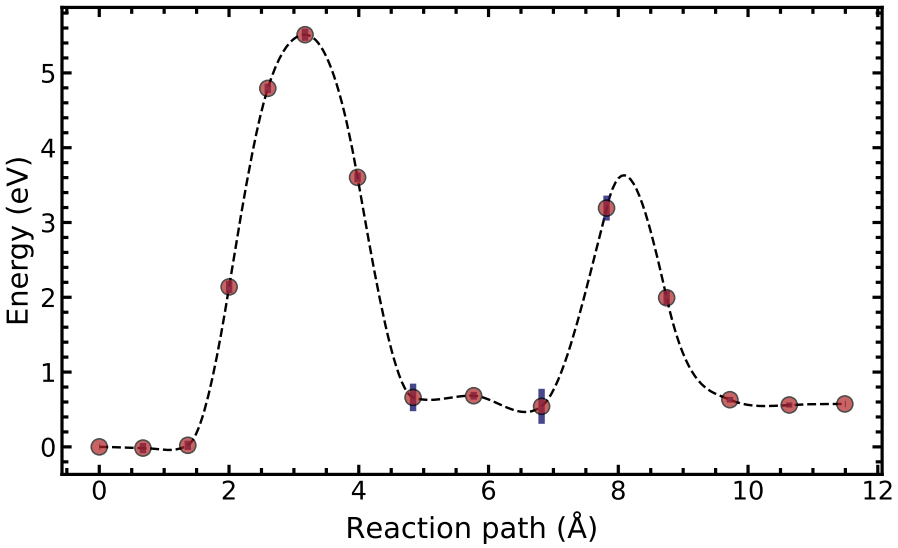
Supplementary Figure 3: The reaction pathway at large separation distances. Each row is the proton transfer reaction pathway, and while each column shows the mechanism at increasing separation distances.

SUPPLEMENTARY NOTE 2: PROTON TRANSFER AT LARGER SEPARATION DISTANCES

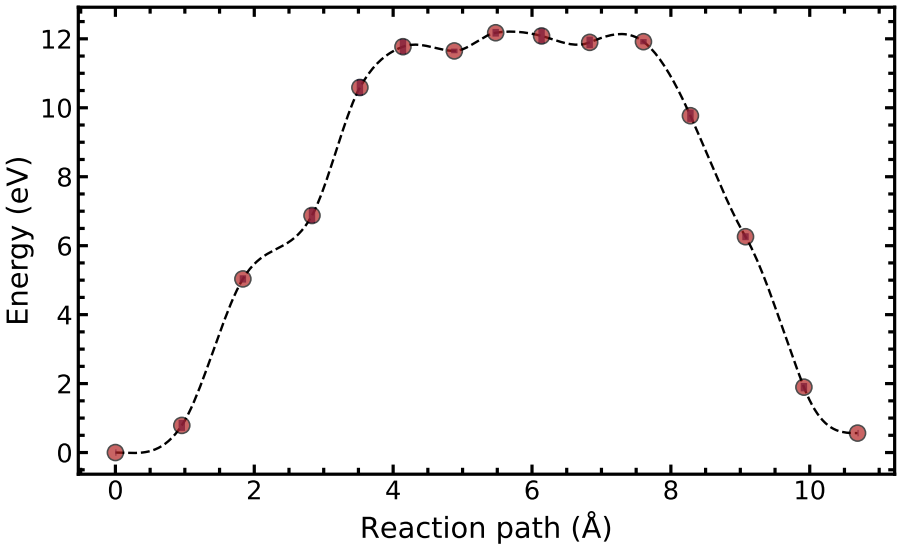
We explore the double proton transfer mechanism at separation distances greater than > 2.0 Å.

Fig. 3 highlights the reaction pathway of two separation distances, 2.5 Å and 5.0 Å. In both pathways, the rotation is diminished as it no longer becomes favourable to rotate to minimise bonding length, but to maintain the tail end of the hydrogen bonds so that the residual bonds are maximised. In the 2.5 Å case, the pathway is comprised of two separate movements of the protons. First, the middle hydrogen transfers, followed by the top proton in a similar mechanism observed before. On the other hand, for 5.0 Å the middle hydrogen begins to transfer before meeting the other halfway between the bases. This switch to a concerted asynchronous mechanism is likely due to the solvent's increasing role when the bases are far apart, since at this distance a water molecule could begin to creep in between the bases.

In Fig. 4 the energy of the reaction paths are shown. At a separation of 2.5 Å, two distinct large barriers are observed. The first barrier corresponds to the middle hydrogen moving. The single proton transfer has a deep well due to the high barriers on the reaction coordinate in both directions. In addition, it has a considerable asymmetry value of 0.668 eV relative to the canonical form. The second barrier corresponds to the top proton transferring. At this separation distance, it is the minimum energy path, but not the most energy efficient mechanism. Instead, we expect the intra-base transfer scheme, which is not considered in the reaction path calculation, to compete with this mechanism since it has a much lower and narrower reaction barrier. While at a separation of 5.0 Å the reaction barrier has doubled in height compared with that of the 2.5 Å separation, and the barriers have merged. We expect both the classical and quantum contribution to the rate to be low, resulting in a slow reaction that is unlikely to form biological products in either of these cases. Consequently, we expect proton transfer to be unlikely to occur above 2.0 Å.



(a) 2.5 Å

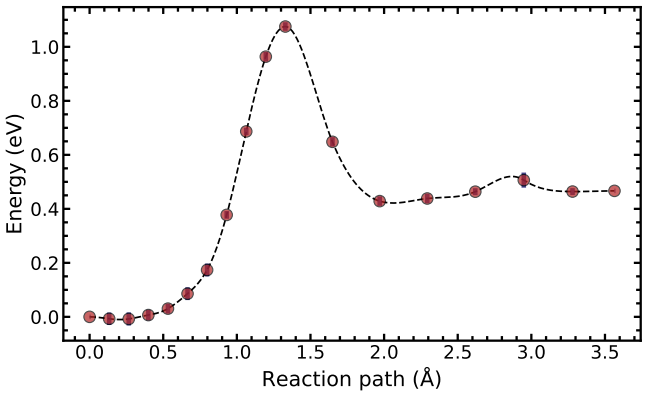


(b) 5.0 Å

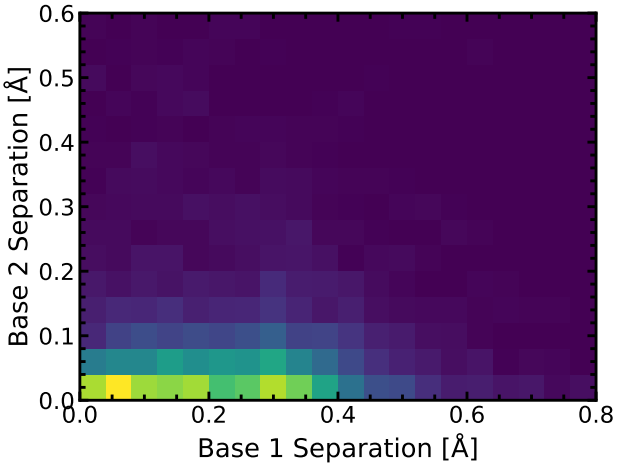
Supplementary Figure 4: Reaction paths at separation distances larger than 2.0 Å.

SUPPLEMENTARY NOTE 3: JUSTIFYING ASSUMPTIONS MADE IN STACKING CALCULATIONS

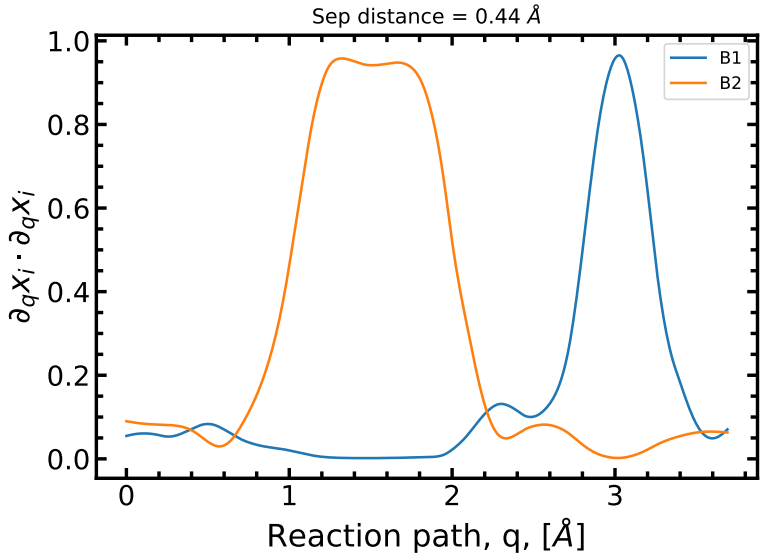
During our quantum mechanical calculations where the effect of stacking interactions between DNA base pairs on the double proton transfer was investigated, the base pair below the base pair of interest was assumed to be frozen during the separation event in the base pair above it. The resulting reaction profile is shown in Fig. 5. This assumption was not taken lightly, and backed up by our molecular dynamics simulations, in which we can see a clear separation of timescales between the opening of one base pair, and the next. This can be deduced from Fig. 6 where the separation of base pair 2 (plotted on the y-axis) remains near-zero during separation events in base pair 1 (x-axis).



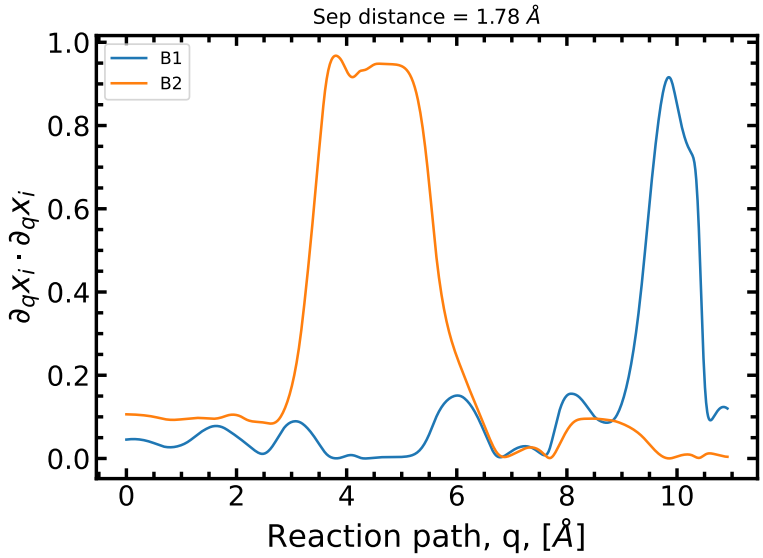
Supplementary Figure 5: The double proton transfer reaction pathway including the stacking interactions from a frozen base pair below. The separation distance is fixed to0.39 Å.



Supplementary Figure 6: The separation of base pair 2 during a separation event in base pair 1 in our Molecular Dynamics simulations of aqueous DNA.



(a) Double proton transfer reaction path at 0.44 Å



(b) Double proton transfer reaction path at 1.78 Å

Supplementary Figure 7: Evaluating the dot product of the derivative of the Cartesian vector with respect to the overall reaction coordinate at a fixed separation distance of 0.44 Å and 1.78 Å.

SUPPLEMENTARY NOTE 4: FURTHER NOTES ON PROTON TRANSFER ASYNCHRONICITY

We determine the asynchronicity to analyse further how the proton transfer mechanism changes during the base dissociation process. As a concept, asynchronicity is defined by a slight separation of the double proton transfer, i.e. one proton transfers, other heavy ions rearrange and then the second proton transfers.

To evaluate the derivatives shown in the methods section "Proton Transfer Asynchronicity", we first pass the Cartesian coordinate vectors into a Savitzky–Golay filter to suppress any spurious noise introduced by the uncertainty in the path which is inherent to the machine-learning approach of finding the reaction path. The filtered Cartesian coordinate vector and the reaction path are then interpolated using cubic splines.

In the top panel of Fig. 7, the first peak demonstrates B2 transferring before the second peak corresponding to

B1. The distance between each peak is used to determine the asynchronicity. In the second panel, the two prominent peaks are observed, along with some additional oscillations of the atoms rearranging. However, here the reaction path is much longer, and there is a clear separation between the two transfer peaks. The separation between the peaks demonstrates the large asynchronicity of the transfer process.

C. Supporting Information: Tautomerisation Mechanisms in the Adenine-Thymine Nucleobase Pair During DNA Strand Separation

Supporting Information: Tautomerisation Mechanisms in the Adenine-Thymine Nucleobase Pair During DNA Strand Separation

Benjamin King,[1,*] Max Winokan,[2,†] Paul Stevenson,[1,‡] Jim Al-Khalili,[1,§] Louie Slocombe,[3,¶] and Marco Sacchi[3,**]

[1] *Department of Physics, University of Surrey, Guildford, GU2 7XH, UK.*
[2] *Leverhulme Quantum Biology Doctoral Training Centre, University of Surrey, Guildford, GU2 7XH, UK.*
[3] *Department of Chemistry, University of Surrey, Guildford, GU2 7XH, UK.*

(Dated: February 16, 2023)

This resource provides supplementary information detailing the investigative procedures and results presented in the manuscript *Tautomerisation Mechanisms in the Adenine-Thymine Nucleobase Pair During DNA Strand Separation*. The data presented in the article, reaction pathways, structures, and analysis source codes are available on Github. Additional information is available from the corresponding authors upon reasonable request.

CONTENTS

* bk00346@surrey.ac.uk
† m.winokan@surrey.ac.uk
‡ p.stevenson@surrey.ac.uk
§ j.al-khalili@surrey.ac.uk
¶ louie.slocombe@surrey.ac.uk
** m.sacchi@surrey.ac.uk

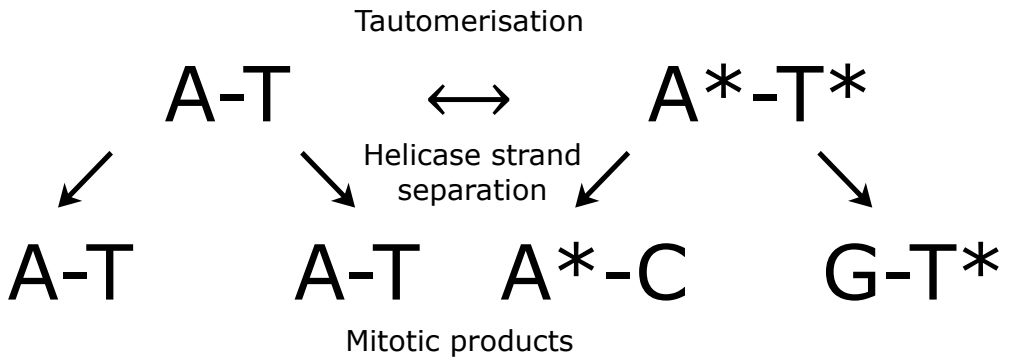


Figure S 1: A three-step description of the spontaneous mutagenesis process driven by tautomerisation of the adenine-thymine base pair.

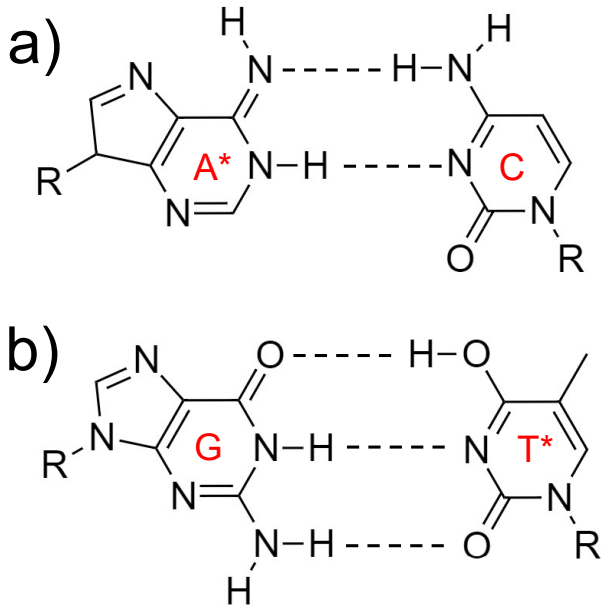


Figure S 2: The possible non-standard, Watson-Crick-like base pairings that can develop due to proton transfer. This is the incorporation of a genetic error by the tautomerisation mechanism.

SUPPLEMENTARY NOTE 1: ADENINE-THYMINE TAUTOMERISATION AND NON-STANDARD NUCLEOBASE PAIRING

The process of the tautomerisation reaction of adenine-thymine (A-T) is detailed in the manuscript. A broad overview of the tautomerisation process is presented in Supplementary Figure 1.

Once the non-standard forms of the bases, A* and T*, are established on isolated DNA strands during mitosis, the next step in the biological process is the formation of the new DNA strands by adding nucleotides at the active site of the DNA polymerase. The mutated bases A* and T* can bond in the non-standard, Watson-Crick-like pairings A*-C and G-T*. These pairings are shown in Supplementary Figure 2.

The consequence of the non-standard base pairings is that an A*-C pair is created in one strand of DNA and a G-T* pair in another instead of two A-T base pairs. These DNA code errors can evade replisome fidelity checkpoints and create equivalent errors through subsequent generations of DNA replication. The tautomerisation of A-T is an asynchronous, step-wise reaction with a zwitterionic intermediate product from a single proton transfer which

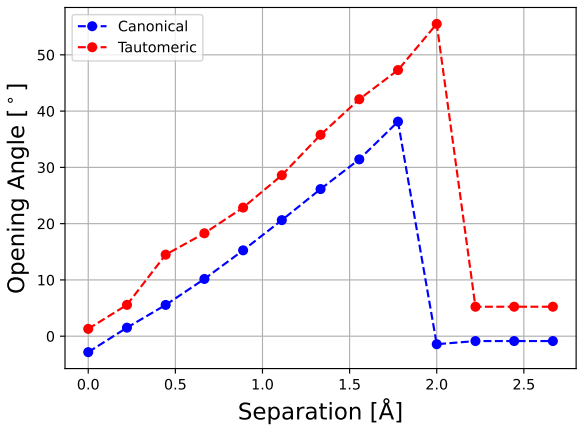


Figure S 3: The opening angles of the base pair for the canonical and tautomeric configurations across the range of DNA strand separation 0.0 Å-2.5 Å.

produces A^+-T^-. The zwitterionic bases can bond in the following pairs: A^+_{anti}-G_{syn}, A^+_{anti}-C_{syn}, and G-T^-_{wobble} (where 'anti' and 'syn' indicate anti-addition and syn-addition wherein the non-standard base bonding occurs upon non-standard faces of the base molecules and 'wobble' indicates a sterical bonding mismatch) [1, 2]. These base pairs are incompatible with the double helix structure, so mitosis will cease when the replication machinery encounters zwitterionic products. Therefore, only double proton transfer tautomerisation products are relevant to spontaneous mutagenesis.

SUPPLEMENTARY NOTE 2: MOLECULAR DYNAMICS SIMULATIONS

We use molecular dynamics (MD) to investigate the detail of the DNA strand separation process. We calculate the occurrences of the opening angles θ across the range of DNA strand separation 0.0 Å - 2.0 Å and estimate the speed at which the strand separation occurs. The MD system is constructed of 14 base pairs within a double strand of DNA, initially in equilibrium and an aqueous environment. In the quantum mechanical investigations, the opening angle smoothly increases during strand separation (see Supplementary Figure 3). We examine whether this observation remains true in the biological ensemble in the MD calculation. We study the opening angle in two scenarios: where the B1 bond is the first to open and where the B2 bond is the first to open. We collect a histogram of opening angles and their occurrences across a 75° range for each scenario of B1 and B2 opening first. We treat a positive angle as opening from the B2 end of the base pair and a negative angle as opening from the B1 end of the base pair.

The software we use to conduct molecular dynamics (MD) simulations is GROMACS 2018 [3]. The system consists of 14 base pairs within a DNA duplex with the base code: T^3TTGTACGTACAAA^5. A 2 nm x 2 nm x 2 nm) simulation box is constructed to surround the DNA system with explicit SPCE solvent and, neutralising sodium ions, a CHARMM36 [4] force field is employed for all the MD simulations. We minimise a group of replica systems, equilibrate them over a scale of 50 ps of NVT ensemble over incremental time steps of 1 fs and simulate 10 different separation forces with a maximum force 12 kJ mol^{-1} nm^{-1}. The system temperature is maintained, by a Nose-Hoover thermostat, at 310 K, using 0.2 ps as a coupling constant. The data is collected over 66 system replicas. We collect the time series statistics of the two hydrogen bond lengths between the adenine and thymine base pair and pass these statistics through a Savitsky-Golay filter.

We define the opening angle between the bases with the scheme presented in Supplementary Figure 4. We differentiate between positive and negative θ values by which end of the base pair opens first. An initial opening of the bond N_T-H_T—N_A (bond B2) is defined by a positive angle, and an initial opening of the bond O_T—H_A-N_A (bond B1) is defined by a negative angle.

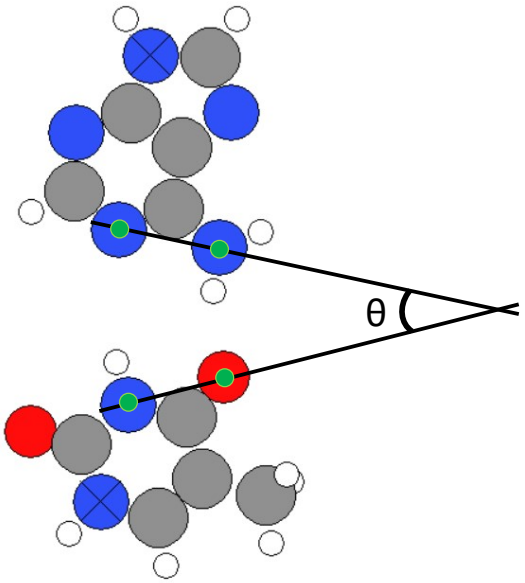


Figure S 4: The scheme by which the base pair's opening angle, θ, is determined. The angle is the dot product of the 3-dimensional vectors that pass through the hydrogen donor and acceptor atoms, indicated with green spots.

SUPPLEMENTARY REFERENCES

[1] A. Gheorghiu, P. Coveney, and A. Arabi, The influence of base pair tautomerism on single point mutations in aqueous DNA, Interface focus **10**, 20190120 (2020).

[2] Y. Kim, F. Bertagna, E. M. D'Souza, D. J. Heyes, L. O. Johannissen, E. T. Nery, A. Pantelias, A. Sanchez-Pedreño Jimenez, L. Slocombe, M. G. Spencer, *et al.*, Quantum biology: An update and perspective, Quantum Reports **3**, 80 (2021).

[3] H. Berendsen, D. van der Spoel, and R. van Drunen, Gromacs: A message-passing parallel molecular dynamics implementation, Computer Physics Communications **91**, 43 (1995).

[4] K. Hart, N. Foloppe, C. M. Baker, E. J. Denning, L. Nilsson, and A. D. MacKerell Jr, Optimization of the charmm additive force field for dna: Improved treatment of the bi/bii conformational equilibrium, Journal of chemical theory and computation **8**, 348 (2012).

D. Supporting Information: Multiscale simulations reveal the role of PcrA helicase in protecting against spontaneous point mutations in DNA

Supporting Information for: Multiscale simulations reveal the role of PcrA helicase in protecting against spontaneous point mutations in DNA

Max Winokan[a,1], **Louie Slocombe**[b,2], **Jim Al-Khalili**[c,3], **and Marco Sacchi**[b,4]

[a]Leverhulme Quantum Biology Doctoral Training Centre, University of Surrey, Guildford, GU2 7XH, UK.
[b]Department of Chemistry, University of Surrey, Guildford, GU2 7XH, UK.
[c]School of Mathematics and Physics, University of Surrey, Guildford, GU2 7XH, UK.
[1]E-mail: m.winokan@surrey.ac.uk
[2]E-mail: louie.slocombe@surrey.ac.uk
[3]E-mail: j.al-khalili@surrey.ac.uk
[4]E-mail: m.sacchi@surrey.ac.uk

June 2023

This document contains the supplementary information for the article entitled: "Multiscale simulations reveal the role of PcrA helicase in protecting against proton transfer in DNA". In addition to the computational methodology (Section 1), additional simulation details (Section 2), and additional simulation results (Section 3), input, parameter, and analysis files sufficient to reproduce the results published in this work are available on Github at https://github.com/mwinokan/GC-tautomerism-in-PcrA-Helicase.

1 Computational Methodology

This section describes the computational methodology used in this work. Broadly, the computational research undertaken in this work involved the mapping of double proton transfer within Guanine-Cytosine in an aqueous DNA duplex (as a control), and in the bacterial PcrA Helicase-DNA complex.

1.1 System Preparation

The duplex B-DNA structure was generated with Avogadro(1) using the sequence shown in Table S1. The atom and residue names were modified for use with the CHARMM36(2–4) force field and the Gromacs(5) software and suitable 3' and 5' termini were added using MolParse(6).

A modified and optimised version of the PcrA Helicase-DNA substrate complex PDB entry 3PJR,(7) was obtained from Yu et al. (8).

With tools included with Gromacs, topologies were generated for these systems using the March 2019 CHARMM36 force field. The systems were solvated in a cubic box of SPCE water ensuring at least 1 nanometre of clearance between any non-solvent atom and the edge of the box. The system was neutralised with the addition of sodium ions.

1.2 Molecular Dynamics Equilibration

All energy minimisation, molecular mechanics, and molecular dynamics were performed using Gromacs 2018(5), with the March 2019 version of the CHARMM36 force field, and SPCE water.

Prior to any dynamics energy minimisation was performed on both these systems using a steepest descent algorithm until the maximum force did not exceed 12 kJ/mol/nm. The molecular dynamics parameters for the geometry optimisation can be found in gromacs_parameters/minim_12.mdp.

Each of these systems was placed simulated in an NVT ensemble using a leap-frog integrator, timestep of 1 fs, and Nose-Hoover temperature coupling with time constant 0.2 ps and reference temperature 310 K, and three-dimensional periodic boundary conditions for a total of 3 ns. Within this time the systems were verified to have been equilibrated as the temperatures and RMSD stabilised. An example molecular dynamics parameter file for the molecular dynamics equilibration can be found in gromacs_parameters/eq500ps_2CX.mdp.

1.3 Hybrid Quantum Mechanics / Molecular Mechanics

Hybrid QM/MM calculations were performed with the Gromacs-CP2K distribution(5, 9). The nucleobases of the relevant GC base pair were defined as the QM region, whose electronic structure would be solved CP2K(9) using density functional theory with the BLYP exchange correlation functional, DZVP-MOLOPT-GTH basis set, GTH-BLYP potential, electrostatic (point charge) embedding, and the VDW3 dispersion correction. For QM/MM dynamics, Gromacs applied the leap-frog algorithm, with forces obtained from DFT in CP2K for the QM region and CHARMM36 elsewhere.

1.4 Umbrella Sampling

Umbrella Sampling was performed with Gromacs-CP2K at a CHARMM36/DFT/BLYP level of theory using four distance reaction coordinates describing the double proton transfer as illustrated in Figure S1. Each umbrella sampling window applied a 20000 kJ/mol/nm^2 harmonic potential along the four reaction coordinates, and was simulated for at least 8ps per replica per window. Depending on the level of equilibration, up to the first 2 ps of each simulation were discarded. An example for the steered molecular dynamics parameter input file for Gromacs can be found in gromacs_parameters/umb_8ps_f20k_2CM_BLYP_notraj.mdp, and a corresponding example CP2K input file in cp2k_parameters/HS41R1_BLYP_VDW_w008.inp.

The double proton transfer pathways that were sampled are illustrated in Figure S2. The concerted/synchronous DPT is a simple linear interpolation between the canonical GC and the G*C* tautomer. The asynchronous pathway was taken from the equilibrium distance GC dimer ML-NEB pathways reported by Slocombe et al.(10)

The potentials of mean force (PMF) along the sampled reaction were obtained by the Weighted Histogram Analysis Method (WHAM) implemented in Gromacs with a tolerance of 1.0×10^{-6} and statistical variance was estimated with 100 bootstrapping samples. The four reaction coordinates were projected onto one dimension in an average relative change from the canonical minimum, which is the default for Gromacs' WHAM. The reaction coordinate z is thus defined as:

$$z = \frac{(\mathrm{RC1} - \mathrm{RC1}_0) + (\mathrm{RC3} - \mathrm{RC3}_0) - (\mathrm{RC2} - \mathrm{RC2}_0) - (\mathrm{RC4} - \mathrm{RC4}_0)}{4} + \frac{\mathrm{RC1}_0 + \mathrm{RC3}_0}{2} \tag{1}$$

Where $\mathrm{RC1}_0$ is the canonical reference point for RC1, etc.

The reaction asymmetries obtained from PMFs were used to calculate the equilibrium constants (K) using Equation 2. Where ΔG is a reaction free energy, R is the gas constant, and T is temperature (310K in this work).

$$\ln K = -\frac{\Delta G}{RT} \tag{2}$$

1.5 Instantaneous Transfer Surface

To determine the instantaneous double proton transfer potential energy surface, snapshots were taken from the QM/MM umbrella sampling simulations corresponding to the canonical GC. 256 replica structures were created using Python and MolParse(6) mapping the two dimensional transfer landscape of the two protons. Each proton had its position interpolated between 0.9 and 2.5 Angstrom distance from its covalently bonded donor atom. The python code in analysis_code/grid.py was used to generate the array of input structures.

For each of these grid structures a single point energy (SPE) was calculated using the previously described electrostatically embedded QM/MM in CP2K with two levels of theory: identical to the umbrella sampling methodology with BLYP/DZVP-MOLOPT-GTH and B3LYP/DZVP-MOLOPT-GTH which also utilised the auxillary density matrix method using the cFIT3 basis. For each SPE calculation a new CP2K input was constructed using the following parameter templates: BLYP cp2k_parameters/template_BLYP.inp and B3LYP cp2k_parameters/template_B3LYP_new.inp.

The grid of SPEs was interpolated using the CloughTocher piecewise cubic, C1 smooth, curvature-minimizing interpolant in two-dimensions inplemented in SciPy.(11) Within this surface local minima were determined using the BFGS optimiser implemented in ASE,(12) and the minimum energy path connecting the canonical and tautomeric states was obtained using the nudged elastic band (NEB) algorithm also implemented in ASE. All of this functionality was achieved the analysis code in analysis_code/plot_grid.py

1.6 G*C* Ensemble decay

To investigate the metastable nature of the G*C* tautomer, post DPT product structures were generated from each instantaneous PES, and these were placed in unbiased 310K NVT QM/MM MD simulations, the parameters for which can be found in gromacs_parameters/md5ps_qmmm_vel.mdp. The velocities were taken from the original US snapshot used to generate the surface.

To quantitatively compare the decays the time at which the system leaves the local potential energy minima has been estimated by considering the point where the two reaction coordinates cross over. For example, in the DC:N4-DC:H41 – DG:O6 transfer described by RC1 and RC3 (see Subsection 1.4) we consider the proton to be in the canonical well as long as RC1 < RC3. Additionally once the vector between the canonical minima and [RC1,RC3] is within 0.1 Angstrom, we consider the system to have 'arrived' at the canonical GC state. These lifetime results are shown in Figure S3.

1.7 Instantaneous PES at non-equilibrium separation

To obtain a strand separation trajectory in the DNA-Helicase complex a short (5ps) steered molecular dynamics simulation was performed on an equilibrated structure. A constant separating force (500 kJ/mol/nm^2) was applied to the DC:N1 and DG:N9 backbone atoms forcing them directly apart. See gromacs_parameters/pull_v1.mdp for exact simulation parameters.

1.8 Mutations of the PcrA Helicase sequence

Sequences for PcrA Helicase from 59 species were analysed around site 624 which is an asparagine (N624) in 51% of cases, other minor populations of 624 include arginine (R624, 24%), glutamine (Q624, 14%), and tyrosine (Y624, 12%). Table S4 shows the statistics of mutations in the sites 622 to 626. When considering vicinal residues to N624 multiple motifs appear. Motifs with N624 remain the most common, with FG**N**IQ, and FG**N**IH making up the 51% majority. When considering sites 622 to 626, a tyrosine at 624 (Y624) is always part of a FG**Y**TN motif (12% frequency), while sequences containing glutamine Q624 and arginine R624 are more diverse, i.e. the vicinal positions 622, 623, 625, and 626 are not strongly correlated to the amino acid in 624.

1.9 Additional Molecular Dynamics trajectories for aqueous DNA and the PcrA Helicase-DNA complex

In addition to the 3 ns of equilibration time, three systems were simulated across a further 30 ns each to observe slower dynamics. The same Molecular Dynamics (MD) parameters were used as described in Subsection 1.2.

2 Additional Simulation Details

This section provides further details regarding the validation of the methods described in Section 1.

2.1 QM/MM level of theory

The exchange correlation functional BLYP with and without VDW3 dispersion corrections was compared to PBE and the semi-empirical PM6 for the concerted DPT in aqueous DNA, and the PMFs are shown in Figure S4. PM6 provided the greatest statistical uncertainty between all tested levels of theory and was thus not a suitable description of this reaction. Between BLYP, PBE, and BLYP+VdW3 there are small differences in the produced concerted double proton transfer barriers. BLYP+VdW3 was chosen for the umbrella sampling due to its faster sampling convergence. See Subsection 2.2.

2.2 Sampling convergence

To ensure that the potentials of mean force produced from the umbrella sampling simulations have converged the RMS difference between the reaction profiles was calculated for a given sampling time. For the chosen level of theory (BLYP VDW) a convergence is reached within 500ps of umbrella sampling, see Figure S5. Table S2 shows the total sampling time for each umbrella sampling PMF obtained in this work. Despite the computational expense of QM/MM umbrella sampling, we have ensured that the asynchronous pathway PMF for each studied system has converged to above 95%.

2.3 GC dimer separation during umbrella sampling simulations

During the umbrella sampling reactions the average donor-acceptor distances of the hydrogen bonds involved in proton transfer varied significantly and are reported in Figure S6. Mathematically the separation can be described as $(RC1 + RC2 + RC3 + RC4)/4$. We report a compression of the dimer during the sampled reaction, as previously observed by Slocombe *et al.* in (13) and (10). For the concerted double proton transfer a single compression is observed, corresponding to the single transition state, except the Helicase N-1 case which has a compressed tautomeric state. For the asynchronous pathway we observe two compression events corresponding to the two transition states resisting each proton transfer, again, the N-1 case was an outlier and did not exhibit this compression.

This compression shows the high energetic cost of the proton transfer is alleviated by molecular degrees of freedom, not directly involved in the proton transfer. This motivates the need for a description of the proton transfer without such atomic rearrangement, especially if the proton transfer is known to occur near-instantaneously due to quantum tunnelling.

3 Additional Simulation Results

This section provides additional results to support discussions in the main text.

3.1 Umbrella sampling results for the synchronous DPT

While the asynchronous double proton transfer (DPT) was found to be energetically preferred with both umbrella sampling (US) and the instantaneous minimum energy path (MEP), we also report the synchronous proton transfer in Table 1 of the main article. The corresponding potentials of mean force (PMF's) are shown in S7.

3.2 Conformational behaviour of aqueous DNA and the PcrA-Helicase in longer molecular dynamics trajectories

Aqueous DNA, the wild type PcrA Helicase-DNA complex (N624), and the point mutation N624A were simulated in longer 33ns NVT MD (see Subsection 1.9).

For Aqueous DNA, an overview of the dynamics is shown in Figure S15, which shows significant fraying at the ends of the duplex, while more stability is maintained in the middle. Figure S18 shows the donor-acceptor distance for the two hydrogen bonds involved in the DPT from these extended MD trajectories. The distribution for both bonds - DGO6-DCN4 (proton 1's donor and acceptor, and sum of RC1 and RC2) and DGN1-DCN3 (proton 2's donor and acceptor, and sum of RC3 and RC4) - centre on a donor-acceptor distance of 3 Å.

For the wild-type PcrA Helicase-DNA complex with asparagine N624, the conformations present in the extended MD are summarised in Figure S16, in which panel A shows significant fluctuations in the single-stranded DNA not passing through the enzyme, but relative stability across the rest of the complex. For base pair N, Panel B shows the conformation used for umbrella sampling and corresponds to the 3 Å peak of the donor-acceptor distance distribution shown in Figure S18. Panel C shows the a common alternative conformation corresponding to the more separated peak around 5.7 Å. While the GC base pair N-1 did open up slightly further than the aqueous DNA reference (deduced from the longer tail in the donor-acceptor distance distribution), the distribution still peaks at the equilibrium value (3 Å).

For the PcrA-DNA complex with the N624A point mutation, the smaller sidechain of alanine compared to asparagine has allows for increased mobility of the single-stranded DNA and a different alternative conformation is formed (see panel C of Figure S17). In this alternative conformation, DG662 and DC701 (base pair N) are no longer the last in the duplex as the DT663-DA702 (pair N+1) can reform. It is also observed that the alternative conformation does not stretch the base pair N (see the donor-acceptor distance in Figure S18), as most of the distribution is still centred around 3 Å. However, likely due to the increased mobility of the ssDNA, base pair N is opens more frequently during the simulations than in aqueous DNA, but not as often as in complex with wild-type PcrA.

3.3 Instantaneous DPT transfer surfaces

The individual instantaneous DPT transfer surfaces described in the main text are shown in Figures S9, S11, and S13 for duplex DNA, wild type PcrA Helicase (base pair N), and for helicase with the N624A mutation, respectively. The minimum energy paths through the surfaces are shown in Figures S10, S12, and S14, for the same three systems.

Table S1. Sequence of duplex DNA used for simulations.

Chain	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	C_5	A	G	T	G	C	A	G	T	G	C	T	C	G	T	T	T	T	T	T	T	T	T	T_3
2	G_3	T	C	A	C	G	T	C	A	C	G	A	G	C	A	A	A	A	A	A	A	A	A	A_5

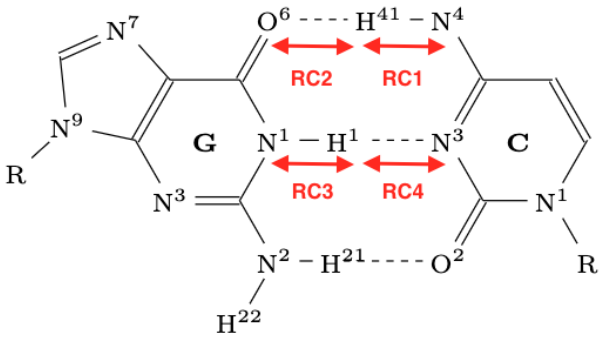


Fig. S1. Reaction coordinates used to map the double proton transfer

From these figures we can determine that the instantaneous PES and MEP through it is highly sensitive to conformational differences between snapshots. The locations and energies of the PES minimas are summarised in Table S3. The locations of the non-canonical minima vary between replicas of the same system, as the length of the hydrogen bonds and separation of the GC dimer fluctuate during the MD and US (see 2.3). Similarly, conformational differences lead to large energetic differences between the non-canonical minima between replicas, by as much as 28% relative standard deviation.

Table S2. QM/MM umbrella sampling: total simulation time for free energy calculations

System	Path	Base Pair	Site 624	Sampling [ps]
Aqueous Duplex DNA	Concerted	-	-	436.0
Aqueous Duplex DNA	Asynchronous	-	-	789.0
PcrA Helicase-DNA Complex	Concerted	N	N	408.0
PcrA Helicase-DNA Complex	Asynchronous	N	N	1008.7
PcrA Helicase-DNA Complex	Concerted	N-1	N	288.0
PcrA Helicase-DNA Complex	Asynchronous	N-1	N	480.0
PcrA Helicase-DNA Complex	Asynchronous	N	A	797.8

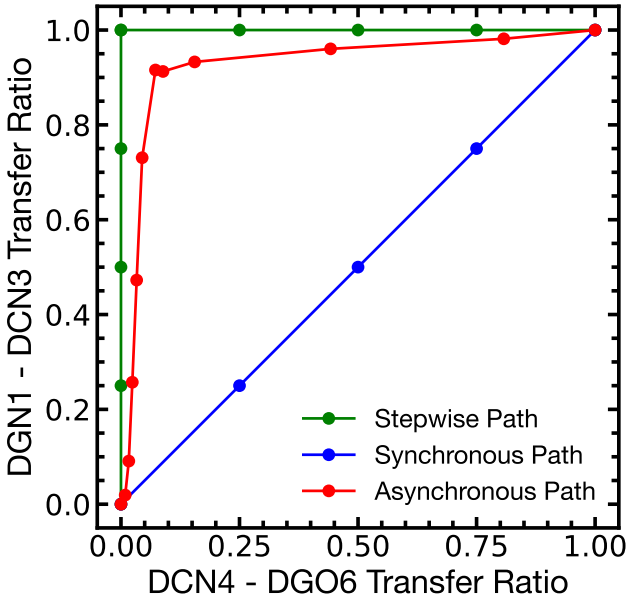


Fig. S2. Reaction paths used in Umbrella Sampling

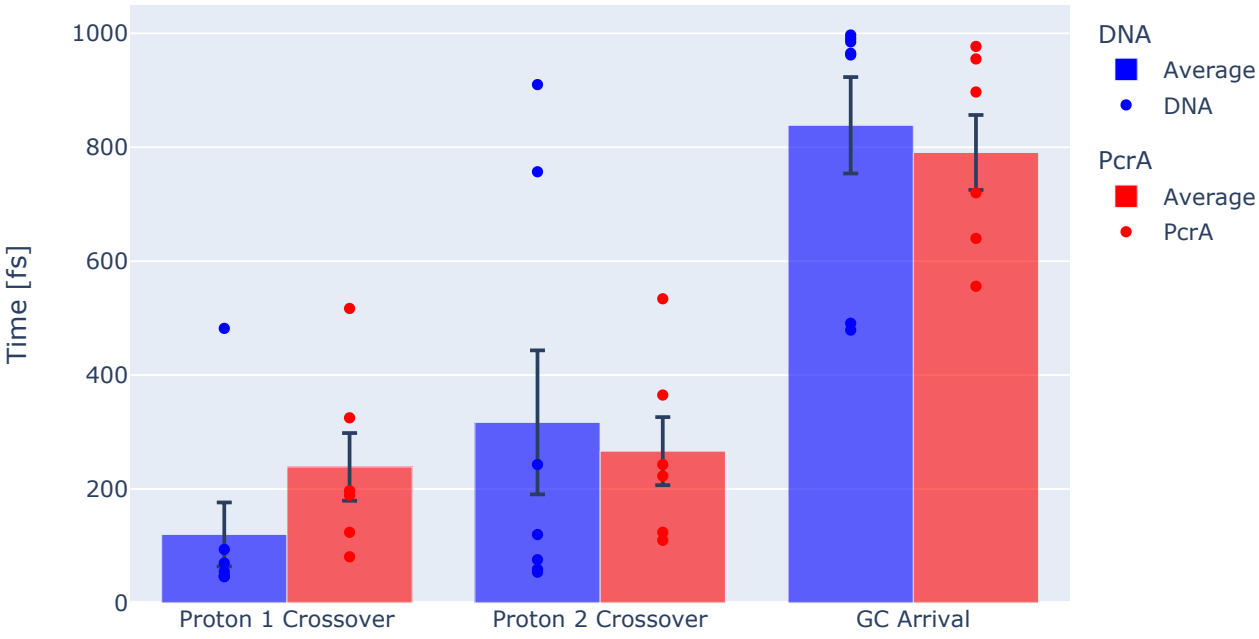


Fig. S3. Lifetime estimation for the G*C* tautomer in duplex DNA and the PcrA Helicase-DNA complex.

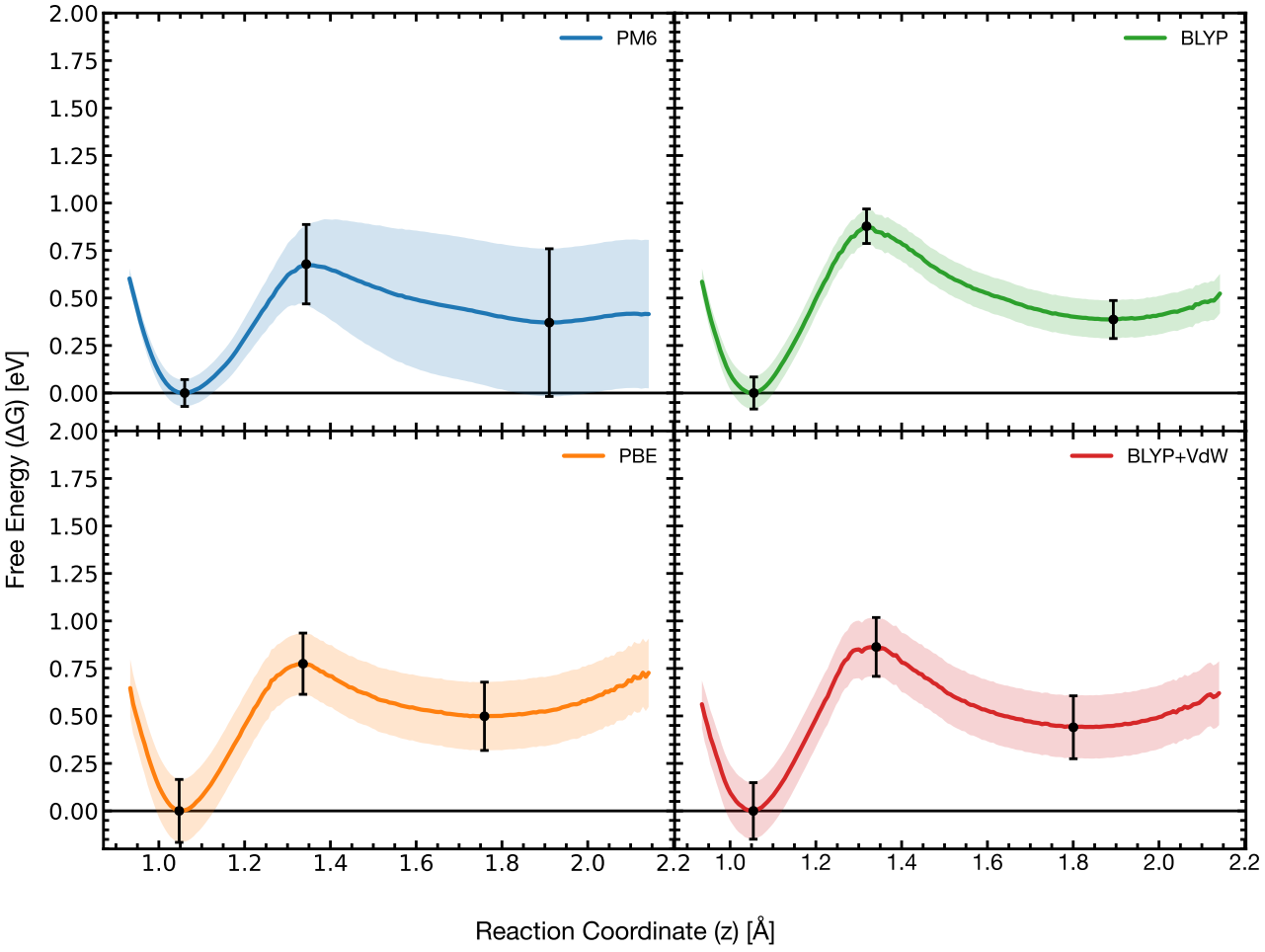


Fig. S4. Umbrella sampling potentials of mean force for different levels of QM theory.

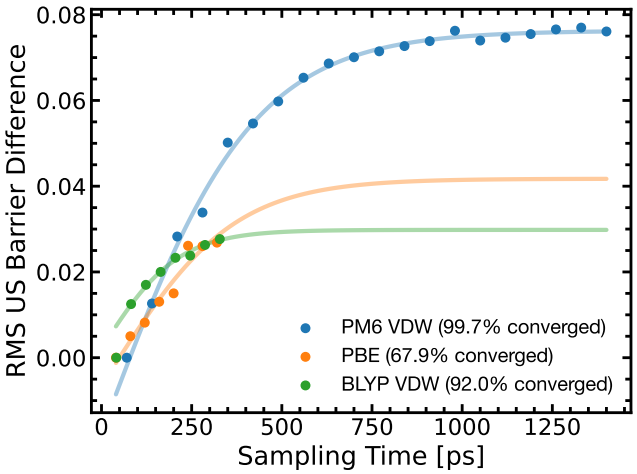


Fig. S5. Convergence of the umbrella sampling reaction profiles with increasing sampling time

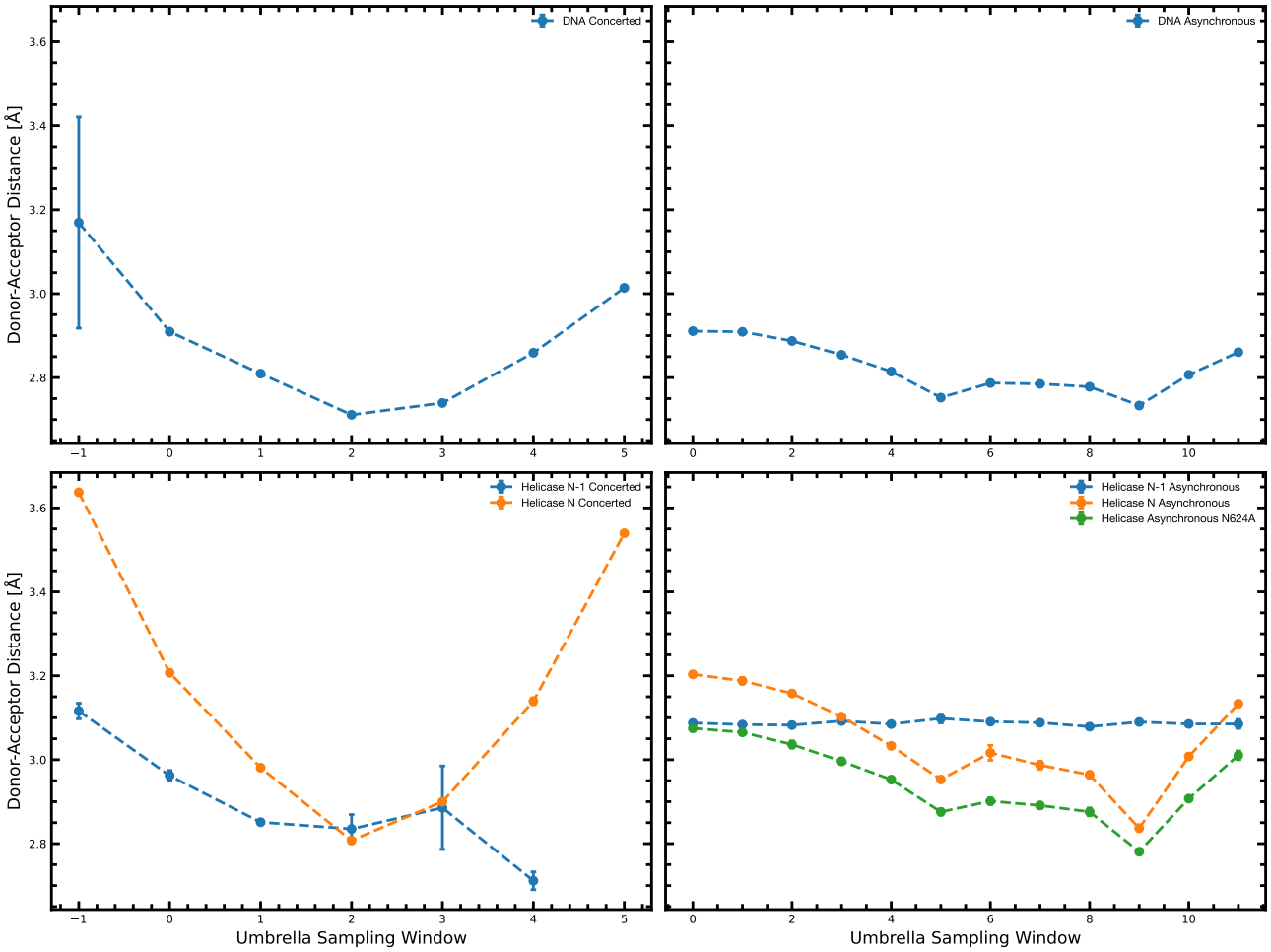


Fig. S6. Average GC dimer hydrogen bond donor-acceptor distance during umbrella sampling.

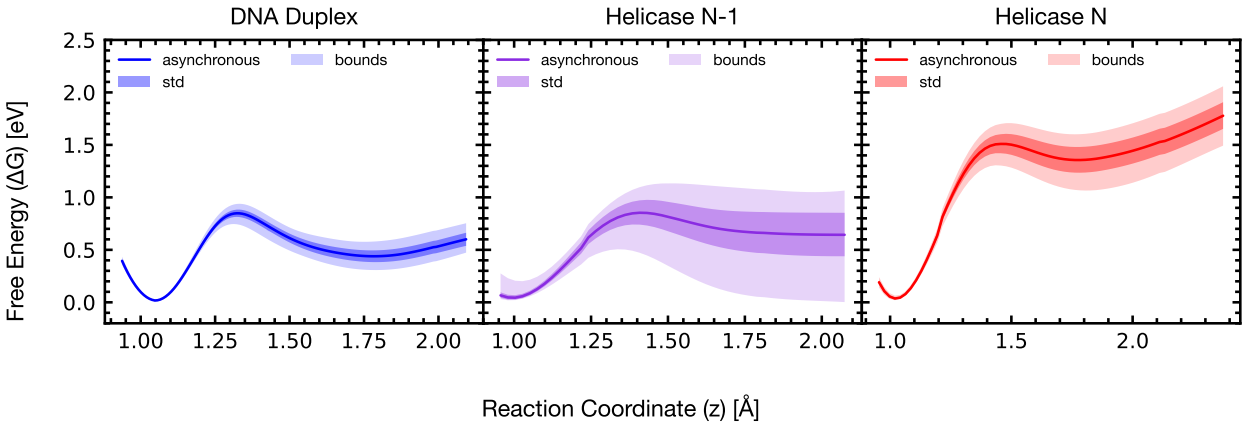


Fig. S7. Potentials of mean force (PMF) for the synchronous double proton transfer in guanine-cytosine obtained from QM/MM Umbrella Sampling (US) simulations for Aqueous Duplex DNA (A) and two base pairs in the wild type PcrA Helicase-DNA Complex, pair N-1 (B) and pair N (C).

Table S3. Properties of energetic minima in instantaneous potential energy surfaces. RC1 and RC3 are the donor-hydrogen distance for each hydrogen bond involved in the DPT (see Figure S1), and ΔE is the potential energy relative to the canonical (GC) minimum.

Duplex DNA	GC			G^-C^+			G^+C^-			G^*C^*		
Property	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)
mid_L01	1.028	1.048	0.000	1.044	1.779	0.650	-	-	-	1.922	1.830	1.303
mid_L02	1.030	1.036	0.000	1.050	2.002	0.953	-	-	-	1.751	2.047	1.461
mid_L03	1.033	1.048	0.000	1.054	1.941	0.482	-	-	-	1.812	1.966	0.918
mid_L05	1.046	1.061	0.000	1.056	1.511	0.426	-	-	-	1.735	1.637	0.689
L03_c4	1.018	1.053	0.000	1.046	1.769	0.697	-	-	-	1.764	1.825	1.344
L05	1.020	1.060	0.000	1.043	1.773	0.456	1.895	1.083	1.618	1.957	1.827	1.032
L11	1.029	1.047	0.000	1.053	1.835	0.592	-	-	-	1.740	1.876	1.168
avg.	1.029	1.050	0.000	1.049	1.801	0.608	1.895	1.083	1.618	1.811	1.858	1.131
std.	0.008	0.008	0.000	0.005	0.145	0.170	0.000	0.000	0.000	0.085	0.119	0.249
std. (%)	0.8	0.8	0.0	0.5	8.1	27.9	0.0	0.0	0.0	4.7	6.4	22.1
PcrA (N624)	GC			G^-C^+			G^+C^-			G^*C^*		
Property	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)
HL01	1.016	1.041	0.000	1.038	1.899	1.500	2.230	1.059	1.964	2.262	1.997	1.735
HL02	1.003	1.046	0.000	1.035	1.984	1.204	2.237	1.062	2.316	2.279	2.041	1.674
HL03	1.018	1.046	0.000	-	-	-	2.113	1.074	1.794	2.161	1.827	1.493
HL04	1.013	1.038	0.000	1.031	1.855	1.685	2.182	1.052	2.267	2.224	1.957	2.280
HL05	1.013	1.038	0.000	1.032	1.864	1.391	2.188	1.058	2.086	2.222	1.936	1.812
HL06	1.041	1.019	0.000	1.041	1.874	1.817	2.273	1.041	1.581	2.282	1.985	1.648
grid_new	1.017	1.048	0.000	1.033	1.984	1.445	2.286	1.055	2.572	2.322	2.081	2.214
avg.	1.017	1.039	0.000	1.035	1.910	1.507	2.216	1.057	2.083	2.250	1.975	1.837
std.	0.011	0.009	0.000	0.004	0.054	0.199	0.055	0.009	0.311	0.049	0.075	0.275
std. (%)	1.0	0.9	0.0	0.4	2.8	13.2	2.5	0.9	14.9	2.2	3.8	15.0
PcrA (N624A)	GC			G^-C^+			G^+C^-			G^*C^*		
Property	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)	RC1(Å)	RC3(Å)	ΔE(eV)
N624A11	1.014	1.043	0.000	1.037	1.897	1.240	2.129	1.066	2.259	2.196	1.970	1.809
N624A12	1.019	1.046	0.000	1.047	1.822	1.030	1.862	1.067	1.622	1.935	1.890	1.368
N624A13	1.016	1.039	0.000	-	-	-	2.286	1.053	2.021	2.303	1.907	1.883
N624A14	1.013	1.041	0.000	1.034	1.889	1.229	2.280	1.056	2.425	2.291	1.962	1.886
avg.	1.015	1.042	0.000	1.039	1.869	1.166	2.139	1.060	2.082	2.181	1.932	1.737
std.	0.002	0.003	0.000	0.006	0.033	0.096	0.172	0.006	0.302	0.148	0.035	0.215
std. (%)	0.2	0.3	0.0	0.5	1.8	8.3	8.0	0.6	14.5	6.8	1.8	12.4

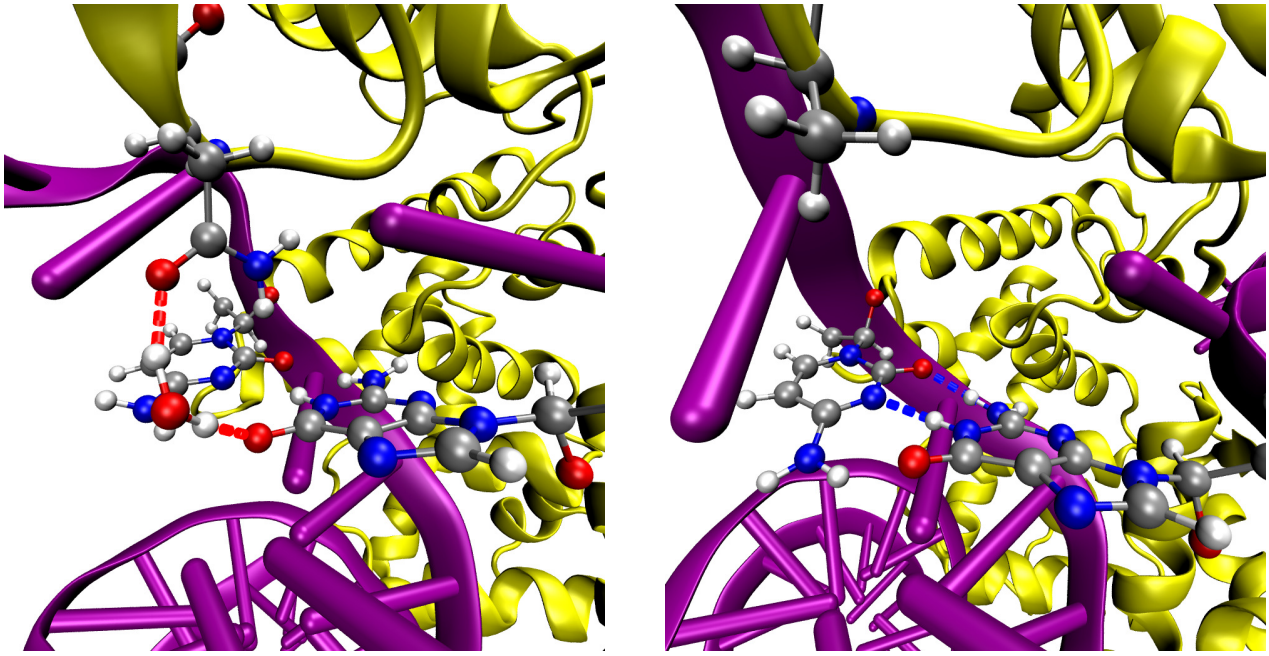


Fig. S8. Archetypical conformations of the PcrA Helicase-DNA duplex with the wild type sequence and N624A point mutation. The base pair N GC dimer and amino acid in site 624 are shown in CPK representation.

Table S4. Amino acid residues in sites 622-626 across 59 species of PcrA Helicase. Asterisks (*) are used as a wildcard denoting any amino acid. N is the number of occurences of the given motif, also given as a percentage in parentheses.

622	623	624	625	626	N (%)
*	*	N	*	*	30 (51%)
*	*	R	*	*	14 (24%)
*	*	Q	*	*	8 (14%)
*	*	Y	*	*	7 (12%)
*	G	N	I	*	30 (51%)
*	G	R	T	*	14 (24%)
*	G	Q	T	*	8 (14%)
*	G	Y	T	*	7 (12%)
F	G	N	I	Q	24 (41%)
F	G	Q	T	H	7 (12%)
F	G	Y	T	N	7 (12%)
F	G	N	I	H	6 (10%)
F	G	R	T	N	5 (8%)
Y	G	R	T	N	2 (3%)
F	G	R	T	G	2 (3%)
F	G	R	T	M	1 (2%)
F	G	R	T	A	1 (2%)
F	G	R	T	T	1 (2%)
F	G	R	T	S	1 (2%)
F	G	Q	T	N	1 (2%)
Y	G	R	T	E	1 (2%)

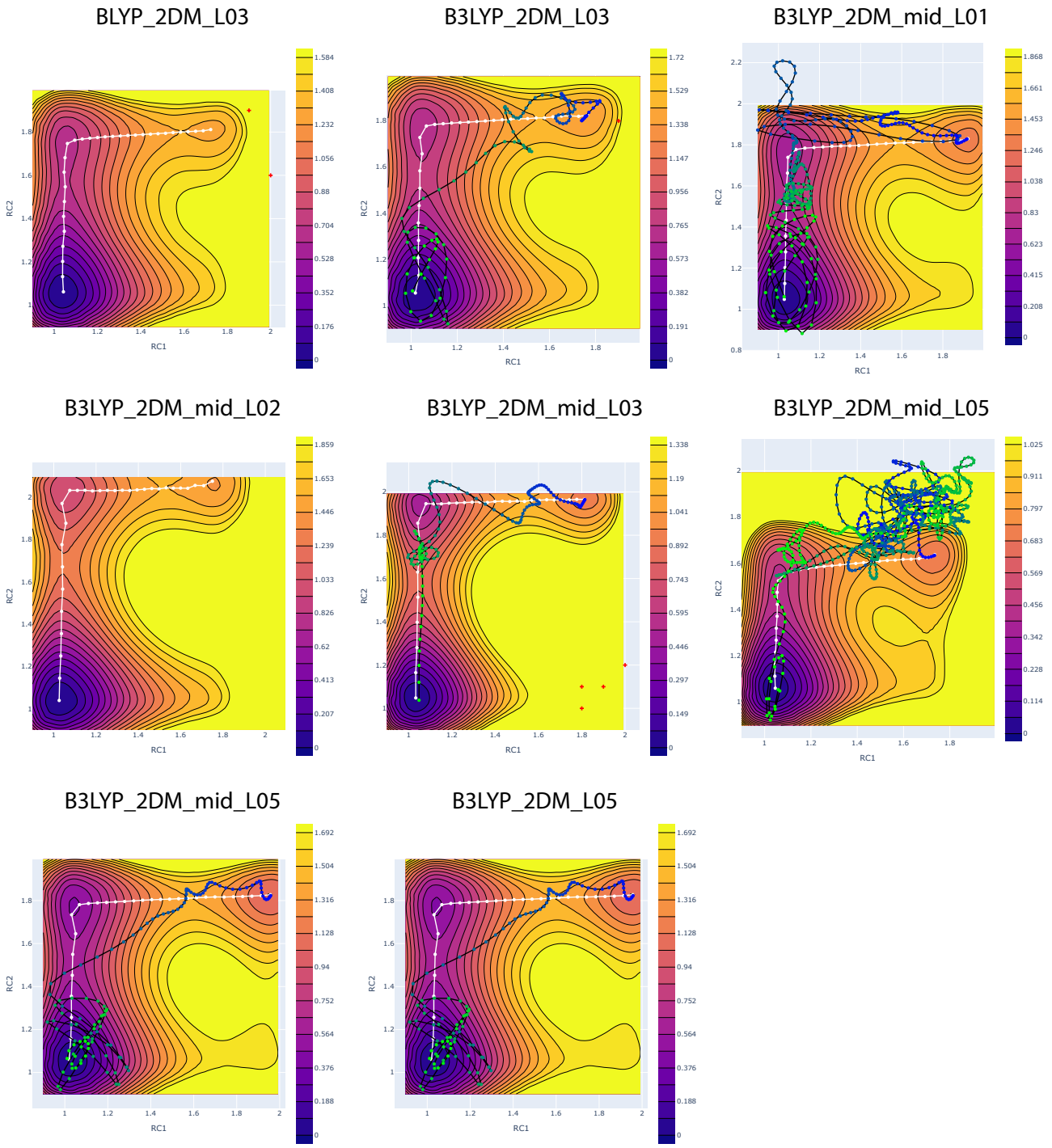


Fig. S9. Instantaneous DPT PES for duplex DNA. Contoured heatmap illustrating the DPT's instantaneous energy surface, the minimum energy pathway within this landscape (white line with circular dots), and the decay path from G*C* to GC (black line with multicoloured circles). For the decay path, the dots are coloured according to the simulation time, with blue corresponding to the start of the simulation and green at the end of the decay.

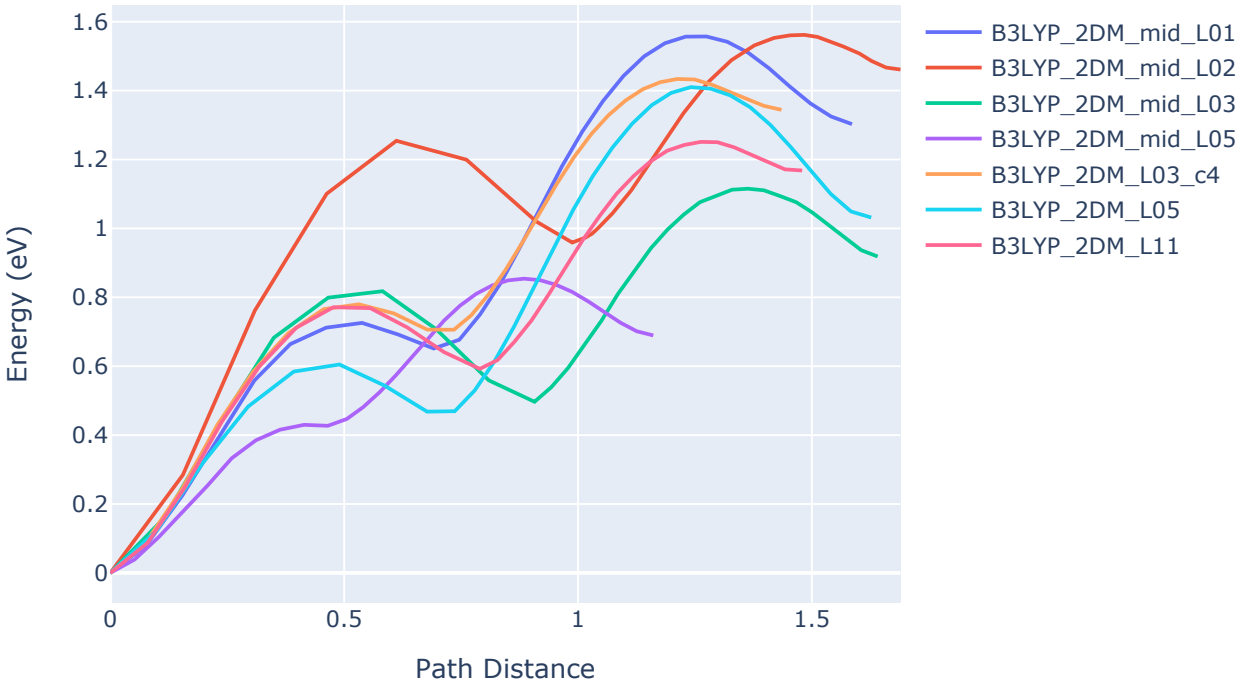


Fig. S10. Minimum energy paths through the instantaneous DPT PES for duplex DNA.

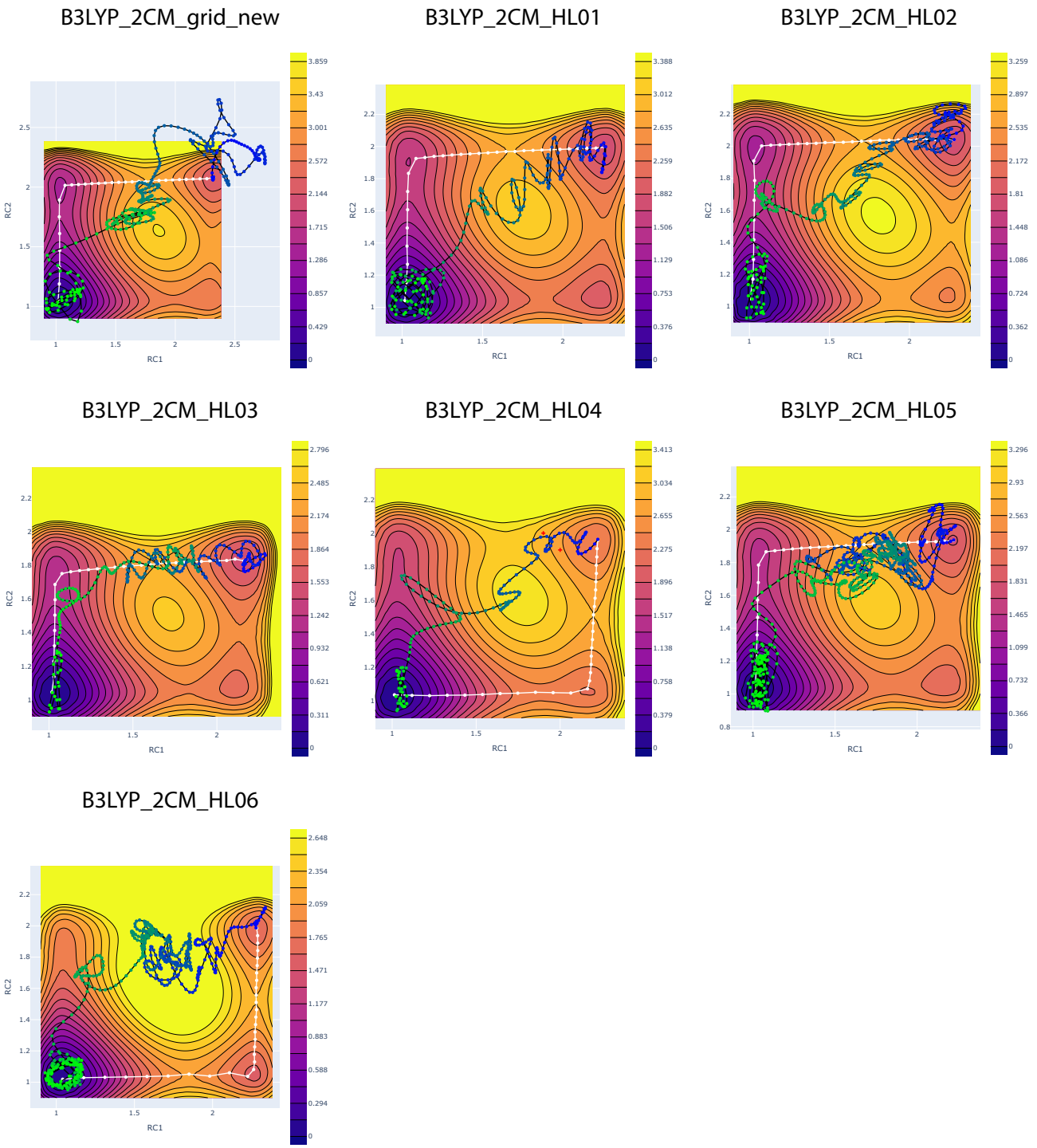


Fig. S11. Instantaneous DPT PES for the PcrA Helicase-DNA complex (base pair N). Contoured heatmap illustrating the DPT's instantaneous energy surface, the minimum energy pathway within this landscape (white line with circular dots), and the decay path from G*C* to GC (black line with multicoloured circles). For the decay path, the dots are coloured according to the simulation time, with blue corresponding to the start of the simulation and green at the end of the decay.

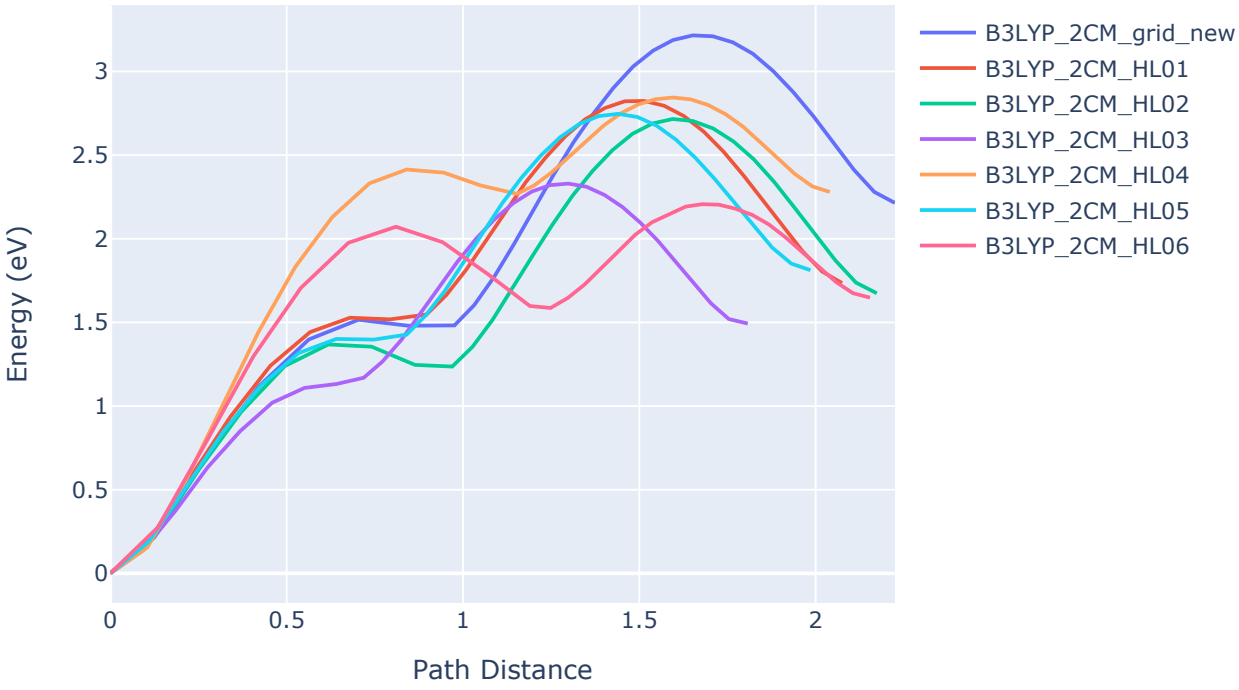


Fig. S12. Minimum energy paths through the instantaneous DPT PES for the PcrA Helicase-DNA complex (base pair N).

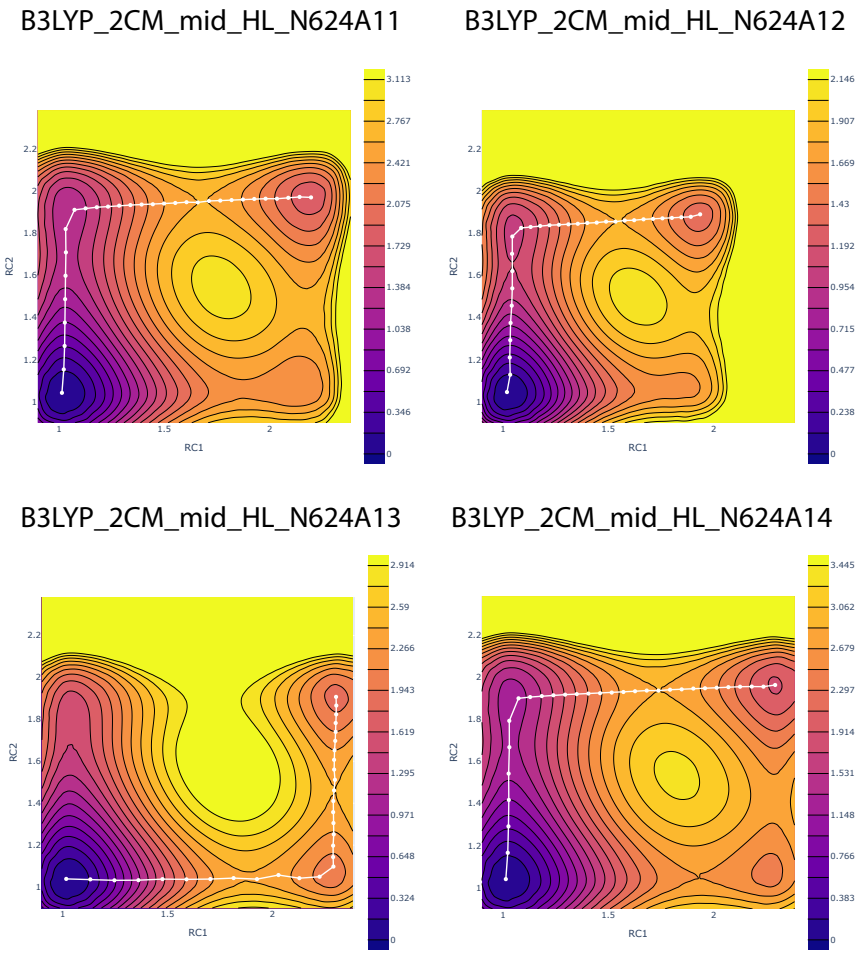


Fig. S13. Instantaneous DPT PES for the PcrA Helicase-DNA complex (base pair N) with the N624A point mutation. Contoured heatmap illustrating the DPT's instantaneous energy surface, the minimum energy pathway within this landscape (white line with circular dots).

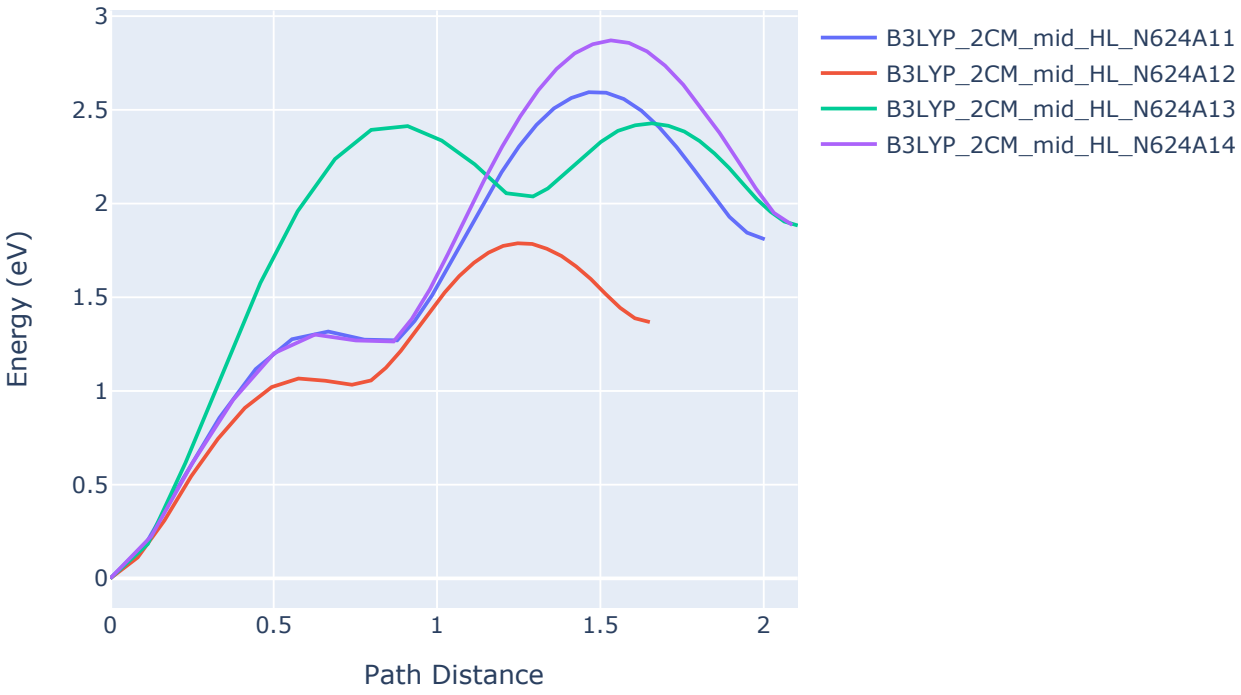


Fig. S14. Minimum energy paths through the instantaneous DPT PES for the PcrA Helicase-DNA complex (base pair N) with the N624A point mutation.

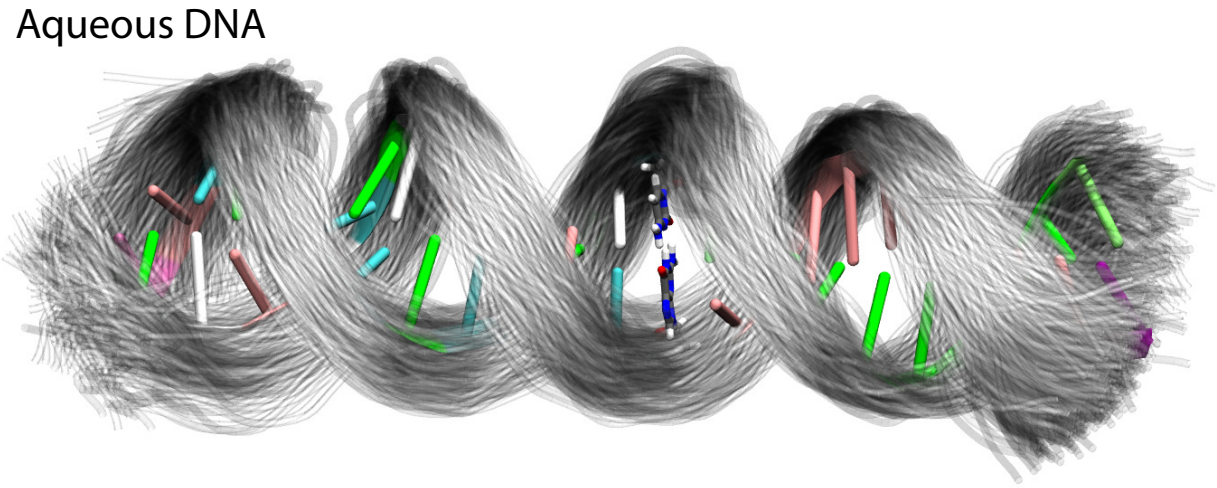


Fig. S15. DNA conformations present during 30ns of NVT MD. One archetypical conformation is shown as a coloured cartoon ladder structure with colours according to the residue: guanine is white, cytosine is cyan, adenine is lime green, and thymine is salmon pink. The GC base pair considered for the DPT is shown as a 'licorice' representation. 600 regularly spaced snapshots from the last 30ns of MD are superimposed as a transparent tube representing the backbone of the DNA.

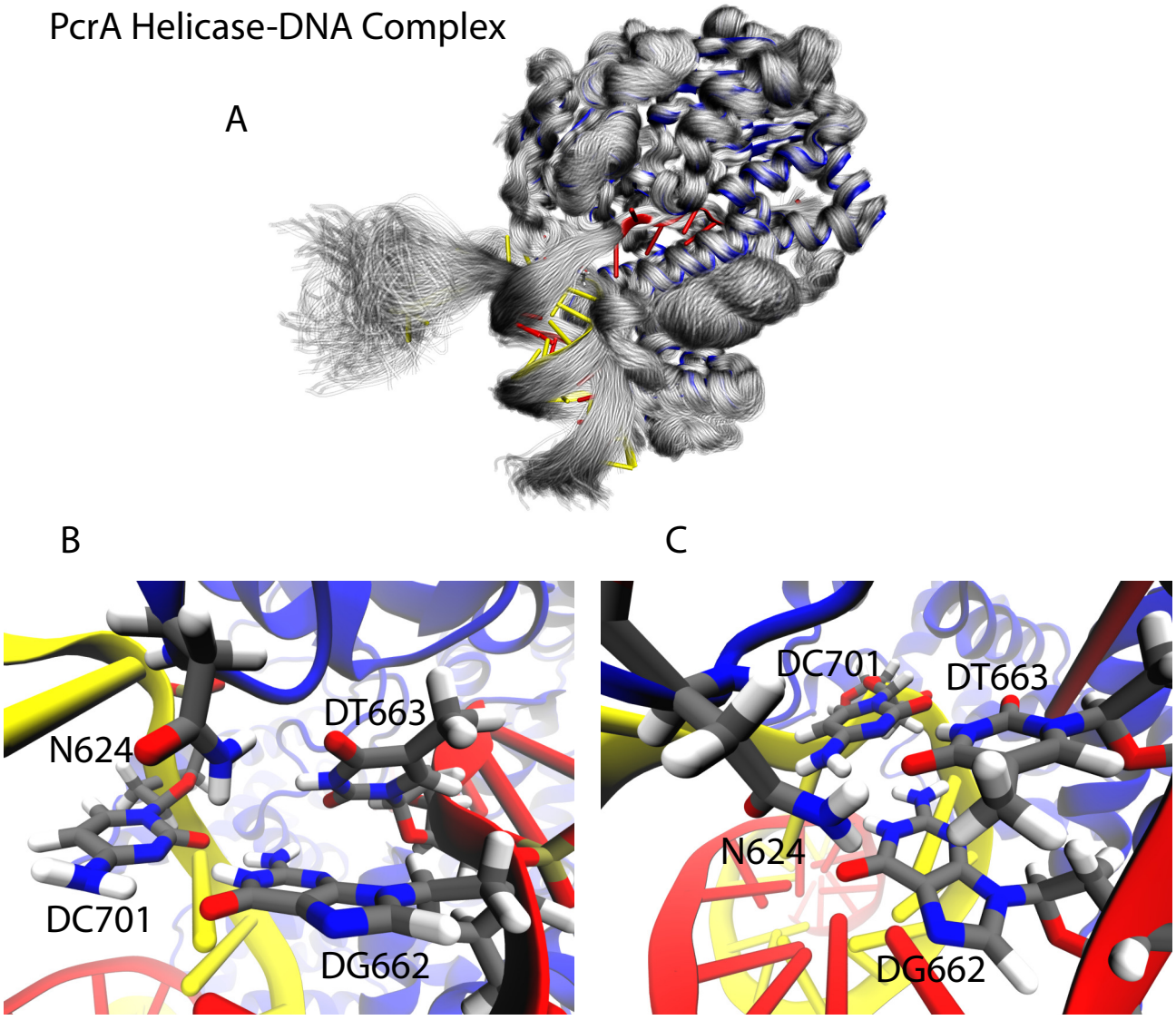


Fig. S16. (A) PcrA Helicase-DNA complex conformations present during 30ns of NVT MD. One archetypical conformation is shown as a coloured cartoon structure with colours according to the chain: DNA is yellow and red, and the protein is blue. 600 regularly spaced snapshots from the last 30ns of MD are superimposed as a transparent tube representing the backbone of the DNA and protein. (B) Primary conformation for the DPT site as used in QM/MM. It can be seen that the asparagine N624's sidechain is in the proximity of the hydrogen bonds between DG662 and DC701. (C) Alternate conformation for the DPT site where the GC tautomerism is presumed irrelevant as DG662 and DC701 have dissociated and a new wobble conformation between DC701 and DT663 is formed. Asparagine N624 has also rotated away from the hydrogen bonds.

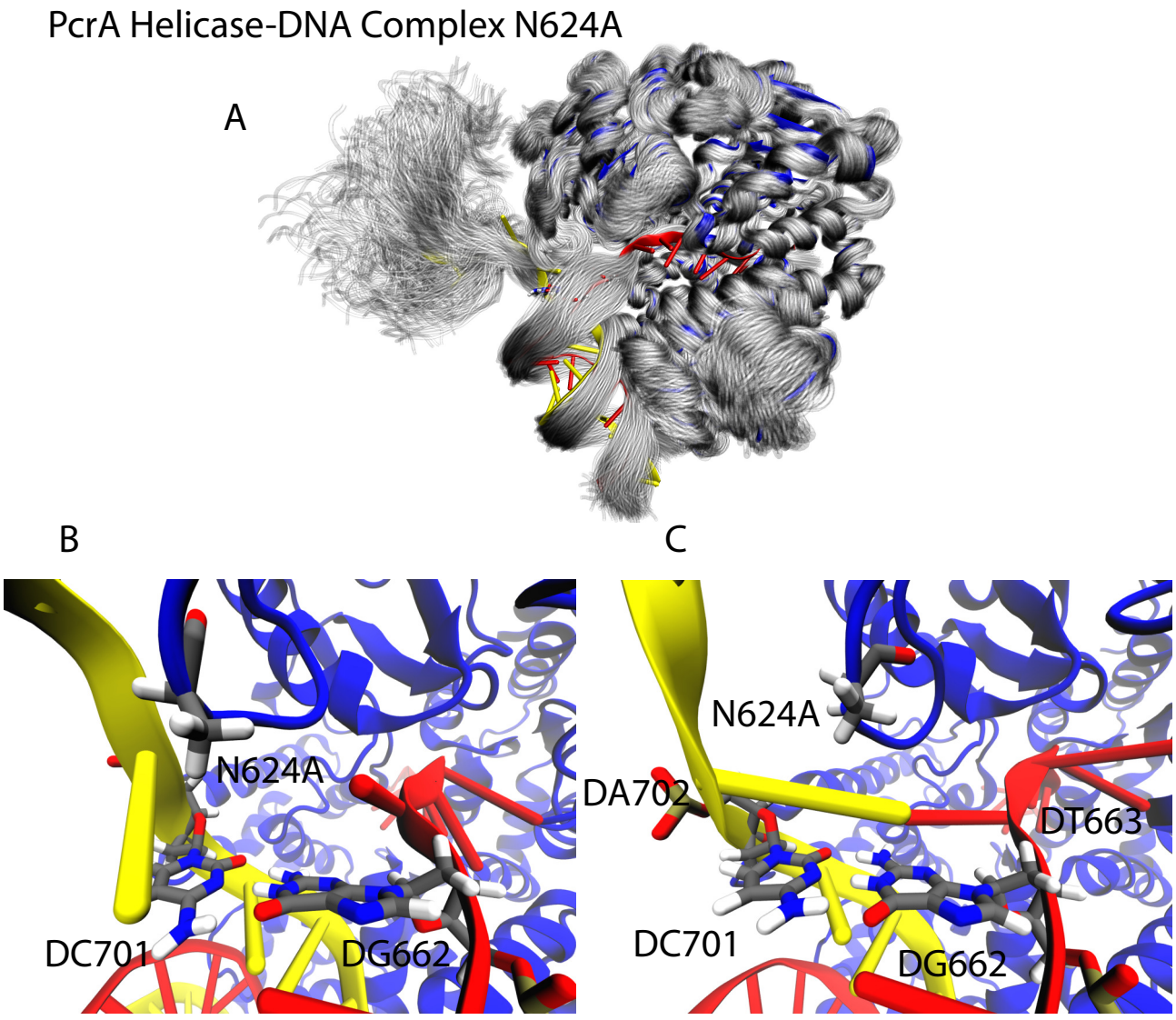


Fig. S17. (A) N624A point mutated PcrA Helicase-DNA complex conformations present during 30ns of NVT MD. One archetypical conformation is shown as a coloured cartoon structure with colours according to the chain: DNA is yellow and red, and the protein is blue. 600 regularly spaced snapshots from the last 30ns of MD are superimposed as a transparent tube representing the backbone of the DNA and protein. (B) Primary conformation for the DPT site as used in QM/MM. The shorter sidechain of the alanine N624A compared to the wild type's asparagine N624 does not interfere with the hydrogen bonding between DG662 and DC701. (C) Alternate conformation for the DPT site where the base pair DG662 and DC701 (base pair N) is no longer the last in the duplex as the pair DT663-DA702 (pair N+1) has reformed.

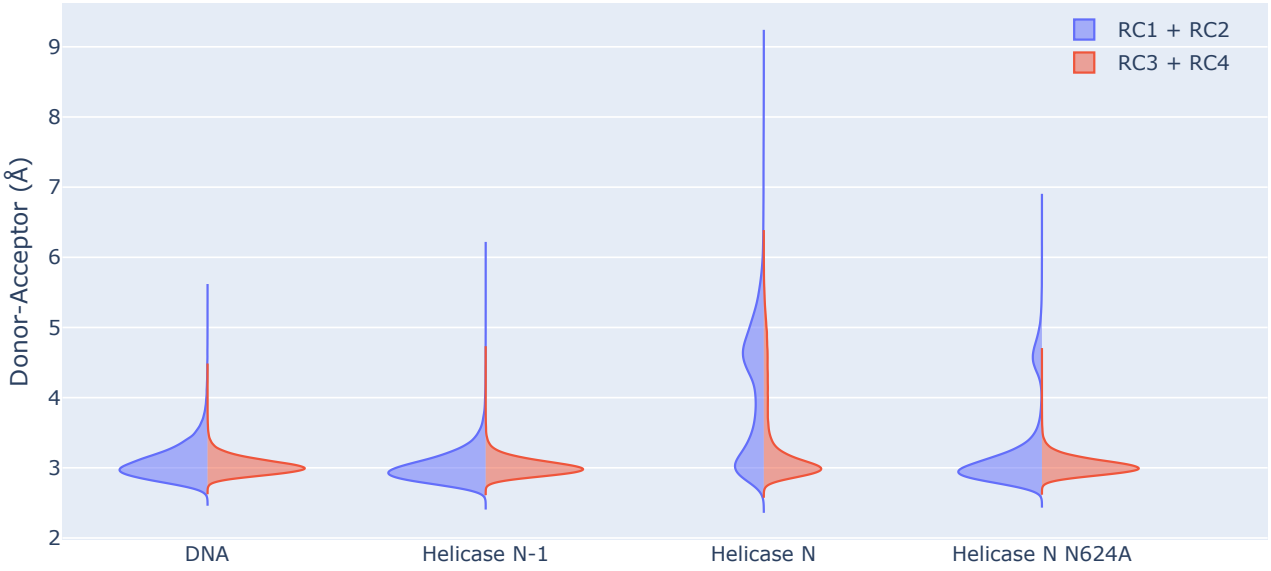


Fig. S18. Donor-acceptor distance statistics for the two hydrogen bonds involved in the DPT of GC in four environments. Each system was simulated for 33ns in NVT Molecular Dynamics.

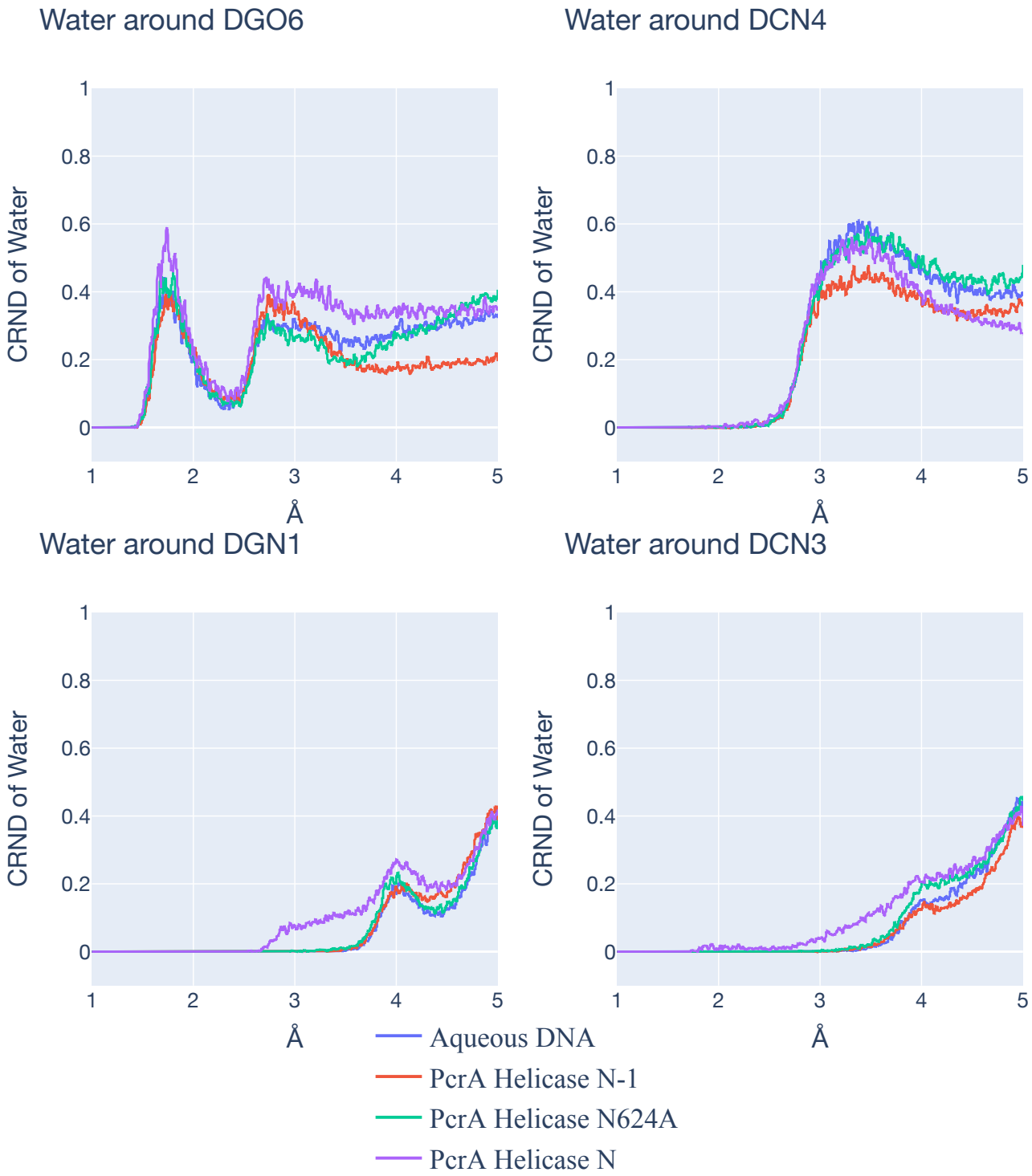


Fig. S19. Radial number density (RND) of water atoms around donor and acceptor atoms involved in the DPT of GC in four environments. Each system was simulated for 33ns in NVT Molecular Dynamics.

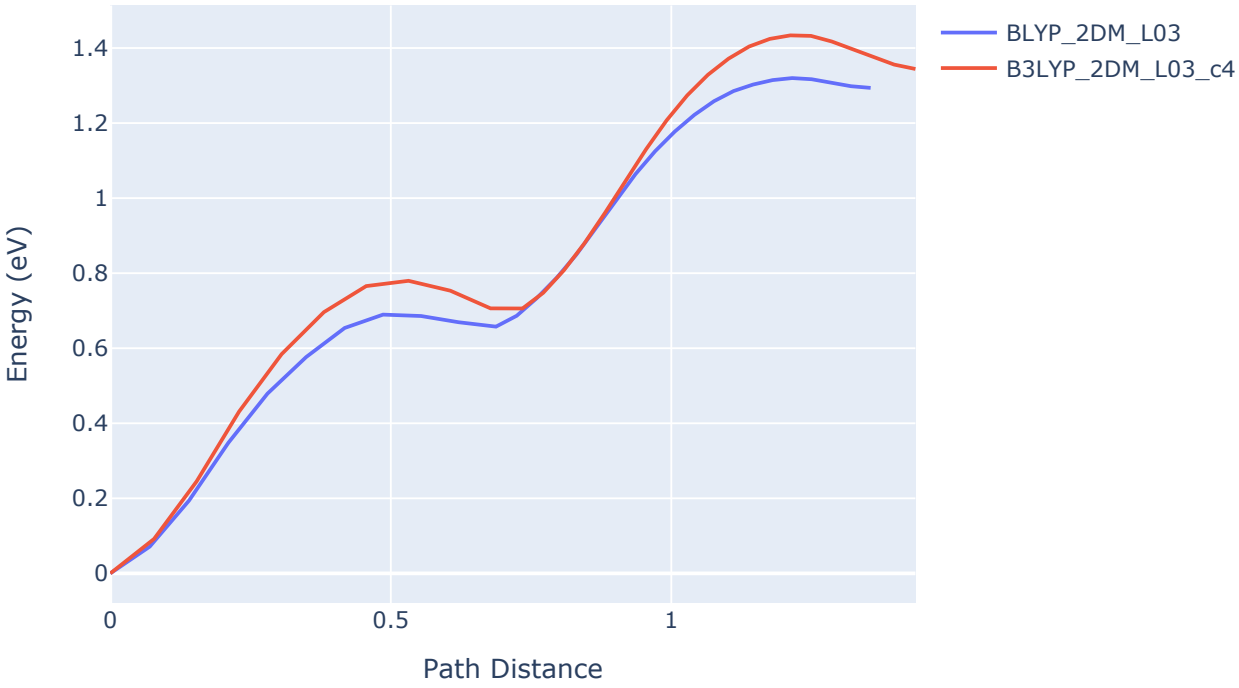


Fig. S20. NEB profiles through instantaneous potential energy surface for the DPT in the same aqueous DNA conformation for BLYP and B3LYP levels of theory.

References

[1] MD Hanwell, et al., Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. cheminformatics* **4**, 1–17 (2012).

[2] RB Best, et al., Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ1 and χ2 dihedral angles. *J. chemical theory computation* **8**, 3257–3273 (2012).

[3] J Huang, et al., Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nat. methods* **14**, 71–73 (2017).

[4] K Hart, et al., Optimization of the charmm additive force field for dna: Improved treatment of the bi/bii conformational equilibrium. *J. chemical theory computation* **8**, 348–362 (2012).

[5] H Berendsen, D van der Spoel, R van Drunen, Gromacs: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).

[6] M Winokan, C Vallee, Molparse: A python package for parsing, modifying, and analysis of molecular structure files. (https://github.com/mwinokan/MolParse) (2023).

[7] SS Velankar, P Soultanas, MS Dillingham, HS Subramanya, DB Wigley, Crystal structures of complexes of pcra dna helicase with a dna substrate indicate an inchworm mechanism. *Cell* **97**, 75–84 (1999).

[8] J Yu, T Ha, K Schulten, Structure-based model of the stepping motor of pcra helicase. *Biophys. journal* **91**, 2097–2114 (2006).

[9] J Hutter, M Iannuzzi, F Schiffmann, J VandeVondele, cp2k: atomistic simulations of condensed matter systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 15–25 (2014).

[10] L Slocombe, M Winokan, J Al-Khalili, M Sacchi, Proton transfer during dna strand separation as a source of mutagenic guanine-cytosine tautomers. *Commun. Chem.* **5**, 144 (2022).

[11] P Virtanen, et al., Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. methods* **17**, 261–272 (2020).

[12] AH Larsen, et al., The atomic simulation environment—a python library for working with atoms. *J. Physics: Condens. Matter* **29**, 273002 (2017).

[13] L Slocombe, JS Al-Khalili, M Sacchi, Quantum and classical effects in dna point mutations. *Phys. Chem. Chem. Phys.* (2021).

E. Supporting Information: Quantum Tunnelling Effects in the Guanine-Thymine Wobble Misincorporation via Tautomerisation

Supporting Information: Quantum Tunnelling Effects in the Guanine-Thymine Wobble Misincorporation via Tautomerisation

Louie Slocombe*
Leverhulme Quantum Biology Doctoral Training Centre, University of Surrey, Guildford, GU2 7XH, UK. and Department of Chemistry, University of Surrey, Guildford, GU2 7XH, UK.

Max Winokan†
Leverhulme Quantum Biology Doctoral Training Centre, University of Surrey, Guildford, GU2 7XH, UK.

Jim Al-Khalili‡
Department of Physics, University of Surrey, Guildford, GU2 7XH, UK.

Marco Sacchi§
Department of Chemistry, University of Surrey, Guildford, GU2 7XH, UK.
(Dated: December 16, 2022)

This document contains supplemental information for the results presented in the manuscript *Quantum Tunnelling Effects in the Guanine-Thymine Wobble Misincorporation via Tautomerisation*. In this file, unless otherwise stated, we use the Hartree atomic unit system, energy is in units of Hartrees E_h, the reduced Planck constant is $\hbar = 1$, and lengths are in Bohr radius a_0.

CONTENTS

* louie.slocombe@surrey.ac.uk
† m.winokan@surrey.ac.uk
‡ J.Al-Khalili@surrey.ac.uk
§ m.sacchi@surrey.ac.uk

SUPPLEMENTARY NOTE 1: DENSITY FUNCTIONAL THEORY METHODS

Obtaining the Reaction Pathway

We performed Density Functional Theory (DFT) calculations with NWChem 7.0.2 [1] at the B3LYP+D3/6-311++G** level of theory. We use the B3LYP exchange-correlation functional [2] with Grimme DFT-D3 dispersion corrections to capture empirical long-range contributions [3, 4]. Based on previous works, this combination of exchange-correlation functional and basis set offers comparable accuracy to Møller–Plesset perturbation theory of the second order while at a fraction of the computing cost [5]. We embed the DNA bases in an implicit continuum solvation model [6–8] describing the influence of the surrounding water molecules, where $\epsilon = 78.4$.

We obtained the potential energy landscapes describing the proton transfer reactions using a machine learning approach to the classical all-nudged elastic band algorithm (ML-NEB) [9, 10]. The ML-NEB approach minimises the number of DFT single-point energy calculations required to depict the minimum energy path (MEP) accurately. In our treatment, we collect the movement of the protons transferring (and other atoms moving to facilitate the transfer) into a single axis. The reaction pathway contains a general description of the transfer process; the energetic landscape of this pathway is then explored using ML-NEB. The ML-NEB algorithm incorporates a Gaussian regression model to produce a surrogate description of the accurate MEP. Thus the uncertainty in the energy points on surrogate MEP becomes the convergence criteria.

The atomic simulation environment (ASE) [11, 12] was used throughout this work to connect NWChem to Python3 and the ML-NEB algorithm. All the structures were optimised using a force tolerance of $0.01\,\mathrm{eV\,\AA^{-1}}$. We adopt the optimised monomeric forms from Ref. [13], which are combined with their target base to form the hydrogen-bonded pair. We introduce free energy contributions by conducting a vibrational analysis at three points along the reaction coordinate: reactant, TS, and product structures; we use the ideal-gas limit to account for the translational and rotational degrees of freedom.

Obtaining an Effective Mass of the System

During the potential energy surface calculations, all degrees of freedom can relax. Consequently, the reaction pathway contains the joint motion of several atoms to facilitate the proton transfer reaction. Using the minimum energy pathway, we construct a reaction coordinate linking the reactant to the product via a transition state. We project the motion of each atom on to reaction coordinate.

To account for the contribution of the masses of the atoms participating in this reaction coordinate, we determine the effective mass μ using [14–18]

$$\mu = \sum_{i=1}^{N} m_i \left(\frac{\partial r_i}{\partial q} \cdot \frac{\partial r_i}{\partial q} \right). \tag{1}$$

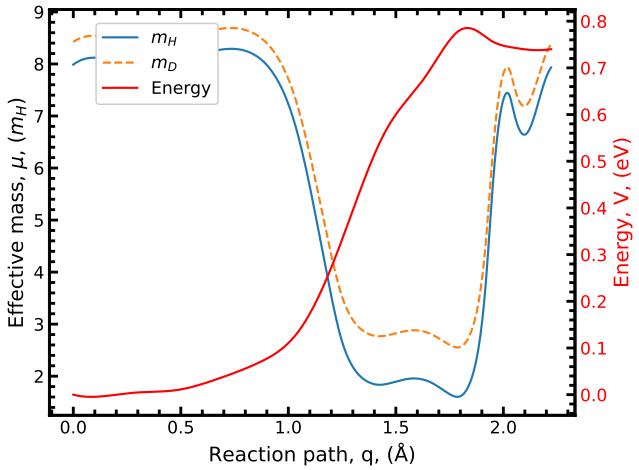
Here the total number of atoms N, and m_i denotes the mass of the atom with index i, $m_i = (m_1, m_1, m_1, m_2, \ldots)$ 3N-sized vector of atomic masses. Whereas r_i is the 3N-sized Cartesian vector describing the change in the coordinates of atoms, with the reaction coordinate defined as q. The partial derivative of the Cartesian vector tracks the motion of atom m_i along the reaction coordinate q. The dot product normalises the motion relative to the collective rearrangement of all atoms. If there is no motion of atom i, it does not contribute to the reaction path.

Conversely, when atom i transfers, the dot product becomes non-trivial. Then by summing over all atoms i, we can obtain an effective mass which contains the contribution from all degrees of freedom. To evaluate the derivatives, we first pass the Cartesian coordinate vectors into a Savitzky–Golay filter to suppress any spurious noise introduced by the uncertainty in the path, which is inherent to the machine-learning approach to finding the reaction path. The filtered Cartesian coordinate vector and the reaction path are then interpolated using cubic splines.

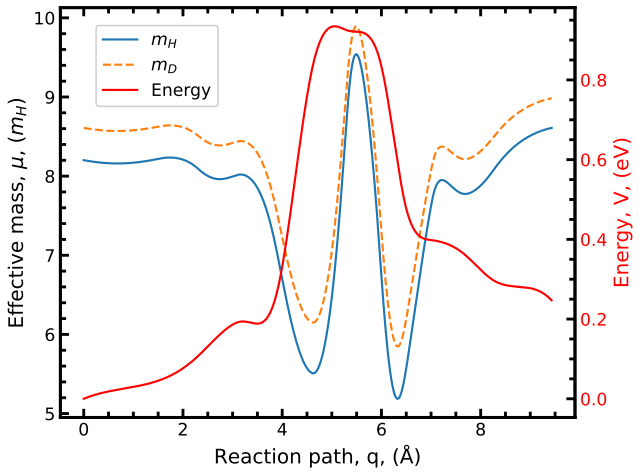
Consequently, we determine the effective mass to account for the contribution of the masses of the atoms participating in this reaction coordinate. The result is shown in Fig. 1.

Double-Well Potential Energy Surface Representations

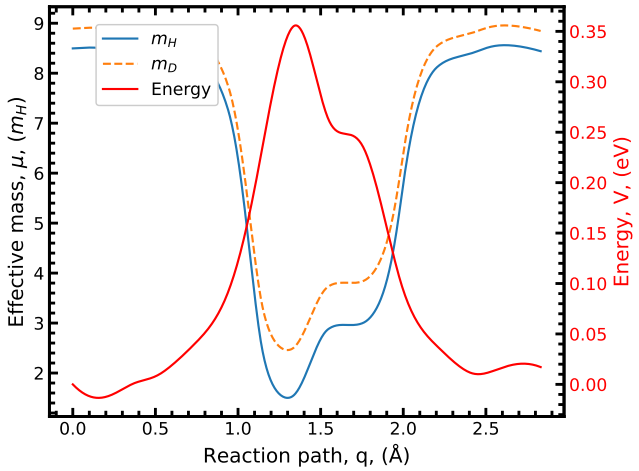
In the open quantum systems approach, we describe the proton transfer reactions using a pseudo-one-dimensional reaction coordinate connecting the reactant and product via a transition state barrier. To adopt the reaction profile



(a) wobble(G-T) ⇌ G*-T*, $m_p = 1.76m_H$ and $m_d = 2.69m_H$

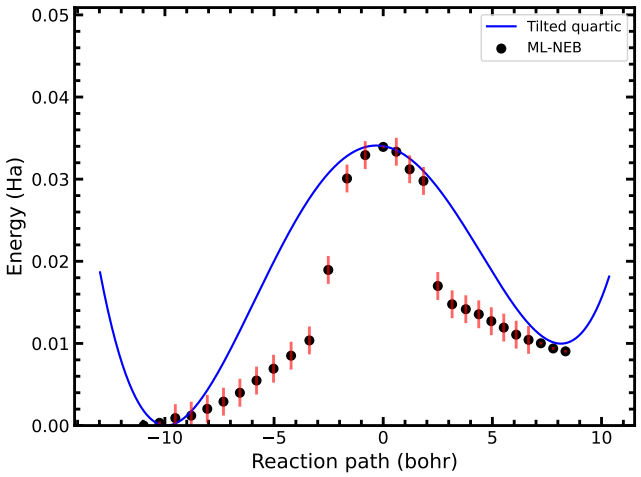


(b) wobble(G-T) ⇌ G-T*, $m_p = 6.77m_H$ and $m_d = 7.32m_H$.

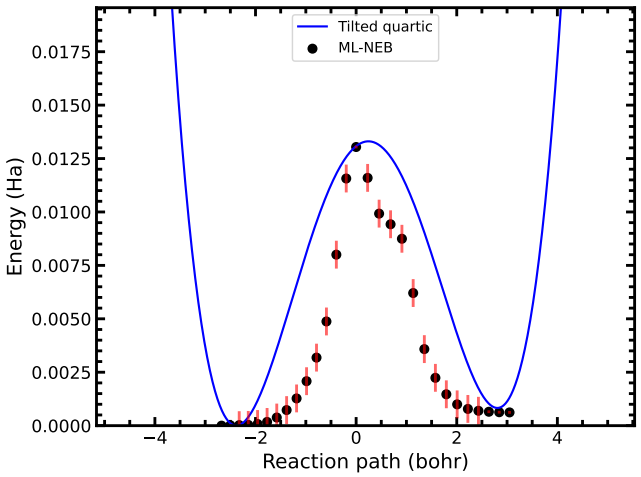


(c) G*-T ⇌ G-T*, $m_p = 1.59m_H$ and $m_d = 2.54m_H$

Supplementary Figure 1: Plots of the effective mass as a function of the reaction path.



(a) wobble(G-T) ⇌ G-T*



(b) G*-T ⇌ G-T*

Supplementary Figure 2: Fits of the proton transfer potential energy surfaces to the model tilted quartic potential.

into the open quantum systems Hamiltonian and to determine the reactive flux passing through the barrier, we describe the potential energy surface using a tilted quartic double-well model potential given by

$$V(q) = \frac{\hbar\omega_0}{2L_0^2}\, q^2 \left((q - q_0)^2 - \frac{L_0^2}{2} \right) + \frac{\Delta E}{L_0} q \tag{2}$$

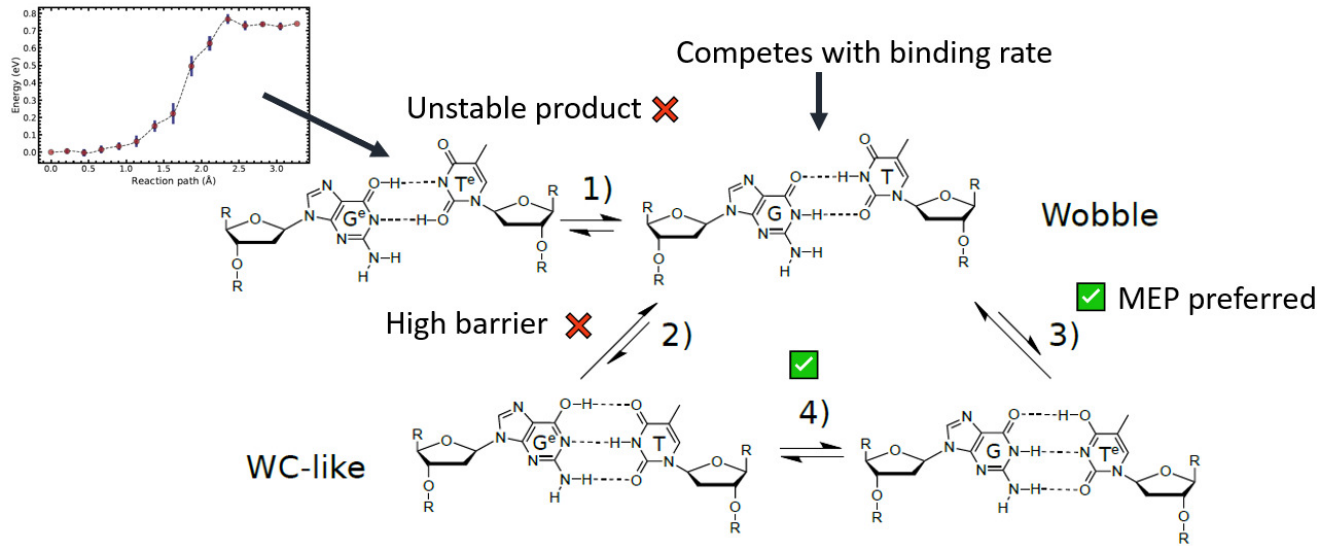
Where q is the position coordinate, ω_0 is the effective spring constant of the barrier, L_0 is the displacement between the well minima, q_0 is the additional tilt parameter, and ΔE is the energy difference between the well minima. Each ML-NEB reaction is fitted to this potential to form the following set of potentials shown in Table I. A constrained least-squares fit is performed to the NEB data points; constraining is required to capture the reaction asymmetry and barrier values correctly. This potential is inserted into our open quantum system Hamiltonian; the result is plotted in Fig. 2.

Tunnelling-Ready State: Assuming the Frozen Approximation

Only the inner barrier (section 2 of Fig. 2 in the manuscript) corresponds to the proton transfer between the bases. In contrast, regions 1 and 3 correspond to overall translations of the bases without significant changes in the

Supplementary Table I: Summary of the potential parameters used to describe the proton transfer reactions. The following parameters are defined: ω_0 spring constant of the barrier, L_0 is the displacement, q_0 is the additional tilt parameter, and ΔE well energy.

Parameter	wobble(G-T) ⇌ G-T*	G*-T ⇌ G-T*
ω_0	0.001 41 AUT	0.007 60 AUT
L_0	$12.76\,a_0$	$3.66\,a_0$
q_0	$-1.40\,a_0$	$0.45\,a_0$
ΔE	$-0.005\,56\,E_\mathrm{h}$	$0.006\,78\,E_\mathrm{h}$



Supplementary Figure 3: Schematic representation of the G-T wobble mispairs and the conversion to a Watson-Crick-like configuration via a proton transfer process.

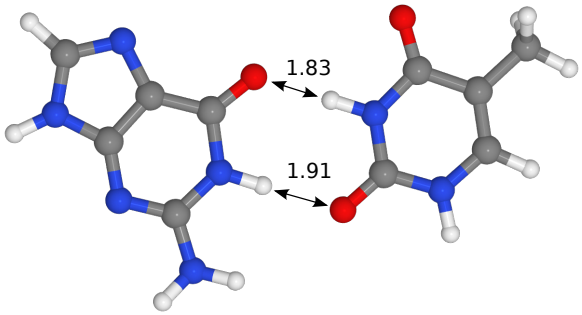
hydrogen bond length. This observation is compatible with a so-called "tunnelling-ready state" along the reaction path, whereby the hydrogen bonds become partly compressed. To model the state, we take the image corresponding to the start of the proton transfer barrier from the wobble to the Watson-Crick pathway and assume that the local environment has induced this conformation change. Then at this point, the proton transfer timescale would be much quicker due to its low mass than the whole base translating. So at this point, the proton could transfer, and the rest of the base would recoil. The fast proton transfer could facilitate the required rearrangement for the base to snap into a Watson-Crick-like shape. We take the image along the reaction path, constrain all but the hydrogen atom, and calculate the minimum energy path the hydrogen would trace out while transferring across. During the transfer, we neglect rearrangements from the rest of the atoms. The result is a reaction path that may not be the minimum energy pathway but has a higher overall rate due to the boost from tunnelling contributions. The first minimum has an energy difference of 0.621 eV compared to the ground state G-T wobble structure. Assuming this configuration is explored with a penalty in the form $\exp(-E/k_BT)$ gives a weighting of $\sim 3.18 \times 10^{-11}$.

Summary of the Proton Transfer Energy Landscapes

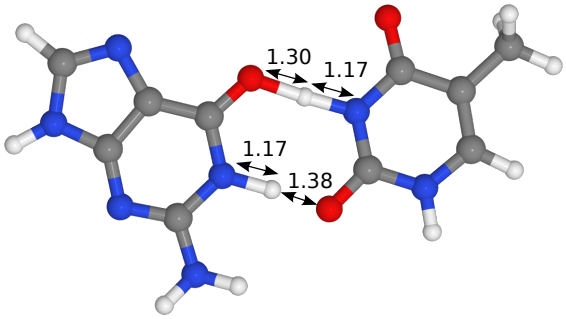
We utilise the methods described before to determine the reaction pathway of the proton transfer in the formation of wobble mismatches. An overview of the reaction pathways is shown in Fig. 3. The results are summarised in table II, and figs. 4 5 6.

Supplementary Table II: Summary of the reactions. With, forward reaction barrier E_f, reverse reaction barrier E_r, reaction asymmetry ΔE, forward reaction free-energy barrier G_f, reverse reaction free-energy barrier G_r, reaction free-energy asymmetry ΔG, and the imaginary frequency at the transition state ν_i.

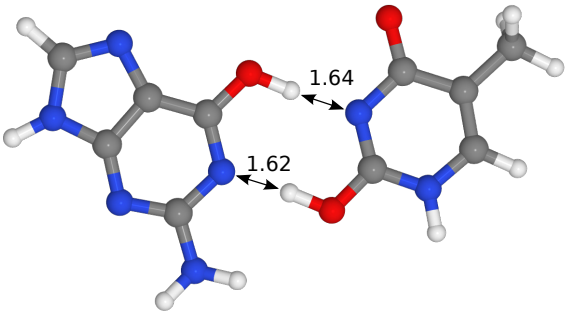
Parameter	wobble(G-T) $\rightleftharpoons$ wobble(G*-T*)	wobble(G-T) $\rightleftharpoons$ G-T*	G*-T $\rightleftharpoons$ G-T*
E_f	0.766 eV	0.926 eV	0.356 eV
E_r	0.026 eV	0.680 eV	0.339 eV
ΔE	0.740 eV	0.246 eV	0.017 eV
G_f	0.497	0.774 eV	0.075 eV
G_r	-0.191	0.465 eV	0.107 eV
ΔG	0.688	0.309 eV	0.032 eV
ν_i	1022.3 cm^{-1}	156.3 cm^{-1}	1172.9 cm^{-1}



(a) wobble form of G-T

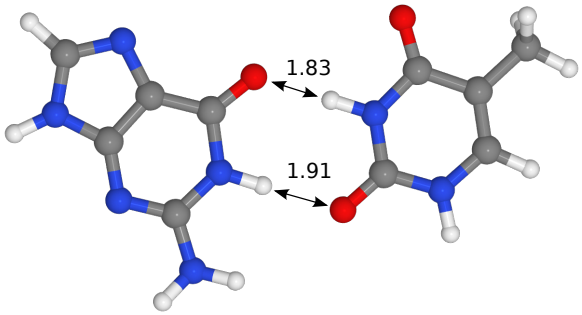


(b) Proton transfer transition state

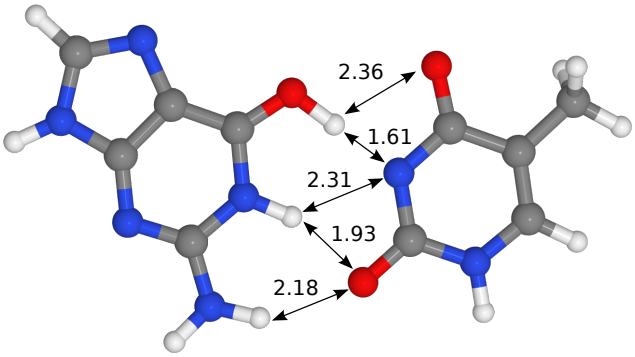


(c) wobble form of G*-T*

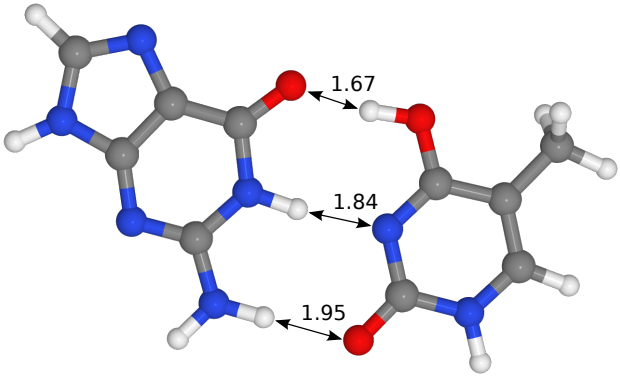
Supplementary Figure 4: Optimised geometries of the proton transfer reaction from the canonical wobble forms to the tautomeric wobble configuration, wobble(G-T) $\rightleftharpoons$ wobble(G*-T*). Here the bases do not translate to facilitate proton transfer but instead transfer along the hydrogen bonds, forming a double proton transfer product.



(a) wobble form of G-T

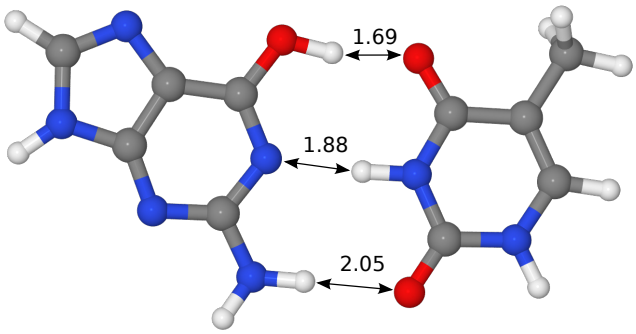


(b) Proton transfer transition state

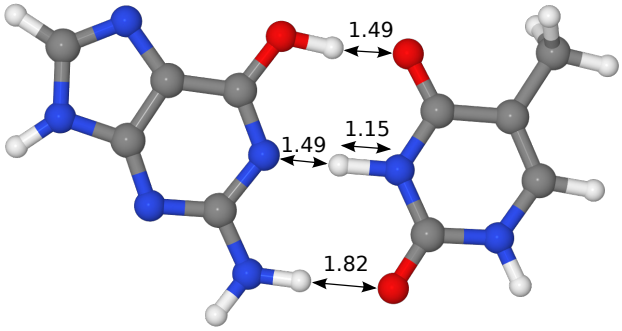


(c) Watson-Crick-like G-T*

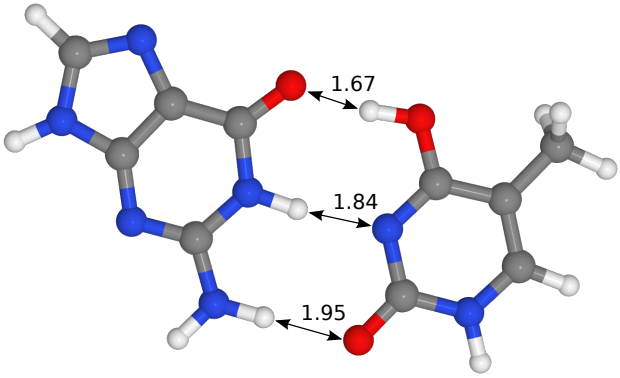
Supplementary Figure 5: Optimised geometries of the proton transfer reaction from the canonical wobble forms to the tautomeric Watson-Crick-like configuration, wobble(G-T) ⇌ G-T*. Here the bases translate relative to each other to facilitate the proton transfer.



(a) Watson-Crick-like G*-T



(b) Transition state



(c) Watson-Crick-like G-T*

Supplementary Figure 6: Optimised geometries of the proton transfer reaction from the Watson-Crick-like forms, G*-T $\rightleftharpoons$ G-T*. Here double proton transfer along the hydrogen bonds swaps the enol form between the two bases.

SUPPLEMENTARY NOTE 2: USING OPEN QUANTUM SYSTEMS TO DESCRIBE TUNNELLING

An ideally isolated quantum system, particularly in biology, is unlikely. Instead, the environment is constantly interacting with the system. In the cellular environment, there is a constant energy flow between the system and the environment through vibrations and collisions with the surrounding solvent and proteins, constantly perturbing the quantum system. Once decoherence sets in, we might expect entirely classical behaviour to emerge. To describe this transition region, we require a theoretical framework to describe the protons in DNA using an open quantum systems approach.

The idea of an open quantum system is to incorporate interactions with the local environment. These interactions significantly change the system's dynamics and result in quantum dissipation and decoherence. The general idea is to couple a system Hamiltonian $\hat{H}_{\rm S}$ with a bath $\hat{H}_{\rm B}$ via an interaction $\hat{H}_{\rm I}$,

$$\hat{H}_{\rm SB} = \hat{H}_{\rm S} + \hat{H}_{\rm B} + \hat{H}_{\rm I}. \tag{3}$$

Here the interaction term generates quantum and classical correlations between the system and the environment [19].

The Wigner-Moyal Caldeira-Leggett Model

Wigner introduced a quantum-mechanical distribution function in phase space by considering the integral transform of the wave function [20]

$$W(q,p,t) = \frac{1}{\pi\hbar}\int \psi^*(q+q')\psi(q-q')e^{2ipq'/\hbar}dq', \tag{4}$$

or, for a mixed quantum state, using a density matrix,

$$W(q,p,t) = \frac{1}{2\pi\hbar}\int e^{-ipq'/\hbar}\left\langle q+\frac{q'}{2}\right|\tilde{\rho}\left|q-\frac{q'}{2}\right\rangle dq'. \tag{5}$$

The Wigner transformation is the Fourier transform of the antidiagonals of the density matrix when that matrix is expressed in the position basis. Taking the partial time derivative and employing the time-dependent Schrödinger equation, we obtain the Wigner-Moyal (WM) equation describing the time evolution of the Wigner function. Explicitly:

$$\frac{\partial}{\partial t}W(q,p,t) = \mathcal{L}(q,p)\,W(q,p,t). \tag{6}$$

The WM equation is a deterministic dynamical equation that encapsulates the uncertainty in position q and momentum p into a quasi-probability density $W\,(q,p,t)$ [20, 21]. The quantum Liouvillian ($\mathcal{L}$) for the Wigner function is given by the kinetic ($\mathcal{K}$) and potential ($\mathcal{V}$) terms

$$\mathcal{L}(q,p) \equiv \mathcal{K}(q,p) + \mathcal{V}(q,p). \tag{7}$$

with terms,

$$\mathcal{K}(q,p)W(q,p) \equiv -\frac{p}{m}\frac{\partial}{\partial q}W(q,p) \tag{8}$$

and

$$\mathcal{V}(q,p)W(q,p) \equiv -\frac{i}{\hbar}\left(\mathcal{V}(q)\star W(p,q) - W(p,q)\star\mathcal{V}(q)\right). \tag{9}$$

Here, we use the star operator, $\star$, corresponding to the Moyal product [22, 23],

$$\star \equiv \exp\left[\frac{i}{2}\left(\underset{\leftarrow_q}{\partial}\underset{\rightarrow_p}{\partial} - \underset{\rightarrow_q}{\partial}\underset{\leftarrow_p}{\partial}\right)\right]. \tag{10}$$

The directional differentiation operators from the left and right appearing here are defined as

$$\underset{\rightarrow_x}{\partial} f(x) = f(x)\underset{\leftarrow_x}{\partial} \equiv \frac{\partial f(x)}{\partial x}. \tag{11}$$

Consequently,

$$\mathcal{V}(q,p)W(q,p) \equiv \frac{\partial V}{\partial q}\frac{\partial W}{\partial p} + \sum_{r=1}^{\infty} \frac{(\mathrm{i}\hbar/2)^{2r}}{(2r+1)!}\frac{\partial^{2r+1}V}{\partial q^{2r+1}}\frac{\partial^{2r+1}W}{\partial p^{2r+1}}. \tag{12}$$

For brevity, we drop the dependence and expand the potential terms,

$$\frac{\partial W}{\partial t} = -\frac{p}{m}\frac{\partial W}{\partial q} + \frac{\partial V}{\partial q}\frac{\partial W}{\partial p} - \frac{\hbar^2}{24}\frac{\partial^3 V}{\partial q^3}\frac{\partial^3 W}{\partial p^3} + \mathcal{O}\left(\hbar^4\right). \tag{13}$$

Here the potential terms have been expanded as the Taylor series truncated to just the first two terms, which is frequently done in literature [21]. The series expansion contains powers of $\hbar$ and introduces quantum effects into the dynamics. The equivalent phase-space formulation of Caldeira-Leggett's model, also known as the Wigner-Moyal Caldeira-Leggett (WM-CL) equation, is written as [20, 24]

$$\frac{\partial W}{\partial t} = \underbrace{-\frac{p}{m}\frac{\partial W}{\partial q} + \frac{\partial V}{\partial q}\frac{\partial W}{\partial p} - \frac{\hbar^2}{24}\frac{\partial^3 V}{\partial q^3}\frac{\partial^3 W}{\partial p^3} + \mathcal{O}\left(\hbar^4\right)}_{\text{Schrödinger dynamics}} \underbrace{+\gamma\frac{\partial pW}{\partial p}}_{\text{Dissipation}} \underbrace{+\gamma m k_{\mathrm{B}} T\frac{\partial^2 W}{\partial p^2}}_{\text{Decoherence}}. \tag{14}$$

The WM-CL equation has a similar form to the WM equation [25]; however, it now contains two additional terms describing dissipation and decoherence arising from the coupling to the quantum bath.

Low-temperature Correction

The WM-CL equation is valid in the weak coupling and high-temperature regime $k_{\mathrm{B}}T \gg E_0$, where E_0 is the zero-point energy of the uncoupled system [24]. However, for this system and biologically relevant temperatures ($T \approx 300\,\mathrm{K}$), the high-temperature limit can no longer be a valid approximation. Consequently, we adopt a temperature correction [26–30],

$$T \to \tilde{T} = \frac{\hbar\Omega}{2k_{\mathrm{B}}}\coth\left(\frac{\hbar\Omega}{2k_{\mathrm{B}}T}\right). \tag{15}$$

The correction modifies the temperature of the bath such that it saturates to the zero-point energy of the system at a low temperature, indicating that the lowest energy the system can take is the zero-point energy of the system. Without the correction, the CL model breaks down and fails to maintain the generalised position-momentum uncertainty principle.

$$\lim_{\tilde{T}\to 0^+}\coth\left(\frac{\hbar\Omega}{2k_{\mathrm{B}}T}\right) \to 1. \tag{16}$$

While in the high-temperature limit, it is simply the leading term in the Taylor expansion of $\coth x = \frac{1}{x} + \frac{x}{3} - \frac{x^3}{45} + \cdots$ [24, 26–30],

$$k_{\mathrm{B}}\tilde{T} = \frac{\hbar\Omega}{2}\coth\left(\frac{\hbar\Omega}{2k_{\mathrm{B}}T}\right) \sim k_{\mathrm{B}}T. \tag{17}$$

The addition of the coth term permits using the WM-CL equation at lower temperatures. Within the harmonic approximation, assuming that at the potential global minimum, the third-order or high terms vanish Ω can be approximated by inspecting the second derivative of the potential,

$$\Omega_{\text{approx.}} = \sqrt{m^{-1}\frac{\partial^2 V\left(q_{\text{min.}}\right)}{\partial q^2}} \tag{18}$$

Where $q_{\text{min.}}$ is the location of the global minimum of the potential. However, the correction limits the system to be in potentials with a low anharmonicity [30]. Note that the original coth term in the influence functional contains the ω of the bath, whereas here, we replace it with a property of the potential.

Numerically Solving

To solve the WM-CL (Eq. 14) equation, we use the Smoluchowski limit. In the Smoluchowski (over-damped) limit, when the bath coupling is much greater than the spring constant at the global minimum ($\gamma >> \Omega$), our system becomes over-damped. Assuming that the oscillations in the bath are much faster than the system dynamics, we approach the Smoluchowski limit ($\gamma = 3900\,\mathrm{cm}^{-1} >> \Omega$). In this limit, the bath induces the separation of timescales between the evolution of position and momentum. We can then take

$$P^{\mathrm{QSE}}(q,t) \equiv \int W(p,q,t)\,dp. \tag{19}$$

In the Smoluchowski limit, the Eq. 14 can be rewritten as [31]

$$\frac{\partial}{\partial t}P^{\mathrm{QSE}}(q,t) = \frac{1}{m\gamma}\frac{\partial}{\partial q}\left[\frac{\partial V}{\partial q} + k_{\mathrm{B}}\tilde{T}\frac{\partial}{\partial q}\right]P^{\mathrm{QSE}}(q,t). \tag{20}$$

Eq. 20 can be solved with the method of lines approach. q and p are discretised to a fixed equally spaced lattice with N_q points in range $[q_{\min}, q_{\max}]$ and N_p equally spaced points in range $[p_{\min}, p_{\max}]$. The partial derivatives are expanded using a second-order central finite difference approach. The outer coefficients are set to zero, corresponding to Dirichlet (reflecting) boundary conditions. We use Feagin's 14 explicit Runge-Kutta algorithms to solve for time [32].

Alternatively, to solve the WM-CL (Eq. 14) equation using the method of lines approach. The partial derivatives are expanded using a second-order central finite difference approach. The outer coefficients are set to zero, corresponding to Dirichlet (reflecting) boundary conditions. To integrate the equations in time, we utilise the VCABM5 algorithm [33], an adaptive fifth-order Adams-Moulton method implemented in the DifferentialEquations.jl ecosystem [32]. We find that VCABM5 offers a good trade-off between accuracy and speed.

Obtaining a Quantum Corrected Reaction Rate

The full forward and reverse reaction rate constants, k_{f} and k_{r}, are obtained from,

$$k_{\mathrm{f,r}} = \frac{\kappa}{h\beta}e^{-G_{\mathrm{f,r}}\,\beta} \tag{21}$$

Where $G_{\mathrm{f,r}}$ corresponds to the Gibbs free energy barrier of the forward and reverse reaction barrier, respectively, the tunnelling factor, κ, encapsulates the quantum-to-classical contribution to the rate, incorporating quantum effects such as tunnelling and non-classical reflections thus, if the quantum contribution is negligible $\kappa \to 1$. On the other hand, if quantum effects dominate $\kappa \gg 1$.

We define κ using,

$$\kappa(T) = 1 + \frac{k_f^{\mathrm{QM}}}{k_f^{\mathrm{CL}}} \tag{22}$$

The classical transition state theory value is determined via

$$k_f^{\mathrm{CL}} = \frac{1}{\beta h}\frac{Q^{\neq}}{Q_R}e^{-\beta E_f} \approx \frac{\omega_0}{2\pi}e^{-\beta E_f}. \tag{23}$$

Where $Q^{\neq}$ and Q_R are the reactant and transition state partition functions, and ω_0 is the harmonic constant at the bottom of the reactant well. The second half of the equation follows previous studies which employed a harmonic approximation to the partition functions [34–37]. Finally, k_{CL} does not depend on the bath modes or the friction. Therefore, the classical rates are independent of the friction value.

At thermodynamic equilibrium, in a canonical ensemble, the populations of the reactant and the product regions have the following stationary values

$$P_r^{\mathrm{eq}} = \frac{1}{Q}\int\int W^{\mathrm{eq}}\,(1-\hat{h}(q))\,dp\,dq \tag{24}$$

$$P_p^{\mathrm{eq}} = \frac{1}{Q}\int\int W^{\mathrm{eq}}\,\hat{h}(q)\,dp\,dq. \tag{25}$$

Where $Q = \int\int W^{\mathrm{eq}}\, dp\, dq$ partition function for the overall system. $\hat{h}(q)$ is a Heaviside step function that projects onto the product side of a transition state dividing surface (reaction barrier). At thermal equilibrium, $W^{\mathrm{eq}}(q,p) = W(q,p,t\to\infty)$, and we can integrate Eq. 14 until it comes to a stationary solution.

In chemical kinetics, this equilibrium can be viewed as dynamically reached in the long-time limit when starting from a non-stationary initial state. If both the forward and backward reactions are governed by rate processes, a relatively simple kinetic equation can be used to describe the population of the reactant (P_r) or the product (P_p)

$$\frac{d}{dt}P_r(t) = -\frac{d}{dt}P_p(t) = -k_f^{\mathrm{QM}}(T)P_r(t) + k_r^{\mathrm{QM}}(T)P_p(t). \tag{26}$$

Where the thermal rate constants for the forward (k_f^{QM}) and reverse (k_f^{QM}) reactive processes. Consequently, we have a detailed balance requirement

$$\frac{k_f^{\mathrm{QM}}(T)}{k_r^{\mathrm{QM}}(T)} = \frac{Q_p(T)}{Q_r(T)}. \tag{27}$$

Where Q_r and Q_p are defined as the reactant and product partition functions

$$Q_r(T) = \int\int W^{\mathrm{eq}}\,(1-\hat{h}(q))\, dp\, dq \tag{28}$$

$$Q_p(T) = \int\int W^{\mathrm{eq}}\,\hat{h}(q)\, dp\, dq. \tag{29}$$

We determine the quantum contribution to the chemical reaction rate by monitoring the flux of the density passing through the transition state barrier. We start the system with a non-stationary initial distribution that models the system at thermal equilibrium in the reactant well

$$W(q,p,t=0) = \frac{1}{\mathcal{N}}e^{-\mathcal{H}\tilde{\beta}}(1-\hat{h}(q)) \tag{30}$$

with $\mathcal{H} = p^2/(2m)+V(q)$ and normalisation constant $\mathcal{N}$. Thus, we monitor the flux of the probability density changes between the left and right-hand well [38–41]

$$\tilde{k}_f^{\mathrm{QM}}(t,T) = -\frac{\dot{P}_r(t)}{P_r(t) - [Q_r(T)/Q_p(T)]\,[1-P_r(t)]} \tag{31}$$

We require that a time-scale separation exists between the reaction and other dynamical processes in the system. After some characteristic time, τ_c [42] the phenomenological rate law can be adopted since the rate plateaus and becomes time-independent

$$k_f^{\mathrm{QM}}(T) = \lim_{t\to\tau_c} \tilde{k}_f^{\mathrm{QM}}(t,T). \tag{32}$$

The equilibrium constant can be calculated using

$$K_{\mathrm{eq}} = \frac{k_f}{k_r} = \exp\left(-\frac{\Delta G}{k_{\mathrm{B}}T}\right). \tag{33}$$

Kinetic Isotope Effect

A strong dependence of the reaction rate on the reduced mass of the system could suggest the involvement of tunnelling [43–45]. Consequently, we determine the kinetic isotope effect (KIE) using

$$\mathrm{KIE} = \frac{k_{\mathrm{f,p}}^{\mathrm{QM}}}{k_{\mathrm{f,d}}^{\mathrm{QM}}}, \tag{34}$$

where $k_{\mathrm{f,p}}^{\mathrm{QM}}$ ($k_{\mathrm{f,d}}^{\mathrm{QM}}$) is the forward rate for a proton (deuteron), obtained from applying Eq. 32.

Supplementary Table III: Summary of the quantum and classical contributions to the reactions. With terms, forward reaction rate k_f, reverse reaction barrier k_r, reactant lifetime τ_f, product lifetime τ_r, chemical equilibrium value K_{eq}, quantum vs classical rate contribution κ, KIE (kinetic isotope effect).

Parameter	wobble(G-T) ⇌ G-T*	TRS wobble(G-T) ⇌ G-T*	G*-T ⇌ G-T*
k_f	$5.244 \times 10^{-1}\,\text{s}^{-1}$	-	$6.090 \times 10^{12}\,\text{s}^{-1}$
k_r	$8.767 \times 10^{4}\,\text{s}^{-1}$	-	$1.753 \times 10^{12}\,\text{s}^{-1}$
τ_f	$1.907\,\text{s}$	-	$1.642 \times 10^{-13}\,\text{s}$
τ_r	$1.141 \times 10^{-5}\,\text{s}$	-	$5.706 \times 10^{-13}\,\text{s}$
K_{eq}	5.982×10^{-6}	-	3.475
κ	1.02	99.00	18.10
KIE	1.10	10.15	4.25

Summary of the Quantum Tunnelling Effects in Proton Transfer

We explore to what degree tunnelling plays a role in each reaction. First, we calculate tunnelling rates on the DFT calculated potentials after fitting them with an analytical function; the supplementary information contains a complete description of the parameters. We then insert this potential into the system Hamiltonian to obtain a tunnelling correction. The results are shown in table III.

Supplementary Table IV: Summary of the potential parameters used to describe the proton transfer reactions. The following parameters are defined: ω_0 spring constant of the barrier, L_0 is the displacement, q_0 is the additional tilt parameter, and ΔE well energy.

Parameter	wobble(G-T) ⇌ G-T*			G*-T ⇌ G-T*		
	Aqueous	B-DNA	Poly-λ	Aqueous	B-DNA	Poly-λ
ω_0	0.001 41 AUT	0.001 41 AUT	0.001 41 AUT	0.001 41 AUT	0.001 41 AUT	0.001 41 AUT
L_0	$12.76\,a_0$	$12.76\,a_0$	$12.76\,a_0$	$12.76\,a_0$	$12.76\,a_0$	$12.76\,a_0$
q_0	$12.76\,a_0$	$12.76\,a_0$	$12.76\,a_0$	$12.76\,a_0$	$12.76\,a_0$	$12.76\,a_0$
ΔE	$-0.005\,56\,E_\mathrm{h}$	$-0.005\,56\,E_\mathrm{h}$	$-0.005\,56\,E_\mathrm{h}$	$-0.005\,56\,E_\mathrm{h}$	$-0.005\,56\,E_\mathrm{h}$	$-0.005\,56\,E_\mathrm{h}$

SUPPLEMENTARY NOTE 3: COMPARING THE EFFECT OF THE ENVIRONMENT

Extracting the Free Energy Pathway

To compare how the environment has an impact on the tunnelling, we extract the free energy pathway data from Li *et al.*[46] using WebPlotDigitizer - a web-based tool to extract numerical data from plots [47]. We then have to scale the free energy pathway so that the energy is a function of the reaction path instead of the image index. To do this, we assume that each reaction follows the same path as the ML-NEB data we report. Consequently, we perform a linear rescaling to map the free energy pathway onto our data.

Next, using the extracted free energy pathway, which is now a function of the reaction path, we perform a constrained least-squares fit to adopt the reaction profile into the open quantum systems Hamiltonian. See supplementary note 1 for further details. The result of the fit is summarised in table IV.

Comparing the Free Energy Pathway

Fig. 7 compares our ML-NEB data of the wobble(G-T)⇌G-T* (panel a) and G*-T⇌G-T* (panel b) reaction with the free energy curve from Li *et al.*[46].

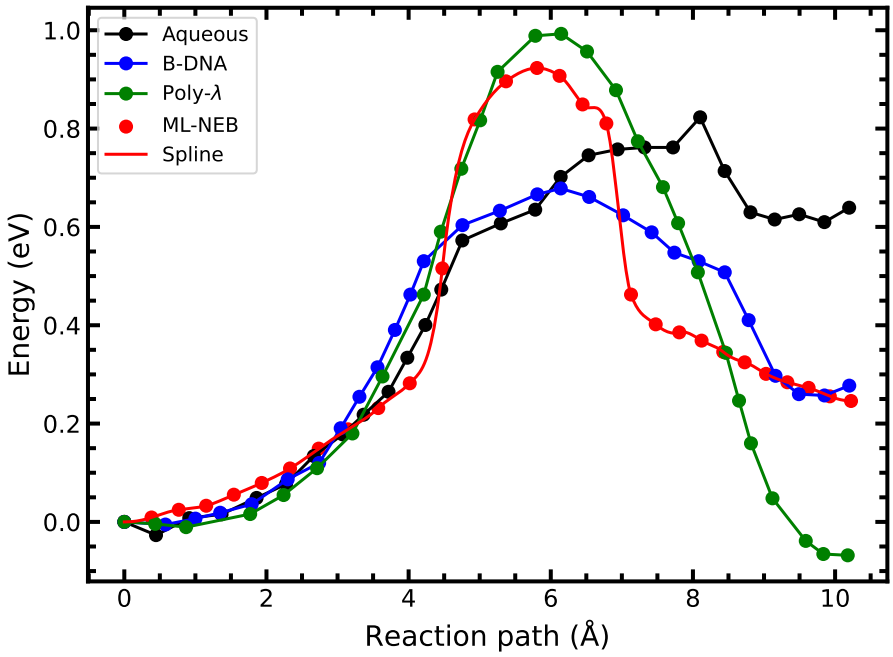
For panel a), the wobble reaction, all reaction paths have a similar initial energy trend, corresponding to the classical sliding and compression of the G-T wobble to facilitate the proton transfer; this mechanism is described in Fig. 2 of the main document. Furthermore, there is little to no variation between the environmental systems suggesting that the initial path is similar irrespective of the local environment. However, our barrier is slightly smaller than the one of the DNA/polymerase system but larger than the isolated B-DNA and DNA/aqueous solution barriers. While our free energy corrected barrier shown in table II is within 14% of the B-DNA. Similarly, the ML-NEB reaction energy matches the B-DNA system within 16%. However, the barrier is significantly higher for the polymerase system, but the reaction energy is much lower. On the other hand, for Fig. 7b), in the Watson-Crick to Watson-Crick reaction, our energy is larger than the free energy profiles. However, with our free energy profiles, our barrier is significantly reduced. Consequently, overall our energy profiles are within reasonable agreement with Li *et al.*[46] B-DNA system.

Comparing Environmental Effects on the Quantum Tunnelling

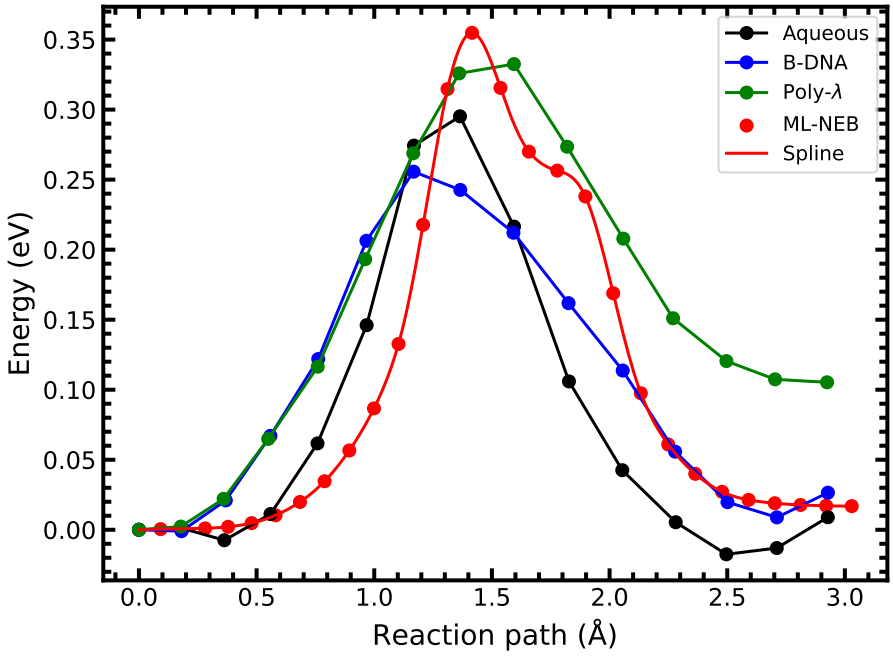
Finally, we use the extracted free energy potentials to determine the rates due to classical over-the-barrier hopping and tunnelling. The results are summarised in table V. Here, we explore the classical and quantum rates and all the previously calculated parameters.

Due to the wide barrier, we find an insignificant amount of tunnelling for the wobble mechanism for all environments. This finding is consistent with our previous finding regarding our ML-NEB potential. Furthermore, as there is little tunnelling, the KIE is also low, again indicating that the reaction is isotopic independent and predominantly classical. Thus if we adopt the Li *et al.* models and ignore the frozen approximation, as we detailed in supplementary note 1, we conclude that there is little dependence on the choice of environments on the tunnelling and instead, the proton transfer is an over-the-barrier classical behaviour. Here we note that the classical B-DNA rate is consistent with the NMR data [48–50].

On the other hand, the reaction weakly depends on the local environment for the Watson-Crick to Watson-Crick reaction, varying from 2.41 to 4.84. Overall, the quantum-to-classical ratio increases in the polymerase compared to



(a) wobble(G-T)⇌G-T*



(b) G*-T⇌G-T*

Supplementary Figure 7: Comparison of the Minimum energy paths of wobble(G-T)⇌G-T* and G*-T⇌G-T* reactions. Here, the aqueous, B-DNA, and poly-λ are taken from Li *et al.*[46]. In red is the data obtained using a machine-learning approach to the nudged elastic band method.

Supplementary Table V: Summary of the quantum and classical contributions to the reactions. With terms, forward reaction rate k_f, reverse reaction barrier k_r, reactant lifetime τ_f, product lifetime τ_r, chemical equilibrium value K_{eq}, quantum vs classical rate contribution κ, KIE (kinetic isotope effect).

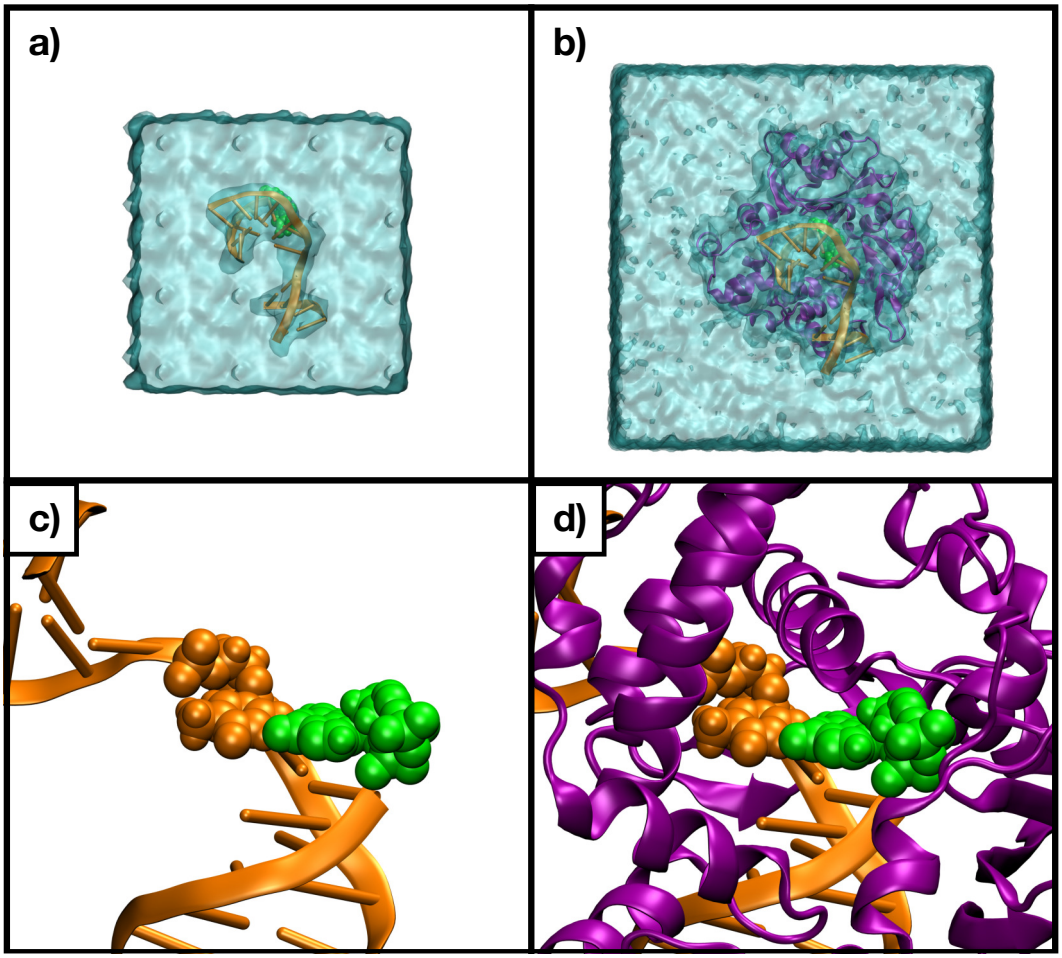
Parameter	wobble(G-T) ⇌ G-T*			G*-T ⇌ G-T*		
	Aqueous	B-DNA	Poly-λ	Aqueous	B-DNA	Poly-λ
k_f	$1.775 \times 10^{-1}\,\text{s}^{-1}$	$8.911\,\text{s}^{-1}$	$3.212 \times 10^{-6}\,\text{s}^{-1}$	$2.493 \times 10^{8}\,\text{s}^{-1}$	$5.959 \times 10^{8}\,\text{s}^{-1}$	$7.494 \times 10^{7}\,\text{s}^{-1}$
k_r	$2.274 \times 10^{10}\,\text{s}^{-1}$	$6.204 \times 10^{5}\,\text{s}^{-1}$	$1.612 \times 10^{-6}\,\text{s}^{-1}$	$2.265 \times 10^{8}\,\text{s}^{-1}$	$7.568 \times 10^{8}\,\text{s}^{-1}$	$4.215 \times 10^{9}\,\text{s}^{-1}$
τ_f	$5.635\,\text{s}^{-1}$	$1.122 \times 10^{-1}\,\text{s}^{-1}$	$3.113 \times 10^{5}\,\text{s}^{-1}$	$4.010 \times 10^{-9}\,\text{s}^{-1}$	$1.678 \times 10^{-9}\,\text{s}^{-1}$	$1.334 \times 10^{-8}\,\text{s}^{-1}$
τ_r	$4.397 \times 10^{-11}\,\text{s}^{-1}$	$1.612 \times 10^{-6}\,\text{s}^{-1}$	$6.205 \times 10^{5}\,\text{s}^{-1}$	$4.415 \times 10^{-9}\,\text{s}^{-1}$	$1.321 \times 10^{-9}\,\text{s}^{-1}$	$2.373 \times 10^{-10}\,\text{s}^{-1}$
K_{eq}	7.804×10^{-12}	1.436×10^{-5}	1.993	1.101	7.874×10^{-1}	1.778×10^{-2}
κ	1.01	1.01	1.03	3.85	2.41	4.84
KIE	1.0	1.0	1.0	2.19	1.67	2.51

the aqueous system due to the classical rate dropping quicker than the quantum rate, as the polymerase system has a higher barrier but a similar width.

SUPPLEMENTARY NOTE 4: QM/MM CALCULATIONS

Ensemble Molecular Dynamics

Classical dynamical simulations were performed in Gromacs 2021.1[51]. A modified input structure was obtained from [46] originating from the experimental crystal structure in PDB entry 3PML[52]. This DNA-enzyme complex contains a wobble(G-T) mismatch involving a 5' thymine as part of a larger DNA molecule and a guanine triphosphate monomer. The topology was generated using the CHARMM36 force field[53, 54], and the SPCE water model [55]. The system was minimised to 12 kJ mol^{-1} nm^{-1} before dynamical NVT simulations were computed with 1 fs timestep at 300 K. The simulation system is shown in 8.



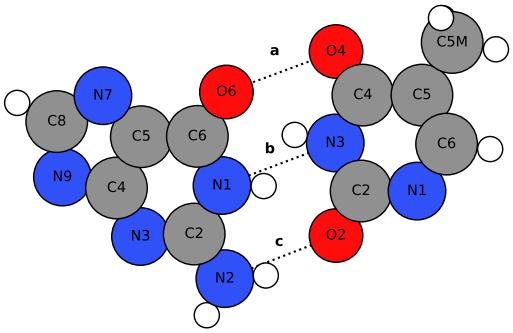
Supplementary Figure 8: The simulation systems for classical MD and hybrid QM/MM MD. Panel a) is the enzyme-less solvated DNA (orange cartoon) system with GTP (green space-filled) bound in the wobble(G-T) configuration, and shown in detail in panel c). Panel b) shows the enzyme-DNA complex of Polymerase-λ (purple cartoon) with the same DNA and GTP representations as panel a). Panel d) shows the wobble(G-T) configuration in the thumb domain of the enzyme.

Ensemble QM/MM MD

Hybrid quantum-classical calculations were performed in Gromacs 2021.1[51], using the interface to quantum chemistry package CP2K[56]. A total of 25 replica QM/MM simulations were performed; per replica, 8000 1 ps timesteps were evaluated within an NVT ensemble at 300 K. For the quantum mechanical region shown in Fig. 9, the BLYP+D3/DZVP-MOLOPT-GTH level of theory was utilised.

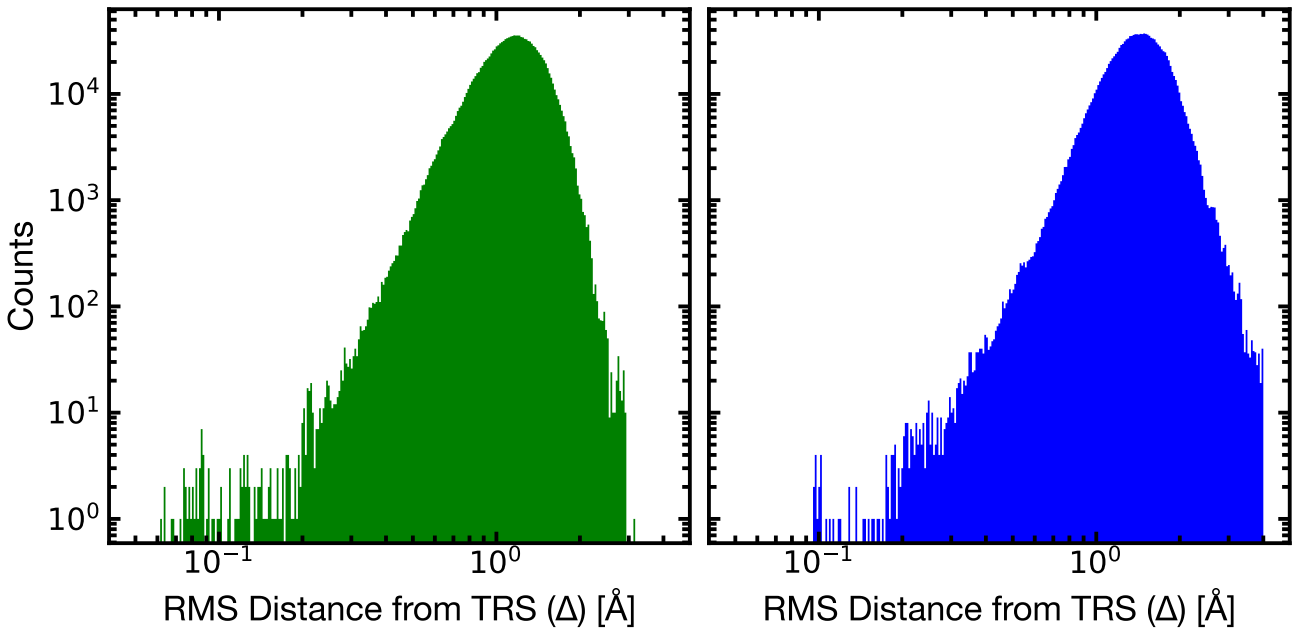
Compression reaction coordinate definition

At each timestep of dynamics, whether MM or QM/MM, the distances a, b, and c shown in Figure 9 were recorded. A root-mean-square distance to the tunnelling ready state can be calculated in terms of these three reaction coordinates as shown in Eq. 35. The reference values a_{TRS} are taken from the tunnelling ready state obtained through ML-NEB calculations described previously.



Supplementary Figure 9: Reaction coordinate definition for the statistical sampling of the tunnelling ready state. a,b, and c are the three reaction coordinates used to quantify the compression of the G-T wobble dimer.

$$\Delta = \sqrt{(a - a_{\mathrm{TRS}})^2 + (b - b_{\mathrm{TRS}})^2 + (c - c_{\mathrm{TRS}})^2} \tag{35}$$



Supplementary Figure 10: Histogram of the compression metric (Δ) for the wobble(G-T) dimer in Polymerase (left panel) and just with DNA (right panel). Each panel corresponds to data aggregated from over 180 ps of ensemble QM/MM MD.

SUPPLEMENTARY REFERENCES

[1] E. Apra, E. J. Bylaska, W. A. De Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. van Dam, Y. Alexeev, J. Anchell, *et al.*, Nwchem: Past, present, and future, The Journal of Chemical Physics **152**, 184102 (2020), https://doi.org/10.1063/5.0004997.

[2] A. D. Becke, Density-functional thermochemistry. iii. the role of exact exchange, The Journal of Chemical Physics **98**, 5648 (1993), https://doi.org/10.1063/1.464913.

[3] S. Grimme, S. Ehrlich, and L. Goerigk, Effect of the damping function in dispersion corrected density functional theory, Journal of Computational Chemistry **32**, 1456 (2011), https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21759.

[4] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu, The Journal of Chemical Physics **132**, 154104 (2010), https://doi.org/10.1063/1.3382344.

[5] O. O. Brovarets' and D. M. Hovorun, Atomistic mechanisms of the double proton transfer in the h-bonded nucleobase pairs: Qm/qtaim computational lessons, Journal of Biomolecular Structure and Dynamics **37**, 1880 (2019).

[6] A. Klamt and G. Schüürmann, Cosmo: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient, J. Chem. Soc., Perkin Trans. 2 , 799 (1993).

[7] D. M. York and M. Karplus, A smooth solvation potential based on the conductor-like screening model, The Journal of Physical Chemistry A **103**, 11060 (1999), https://doi.org/10.1021/jp992097l.

[8] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions, The Journal of Physical Chemistry B **113**, 6378 (2009), pMID: 19366259, https://doi.org/10.1021/jp810292n.

[9] M. H. Hansen, J. A. G. Torres, P. C. Jennings, *et al.*, An atomistic machine learning package for surface science and catalysis, arXiv preprint arXiv:1904.00904 (2019).

[10] J. A. G. Torres, P. C. Jennings, M. H. Hansen, *et al.*, Low-scaling algorithm for nudged elastic band calculations using a surrogate machine learning model, Phys. Rev. Lett. **122**, 156001 (2019).

[11] A. H. Larsen, J. J. Mortensen, J. Blomqvist, *et al.*, The atomic simulation environment—a python library for working with atoms, J. Phys.: Condens. Matter **29**, 273002 (2017).

[12] S. R. Bahn and K. W. Jacobsen, An object-oriented scripting interface to a legacy electronic structure code, Comput. Sci. Eng. **4**, 56 (2002).

[13] P. Jurečka, J. Šponer, J. Černý, and P. Hobza, Benchmark database of accurate (mp2 and ccsd(t) complete basis set limit) interaction energies of small model complexes, dna base pairs, and amino acid pairs, Phys. Chem. Chem. Phys. **8**, 1985 (2006).

[14] D. S. Tikhonov, A simplistic computational procedure for tunneling splittings caused by proton transfer, Structural Chemistry **33**, 351 (2022).

[15] S. Schweiger, B. Hartke, and G. Rauhut, Double proton transfer reactions at the transition from a concerted to a stepwise mechanism: a comparative ab initio study, Physical Chemistry Chemical Physics **7**, 493 (2005).

[16] S. Schweiger and G. Rauhut, Plateau reactions: Double proton-transfer processes with structureless transition states, The Journal of Physical Chemistry A **107**, 9668 (2003).

[17] R. Meyer and H. H. Günthard, General internal motion of molecules, classical and quantum-mechanical hamiltonian, The Journal of Chemical Physics **49**, 1510 (1968).

[18] R. Meyer and H. H. Günthard, Internal rotation and vibration in ch2= ccl–ch2d, The Journal of Chemical Physics **50**, 353 (1969).

[19] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University Press on Demand, 2007).

[20] E. Wigner, On the quantum correction for thermodynamic equilibrium, Phys. Rev. **40**, 749 (1932).

[21] J. Weinbub and D. K. Ferry, Recent advances in wigner function approaches, Applied Physics Reviews **5**, 041104 (2018), https://doi.org/10.1063/1.5046663.

[22] J. E. Moyal, Quantum mechanics as a statistical theory, Mathematical Proceedings of the Cambridge Philosophical Society **45**, 99–124 (1949).

[23] K. İmre, E. Özizmir, M. Rosenbaum, and P. F. Zweifel, Wigner method in quantum statistical mechanics, Journal of Mathematical Physics **8**, 1097 (1967), https://doi.org/10.1063/1.1705323.

[24] A. O. Caldeira and A. J. Leggett, Path integral approach to quantum brownian motion, Physica A: Statistical mechanics and its Applications **121**, 587 (1983).

[25] C. Trahan and R. Wyatt, *Quantum Dynamics with Trajectories: Introduction to Quantum Hydrodynamics*, Interdisciplinary Applied Mathematics (Springer New York, 2006).

[26] I. Burghardt and K. B. Møller, Quantum dynamics for dissipative systems: A hydrodynamic perspective, The Journal of chemical physics **117**, 7409 (2002).

[27] G. Agarwal, Brownian motion of a quantum oscillator, Physical Review A **4**, 739 (1971).

[28] F. Haake and R. Reibold, Strong damping and low-temperature anomalies for the harmonic oscillator, Physical Review A **32**, 2462 (1985).

[29] F. Haake, H. Risken, C. Savage, and D. Walls, Master equation for a damped nonlinear oscillator, Physical Review A **34**, 3969 (1986).

[30] K. H. Hughes, Dissipative quantum phase space dynamics on dynamically adapting grids, The Journal of chemical physics **122**, 074106 (2005).

[31] T. Ikeda and Y. Tanimura, Low-temperature quantum fokker–planck and smoluchowski equations and their extension to multistate systems, Journal of Chemical Theory and Computation **15**, 2517 (2019).

[32] C. Rackauckas and Q. Nie, Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in julia, J. Open Res. Softw. **5**, 15 (2017).

[33] J. C. Butcher, Numerical methods for ordinary differential equations in the 20th century, J. Comput. Appl. Math **125**, 1 (2000).

[34] M. Topaler and N. Makri, Quantum rates for a double well coupled to a dissipative bath: Accurate path integral results and comparison with approximate theories, The Journal of Chemical Physics **101**, 7500 (1994), https://doi.org/10.1063/1.468244.

[35] A. Pomyalov and D. J. Tannor, The non-markovian quantum master equation in the collective-mode representation: Application to barrier crossing in the intermediate friction regime, The Journal of Chemical Physics **123**, 204111 (2005), https://doi.org/10.1063/1.2121649.

[36] I. R. Craig, M. Thoss, and H. Wang, Proton transfer reactions in model condensed-phase environments: Accurate quantum dynamics using the multilayer multiconfiguration time-dependent hartree approach, The Journal of Chemical Physics **127**, 144503 (2007), https://doi.org/10.1063/1.2772265.

[37] S. Y. Kim and S. Hammes-Schiffer, Hybrid quantum/classical molecular dynamics for a proton transfer reaction coupled to a dissipative bath, The Journal of chemical physics **124**, 244102 (2006).

[38] Y. Tanimura and P. G. Wolynes, Quantum and classical fokker-planck equations for a gaussian-markovian noise bath, Phys. Rev. A **43**, 4131 (1991).

[39] Y. Tanimura and P. G. Wolynes, The interplay of tunneling, resonance, and dissipation in quantum barrier crossing: A numerical study, The Journal of chemical physics **96**, 8485 (1992).

[40] J. Zhang, R. Borrelli, and Y. Tanimura, Proton tunneling in a two-dimensional potential energy surface with a non-linear system–bath interaction: Thermal suppression of reaction rate, The Journal of Chemical Physics **152**, 214114 (2020).

[41] A. Ishizaki and Y. Tanimura, Multidimensional vibrational spectroscopy for tunneling processes in a dissipative environment, The Journal of chemical physics **123**, 014503 (2005).

[42] P. Hänggi, P. Talkner, and M. Borkovec, Reaction-rate theory: fifty years after kramers, Reviews of modern physics **62**, 251 (1990).

[43] J. P. Klinman and A. R. Offenbacher, Understanding biological hydrogen transfer through the lens of temperature dependent kinetic isotope effects, Accounts of chemical research **51**, 1966 (2018).

[44] L. O. Johannissen, S. Hay, and N. S. Scrutton, Nuclear quantum tunnelling in enzymatic reactions – an enzymologist's perspective, Phys. Chem. Chem. Phys. **17**, 30775 (2015).

[45] L. O. Johannissen, A. I. Iorgu, N. S. Scrutton, and S. Hay, What are the signatures of tunnelling in enzyme-catalysed reactions?, Faraday Discuss. **221**, 367 (2020).

[46] P. Li, A. Rangadurai, H. M. Al-Hashimi, and S. Hammes-Schiffer, Environmental effects on guanine-thymine mispair tautomerization explored with quantum mechanical/molecular mechanical free energy simulations, Journal of the American Chemical Society **142**, 11183 (2020), pMID: 32459476, https://doi.org/10.1021/jacs.0c03774.

[47] A. Rohatgi, Webplotdigitizer: Version 4.6 (2022).

[48] I. J. Kimsey, E. S. Szymanski, W. J. Zahurancik, A. Shakya, Y. Xue, C.-C. Chu, B. Sathyamoorthy, Z. Suo, and H. M. Al-Hashimi, Dynamic basis for dg• dt misincorporation via tautomerization and ionization, Nature **554**, 195 (2018).

[49] I. J. Kimsey, K. Petzold, B. Sathyamoorthy, Z. W. Stein, and H. M. Al-Hashimi, Visualizing transient watson–crick-like mispairs in dna and rna duplexes, Nature **519**, 315 (2015).

[50] A. Rangadurai, E. S. Szymanski, I. Kimsey, H. Shi, and H. M. Al-Hashimi, Probing conformational transitions towards mutagenic watson–crick-like g· t mismatches using off-resonance sugar carbon r 1ρ relaxation dispersion, Journal of Biomolecular NMR , 1 (2020).

[51] H. Bekker, H. Berendsen, E. Dijkstra, S. Achterop, R. Vondrumen, D. Vanderspoel, A. Sijbers, H. Keegstra, and M. Renardus, Gromacs-a parallel computer for molecular-dynamics simulations, in *4th International Conference on Computational Physics (PC 92)* (World Scientific Publishing, 1993) pp. 252–256.

[52] K. Bebenek, L. C. Pedersen, and T. A. Kunkel, Replication infidelity via a mismatch with watson–crick geometry, Proceedings of the National Academy of Sciences **108**, 1862 (2011).

[53] K. Hart, N. Foloppe, C. M. Baker, E. J. Denning, L. Nilsson, and A. D. MacKerell Jr, Optimization of the charmm additive force field for dna: Improved treatment of the bi/bii conformational equilibrium, Journal of chemical theory and computation **8**, 348 (2012).

[54] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. MacKerell Jr, Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ1 and χ2 dihedral angles, Journal of chemical theory and computation **8**, 3257 (2012).

[55] H. Berendsen, J. Grigera, and T. Straatsma, The missing term in effective pair potentials, Journal of Physical Chemistry **91**, 6269 (1987).

[56] T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, *et al.*, Cp2k: An electronic structure and molecular dynamics software package-quickstep: Efficient and accurate electronic structure calculations, The Journal of Chemical Physics **152**, 194103 (2020).