

**COMPUTATIONAL STUDIES OF INTERSTRAND CROSSLINK FORMATION IN
DNA**

NATHANIA TAKYI
Bachelor of Science, University of Lethbridge, 2020

A thesis submitted
In partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

CHEMISTRY

Department of Chemistry and Biochemistry
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

© Nathania Takyi, 2022

**COMPUTATIONAL STUDIES OF INTERSTRAND CROSSLINK FORMATION IN
DNA**

NATHANIA TAKYI

Date of Defence: November 23, 2022

Dr. S. Wetmore
Thesis Supervisor

Professor

Ph.D.

Dr. T. Patel
Thesis Examination Committee Member

Associate Professor

Ph.D.

Dr. P. Hayes
Thesis Examination Committee Member

Professor

Ph.D.

Dr. M. Gerken
Chair, Thesis Examination Committee

Professor

Ph.D.

Abstract

Interstrand crosslinks (ICLs) are toxic DNA lesions that can result in cell death if left unrepaired. ICLs can form following the reaction of natural DNA nucleobases with reactive chemical species generated during DNA repair. This thesis uses computational methods to investigate ICL formation resulting from the reaction of DNA repair intermediates with DNA nucleobases. Molecular dynamics (MD) simulations are used to provide an explanation for the effects of sequence context on the yield of ICLs that form following attempted repair of 1,N⁶-ethenoadenosine by alpha-ketoglutarate dependent dioxygenase homolog 2 (ALKBH2). Subsequently, an explanation for the preference of different Ap-derived ICLs is provided using density functional theory (DFT) calculations and MD simulations. Finally, the effect of sequence context on the dA-Ap ICL yield is investigated using MD simulations. Overall, this thesis uses molecular modelling approaches to rationalize how sequence context affects ICL yield and reveals the role of the structure of damaged DNA in ICL formation.

Preface

This thesis contains research conducted as part of a collaboration led by Dr. Stacey Wetmore at the University of Lethbridge and Dr. Deyu Li at the University of Rhode Island. The *in vitro* work in chapter 2 was completed by the following collaborators: Dr. Jie Wang at the University of Rhode Island, Qi Tang at the University of Rhode Island, Dr. Yi-Tzai Chen at the University of Rhode Island, and Dr. Deyu Li at the University of Rhode Island. Though *in vivo*, data is not presented in this thesis, *in vivo* work for the collaborative project was completed by: Yun-Chung Hsiao at the University of North Carolina at Chapel Hill, Dr. Chih-Wei Liu at the University of North Carolina at Chapel Hill, Dr. Kun Lu at the University of North Carolina at Chapel Hill, and Dr. John M. Essigmann at the Massachusetts Institute of Technology. For this collaboration (in chapter 2), I conducted all *in silico* studies, and participated in the writing and editing of the draft manuscript. Chapters 1, 3, and 4 of this thesis is all my original work, and has not been previously published.

Acknowledgements

I would like to thank my supervisor Dr. Stacey Wetmore. Your mentorship and support has been invaluable, and I will miss working with you. I would also like to thank my collaborators Dr. Deyu Li and Dr. Jie Wang. I would also like to thank my supervisory committee members, Dr. Paul Hayes and Dr. Trushar Patel for providing their advice and insight into the research that I pursued during my degree. I would also like to thank Dr. Michael Gerken for chairing my defense.

I have really enjoyed my time being part of the Wetmore lab and would like to thank past and present members for contributing to a positive and fun work environment (Dr. Preethi Seelam Prabhakar, Dr. James McFarlane, Dr. Ryan Kung, Priya Bhutani, Cynthia Fonderson, Prakash Chukka, Hanzala Hussein, Briana Boychuk, Rajwinder Kaur, Dylan Nikkel, Makay Murray, Rebecca Jeong, Bimaldeep Singh, Bhadra Pandya, Trinity Deak, Kohl Yee, Michaela Reitsma, Chalalai Phawaphootanon, Zaki Olvis, Allison Baptista, Saanya Yadav, Pousali Das, and Pardeepak Sandhu). Thank you for offering your support and jokes. I would also like to thank my friends Travis, Craig, and DJ.

I would like to thank my family members, especially my mother, father, sister, and brother for your prolonged support during my whole education. Thank you to all my friends that I have made before and during this degree. And finally, thank you to the people at the Coffee Company. Each of you have made my time in graduate school memorable and enjoyable. Finally, I would like to thank God for giving me the opportunity to meet all these wonderful people and allowing me to make it through to this point in my education.

Table of Contents

Abstract	iii
Preface	iv
Acknowledgements	v
List of Tables	ix
List of Figures	x
List of Abbreviations	xiv
Chapter 1: Introduction	1
1.1 Overview	1
1.2 The Biological Significance of DNA	2
1.3 Causes and Biological Implications of 1,N ⁶ -Ethenoadenosine	4
1.4 Repair of 1,N ⁶ -Ethenoadenosine	6
1.5 Interstrand Crosslink Formation as a Result of Attempted Repair	9
1.6 The Use of Computational Chemistry to Derive a Structural Rationale for ICL Formation	11
1.7 Thesis Overview	12
1.8 References	15
Chapter 2: A Structural Explanation of Sequence Effects on ALKBH2-mediated Interstrand Crosslink Formation	20
2.1 Introduction	20
2.2 Computational Methodology	23
2.2.1 MD Model Construction	23
2.2.2 MD Simulation Protocol	25

2.2.3 MD Analysis	26
2.3 Results	27
2.3.1 Primary Positions	27
2.3.2 Base Screening Sequences	28
2.3.3 Well-Matched Sequences	32
2.3.3.1 5'-XT-3'/3'-TA-5' Sequence Context	33
2.3.3.2 5'-XG-3'/3'-TC-5' Sequence Context	33
2.3.3.3 5'-XA-3'/3'-TT-5' Sequence Context	37
2.3.3.4 5'-XC-3'/3'-TG-5' Sequence Context	38
2.3.4 Mismatched Sequences	42
2.4 Discussion	45
2.5 Conclusions	50
2.6 References	52
Chapter 3: A Multiscale Computational Investigation of Ap-derived ICL Formation	55
3.1: Introduction	55
3.2 Computational Methods	59
3.2.1 Density Functional Theory Calculations	59
3.2.2 Molecular Dynamics Simulations	59
3.2.2.1 MD Model Construction	59
3.2.2.2 MD Simulation Protocol	60
3.2.2.3 MD Analysis	61
3.3 Results and Discussion	62
	62

3.3.1 Schiff-Base Formation is the Rate-Limiting Step of dA-Ap and dG-Ap ICL Formation	68
3.3.2 The dA-Ap ICL is Intrinsically Preferred to Form over the dG-Ap ICL from a Kinetic and Thermodynamic Standpoint	69
3.3.3 Increasing Solvent Polarity Slightly Decreases Reaction Barriers and Slightly Decreases Product Stability	71
3.3.4 In DNA Duplexes, the dA-Ap ICL Results in Less Helical Distortion and More Stabilizing NCIs than the dG-Ap ICL	83
3.3.5 The dA-Ap ICL causes Similar Structural Distortion in the 5'-AAG-3'/3'-TXC-5' and 5'-AAC-3'/3'-TXG-5' Sequence Contexts, but More Changes to Lesion Site Dynamics in the 5'-AAG-3'/3'-TXC-5' Sequence Context	
3.4 Conclusions	88
3.5 References	90
Chapter 4: Conclusions	94
4.1 Thesis Overview	94
4.2 Applications of Work	97
4.3 Future Work	98
4.4 Final Remarks	99
4.5 References	100
Appendix A: Supplementary Information for Chapter 2	102
Appendix B: Supplementary Information for Chapter 3	130

List of Tables

Table 2.1: Base screening experiments for ϵ A-induced crosslink by ALKBH2 repair. N.D. indicates not detectable under the reaction condition. ²	22
Table 2.2: Yield of interstrand crosslink generated from ϵ A repaired by ALKBH2 in well-matched and mismatched sequences. N.D. indicates not detectable under the reaction condition. ²	23
Table 3.1: Dipole moments of the RC and PC for dA-Ap and dG-Ap ICL formation. ^a	69
Table 3.2: B-DNA backbone, minor groove, and glycosidic torsion structural similarity for the dA-Ap AAG, dG-Ap AAG, or dA-Ap AAC duplexes. ^a	82

List of Figures

Figure 1.1: (A) A DNA nucleotide, B) Complementary A:T and G:C Watson-Crick base pairs (with atomic ie. numbering), and C) B-DNA duplex structure (left), with the major groove highlighted in blue and the minor groove in orange (right).	4
Figure 1.2: (A) Chemical structure and atomic numbering of ϵ A with the adenine portion of the modified nucleobase in black and the etheno modification highlighted in red and (B) natural adenine.	5
Figure 1.3: The proposed DRR pathway for ϵ A by ALKBH2 (and AlkB), with ϵ A bound in the active site of AlkB (middle; PDB ID: 3O1P). The reaction begins at the top left with an asparate residue, two histidine residues, and three waters are coordinated to an Fe^{2+} ion.	7
Figure 1.4: The proposed BER pathway for ϵ A by repair AAG, APE1, DNA polymerase β , and DNA ligase.	8
Figure 2.1: Epoxide intermediate-derived ICL structures proposed by this work, with DNA1 being the ϵ A lesion containing strand and DNA2 being the complementary strand.	21
Figure 2.2: Geometric cutoffs used to identify reactive conformations, including the nucleophile–electrophile distance ($r(\text{N}_{\text{nuc1}}\text{C}_{\text{elec}})$) and the corresponding angle of attack ($\angle(\text{N}_{\text{nuc1}}\text{C}_{\text{elec}}\text{O}_{\text{leave}})$).	27
Figure 2.3: The position of different nucleophilic sites in nucleobases neighboring the epoxide intermediate considered in the present work for a representative segment of DNA taken from the 5'-TXG-3'/3'-ATC-5' sequence context.	28
Figure 2.4: Frequency of reactive conformations for ICL in the base-screening sequences, when (A) adenine, (B) cytosine, or (C) guanine is the primary interstrand nucleophile. Reactive conformations were considered for C10 (blue) or C11 (orange) in the epoxide, as well as the total for both sites (purple).	30
Figure 2.5: MD representative structures of the reactive conformations for crosslink formation between N6 of the complementary adenine (left) or N4 of the complementary cytosine (right) and C10 of the epoxide intermediate (X) for DNA duplexes containing the base screening sequences. Average electrophile–nucleophile distance ($r(\text{N}_{\text{nuc1}}\text{C}_{\text{elec}})$, Å) and angle of attack ($\angle(\text{N}_{\text{nuc1}}\text{C}_{\text{elec}}\text{O}_{\text{leave}})$, deg.) across all replicas provided.	31
Figure 2.6: MD representative structures of the reactive conformations for crosslink formation between N6 of the 5'-interstrand adenine (left) or N4 of the 5'-interstrand cytosine (right) and C11 of the epoxide intermediate (X) for DNA duplexes containing the base screening sequences. Average electrophile–nucleophile distance ($r(\text{N}_{\text{nuc1}}\text{C}_{\text{elec}})$, Å) and angle of attack ($\angle(\text{N}_{\text{nuc1}}\text{C}_{\text{elec}}\text{O}_{\text{leave}})$, deg.) across all replicas provided.	31

Figure 2.7: Frequency of reactive conformations suitable for interstrand (left) or intrastrand (right) crosslink formation for 5'-XT/5'-AT (A and B) or 5'-XG/5'-CT (C and D) sequence contexts. Reactive conformations involving C10 (blue) or C11 (orange) of the epoxide are provided, as well as the total for either site (purple). 32

Figure 2.8: MD representative structures of the reactive conformations for crosslink formation between N6 of the 5'-interstrand adenine and C10 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XT/5'-AT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided. 34

Figure 2.9: MD representative structures of the reactive conformations for crosslink formation between N6 of the 5'-interstrand adenine and C11 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XT/5'-AT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided. 35

Figure 2.10: MD representative structures of the reactive conformations for crosslink formation between N4 of the 5'-interstrand cytosine and C10 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XG/5'-CT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided. 36

Figure 2.11: MD representative structures of the reactive conformations for crosslink formation between N4 of the 5'-interstrand cytosine and C11 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XG/5'-CT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided. 37

Figure 2.12: Frequency of reactive conformations suitable for interstrand (left) or intrastrand (right) crosslink formation for 5'-XC/5'-GT (A and B) or 5'-XA/5'-TT (C and D) sequence contexts. Reactive conformations involving C10 (blue) or C11 (orange) of the epoxide are provided, as well as the total for either site (purple). 38

Figure 2.13: MD representative structures of the reactive conformations for crosslink formation between N6 of the 3'-intrastrand adenine and C10 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XA/5'-TT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided. 39

Figure 2.14: MD representative structures of the reactive conformations for crosslink formation between N6 of the 3'-intrastrand adenine and C11 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XA/5'-TT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided. 40

Figure 2.15: MD representative structures of the reactive conformations for crosslink formation between N4 of the 3'-intrastrand cytosine and C10 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XC/5'-GT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.	41
Figure 2.16: MD representative structures of the reactive conformations for crosslink formation between N4 of the 3'-intrastrand cytosine and C11 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XC/5'-GT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.	42
Figure 2.17: Frequency of reactive conformations for the formation of ICL in the mismatched sequences. Reactive conformations for a nucleophile in the 5'-interstrand nucleotide and C10 (blue) or C11 (orange) of the epoxide, a nucleophile at the direct opposite position in the complementary nucleotide and C10 (yellow) or C11 (green) of the epoxide, and the total reactive conformations (purple) are provided.	43
Figure 2.18: MD representative structures of the reactive conformations for crosslink formation between N6 of the 5'-interstrand adenine and C10 (top) or N6 of the 5'-interstrand adenine and C11 (bottom) of the epoxide intermediate (X) for DNA duplexes containing mismatches. Average electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.	44
Figure 2.19: MD representative structures of the reactive conformations for crosslink formation between N4 of the complementary cytosine (left) or N6 of the complementary adenine (right) and C11 of the epoxide intermediate (X) for DNA duplexes containing mismatches. Average electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.	45
Figure 2.20: MD representative structures of the lesion site in the 5'-XC/5'-GT sequence context, displaying that the amino group of the interstrand guanine and the epoxide intermediate is too far to react for each 5'-XC/5'-GT duplex. Average electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.	47
Figure 3.1: A representative example of the proposed Ap-derived ICL formation reaction (C1' of an Ap site and N6 of A).5.....	55
Figure 3.2: (top) Ap-derived ICLs: dA-Ap (left), dC-Ap (right), dG-Ap (left), and (bottom) the sequence contexts in which they form. ^{2, 4-5}	57
Figure 3.3: Structures and relative energies for stationary points of A) dA-Ap and B) dG-Ap ICL formation calculated using M06-2X/6-311+G(2df,2p)//B3LYP/6-31G(d) in the gas phase.	63

Figure 3.4: (A) Important angles (deg.) and distances (Å) for (A) step 1, (B) step 2, and (C) step 3 of dA-Ap ICL formation.	65
Figure 3.5: (A) Important angles (deg.) and distances (Å) for (A) step 1, (B) step 2, and (C) step 3 of dG-Ap ICL formation.	66
Figure 3.6: Relative Gibbs energies of stationary points for (A) dA-Ap and (B) dG-Ap ICL formation calculated using B3LYP, B3LYP-D3(BJ), M06-2X, and MP2 with the 6-311+G(2df,2p) basis set in the gas phase.	67
Figure 3.7: Relative Gibbs energies of stationary points for (A) dA-Ap and (B) dG-Ap ICL formation calculated using M06-2X/6-311+G(2df,2p) basis set in the gas phase, implicit 1-bromopropane, and implicit water.	70
Figure 3.8: C1'-C1' distances for the (A) dA-Ap ICL and (B) dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' sequence context and (C) the dA-Ap ICL in the 5'-AAC-3'/3'-TXG-5' sequence context, along with the (D, E, F) corresponding C1'-C1' distances in natural DNA.	72
Figure 3.9: Lesions sites for the (A) dA-Ap AAG, (B) dG-Ap AAG, (C) dA-Ap AAC, (D) AAG natural, and (E) AAC natural models. Crosslinking residues have been bolded.	73
Figure 3.10: Overlay of the sugar at residue 23 in natural DNA (black) and (A, dark blue) dA-Ap AAG (B, purple) dG-Ap AAG, and (C, yellow) dA-Ap AAC.	75
Figure 3.11: Change in RMSF (in Å) for lesion site residues of dA-Ap AAG (yellow), dG-Ap AAG (purple), and dA-Ap AAC (blue) relative to natural DNA values.	79
Figure 3.12: Average and standard deviation of helical parameters at base steps 4 and 12 in a natural DNA duplex (black circle), a DNA duplex containing the dA-Ap ICL (dark blue diamond), or the dG-Ap ICL (purple diamond) in the 5'-AAG-3'/3'-TXC-5' sequence context.	80
Figure 3.13: Average and standard deviation of helical parameters at base steps 4 and 12 in a natural DNA duplex (black circle) or a DNA duplex the dA-Ap ICL (yellow diamond).	86
Figure 3.14: Representative structures for: (A) AAG dA-Ap, (B) AAG dG-Ap, and (C) AAC dA-Ap ICL, with (from left, to right) severity of backbone glycosidic torsion, and minor groove distortion indicated by the intensity of red.	87

List of Abbreviations

3' inter	3'-interstrand
3' intra	3'-intrastrand
5' inter	5'-interstrand
5' intra	5'-intrastrand
A	adenine
AlkB	alpha-ketoglutarate dependent dioxygenase
ALKBH2	alpha-ketoglutarate dependent dioxygenase homolog 2
Ap	apurinic/aprimidinic
Asp	aspartate
BER	base excision repair
C	cytosine
DNA	deoxyribonucleic acid
DFT	density functional theory
DRR	direct reversal repair
dA-Ap	adenine-apurinic/aprimidinic site crosslink
dC-Ap	cytosine-apurinic/aprimidinic site crosslink
dG-Ap	guanine-apurinic/aprimidinic site crosslink
ϵ A	1,N ⁶ -ethenoadenosine
FA	Fanconi anemia
FRC	frequency of reactive conformations
G	guanine
GAFF	generalized AMBER force field
His	histidine
IC1	intermediate complex 1
IC2	intermediate complex 2
ICL	interstrand crosslink
MD	molecular dynamics
MM	molecular mechanics
NCI	noncovalent interactions
NCP	nucleosome core particle
P*	purine
PC	product complex
RONS	reactive oxygen and nitrogen species
RC	reactive complex
T	thymine
TS1	transition state 1
TS2	transition state 2
TS3	transition state 3
U	uracil
UC	ulcerative colitis
WC	Watson-Crick
QM/MM	quantum mechanics/molecular mechanics

Chapter 1: Introduction

1.1 Overview

Deoxyribonucleic acid (DNA) is a macromolecule that functions as the starting point for cellular development as it encodes RNA transcripts, which are used to synthesize proteins.¹ While some of these transcripts are noncoding, other transcripts can be translated into proteins that play vital roles in cellular structure and function.¹ A DNA nucleotide is made up of a phosphate, a sugar, and one of the four DNA nucleobases: adenine (A), cytosine (C), guanine (G), or thymine (T).¹ When DNA is damaged, the structural and functional integrity of a cell can be compromised, resulting in diseases.² 1,N⁶-ethenoadenosine (ϵ A) is a type of DNA damage that forms when DNA is exposed to byproducts of lipid metabolism or pollutants generated during polyvinyl chloride production and water chlorination.²⁻⁴ ϵ A is known to cause A \rightarrow T, A \rightarrow C, and A \rightarrow G substitution mutations, and has been connected to a number of cancers (gastric, colon, and lung cancers).² Therefore, it is imperative that this lesion is repaired.² In human cells, ϵ A can be repaired through two distinct pathways: direct reversal repair (DRR) and base excision repair (BER).² While DRR of ϵ A is executed by ALKBH2 and involves the removal of the alkene group directly from the nucleobase, BER is performed by multiple enzymes and involves removal of the damaged nucleobase through cleavage of the glycosidic bond to form an abasic site.² In BER, glycosidic bond cleavage is followed by removal of the remaining sugar and phosphate, and insertion of the correct canonical nucleotide.² While these repair pathways are distinct, both DRR and BER involve the creation of reactive intermediates which can interact with the amino groups of nearby nucleobases to yield interstrand crosslinks (ICLs). It has been observed that the amount of the resulting ICLs formed depends on sequence context, and the underlying reason for this is unknown.⁵⁻⁷

ICLs are particularly toxic DNA lesions that can lead to cell death or senescence if not repaired. In addition, a suppressed ability to repair ICLs has been connected to Fanconi anemia (FA), a condition that leads to organ dysfunction, problems with bone marrow development, and increased risk of cancer.⁸ On the other hand, the toxicity of ICLs have been proven useful as the induction of ICLs has been utilized as a chemotherapeutic technique.⁹ To understand how to better utilize ICLs as a therapeutic technique and learn more about the role that ICLs play in the development of a variety of cancers following FA diagnosis, I must first understand how ICLs form in different environments, and how these ICLs affect the structure of DNA. As computational chemistry has been a useful and efficient technique to provide atomic level descriptions of biochemical reactions, the role of macromolecule structure in the propagation of reactions, and the effect of DNA damage on macromolecule structure and function, this thesis utilizes computational chemistry techniques including molecular dynamics simulations (MD) and density functional theory (DFT) methods to understand the formation of different ICLs and how ICLs affect the structure of DNA, leading to hypotheses regarding the thermal stability of DNA containing different ICLs.¹⁰⁻¹⁸ This chapter documents the context necessary to understand the results and implications of this work including the structure of DNA, the biological implications of ϵ A damage, how attempted repair of DNA results in ICLs through the creation of reactive intermediates, the deviations in experimentally observed ICL yields in response to sequence context, and the biological implications of ICL formation.

1.2 The Biological Significance of DNA

As the storage of genetic information is required for competent cellular function, DNA is a crucial macromolecule for the propagation of life. A DNA molecule is made up of a sequence

of monomers called nucleotides, each of which are composed of a phosphate group, deoxyribose, and a nucleobase. The four natural (unmodified) DNA nucleotides differ from each other in the identity of their nucleobase, namely A, G, C, or T (Figure 1.1). Information is encoded in DNA through the sequence of the nucleotides. The permutation of nucleotides contained in each gene is replicated, then passed to daughter cells during cellular division. While the quality of replication depends on the structure of DNA, healthy cells have been found to have lower rates of error in replication when compared to unhealthy cells.¹⁹ While the structure of DNA can differ based on cellular conditions, the majority of DNA adopts a B-DNA structure that involves the pairing of two complementary DNA strands through Watson-Crick hydrogen bonding of A to T and G to C, with a right-handed twist around the helical axis (Figure 1.1).¹⁹ B-DNA also contains a major groove, which can be described as the side of the DNA where the backbones of the independent strands are far apart, and a minor groove where the backbones of the two strands are closer together.¹⁹ The structure and stability of DNA can be compromised when DNA is damaged, which may affect important biochemical processes.¹ Deleterious nucleobase modifications, such as ϵ A (Figure 1.2), have been proven to compromise cellular health,² which will be discussed in the following subsection.

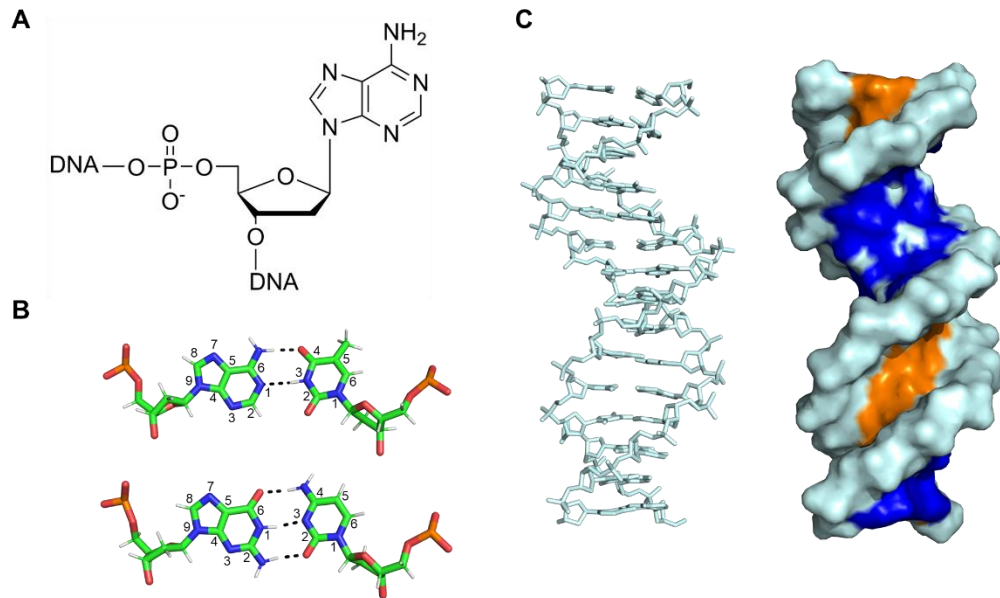


Figure 1.1: (A) A DNA nucleotide, B) Complementary A:T and G:C Watson-Crick base pairs (with atomic ie. numbering), and C) B-DNA duplex structure (left), with the major groove highlighted in blue and the minor groove in orange (right).

1.3 Causes and Biological Implications of 1,N⁶-Ethenoadenosine

ϵ A is formed when A is exposed to lipid metabolism byproducts or vinyl chloride.²

Experimental investigations revealed the presence of etheno modifications in chronically inflamed human tissues, as reactive oxygen and nitrogen species (RONS) generated in response to cellular stress break down phospholipids, and create aldehyde byproducts that can react with the DNA nucleobases.² Etheno modifications have also been found in humans following chronic exposure to vinyl chloride, as vinyl chloride can be activated by cytochrome P450, resulting in chloroethylene oxide and chloroacetaldehyde.² Extensive studies on the formation of ϵ A adducts have revealed a number of other possible causes of ϵ A. For example, connections between ϵ A and ethanol consumption have been found, as the presence of ϵ A was confirmed in liver cells

following incubation in ethanol.² Mucochloric acid, which can be found in drinking water as a byproduct of chlorination, was discovered to react with nucleotides *in vitro* to form ethenoadducts.^{4, 20} In addition, the carcinogen ethyl carbamate, formed as a byproduct of fermentation of food and alcohol, can be oxidized into an epoxide that reacts with nucleic acids to form ethenoadducts.²¹⁻²²

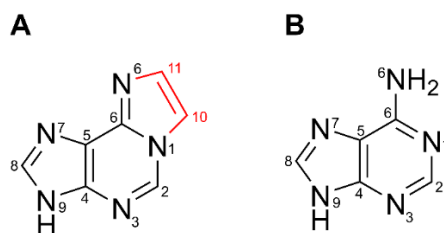


Figure 1.2: (A) Chemical structure and atomic numbering of εA with the adenine portion of the modified nucleobase in black and the etheno modification highlighted in red and (B) natural adenine.

The alkene group of εA is located on the Watson-Crick face of A, which changes hydrogen bonding between A and T.² This may contribute to the fact that εA is highly mutagenic, as εA was found to cause A → T, A → G, and A → C mutations in *E. coli* and human cells, and A → T, A → G, A → C, and -1 frameshift mutations *in vitro*.² In bacteria, the transfection of εA increased the chance of genome mutation by 65%.²³ While there indeed appears to be a relationship between εA and inflamed tissues, εA has also been connected to cancer, with decreased levels of εA repair in the lung adenocarcinoma leukocytes of cancer patients, and the presence of εA linked to the development of lung, colon, and liver cancer.^{2, 24} In addition, this lesion has been located in healthy tissues, with previous studies documenting 2.3 – 2.5 adducts per 10⁶ A bases in human placental DNA.²⁵ Because this lesion is highly mutagenic, εA repair is pivotal for healthy cellular function and the avoidance of disease.²

1.4 Repair of 1,N⁶-Ethenoadenosine

In vivo, ϵ A is mainly repaired through two independent pathways: DRR and BER.⁴ DRR of ϵ A involves a single enzyme: namely AlkB in *E. coli* and ALKBH2 in humans.²⁶ At the beginning of this repair process, ALKBH2 (or AlkB) searches the DNA strand and flips each nucleobase into its active site.² In ALKBH2, the base flipping process is driven by the intercalation of Val101 and Phe102 into the helix, allowing ϵ A to reside in the active site, stacked between Phe124 and His171.²⁷ The significant difference in favorability for the base-flipping of ϵ A compared to natural nucleotides contributes to the high specificity of ALKBH2 for damaged DNA.²⁷ The catalytically important residues in the active site include two histidine (His171 and 236) residues and an aspartate (Asp173) residue.²⁸ In addition to three water molecules, His171, His236 and Asp173 are coordinated to an Fe(II) ion (Figure 1.3, step 1). The removal of the alkene group begins with the displacement of two water molecules from the Fe(II) complex upon α -ketoglutarate (2OG) coordination (Figure 1.3, step 2). In Figure 1.3, step 3, an O₂ molecule oxidizes Fe(II) into Fe(IV) and attacks a carbonyl on α -ketoglutarate. As a result of this oxygen activation, carbon dioxide, succinate, His171, His236, and Asp173 are coordinated to the newly formed Fe(IV) oxo group (Figure 1.3, step 4). The damaged ϵ A substrate is oriented towards the Fe(IV) complex and the alkene group of ϵ A is oxidized into an epoxide intermediate through the Fe(IV) oxo ligand (Figure 1.3, step 5). The alkene group is then hydrolyzed into a glycol (Figure 1.3, step 6), which is subsequently deprotonated at each hydroxyl group, and doubly protonated at nitrogen to result in natural A and glyoxal (Figure 1.3, step 7).²⁹

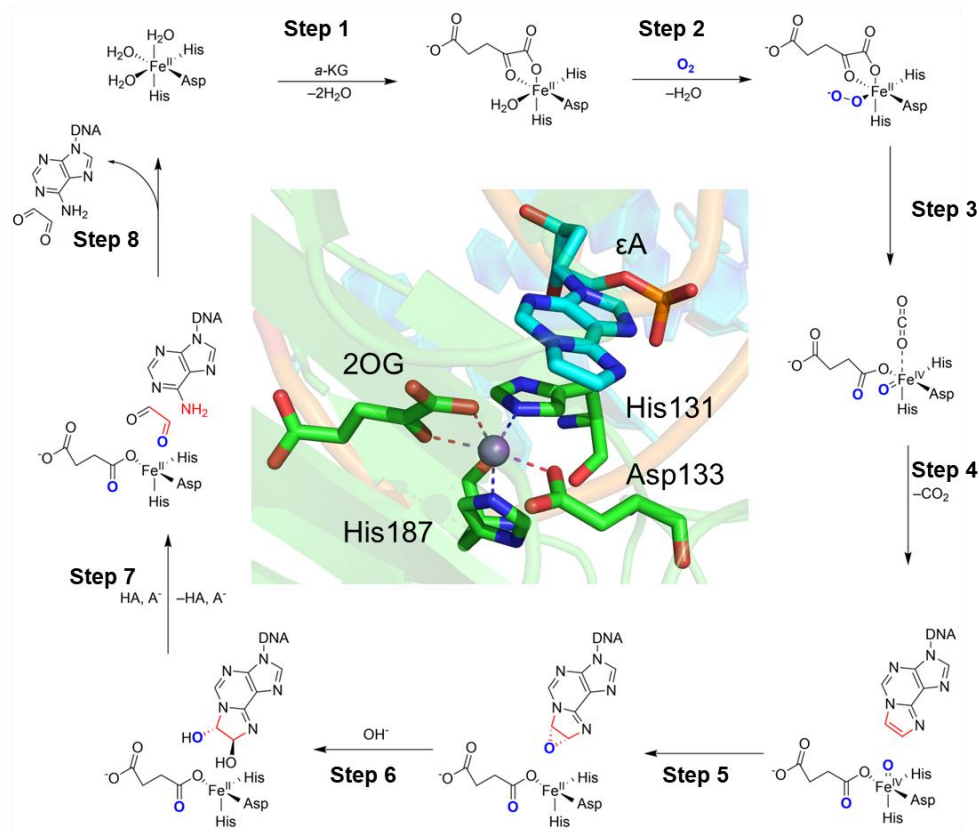


Figure 1.3: The proposed DRR pathway for ϵ A by ALKBH2 (and AlkB), with ϵ A bound in the active site of AlkB (middle; PDB ID: 3O1P). The reaction begins at the top left with an aspartate residue, two histidine residues, and three waters coordinated to an Fe^{2+} ion.

During the BER pathway, a DNA glycosylase flips ϵ A outside of the DNA helix, cleaves the glycosidic bond of the damaged nucleotide and results in an apurinic/apyrimidinic (Ap) site.⁴ The main glycosylase for this process is AlkA in *E. coli* and AAG in humans.⁴ After searching the DNA strand for damage, AAG intercalates Tyr162 into the DNA minor groove, flipping ϵ A into its active site. Tyr159, Met164, and Tyr165 aid this process by forming van der Waals interactions with the sugars of the damaged nucleotide pair, resulting in an increase in the major groove width. ϵ A stacks between Tyr127 and His136 in the AAG active site (Figure 1.4). In addition to stacking with the ϵ A substrate, Tyr127 hydrogen bonds to Glu125. Like ALKBH2, the change in hydrogen bonding at the damaged base pair upon ϵ A formation aids the high

selectivity of AAG for damaged nucleobases. Arg182 hydrogen bonds with the phosphate 3' with respect to ϵ A, and to an active site water. Hydrolysis of the glycosidic bond occurs as Glu125 deprotonates the active site water. The resulting hydroxide attacks C1' of ϵ A, leading to an Ap site and an ϵ A anion.³⁰⁻³¹ Following glycosidic bond cleavage, APE1 cleaves the sugar phosphate bond at the 5' end of the newly formed Ap site, polymerase β cleaves the 3' end of the Ap site and replaces the removed nucleotide with a natural nucleotide, and a ligase seals the backbone.²

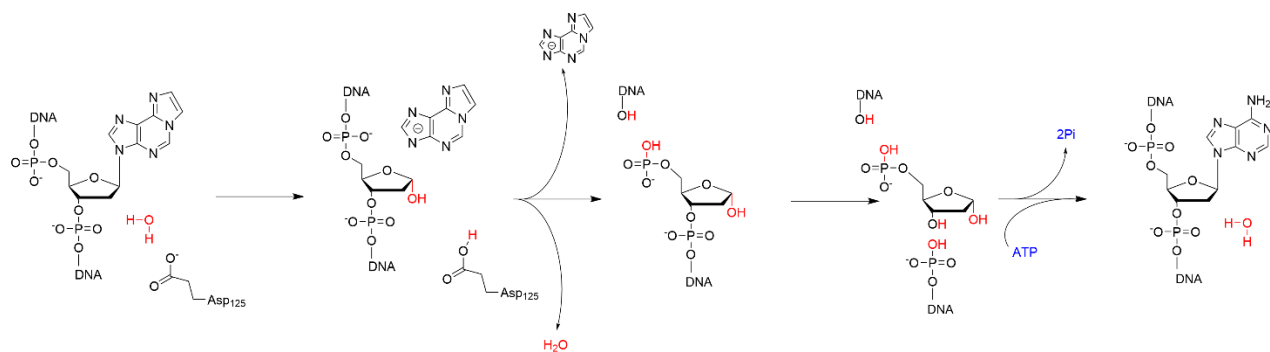


Figure 1.4: The proposed BER pathway for ϵ A by repair AAG, APE1, DNA polymerase β , and DNA ligase.

Both DRR and BER have been shown to play an important role in ϵ A repair within the cell. Experimental studies combined with structural analysis of a nucleosome core particle (NCP) crystal structure revealed that while ALKBH2 had the ability to repair ϵ A lesions located at the exterior, middle, and interior of a NCP particle, AAG exclusively repaired the ϵ A lesions present at the exterior of the NCP.³² However, AAG repaired the exterior ϵ A lesions at a significantly higher efficiency compared to ALKBH2.³² Regardless, both DRR and BER appear to be critical for cell survival in healthy and malignant human tissues.³² Moreover, ALKBH2 appears to be important for the survival of cancerous cells.³² For example, downregulation of ALKBH2 increased the sensitivity of bladder cancer cells to chemotherapeutics.³² Furthermore, the

overexpression of ALKBH2 has been found in human glioblastoma cell lines and has been linked to chemotherapeutic resistance.³³ ALKBH2 knockdown and chemotherapy has been found to be effective in the destruction of lung cancer cells and human colorectal cancer.^{9, 34} This supported the sentiment that DRR plays a significant role in the survival of malignant cells.³⁵

Comparable to ALKBH2, AAG has been related to genetic instability and disease in humans. For example, AAG activity has been connected to the presence of frameshift mutations and microsatellite instability.³⁶ Ulcerative colitis (UC) has been found to increase the risk of microsatellite instability and colorectal cancer, and patients with UC were found to have increased levels of AAG expression and activity in their inflamed tissue.³⁶ Increased AAG activity has also been correlated to lung cancer as patients with high AAG activity were reported to have a greater likelihood of lung cancer.³⁷ Like ALKBH2, AAG has been suspected to play a role in chemotherapeutic-resistant carcinogenicity, as ovarian cancer cells were more susceptible to alkylating agents and chemotherapeutics following AAG knockdown.³⁶ In addition, the significantly increased activity of AAG has been linked to the progression of other conditions, such as cerebellar toxicity and retinal degeneration.³⁶ While both ALKBH2 and AAG have been proven to be important for survival of both healthy and diseased cells, it has been discovered that these repair pathways can result in additional damage.

1.5 Interstrand Crosslink Formation as a Result of Attempted Repair

Both DRR and BER involve the creation of reactive electrophilic intermediates such as epoxides or aldehydes, which can be attacked by local nucleophilic amino groups on the complementary DNA strand to form ICLs.³² My collaborators in the Li lab at the University of

Rhode Island have discovered the formation of ICLs during the attempted repair of ϵ A by ALKBH2 *in vitro* and *in vivo*. They have also discovered that the relative ICL yield varies based on the identity of the nucleotides flanking ϵ A. It is suspected that the epoxide intermediate is formed during DRR, then flipped out of the ALKBH2 active site and back into the DNA duplex and attacked by the amino groups of nearby nucleobases to result in ICLs. Indeed, *in vitro* studies have found the largest ICL yields for DNA duplexes containing the 5'-XT/5'-AT sequence context, followed by duplexes containing the 5'-XG/5'-CT sequence context. On the other hand, duplexes containing the 5'-XC/5'-GT and 5'-XA/5'-TT sequence contexts result in the smallest (in many cases negligible) ICL yields.

ICL formation has also been observed through the electrophilic attack of an Ap site (an intermediate formed during BER) by nearby exocyclic amino groups.³⁸⁻⁴⁰ Specifically, ICLs are exclusively formed between an Ap site and A (dA-Ap) when A is present on the complementary strand offset in the 3' direction relative to the Ap site. Additionally, ICLs between an Ap site and C (dC-Ap) do not form when C is involved in a natural Watson–Crick pair, and ICLs between an Ap site and G (dG-Ap) only form when G is present on the strand complementary to the Ap site offset in the 5' direction relative to the Ap site.³⁸⁻⁴⁰ Previous work has found that in the 5'-AAG-3'/3'-TXC-5' (X = Ap site) sequence context, the dA-Ap ICL is the major product and the dG-Ap ICL is the minor product. In addition, the authors found that the dA-Ap ICL has a 70% yield in the 5'-AAG-3'/3'-TXC-5' sequence context and a 13-18% yield in the 5'-AAC-3'/3'-TXG-5' sequence context.²⁴

Because ICLs prevent DNA strands from being separated, ICL formation can prevent important cellular processes.⁴¹ For example, if a single ICL is not repaired, replication and protein synthesis will be blocked, leading the cell to proceed through apoptosis.⁴² ICLs have

been found to play a role in FA, a condition that causes an increased susceptibility to organ dysfunction, and an increased predisposition for cancer. FA increases cellular susceptibility to crosslinking agents, leading to the connection between the ability to repair ICLs and FA.⁸

The discoveries that different repair pathways can cause ICLs provide an opportunity for the expansion and fine-tuning of chemotherapeutic options, as the induction of crosslinks *via* administration of exogenous crosslinking agents (such as cisplatin, carboplatin, oxaliplatin, satraplatin, or cyclophosphamide) is a commonly used chemotherapeutic technique.^{28, 43} In addition, a deeper understanding of FA and the role that ϵ A may play in the morbidity of patients with FA is needed. To achieve this goal, a thorough understanding of why and how these crosslinks form, the effect of sequence context on ICL formation, and how the crosslinks affect the structure and dynamics of DNA is needed.

1.6 The Use of Computational Chemistry to Derive a Structural Rationale for ICL Formation

The atomic level insight provided by molecular modeling methods have often been used to understand the formation and structure of crosslinks in DNA. Using these *in silico* methods, chemical and structural rationales have been provided for a number of experimental observations. Methods such as MD simulations and QM/MM have been utilized to study intrastrand crosslinks and ICLs derived through the exposure of DNA to nitrogen mustards, chemotherapeutics, and other crosslinking agents.^{6, 44-47} As ALKBH2 mediated ICLs are newly discovered, there have been no previous computational studies centered on these ICLs. Furthermore, only a few works have focused on Ap-derived ICLs with QM and MD models.

Specifically, density functional theory (DFT) was used to investigate the mechanism of ICL formation *via* nucleophilic attack of Ap sites in which C4' is oxidized (C4' Ap sites) and discovered the order of reactivity for ICL formation to be: dA-Ap > dG-Ap > dC-Ap in terms of kinetic favorability, and dA-Ap > dC-Ap > dG-Ap in terms of thermodynamic favorability.⁶ Interestingly, the authors identified that the dA-Ap ICL adopted geometries that could not be accommodated by B-DNA without severely distorting the duplex.⁴⁸ Further MD simulations and DFT studies on C4' Ap sites found that the nucleobase opposing the damage can help stabilize the Ap site, making it more favorable for attack by nearby nucleobases to form ICLs.⁴⁹ To elucidate the reason for the variation in ICL yields in response to dA-Ap and dG-Ap ICL formation, another study on native Ap sites utilized MD and QM calculations on DNA duplexes containing the Ap site.⁵⁰ This study led to speculations that the positioning of the ICL within the DNA duplex may play an integral role in ICL formation.⁵⁰ Despite this previous computational work on the formation of Ap-derived ICLs, the formation mechanisms of native Ap-derived crosslinks have not been mapped and a detailed analysis of the local and global structure and dynamics of DNA duplexes containing Ap-derived ICLs has not been done. Nevertheless, these examples support the use of computational chemistry to study ICL formation and the structural effects that ICLs have on the DNA duplex. As such, these approaches are used in this thesis to provide a structural explanation of experimentally observed ICL yields.

1.7 Thesis Overview

The overarching goal of this thesis is to explain the role of structure and sequence context in ICL formation following attempted repair of ϵ A by ALKBH2 or AAG using structural and/or mechanistic explanations derived from computational chemistry. This knowledge will provide a

baseline for understanding how these ICLs form and their impact on the structure of DNA, which will allow for further studies of repair and product stability. This knowledge can contribute to our understanding of the roles ICLs play in diseases such as FA, as well as aid in chemotherapeutic development.

As members of the Li lab at the University of Rhode Island discovered that attempted repair of ϵ A by ALKBH2 resulted in ICL formation, and that the yield of ICL formed varied based on sequence context, Chapter 2 employs MD simulations on DNA 16mers containing an epoxide intermediate in sequence contexts complementary to the sequences studied by my collaborators. The frequency that nucleophilic amino groups and electrophilic epoxide carbon atoms adopt a relative orientation conducive to nucleophilic attack is used to provide a rationale for the role of sequence in ALKBH2-mediated ICL formation. This will afford a structural explanation for the link between sequence context and ICL yields following attempted ALKBH2 repair.

In chapter 3, truncated QM models of ICLs are first employed to understand the relative fundamental reactivity of the amino groups of A and G towards a ring-opened Ap site. Next, MD simulations on DNA containing the major dA-Ap or dG-Ap ICL products are used to determine how ICLs affect the local and global structure of the DNA duplex. This multiscale investigation of Ap-derived ICL formation provides insight into the observed yields of ICLs in different sequence contexts.⁵¹

Chapter 4 will highlight the major conclusions from these studies, discuss the biological implications, and provide possible directions for future projects. Overall, this thesis provides insight into the causes of ICLs, along with structural descriptions of the ICL products, which will be useful when performing future *in silico*, *in vitro*, and *in vivo* studies to investigate the relative

thermal stability and reparability of these ICLs. The knowledge that has been gained in this thesis can be applied towards developing novel chemotherapeutics and other pharmaceuticals through the induction or prevention of ICLs. In addition, a greater understanding of the relationship between the magnitude and type of DNA damage and morbidity in patients with FA may be achieved.

1.7 References

1. Duzdevich, D.; Redding, S.; Greene, E. C., DNA Dynamics and Single-Molecule Biology. *Chem. Rev.* **2014**, *114* (6), 3072-3086.
2. Rioux, K. L.; Delaney, S., 1,N6-Ethenoadenine: From Molecular to Biological Consequences. *Chem. Res. Toxicol.* **2020**, *33* (11), 2688-2698.
3. Boudewijns, T.; Piccinini, M.; Degraeve, P.; Liebens, A.; De Vos, D., Pathway to Vinyl Chloride Production via Dehydrochlorination of 1,2-Dichloroethane in Ionic Liquid Media. *ACS Catal.* **2015**, *5* (7), 4043-4047.
4. Kronberg, L.; Sjöholm, R.; Karlsson, S., Formation of 3,N4-Ethenocytidine, 1,N6-Ethenoadenosine, and 1,N2-Ethenoguanosine in Reactions of Mucochloric Acid with Nucleosides. *Chem. Res. Toxicol.* **1992**, *5* (6), 852-855.
5. Garrec, J.; Patel, C.; Rothlisberger, U.; Dumont, E., Insights into Intrastrand Cross-Link Lesions of DNA from QM/MM Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2012**, *134* (4), 2111-2119.
6. Churchill, C. D. M.; Eriksson, L. A.; Wetmore, S. D., Formation Mechanism and Structure of a Guanine–Uracil DNA Intrastrand Cross-Link. *Chem. Res. Toxicol.* **2011**, *24* (12), 2189-2199.
7. Dong, Q.; Barsky, D.; Colvin, M. E.; Melius, C. F.; Ludeman, S. M.; Moravek, J. F.; Colvin, O. M.; Bigner, D. D.; Modrich, P.; Friedman, H. S., A Structural Basis for a Phosphoramidate Mustard-Induced DNA Interstrand Cross-Link at 5'-d(GAC). *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92* (26), 12170-12174.
8. Cybulski, K. E.; Howlett, N. G., FANCP/SLX4: A Swiss army knife of DNA interstrand crosslink repair. *Cell Cycle* **2011**, *10* (11), 1757-1763.
9. Ke, B.; Ye, K.; Cheng, S., ALKBH2 Inhibition Alleviates Malignancy in Colorectal Cancer by Regulating BMI1-Mediated Activation of NF- κ B Pathway. *World J. Surg. Oncol.* **2020**, *18* (1), 328.
10. Suazo, K. F.; Park, K.-Y.; Distefano, M. D., A Not-So-Ancient Grease History: Click Chemistry and Protein Lipid Modifications. *Chem. Rev.* **2021**, *121* (12), 7178-7248.
11. Wang, B.; Feig, M.; Cukier, R. I.; Burton, Z. F., Computational Simulation Strategies for Analysis of Multisubunit RNA Polymerases. *Chem. Rev.* **2013**, *113* (11), 8546-8566.
12. Harrison, J. G.; Zheng, Y. B.; Beal, P. A.; Tantillo, D. J., Computational Approaches to Predicting the Impact of Novel Bases on RNA Structure and Stability. *ACS Chem. Biol.* **2013**, *8* (11), 2354-2359.

13. Gossett, J. J.; Harvey, S. C., Computational Screening and Design of DNA-Linked Molecular Nanowires. *Nano Lett.* **2011**, *11* (2), 604-608.
14. Nørby, M. S.; Steinmann, C.; Olsen, J. M. H.; Li, H.; Kongsted, J., Computational Approach for Studying Optical Properties of DNA Systems in Solution. *J. Chem. Theory Comput.* **2016**, *12* (10), 5050-5057.
15. Aronica, P. G. A.; Reid, L. M.; Desai, N.; Li, J.; Fox, S. J.; Yadahalli, S.; Essex, J. W.; Verma, C. S., Computational Methods and Tools in Antimicrobial Peptide Research. *J. Chem. Inf. Model.* **2021**, *61* (7), 3172-3196.
16. Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz, K. M.; Metz, T. O.; Nunez, J. R.; Tantillo, D. J.; Wang, L.-P.; Wang, S.; Renslow, R. S., Quantum Chemistry Calculations for Metabolomics. *Chem. Rev.* **2021**, *121* (10), 5633-5670.
17. Marrink, S. J.; Corradi, V.; Souza, P. C. T.; Ingólfsson, H. I.; Tieleman, D. P.; Sansom, M. S. P., Computational Modeling of Realistic Cell Membranes. *Chem. Rev.* **2019**, *119* (9), 6184-6226.
18. Baum, Z. J.; Yu, X.; Ayala, P. Y.; Zhao, Y.; Watkins, S. P.; Zhou, Q., Artificial Intelligence in Chemistry: Current Trends and Future Directions. *J. Chem. Inf. Model.* **2021**, *61* (7), 3197-3212.
19. Zhao, J.; Bacolla, A.; Wang, G.; Vasquez, K. M., Non-B DNA Structure-Induced Genetic Instability and Evolution. *Cell. Mol. Life Sci.* **2010**, *67* (1), 43-62.
20. Gómez-Bombarelli, R.; González-Pérez, M.; Calle, E.; Casado, J., Reactivity of Mucohalic Acids in Water. *Water Res.* **2011**, *45* (2), 714-720.
21. Guengerich, F. P.; Kim, D. H., Enzymic Oxidation of Ethyl Carbamate to Vinyl Carbamate and Its Role as an Intermediate in the Formation of 1,N6-Ethenoadenosine. *Chem. Res. Toxicol.* **1991**, *4* (4), 413-421.
22. Jiao, Z.; Dong, Y.; Chen, Q., Ethyl Carbamate in Fermented Beverages: Presence, Analytical Chemistry, Formation Mechanism, and Mitigation Proposals. *Compr. Rev. Food Sci. Food Saf.* **2014**, *13* (4), 611-626.
23. Basu, A. K.; Wood, M. L.; Niedernhofer, L. J.; Ramos, L. A.; Essigmann, J. M., Mutagenic and Genotoxic Effects of Three Vinyl Chloride-Induced DNA Lesions: 1,N6-Ethenoadenine, 3,N4-Ethenocytosine, and 4-amino-5-(imidazol-2-yl)imidazole. *Biochem.* **1993**, *32* (47), 12793-12801.
24. Choudhury, S.; Adhikari, S.; Cheema, A.; Roy, R., Evidence of Complete Cellular Repair of 1,N6-Ethenoadenine, a Mutagenic and Potential Damage for Human Cancer, Revealed by a Novel Method. *Mol. Cell. Biochem.* **2008**, *313* (1), 19-28.

25. Chen, H.-J. C.; Chiang, L.-C.; Tseng, M.-C.; Zhang, L. L.; Ni, J.; Chung, F.-L., Detection and Quantification of 1,N6-Ethenoadenine in Human Placental DNA by Mass Spectrometry. *Chem. Res. Toxicol.* **1999**, *12* (12), 1119-1126.
26. Liu, M.; Xu, M.; Lee, J. K., The Acidity and Proton Affinity of the Damaged Base 1,N6-Ethenoadenine in the Gas Phase Versus in Solution: Intrinsic Reactivity and Biological Implications. *J. Org. Chem.* **2008**, *73* (15), 5907-5914.
27. Yi, C.; Chen, B.; Qi, B.; Zhang, W.; Jia, G.; Zhang, L.; Li, C. J.; Dinner, A. R.; Yang, C.-G.; He, C., Duplex Interrogation by a Direct DNA Repair Protein in Search of Base Damage. *Nat. Struct. Mol. Biol.* **2012**, *19* (7), 671-676.
28. Müller, T. A.; Hausinger, R. P., Chapter 8: AlkB and Its Homologues – DNA Repair and Beyond. In *2-Oxoglutarate-Dependent Oxygenases*, The Royal Society of Chemistry: 2015; pp 246-262.
29. Zheng, G.; Fu, Y.; He, C., Nucleic Acid Oxidation in DNA Damage Repair and Epigenetics. *Chem. Rev.* **2014**, *114* (8), 4602-4620.
30. Lau Albert, Y.; Wyatt Michael, D.; Glassner Brian, J.; Samson Leona, D.; Ellenberger, T., Molecular Basis for Discriminating Between Normal and Damaged Bases by the Human Alkyladenine Glycosylase, AAG. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (25), 13573-13578.
31. Rutledge, L. R.; Wetmore, S. D., Modeling the Chemical Step Utilized by Human Alkyladenine DNA Glycosylase: A Concerted Mechanism Aids in Selectively Excising Damaged Purines. *J. Am. Chem. Soc.* **2011**, *133* (40), 16258-16269.
32. Caffrey, P. J.; Kher, R.; Bian, K.; Li, D.; Delaney, S., Comparison of the Base Excision and Direct Reversal Repair Pathways for Correcting 1,N6-Ethenoadenine in Strongly Positioned Nucleosome Core Particles. *Chem. Res. Toxicol.* **2020**, *33* (7), 1888-1896.
33. Johannessen, T.-C. A.; Prestegarden, L.; Grudic, A.; Hegi, M. E.; Tysnes, B. B.; Bjerkvig, R., The DNA Repair Protein ALKBH2 Mediates Temozolomide Resistance in Human Glioblastoma Cells. *Neuro-Oncol.* **2013**, *15* (3), 269-278.
34. Wu, S.-s.; Xu, W.; Liu, S.; Chen, B.; Wang, X.-l.; Wang, Y.; Liu, S.-f.; Wu, J.-q., Down-Regulation of ALKBH2 Increases Cisplatin Sensitivity in H1299 Lung Cancer Cells. *Acta Pharmacol. Sin.* **2011**, *32* (3), 393-398.
35. Gutierrez, R.; O'Connor, T. R., DNA Direct Reversal Repair and Alkylating Agent Drug Resistance. *Cancer Drug Resist.* **2021**, *4* (2), 414-423.
36. Dixon, M.; Woodrick, J.; Gupta, S.; Karmahapatra, S. K.; Devito, S.; Vasudevan, S.; Dakshanamurthy, S.; Adhikari, S.; Yenugonda, V. M.; Roy, R., Naturally Occurring

- Polyphenol, Morin Hydrate, Inhibits Enzymatic Activity of N-Methylpurine DNA Glycosylase, a DNA Repair Enzyme with Various Roles in Human Disease. *Bioorg. Med. Chem.* **2015**, *23* (5), 1102-1111.
37. Crosbie, P. A. J.; Watson, A. J.; Agius, R.; Barber, P. V.; Margison, G. P.; Povey, A. C., Elevated N3-Methylpurine-DNA Glycosylase DNA Repair Activity is Associated with Lung Cancer. *Mutat. Res-Find Mol. M.* **2012**, *732* (1), 43-46.
 38. Kellum, A. H.; Qiu, D. Y.; Voehler, M. W.; Martin, W.; Gates, K. S.; Stone, M. P., Structure of a Stable Interstrand DNA Cross-Link Involving a β -N-Glycosyl Linkage Between an N6-dA Amino Group and an Abasic Site. *Biochem.* **2021**, *60* (1), 41-52.
 39. Varela, J. G.; Pierce, L. E.; Guo, X.; Price, N. E.; Johnson, K. M.; Yang, Z.; Wang, Y.; Gates, K. S., Interstrand Cross-Link Formation Involving Reaction of a Mismatched Cytosine Residue with an Abasic Site in Duplex DNA. *Chem. Res. Toxicol.* **2021**, *34* (4), 1124-1132.
 40. Johnson, K. M.; Price, N. E.; Wang, J.; Fekry, M. I.; Dutta, S.; Seiner, D. R.; Wang, Y.; Gates, K. S., On the Formation and Properties of Interstrand DNA–DNA Cross-Links Forged by Reaction of an Abasic Site with the Opposing Guanine Residue of 5'-CAP Sequences in Duplex DNA. *J. Am. Chem. Soc.* **2013**, *135* (3), 1015-1025.
 41. Hashimoto, S.; Anai, H.; Hanada, K., Mechanisms of Interstrand DNA Crosslink Repair and Human Disorders. *Genes Environ.* **2016**, *38* (1), 9.
 42. Deans, A. J.; West, S. C., DNA Interstrand Crosslink Repair and Cancer. *Nat. Rev. Cancer* **2011**, *11* (7), 467-480.
 43. Romero, R. M.; Rojsitthisak, P.; Haworth, I. S., DNA Interstrand Crosslink Formation by Mechlorethamine at a Cytosine–Cytosine Mismatch Pair: Kinetics and Sequence Dependence. *Arch. Biochem.* **2001**, *386* (2), 143-153.
 44. Guainazzi, A.; Campbell, A. J.; Angelov, T.; Simmerling, C.; Schäfer, O. D., Synthesis and Molecular Modeling of a Nitrogen Mustard DNA Interstrand Crosslink. *Chem. Eur. J.* **2010**, *16* (40), 12100-12103.
 45. Upadhyaya, A.; Nath, S.; Kumar, S., Force-Induced Rupture of Double-Stranded DNA in the Absence and Presence of Covalently Bonded Anti-Tumor Drugs: Insights from Molecular Dynamics Simulations. *J. Chem. Phys.* **2018**, *148* (21), 215105.
 46. Bueren-Calabuig, J. A.; Giraudon, C.; Galmarini, C. M.; Egly, J. M.; Gago, F., Temperature-Induced Melting of Double-Stranded DNA in the Absence and Presence of Covalently Bonded Antitumour Drugs: Insight from Molecular Dynamics Simulations. *Nucleic Acids Res.* **2011**, *39* (18), 8248-8257.

47. Spiegel, K.; Magistrato, A., Modeling Anticancer Drug–DNA Interactions via Mixed QM/MM Molecular Dynamics Simulations. *Org. Biomol. Chem.* **2006**, *4* (13), 2507-2517.
48. Grüber, R.; Dumont, É., DFT Investigation of the Formation of Linear Aminols as the First Step Toward the Induction of Oxidatively Generated Interstrand Cross-Link DNA Lesions. *Theor. Chem. Acc.* **2015**, *134* (3), 26.
49. Patel, C.; Dršata, T.; Lankaš, F.; Dumont, E., Structure, Dynamics, and Interactions of a C4'-Oxidized Abasic Site in DNA: A Concomitant Strand Scission Reverses Affinities. *Biochem.* **2013**, *52* (45), 8115-8125.
50. Bignon, E.; Dršata, T.; Morell, C.; Lankaš, F.; Dumont, E., Interstrand Cross-Linking Implies Contrasting Structural Consequences for DNA: Insights from Molecular Dynamics. *Nucleic Acids Res.* **2017**, *45* (4), 2188-2195.
51. Price, N. E.; Johnson, K. M.; Wang, J.; Fekry, M. I.; Wang, Y.; Gates, K. S., Interstrand DNA–DNA Cross-Link Formation Between Adenine Residues and Abasic Sites in Duplex DNA. *J. Am. Chem. Soc.* **2014**, *136* (9), 3483-3490.

Chapter 2. A Structural Explanation of Sequence Effects on ALKBH2-mediated Interstrand Crosslink Formation

2.1 Introduction

The aim of this thesis chapter is to provide a structural explanation for the effects of DNA sequence context on the experimentally quantified yield of ICL due to attempted ϵ A repair by ALKBH2. To the best of my knowledge, there has not been documentation of ICL formation as the result of attempted repair by a DRR enzyme before discovery by the Li lab at the University of Rhode Island. As mentioned in the introduction of this thesis, a multitude of sequence contexts were explored by my collaborators using *in vitro* enzymatic reactions, electrophoresis, and LC-MS studies. As ALKBH2 repair of ϵ A is known to generate an epoxide intermediate, the detected ICLs are proposed to form as the result of the epoxide intermediate prematurely leaving the active site of ALKBH2, being flipped into the DNA duplex, and aminolysis of the epoxide intermediate with the amino groups of A, G, and C acting as the nucleophiles.¹ Interestingly, a trend of ICL yield was detected based on the relative helical locations of A, G, or C to ϵ A, which represents where the epoxide intermediate would subsequently be formed.

Twelve sequences have been studied by my collaborators through experimental techniques (Table 2.1 and 2.2), each of which contain ϵ A flanked by A to the 5' and 3' directions. With the exception of sequence 1, a singular A, G, C, uracil (U), or purine (P*) nucleotide was placed on the complementary strand in order to understand: whether an amino group is necessary to form an ICL, and the effect that the position of A, G, and C has on the presence of the observed ICL. As it was observed that the presence of U and P* on the complementary strand did not result in the formation of ICLs, an exocyclic amino group is deemed necessary for ICL formation (Figure 2.1).

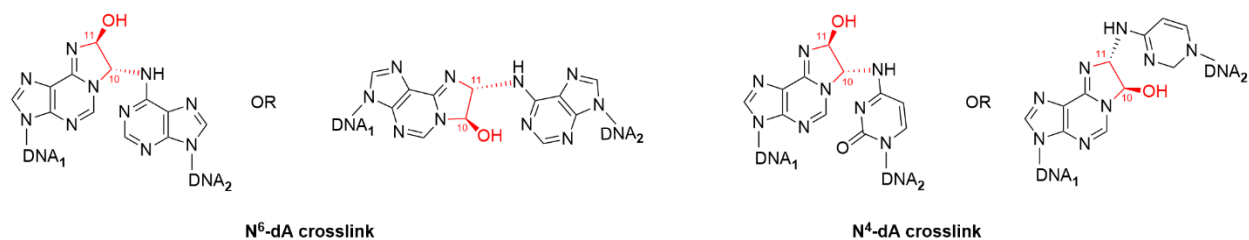


Figure 2.1: Epoxide intermediate-derived ICL structures proposed by this work, with DNA₁ being the ϵ A lesion containing strand and DNA₂ being the complementary strand.

DNA duplexes with complementary pairing throughout the helix, or a single mismatch at the site complementary to ϵ A were also studied using experimental techniques, as this lesion was observed to cause A:A, A:C, and A:G mismatches.¹ The *pre-in silico* analysis of the sequences revealed that the yield of ICL varied when the nucleotides flanking ϵ A were manipulated (Table 2.1).² However, more information is required in terms of: 1) how the structure of DNA and the position of various nucleobases affect which sequence contexts result in ICL formation, 2) which epoxide carbon is primarily attacked to form ICLs, and 3) why G does not result in ICLs when in close proximity ϵ A. This work utilizes MD simulations of DNA duplexes containing the epoxide intermediate in sequences analogous to those studied by the Li lab (Tables 2.1 and 2.2) to answer the above questions.

Table 2.1: Base screening experiments for ϵ A-induced crosslink by ALKBH2 repair. N.D. indicates not detectable under the reaction conditions.²

entry	sequence	crosslink yield (%)
1	5'-AXA-3' 3'-TTT-5'	N.D.
2	5'-AXA-3' 3'-ATT-5'	0.1±0.0
3	5'-AXA-3' 3'-TAT-5'	0.7±0.2
4	5'-AXA-3' 3'-TTA-5'	4.6±0.3
5	5'-AXA-3' 3'-TTA*-5'	N.D.
6	5'-AXA-3' 3'-CTT-5'	0.4±0.0
7	5'-AXA-3' 3'-TCT-5'	2.2±0.2
8	5'-AXA-3' 3'-TTC-5'	3.2±0.6
9	5'-AXA-3' 3'-TUT-5'	N.D.
10	5'-AXA-3' 3'-GTT-5'	N.D.
11	5'-AXA-3' 3'-TGT-5'	N.D.
12	5'-AXA-3' 3'-TTG-5'	N.D.

Table 2.2: Yield of interstrand crosslink generated from ϵ A repaired by ALKBH2 in well-matched and mismatched sequences. N.D. indicates not detectable under the reaction conditions.²

entry	sequence	crosslink yield (%)	entry	sequence	crosslink yield (%)
1	5'-TXT-3' 3'-ATA-5'	1.3±0.0	11	5'-CXC-3' 3'-GTG-5'	N.D.
2	5'-AXT-3' 3'-TTA-5'	1.3±0.1	12	5'-GXC-3' 3'-CTG-5'	N.D.
3	5'-CXT-3' 3'-GTA-5'	1.7±0.2	13	5'-TXG-3' 3'-ATC-5'	<0.1
4	5'-GXT-3' 3'-CTA-5'	0.5±0.1	14	5'-AXG-3' 3'-TTC-5'	N.D.
5	5'-TXA-3' 3'-ATT-5'	<0.1	15	5'-CXG-3' 3'-GTC-5'	<0.1
6	5'-AXA-3' 3'-TTT-5'	N.D.	16	5'-GXG-3' 3'-CTC-5'	<0.1
7	5'-CXA-3' 3'-GTT-5'	N.D.	17	5'-TXT-3' 3'-ACA-5'	2.6±0.1
8	5'-GXA-3' 3'-CTT-5'	N.D.	18	5'-TXT-3' 3'-AAA-5'	5.7±0.2
9	5'-TXC-3' 3'-ATG-5'	N.D.	19	5'-TXT-3' 3'-AGA-5'	1.1±0.1
10	5'-AXC-3' 3'-TTG-5'	N.D.			

2.2 Computational Methodology

2.2.1 MD Model Construction

Prior to pre-production, DNA complementary duplexes corresponding to all experimentally considered fully-matched sequences were built using the Nucleic Acid Builder (NAB) module of AMBER18. The epoxide intermediate was incorporated into each helix using GaussView 6.0.16³, while mismatched base pairs were introduced into the relevant mismatched helices using PyMOL.⁴⁻⁶ Since different combinations of nucleotide *anti* and *syn* glycosidic orientations have been identified in crystal structures of DNA containing A:A, A:C, and A:G

mismatches,⁷⁻¹² molecular mechanics (MM) minimizations were performed on all conformational combinations for the mismatched A:A, A:G, and A:C pairs in each sequence context where the mismatch flanked the epoxide intermediate. These minimizations were performed in a TIP3P water box where the solvent was at least 8 Å away from the edge of the box and similar force field parameters used for the MD simulations were used to describe DNA (described below). Following MM minimization, all atoms were stripped from the model except for the mismatched and two flanking nucleobase pairs. The positions of the hydrogen atoms added at the sugar–phosphate truncation point in each model were optimized with the same MM force field. Single-point energy calculations were then performed in implicit water using SMD-B3LYP-D3(BJ)/6-31+G(d,p) to determine the most energetically preferred base pairing conformations, which were used in subsequent MD models. See Table A–1 for a summary of the base pairing models considered, relative energies and identification of the final pairing combinations considered in the subsequent MD simulations.

The AMBER OL15 force field was used to describe canonical DNA nucleotides, with parameters for the epoxide intermediate supplemented by generalized AMBER force field (GAFF) parameters. Charges compatible with AMBER were generated for the epoxide intermediate using the PyRED server.¹³⁻¹⁶ Each modified duplex was solvated in a rectangular TIP4-Ew water box, with the solute at least 12 Å from the edge of the box. Each system was then neutralized with Na⁺ ions, and Na⁺ and Cl⁻ ions were subsequently added to yield an overall salt concentration of 150 mM in order to complement physiological conditions. MD input models for the production and pre-production steps were further prepared using the tLEaP module of AMBER18.

2.2.2 MD Simulation Protocol

Each system was minimized in four steps. First, 5000 cycles of minimization were performed, including 2500 steps of the steepest descent algorithm followed by 2500 steps of the conjugate gradient algorithm, while a restraint was applied to all DNA atoms using a 100 kcal/(mol \AA^2) force constant. Subsequently, the same minimization steps were performed for the substrate, while the restraint was applied to the water and ions. Next, 3000 minimization steps were completed, including 1500 steps of the steepest descent method and 1500 steps of the conjugate gradient method, while restraining the DNA with 100 kcal/(mol \AA^2). In the final minimization step, no restraints were applied, and the entire system was minimized using 2000 steps of the steepest descent algorithm and 1000 steps of the conjugate gradient algorithm. Subsequently, each minimized system was heated in six segments from 0 to 310 K over 1200 ps. Equilibration was then performed with a 20 kcal/(mol \AA^2) restraint initially imposed on the DNA at a constant temperature (Langevin thermostat) of 310 K, which was sequentially decreased (20, 15, 10, 5, and 1.5 kcal/(mol \AA^2)). During pre-production, a nonbonded cutoff of 10.0 \AA was used in every step.

Finally, 1.1 μs MD production simulations were performed on each system in triplicate, with the first 100 ns of each simulation treated as an extra equilibration step. The system was considered at a constant pressure of 1.0 bar using isotropic position scaling. 2.0 ps were allowed for pressure relaxation and the nonbonded cutoff remained at 10 \AA . The SHAKE algorithm was employed, and a distance restraint was enforced on the heavy atoms participating in Watson-Crick hydrogen bonding at the terminal ends of the duplex.¹⁷ The temperature was held at 310 K using Langevin dynamics with a collision frequency of 3.0. A step size of 0.002 ps was used throughout. The atomic coordinates were saved every 5000 steps for analysis.

2.2.3 MD Analysis

Trajectory visualization was performed using PyMOL, while all quantitative analysis was conducted using the CPPTRAJ module of AMBERTOOLS 20. Among the 100,000 frames collected for each system, every fifth frame was analyzed (a total of 20,000 frames per trajectory). Root-mean square deviations (RMSD) were calculated using DNA backbone atoms (P, O3', O5', C3', C4', and C5' atoms of each nucleotide), and were evaluated with respect to the first frame of the corresponding simulation to ensure DNA conformational stability (Table A–2). To determine the frequency of the reactive conformation, the distance between a given nucleophilic amine in a neighboring nucleobase and an electrophilic epoxide carbon atom (C10 or C11; denoted $r(N_{\text{nucl}}C_{\text{elec}})$), as well as the angle between the nucleophilic site, electrophilic carbon, and leaving oxygen ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$) were measured (Figure 2.2). A reactive conformation was defined to occur when $r(N_{\text{nucl}}C_{\text{elec}})$ is less than 3.3 Å, which represents the sum of the average van der Waals radii, and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ falls between 130 and 180°, which is consistent with the range of angles occupied by reactant complexes and transition states in a previous quantum mechanical study on the aminolysis of ethylene oxide by the amino groups of the canonical DNA nucleobases.¹⁸ The frequency of the reactive conformation (FRC) was reported as the number of identified reactive conformations compared to the total number of MD frames analyzed (Tables A–3–A–15). The average $r(N_{\text{nucl}}C_{\text{elec}})$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ during the adopted reactive conformation were calculated using the distance and angle commands in CPPTRAJ (Tables A–16–A–26). Representative structures of each DNA sequence in the reactive conformation were generated using the filter command (with the cutoffs consistent with the ones used to determine the reactive conformations). The cluster command in CPPTRAJ was used with the hieraggl

algorithm to cluster with respect to the RMSD of the DNA backbone atoms (P, O3', O5', C3', C4', and C5' atoms of each nucleotide), with one cluster requested.

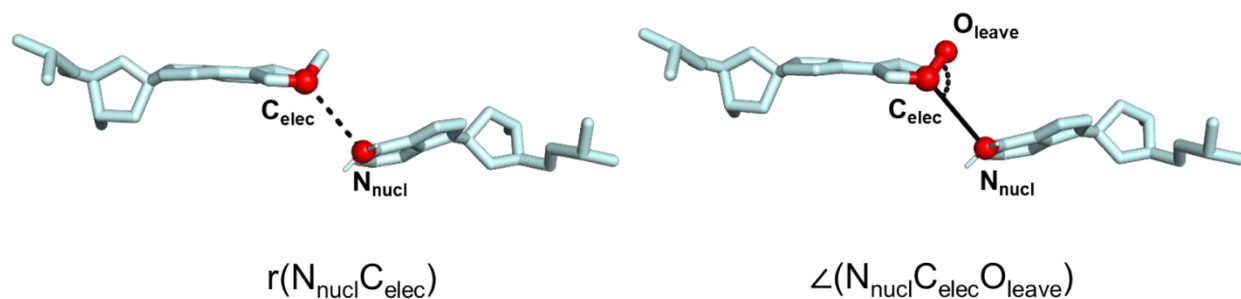


Figure 2.2: Geometric cutoffs used to identify reactive conformations, including the nucleophile–electrophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$) and the corresponding angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$).

2.3 Results

2.3.1 Primary Positions

As aminolysis of the epoxide intermediate by nearby exocyclic amino groups was suspected to be the cause of ICL formation, the relative orientation of nucleophiles (exocyclic amino nitrogen atoms) and electrophiles (epoxide carbon atoms) were necessary to analyze. Because the identity of an amino group offset by more than one base pair step relative to the epoxide did not appear to affect ICL formation (Table 2.3), only amino groups within the proximity of one base pair step were considered. In total, five nucleophilic positions were considered (Figure 2.3). Specifically, while the 5'-intrastrand position with respect to the epoxide (5' intra) were considered, the 3'-intrastrand position relative to the epoxide (3' intra), the complementary position to the epoxide (complementary), the 5'-interstrand position relative to the epoxide (5' inter), and 3'-interstrand position relative to the epoxide (3' inter) have also been

considered. To identify the positioning relative to the epoxide, the identity of the nucleotide, and atom of potential nucleophiles in each sequence context, the following format was used: $a b^{(Y)}$ (where a = complementary, 5' intra, 3' inter, 5' inter, or 3' intra, b = A, C, or G, and Y = N6, N4, or N2). Likewise, as an epoxide contains two electrophilic sites, the notation $X^{(Z)}$ (where X = the epoxide intermediate and Z = C10 or C11) was used to identify each electrophilic site. For example, 5' inter A^(N6) represents a potential ICL that involves the amino group of A in the 5' inter position, and X^(C10) represents a potential ICL that involves C10 of the epoxide intermediate.

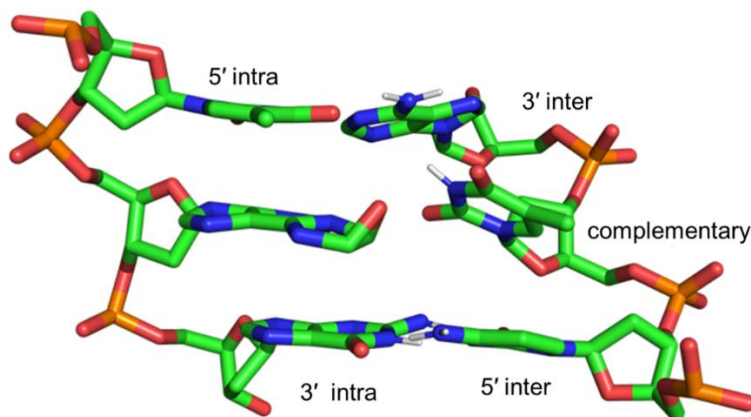


Figure 2.3: The position of different nucleophilic sites in nucleobases neighboring the epoxide intermediate considered in the present work for a representative segment of DNA taken from the 5'-TXG-3'/3'-ATC-5' sequence context.

2.3.2 Base Screening Sequences

To identify the interstrand positions that can result in ICL formation (3' inter, complementary, and 5' inter), sequences that contained a single A, G, or C on the complementary strand were constructed and simulated (base screening sequences, Table 2.2²). Each of the base screening sequences contained A in the 5' intra and 3' intra positions. On the opposing strand, the

corresponding sequence were either 5'-YTT-3', 5'-TYT-3', or 5'-TTY-3' (Y = A, G, or C). When A, C, or G is in the 3' inter position, no evidence of ICLs was observed in the *in vitro* experiments. However, a small and almost negligible frequency of reactive conformations (FRC) of 0.001 was observed between C11 and 3' inter G^(N2) (Figure 2.4). These observations support the hypothesis that the residue in the 3' inter position does not play a significant (in comparison to the complementary and 5' inter positions) role in ICL formation with the epoxide. In the case of A^(N6) and C^(N4), the largest interstrand FRC values were observed when A^(N6) or C^(N4) was in the 5' inter position. When A^(N6) was in the 5' inter position the interstrand FRC was 0.007, with all interstrand reactive conformations formed between A^(N6) and X^(C11). The average angle and distance of reactivity for 5' inter A^(N6) were 3.2 Å and 127.9° (Figure 2.5). When A^(N6) was in the complementary position, the interstrand FRC was smaller with a value of 0.002, with all reactive conformations formed between X^(C10) and 5' inter A^(N6). The average distance and angle for reactive conformations formed between complementary A^(N6) and X^(C10) were 3.2 Å and 125.0°. For reactive conformations between 5' inter C^(N4) and X^(C11), the average reactive angle and distance was 0.007. The interaction between X^(C11) and 5' inter C^(N4) made up most of the total reactive conformations in this model, with an interstrand FRC of 0.006, and an average reactive distance and angle of 3.2 Å and 129.0°. The FRC between 5' inter C^(N4) in the 5' inter and X^(C10) was 0.001. When the primary nucleophile was complementary C^(N4), the total interstrand FRC was 0.001, and all reactive conformations were adopted between X^(C10) and complementary C^(N4). In addition, the average reactive angle and distance for reactive conformations between X^(C10) and complementary C^(N4) was 3.2 Å and 125.1° (Figure 2.6).

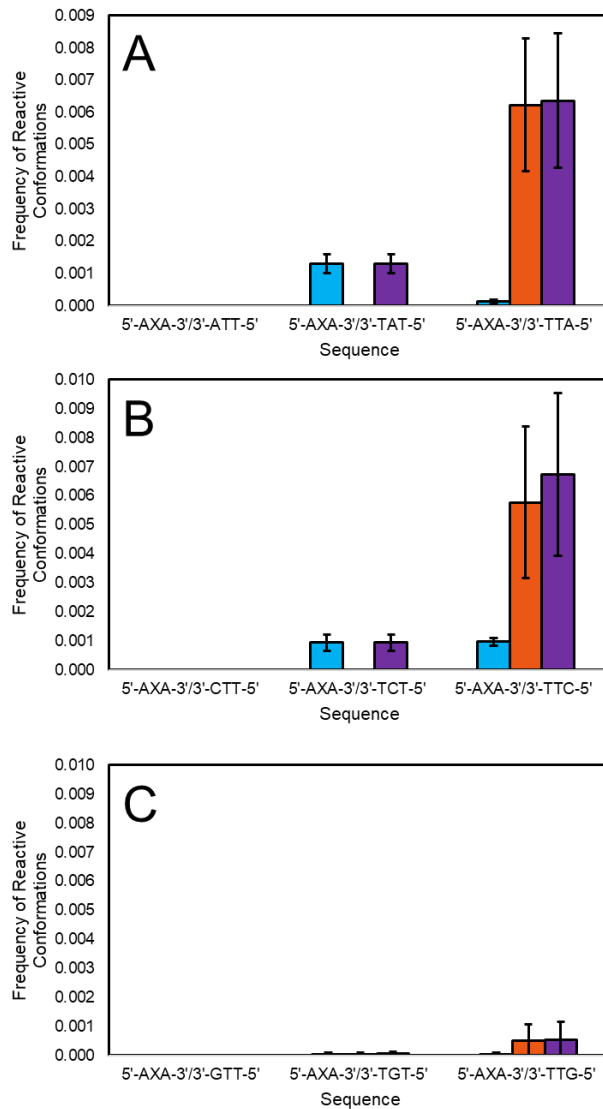


Figure 2.4: Frequency of reactive conformations suitable for interstrand ICL formation in the base-screening sequences, when (A) adenine, (B) cytosine, or (C) guanine is the primary interstrand nucleophile. Reactive conformations were considered for C10 (blue) or C11 (orange) in the epoxide, as well as the total for both sites (purple).

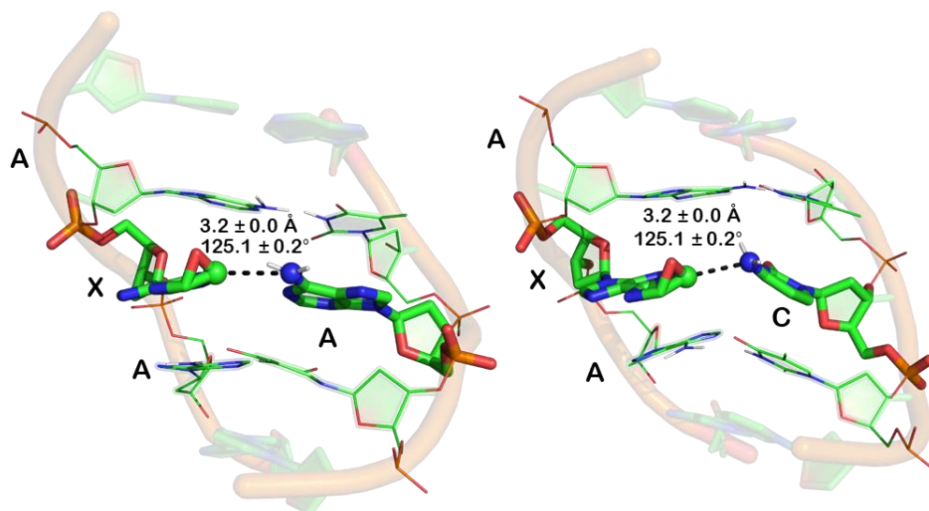


Figure 2.5: MD representative structures of the reactive conformations for crosslink formation between N6 of the complementary adenine (left) or N4 of the complementary cytosine (right) and C10 of the epoxide intermediate (X) for DNA duplexes containing the base screening sequences. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

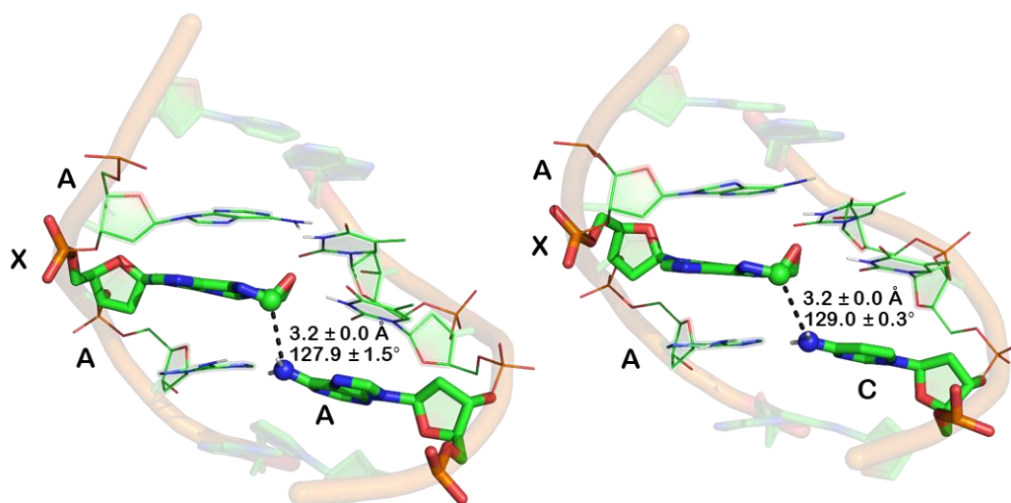


Figure 2.6: MD representative structures of the reactive conformations for crosslink formation between N6 of the 5'-interstrand adenine (left) or N4 of the 5'-interstrand cytosine (right) and C11 of the epoxide intermediate (X) for DNA duplexes containing the base screening sequences. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

2.3.3 Well-Matched Sequences

ICL formation was then investigated in sequence contexts that contained natural nucleotide pairing. Sixteen sequences were investigated and arranged into four groups based off of the identities of the 5' inter and 3' intra nucleotides. The 5'-XT-3'/3'-TA-5' group was composed of sequences 1–4 in Table 2.1², and contained a 5' inter A^(N6). This group of sequences were found to have the highest yields of ICL relative to the other groups. The 5'-XG-3'/3'-TC-5' sequences (Table 2.1², sequences 13–16) usually had some ICL detected, however quantitatively less than the 5'-XT-3'/3'-TA-5' group. These sequences contained a 3' intra G^(N2) and 5' inter C^(N4). The 5'-XA-3'/3'-TT-5' sequences (Table 2.1², 5–8) contained a 3' intra A^(N6), and usually did not result in any detected ICL. The 5'-XC-3'/3'-TG-5' group (Table 2.1², sequences 9–12) resulted in no ICL and featured a 5' inter G^(N2) and 3' intra C^(N4).

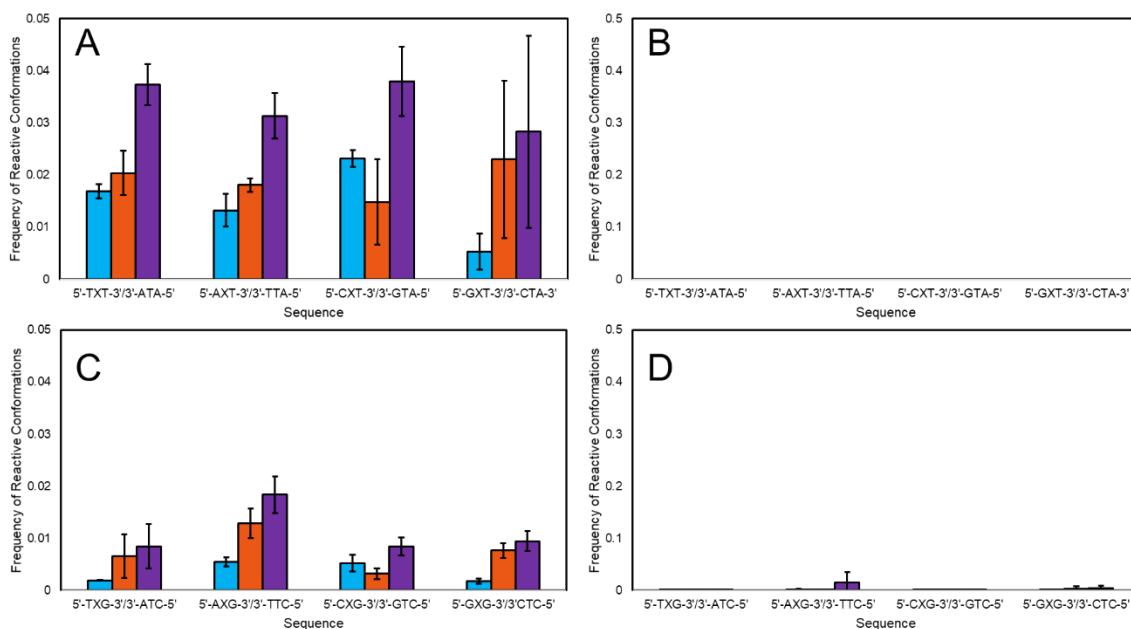


Figure 2.7: Frequency of reactive conformations suitable for interstrand (left) or intrastrand (right) crosslink formation for 5'-XT/5'-AT (A and B) or 5'-XG/5'-CT (C and D) sequence contexts. Reactive conformations involving C10 (blue) or C11 (orange) of the epoxide are provided, as well as the total for either site (purple).

2.3.3.1 5'-XT-3'/3'-TA-5' Sequence Context

In these sequences, the primary nucleophile was 5' inter A^(N6). Reactive conformations with 5' inter A^(N6) and X^(C10) had FRC values that ranged from 0.006 to 0.023 (Figure 2.7A), with the average reactive distances consistently being 3.2 Å, and average angles ranged from 136.3° to 138.3° (Figure 2.8). The reactive frequencies for 5' inter A^(N6) with X^(C11) ranged from 0.015 to 0.023. The average reactive distance between 5' inter A^(N6) and X^(C11) reactive conformations was 3.2 Å, and the average angle of the conformations ranged from 133.3° to 135.5° (Figure 2.9). The ranges in FRC did not reveal a clear preference between C10 and C11 for nucleophilic aminolysis of the epoxide by 5' inter A^(N6) in the 5'-XT-3'/3'-TA-5' group. Overall, the combined C10 and C11 FRC for the 5'-XT-3'/3'-TA-5' group ranged from 0.029 to 0.038. As there was no nucleobase with an exocyclic amino group in the 3' intra position, no reactive conformations for intrastrand crosslink formation could be detected in my simulations of these sequences.

2.3.3.2 5'-XG-3'/3'-TC-5' Sequence Context

The 5'-XG-3'/3'-TC-5' sequences featured 3' intra G^(N2) and 5' inter C^(N4) as the primary potential nucleophiles. For reactivity between 5' inter C^(N4) and the epoxide, FRC values did not support a preference for either X^(C10) or X^(C11), as the ranges of FRC values were from 0.002 to 0.006 for X^(C11), and 0.003 to 0.013 with X^(C10). Combined X^(C10) and X^(C11) FRC values ranged from 0.010 to 0.019. The average reactive conformational geometrical reactive angles ranged from 132.2° to 134.0° for X^(C10) (Figure 2.10), and 134.0° to 135.6° for X^(C11) (Figure 2.11).

Reactive conformations between the epoxide and 3' intra G^(N2) were substantially less frequent in the 5'-XG-3'/3'-TC-5' sequences, when compared to the quantity of reactive

conformations between the epoxide and 5' inter C^(N4). In particular, average FRC values ranged from 0.000 to 0.002 between 3' intra G^(N2) and X^(C10), 0.000 to 0.004 between 3' intra G^(N2) and X^(C11), and 0.001 to 0.005 between 3' intra G^(N2) and both carbon atoms of the epoxide intermediate.

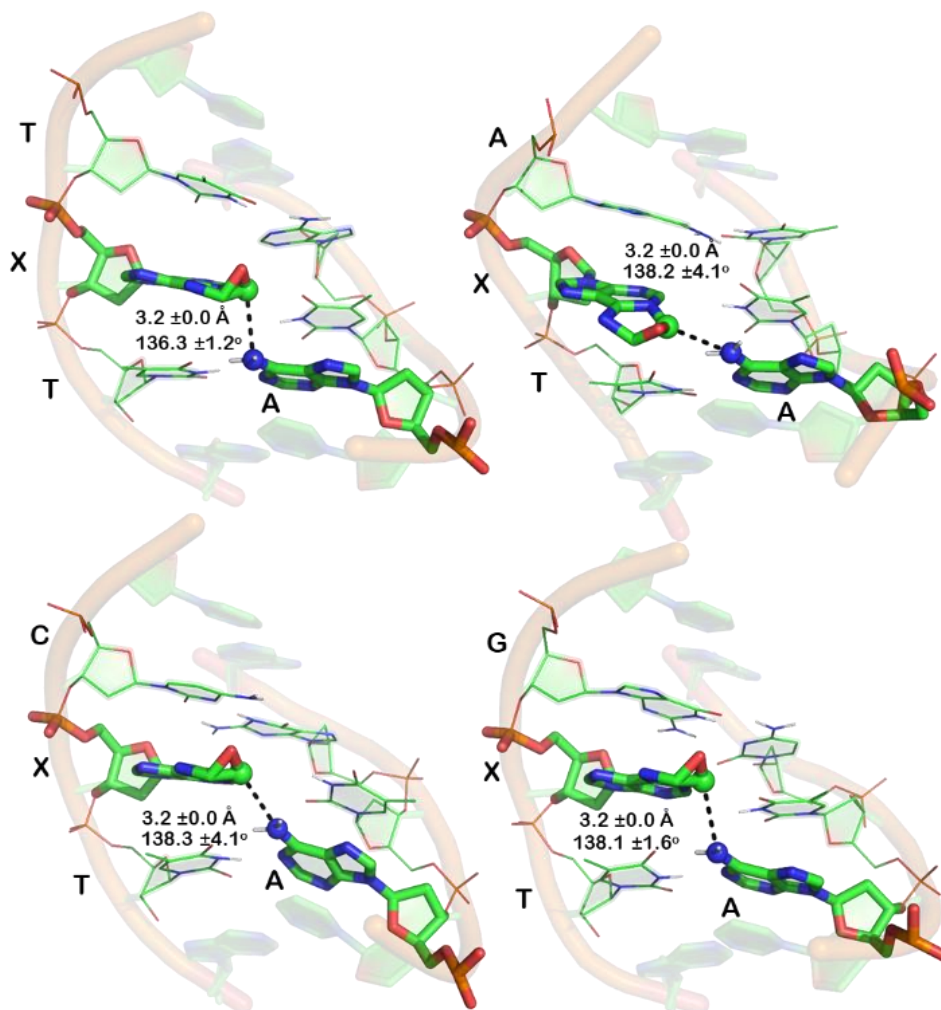


Figure 2.8: MD representative structures of the reactive conformations for crosslink formation between N6 of the 5'-interstrand adenine and C10 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XT/5'-AT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

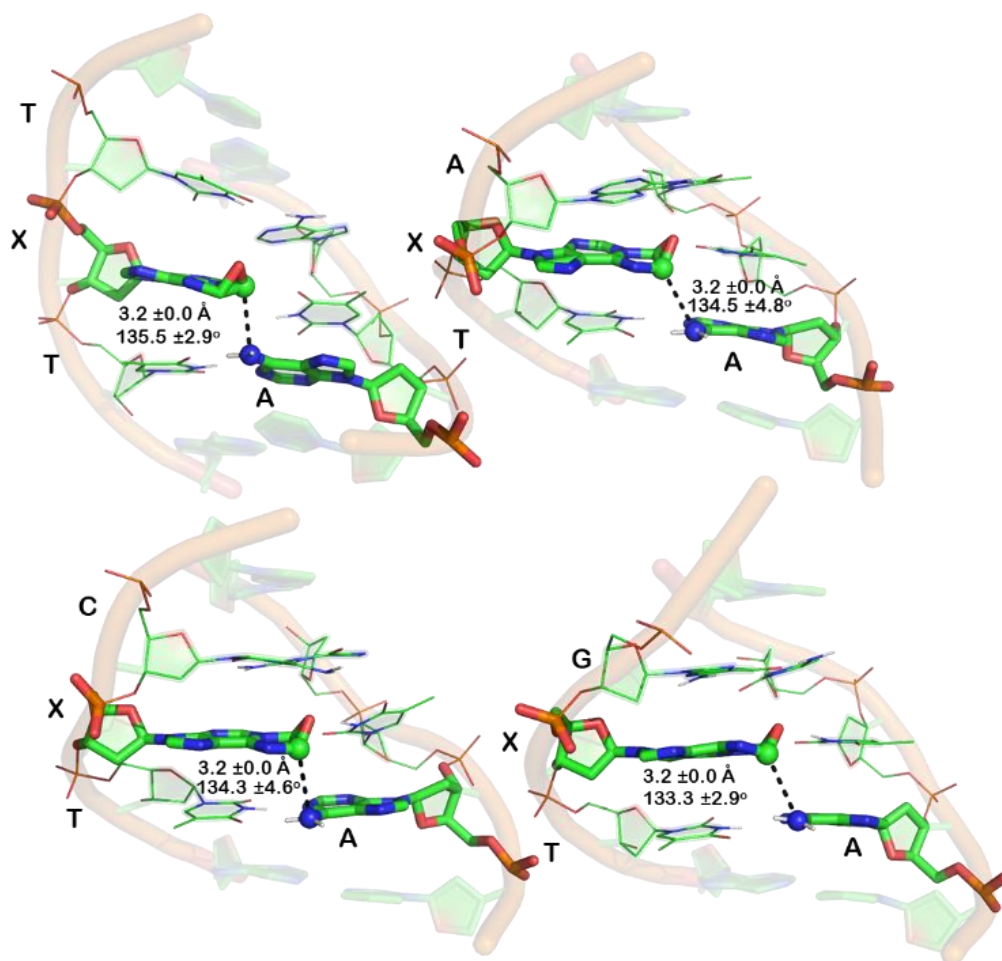


Figure 2.9: MD representative structures of the reactive conformations for crosslink formation between N6 of the 5'-interstrand adenine and C11 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XT/5'-AT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

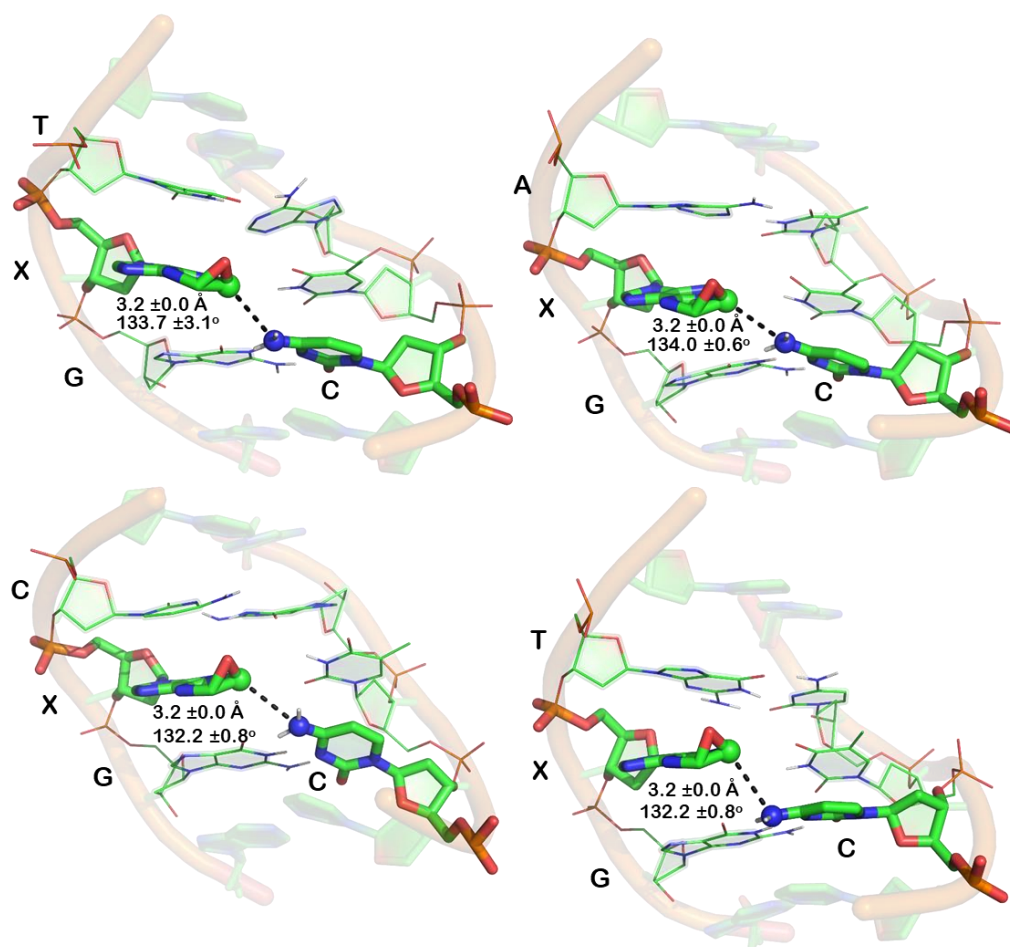


Figure 2.10: MD representative structures of the reactive conformations for crosslink formation between N4 of the 5'-interstrand cytosine and C10 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XG/5'-CT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

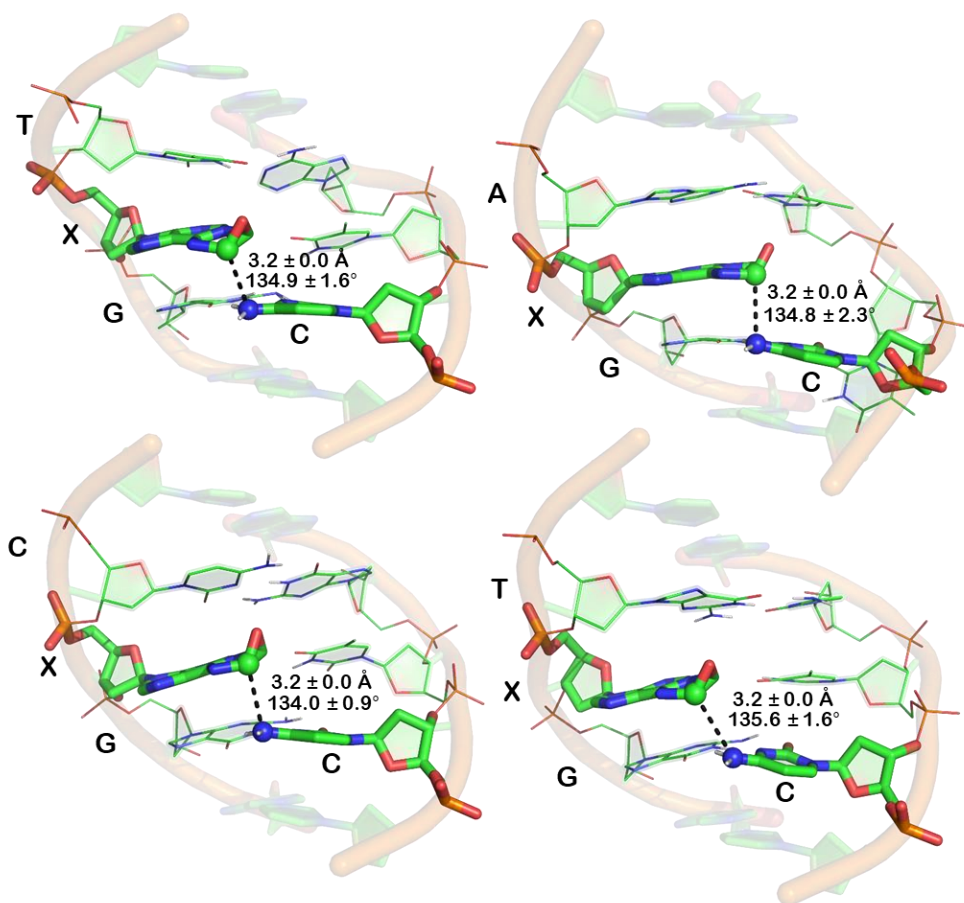


Figure 2.11: MD representative structures of the reactive conformations for crosslink formation between N4 of the 5'-interstrand cytosine and C11 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XG/5'-CT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nuc}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

2.3.3.3 5'-XA-3'/3'-TT-5' Sequence Context

While the 5'-XA-3'/3'-TT-5' did not have a nucleobase with an exocyclic amino group in the 5' inter position, a significant number of reactive conformations were identified between the epoxide intermediate, and 3' intra A^(N6) (Figure 2.12B). While FRC values between 3' intra A^(N6) and X^(C10) ranged from 0.017 to 0.040, there appeared to be a significant preference for reactive conformations formed between X^(C11) and 3' intra A^(N6), as FRC values ranged from 0.365 to

0.439. The overall FRC values between 3' intra A^(N6) and the epoxide intermediate ranged from 0.382 to 0.463. The average reactive geometrical angles ranged from 127.8° to 129.9° for reactive conformations between 3' intra A^(N6) (Figure 2.13) and X^(C10), and 139.6° to 143.4° for conformations between 3' intra A^(N6) and X^(C11). All average reactive distances remained at 3.2 Å (Figure 2.14).

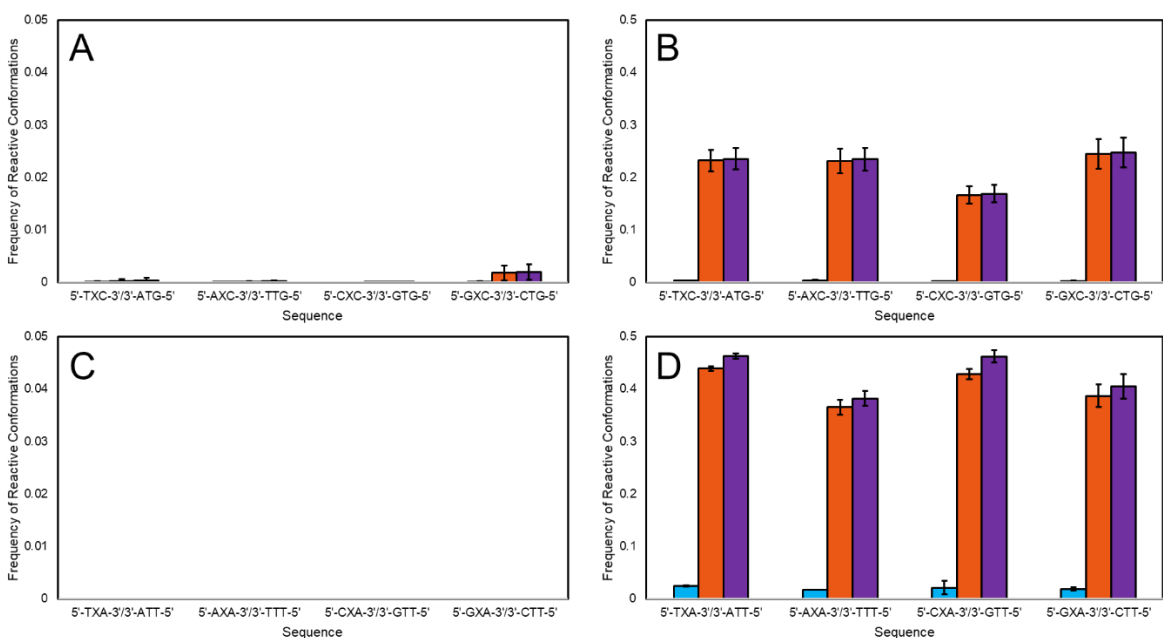


Figure 2.12: Frequency of reactive conformations suitable for interstrand (left) or intrastrand (right) crosslink formation for 5'-XC/5'-GT (A and B) or 5'-XA/5'-TT (C and D) sequence contexts. Reactive conformations involving C10 (blue) or C11 (orange) of the epoxide are provided, as well as the total for either site (purple).

2.3.3.4 5'-XC-3'/3'-TG-5' Sequence Context

As shown in Figure 2.12A, interstrand reactive conformations were not observed in my simulations of these sequences, with reactive conformations between X^(C10) and 5' inter G^(N2) remaining at 0.000 for the entirety of each of the simulations in this group. Reactive

conformations between 5' inter G^(N2) and X^(C11) ranged from 0.000 to 0.002. Intrastrand reactive conformations, however, were significantly more frequent with reactive conformations between 3' intra C^(N4) and X^(C10) ranging from 0.017 to 0.040, with the average reactive angle ranging from 128.0° to 131.3° (Figure 2.15). The FRC between 3' intra C^(N4) and X^(C11) ranged from 0.167 to 0.245, with the average reactive angle ranging from 146.3° to 148.5° (Figure 2.16). These values pointed to a preference for C11 attack for intrastrand crosslink formation between the epoxide intermediate and C. The overall FRC values for this group ranged from 0.169 to 0.247 and all average reactive distances remained at 3.2 Å.

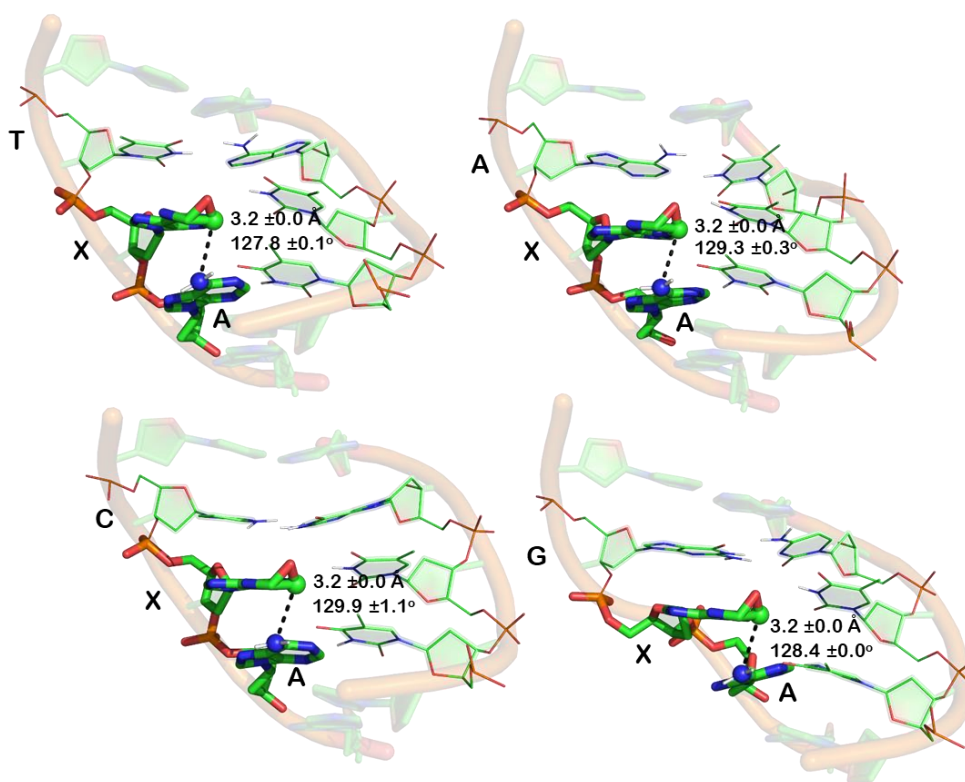


Figure 2.13: MD representative structures of the reactive conformations for crosslink formation between N6 of the 3'-intrastrand adenine and C10 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XA/5'-TT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

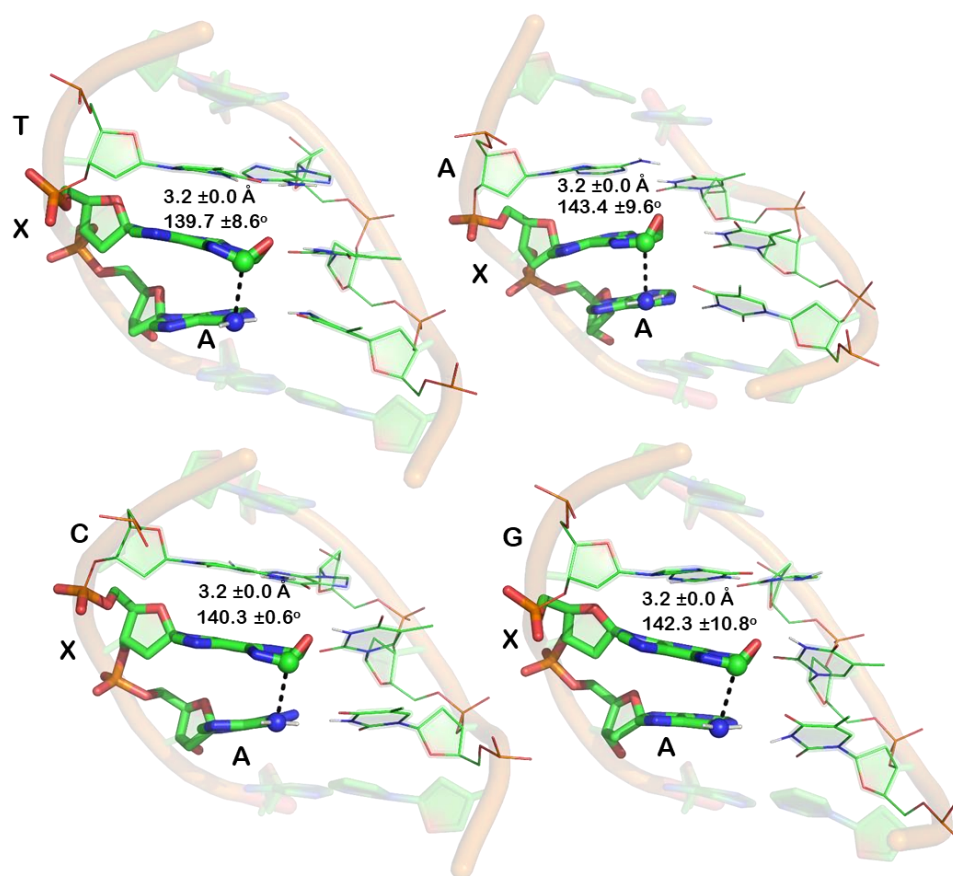


Figure 2.14: MD representative structures of the reactive conformations for crosslink formation between N6 of the 3'-intrastrand adenine and C11 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XA/5'-TT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nuc}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

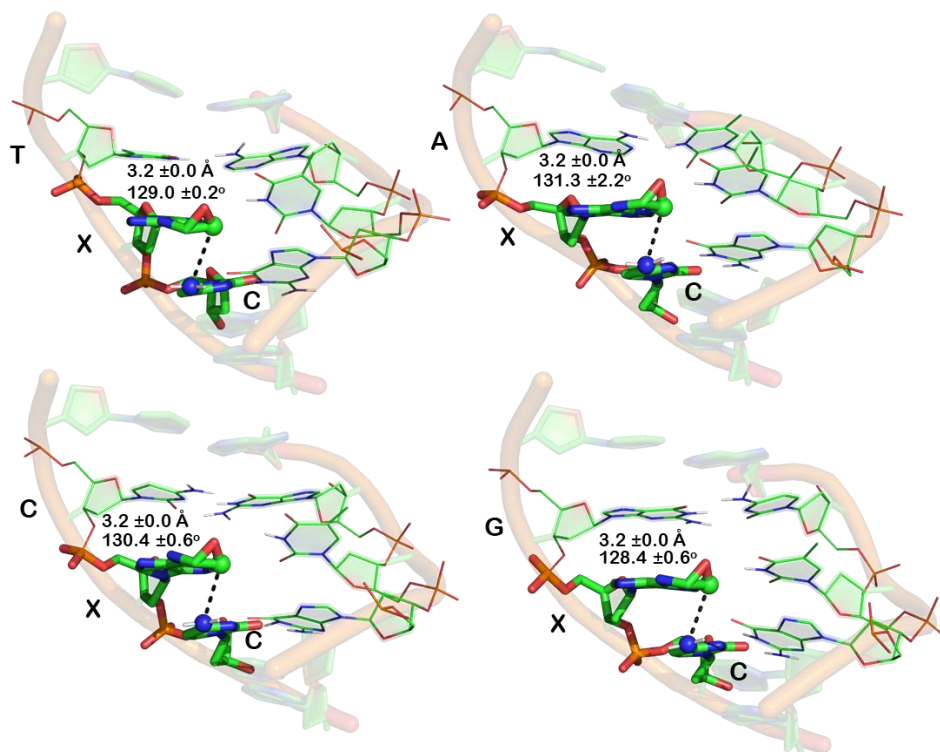


Figure 2.15: MD representative structures of the reactive conformations for crosslink formation between N4 of the 3'-intrastrand cytosine and C10 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XC/5'-GT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

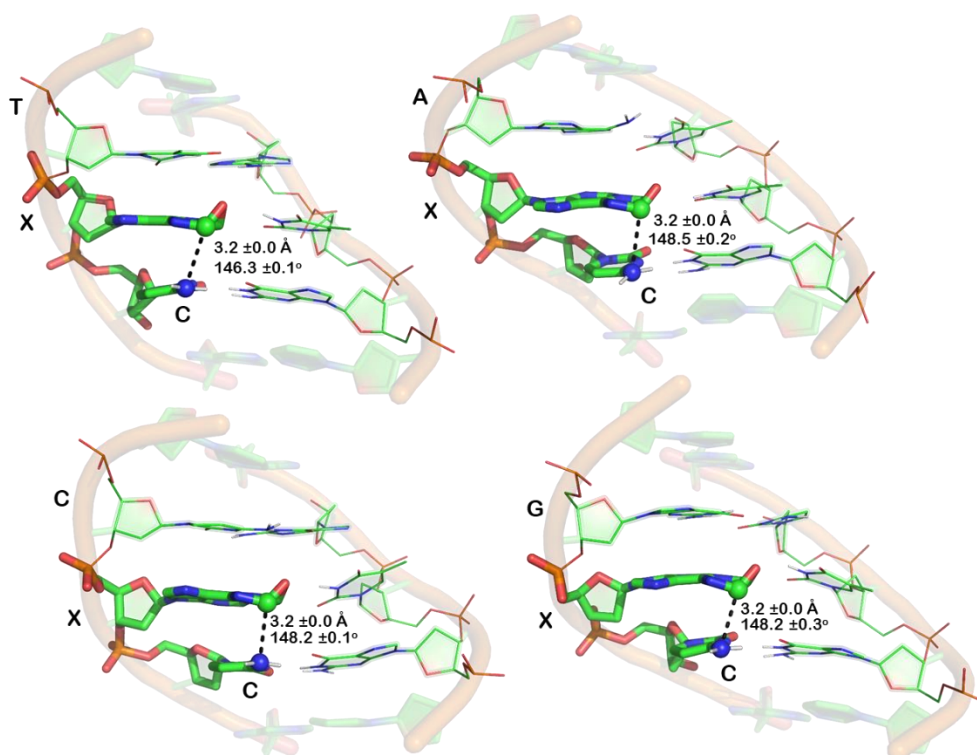


Figure 2.16: MD representative structures of the reactive conformations for crosslink formation between N4 of the 3'-intrastrand cytosine and C11 of the epoxide intermediate (X) for DNA duplexes containing the 5'-XC/5'-GT sequence context. Average electrophile–nucleophile distance ($r(N_{\text{nuc}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

2.3.4 Mismatched Sequences

While each mismatched sequence contained a 5' inter A^(N6) similar to sequence 1 (Table 2.1²), each mismatched sequence had a potential nucleophile in the complementary position relative to the epoxide, contrary to the well-matched sequences, which all contained T at the complementary position. In the mismatched groups, reactive conformations between the 5' inter A^(N6) and X^(C11) were frequently formed with FRC values ranging from 0.011 to 0.035 (Figure 2.17), with average reactive angle values ranging from 131.3° to 143.1° (Figure 2.18). It appeared that the addition of a complementary C or A helps position the epoxide to better

accommodate C11 attack by the 5' inter A^(N6). The reactive conformations between 5' inter A^(N6) and X^(C10) of the epoxide were also frequently formed with FRC values ranging from 0.020 to 0.033, with average reactive angle values ranging from 133.8° to 140.3° (Figure 2.18).

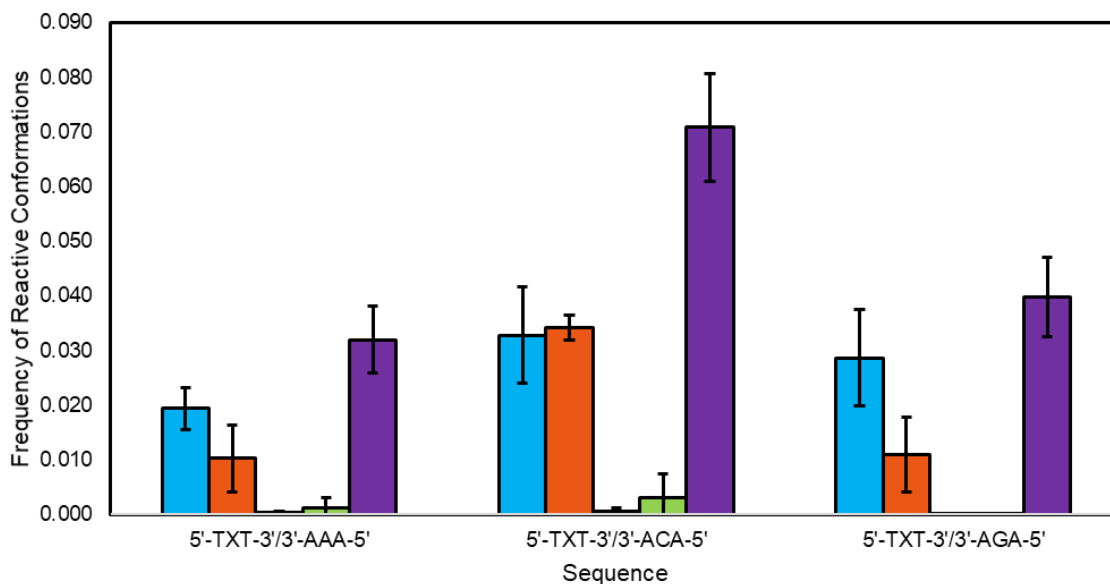


Figure 2.17: Frequency of reactive conformations for the formation of ICLs in the mismatched sequences. Reactive conformations for a nucleophile in the 5'-interstrand nucleotide and C10 (blue) or C11 (orange) of the epoxide, a nucleophile at the direct opposite position in the complementary nucleotide and C10 (yellow) or C11 (green) of the epoxide, and the total reactive conformations (purple) are provided.

While it was expected that the presence of a potential nucleophile in the complementary position in addition to the 5' inter position would increase reactive conformations as the result of the formation of reactive conformations between the epoxide and complementary nucleobase, analysis of reactive conformations has revealed that this reaction is not likely. Compared to the 5' inter A^(N6), complementary A^(N6) or complementary C^(N4) rarely adopted reactive conformations with the epoxide. FRC values between X^(C10) complementary A^(N6) or complementary C^(N4) remained negligible for the duration of each simulation. The FRC values between X^(C11) and

complementary A^(N6) or complementary C^(N4) ranged from 0.001 to 0.001, with average reactive angle values ranging from 147.4° to 151.3° (Figure 2.19). The highest incidence of reactive conformations in total was observed when C was complementary to the epoxide, with a total FRC of 0.071. When A^(N6) or G^(N2) was inserted across from the epoxide, the total frequencies of reactive conformations were 0.032 and 0.040, respectively. In addition, all reactive distances remained at 3.2 Å.

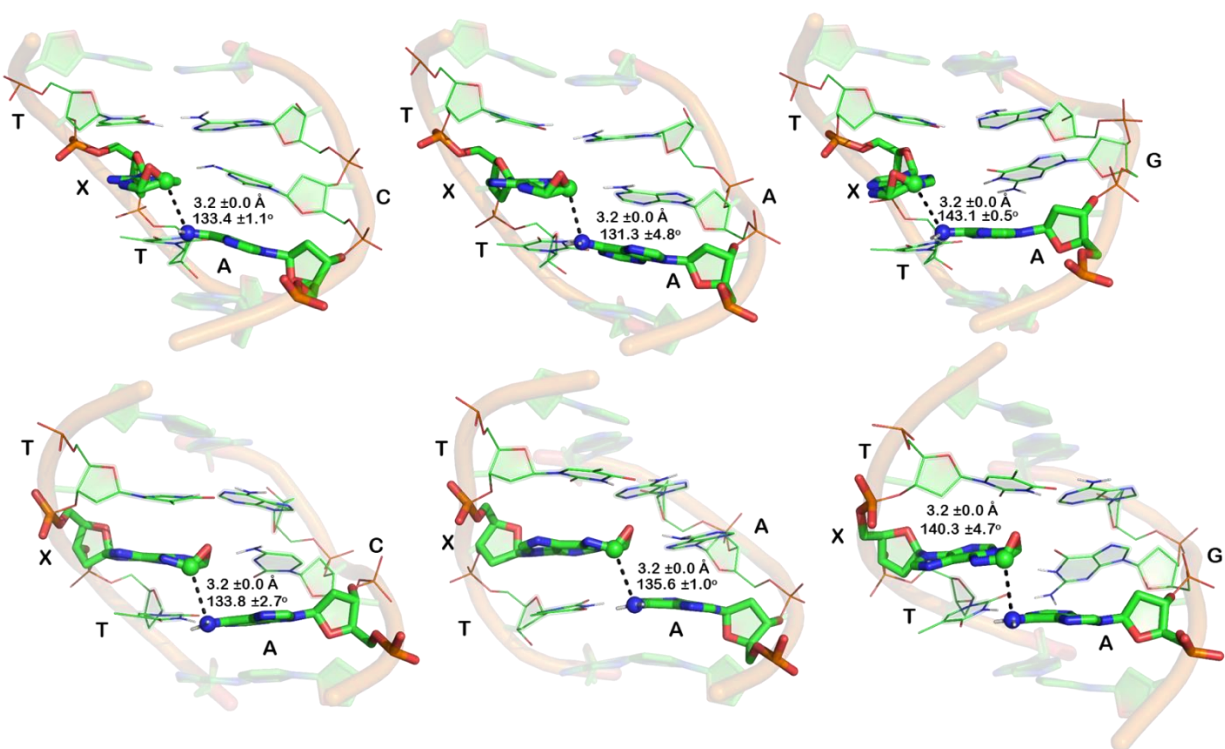


Figure 2.18: MD representative structures of the reactive conformations for crosslink formation between N6 of the 5'-interstrand adenine and C10 (top) or N6 of the 5'-interstrand adenine and C11 (bottom) of the epoxide intermediate (X) for DNA duplexes containing mismatches. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

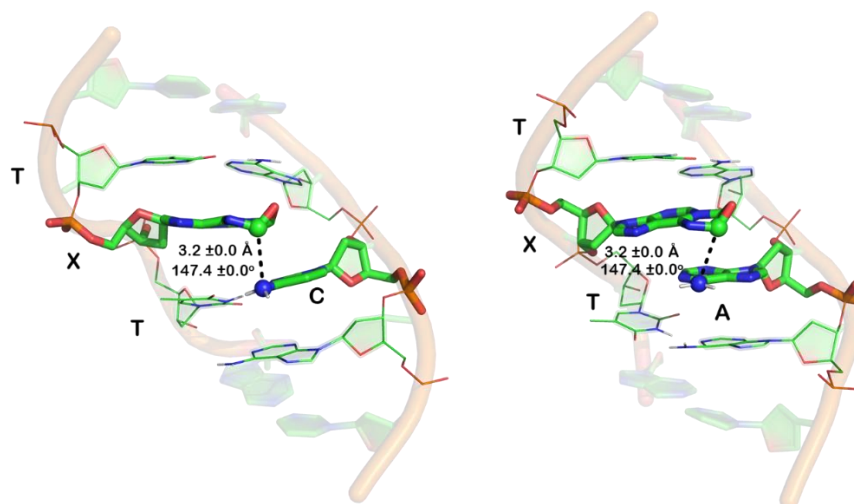


Figure 2.19: MD representative structures of the reactive conformations for crosslink formation between N4 of the complementary cytosine (left) or N6 of the complementary adenine (right) and C11 of the epoxide intermediate (X) for DNA duplexes containing mismatches. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

2.4 Discussion

As the details and observations of my computational study have been discussed in the results section, the correlations between my MD findings and the LC-MS experimental findings of the Li lab will be discussed in this section. The sequences discussed in the results section were analogous to sequences studied by the Li lab (Table 2.1 and 2.2). While the data provided by the Li lab demonstrated a sequence dependence of ICL formation, my work confirms the origin of this sequence context is the geometrical alignment of the electrophilic epoxide carbon atoms and nucleophilic amino groups, and identifies the atoms involved in ICL formation.

To verify which nucleobase and which heteroatom on the nucleobase reacts with the ϵ A epoxide intermediate in the opposing strand to form an ICL, base screening experiments (with only one A/G/C in the complementary strand) were performed (Table 2.1). This dataset showed that no

crosslink was generated in the duplex (5'-AXA-3'/5'-TTT-3'; entry 1), likely due to the lack of an exocyclic amino group in thymine to attack the ϵ A epoxide intermediate. For A and C in the complementary strand, obvious crosslink signals were detected (entries 2-4, 6-8). However, when A was replaced by purine (Table 2.1, dA*, without the 6-amino group of A), no ICL was detected (entry 7). Similarly, the replacement of C with U (Table 2.1, 4-amino group of C is replaced by a carbonyl group) completely abrogated crosslink formation. The results from the structural analogs that lack the exocyclic amino groups show that the N6-amino group in A and N4-amino group in C are necessary for the formation of ICLs. This further supports my analysis approach for the MD simulations, focusing on the amino group nucleophiles, as well as literature showing that the exocyclic amino groups are involved in crosslink formation in other contexts.¹⁹⁻²¹ In contrast to A/C, no crosslink could be identified for the N2-amino group of G (Table 2.1, entry 9 to 13, Figure 2.12A).² This supports my suggestion that the N2-amino group of G in the minor groove is too far from C10/C11 of the epoxide, thus preventing the ICL formation (entry 10-12).

To understand the relative positioning of potential nucleophiles on the nucleobase with respect to the highly reactive epoxide intermediate and thereby shed light on the observed trends in interstrand crosslink formation, MD simulations were performed on 15-mer complementary DNA duplexes containing the epoxide intermediate in each of the 16 possible canonical DNA sequence contexts (Table 2.2, entries 1–16).² The structural data was analyzed to determine the frequency in which the observed helical conformation correctly aligned any nucleophilic nitrogen centers in any of the neighboring or opposing canonical nucleobases with respect to either electrophilic center on the epoxide (C10 or C11, Figure 2.2) to afford potential ICL formation (Figures 2.4-2.18, Tables A-4-A-15). Because no crosslink products were observed when G or T was opposite the epoxide intermediate, I excluded the possibility of carbonyl oxygen atoms (such

as O6 in G and O2/O4 in C) or endocyclic amine nitrogen atoms (such as N3/N7 in the purines) as nucleophiles to react with the epoxide thus forming the ICL. Indeed, only the exocyclic amino groups of the A, C and G nucleobases were considered since these have been previously shown to act as nucleophiles and form crosslinks around other electrophiles such as Ap-derived ICLs.¹⁹⁻²¹

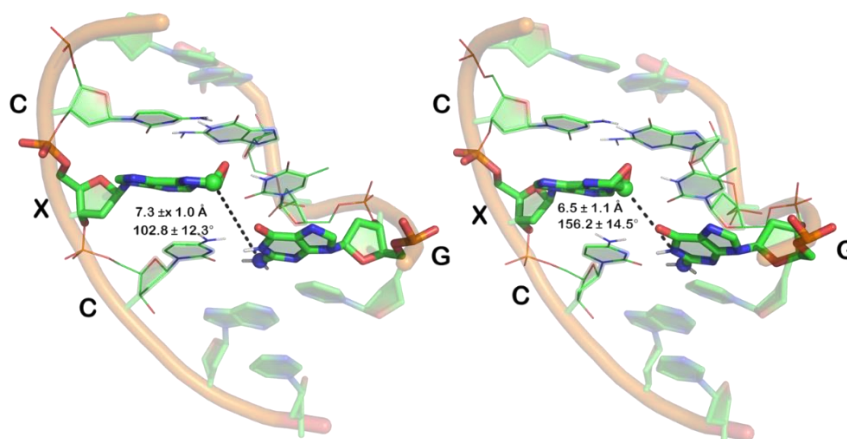


Figure 2.20: MD representative structures of the lesion site in the 5'-XC/5'-GT sequence context, displaying that the amino group of the interstrand guanine and the epoxide intermediate is too far to react for each 5'-XC/5'-GT duplex. Average electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) across all replicas provided.

In direct correlation with the experimental data, many reactive conformations were identified for the sequences containing 5'-XT/5'-AT, with the N6 amino group of the 5' inter A being aligned with respect to both C10 and C11 of the epoxide for ICL formation (Figure 2.7A). Although ICL formation was likely to occur between the amino group of the 5' inter C and both C10 and C11 of the epoxide for the 5'-XG/5'-CT sequences, the overall number of reactive conformations was significantly less (Figure 2.7C), which correlated with the lower experimental yields of C than A (Table 2.2). In contrast, a negligible number of conformations that could result in ICL formation were observed for the 5'-XC/5'-TT and 5'-XA/5'-GT sequences (Figure 2.12A and 2.11C), supporting the suggestion from experiments that neither T nor G opposite ϵ A in the

complementary strand is conducive for crosslink formation. Indeed, while G has an exocyclic amino group, the minor groove location places the exocyclic amino group too far from C10 or C11 of the epoxide to make the reaction feasible (Figure 2.20, Table A-4). Besides the nucleophiles in the complementary strand, the nucleophiles from the neighboring bases in the same strand can form intrastrand crosslinks. Interestingly, the 3' intra A or C in the corresponding sequences was aligned to react with the epoxide (at the C11 site in particular), which may result in an intrastrand crosslink. This contrasted the stark lack of conformations conducive for intrastrand crosslink formation for the 5'-XT/5'-AT and 5'-XG/5'-CT sequence contexts. Furthermore, amino groups present in 5' intra A, C, or G were not aligned for attack of the epoxide. Overall, the exocyclic amino groups of the 5' inter A, and to a lesser extent the 5' inter C (both in the major groove) were aligned to yield ICLs in DNA duplexes (Figure 2.8 to 2.10), providing a structural rationalization for the experimental ICL yields. Interestingly, amino groups present on nucleobases in the 3' inter positions (A, C, and G) did not significantly contribute to the frequency of reactive conformations.

Previous work has shown that ϵ A can induce 35% mutagenicity in cells without repair by the AlkB protein, yielding 25% A→T, 5% A→G and 5% A→C mutations.²² Therefore, three mismatched DNA sequences containing A/C/G instead of T opposite ϵ A in the complementary strand were generated (Table 2.2, entry 17 to 19) to study the crosslink to mimic ALKBH2 repair activity after misincorporation of the nucleobases during replication. The 5'-TXT-3' sequence was chosen for this analysis to rule out the possible formation of intrastrand crosslinks. Interestingly, crosslinks were detected in all three mismatched sequences (Table 2.2, entry 17 to 19). It was identified that the sequence containing 5'-TXT-3'/5'-AAA-3' could obtain a 5.7% yield quantified by LC-MS as the top hit (entry 18). For each mismatched sequence, MD simulations revealed that the 5' inter A (not the A opposite or 3' inter with respect to the epoxide) is frequently aligned with

respect to either C10 or C11 of the epoxide for ICL formation (Figure 2.18). Even for the mismatched strand containing G opposite ϵ A, MD simulations indicated that the formation of a crosslink (entry 19) is generated from the 5'-A, but not from G in the opposing (complementary) position. Furthermore, in each case, a smaller number of reactive conformations occur that involve the complementary A or C compared to 5'-A. Although the overall kinetics and thermodynamics of crosslink formation are not considered in my computational model, MD structural predictions regarding the alignments of nucleobase amino groups relative to the epoxide intermediate in DNA duplexes correlated with the experimentally determined trends in crosslink yields with respect to sequence context (entry 1 to 19 in Table 2.2). Therefore, the MD structural data rationalizes the observed ICL formation in mismatched sequences, with overall larger yields than the corresponding matched sequence likely occurring due to the potential for attack of the epoxide by the complementary nucleobase.

As discussed previously, MD simulations corroborated that no nucleophilic amino groups in the neighboring nucleobases in the complementary strand are appropriately aligned to yield an ICL at the epoxide center for the 5'-AXA-3'/5'-TTT-3' sequence (Figure 2.12). When A or C is considered in each of the 3 positions on the opposing strand, reactive conformations are observed, with the frequency decreasing in the order of 5' inter > complementary >>> 3' inter for both nucleobases (Figure S23 and 2.4). Interestingly, the majority of the reactive conformations align the amino group of the 5' inter A or C and C11 of the epoxide, or the amino group of the complementary A or C and C10 of the epoxide. In contrast to A or C, negligible reactive conformations occur with an interstrand G, regardless of its position relative to the epoxide. This structural data strongly correlates with the experimental yields, providing further support for the

involvement of the N6-amino group of A and N4-amino group of C in ICL formation, and rationalizes the limited ICL formation for G.

This work provides a detailed structural explanation for the role of sequence context in the experimentally observed ICL yields because of attempted repair by ALKBH2. Specifically, I have confirmed that while the exocyclic amino group is necessary for ICL formation, the minor groove placement of the exocyclic amino group prevents G from being an effective nucleophile in the aminolysis of the epoxide intermediate. Additionally, I have shown that while the major groove placement allows the exocyclic amino groups of A and C to align with the carbons of the epoxide in a manner conducive to reactivity, A is the superior nucleophile. I have also uncovered that the 5' inter position is the most favorable site for a potential nucleophile to attack the epoxide. My detailed analysis also reveals that intrastrand crosslinks can be formed when A or C is in the 3' intra position, with both carbon atoms being preferred by the nucleophiles in the 5' inter position and C10 being preferred by nucleophiles in the complementary position. This finding is significant because intrastrand crosslinks have been discovered by my collaborators when looking at ALKBH2 mediated crosslink formation in HeLa cells.

2.5 Conclusions

The Li lab discovered a novel DNA ICL *in vitro* as the result of attempted repair of ϵ A by ALKBH2 and found that formation of the observed ICL had a dependence on sequence context. My MD simulations revealed that the presence of A or C in the 5' inter position resulted in the adoption of reactive conformations by the damaged DNA, which was conducive to ICL formation. Relative to the 5' inter position, nucleotides in the 3' inter and complementary positions did not

form a large number of reactive conformations. This work supports the use of computational chemistry to characterize and quantify ICL formation within DNA duplexes. Further work must be done to map the mechanism of ICL formation in DNA duplex and NCP models, as ALKBH2 has the ability to access the ϵ A lesion even when occluded in the NCP.²³ Similarly, ICL formation in various locations and sequence contexts in the NCP model must be studied. Subsequently, the mechanisms of repair for these ICLs must be investigated to determine the toxicity of epoxide intermediate-derived ICLs relative to other ICLs (i.e., Ap-derived and drug induced ICLs).

2.6 References

1. Rioux, K. L.; Delaney, S., 1,N6-Ethenoadenine: From Molecular to Biological Consequences. *Chem. Res. Toxicol.* **2020**, *33* (11), 2688-2698.
2. Wang, J., Takyi, N. A., Hsiao, Y., Tang, Q., Chen, Y., Liu, C., Lu, K., Essigmann, J. M., Wetmore, S. D., Li, D., Stable Interstrand Crosslinks Generated from the Repair of 1,N6-Ethenoadenine in DNA by the α -Ketoglutarate/Fe(II)-Dependent Dioxygenase ALKBH2. *In preparation* **2022**.
3. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.
4. Schrodinger, LLC, The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8. 2015.
5. Schrodinger, LLC, The JyMOL Molecular Graphics Development Component, Version 1.8. 2015.
6. Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.8. 2015.
7. Fairlamb, M. S.; Whitaker, A. M.; Freudenthal, B. D., Apurinic/apyrimidinic (AP) Endonuclease 1 Processing of Ap sites with 5' Mismatches. *Acta Crystallogr. D* **2018**, *74* (8), 760-768.
8. Batra, Vinod K.; Beard, William A.; Pedersen, Lars C.; Wilson, Samuel H., Structures of DNA Polymerase Mispaiored DNA Termini Transitioning to Pre-catalytic Complexes Support an Induced-Fit Fidelity Mechanism. *Structure* **2016**, *24* (11), 1863-1875.
9. Batra, V. K.; Beard, W. A.; Shock, D. D.; Pedersen, L. C.; Wilson, S. H., Nucleotide-Induced DNA Polymerase Active Site Motions Accommodating a Mutagenic DNA Intermediate. *Structure* **2005**, *13* (8), 1225-1233.
10. Natrajan, G.; Lamers, M. H.; Enzlin, J. H.; Winterwerp, H. H. K.; Perrakis, A.; Sixma, T. K., Structures of Escherichia coli DNA Mismatch Repair Enzyme MutS in Complex with

- Different Mismatches: a Common Recognition Mode for Diverse Substrates. *Nucleic Acids Res.* **2003**, *31* (16), 4814-4821.
11. Wang, W.; Hellinga, H. W.; Beese, L. S., Structural Evidence for the Rare Tautomer Hypothesis of Spontaneous Mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (43), 17644-17648.
 12. Gao, Y.-G.; Robinson, H.; Sanishvili, R.; Joachimiak, A.; Wang, A. H. J., Structure and Recognition of Sheared Tandem G·A Base Pairs Associated with Human Centromere DNA Sequence at Atomic Resolution. *Biochem.* **1999**, *38* (50), 16452-16460.
 13. Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F.-Y., R.E.D. Server: A Web Service for Deriving RESP and ESP Charges and Building Force Field Libraries for New Molecules and Molecular Fragments. *Nucleic Acids Res.* **2011**, *39* (suppl_2), W511-W517.
 14. F. Wang, J.-P. B., P. Cieplak & F.-Y. Dupradeau *R.E.D. Python: Object oriented programming for Amber force fields*, Sanford Burnham Prebys Medical Discovery Institute: 2013.
 15. Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P., The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.* **2010**, *12* (28), 7821-7839.
 16. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A Well-Behaved Electrostatic Potential Based Method using Charge Restraints for Deriving Atomic Charges: the RESP Model. *J. Phys. Chem.* **1993**, *97* (40), 10269-10280.
 17. Zgarbová, M.; Otyepka, M.; Šponer, J.; Lankaš, F.; Jurečka, P., Base Pair Fraying in Molecular Dynamics Simulations of DNA and RNA. *J. Chem. Theory Comput.* **2014**, *10* (8), 3177-3189.
 18. Vijayalakshmi, K. P.; Mohan, N.; Ajitha, M. J.; Suresh, C. H., Mechanism of Epoxide Hydrolysis in Microsolvated Nucleotide Bases Adenine, Guanine and Cytosine: A DFT Study. *Org. Biomol. Chem.* **2011**, *9* (14), 5115-5122.
 19. Varela, J. G.; Pierce, L. E.; Guo, X.; Price, N. E.; Johnson, K. M.; Yang, Z.; Wang, Y.; Gates, K. S., Interstrand Cross-Link Formation Involving Reaction of a Mispaiored Cytosine Residue with an Abasic Site in Duplex DNA. *Chem. Res. Toxicol.* **2021**, *34* (4), 1124-1132.

20. Kellum, A. H.; Qiu, D. Y.; Voehler, M. W.; Martin, W.; Gates, K. S.; Stone, M. P., Structure of a Stable Interstrand DNA Cross-Link Involving a β -N-Glycosyl Linkage Between an N6-dA Amino Group and an Abasic Site. *Biochem.* **2021**, *60* (1), 41-52.
21. Housh, K.; Jha, J. S.; Yang, Z.; Haldar, T.; Johnson, K. M.; Yin, J.; Wang, Y.; Gates, K. S., Formation and Repair of an Interstrand DNA Cross-Link Arising from a Common Endogenous Lesion. *J. Am. Chem. Soc.* **2021**, *143* (37), 15344-15357.
22. Delaney, J. C.; Smeester, L.; Wong, C.; Frick, L. E.; Taghizadeh, K.; Wishnok, J. S.; Drennan, C. L.; Samson, L. D.; Essigmann, J. M., AlkB Reverses Etheno DNA Lesions Caused by Lipid Oxidation In Vitro and In Vivo. *Nat. Struct. Mol. Biol.* **2005**, *12* (10), 855-860.
23. Caffrey, P. J.; Kher, R.; Bian, K.; Li, D.; Delaney, S., Comparison of the Base Excision and Direct Reversal Repair Pathways for Correcting 1,N6-Ethenoadenine in Strongly Positioned Nucleosome Core Particles. *Chem. Res. Toxicol.* **2020**, *33* (7), 1888-1896.

Chapter 3.0: A Multiscale Computational Investigation of Ap-derived ICL Formation

3.1: Introduction

While the previous chapter documented a novel type of ICL formed through nucleophilic attack of an epoxide by exocyclic amino groups of DNA nucleobases, this chapter centers around a well-known type of ICL, namely the apurinic/apyrimidinic (Ap)-derived ICL. Experimental investigations have uncovered that a native Ap site exists in an equilibrium between a ring-closed hemiacetal form and a ring-opened aldehyde form in physiological conditions (Figure 3.1).¹ It has been proposed that the electrophilic carbon atom of the aldehyde is attacked by a nearby nucleophilic exocyclic amino group (present on the A, C, or G nucleobase) in the DNA helix to yield an ICL.² This step results in the formation of a hemiaminal group, which links the complementary DNA strands. Subsequently, the rate-limiting step involves Schiff base formation, where an imine is formed as the linker between the two strands. Finally, O4' attacks C1' to form an aminoglycoside linkage between the complementary strands.

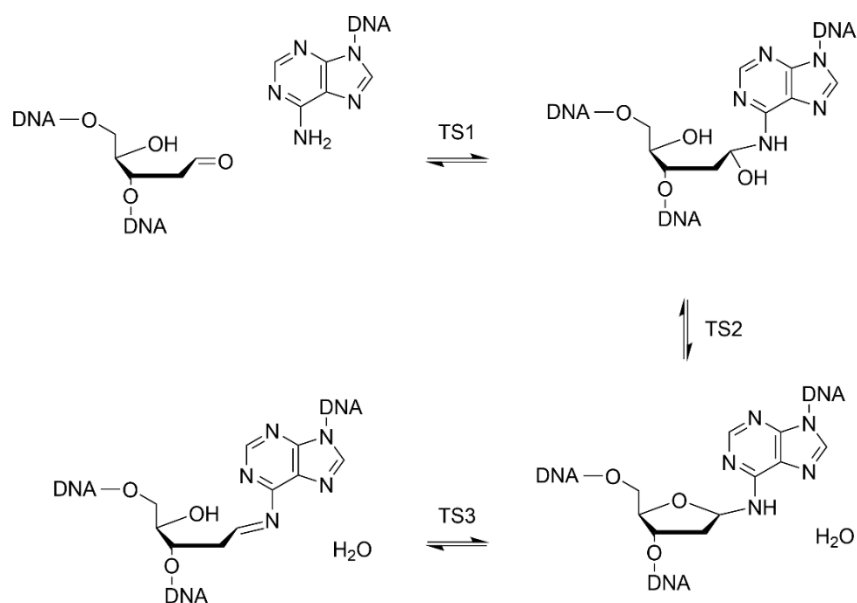


Figure 3.1: A representative example of the proposed Ap-derived ICL formation reaction (C1' of an Ap site and N6 of A).⁵

Because ICLs covalently link complementary DNA strands, they can block processes that are integral to cell survival and division, such as transcription and DNA replication.² Previous experimental studies have documented that Ap-derived ICLs possess the ability to block the ϕ 29 DNA polymerase *in vitro* regardless of whether the Ap portion of the Ap-derived ICL resides on the leading or lagging strand.² In addition, the dG-Ap ICL was observed to block DNA replication in HEK293T cells or result in nucleobase substitution mutations largely at the Ap portion of the ICL during attempted replication.³ As they disrupt these important processes, ICLs are exceedingly toxic to cells and have the ability to induce cellular apoptosis if left unrepaired.¹⁵ Due to the potentially catastrophic effects that ICLs have on the integrity of a cell, the mechanisms under which ICLs form and how they affect the structure of DNA must be studied.

Experimental studies have revealed important information about the effect of sequence on ICL formation. ICLs have been observed to form between C1' of an Ap site and the exocyclic amino groups of G (dG-Ap), C (dC-Ap), and A (dA-Ap; Figure 3.2).^{2, 4-5} However, for each type of Ap-derived ICL, each nucleophile (i.e., the amino group of A, C, or G) must be present in a certain position in the complementary strand relative to an Ap site. Specifically, a dA-Ap ICL was only found to form when A is offset in the 3' direction of an Ap site, a dG-Ap ICL occurred when G is offset in the 5' direction of an Ap site, and a dC-Ap exclusively formed when C in a C:A mispair is offset in the 3' direction of an Ap (Figure 3.2).^{2, 4-5} In addition, each of these ICLs formed in different relative yields, with the dA-Ap ICL being the most abundant independent of the possibility of other Ap-derived ICLs. Previous work that investigated the effect of sequence context on ICL formation found that the dA-Ap ICL preferentially formed over the dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' sequence context.⁶ In addition, the same study documented that the dA-Ap ICL formed at a 70% yield in the 5'-AAG-3'/3'-TXC-5' sequence context compared to a

13-18% yield in 5'-AAC-3'/3'-TXG-5'.⁶ As the reasons for these findings are not yet understood, structural and chemical descriptions of these systems obtained using computational chemistry may be useful for rationalizing these requirements for Ap-derived ICL formation.

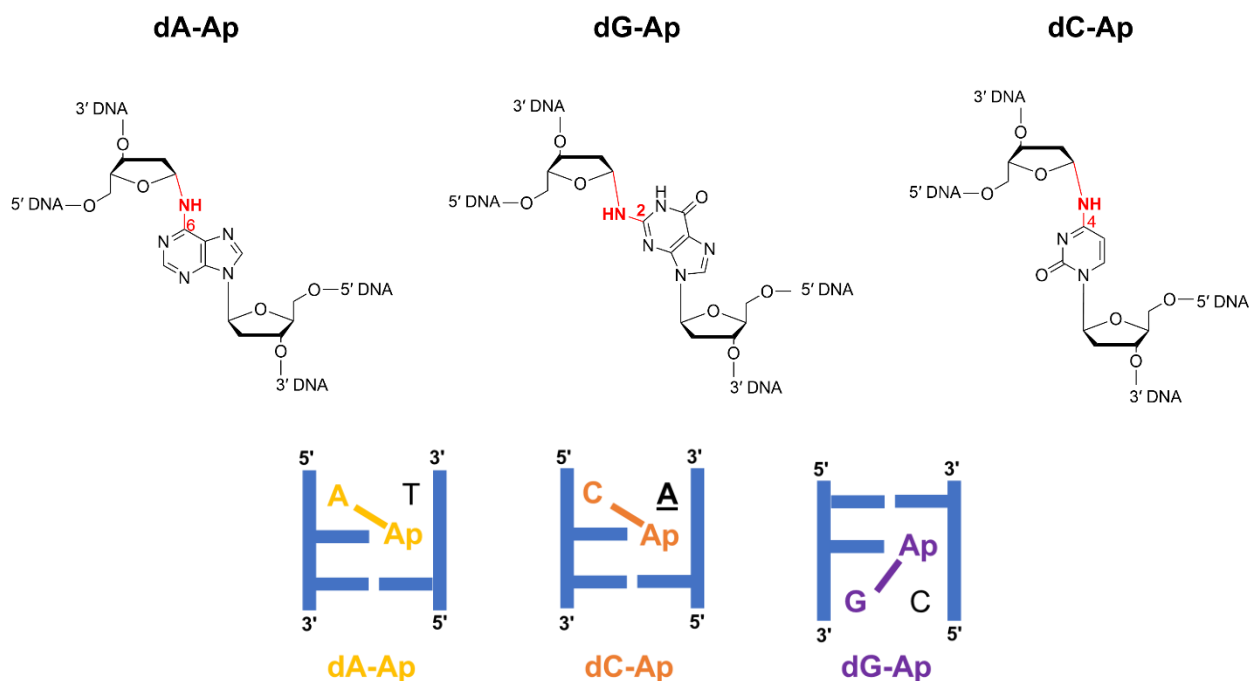


Figure 3.2: Ap-derived ICLs (top): dA-Ap (left), dG-Ap (left), dC-Ap (center), and the sequence contexts in which they form (bottom).^{2, 4-5}

Some *in silico* studies have already been performed on Ap-derived ICLs. When DFT was used to look into the fundamental relative reactivity of the exocyclic amino groups of the nucleobases towards C4' oxidized Ap sites, the barriers were found to decrease in the order: dA-Ap > dG-Ap > dC-Ap, and the product complex stability was found to decrease in the order: dA-Ap > dC-Ap > dG-Ap.⁷ However, in the same study, the product complex for dA-Ap as well as each transition state involved relative orientations of chemical species that are not realistic for B-DNA or would drastically distort a B-DNA helix.⁷ Another study that compared dA-Ap to dG-Ap ICL formation in the DNA duplex environment employed MD simulations on DNA duplexes

containing Ap sites and ICLs to understand the difference in ICL formation yield and speculated that the amount of DNA duplex distortion necessary to form an ICL product may play an integral role in ICL formation.⁸ However, while different sequences were used compared to the sequences used in this thesis, the structural analyses of the duplexes was limited to averaged helical parameters and hydrogen bond distances.⁸ While DFT and atomistic MD simulations have been used to study Ap-derived ICL formation, a multiscale investigation of native (C4' is not oxidized) Ap-derived ICL formation has not yet been completed. Regardless of the number of discoveries that have been made about Ap-derived ICLs, information regarding why the dA-Ap ICL is preferred over the dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' sequence context, and why the dA-Ap ICL yield decreases in the 5'-AAC-3'/3'-TXG-5' sequence context is missing.

This work employs a multiscale computational approach to investigate Ap-derived ICL formation in the 5'-AAG-3'/3'-TXC-5' and 5'-AAC-3'/3'-TXG-5' sequence contexts. Specifically, the relative reactivity of the exocyclic amines in the A and G nucleobases towards native Ap sites is explored using DFT models. Subsequently, the effects of an Ap site on the B-DNA structure and the effect of each ICL on the DNA structure will be studied in their respective sequence contexts using MD simulations. This work demonstrates and supports the application of computational chemistry to provide molecular level insight into experimentally determined results, as well as the role that structure and implicit reactivity play in Ap-derived ICL formation.

3.2 Computational Methods

3.2.1 Density Functional Theory Calculations

The proposed mechanisms for dA-Ap and dG-Ap ICL formation were investigated using DFT calculations. Initial nucleobase-Ap site models (Figure B-1) were constructed with the intent to orient the chemical species in a manner relevant to B-DNA. B3LYP/6-31G(d) gas-phase geometry optimizations were implemented to search the potential energy surfaces. Frequency calculations were used to confirm the nature of minima and transition states while IRC calculations were used to confirm a continuous reaction pathway. Single-point energy calculations were performed at the B3LYP, B3LYP-D3(BJ), M06-2X, and MP2 levels of theory with a 6-311+G(2df,2p) basis set. Additionally, M06-2X/6-311+G(2df,2p) single-point energy calculations were performed in implicit solvent (1-bromopropane and water) using the IEF-PCM solvation model. Gibbs energies were computed using the aggregated corresponding single-point energy values and Gibbs energy corrections. Key distances and angles were measured using GaussView 6.0.⁹ All DFT calculations were executed using Gaussian 16 (B.01), and all input files were constructed using GaussView 6.0.⁹

3.2.2 Molecular Dynamics Simulations

3.2.2.1 MD Model Construction

Unmodified B-DNA double helices were constructed using the Nucleic Acid Builder (NAB) module in AMBER18,¹⁰ with the 5'-GCGTCGAACAGTCGG-3'/3'-CGCAGCTTGTCAGCC-5' and 5'-GCGTCGAAGAGTCGG-3'/3'-CGCAGCTTCTCAGCC-5' sequences. To create models that were representative of the experimentally studied products,

these structures were used to create DNA duplexes containing ICLs (possible sites bolded in the previous sequences) and natural DNA duplex models. For each damaged DNA model, the modelled ICL was incorporated into the DNA duplex using GaussView 6.0⁹ and the chirality of each ICL was reproduced based on previously resolved experimental structures.^{5, 11} Natural DNA nucleotides were described using the AMBER OL15 force field,^{10, 12} and missing parameters for the ICLs were supplemented using Generalized AMBER Force Field (GAFF) parameters.^{10, 12} The charges for crosslinked DNA were supplied by the pyRED server.¹³⁻¹⁶ The duplex was placed in a TIP4P-EW water box such that the solute was situated at least 10 Å away from the edge of the box. Subsequently, the system was neutralized with Na⁺, and Na⁺ and Cl⁻ was added to achieve an overall salt concentration of 150 mM. Input files for MD simulations were prepared using the tleap module in AMBER18.

3.2.2.2 MD Simulation Protocol

Pre-production simulations on each system were initiated with four stages of minimization. The first minimization stage consisted of 2500 steepest descent steps and a subsequent 2500 conjugate gradient steps, with the DNA restrained using a 100 kcal/(mol Å²) restraint. The second stage involved a 100 kcal/(mol Å²) restraint on all water and ions, and the same number of steepest descent and conjugate gradient steps as stage one. In stage three, a 100 kcal/(mol Å²) restraint was imposed on the DNA and 1500 steepest descent steps were performed, followed by 1500 conjugate gradient steps. The fourth stage did not involve any restraints and consisted of 2000 steepest descent steps and 1000 conjugate gradient steps. Using a Langevin thermostat, the system was then slowly heated in six stages from 0 to 310 K. Next,

the system was equilibrated with a decreasing (20, 15, 10, 5, and 1.5 kcal/(mol Å²)) restraint on DNA for 10 ps in total. During equilibration, a 10.0 Å nonbonded cutoff was employed.

Each system was run for 1.1 μs of molecular dynamics production in triplicate, with a 0.002 ps step size and a nonbonded cutoff of 10.0 Å. Isotropic positioning scaling was utilized to keep the system at a constant pressure of 1.0 bar. Pressure relaxation occurred every 2.0 ps and SHAKE bond length constraints were used to constrain all bonds containing hydrogen. In addition, all bond information involving hydrogen was omitted in the force evaluation. The temperature was held at 310 K using Langevin dynamics, with a collision frequency of 3.0. The terminal ends of DNA were restrained such that when Watson-Crick heavy atom pair distances exceeded the range of 2–4 Å, a 25 kcal/(molÅ²) of force was applied to the heavy atoms participating in terminal Watson-Crick hydrogen bonding. The final 1 μs MD simulation was extracted and every 5000th step was analyzed.

3.2.2.3 MD Analysis

All quantitative analyses, apart from helical parameter calculations, were completed using the CPPTRAJ module in AMBERTOOLS/20.¹² Root-mean-square deviations (RMSD) were calculated with respect to all heavy atoms in the DNA backbone (namely: P, O3', O5', C3', C4', and C5'). Representative structures were generated using the hieraglo algorithm according to the RMSD of the DNA backbone heavy atoms. Minor groove sizes were calculated using the P–P distance between complementary residues (Figure B–2), while glycosidic torsion angles were calculated using the O4', C1', N1, and C2 atoms (Figure B–2) for pyrimidines, and O4', C1', N9, and C4 for purines (Figure B–2). Backbone torsion angles (Figure B–3) were defined as follows:

$\alpha = \angle(\text{O}3'_{n-1}, \text{P}_n, \text{O}5'_n, \text{C}5'_n)$, $\beta = \angle(\text{P}_n, \text{O}5'_n, \text{C}5'_n, \text{C}4'_n)$, $\gamma = \angle(\text{O}5'_n, \text{C}5'_n, \text{C}4'_n, \text{C}3'_n)$, $\delta = \angle(\text{C}5'_n, \text{C}4'_n, \text{C}3'_n, \text{O}3'_n)$, $\varepsilon = \angle(\text{C}4'_n, \text{C}3'_n, \text{O}3'_n, \text{P}_{n-1})$, and $\zeta = \angle(\text{C}3'_n, \text{O}3'_n, \text{P}_{n+1}, \text{O}5'_{n+1})$. Nucleobase–nucleobase noncovalent interactions (NCIs; Figure B–4) and DNA–solvent interaction energies were calculated using the linear interaction energy method. While all DNA and solvent atoms were considered for DNA–solvent interaction energies, only nucleobase atoms were considered for nucleobase–nucleobase NCI energies. Nucleic acid helical parameters were calculated for each duplex using the CURVES/CANAL+ software package.¹⁷ PyMOL was used to visualize trajectories and representative structures.¹⁸⁻¹⁹

3.3 Results and Discussion

3.3.1 Schiff-Base Formation is the Rate-Limiting Step of dA-Ap and dG-Ap ICL Formation

As previously stated, native Ap sites exist in two forms: a ring-closed hemiacetal form and a ring-open aldehyde form, the latter of which is known to be significantly reactive (Figure 3.3).^{2, 4-5} While the mechanism of ICL formation as the result of nucleophilic attack by nearby DNA nucleobases has been previously proposed,^{2, 4-5} important geometries between nucleophiles and electrophiles have not been described along the reaction pathway. The intent of this section is to describe the orientations of nucleophiles and electrophiles as the reaction proceeds and to determine the rate-determining step. Ap-derived ICL formation has been proposed to be initiated with the nucleophilic attack at the carbonyl of a ring-opened Ap site (RC) by a nearby exocyclic amino group of a DNA nucleobase, resulting in the formation of IC1, in which the two DNA strands are linked by a hemiaminal group.^{2, 4-5} During this step, a water molecule deprotonates the amino group, and protonates the alcohol. In my DFT calculations the distance between the

electrophile and nucleophile in this initial S_N2 step (R(C1'–N)) was 6.166 Å in RC, 1.661 Å in TS1, and 1.489 Å in IC1 for dA-Ap ICL formation (Figure 3.4), and 6.138 Å in RC, 1.679 Å in TS1, and 1.495 Å in IC1 for dG-Ap ICL formation (Figure 3.5). Additionally, the distance between the electrophile and the leaving group (R(C1'–O1')) for this step was 1.220 Å in RC, 1.340 Å in TS1, and 1.414 Å in IC1 for dA-Ap ICL formation, and 1.222 Å in RC, 1.333 Å in TS1, and 1.411 Å in IC1 for dG-Ap ICL formation. The angle of attack ∠(N,C1',O1')) for this step was 40.6° in RC, 104.4° in TS2, and 106.0° in IC1 for dA-Ap ICL formation, and 42.7° in RC, 104.6° in TS2, and 106.1° in IC1 for dG-Ap ICL formation. Thus, the key angles and distances for the first reaction step did not differ by a large amount between dA-Ap and dG-Ap ICL formation.

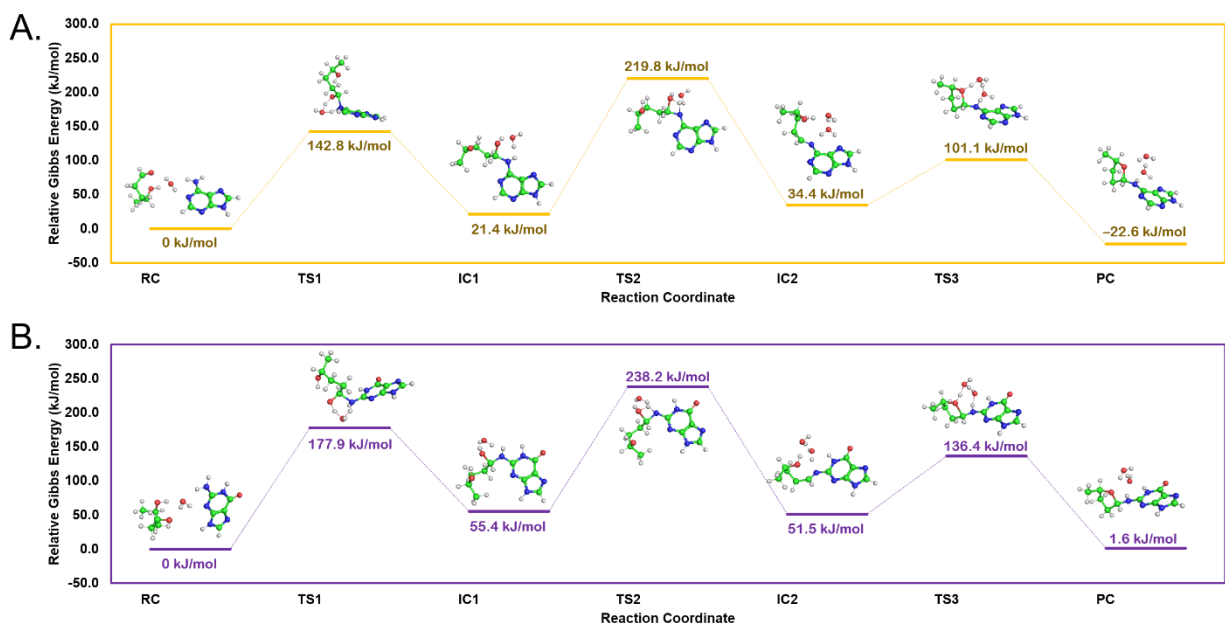


Figure 3.3: Structures and relative energies for stationary points of A) dA-Ap and B) dG-Ap ICL formation calculated using M06-2X/6-311+G(2df,2p)//B3LYP/6-31G(d) in the gas phase.

During the second step of ICL formation, the nitrogen group of the hemiaminal is deprotonated by the alcohol group on C1', which forms a sterically strained four-membered ring as the transition state, and the resulting water leaves.^{2, 4-5} This results in IC2 in which the two DNA strands are linked through an imine group.^{2, 4-5} Because the formation of the imine group generates an additional water, two water molecules are present in IC2. The distance between the electrophile and nucleophile for this imine-yielding step ($R(C1'-N)$) was 1.440 Å in TS2 and 1.287 Å in IC2 for dA-Ap ICL formation. The respective distances for dG-Ap ICL formation were: 1.443 Å in TS2 and 1.288 Å in IC2. $R(C1'-O1')$ for this step was 1.563 Å in TS2 and 4.000 Å in IC2 for dA-Ap ICL formation, and 1.549 Å in TS2 and 4.314 Å for dG-Ap ICL formation. Additionally, $\angle(N,C1',O1')$ for this step was 92.8° in TS2 and 42.2° in IC2 for dA-Ap ICL formation, and 93.4° in TS2 and 36.4° in IC2 for dG-Ap ICL formation. Like in step one, key reactive angles and distances remained similar in magnitude when step 2 of dA-Ap and dG-Ap ICL formation were compared.

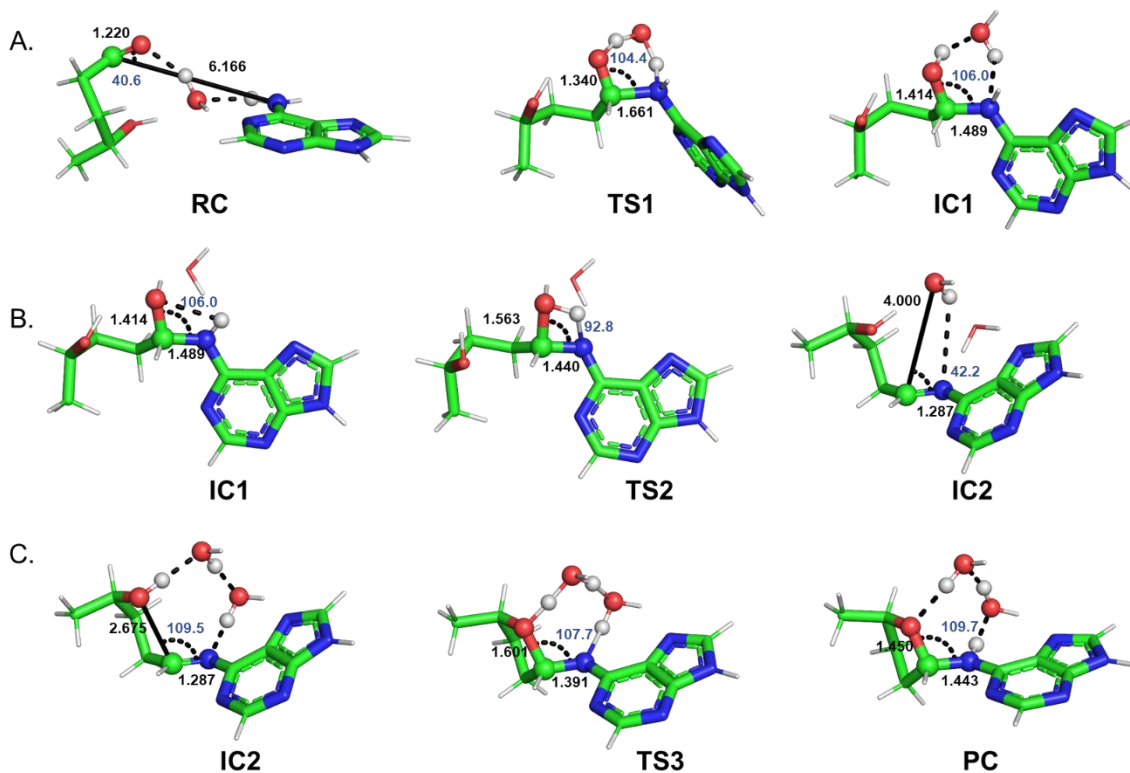


Figure 3.4: (A) Important angles (deg.) and distances (Å) for (A) step 1, (B) step 2, and (C) step 3 of dA-Ap ICL formation provided in Figure 3.

Finally, O4' of an Ap site is deprotonated while attacking C1' and the nitrogen atom on the imine is protonated, resulting in an aminoglycoside (PC).^{2, 4-5} In this step, both water molecules facilitate the deprotonation of O4' and the protonation of the nitrogen atom. The distance between the electrophile and nucleophile in this step ($R(O4'-C1')$) was 2.675 Å in IC2, 1.601 Å in TS3, and 1.450 Å in PC for dA-Ap ICL formation, and 2.684 Å in IC2, 1.905 Å in TS3, and 1.458 Å in PC for dG-Ap ICL formation. The distance between the electrophile and the leaving group ($R(O4'-N)$) for this step was 1.287 Å in IC2, 1.391 Å in TS3, and 1.443 Å in PC for dA-Ap ICL formation, as well as 1.288 Å in IC2, 1.336 Å in TS3, and 1.440 Å for dG-Ap ICL formation. The angle of attack $\angle(N,C1',O4')$ for this step was 109.5° in IC2, 107.7° in TS3, and 109.7° in PC for dA-Ap ICL formation, and 107.7° in IC2, 108.4° in TS3, and 111.3° in PC

for dG-Ap ICL formation. Again, in step 3, key reactive distances and angles were nearly equal when dA-Ap and dG-Ap ICL formation were compared. Thus, there are only slight geometrical differences between the key angles and distances along the entire reaction pathway for formation of dA-Ap and dG-Ap ICLs.

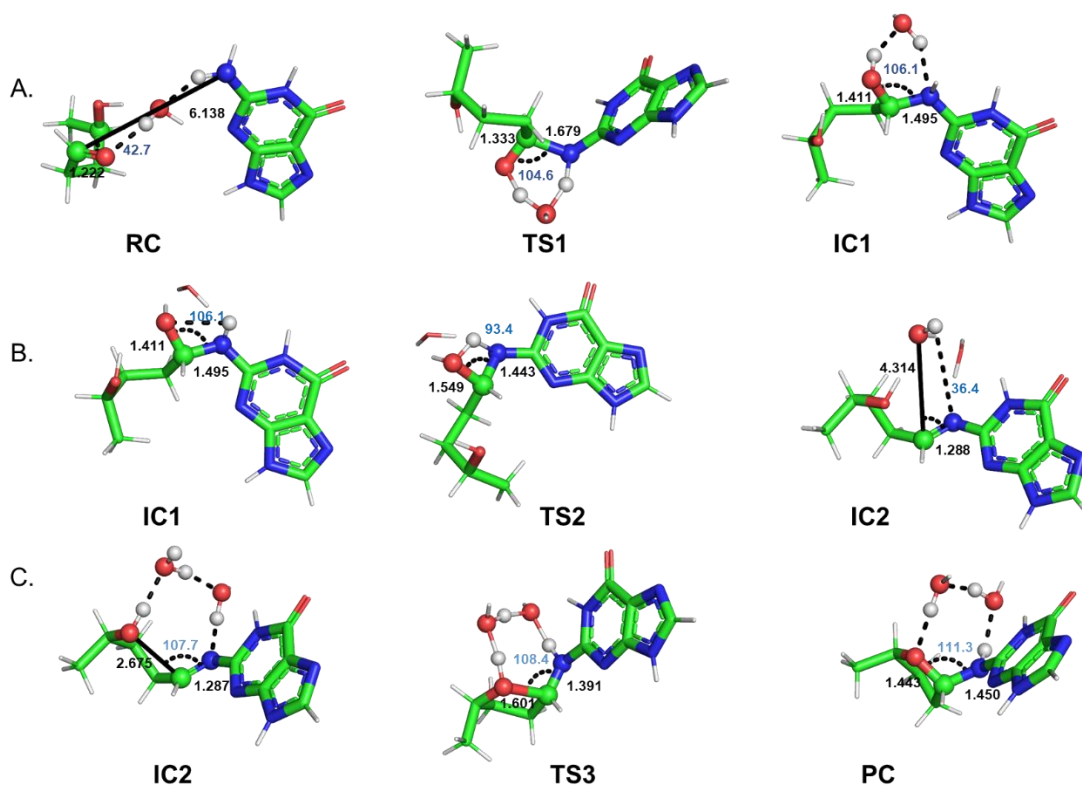


Figure 3.5: (A) Important angles (deg.) and distances (Å) for (A) step 1, (B) step 2, and (C) step 3 of dG-Ap ICL formation provided in Figure 3.

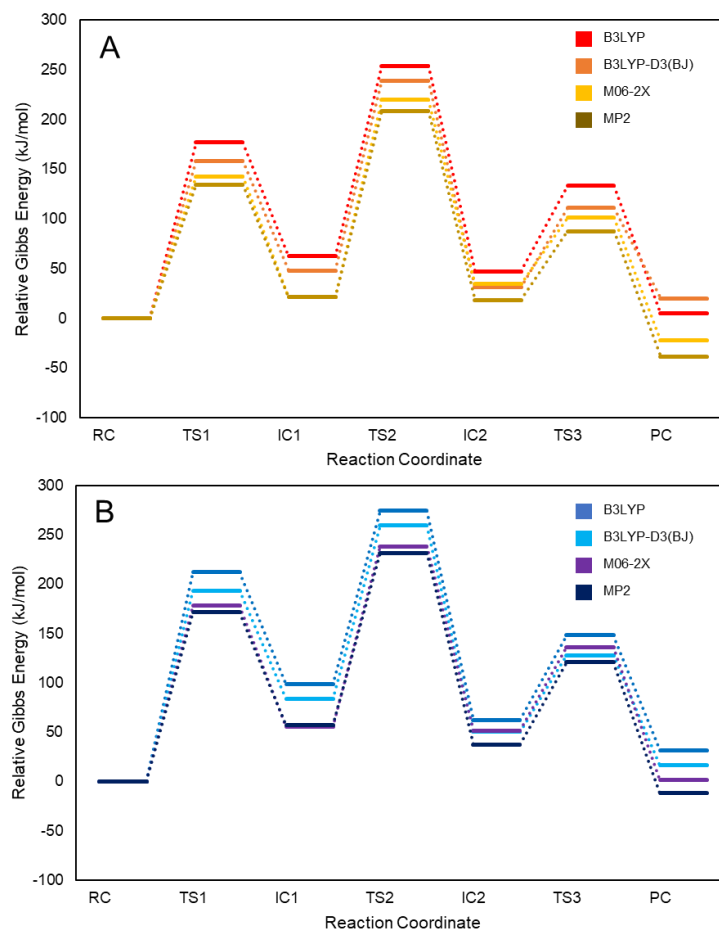


Figure 3.6: Relative Gibbs energies of stationary points for (A) dA-Ap and (B) dG-Ap ICL formation calculated using B3LYP, B3LYP-D3(BJ), M06-2X, and MP2 with the 6-311+G(2df,2p) basis set in the gas phase.

After the structures of the stationary points along the dA-Ap and dG-Ap ICL formation reactions were determined, B3LYP, B3LYP-D3(BJ), M06-2X, and MP2 single-point calculations using a 6-311+G(2df,2p) basis set were employed to gain an understanding of the energetics of dA-Ap and dG-Ap ICL formation (Figure 3.6). Regardless of the nucleobase considered or the method used to calculate the relative energies, the trend in the relative energies for the transition states was $TS2 > TS1 > TS3$ (Table B-8). Thus, the rate-determining step is imine formation from the hemiaminal group (TS2). However, the TS2 barriers may have been

artificially high due to the strained transition state (as it is a four-membered ring). The dA-Ap and dG-Ap ICL formation reaction kinetics and thermodynamics will be compared in the following subsection. As the M06-2X method performed most similarly to the more computationally demanding MP2 calculations in terms of relative energy values, subsequent calculations in implicit 1-bromopropane and water were performed using M06-2X. As such, energy values discussed in the next two subsections have been obtained using M06-2X.

3.3.2 The dA-Ap ICL is Intrinsically Preferred over the dG-Ap ICL from a Kinetic and Thermodynamic Standpoint

The rate-limiting barrier (corresponding to TS2; Figure 3.3) revealed that dA-Ap ICL formation is kinetically preferred over dG-Ap ICL formation. The gas-phase rate-limiting barrier for dA-Ap ICL formation was 18.4 kJ/mol lower than the rate-limiting barrier for dG-Ap ICL formation. This correlated with the reported larger proton affinity for exocyclic amino group of A than G,²⁰ making it a better nucleophile than G. This was due to the presence of an electron-withdrawing carbonyl at O6 of G.²¹ In addition, DFT calculations showed that dA-Ap ICL formation is exergonic (-22.6 kJ/mol), while dG-Ap ICL formation is endergonic (1.6 kJ/mol). This suggests that dA-Ap ICL formation is thermodynamically preferred, as consistent with experimentally observed dA-Ap ICL yields. One reason for the thermodynamic preference of dA-Ap ICL formation was the greater delocalization of charge upon dA-Ap ICL formation compared to that of dG-Ap ICL formation (Table 3.1). Specifically, the magnitude of the dipole moment of RC was greater than the dipole moment of PC for dA-Ap ICL formation ($\Delta\mu = -1.2135$ Debye), while the opposite was true for dG-Ap ICL formation ($\Delta\mu = 1.5421$ Debye). As previous studies have indicated that solvent conditions can affect reaction energetics,²²⁻²⁵ the

relative energies of each reaction step were also calculated in different implicit solvents to investigate the effects on the energy barrier.

Table 3.1: Dipole moments of the RC and PC for dA-Ap and dG-Ap ICL formation.^a

System	Dipole Vector Magnitude (Debye)		$\Delta\mu$
	RC	PC	
dA-Ap	3.046	1.832	-1.214
dG-Ap	6.559	8.101	1.542

^a Dipole moments were calculated using M06-2X/6-311+G(2df,2p)//B3LYP/6-31G(d) gas-phase calculations.

3.3.3 Increasing Solvent Polarity Slightly Decreases Reaction Barriers and Slightly Decreases Product Stability

To understand how the environment affects ICL formation, the dA-Ap and dG-Ap ICL reaction energies were investigated in 1-bromopropane ($\epsilon = 8.1$), chosen due to having a similar dielectric constant to DNA, and water ($\epsilon = 80.4$).²⁶⁻²⁸ An increase in the dielectric constant of the surroundings decreased the barriers associated with TS1 and TS2 (Figure 3.7). For example, the rate-determining barrier for dA-Ap ICL formation was 219.8 kJ/mol in the gas phase, 212.6 kJ/mol in 1-bromopropane, and 210.8 kJ/mol in water. Similarly, the rate-determining barrier for dG-Ap ICL formation was 238.2 kJ/mol in the gas phase, 231.7 kJ/mol in 1-bromopropane, and 230.4 kJ/mol in water. The largest decrease in barriers in response to an increase in the dielectric constant occurred for TS1 for both dA-Ap and dG-Ap ICL formation, which corresponded to a decrease in the barrier of 14.1 kJ/mol and 8.5 kJ/mol respectively. Increasing the dielectric constant of the surrounding environment resulted in the smallest change of the barrier height for TS3 for both dA-Ap (-1.8 kJ/mol in 1-bromopropane) and dG-Ap ICL formation (-3.6 kJ/mol in

1-bromopropane), which were almost negligible. Additionally, larger dielectric constants resulted in a slightly less stable product complex (increase by 0.9 kJ/mol for dA-Ap ICL formation and 7 kJ/mol for dG-Ap ICL formation). As a larger implicit solvent dielectric constant tends to weaken hydrogen bonds, this destabilization of the PC could be in part due to the PC containing one more hydrogen bond than the RC in both cases.

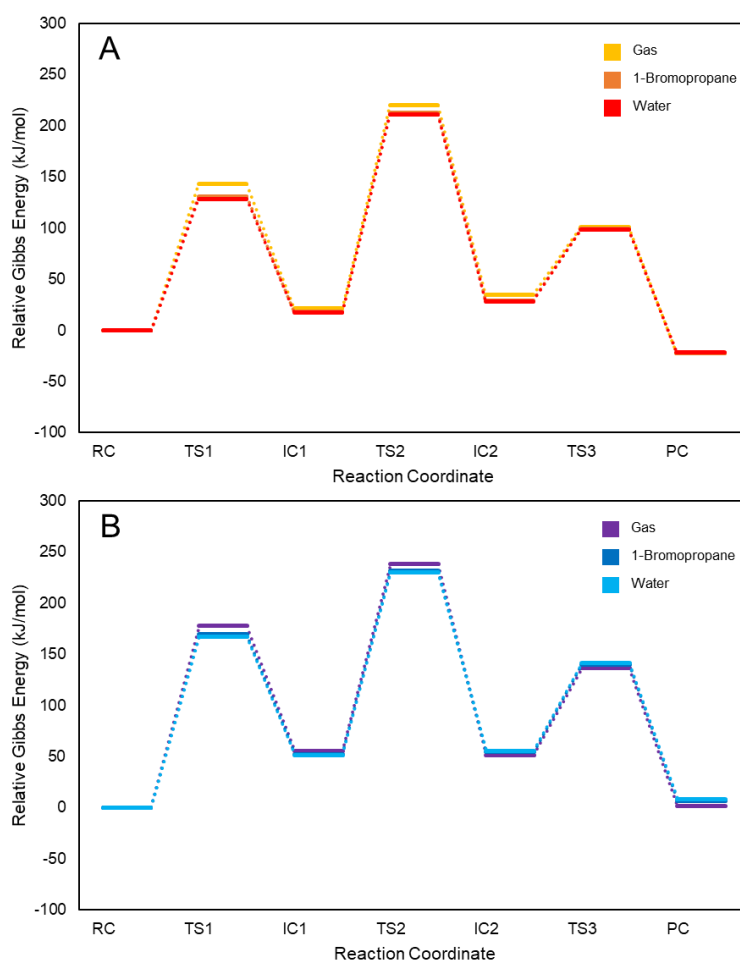


Figure 3.7: Relative Gibbs energies of stationary points for (A) dA-Ap and (B) dG-Ap ICL formation calculated using M06-2X/6-311+G(2df,2p)//B3LYP/6-31G(d) in the gas phase, implicit 1-bromopropane, and implicit water.

While changing the dielectric constant of the implicit solvent affected the magnitude of the barriers and the reaction thermodynamics, TS2 remained the rate-limiting barrier, and dA-Ap ICL formation was both kinetically and thermodynamically preferred to dG-Ap ICL formation, in all environmental conditions. Though these observations provide a rationale for the observed preference of dA-Ap over dG-Ap from the perspective of intrinsic reactivity, the products must be studied in the context of a DNA duplex for a comprehensive understanding of why dA-Ap is preferred over dG-Ap in the 5'-AAG-3'/3'-TXC-5' sequence context.

3.3.4 In DNA Duplexes, the dA-Ap ICL Results in Less Helical Distortion and More Stabilizing NCIs than the dG-Ap ICL

To gain insight into why the dA-Ap ICL was the major product in the 5'-AAG-3'/3'-TXC-5' sequence context, I compared the data from MD simulations of a DNA duplex containing a dA-Ap ICL in the 5'-AAG-3'/3'-TXC-5' (dA-Ap AAG) and a dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' (dG-Ap AAG) sequence contexts. Because the dA-Ap ICL has been experimentally resolved in a duplex and the dG-Ap ICL has been resolved as a fragment (nucleoside model), the ICL products have been selected to be investigated.^{5, 11}

Within the 5'-AAG-3'/3'-TXC-5' sequence context, dA-Ap and dG-Ap both involved complete loss of Watson-Crick hydrogen bonding at the base pairs where ICL formation occurred (Table B-9). While dA-Ap formation required an A:T pair to be broken, dG-Ap formation required the breaking of a G:C pair, which was more energetically costly. While the dA-Ap ICL connected an Ap site to the interstrand nucleotide in the 3' direction, dG-Ap connected the Ap site to the interstrand nucleotide in the 5' direction (Figure B-6). While the

distance between the sugars of the dA-Ap ICL crosslinking residues was 7.3 (± 0.1) Å (Figure 3.8), the sugar distance of the dG-Ap ICL was 6.7 (± 0.1) Å. These sugar distances were short in comparison to natural DNA values and caused a significant amount of distortion around each of the crosslinks formed. Compared to the sugar distance of the dA-Ap ICL, the short sugar distance of the dG-Ap ICL appeared to have a large affect the global helical structure of the DNA. However, both ICLs caused a number of structural rearrangements at the base pairs involving and surrounding each ICL.

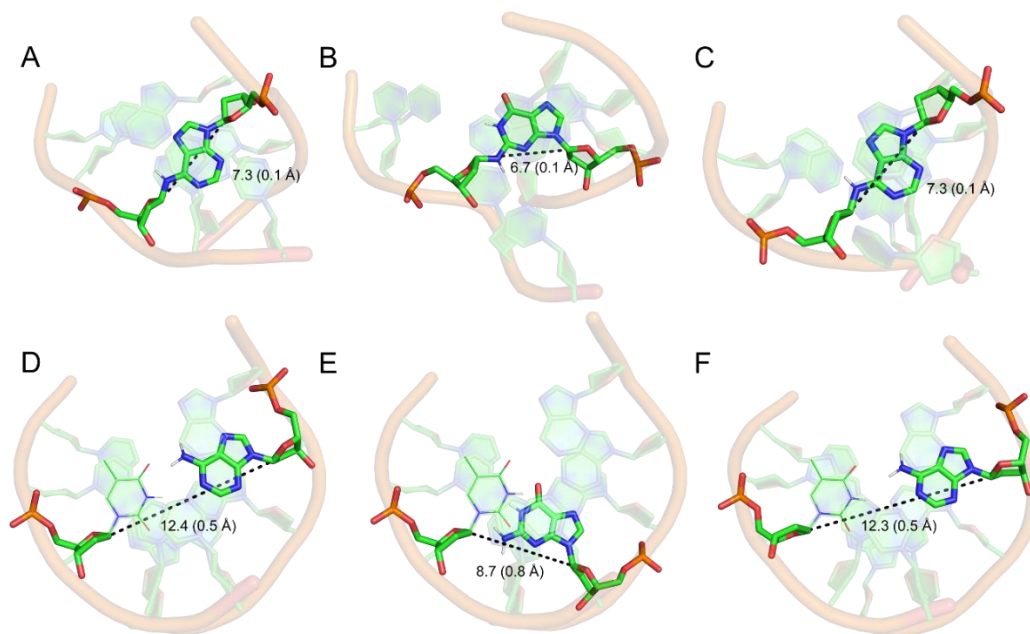


Figure 3.8: C1' – C1' distances for the (A) dA-Ap ICL and (B) dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' sequence context and (C) the dA-Ap ICL in the 5'-AAC-3'/3'-TXG-5' sequence context, along with the (D, E, F) corresponding C1'-C1' distances in natural DNA.

In comparison to natural DNA, the lesion site (base pairs 6 – 10) of the dA-Ap AAG helix had many structural differences (Figure 3.9). As dA-Ap involved the covalent linkage of residues 7 (in base pair 7) and 23 (in base pair 8), changes to the hydrogen bonding and stacking

arrangements were visible within these base pairs. While no high occupancy hydrogen bonds were present between dA-Ap and nearby residues, residues 8 and 24 (previously complementary partners of residues 7 and 23) lacked a base pairing partner and stacked with the dA-Ap ICL. Residue 24 in particular intercalated between dA-Ap and residue 6 (base pair 6). Interestingly, the base pairs that directly flanked the ICL were not consistently affected. While base pair 6 involved a loss of Watson-Crick hydrogen bonding between complementary residues 6 and 25, Watson-Crick hydrogen-bonding patterns and stacking arrangements were maintained at base pair 9. Because residue 23 was lacking a nucleobase, there was more room for residue 8 to stack with the dA-Ap without requiring rearrangement of the adjacent base pair 9. Overall, the dA-Ap ICL caused several rearrangements at the lesion site, with the most significant changes occurring at the crosslinked base pairs (7 and 8), and adjacent base pair 6.

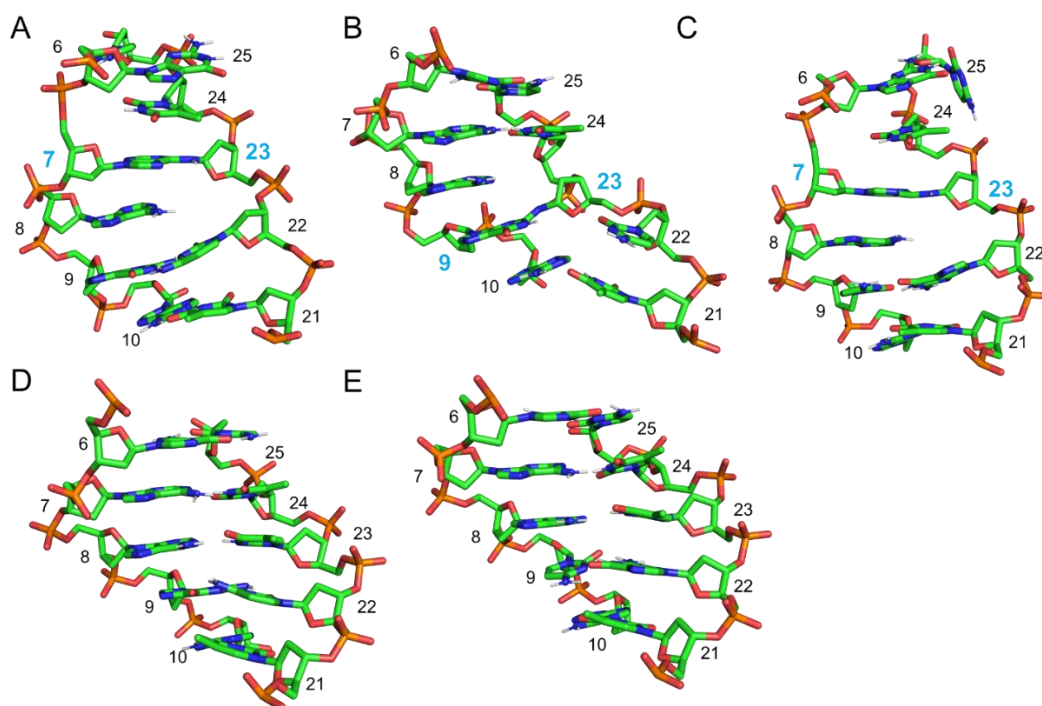


Figure 3.9: Lesions sites from the representative structures of the (A) dA-Ap AAG, (B) dG-Ap AAG, (C) dA-Ap AAC, (D) AAG natural, and (E) AAC natural models. Crosslinked residues have been bolded and coloured blue.

Like the dA-Ap ICL, the dG-Ap ICL caused a large amount of structural distortion. dG-Ap acted as a donor in a hydrogen bond with the phosphate of residue 24. This hydrogen bond appeared to cause a kink in the backbone, which propagated backbone distortion into the lesion site of dG-Ap AAG (Figure 3.9). Another interesting occurrence was the change in the sugar puckering at the Ap site portion of dG-Ap to C3'-endo (Figure 3.8). While residue 8 was unpaired and did not form strong hydrogen bonds with nearby residues, it stacked with dG-Ap and residue 7 (base pair 7). Residue 22 (base pair 9), on the other hand, was pushed outside of the helix and did not form any WC hydrogen bonds with its complementary base. While the arrangement of residues at base pair 7 was not affected in dG-Ap AAG, Watson-Crick hydrogen bonding at base pair 10 was distorted (Table B-9). In summary, the dG-Ap ICL caused severe backbone distortion at the ICL site, a change in sugar puckering at the Ap site, loss of WC hydrogen-bonding at the crosslinking base pairs, and distortion of WC hydrogen bonding at base pair 10.

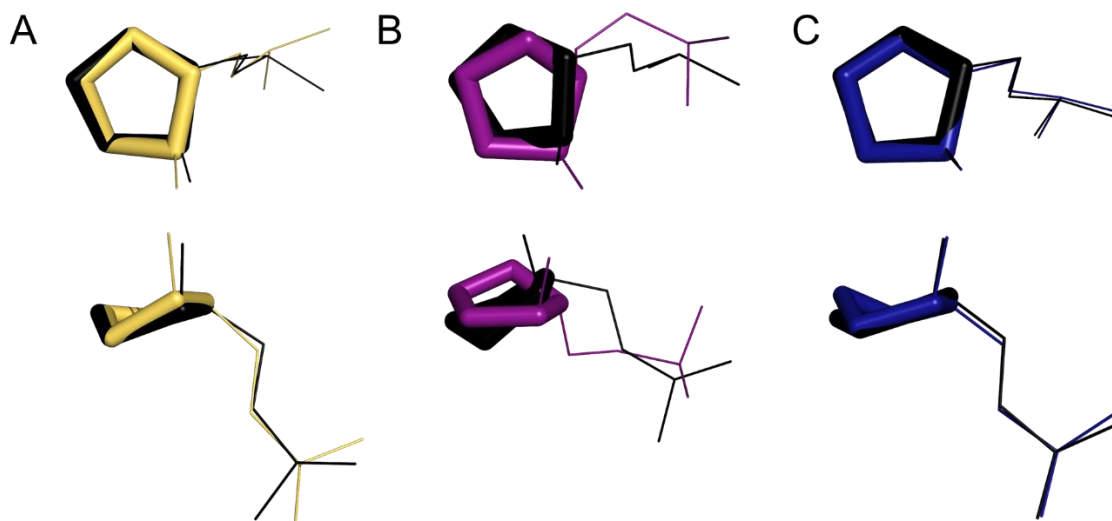


Figure 3.10: Overlay of the sugar at residue 23 in natural DNA (black) and (A, yellow) dA-Ap AAG (B, purple), dG-Ap AAG, or (C, dark blue) dA-Ap AAC.

As previous studies have correlated product formation and helical distortion, I intended to make similar connections for duplexes containing Ap-derived ICLs.^{8, 29-31} Specifically, multiple studies that discuss the structure of DNA damage have connected the preference for a DNA damage product to the degree of structural similarity between the damage product and natural DNA.^{8, 29-31} As such, products that lead to highly distorted DNA duplexes have been reported to have lower thermal stability and were typically not the preferred product. For example, a study investigating ICL formation within DNA duplexes found that DNA ICL products that more significantly distorted the duplex formed in smaller yields.⁸ Another work found that C4' oxidized Ap-derived ICL species that had previously been identified as preferred products caused smaller amounts of helical distortion and were more thermally stable.²⁹ Finally, a study of the structure of a DNA duplex containing an ϵ A:T base paired in different sequence contexts highlighted a correlation between helical distortion and the thermal stability of the duplex.³¹

Properties such as glycosidic torsion angle values, groove width, backbone torsion angles, and helical parameters were commonly used to describe B-DNA structure and differentiate B-DNA from other forms of DNA.³²⁻³⁴ As such, in the next two subsections, the ability of each damaged duplex to mimic B-DNA has been evaluated through comparison of specific structural properties (such as backbone torsion angles, glycosidic torsion angles, minor groove size, and helical parameters) for damaged DNA. Histograms were created for the minor groove size at each base pair (excluding terminal base pairs), and the α , β , γ , δ , ϵ , ζ , and χ torsion angles for every residue (excluding terminal residues) in natural and damaged duplexes. Corresponding natural and damaged histograms were then overlaid and the areas of overlap were calculated and converted (multiplied by 100%) into a value representing the B-DNA structural similarity for each property. A lesion site B-DNA similarity value was calculated for each property by averaging corresponding B-DNA similarity values for the entire lesion site (base pairs 6 – 10).

Among all properties analyzed at the lesion site for the dA-Ap AAG duplex, the minor groove was most significantly affected (base pairs 6 – 10). The calculated lesion site minor groove reproducibility value was found to be 31.3% in the dA-Ap AAG duplex (Table 3.2). The most severe minor groove distortion occurred at base pairs 6, 7, and 8, where the minor groove site was often smaller compared to natural DNA (Figure B-7 and B-8). The backbone torsion angles were second most distorted at the lesion site of dA-Ap AAG, with a backbone reproducibility of 75.3%. For dA-Ap AAG, the most severe backbone distortion was found at the crosslinking base pairs 7 (backbone torsional reproducibility value averaging 65.7%) and 8 (average reproducibility value of 73.1%; Figure B-9). However, backbone distortion decreased with an increase in the distance from dA-Ap. For example, the average backbone reproducibility

was 78.3% at base pair 6 and 85.4% at base pair 10. The most distortion to glycosidic torsional orientations occurred at the crosslinking base pairs or directly adjacent (7, 8, 9, and 22). While these residues had B-DNA glycosidic structural similarity values ranging from 46.1 – 67.5% (Figure B–10 and B–11), the average lesion site B-DNA glycosidic torsion angle similarity was 76.0%. As the distortion that occurs in the dA-Ap AAG duplex has been quantitatively described, an important observation was that for each property the most severe distortion occurred directly at or adjacent to the crosslinked base pairs. This distortion decreased further from the crosslinking base pairs.

As previously stated, the backbone kink at the crosslink site of dG-Ap AAG caused severe backbone distortion along the whole lesion site (base pairs 6 – 10), with 63.0% B-DNA backbone similarity. While crosslinking base pairs 8 and 9 had the most severe distortion (51.0 and 46.6% respectively), non-crosslinking base pairs in the lesion site also had severe distortion. For example, the average B-DNA backbone torsional similarity values were 75.3% for base pair 6 and 73.6% for base pair 10. In comparison to dA-Ap AAG, dG-Ap AAG had more severe lesion site backbone distortion (by 12.3%). In addition, the glycosidic torsional angles at the lesion site were more distorted in dG-Ap by 13.2% when compared to dA-Ap AAG. The most severe glycosidic torsional distortion occurred in residues 6, 9, 10, 24, and 22 (Figure B–12). The B-DNA glycosidic torsional similarity values for these residues ranged from 34.6 to 61.4%. In contrast to the DNA backbone and glycosidic torsional angles at the lesion site, the minor groove at the lesion site for dG-Ap AAG was less distorted than dA-Ap AAG (dA-Ap has larger B-DNA minor groove similarity value by 14.2%). Like the dA-Ap AAG duplex, the minor groove was distorted along the whole lesion site, with minor groove similarity values ranged from 28.0 to 70.8% (Figure B–8).

The alternative arrangement of the lesion site was evident through analysis of nucleobase–nucleobase non-covalent interaction (NCI) energies. While residues 6 and 7 were directly stacked in dG-Ap AAG, residue 24 in dA-Ap AAG was intercalated between residues 6 and 7. This difference in nucleobase arrangements was reflected in the nucleobase–nucleobase NCI energy between 6 and 7 for each model. The nucleobase–nucleobase NCI energy between 6 and 7 was smaller in the dA-Ap AAG duplex (−5.9 kJ/mol) than the dG-Ap AAG duplex (−38.5 kJ/mol; Figure B–14). Because residue 24 was intercalated between residue 6 and 7, more stabilization occurred between residues 6 and 24 in the dA-Ap AAG duplex (−18.8 kJ/mol; Figure B–15) than the dG-Ap AAG duplex (−0.4 kJ/mol). The largest difference in nucleobase–nucleobase NCI energies between dA-Ap AAG and dG-Ap AAG was the interaction between residues 9 and 22, which was −119.2 kJ/mol in dA-Ap AAG and 0.4 kJ/mol in dG-Ap AAG. These residues belonged to base pair 9. While in the dA-Ap AAG duplex the residues of this base pair were interacting through WC hydrogen bonds, in the dG-Ap AAG duplex there were no maintained WC hydrogen bonds at base pair 9. In addition, base pair 9 was more severely distorted in the dG-Ap AAG duplex as the glycosidic B-DNA similarity was 16.6% and 9.8% less (compared to the values in the dA-Ap AAG model) at residues 9 and 22 (Figure B–11). Additionally, the minor groove B-DNA similarity at base pair 9 was 34.2% less and the average glycosidic torsional distortion was 27.5% less in dG-Ap than in dA-Ap. Here, I saw a correlation between duplex distortion and the loss of nucleobase–nucleobase NCIs. Additionally, to get a sense of how the lesion site dynamics were affected by the presence of the dA-Ap and dG-Ap ICLs, the change in RMSF values ($\Delta\text{RMSF} = \text{RMSF}_{\text{crosslinked}} - \text{RMSF}_{\text{natural}}$) of lesion site residues were calculated. Analysis of lesion site ΔRMSF values showed that residue 22 had a significantly greater increase in mobility in the dG-Ap AAG duplex (increase of 5.6 Å; Figure

3.11) than the dA-Ap AAG duplex (increase of 1.1 Å). This increase in mobility was due to a lack of attractive interactions between residue 22 and nearby residues. This significant increase in mobility at residue 22 was related to both the lack of stabilization through the nucleobase–nucleobase interactions between residues 9 and 22, and the increased amount of helical distortion found at base pair 9. Overall, differences in nucleobase–nucleobase noncovalent interactions (between dA-Ap AAG and dG-Ap AAG) have been described and correlated with structural distortion and/or changes in lesion site dynamics. The presented data suggests that in addition to structural distortion, large changes to lesion site dynamics, and lack of stabilizing nucleobase–nucleobase NCIs make the dG-Ap ICL a less favorable product compared to the dA-Ap ICL.

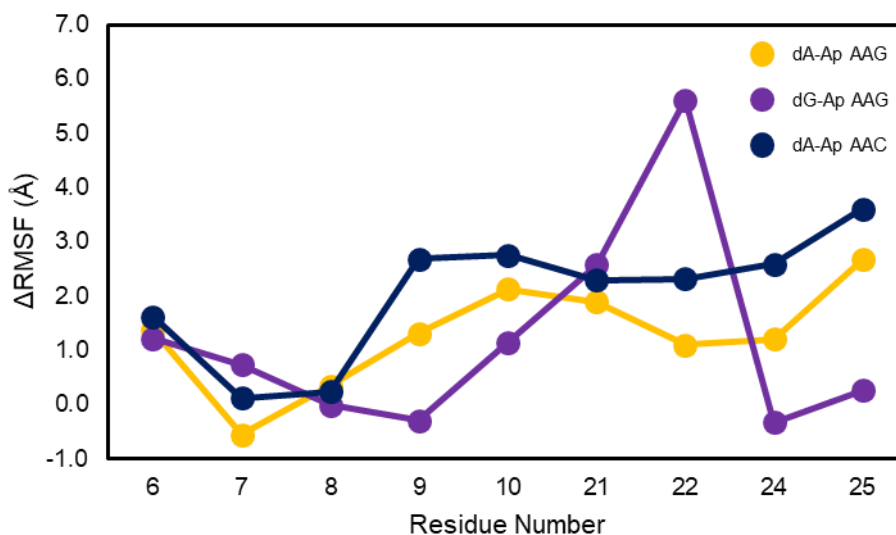


Figure 3.11: Change in RMSF (in Å) for lesion site residues of dA-Ap AAG (yellow), dG-Ap AAG (purple), and dA-Ap AAC (blue) relative to natural DNA values.

To explore the global helical distortion that occurred due to the presence of the dA-Ap or dG-Ap ICLs, the helical parameters at base pairs 4 and 12 were analyzed (Figure 3.12). These

base pairs were selected because they are at least two base pairs from the ICL and the terminal ends (which have been restrained) of the duplex. While there were differences between the averages of natural and crosslinked DNA for certain helical parameters, such as x-displacement, twist, propellor, and inclination at base pair 4, and x-displacement and buckle at base pair 12, none were statistically significant as all standard deviations show overlap. As it was clear in Figure 3.12 that the presence of any ICL caused helical distortion, the lack of statistical significance (due to overlapping standard deviations) did not allow for a strong rationale of the role that the global structure played in the preferred ICL product in the 5'-AAG-3'/3'-TXC-5' sequence context.

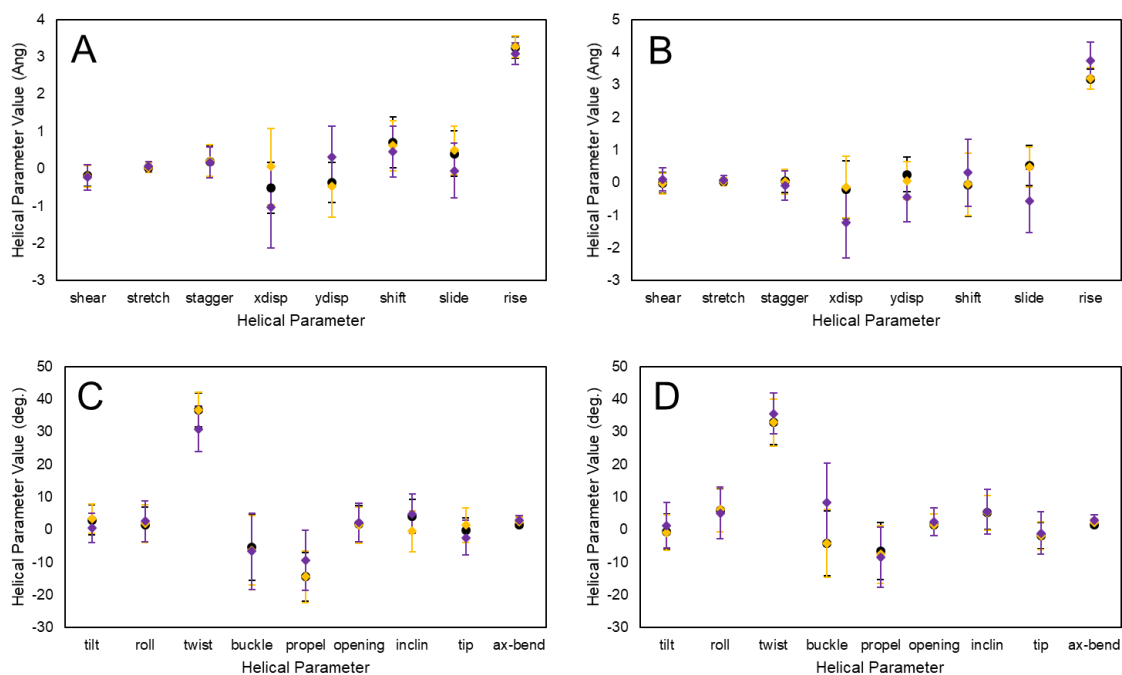


Figure 3.12: Average and standard deviation of helical parameters at base steps 4 (A, C) and 12 (B, D) in a natural DNA duplex (black circle), a DNA duplex containing the dA-Ap ICL (dark blue diamond), or the dG-Ap ICL (purple diamond) in the 5'-AAG-3'/3'-TXC-5' sequence context.

To get a detailed sense of how far the duplex distortion propagates, and how well the overall duplex structure and dynamics resembled natural DNA, B-DNA similarity values were also calculated at base pairs outside of the lesion site. Figure B-7 shows that severe minor groove distortion propagated along the duplex in the case of both ICLs. While the most severe minor groove distortion for the dA-Ap AAG duplex remained at the lesion site, the minor groove distortion outside of the lesion site was still large (52.7 – 87.9%). Similarly, the most severe minor groove distortion for the dG-Ap AAG duplex was at the lesion site. However, compared to the dA-Ap AAG duplex, there was more minor groove distortion outside of the lesion site in the dG-Ap AAG duplex (48.1 – 73.7%). While in the dA-Ap AAG duplex, backbone distortion dramatically dissipated outside of the lesion site (91.1 – 98.7%), backbone distortion in the dG-Ap AAG duplex remained outside of the lesion site (80.1 – 86.2%). Similarly, glycosidic torsion angles outside of the lesion site showed high B-DNA similarity values in the dA-Ap AAG duplex (90.3 – 99.3%), and comparatively lower values in the dG-Ap AAG duplex (39.2 – 97.1%). Overall, the structural distortion was contained near the ICL in the dA-Ap AAG duplex and spread along the helix in the dG-Ap AAG duplex.

Finally, a total duplex B-DNA similarity was calculated (minor groove size, glycosidic torsion angles, and backbone torsion angles) for each duplex by averaging the corresponding B-DNA similarity values of all base pairs for each property. Total duplex analysis of B-DNA structural similarity revealed that the dG-Ap AAG helix was significantly more distorted than the dA-Ap AAG helix. In particular, the dA-Ap AAG helix had a larger overall backbone B-DNA similarity by 13.0% (Table 3.2), a larger overall glycosidic torsion B-DNA similarity by 18.0%, and a larger overall minor groove similarity value by 5.7%. As the overall B-DNA structural similarity values were larger than the lesion site B-DNA structural similarity values, this shows

that the most severe distortion was present at the lesion site for both ICLs. Nevertheless, the distortion propagated further along the helix (Figures B–8, B–9, and B–11) when the dG-Ap ICL was present compared to that of the dA-Ap ICL, suggesting that the smaller amount of structural distortion caused as a result of dA-Ap ICL formation likely contributes to why the dA-Ap ICL was the preferred product.

Table 3.2: B-DNA backbone, minor groove, and glycosidic torsion structural similarity for the dA-Ap AAG, dG-Ap AAG, or dA-Ap AAC duplexes. ^a

Lesion	B-DNA Structural Similarity (%)					
	Lesion Site			Whole Duplex		
	Backbone	Minor Groove	Glycosidic Torsion	Backbone	Minor Groove	Glycosidic Torsion
AAG dA-Ap	75.3	31.3	76.0	88.3	59.9	89.2
AAG dG-Ap	63.0	45.5	62.8	75.3	54.2	71.2
AAC dA-Ap	73.1	21.5	79.4	87.2	56.7	90.3

^a These values were calculated through summation of histogram overlap for the property in natural DNA and damaged DNA duplexes.

In summary, formation of the dA-Ap ICL required the breaking of an A:T WC pair and subsequently the pairing T stacked with dA-Ap. In contrast, dG-Ap ICL formation broke a G:C and the pairing C flipped outside of the helix as a result of the greater distortion at the crosslinking base pairs. While both ICLs caused a large amount of structural distortion at the lesion site, distortion resulting from dG-Ap formation propagated further along the helix compared to dA-Ap ICL formation. Specifically, dG-Ap ICL formation resulted in significantly more glycosidic and backbone distortion throughout the helix. This structural distortion makes the structure less energetically favorable and contributes to why the dG-Ap ICL was the less preferred product.

3.3.5 The dA-Ap ICL causes Similar Structural Distortion in the 5'-AAG-3'/3'-TxC-5' and 5'-AAC-3'/3'-TXG-5' Sequence Contexts, but More Changes to Lesion Site Dynamics in the 5'-AAG-3'/3'-TxC-5' Sequence Context

To understand the role that duplex structure and dynamics play in the relative stability of the dA-Ap ICL in the 5'-AAG-3'/3'-TxC-5' (dA-Ap AAG) and 5'-AAC-3'/3'-TXG-5' (dA-Ap AAC) sequence contexts, MD simulations were performed on DNA duplexes containing the dA-Ap ICL in both aforementioned sequence contexts. Similar to the previous section, because the dA-Ap ICL has been experimentally resolved in a duplex (providing valuable information about connectivity and chirality), the product complexes have been selected to be modelled in this section. As the dA-Ap AAG duplex has been discussed in the previous section, the dA-Ap AAC duplex will be discussed in detail and directly compared to the dA-Ap AAG duplex.

Similar to the dA-Ap AAG duplex (7.3 Å), the distance between the sugars of the dA-Ap ICL was 7.3 Å in the dA-Ap AAC duplex, which was smaller than the corresponding value in natural DNA. Like in the dA-Ap AAG duplex, the dA-Ap (AAC) ICL did not form any strong hydrogen bonds with surrounding residues. Complementary residues 8 and 24 stacked with dA-Ap (AAC) similar to dA-Ap AAG. The Watson-Crick hydrogen bonding at base pair 6 was weakened as the hydrogen bond occupancies were lower at this site in comparison to natural DNA (by 52.7, 51.8, and 65.0 %; Table B–10). Residue 25 was flipped out of the helix and N4 of the nucleobase acted as a hydrogen-bond donor to the acceptor phosphate (non-bridging oxygen) of residue 24. The arrangement at base pair 10 was not affected as the Watson-Crick hydrogen bonding was retained and strong as the hydrogen bond occupancies were above 86.5%.

Overall, lesion site distortion was similar in magnitude between the dA-Ap AAC and dA-Ap AAG duplexes. While the lesion site B-DNA backbone and minor groove similarities were

2.2% and 8.8% less in the dA-Ap AAC duplex than the dA-Ap AAG duplex, respectively (Table 3.2), the lesion site B-DNA glycosidic torsional similarity was 3.8% higher in dA-Ap AAC. In the dA-Ap AAC duplex, the most severely distorted residues were present at the dA-Ap ICL or at residues flanking the ICL. For example, severe backbone, minor groove, and glycosidic distortion was present at base pair 6. Interestingly, the most severe backbone distortion was visible at residues 6 (base pair 6) and 8 (base pair 8, Figure B–8). The backbones of these residues were more distorted (6.1% and 6.3% less B-DNA similarity) compared to in the dA-Ap AAG duplex. The most severe minor groove distortion was observed at base pairs 6 (14.1% less B-DNA similarity) and 9 (26.5% less B-DNA similarity; Figure B–7). Additionally, all base pairs in the lesion site of the dA-Ap AAC duplex had lower B-DNA minor groove similarity values than in the dA-Ap AAG duplex. While residues 6 and 24 in the dA-Ap AAC duplex had lower B-DNA glycosidic similarity values than in the dA-Ap AAG duplex, residues 7, 8, and 9 had higher B-DNA glycosidic similarity (by 22.7%, 10.6%, and 20.0%, respectively; Figure B–11).

Differences in nucleobase–nucleobase NCI energies between the dA-Ap AAC and dA-Ap AAG were found between residues 8 and 9, 9 and 10, 8 and 22, and 10 and 22. Each of these interactions were more stable in the dA-Ap AAG duplex. Interestingly, all of these interactions involved base pair 9 (residues 9 and 22). While analysis of only B-DNA structural distortion did not directly provide answers to why the dA-Ap ICL had a larger yield in the 5'-AAG-3'/3'-TXC-5' sequence context compared to the 5'-AAC-3'/3'-TXG-5' sequence context, investigation of lesion site dynamics provided an explanation. Analysis of changes to the nucleobase RMSF in response to the presence of the ICL (Figure 3.11) showed a greater increase in mobility of residues 9 and 24 in the dA-Ap AAC duplex than the dA-Ap AAG duplex. This greater mobility

was related to the less stabilizing nucleobase–nucleobase NCIs between residues 8 and 9, 9 and 10, 8 and 22, and 10 and 22 in the dA-Ap AAC compared to the dA-Ap AAG duplex. The largest difference in NCI energies involved interactions between residues 8 and 9 (-5.0 kJ/mol in the dA-Ap AAC duplex and -25.9 kJ/mol in the dA-Ap AAG duplex; Figure B–14), and interactions between residues 9 and 10 (-3.7 kJ/mol in the dA-Ap AAC duplex and -41.0 kJ/mol in the dA-Ap AAG duplex). Nucleobase–nucleobase NCIs between 8 and 9 and 9 and 10 each involved intrastrand stacking. Previous quantum mechanical (QM) calculations on DNA nucleobase stacking suggested that stacking interactions between A and G are more thermodynamically stable than stacking interactions between A and C.³⁵ In comparison to the A and C stacking interactions present in the 5'-AAG-3'/3'-TXC-5' sequence context, the dA-Ap ICL, may have been more stabilized in the 5'-AAG-3'/3'-TXC-5' sequence context because of the A and G stacking interactions involving residues 8, 9 and 10.

Similar to the helical parameter analysis in the previous section, helical parameter distortion at base pairs 4 and 12 appeared to be present in the dA-Ap AAC helix. However, none of the observed differences in helical parameters were statistically significant.

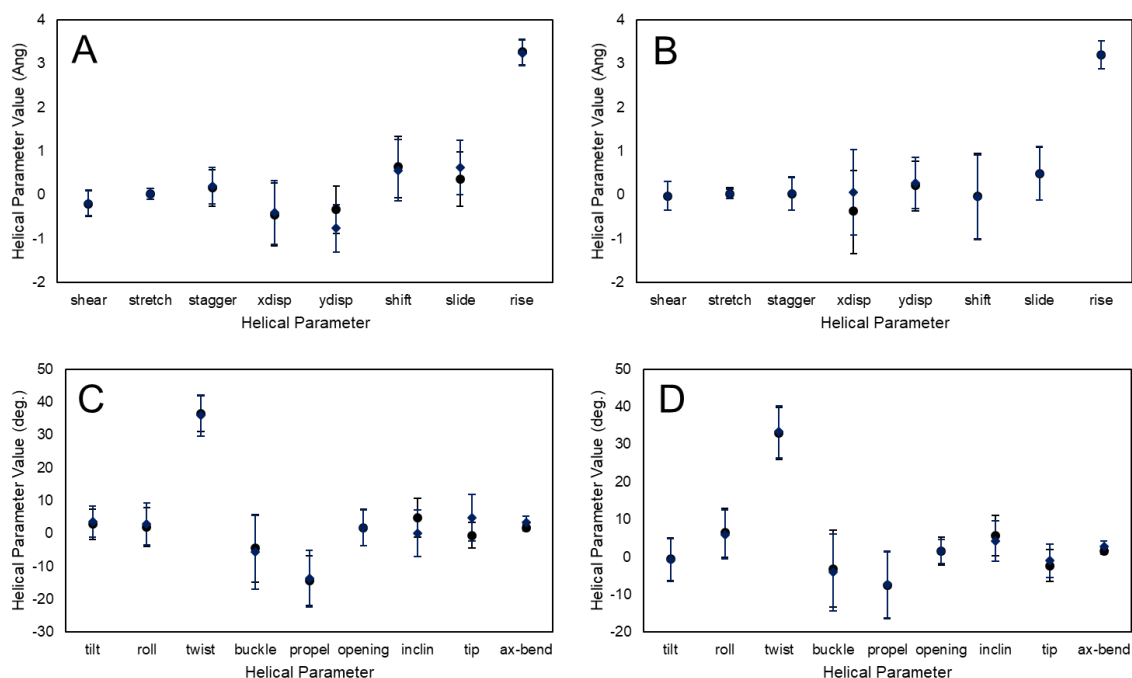


Figure 3.13: Average and standard deviation of helical parameters at base steps 4 and 12 in a natural DNA duplex (black circle) or a DNA duplex the dA-Ap ICL (yellow diamond).

Investigation of B-DNA structural similarity outside of the lesion site showed similar trends of distortion between the dA-Ap AAC and dA-Ap AAG duplexes (Figure 3.14, B–8, B–9, B–11). Specifically, the most severe backbone, minor groove, and glycosidic torsion distortion was present at and around the lesion site, and the distortion decreased progressively as the distance from the lesion site increased. Overall, trends of distortion outside of the lesion site were similar between the dA-Ap AAC and dA-Ap AAG duplexes.

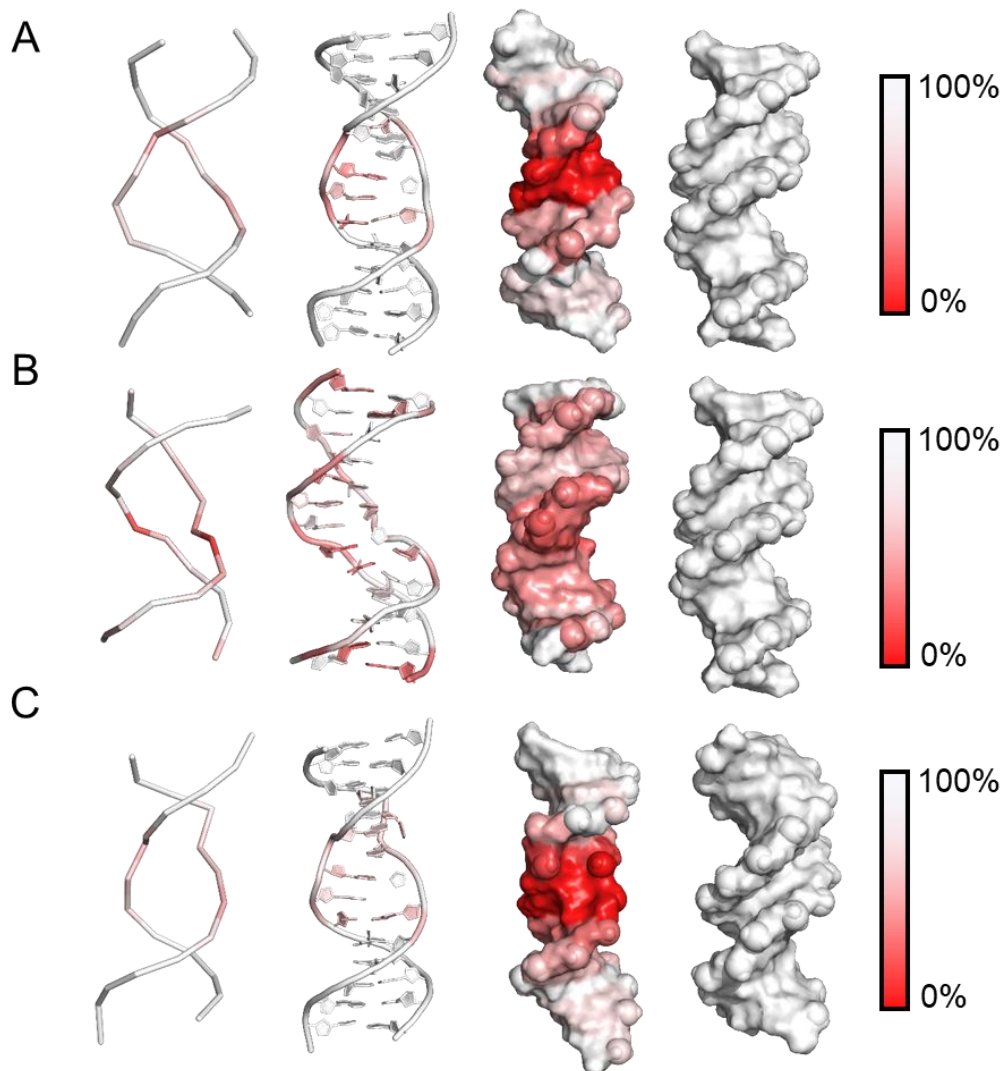


Figure 3.14 Representative structures for the: (A) dA-Ap AAG, (B) dG-Ap AAG, and (C) dA-Ap ICL AAC, with (from left, to right) severity of backbone glycosidic torsion and minor groove distortion indicated by the intensity of red.

In the dA-Ap AAC duplex, similar amounts of global helical distortion to the dA-Ap AAG duplex were observed. Between the dA-Ap AAC and dA-Ap AAG duplexes, the observed overall backbone, minor groove, and glycosidic torsion B-DNA structural similarity values differed by 1.1, 3.2, and 1.1%, respectively. This shows that the dA-Ap ICL causes similar global distortion in the two sequence contexts.

In summary, while the dA-Ap ICL causes similar amounts of helical distortion in the 5'-AAG-3'/3'-TXC-5' and 5'-AAC-3'/3'-TXG-5' sequence contexts, the change in sequence context from AAG-3'/3'-TXC-5' to 5'-AAC-3'/3'-TXG-5' alters the local noncovalent interactions at the lesion site (as correlated with the difference in magnitude of the nucleobase–nucleobase NCI energies). In addition, greater lesion site dynamics are present when the dA-Ap ICL is present in the 5'-AAC-3'/3'-TXG-5' sequence context.

3.4 Conclusions

To gain atomic level rationalization for the preference of dA-Ap ICL over dG-Ap formation and the sequence dependence of dA-Ap ICL formation, a multiscale computational investigation of the Ap-derived ICLs was performed. DFT calculations were utilized to map the mechanism of dA-Ap and dG-Ap formation outside of the DNA duplex, which revealed both an intrinsic thermodynamic and kinetic preference for dA-Ap formation. MD simulations of DNA duplexes containing the dA-Ap and dG-Ap ICLs in the 5'-AAG-3'/3'-TXC-5' sequence context showed that dA-Ap causes less severe helical distortion and less disruption of lesion site dynamics and noncovalent interactions, supporting the preference for the dA-Ap ICL over the dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' sequence context. While the decreased yield of dA-Ap in the 5'-AAC-3'/3'-TXG-5' sequence context can not be explained through differences in duplex distortion alone, the presence of the dA-Ap ICL in the 5'-AAC-3'/3'-TXG-5' sequence context resulted in a duplex with less stabilizing nucleobase–nucleobase noncovalent interactions when compared to a duplex with the 5'-AAG-3'/3'-TXC-5' sequence context. To investigate how the structure of DNA affects the mechanism of Ap-derived ICL formation, future endeavors should

consider the dA-Ap and dG-Ap ICL formation mechanisms in the context of a DNA helix and NCP using QM/MM approaches. Additionally, as ICLs have been proven to be toxic to cells,¹⁷ a study of the repair of Ap-derived ICLs must also be conducted.

3.5 References

1. Gamboa Varela, J.; Gates, K. S., A Simple, High-Yield Synthesis of DNA Duplexes Containing a Covalent, Thermally Cleavable Interstrand Cross-Link at a Defined Location. *Angew. Chem., Int. Ed. Engl.* **2015**, *54* (26), 7666-7669.
2. Varela, J. G.; Pierce, L. E.; Guo, X.; Price, N. E.; Johnson, K. M.; Yang, Z.; Wang, Y.; Gates, K. S., Interstrand Cross-Link Formation Involving Reaction of a Mismatched Cytosine Residue with an Abasic Site in Duplex DNA. *Chem. Res. Toxicol.* **2021**, *34* (4), 1124-1132.
3. Price, N. E.; Li, L.; Gates, K. S.; Wang, Y., Replication and Repair of a Reduced 2'-Deoxyguanosine-Abasic Site Interstrand Cross-Link in Human Cells. *Nucleic Acids Res.* **2017**, *45* (11), 6486-6493.
4. Johnson, K. M.; Price, N. E.; Wang, J.; Fekry, M. I.; Dutta, S.; Seiner, D. R.; Wang, Y.; Gates, K. S., On the Formation and Properties of Interstrand DNA–DNA Cross-Links Forged by Reaction of an Abasic Site with the Opposing Guanine Residue of 5'-CAP Sequences in Duplex DNA. *J. Am. Chem. Soc.* **2013**, *135* (3), 1015-1025.
5. Kellum, A. H.; Qiu, D. Y.; Voehler, M. W.; Martin, W.; Gates, K. S.; Stone, M. P., Structure of a Stable Interstrand DNA Cross-Link Involving a β -N-Glycosyl Linkage Between an N6-dA Amino Group and an Abasic Site. *Biochem.* **2021**, *60* (1), 41-52.
6. Price, N. E.; Johnson, K. M.; Wang, J.; Fekry, M. I.; Wang, Y.; Gates, K. S., Interstrand DNA–DNA Cross-Link Formation Between Adenine Residues and Abasic Sites in Duplex DNA. *J. Am. Chem. Soc.* **2014**, *136* (9), 3483-3490.
7. Grüber, R.; Dumont, É., DFT Investigation of the Formation of Linear Aminols as the First Step Toward the Induction of Oxidatively Generated Interstrand Cross-Link DNA Lesions. *Theor. Chem. Acc.* **2015**, *134* (3), 26.
8. Bignon, E.; Dršata, T.; Morell, C.; Lankaš, F.; Dumont, E., Interstrand Cross-Linking Implies Contrasting Structural Consequences for DNA: Insights From Molecular Dynamics. *Nucleic Acids Res.* **2017**, *45* (4), 2188-2195.
9. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi,

- R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.
10. D.A. Case, I. Y. B.-S., S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden,; R.E. Duke, D. G., M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang,; S. Izadi, A. K., T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J.; Mermelstein, K. M. M., Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R.; Qi, D. R. R., A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman *AMBER 2018*, University of California: San Francisco, 2018.
 11. Catalano, M. J.; Ruddraraju, K. V.; Barnes, C. L.; Gates, K. S., Crystal Structure of a Nucleoside Model for the Interstrand Cross-Link Formed By the Reaction of 2'-Deoxyguanosine and an Abasic Site in Duplex DNA. *Acta Crystallographica Section E* **2016**, 72 (5), 624-627.
 12. D.A. Case, K. B., I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro,; T.A. Darden, R. E. D., G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu,; T. Luchko, R. L., V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.; Pan, S. P., R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.; Skrynnikov, J. S., J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue,; Kollman, D. M. Y. a. P. A. *AMBER 2020*, University of California: San Francisco, 2020.
 13. Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P., The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.* **2010**, 12 (28), 7821-7839.
 14. F. Wang, J.-P. B., P. Cieplak & F.-Y. Dupradeau *R.E.D. Python: Object Oriented Programming for Amber Force Fields*, Sanford Burnham Prebys Medical Discovery Institute: 2013.
 15. Vanquelef, E.; Simon, S.; Marquand, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F.-Y., R.E.D. Server: A Web Service For Deriving RESP and ESP Charges and Building Force Field Libraries For New Molecules and Molecular Fragments. *Nucleic Acids Res.* **2011**, 39 (suppl_2), W511-W517.
 16. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints For Deriving Atomic Charges: the RESP Model. *J. Phys. Chem.* **1993**, 97 (40), 10269-10280.

17. Blanchet, C.; Pasi, M.; Zakrzewska, K.; Lavery, R., CURVES+ Web Server For Analyzing and Visualizing the Helical, Backbone and Groove Parameters of Nucleic Acid Structures. *Nucleic Acids Res.* **2011**, *39* (suppl_2), W68-W73.
18. Schrodinger, LLC, The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8. 2015.
19. Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.8. 2015.
20. Nguyen, M. T.; Uchimaru, T.; Zeegers-Huyskens, T., Protonation and Deprotonation Enthalpies of Guanine and Adenine and Implications for the Structure and Energy of Their Complexes with Water: Comparison with Uracil, Thymine, and Cytosine. *J. Phys. Chem. A* **1999**, *103* (44), 8853-8860.
21. Morell, C.; Ayers, P. W.; Grand, A.; Gutiérrez-Oliva, S.; Toro-Labbé, A., Rationalization of Diels–Alder Reactions Through the Use of the Dual Reactivity Descriptor $\Delta f(r)$. *Phys. Chem. Chem. Phys.* **2008**, *10* (48), 7239-7246.
22. Przybylski, J. L.; Wetmore, S. D., Designing an Appropriate Computational Model for DNA Nucleoside Hydrolysis: A Case Study of 2'-Deoxyuridine. *J. Phys. Chem. B* **2009**, *113* (18), 6533-6542.
23. Przybylski, J. L.; Wetmore, S. D., Modeling the Dissociative Hydrolysis of the Natural DNA Nucleosides. *J. Phys. Chem. B* **2010**, *114* (2), 1104-1113.
24. Zhang, X.; Gao, H.; Xu, H.; Xu, J.; Chao, H.; Zhao, C., A Density Functional Theory Study of the Hydrolysis Mechanism of Phosphodiester Catalyzed By a Mononuclear Zn(II) Complex. *J. Mol. Catal. A Chem.* **2013**, *368-369*, 53-60.
25. Ribeiro, A. J. M.; Ramos, M. J.; Fernandes, P. A., Benchmarking of DFT Functionals for the Hydrolysis of Phosphodiester Bonds. *J. Chem. Theory Comput.* **2010**, *6* (8), 2281-2292.
26. Katō, M.; Taniguchi, Y., Effect of Pressure on Conformational Equilibria of Liquid 1-Chloropropane and 1-Bromopropane. *J. Chem. Phys.* **1990**, *93* (6), 4345-4348.
27. Cuervo, A.; Dans, P. D.; Carrascosa, J. L.; Orozco, M.; Gomila, G.; Fumagalli, L., Direct Measurement of the Dielectric Polarization Properties of DNA. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111* (35), E3624-E3630.
28. Kaatze, U., The Dielectric Properties of Water in its Different States of Interaction. *J. Solution Chem.* **1997**, *26* (11), 1049-1112.

29. Ghosh, S.; Greenberg, M. M., Correlation of Thermal Stability and Structural Distortion of DNA Interstrand Cross-Links Produced from Oxidized Abasic Sites with Their Selective Formation and Repair. *Biochem.* **2015**, *54* (40), 6274-6283.
30. Li, C.; Zhao, X.; Liu, W.; Yin, F.; Hu, J.; Zhang, G.; Chen, G., DNA Structural Distortions Induced by a Monofunctional Trinuclear Platinum Complex with Various Cross-Links Using Molecular Dynamics Simulation. *J. Chem. Inf. Model.* **2020**, *60* (3), 1700-1708.
31. Guliaev, A. B.; Sági, J.; Singer, B., Sequence-Dependent Conformational Perturbation in DNA Duplexes Containing an ϵ A•T Mismatch Using Molecular Dynamics Simulation. *Carcinogenesis* **2000**, *21* (9), 1727-1736.
32. Jurečka, P.; Zgarbová, M.; Černý, F.; Salomon, J., Continuous B- to A-Transition in Protein-DNA Binding - How Well Is It Described by Current AMBER Force Fields? *bioRxiv* **2022**, 2022.01.13.476176.
33. Wu, M.; Yan, S.; Patel, D. J.; Geacintov, N. E.; Broyde, S., Relating Repair Susceptibility of Carcinogen-Damaged DNA With Structural Distortion and Thermodynamic Stability. *Nucleic Acids Res.* **2002**, *30* (15), 3422-3432.
34. Luo, P.; Leitzel, J. C.; Zhan, Z.-Y. J.; Lynn, D. G., Analysis of the Structure and Stability of a Backbone-Modified Oligonucleotide: Implications for Avoiding Product Inhibition in Catalytic Template-Directed Synthesis. *J. Am. Chem. Soc.* **1998**, *120* (13), 3019-3031.
35. Šponer, J.; Leszczynski, J.; Hobza, P., Electronic Properties, Hydrogen Bonding, Stacking, and Cation Binding of DNA and RNA Bases. *Biopolymers* **2001**, *61* (1), 3-31.

Chapter 4: Conclusions

4.1 Thesis Overview

The central focus of this thesis was to provide structural insight into how ICLs form following the incomplete DRR or BER of ϵ A, along with how they affect the structure of DNA and the role that sequence context plays in the formation of ICLs. Both DRR of ϵ A by ALKBH2 and BER repair initiated by AAG involve the creation of electrophilic intermediates. Specifically, ALKBH2 creates an epoxide intermediate while cleavage of the glycosidic bond by AAG yields an Ap site.¹⁻² Both intermediates can be attacked by nearby exocyclic amino groups of DNA nucleobases to form ICLs, which have been observed *in vitro* and *in vivo*.³⁻⁵ As ICLs have been characterized to be extremely toxic to cells if not repaired, it is important to understand how they form, how they affect the structure of DNA, and the role that sequence context plays in ICL formation prior to understanding the chemistry of the ICL repair process.⁶ Additionally, as ICL formation has been proven as a viable chemotherapeutic technique, the information about crosslinked DNA structure and formation can aid in the design of new therapeutics.⁶

In chapter 2, MD simulations were used to explain the effects of sequence context on trends in ALKBH2-mediated ICL yields. The Li lab at the University of Rhode Island found that attempted ALKBH2 repair of ϵ A results in ICL formation *in vitro* and *in vivo*.⁷ To derive information about how DNA sequence affects the alignment of potential electrophiles and nucleophiles, as well as connect those findings to the relative yields of ICLs, MD simulations of DNA duplexes containing the epoxide intermediate in a variety of sequence contexts were analyzed. While MD simulations revealed the presence of A, C, or G in the complementary position did not result in a significant number of reactive orientations between the epoxide and

complementary nucleobase, potentially nucleophilic exocyclic amino groups were better positioned for attack in the 5'-interstrand position. In contrast, the 3'-interstrand position was rarely aligned for ICL formation. Because no ICL was detected for the 5'-AXA-3'/3'-TTT-5' sequence context, it was concluded that due to its lack of an exocyclic amino group, thymine was not able to act as a nucleophile in ALKBH2-mediated ICL formation. The simulations revealed that the identity of the nucleobase in the 5'-interstrand position plays a dominant role in ICL yield, with the relative frequency of proper electrophile–nucleophile alignment being $A > C \gg G$. Additionally, both epoxide carbons were found to be aligned for attack. Finally, MD simulations showed that ICL yield is likely influenced by competing intrastrand crosslink formation reactions as the epoxide intermediate also forms reactive conformations with the exocyclic amino groups of A and C nucleobases present on the same strand as the epoxide.

MD simulations also revealed the possibility of intrastrand crosslink formation. When A or C was in the 3'-intrastrand position, reactive conformations conducive to intrastrand crosslink formation were frequently identified. Interestingly, when a low interstrand FRC was observed, a high intrastrand FRC was identified. Overall, the MD findings correlate with experimental data, supporting the hypothesis that the exocyclic amino groups attack the epoxide intermediate. These findings provide a connection between DNA structure and the effects of sequence context on observed ICL yield. This work justifies the use of molecular modelling to rationalize, characterize, and quantify ICL formation following attempted repair.

In chapter 3, DFT calculations and MD simulations were utilized to explain previously observed nucleobase and sequence effects on Ap-derived ICLs. A structural and chemical rationale was uncovered regarding why the dA-Ap ICL is the major product and the dG-Ap ICL is the minor product in the 5'-AAG-3'/3'-TXC-5' sequence. First, DFT calculations of truncated

nucleobase-Ap site models were used to map the dA-Ap and dG-Ap ICL formation mechanisms and revealed that dA-Ap ICL formation is intrinsically kinetically and thermodynamically preferred. Second, MD simulations of DNA duplexes containing the dA-Ap or dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' ICLs showed while both lesions cause severe distortion in the local lesion site, the dA-Ap ICL leads to less overall structural distortion of the helix, the dG-Ap ICL leads to global distortion throughout the helix. As helical distortion makes the product less thermally stable, the smaller amount of distortion caused by the dA-Ap ICL likely contributes to dA-Ap being the preferred product. In addition, MD simulations of DNA duplexes containing the dA-Ap or dG-Ap ICL in 5'-AAG-3'/3'-TXC-5' confirmed that complete loss of WC hydrogen bonding occurred upon the formation of each crosslink. While the dA-Ap ICL required the loss of an A:T WC pair, dG-Ap required the loss of a G:C WC pair, which was more energetically costly. Overall, the preferred dA-Ap ICL leads to less helical distortion, required the loss of a less energetically costly A:T WC pair.

Following the rationalization of the preference for the dA-Ap over dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' sequence context, an investigation into sequence context effects on the dA-Ap ICL yield was conducted. Specifically, a structural rationalization was provided for why the dA-Ap ICL had a 70% yield in the 5'-AAG-3'/3'-TXC-5' sequence context and a 13-18% yield in the 5'-AAC-3'/3'-TXG-5' sequence context.⁸ While there were not significant differences in duplex distortion when the dA-Ap ICL was present in each sequence context, there was suboptimal stacking at the lesion site between the residue 9 and the flanking residues in the 5'-AAC-3'/3'-TXG-5' sequence context. This loss of stabilization was likely responsible for the observed lower dA-Ap ICL yield in 5'-AAC-3'/3'-TXG-5' compared to 5'-AAG-3'/3'-TXC-5'. Overall, a multiscale computational investigation was successfully used in Chapter 3 to explain

the roles of DNA structure in the experimentally observed preference for the dA-Ap ICL over the dG-Ap ICL in the 5'-AAG-3'/3'-TXC-5' sequence, and the effects of sequence context on dA-Ap ICL formation.

In summary, this thesis demonstrates the ability of computational chemistry to explain experimental findings regarding DNA damage. More specifically, this work highlights that the structure of damaged DNA affects the relative yields of ICLs derived from attempted DNA repair. In chapter 2, the relative arrangement of the epoxide and nearby exocyclic amino groups affected observed ICL yields. In chapter 3, the intrinsic reactivity of the nucleobase toward the Ap site and the structure of the damaged product played roles in ICL yield.

4.2 Applications of Work

ICL repair has been linked to several disorders including Cockayne syndrome, trichothiodystrophy, cerebro-oculo-facio-skeletal syndrome, xeroderma pigmentosum, and FA.⁹ To understand these hereditary disorders, through the lens of various ICL repair mechanisms, it is important to first understand the structure and formation mechanism of the damage.⁶ Additionally, the creation of ICLs is a chemotherapeutic technique, with crosslinking agents such as melphalan, cyclophosphamide, cisplatin, and carboplatin being used in the treatment of various cancers including leukemia, breast, melanoma, testicular, lung, ovarian, esophageal, and bladder cancer.⁶ As a number of these compounds are toxic and have detrimental side effects such as dermatitis and bone marrow suppression, alternative methods of ICL formation that cause less concerning side effects may be valuable therapeutic approaches.⁶ To do so, the

structures and formations of known and novel ICLs must be understood.⁶ This thesis provides fundamental information about the structures and formation mechanisms of different ICLs.

4.3 Future Work

While this thesis has explained how different ICLs are formed as a result of attempted repair of ϵ A, the possibility of ICL formation resulting from attempted repair of other DNA etheno lesions by ALKBH2 should be investigated. Though ICLs have not yet been discovered along the attempted repair of 3,N⁴-ethenocytosine, 1,N²-ethenoguanine, and N²,3-ethenoguanine by ALKBH2,¹⁰ a structural investigation regarding the reactive epoxide intermediates of these lesions would provide insight into the possibility of ICL formation for each ethenoadduct. Also, as the effect of sequence context on the binding of ALKBH2 to damaged DNA has not yet been uncovered, MD studies investigating the structure and dynamics of ALKBH2 to DNA containing the epoxide intermediate can be performed (in a variety of sequence contexts). Additionally, the reaction mechanisms of ICL formation can be studied in the DNA duplex using QM and QM/MM calculations to uncover which carbon of the epoxide is preferred for nucleophilic attack by exocyclic amino groups and the energetic differences for ICL formation with different nucleobases acting as the nucleophile. Much like epoxide-derived ICL formation, the dA-Ap and dG-Ap ICL formation reactions can be investigated in a DNA duplex using either QM or QM/MM techniques. Subsequently, as various enzymes that have the ability to remove ICLs have been structurally resolved while bound to DNA, repair mechanisms of ALKBH2-mediated and Ap-derived ICLs by enzymes such as Mus81–Eme1¹¹, ERCC1¹², XPF¹², NEIL3¹³, and FANC/BRCA¹⁴ can be investigated and proposed using QM, QM/MM, and/or MD simulations.¹¹⁻¹⁴ While experimental techniques have documented the ability of cisplatin,

psoralen, and Ap site induced ICLs to block transcription and/or replication, the effects of ALKBH2-induced ICLs on transcription and replication should also be investigated in an experimental setting.^{3, 15-16} Furthermore, an atomic description of how the Ap site and ALKBH2-induced ICLs affect transcription and replication should be obtained using QM, QM/MM, and/or MD simulation. In summary, future projects can focus on the possibility of ICL formation in response to attempted repair of different etheno lesions by ALKBH2, studying the epoxide and Ap-derived ICL formation mechanisms in the context of a DNA duplex, and consider the effects of ICL formation on replication and transcription can be studied from a structural standpoint.

4.4 Final Remarks

This thesis provides insight into the formation of ICLs and the effect of ICL formation on the structure of DNA and highlights different properties that mediate ICL formation in a DNA duplex, allowing for a preliminary understanding of the reparability of these ICLs. While proximity and alignment of the electrophilic epoxide carbon atoms and exocyclic amino groups in a DNA duplex play important roles in the effects of sequence context on ALKBH2-mediated ICL formation, noncovalent interactions such as the maintenance of stacking and hydrogen-bonding interactions upon product formation also significantly affect Ap derived ICL yield in the DNA duplex. Overall, this knowledge can provide direction for further studies in the understanding of ICL-repair related diseases and the development of therapeutics that utilize ICL formation.

4.5 References

1. Rioux, K. L.; Delaney, S., 1,N6-Ethenoadenine: From Molecular to Biological Consequences. *Chem. Res. Toxicol.* **2020**, *33* (11), 2688-2698.
2. Lau Albert, Y.; Wyatt Michael, D.; Glassner Brian, J.; Samson Leona, D.; Ellenberger, T., Molecular Basis for Discriminating Between Normal and Damaged Bases by the Human Alkyladenine Glycosylase, AAG. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (25), 13573-13578.
3. Kellum, A. H.; Qiu, D. Y.; Voehler, M. W.; Martin, W.; Gates, K. S.; Stone, M. P., Structure of a Stable Interstrand DNA Cross-Link Involving a β -N-Glycosyl Linkage Between an N6-dA Amino Group and an Abasic Site. *Biochem.* **2021**, *60* (1), 41-52.
4. Romero, R. M.; Rojsitthisak, P.; Haworth, I. S., DNA Interstrand Crosslink Formation by Mechlorethamine at a Cytosine–Cytosine Mismatch Pair: Kinetics and Sequence Dependence. *Arch. Biochem.* **2001**, *386* (2), 143-153.
5. Johnson, K. M.; Price, N. E.; Wang, J.; Fekry, M. I.; Dutta, S.; Seiner, D. R.; Wang, Y.; Gates, K. S., On the Formation and Properties of Interstrand DNA–DNA Cross-Links Forged by Reaction of an Abasic Site with the Opposing Guanine Residue of 5'-CAP Sequences in Duplex DNA. *J. Am. Chem. Soc.* **2013**, *135* (3), 1015-1025.
6. Huang, Y.; Li, L., DNA Crosslinking Damage and Cancer - A Tale of Friend and Foe. *Transl. Cancer Res.* **2013**, *2* (3).
7. Wang, J., Takyi, N. A., Hsiao, Y., Tang, Q., Chen, Y., Liu, C., Lu, K., Essigmann, J. M., Wetmore, S. D., Li, D., Stable Interstrand Crosslinks Generated from the Repair of 1,N6-Ethenoadenine in DNA by the α -Ketoglutarate/Fe(II)-Dependent Dioxygenase ALKBH2. *In preparation* **2022**.
8. Price, N. E.; Johnson, K. M.; Wang, J.; Fekry, M. I.; Wang, Y.; Gates, K. S., Interstrand DNA–DNA Cross-Link Formation Between Adenine Residues and Abasic Sites in Duplex DNA. *J. Am. Chem. Soc.* **2014**, *136* (9), 3483-3490.
9. Hashimoto, S.; Anai, H.; Hanada, K., Mechanisms of Interstrand DNA Crosslink Repair and Human Disorders. *Genes Environ.* **2016**, *38* (1), 9.
10. Lenz, S. A. P.; Li, D.; Wetmore, S. D., Insights into the Direct Oxidative Repair of Etheno Lesions: MD and QM/MM Study on the Substrate Scope of ALKBH2 and AlkB. *DNA Repair* **2020**, *96*, 102944.
11. Hanada, K.; Budzowska, M.; Modesti, M.; Maas, A.; Wyman, C.; Essers, J.; Kanaar, R., The Structure-Specific Endonuclease Mus81–Eme1 Promotes Conversion of Interstrand DNA Crosslinks into Double-Strands Breaks. *EMBO J.* **2006**, *25* (20), 4921-4932.

12. Wood, R. D., Mammalian Nucleotide Excision Repair Proteins and Interstrand Crosslink Repair. *Environ. Mol. Mutagen.* **2010**, *51* (6), 520-526.
13. Imani Nejad, M.; Housh, K.; Rodriguez, A. A.; Haldar, T.; Kathe, S.; Wallace, S. S.; Eichman, B. F.; Gates, K. S., Unhooking of an Interstrand Cross-link at DNA Fork Structures by the DNA Glycosylase NEIL3. *DNA Repair* **2020**, *86*, 102752.
14. Renaudin, X.; Rosselli, F., The FANC/BRCA Pathway Releases Replication Blockades by Eliminating DNA Interstrand Cross-Links. *Genes* **2020**, *11* (5).
15. Zhu, G.; Song, L.; Lippard, S. J., Visualizing Inhibition of Nucleosome Mobility and Transcription by Cisplatin–DNA Interstrand Crosslinks in Live Mammalian Cells. *Cancer Res.* **2013**, *73* (14), 4451-4460.
16. Derheimer, F. A.; Hicks, J. K.; Paulsen, M. T.; Canman, C. E.; Ljungman, M., Psoralen-Induced DNA Interstrand Cross-Links Block Transcription and Induce p53 in an Ataxia-Telangiectasia and Rad3-Related-Dependent Manner. *Mol. Pharmacol.* **2009**, *75* (3), 599.

**Appendix A: Supplementary Information for Chapter 2: A Structural Explanation of
Sequence Effects on ALKBH2-mediated Interstrand Crosslink Formation**

Tables A-1 – A-27

Table A–1. Relative energies of truncated DNA duplexes containing an A:A, A:C, or A:G mismatch in different sequence contexts.^a

Mismatch Type	Mismatch Conformation	Sequence Context	
		5'-AAG-3'	5'-CAA-3'
A:A	<i>A_{anti}:A_{anti}</i>	0.0	9.6
	<i>A_{anti}:A_{syn}</i>	10.3	36.0
	<i>A_{syn}:A_{anti}</i>	33.8	0.0
A:C	<i>A_{anti}:C_{anti}</i>	0.0	0.0
	<i>A_{anti}:C_{syn}</i>	26.0	29.4
	<i>A_{syn}:C_{anti}</i>	85.9	16.9
A:G	<i>A_{anti}:G_{anti}</i>	0.0	0.0
	<i>A_{anti}:G_{syn}</i>	40.8	6.7
	<i>A_{syn}:G_{anti}</i>	17.6	27.6

^aSMD-B3LYP-D3(BJ)/6-31+G(d,p) single-point calculations on DNA duplex fragments optimized with AMBEROL15 (see Computational Details for additional information). The mismatch conformations selected for subsequent MD simulations are bolded.

Table A–2. Average RMSD for each MD trajectory with respect to the first frame of the corresponding simulation.^a

Sequence	Rep 1		Rep 2		Rep 3	
	Average (Å)	Standard Deviation (Å)	Average (Å)	Standard Deviation (Å)	Average (Å)	Standard Deviation (Å)
5'-TXT-3'/3'-ATA-5'	2.611	0.796	3.676	0.832	2.697	0.727
5'-AXT-3'/3'-TTA-5'	2.477	0.690	2.542	0.631	2.660	0.679
5'-CXT-3'/3'-GTA-3'	2.867	0.702	2.538	0.643	2.948	0.658
5'-GXT-3'/3'-CTA-5'	3.191	0.766	2.830	0.700	2.401	0.946
5'-TXA-3'/3'-ATT-5'	2.344	0.513	2.136	0.487	2.298	0.530
5'-AXA-3'/3'-TTT-5'	2.373	0.537	2.553	0.545	3.188	0.773
5'-CXA-3'/3'-GTT-5'	2.646	0.706	3.415	0.487	2.447	0.503
5'-GXA-3'/3'-CTT-5'	2.857	0.662	2.357	0.491	2.713	0.732
5'-TXC-3'/3'-ATG-5'	2.785	0.619	2.844	0.563	3.317	0.760
5'-AXC-3'/3'-TTG-5'	2.656	0.471	2.570	0.552	3.335	0.701
5'-CXC-3'/3'-GTG-5'	2.677	0.546	2.554	0.570	1.775	0.351
5'-GXC-3'/3'-CTG-5'	2.767	0.633	2.707	0.734	2.832	0.694
5'-TXG-3'/3'-ATC-5'	2.725	0.636	6.023	0.950	2.444	0.603
5'-AXG-3'/3'-TTC-5'	2.794	0.534	2.928	0.636	2.756	0.780
5'-CXG-3'/3'-GTC-5'	3.578	0.867	2.669	0.611	2.655	0.563
5'-GXG-3'/3'-CTC-5'	2.768	0.663	3.798	0.794	2.693	0.659
5'-TXT-3'/3'-ACA-5'	2.771	0.690	2.829	0.827	3.677	1.059
5'-TXT-3'/3'-AAA-5'	2.530	0.582	2.500	0.695	3.394	0.935
5'-TXT-3'/3'-AGA-5'	3.064	0.850	2.619	0.570	3.737	1.203
5'-AXA-3'/3'-ATT-5'	3.203	0.772	3.803	0.804	3.380	0.756
5'-AXA-3'/3'-TAT-5'	2.524	0.545	2.927	0.687	2.797	0.844
5'-AXA-3'/3'-TTA-5'	2.973	0.748	3.047	0.711	2.671	0.563
5'-AXA-3'/3'-CIT-5'	2.528	0.614	3.110	0.715	2.946	0.729
5'-AXA-3'/3'-TCT-5'	2.649	0.583	2.789	0.663	2.723	0.615
5'-AXA-3'/3'-TTC-5'	3.709	0.859	2.932	0.669	2.748	0.676
5'-AXA-3'/3'-GTT-5'	2.663	0.677	2.537	0.672	4.510	0.914
5'-AXA-3'/3'-TGT-5'	3.112	0.663	3.089	0.790	2.807	0.698
5'-AXA-3'/3'-TTG-5'	2.631	0.578	2.526	0.555	2.620	0.587

^aRMSD values were calculated over each 1 μ s of MD production simulation in reference to the corresponding first frame of production.

Table A–3. Frequency of reactive conformations (in reactive conformations/ns) for the N6 nucleophilic position of adenine in the interstrand position attacking C10 or C11 of the epoxide intermediate (X) for sequences containing 5'-XT/5'-AT.^a

Sequence ^b		C10	C11	Combined ^c
5'-T X T-3'/3'-ATA-5'	Rep 1	0.019	0.023	0.042
	Rep 2	0.015	0.024	0.039
	Rep 3	0.018	0.015	0.032
	Average	0.017	0.021	0.038
	SD	0.002	0.004	0.004
5'-A X T-3'/3'-TTA-5'	Rep 1	0.018	0.020	0.038
	Rep 2	0.012	0.018	0.030
	Rep 3	0.010	0.017	0.027
	Average	0.013	0.018	0.032
	SD	0.003	0.002	0.005
5'-C X T-3'/3'-GTA-5'	Rep 1	0.022	0.021	0.043
	Rep 2	0.026	0.003	0.029
	Rep 3	0.022	0.021	0.043
	Average	0.023	0.015	0.038
	SD	0.002	0.008	0.007
5'-G X T-3'/3'-CTA-5'	Rep 1	0.009	0.034	0.044
	Rep 2	0.006	0.034	0.040
	Rep 3	0.001	0.002	0.003
	Average	0.006	0.023	0.029
	SD	0.004	0.015	0.019

^aReactive conformations have $r(\text{N}_{\text{nuc1}}\text{C}_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(\text{N}_{\text{nuc1}}\text{C}_{\text{elec}}\text{O}_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–4. Frequency of reactive conformations (in reactive conformations/ns) for the N2 nucleophilic position of guanine in the interstrand position attacking C10 or C11 of the epoxide intermediate (X) for sequences containing 5'-XC/5'-GT.^a

Sequence ^b		C10	C11	Combined ^c
5'- TXC -3'/3'-ATG-5'	Rep 1	0.000	0.000	0.000
	Rep 2	0.001	0.001	0.002
	Rep 3	0.000	0.000	0.000
	Average	0.000	0.000	0.001
	SD	0.000	0.001	0.001
5'- AXC -3'/3'-TTG-5'	Rep 1	0.000	0.001	0.001
	Rep 2	0.000	0.000	0.000
	Rep 3	0.000	0.000	0.000
	Average	0.000	0.000	0.000
	SD	0.000	0.000	0.000
5'- CXC -3'/3'-GTG-5'	Rep 1	0.000	0.000	0.000
	Rep 2	0.000	0.000	0.000
	Rep 3	0.000	0.000	0.000
	Average	0.000	0.000	0.000
	SD	0.000	0.000	0.000
5'- GXC -3'/3'-GTG-5'	Rep 1	0.000	0.000	0.000
	Rep 2	0.000	0.004	0.004
	Rep 3	0.000	0.002	0.002
	Average	0.000	0.002	0.002
	SD	0.000	0.002	0.002

^a Reactive conformations have $r(N_{\text{nuc}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nuc}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–5. Frequency of reactive conformations (in reactive conformations/ns) for the N4 nucleophilic position of cytosine in the interstrand position attacking C10 or C11 of the epoxide intermediate (X) for sequences containing 5'-XG/3'-CT.^a

Sequence ^b		C10	C11	Combined ^c
5'- TXG -3'/3'-ATC-5'	Rep 1	0.002	0.012	0.014
	Rep 2	0.002	0.002	0.004
	Rep 3	0.002	0.006	0.008
	Average	0.002	0.007	0.009
	SD	0.000	0.004	0.004
5'- AXG -3'/3'-TTC-5'	Rep 1	0.006	0.010	0.016
	Rep 2	0.007	0.017	0.024
	Rep 3	0.005	0.012	0.017
	Average	0.006	0.013	0.019
	SD	0.001	0.003	0.004
5'- CXG -3'/3'-GTC-5'	Rep 1	0.005	0.002	0.007
	Rep 2	0.004	0.005	0.008
	Rep 3	0.008	0.004	0.011
	Average	0.005	0.003	0.009
	SD	0.002	0.001	0.002
5'- GXG -3'/3'-CTC-5'	Rep 1	0.002	0.009	0.011
	Rep 2	0.001	0.006	0.007
	Rep 3	0.002	0.009	0.011
	Average	0.002	0.008	0.010
	SD	0.001	0.002	0.002

^a Reactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–6. Frequency of reactive conformations (in reactive conformations/ns) for the N6 nucleophilic position of adenine in the intrastrand position attacking C10 or C11 of the epoxide intermediate (X) for sequences containing 5'-XA/5'-TT.^a

Sequence ^b		C10	C11	Combined ^c
5'- TXA -3'/3'-ATT-5'	Rep 1	0.025	0.437	0.462
	Rep 2	0.025	0.445	0.470
	Rep 3	0.023	0.435	0.458
	Average	0.024	0.439	0.463
	SD	0.001	0.004	0.005
5'- AXA -3'/3'-TTT-5'	Rep 1	0.017	0.348	0.364
	Rep 2	0.017	0.382	0.399
	Rep 3	0.016	0.366	0.382
	Average	0.017	0.365	0.382
	SD	0.001	0.014	0.014
5'- CXA -3'/3'-GTT-5'	Rep 1	0.045	0.432	0.477
	Rep 2	0.052	0.409	0.461
	Rep 3	0.021	0.428	0.449
	Average	0.040	0.423	0.462
	SD	0.014	0.010	0.012
5'- GXA -3'/3'-CTT-5'	Rep 1	0.022	0.387	0.342
	Rep 2	0.019	0.414	0.432
	Rep 3	0.015	0.360	0.375
	Average	0.019	0.387	0.383
	SD	0.003	0.022	0.037

^a Reactive conformations have $r(N_{\text{nuc1}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–7. Frequency of reactive conformations (in reactive conformations/ns) for the N4 nucleophilic position of cytosine in the intrastrand position attacking C10 or C11 of the epoxide intermediate (X) for sequences containing 5'-XC/5'-GT.^a

Sequence ^b		C10	C11	Combined ^c
5'- TXC -3'/3'-ATG-5'	Rep 1	0.003	0.245	0.248
	Rep 2	0.003	0.204	0.207
	Rep 3	0.004	0.248	0.252
	Average	0.003	0.232	0.235
	SD	0.001	0.020	0.021
5'- AXC -3'/3'-TTG-5'	Rep 1	0.002	0.264	0.266
	Rep 2	0.004	0.216	0.219
	Rep 3	0.005	0.214	0.219
	Average	0.003	0.231	0.235
	SD	0.001	0.023	0.022
5'- CXC -3'/3'-GTG-5'	Rep 1	0.003	0.172	0.174
	Rep 2	0.002	0.145	0.147
	Rep 3	0.003	0.184	0.186
	Average	0.003	0.167	0.169
	SD	0.000	0.016	0.017
5'- GXC -3'/3'-GTC-5'	Rep 1	0.003	0.267	0.270
	Rep 2	0.002	0.206	0.207
	Rep 3	0.002	0.263	0.264
	Average	0.002	0.245	0.247
	SD	0.001	0.028	0.028

^a Reactive conformations have $r(N_{\text{nuc1}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–8. Frequency of reactive conformations (in reactive conformations/ns) for the N2 nucleophilic position of guanine in the intrastrand position attacking C10 or C11 of the epoxide intermediate (X) for sequences containing 5'-XG/5'-CT.^a

Sequence ^b		C10	C11	Combined ^c
5'- TXG -3'/3'-ATC-5'	Rep 1	0.001	0.000	0.001
	Rep 2	0.000	0.000	0.001
	Rep 3	0.001	0.001	0.002
	Average	0.001	0.001	0.001
	SD	0.001	0.000	0.001
5'- AXG -3'/3'-TTC-5'	Rep 1	0.002	0.001	0.003
	Rep 2	0.004	0.000	0.001
	Rep 3	0.001	0.001	0.011
	Average	0.002	0.001	0.005
	SD	0.001	0.001	0.005
5'- CXG -3'/3'-GTC-5'	Rep 1	0.001	0.001	0.001
	Rep 2	0.000	0.000	0.001
	Rep 3	0.000	0.000	0.001
	Average	0.000	0.000	0.001
	SD	0.000	0.000	0.000
5'- GXG -3'/3'-CTC-5'	Rep 1	0.002	0.001	0.003
	Rep 2	0.001	0.000	0.001
	Rep 3	0.001	0.010	0.011
	Average	0.001	0.004	0.005
	SD	0.001	0.005	0.005

^a Reactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–9. Frequency of reactive conformations (in reactive conformations/ns) for the N6 nucleophilic position of adenine in the interstrand position attacking C10 or C11 of the epoxide intermediate (X) for sequences containing mismatches.^a

Sequence ^b	5' inter (Reactive Conformations/ns)		Complementary (Reactive Conformations/ns)		Combined ^c	
	C10	C11	C10	C11		
5'-T X T-3'/3'-ACA-5'	Rep 1	0.045	0.037	0.000	0.001	0.083
	Rep 2	0.030	0.032	0.000	0.000	0.072
	Rep 3	0.024	0.034	0.000	0.000	0.059
	Average	0.033	0.035	0.000	0.000	0.071
	SD	0.004	0.006	0.000	0.001	0.006
5'-T X T-3'/3'-AAA-5'	Rep 1	0.023	0.002	0.000	0.000	0.025
	Rep 2	0.024	0.016	0.000	0.002	0.039
	Rep 3	0.015	0.014	0.000	0.001	0.033
	Average	0.020	0.011	0.000	0.001	0.032
	SD	0.004	0.006	0.000	0.001	0.006
5'-T X T-3'/3'-AGA-5'	Rep 1	0.020	0.010	0.000	0.001	0.030
	Rep 2	0.041	0.004	0.000	0.001	0.044
	Rep 3	0.026	0.020	0.000	0.000	0.046
	Average	0.029	0.011	0.000	0.001	0.040
	SD	0.009	0.007	0.000	0.000	0.008

^a Reactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^b The primary potentially reactive residues considered are bolded. ^c The total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–10. Frequency of reactive conformations (in reactive conformations/ns) for the N6 nucleophilic position of adenine in the intrastrand position attacking C10 or C11 of the epoxide intermediate (X) for base screening sequences, where adenine is the primary interstrand nucleophile.^a

Sequence ^b		C10	C11	Combined ^c
5'- AXA -3'/3'-ATT-5'	Rep 1	0.013	0.129	0.142
	Rep 2	0.015	0.117	0.132
	Rep 3	0.007	0.05	0.056
	Average	0.011	0.099	0.11
	SD	0.004	0.035	0.039
5'- AXA -3'/3'-TAT-5'	Rep 1	0.028	0.273	0.301
	Rep 2	0.026	0.263	0.288
	Rep 3	0.053	0.236	0.289
	Average	0.036	0.257	0.293
	SD	0.013	0.016	0.006
5'- AXA -3'/3'-TTA-5'	Rep 1	0.002	0.17	0.172
	Rep 2	0.002	0.127	0.129
	Rep 3	0.002	0.181	0.183
	Average	0.002	0.16	0.162
	SD	0.000	0.023	0.023

^a Reactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–11. Frequency of reactive conformations (in reactive conformations/ns) for the N6 nucleophilic position of adenine in the intrastrand position attacking C10 or C11 of the epoxide intermediate (X) for base screening sequences, where cytosine is the primary interstrand nucleophile.^a

Sequence ^b		C10	C11	Combined ^c
5'-AXA-3'/3'-CTT-5'	Rep 1	0.000	0.184	0.201
	Rep 2	0.006	0.063	0.068
	Rep 3	0.006	0.06	0.065
	Average	0.004	0.102	0.111
	SD	0.003	0.058	0.063
5'-AXA-3'/3'-TCT-5'	Rep 1	0.016	0.346	0.361
	Rep 2	0.014	0.35	0.364
	Rep 3	0.016	0.362	0.377
	Average	0.015	0.352	0.367
	SD	0.001	0.007	0.007
5'-AXA-3'/3'-TTC-5'	Rep 1	0.003	0.195	0.198
	Rep 2	0.000	0.138	0.138
	Rep 3	0.002	0.133	0.134
	Average	0.002	0.155	0.157
	SD	0.001	0.029	0.030

^a Reactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–12. Frequency of reactive conformations (in reactive conformations/ns) for the N6 nucleophilic position of adenine in the intrastrand position attacking C10 or C11 of the epoxide intermediate (X) for base screening sequences, where guanine is the primary interstrand nucleophile.^a

Sequence ^b		C10	C11	Combined ^c
5'- AXA -3'/3'-TTG-5'	Rep 1	0.003	0.115	0.118
	Rep 2	0.006	0.181	0.187
	Rep 3	0.004	0.069	0.072
	Average	0.004	0.122	0.126
	SD	0.002	0.046	0.047
5'- AXA -3'/3'-TGT-5'	Rep 1	0.011	0.018	0.028
	Rep 2	0.009	0.019	0.028
	Rep 3	0.010	0.02	0.03
	Average	0.010	0.019	0.028
	SD	0.001	0.001	0.001
5'- AXA -3'/3'-GTT-5'	Rep 1	0.003	0.364	0.367
	Rep 2	0.003	0.37	0.374
	Rep 3	0.004	0.363	0.367
	Average	0.004	0.366	0.369
	SD	0.000	0.003	0.003

^a Reactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–13. Frequency of reactive conformations (in reactive conformations/ns) for the N6 nucleophilic position of adenine in the interstrand position attacking C10 or C11 of the epoxide intermediate (X) for base screening sequences, where adenine is the primary interstrand nucleophile.^a

Sequence ^b		C10	C11	Combined ^c
5'- AXA -3'/3'-ATT-5'	Rep 1	0.000	0.000	0.000
	Rep 2	0.000	0.000	0.000
	Rep 3	0.000	0.000	0.000
	Average	0.000	0.000	0.000
	SD	0.000	0.000	0.000
5'- AXA -3'/3'-TAT-5'	Rep 1	0.002	0.000	0.002
	Rep 2	0.001	0.000	0.001
	Rep 3	0.002	0.000	0.002
	Average	0.002	0.000	0.002
	SD	0.001	0.000	0.001
5'- AXA -3'/3'-TTA-5'	Rep 1	0.000	0.006	0.006
	Rep 2	0.000	0.004	0.004
	Rep 3	0.000	0.009	0.009
	Average	0.000	0.006	0.007
	SD	0.000	0.002	0.002

^a Reactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–14. Frequency of reactive conformations (in reactive conformations/ns) for the N4 nucleophilic position of cytosine in the interstrand position attacking C10 or C11 of the epoxide intermediate (X) for base screening sequences, where adenine is the primary interstrand nucleophile.^a

Sequence ^b		C10	C11	Combined ^c
5'-AXA-3'/3'-CTT-5'	Rep 1	0.000	0.000	0.000
	Rep 2	0.000	0.000	0.000
	Rep 3	0.000	0.000	0.000
	Average	0.000	0.000	0.000
	SD	0.000	0.000	0.000
5'-AXA-3'/3'-TCT-5'	Rep 1	0.001	0.000	0.001
	Rep 2	0.002	0.000	0.002
	Rep 3	0.001	0.000	0.001
	Average	0.001	0.000	0.001
	SD	0.001	0.000	0.001
5'-AXA-3'/3'-TTC-5'	Rep 1	0.001	0.002	0.003
	Rep 2	0.001	0.008	0.009
	Rep 3	0.001	0.007	0.008
	Average	0.001	0.006	0.007
	SD	0.000	0.003	0.003

^a Reactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–15. Frequency of reactive conformations (in reactive conformations/ns) for the N2 nucleophilic position of guanine in the interstrand position attacking C10 or C11 of the epoxide intermediate (X) for base screening sequences, where adenine is the primary interstrand nucleophile.^a

Sequence ^b		C10	C11	Combined ^c
5'- AXA -3'/3'- GTT -5'	Rep 1	0.000	0.000	0.000
	Rep 2	0.000	0.002	0.002
	Rep 3	0.000	0.000	0.000
	Average	0.000	0.001	0.001
	SD	0.000	0.001	0.001
5'- AXA -3'/3'- TGT -5'	Rep 1	0.000	0.000	0.000
	Rep 2	0.000	0.000	0.000
	Rep 3	0.000	0.000	0.000
	Average	0.000	0.000	0.000
	SD	0.000	0.000	0.000
5'- AXA -3'/3'- TTG -5'	Rep 1	0.000	0.000	0.000
	Rep 2	0.000	0.000	0.000
	Rep 3	0.000	0.000	0.000
	Average	0.000	0.000	0.000
	SD	0.000	0.000	0.000

^aReactive conformations have $r(N_{\text{nucl}}C_{\text{elec}}) < 3.3 \text{ \AA}$ and $\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ between 130 and 180°. ^bThe primary potentially reactive residues considered are bolded. ^cThe total frequency of reactive conformations for both C10 and C11 of the epoxide intermediate.

Table A–16. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) for interstrand reactive conformations in the 5'-XT/5'-AT containing sequences (X = epoxide intermediate).

Sequence ^a	C10					C11			
	$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'- TXT -3'/3'- ATA -5'	Rep 1	3.2	0.1	137.1	11.7	3.2	0.1	135.2	11.2
	Rep 2	3.2	0.1	134.6	11.9	3.2	0.1	139.1	11.4
	Rep 3	3.2	0.1	137.3	11.8	3.2	0.1	132.1	10.4
	Average	3.2	0.0	136.3	1.2	3.2	0.0	135.5	2.9
5'- AXT -3'/3'- TTA -5'	Rep 1	3.2	0.1	141.8	11.9	3.2	0.1	131.2	9.2
	Rep 2	3.2	0.1	132.5	9.7	3.2	0.1	142.1	11.5
	Rep 3	3.2	0.1	140.4	12.5	3.2	0.1	132.9	10.0
	Average	3.2	0.0	138.2	4.1	3.2	0.0	135.4	4.8
5'- CXT -3'/3'- GTA -5'	Rep 1	3.2	0.1	140.4	12.3	3.2	0.1	129.1	8.6
	Rep 2	3.2	0.1	132.5	10.4	3.2	0.1	140.2	11.3
	Rep 3	3.2	0.1	141.9	11.7	3.2	0.1	133.6	11.2
	Average	3.2	0.0	138.3	4.1	3.2	0.0	134.3	4.6
5'- GXT -3'/3'- CTA -5'	Rep 1	3.2	0.1	140.4	12.4	3.2	0.1	135.0	11.8
	Rep 2	3.2	0.1	137.4	11.6	3.2	0.1	135.7	11.1
	Rep 3	3.2	0.1	136.6	11.8	3.2	0.1	129.2	9.0
	Average	3.2	0.0	138.1	1.6	3.2	0.0	133.3	2.9

^a The primary potentially reactive residues considered are bolded.

Table A–17. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) for intrastrand reactive conformations in the 5'-XA/5'-TT containing sequences (X = epoxide intermediate).

Sequence ^a	C10					C11			
	$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'- TXA -3'/3'-ATT-5'	Rep 1	3.2	0.1	127.9	6.0	3.2	0.1	139.6	8.7
	Rep 2	3.2	0.1	127.8	5.8	3.1	0.1	139.4	8.6
	Rep 3	3.2	0.1	127.7	5.9	3.2	0.1	139.7	8.6
	Average	3.2	0.0	127.8	0.1	3.2	0.0	139.6	0.1
5'- AXA -3'/3'-TTT-5'	Rep 1	3.2	0.1	129.7	8.0	3.2	0.9	143.5	9.6
	Rep 2	3.2	0.1	129.1	7.4	3.2	1.2	143.4	9.6
	Rep 3	3.2	0.1	129.2	7.8	3.2	0.6	143.3	9.7
	Average	3.2	0.0	129.3	0.3	3.2	0.0	143.4	0.1
5'- CXA -3'/3'-GTT-5'	Rep 1	3.2	0.1	130.4	8.8	3.1	0.1	139.9	8.7
	Rep 2	3.2	0.1	130.9	9.1	3.1	0.1	139.8	8.6
	Rep 3	3.2	0.1	128.4	6.7	3.2	0.1	141.2	8.9
	Average	3.2	0.0	129.9	1.1	3.1	0.0	140.3	0.6
5'- GXA -3'/3'-CTT-5'	Rep 1	3.2	0.1	128.3	7.0	3.1	0.1	141.3	9.3
	Rep 2	3.2	0.1	128.4	6.4	3.2	0.1	142.5	9.5
	Rep 3	3.2	0.1	128.4	6.6	3.2	0.1	143.2	9.7
	Average	3.2	0.0	128.4	0.0	3.2	0.0	142.3	0.8

^a The primary potentially reactive residues considered are bolded.

Table A–18. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) for intrastrand reactive conformations in the 5'-XC/5'-GT containing sequences (X = epoxide intermediate).

Sequence ^a	C10					C11				
	$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)			
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.		
5'- TXC -3'/3'-ATG-5'	Rep 1	3.2	0.1	128.7	7.2	3.2	0.1	146.2	8.6	
	Rep 2	3.2	0.1	129.1	7.6	3.2	0.1	146.4	8.6	
	Rep 3	3.2	0.1	129.1	8.0	3.2	0.1	146.4	8.6	
	Average	3.2	0.0	129.0	0.2	3.2	0.0	146.3	0.1	
5'- AXC -3'/3'-TTG-5'	Rep 1	3.2	0.1	128.2	7.3	3.2	0.1	148.2	8.5	
	Rep 2	3.2	0.1	132.6	9.3	3.2	0.1	148.8	8.7	
	Rep 3	3.2	0.1	133.1	9.2	3.2	0.1	148.6	8.6	
	Average	3.2	0.0	131.3	2.2	3.2	0.0	148.5	0.2	
5'- CXC -3'/3'-GTG-5'	Rep 1	3.2	0.1	130.9	8.2	3.2	0.9	148.0	8.3	
	Rep 2	3.2	0.1	130.7	7.8	3.2	0.1	148.3	8.5	
	Rep 3	3.2	0.1	129.5	8.1	3.2	0.9	148.2	8.4	
	Average	3.2	0.0	130.4	0.6	3.2	0.0	148.2	0.1	
5'- GXC -3'/3'-GTC-5'	Rep 1	3.2	0.1	129.2	6.9	3.2	0.1	148.1	8.5	
	Rep 2	3.2	0.1	128.3	6.6	3.2	0.1	147.9	8.6	
	Rep 3	3.2	0.1	127.8	6.0	3.2	0.1	148.6	8.6	
	Average	3.2	0.0	128.4	0.6	3.2	0.0	148.2	0.3	

^aThe primary potentially reactive residues considered are bolded.

Table A–19. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) for interstrand reactive conformations in the 5'-XG/5'-CT containing sequences (X = epoxide intermediate).

Sequence ^a	C10					C11			
	$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'-TXG-3'/3'-ATC-5'	Rep 1	3.2	0.1	131.3	8.7	3.2	0.1	135.8	12.5
	Rep 2	3.2	0.1	138.1	11.8	3.2	0.0	132.6	10.3
	Rep 3	3.2	0.1	131.8	8.1	3.2	0.1	136.3	11.3
	Average	3.2	0.0	133.7	3.1	3.2	0.0	134.9	1.6
5'-AXG-3'/3'-TTC-5'	Rep 1	3.2	0.1	133.5	9.4	3.2	0.1	137.7	10.7
	Rep 2	3.2	0.1	133.7	10.0	3.2	0.1	132.1	8.9
	Rep 3	3.2	0.1	134.9	8.5	3.2	0.1	134.6	10.1
	Average	3.2	0.0	134.0	0.6	3.2	0.0	134.8	2.3
5'-CXG-3'/3'-GTC-5'	Rep 1	3.2	0.1	131.6	8.1	3.2	0.1	135.3	12.4
	Rep 2	3.2	0.1	133.4	11.1	3.2	0.1	133.6	8.0
	Rep 3	3.2	0.1	131.6	7.6	3.2	0.1	133.1	9.8
	Average	3.2	0.0	132.2	0.8	3.2	0.0	134.0	0.9
5'-GXG-3'/3'-CTC-5'	Rep 1	3.2	0.1	133.0	9.4	3.2	0.1	137.3	10.9
	Rep 2	3.2	0.1	136.5	11.3	3.2	0.1	133.4	9.1
	Rep 3	3.2	0.0	130.8	8.6	3.2	0.1	136.2	11.3
	Average	3.2	0.0	133.4	2.3	3.2	0.0	135.6	1.6

^a The primary potentially reactive residues considered are bolded.

Table A–20. Average (with standard deviation) electrophile–nucleophile (5'-interstrand) distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) for interstrand reactive conformations in the sequences containing mismatches (X = epoxide intermediate).

Sequence ^a	C10					C11			
	$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'-T X T-3'/3'-ACA-5'	Rep 1	3.2	0.1	134.6	10.6	3.2	0.1	131.4	9.1
	Rep 2	3.2	0.1	131.9	9.3	3.2	0.1	132.3	9.3
	Rep 3	3.2	0.1	133.7	10.0	3.2	0.1	137.6	13.1
	Average	3.2	0.0	133.4	1.1	3.2	0.0	133.8	2.7
5'-T X T-3'/3'-AAA-5'	Rep 1	3.2	0.1	130.6	8.6	3.2	0.1	134.2	11.4
	Rep 2	3.2	0.1	125.8	5.6	3.2	0.1	136.6	12.5
	Rep 3	3.2	0.1	137.6	12.3	3.2	0.1	136.0	11.7
	Average	3.2	0.0	131.3	4.8	3.2	0.0	135.6	1.0
5'-T X T-3'/3'-AGA-5'	Rep 1	3.2	0.1	142.6	10.8	3.2	0.1	133.6	10.3
	Rep 2	3.2	0.1	143.8	11.0	3.2	0.1	143.0	11.6
	Rep 3	3.2	0.1	143.0	11.6	3.2	0.1	144.2	12.2
	Average	3.2	0.0	143.1	0.5	3.2	0.0	140.3	4.7

^aThe primary potentially reactive residues considered are bolded.

Table A–21. Average (with standard deviation) electrophile–nucleophile (complementary) distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) for interstrand reactive conformations in the sequences containing mismatches (X = epoxide intermediate).

Sequence ^a	C10				C11				
	$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'- TXT -3'/3'-ACA-5'	Rep 1	—	—	—	—	—	—	—	—
	Rep 2	—	—	—	—	—	—	—	—
	Rep 3	—	—	—	—	3.2	0.1	151.3	11.1
	Average	—	—	—	—	3.2	0.0	151.3	0.0
5'- TXT -3'/3'-AAA-5'	Rep 1	—	—	—	—	—	—	—	—
	Rep 2	—	—	—	—	3.2	0.1	147.4	10.8
	Rep 3	—	—	—	—	—	—	—	—
	Average	—	—	—	—	3.2	0.0	147.4	0.0
5'- TXT -3'/3'-AGA-5'	Rep 1	—	—	—	—	—	—	—	—
	Rep 2	—	—	—	—	—	—	—	—
	Rep 3	—	—	—	—	—	—	—	—
	Average	—	—	—	—	—	—	—	—

^a The primary potentially reactive residues considered are bolded.

Table A–22. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) for interstrand reactive conformations in the base screening sequences, where adenine is the primary interstrand nucleophile (X = epoxide intermediate).

Sequence ^a	C10				C11				
	$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'- AXA -3'/3'-ATT-5'	Rep 1	—	—	—	—	—	—	—	—
	Rep 2	—	—	—	—	—	—	—	—
	Rep 3	—	—	—	—	—	—	—	—
	Average	—	—	—	—	—	—	—	—
5'- AXA -3'/3'-TAT-5'	Rep 1	3.2	0.1	125.2	4.5	—	—	—	—
	Rep 2	3.2	0.1	126.1	5.4	—	—	—	—
	Rep 3	3.2	0.1	125.8	5.0	—	—	—	—
	Average	3.2	0.0	125.7	0.4	—	—	—	—
5'- AXA -3'/3'-TTA-5'	Rep 1	—	—	—	—	3.2	0.1	125.8	5.0
	Rep 2	—	—	—	—	3.2	0.1	128.5	7.8
	Rep 3	—	—	—	—	3.2	0.1	129.3	8.1
	Average	—	—	—	—	3.2	0.0	127.9	1.5

^aThe primary potentially reactive residues considered are bolded.

Table A–23. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) for interstrand reactive conformations in the base screening sequences, where cytosine is the primary interstrand nucleophile (X = epoxide intermediate).

Sequence ^a	C10				C11				
	$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'- AXA -3'/3'- CTT -5'	Rep 1	—	—	—	—	—	—	—	—
	Rep 2	—	—	—	—	—	—	—	—
	Rep 3	—	—	—	—	—	—	—	—
	Average	—	—	—	—	—	—	—	—
5'- AXA -3'/3'- TCT -3'	Rep 1	3.2	0.1	125.3	4.5	—	—	—	—
	Rep 2	3.2	0.1	124.8	4.7	—	—	—	—
	Rep 3	3.2	0.1	125.3	5.1	—	—	—	—
	Average	3.2	0.0	125.1	0.2	—	—	—	—
5'- AXA -3'/3'- TTC -5'	Rep 1	—	—	—	—	3.2	0.1	128.7	8.2
	Rep 2	—	—	—	—	3.2	0.1	129.5	7.9
	Rep 3	—	—	—	—	3.2	0.1	128.9	7.7
	Average	—	—	—	—	3.2	0.0	129.0	0.3

^a The primary potentially reactive residues considered are bolded.

Table A–24. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nuc1}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$, deg.) for intrastrand reactive conformations in the base screening sequences, where adenine is the primary interstrand nucleophile (X = epoxide intermediate).

Sequence ^a	C10					C11			
	$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nuc1}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nuc1}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'- AXA -3'/3'-ATT-5'	Rep 1	3.2	0.1	130.3	9.7	3.1	0.1	142.0	11.2
	Rep 2	3.2	0.1	130.3	9.0	3.1	0.1	138.1	8.3
	Rep 3	3.2	0.1	133.7	12.1	3.2	0.1	140.8	9.4
	Average	3.2	0.0	131.4	1.6	3.1	0.0	140.3	1.6
5'- AXA -3'/3'-TAT-5'	Rep 1	3.3	0.3	129.9	9.1	3.2	0.1	141.3	8.8
	Rep 2	3.3	0.3	130.0	9.4	3.2	0.1	141.7	9.1
	Rep 3	3.2	0.2	133.5	11.0	3.2	0.1	141.1	9.7
	Average	3.3	0.0	131.1	1.7	3.2	0.0	141.4	0.2
5'- AXA -3'/3'-TTA-5'	Rep 1	3.2	0.1	129.0	8.6	3.2	0.3	151.6	11.1
	Rep 2	3.2	0.1	129.3	7.8	3.2	0.2	149.4	11.3
	Rep 3	3.2	0.1	128.4	7.0	3.2	0.3	152.0	11.1
	Average	3.2	0.0	128.9	0.4	3.2	0.0	151.0	1.1

^a The primary potentially reactive residues considered are bolded.

Table A–25. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) for intrastrand reactive conformations in the base screening sequences, where cytosine is the primary interstrand nucleophile ($X = \text{epoxide intermediate}$).

Sequence ^a	C10					C11			
	$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'- AXA -3'/3'-CTT-5'	Rep 1	3.2	0.1	133.7	10.3	3.2	0.1	144.8	9.9
	Rep 2	3.2	0.1	135.3	12.1	3.2	0.1	146.0	10.9
	Rep 3	3.2	0.1	134.5	11.1	3.2	0.1	145.1	10.6
	Average	3.2	0.0	134.5	0.7	3.2	0.0	145.3	0.5
5'- AXA -3'/3'-TCT-3'	Rep 1	3.3	0.3	128.7	7.3	3.2	0.1	143.7	9.2
	Rep 2	3.3	0.3	128.8	7.1	3.2	0.1	143.7	9.1
	Rep 3	3.3	0.3	128.8	7.2	3.2	0.1	143.5	9.0
	Average	3.3	0.0	128.8	0.0	3.2	0.0	143.6	0.1
5'- AXA -3'/3'-TTC-5'	Rep 1	3.2	0.1	136.4	12.3	3.2	0.2	157.0	9.6
	Rep 2	3.2	0.1	127.6	5.3	3.3	0.4	159.0	8.7
	Rep 3	3.2	0.1	131.0	7.3	3.3	0.4	158.1	9.5
	Average	3.2	0.0	131.7	3.6	3.3	0.1	158.0	0.8

^a The primary potentially reactive residues considered are bolded.

Table A–26. Average (with standard deviation) electrophile–nucleophile distance ($r(N_{\text{nucl}}C_{\text{elec}})$, Å) and angle of attack ($\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$, deg.) for intrastrand reactive conformations in the base screening sequences, where guanine is the primary interstrand nucleophile (X = epoxide intermediate).

Sequence ^a	C10					C11			
	$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		$r(N_{\text{nucl}}C_{\text{elec}})$ (Å)		$\angle(N_{\text{nucl}}C_{\text{elec}}O_{\text{leave}})$ (deg.)		
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	
5'- AXA -3'/3'-GTT-5'	Rep 1	3.2	0.1	129.2	7.6	3.2	0.1	147.6	9.8
	Rep 2	3.2	0.1	128.6	6.3	3.2	0.1	146.2	9.7
	Rep 3	3.2	0.1	130.5	8.8	3.2	0.1	145.2	9.9
	Average	3.2	0.0	129.4	0.8	3.2	0.0	146.3	1.0
5'- AXA -3'/3'-TGT-5'	Rep 1	3.2	0.1	142.7	12.0	3.2	0.1	143.3	11.2
	Rep 2	3.2	0.1	141.8	12.6	3.2	0.1	142.9	10.3
	Rep 3	3.2	0.1	142.4	12.2	3.2	0.1	143.7	11.4
	Average	3.2	0.0	142.3	0.4	3.2	0.0	143.3	0.3
5'- AXA -3'/3'-TTG-5'	Rep 1	3.2	0.1	126.9	5.0	3.2	0.1	147.6	9.7
	Rep 2	3.2	0.1	127.0	4.8	3.2	0.1	147.6	9.7
	Rep 3	3.2	0.1	126.8	5.0	3.2	0.1	147.5	9.7
	Average	3.2	0.0	126.9	0.1	3.2	0.0	147.6	0.0

^a The primary potentially reactive residues considered are bolded.

Table A-27. Atom types and partial charges for the epoxide intermediate used in MD simulations.

Atom Name	Atom Type	Atom Charge
O5'	OS	-0.4887
P	P	1.2123
OP1	O2	-0.7919
OP2	O2	-0.7919
O3'	OS	-0.519
N1	N*	-0.0307
C2	CQ	0.2535
H2	H5	0.1653
N3	NC	-0.6675
C4	CN	0.3634
C5	CN	0.0869
C6	CN	0.4069
N6	NB	-0.5687
N7	NB	-0.5417
C8	CK	0.0973
H8	H5	0.1981
N9	N*	0.0117
C10	CT	-0.1321
H10	H2	0.1631
C11	CT	0.2488
H11	H2	0.1181
O12	OS	-0.2833
C1'	CT	0.1036
H1'	H2	0.1723
O4'	OS	-0.3717
C2'	CT	-0.1322
H2'	HC	0.0693
H2''	HC	0.0693
C4'	CT	0.1191
H4'	H1	0.1281
C3'	CT	0.0586
H3'	H1	0.1273
C5'	CI	-0.0232
H5'	H1	0.0848
H5''	H1	0.0848

Appendix B: Supplementary Information for Chapter 3: A multiscale computational investigation of Ap-derived ICL formation

Figures B-1 – B-18 and Tables B-1 – B-12

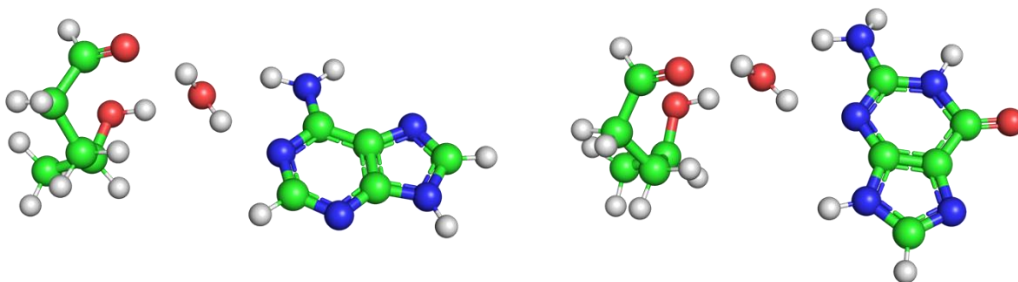


Figure B–1. DFT models used to map the reaction mechanism between an Ap site and DNA nucleobase, with (A) adenine or (B) guanine as acting nucleophile.

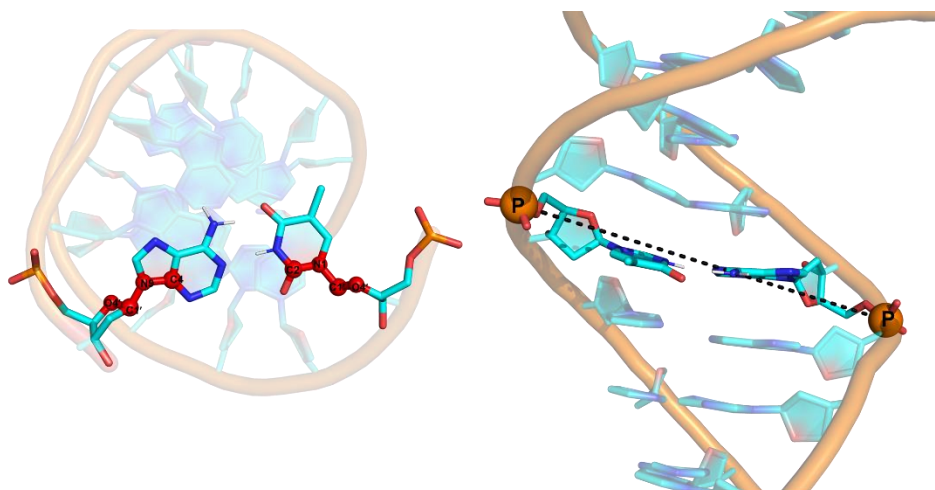


Figure B–2. Definition of DNA glycosidic torsional angle (left: $\angle(O4',C1',N1,C2)$ for pyrimidines; $\angle(O4',C1',N9,C4)$ for purines) and minor groove (right: complementary nucleotide phosphate atom–phosphate atom distances).

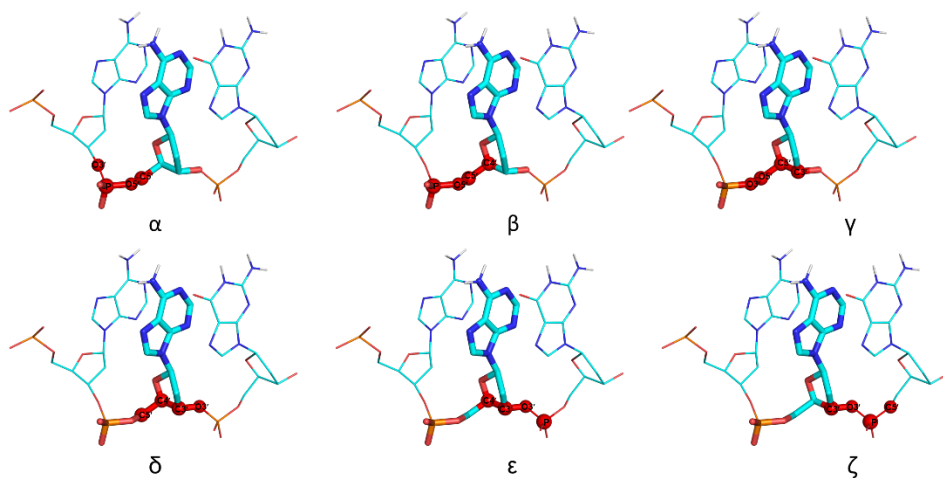


Figure B–3. Definitions of backbone torsional angles α $\angle(\text{O3}'_{n-1}, \text{P}_n, \text{O5}'_n, \text{C5}'_n)$, β $\angle(\text{P}_n, \text{O5}'_n, \text{C5}'_n, \text{C4}')$, γ $\angle(\text{O5}'_n, \text{C5}'_n, \text{C4}'_n, \text{C3}'_n)$, δ $\angle(\text{C5}'_n, \text{C4}'_n, \text{C3}'_n, \text{O3}'_n)$, ϵ $\angle(\text{C4}'_n, \text{C3}'_n, \text{O3}'_n, \text{P}_{n-1})$, and ζ $\angle(\text{C3}'_n, \text{O3}'_n, \text{P}_{n+1}, \text{O5}'_{n+1})$.

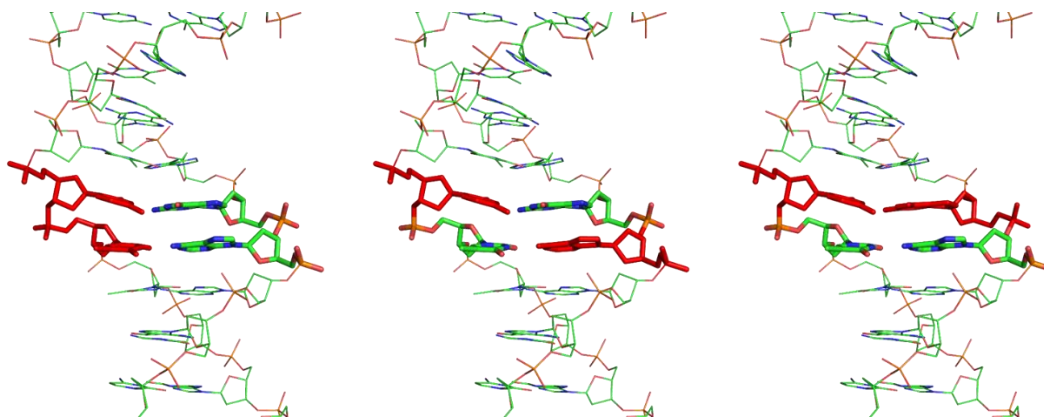


Figure B–4. Definitions of intrastrand (left), interstrand (center), and complementary (right) nucleobase–nucleobase noncovalent interactions.

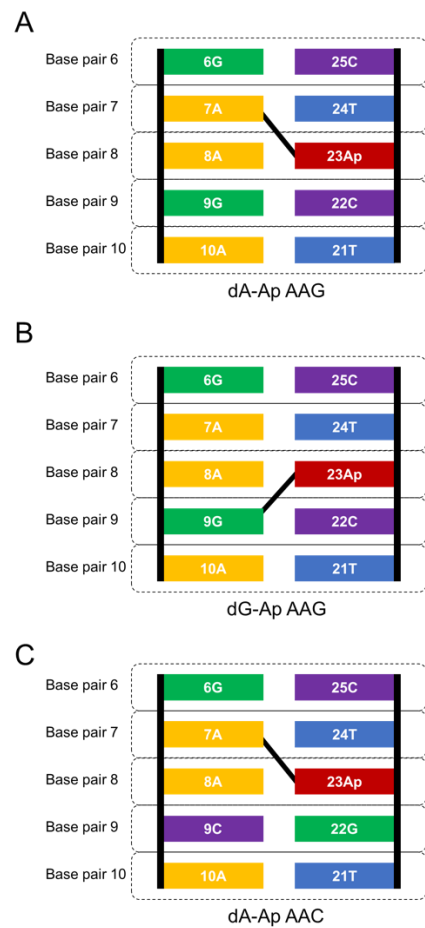


Figure B-5. Lesion site schematics for the (A) dA-Ap AAG, (B) dG-Ap AAG, and (C) dA-Ap AAC containing duplexes. Residue numbering and base pair steps have been labelled.

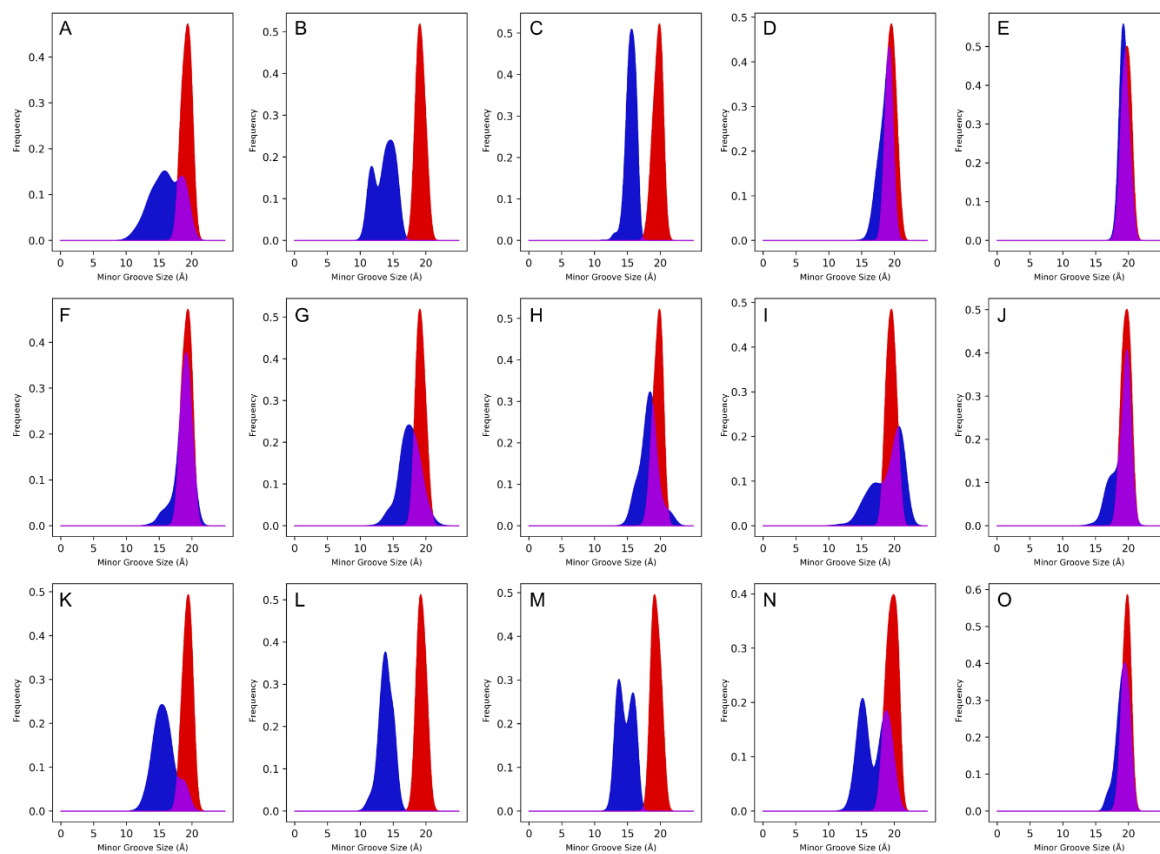


Figure B-6. Minor groove size histograms for base pairs 6–10 for natural DNA and (A-E) dA-Ap AAG, (F-J) dG-Ap AAG, and (H-L) dA-Ap AAC. The natural DNA histogram is shown in red, the damaged DNA in blue, and areas of overlap between the two are shown in purple.

Base pair	dA-Ap AAG	dG-Ap AAG	dA-Ap AAC
2–29	87.1	64.8	86.5
3–28	77.3	54.5	76.9
4–27	87.9	63.6	86.8
5–26	65.8	73.7	62.3
6–25	30.2	70.8	16.1
7–24	0.3	37.4	0.1
8–23	0.4	39.1	0.4
9–22	62.2	28	35.7
10–21	63.4	52.1	55.4
11–20	52.7	50.1	67.6
12–19	86.6	54.5	86.3
13–18	81.8	67.4	81.7
14–17	83	48.1	81.5

Figure B–7. Heat map showing structural similarity between minor groove values in natural DNA and dA-Ap AAG, dG-Ap AAG, or dA-Ap AAC. The intensity of the red is proportional to the amount of distortion at each specific base pair.

A

	2	3	4	5	6	7	8	9	10	11	12	13	14
α	90.5	99.2	94.3	98.2	96.2	32.7	85.4	85.4	97.1	94.3	96.7	97.4	98.3
β	96.5	98.8	97.2	97.7	84.4	61.5	76.2	85.2	62.6	81.5	96.0	98.5	96.4
γ	93.5	99.3	94.4	97.8	93.1	19.0	88.0	85.8	97.5	97.5	95.8	98.0	98.5
δ	99.3	99.1	95.3	95.7	79.9	76.7	93.7	91.4	94.7	85.9	99.4	98.5	98.2
ϵ	98.3	98.1	97.5	78.9	81.9	59.0	82.6	61.1	74.3	96.7	97.1	96.8	97.1
ζ	97.9	98.1	99.0	75.0	71.2	63.1	82.9	59.2	72.0	98.8	98.9	96.4	98.6
α	99.3	98.6	97.0	86.3	65.9	89.2	28.2	79.7	94.4	98.6	98.3	97.6	98.0
β	98.3	98.9	96.0	93.6	77.3	80.6	70.4	79.6	94.1	93.3	97.6	98.2	97.7
γ	98.2	99.1	97.2	90.1	77.2	96.4	23.4	98.1	98.8	97.6	98.3	99.0	98.3
δ	98.5	97.7	99.2	94.1	81.6	86.1	78.1	80.3	85.3	97.5	99.4	99.0	99.2
ϵ	97.3	98.8	97.5	93.0	71.6	49.5	88.5	37.4	70.2	91.0	95.1	96.7	98.6
ζ	98.6	98.9	98.1	93.0	59.5	74.8	79.3	47.0	83.7	97.3	95.2	97.1	97.9
Average	97.2	98.7	96.9	91.1	78.3	65.7	73.1	74.2	85.4	94.2	97.3	97.8	98.1

B

	2	3	4	5	6	7	8	9	10	11	12	13	14
α	82.4	97.2	84.9	84.6	92.3	85.5	79.1	66.5	27.9	66.2	80.5	83.3	84.4
β	65.6	97.3	95.0	93.5	86.8	72.0	69.0	8.9	77.9	66.4	93.8	92.2	92.1
γ	97.8	84.6	96.9	94.8	82.3	90.2	86.6	4.8	76.0	68.7	95.8	95.9	79.1
δ	94.0	73.7	90.6	92.2	68.7	93.1	89.0	72.6	94.5	89.1	92.3	94.8	64.5
ϵ	90.4	71.4	81.7	88.9	68.5	74.8	69.0	22.4	56.2	72.8	78.8	92.7	58.8
ζ	89.9	87.5	88.3	88.2	58.4	82.8	84.1	30.7	58.1	93.3	83.1	91.0	52.2
α	87.8	87.6	84.3	91.0	68.7	16.5	45.7	39.4	88.8	84.5	83.2	87.7	83.0
β	86.6	79.0	58.0	74.6	69.5	44.0	6.1	74.8	94.4	77.9	57.8	86.6	85.4
γ	80.9	94.4	96.5	81.7	81.0	64.9	8.6	74.4	96.6	93.4	96.7	81.1	92.9
δ	82.8	93.7	91.7	78.1	90.7	77.5	63.9	87.7	76.4	87.8	94.0	81.7	93.0
ϵ	82.1	86.3	75.6	54.9	71.9	55.8	3.6	24.3	69.1	84.2	70.8	55.7	91.6
ζ	80.5	82.1	73.3	43.7	64.2	68.2	6.6	53.2	67.5	94.8	72.0	42.1	84.6
Average	85.1	88.2	84.7	80.5	75.3	68.8	50.9	46.6	73.6	81.6	83.2	82.1	80.1

C

	2	3	4	5	6	7	8	9	10	11	12	13	14
α	94.2	99.0	94.6	85.3	73.6	28.0	85.7	76.7	76.6	97.4	96.5	97.5	95.0
β	97.8	97.1	97.9	95.5	71.9	58.9	80.5	74.0	84.7	96.5	98.0	98.7	98.8
γ	95.0	98.2	98.3	91.7	68.7	21.7	84.2	89.3	86.0	99.1	97.0	98.5	97.8
δ	98.7	95.7	92.6	85.6	69.4	78.9	91.1	84.8	92.3	97.8	98.9	99.1	99.0
ϵ	98.9	96.2	90.8	88.7	80.4	81.8	41.6	71.3	94.4	98.9	99.0	98.2	96.0
ζ	98.7	96.6	94.8	79.2	75.6	70.1	61.7	60.7	95.0	98.7	99.3	98.4	97.9
α	97.2	97.3	96.3	62.8	56.5	80.1	19.7	90.0	98.2	93.4	96.7	98.9	93.2
β	98.6	96.9	93.6	89.9	82.3	74.8	88.6	83.8	98.8	97.1	99.0	98.0	97.6
γ	98.1	99.1	96.2	76.4	78.2	76.7	18.7	97.0	96.5	95.8	97.2	98.5	95.8
δ	98.3	98.0	99.4	92.3	86.3	86.8	71.1	95.6	90.9	99.1	99.5	98.3	98.5
ϵ	96.9	97.3	93.3	94.3	70.4	32.5	78.8	38.5	79.0	95.5	99.2	99.0	97.6
ζ	97.6	96.3	96.2	91.9	53.4	48.7	79.4	58.3	69.4	97.4	99.4	99.0	97.9
Average	97.5	97.3	95.3	86.1	72.2	61.6	66.8	76.7	88.5	97.2	98.3	98.5	97.1

Figure B–8. Heat maps showing structural similarity between backbone torsion angle values of natural DNA and (A) dA-Ap AAG (B) dG-Ap (AAG) or (C) dA-Ap AAC. The intensity of the red is proportional to the amount of distortion at each specific torsion angle.

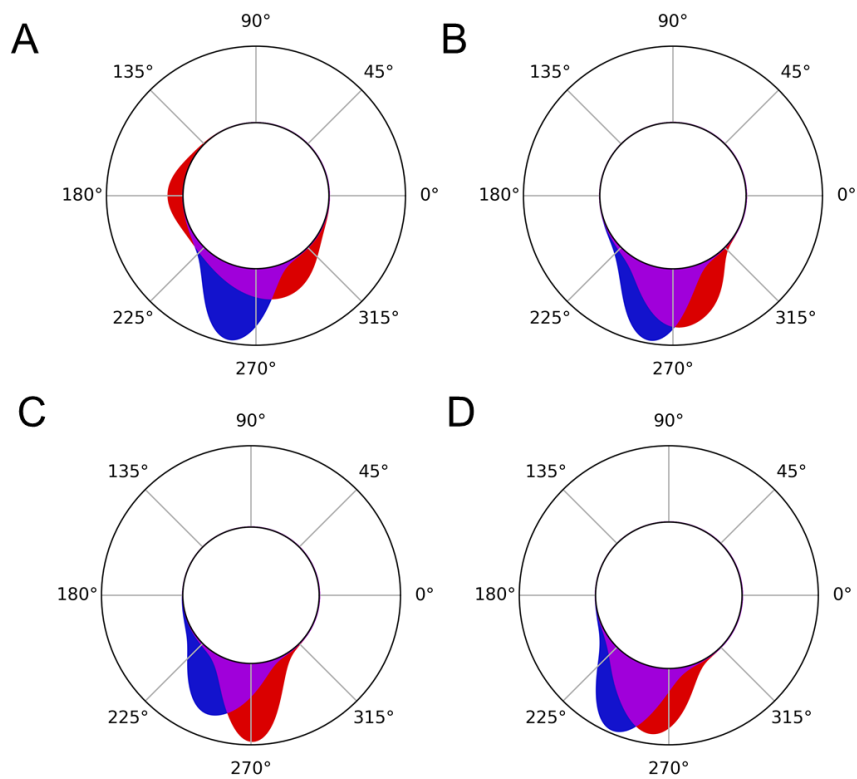


Figure B–9. Polar plots of the glycosidic torsion angles for the residues with the lowest B-DNA glycosidic torsional similarity in the dA-Ap AAG duplex: residues (A) 7, (B) 8, (C) 9, and (D) 22. The natural DNA histogram is shown in red, the damaged DNA in blue, and areas of overlap between the two are shown in purple.

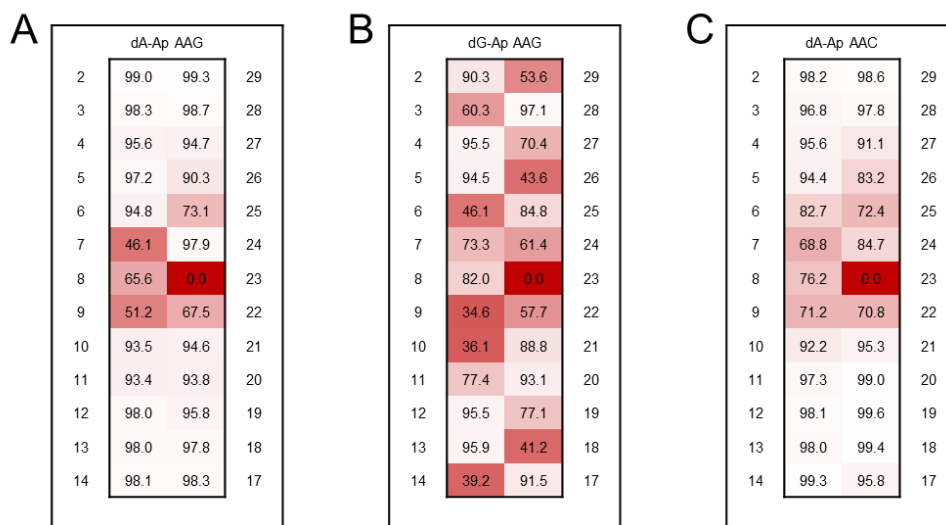


Figure B–10. Heat maps showing structural similarity between glycosidic torsion angle values of natural DNA and (A) dA-Ap AAG (B) dG-Ap (AAG) or (C) dA-Ap AAC. The intensity of the red is proportional to the amount of distortion at each specific torsion angle.

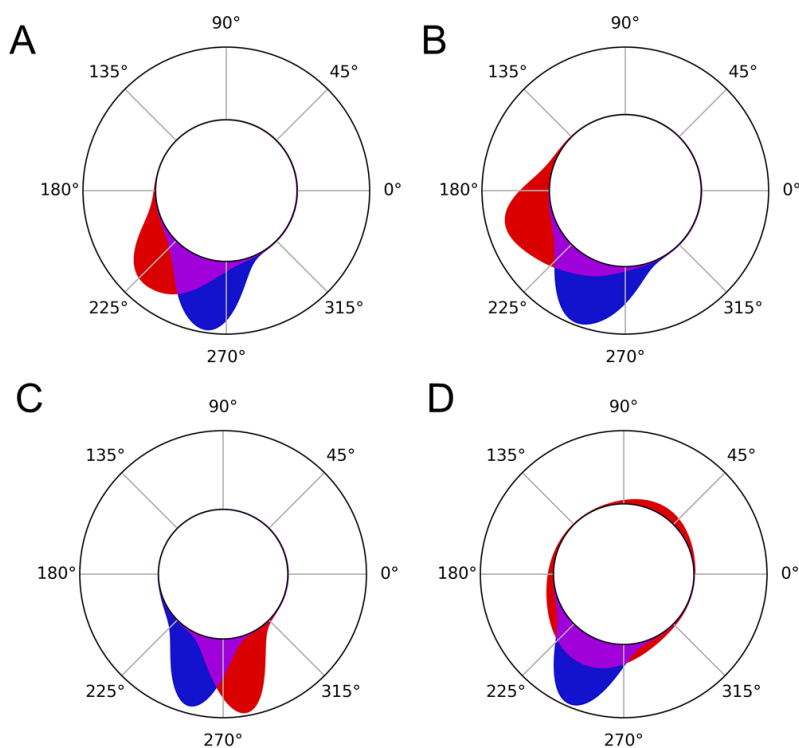


Figure B–11. Polar plots of the glycosidic torsion angles for the residues with the lowest B-DNA glycosidic torsional similarity in the dG-Ap AAG duplex: residues (A) 6, (B) 9, (C) 10, and (D) 22. The

natural DNA histogram is shown in red, the damaged DNA in blue, and areas of overlap between the two are shown in purple.

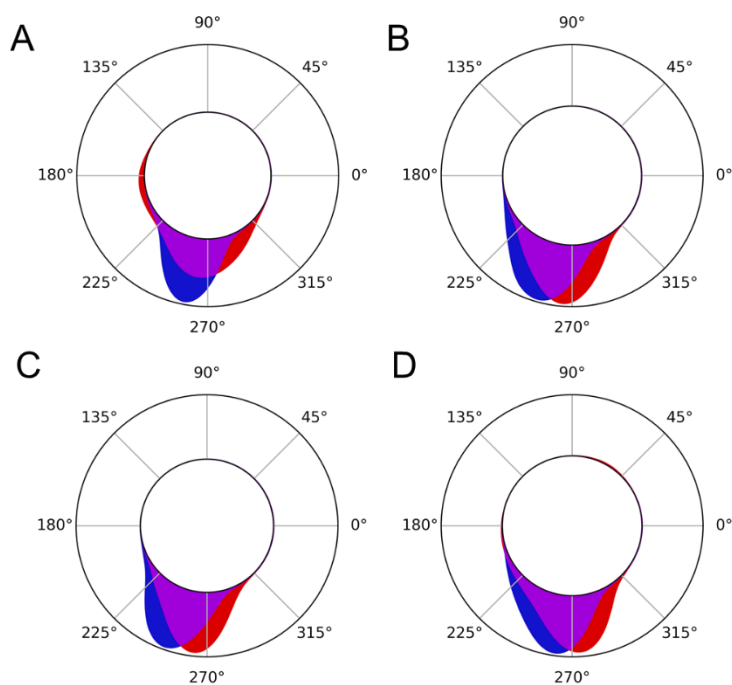


Figure B–12. Polar plots of the glycosidic torsion angles for the residues with the lowest B-DNA glycosidic torsional similarity in the dA-Ap AAG duplex: residues (A) 7, (B) 9, (C) 22, and (D) 25. The natural DNA histogram is shown in red, the damaged DNA in blue, and areas of overlap between the two are shown in purple.

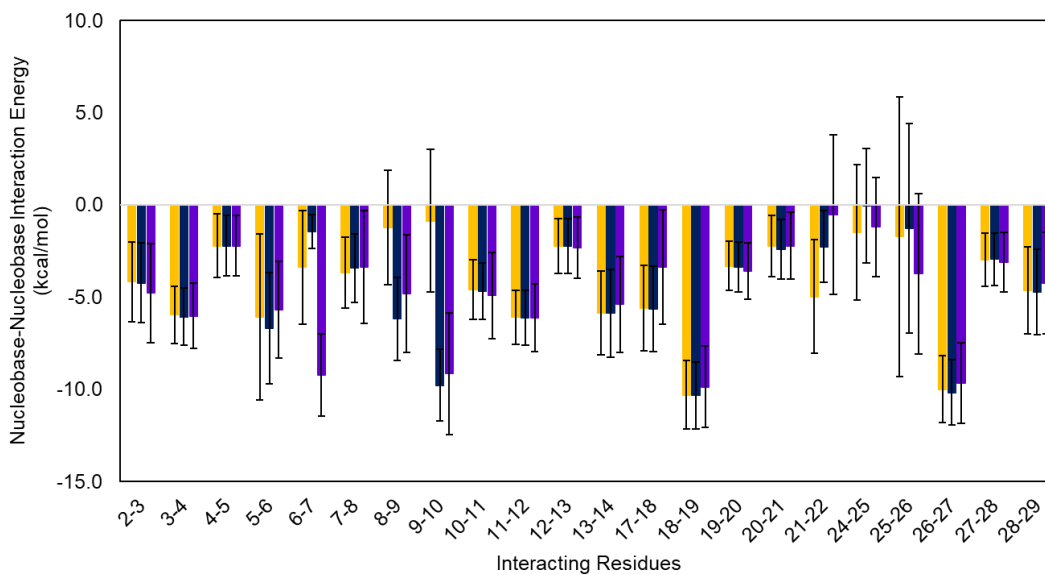


Figure B–13. Intrastrand nucleobase–nucleobase NCIs for dA-Ap AAG (yellow), dA-Ap AAC (dark blue), and dG-Ap AAG (purple).

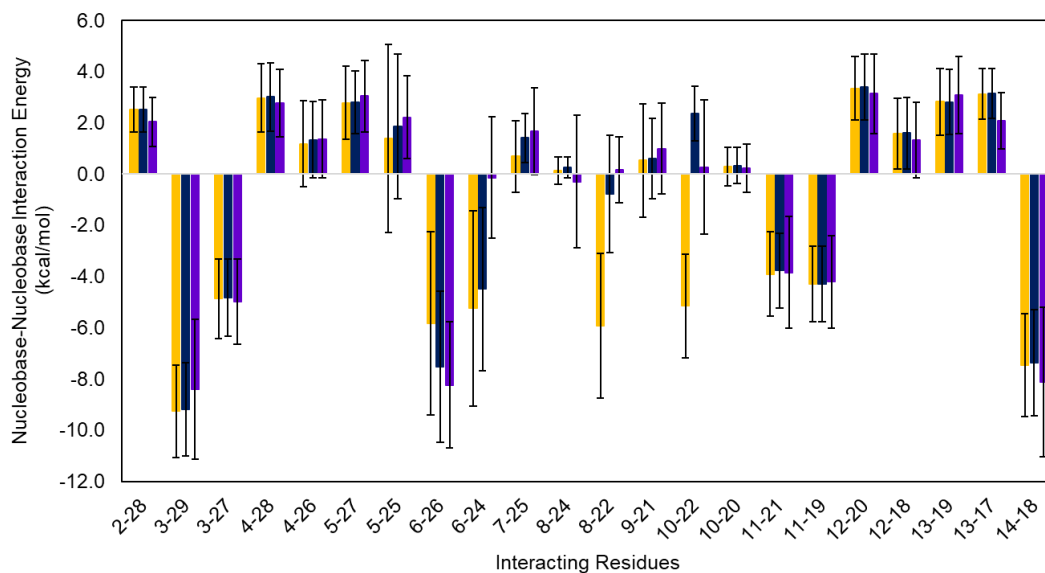


Figure B–14. Interstrand nucleobase–nucleobase NCIs for dA-Ap AAG (yellow), dA-Ap AAC (dark blue), and dG-Ap AAG (purple).

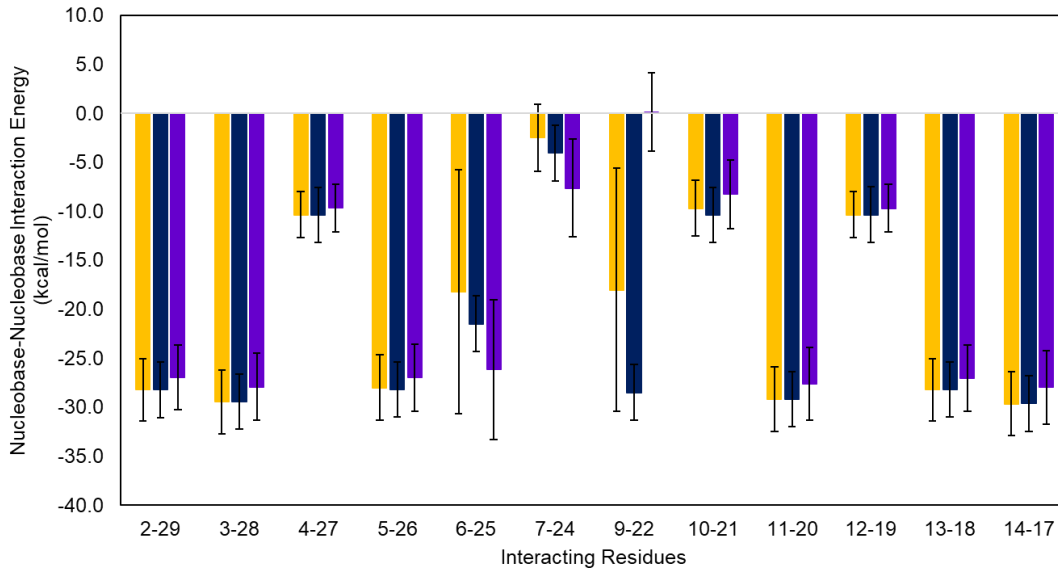


Figure B-15. Complementary nucleobase–nucleobase NCIs for dA-Ap AAG (yellow), dA-Ap AAC (dark blue), and dG-Ap AAG (purple).

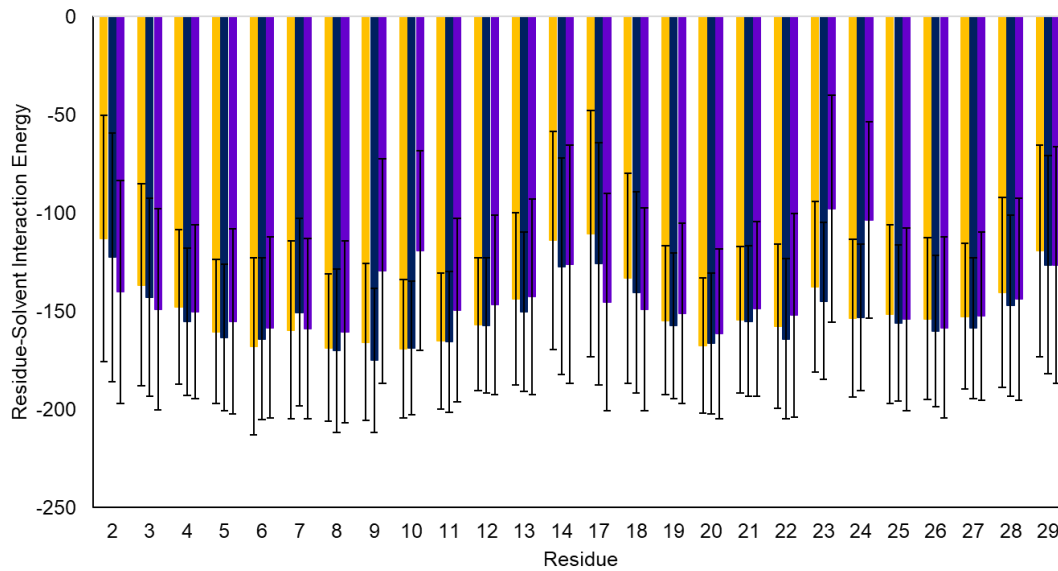


Figure B-16. DNA-solvent NCIs for dA-Ap AAG (yellow), dA-Ap AAC (dark blue), and dG-Ap AAG (purple).

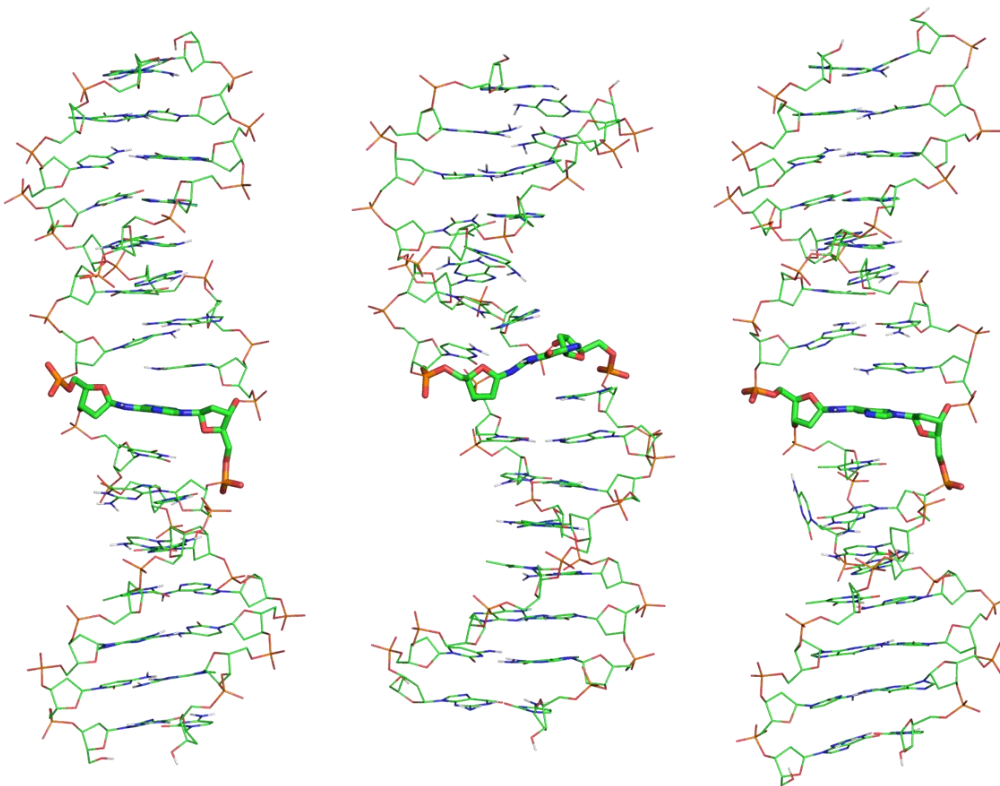


Figure B-17. Representative structures for (left) dA-Ap AAG, (center) dG-Ap AAG, and (right) dA-Ap AAG duplexes.

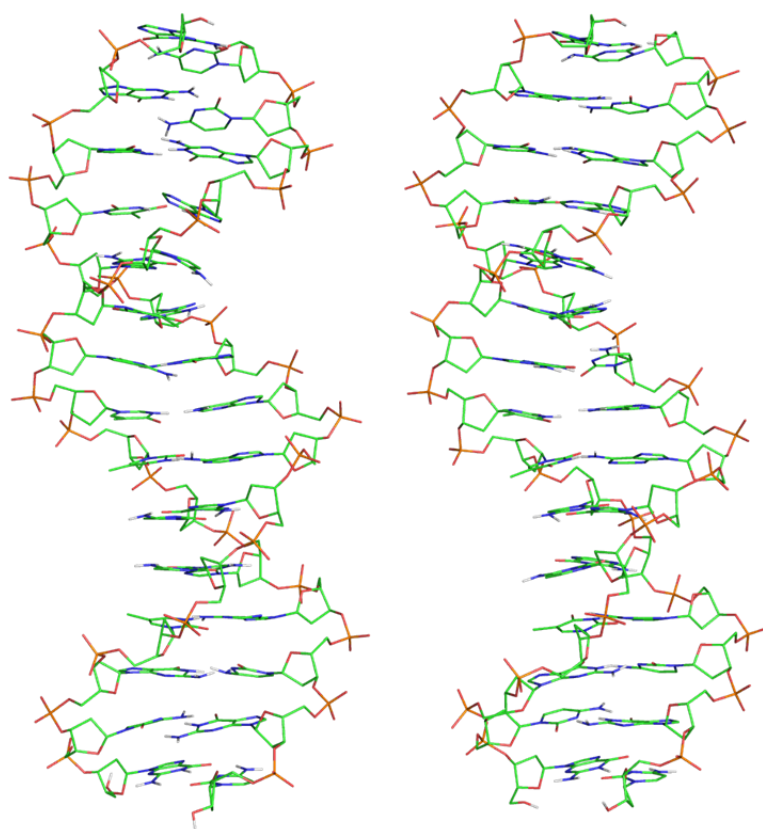


Figure B-18. Representative structures for natural (left) AAG and (right) AAC duplexes.

Table B–1. Important distances (Å) for step 1 of dA-Ap and dG-Ap ICL formation.^a

System	Stationary Point	Distance (Å)					
		R(C1'-N)	R(C1'-O1')	R(N-HN1)	R(O-HN1)	R(O-H1)	R(O1'-H1)
dA-Ap	RC	6.166	1.220	1.020	2.004	0.977	1.926
	TS1	1.661	1.340	1.211	1.324	1.182	1.267
	IC1	1.489	1.414	2.200	0.977	1.896	0.985
dG-Ap	RC	6.138	1.222	1.022	1.967	0.978	1.909
	TS1	1.679	1.333	1.231	1.297	1.168	1.285
	IC1	1.495	1.411	2.187	0.977	1.890	0.985

^a Geometries were obtained using gas-phase B3LYP/6-31G(d) optimizations.

Table B–2. Important distances (Å) for step 2 of dA-Ap and dG-Ap ICL formation.^a

System	Stationary Point	Distance (Å)			
		R(C1'-O1')	R(C1'-N)	R(N-HN)	R(O1'-HN2)
dA-Ap	IC1	1.414	1.489	1.018	2.332
	TS2	1.563	1.440	1.476	1.142
	IC2	4.000	1.287	3.851	0.969
dG-Ap	IC1	1.411	1.495	1.019	2.308
	TS2	1.549	1.443	1.470	1.153
	IC2	4.314	1.288	4.297	0.097

^a Geometries were obtained using gas-phase B3LYP/6-31G(d) optimizations.

Table B–3. Important distances (Å) for step 3 of dA-Ap and dG-Ap ICL formation.^a

System	Stationary Point	Distance (Å)							
		R(N-C1')	R(N-HN1)	R(HN1-O)	R(O-H1)	R(H1-O1')	R(O4'-C1')	R(O4'-H4')	R(H4'-O1')
dA-Ap	IC2	1.287	1.989	0.983	1.753	0.996	2.675	0.991	1.768
	TS3	1.391	1.674	1.026	1.392	1.098	1.601	1.229	1.208
	PC	1.443	3.505	0.970	0.984	1.870	1.450	1.860	0.983
dG-Ap	IC2	1.288	1.831	0.993	1.807	0.984	2.684	0.987	1.796
	TS3	1.336	1.098	1.492	1.364	1.099	1.905	1.078	1.429
	PC	1.440	1.018	1.955	0.992	1.747	1.458	1.834	0.986

^a Geometries were obtained using gas-phase B3LYP/6-31G(d) optimizations.

Table B–4. Important angles (deg.) for RC, TS1, and IC1 for the first step of dA-Ap and dG-Ap ICL formation. ^a

System	Atoms	Stationary Point (deg.)		
		RC	TS1	IC1
dA-Ap	$\angle(\text{N},\text{C1}',\text{O1}')$	40.6	104.4	106.0
	$\angle(\text{C1}',\text{O1}',\text{H1})$	120.2	105.1	106.8
	$\angle(\text{O1}',\text{H1},\text{O})$	149.1	155.4	151.4
	$\angle(\text{H1},\text{O},\text{HN1})$	114.8	83.7	88.8
	$\angle(\text{O},\text{HN1},\text{N})$	143.3	153.5	134.4
	$\angle(\text{HN1},\text{N},\text{C1}')$	22.9	96.9	95.0
dG-Ap	$\angle(\text{N},\text{C1}',\text{O1}')$	42.7	104.6	106.1
	$\angle(\text{C1}',\text{O1}',\text{H1})$	118.8	105.3	107.2
	$\angle(\text{O1}',\text{H1},\text{O})$	147.7	154.8	150.5
	$\angle(\text{H1},\text{O},\text{HN1})$	115.5	84.9	89.8
	$\angle(\text{O},\text{HN1},\text{N})$	147.8	153.2	132.9
	$\angle(\text{HN1},\text{N},\text{C1}')$	19.6	96.7	95.0

^a Geometries were obtained using gas-phase B3LYP/6-31G(d) optimizations.

Table B–5. Important angles (deg.) for IC1, TS2, and IC2 for the second step of dA-Ap and dG-Ap ICL formation. ^a

System	Atoms	Stationary Point (deg.)		
		IC1	TS2	IC2
dA-Ap	$\angle(\text{N},\text{C1}',\text{O1}')$	106.0	92.8	42.2
	$\angle(\text{C1}',\text{O1}',\text{HN1})$	61.6	78.0	21.2
	$\angle(\text{O1}',\text{HN1},\text{N})$	76.7	111.8	153.1
	$\angle(\text{HN1},\text{N},\text{C1}')$	110.2	72.9	21.2
	$\angle(\text{N},\text{C1}',\text{O1}')$	106.1	93.4	36.4
dG-Ap	$\angle(\text{C1}',\text{O1}',\text{HN1})$	61.4	77.9	20.4
	$\angle(\text{O1}',\text{HN1},\text{N})$	78.1	111.7	151.2
	$\angle(\text{HN1},\text{N},\text{C1}')$	107.7	72.7	123.7

^a Geometries were obtained using gas-phase B3LYP/6-31G(d) optimizations.

Table B–6. Important angles (deg.) for IC2, TS3, and PC for the third step of dA-Ap and dG-Ap ICL formation.^a

System	Atoms	Stationary Point (deg.)		
		IC2	TS3	PC
dA-Ap	∠(N,C1',O4')	109.5	107.7	109.7
	∠(C1',O4',H4')	96.3	110.2	115.0
	∠(O4',H4',O1')	175.4	173.8	164.1
	∠(H4',O1',H1)	107.9	103.3	98.7
	∠(O1',H1,O)	169.9	166.3	158.8
	∠(H1,O,HN1)	98.4	91.5	90.4
	∠(O,HN1,N)	153.1	155.9	167.0
	∠(HN1,N,C1')	117.7	121.4	116.9
dG-Ap	∠(N,C1',O4')	107.7	108.4	111.3
	∠(C1',O4',H4')	98.6	105.5	116.2
	∠(O4',H4',O1')	170.4	167.6	163.7
	∠(H4',O1',H1)	109.2	106.3	102.5
	∠(O1',H1,O)	165.8	166.1	161.2
	∠(H1,O,HN1)	103.3	95.7	94.7
	∠(O,HN1,N)	151.2	151.9	147.9
	∠(HN1,N,C1')	123.7	120.3	116.2

^a Geometries were obtained using gas-phase B3LYP/6-31G(d) optimizations.

Table B–7. Relative energies (in kJ/mol) of the stationary points for dA-Ap and dG-Ap formation calculated using various methods and environments. ^a

Stationary Point	Solvent Environment	Method	Relative Energy (kJ/mol)	
			dA-Ap	dG-Ap
RC	Gas	B3LYP	0.0	0.0
		B3LYP-D3(BJ)	0.0	0.0
		M06-2X	0.0	0.0
	1-Bromopropane	MP2	0.0	0.0
		M06-2X	0.0	0.0
		M06-2X	0.0	0.0
TS1	Gas	B3LYP	176.9	212.2
		B3LYP-D3(BJ)	158.4	193.6
		M06-2X	142.8	177.9
	1-Bromopropane	MP2	134.6	171.3
		M06-2X	131.1	169.4
		M06-2X	128.7	167.5
IC1	Gas	B3LYP	62.5	98.6
		B3LYP-D3(BJ)	48.2	83.7
		M06-2X	21.4	55.4
	1-Bromopropane	MP2	21.1	57.2
		M06-2X	18.1	52.3
		M06-2X	17.7	51.6
TS2	Gas	B3LYP	253.8	274.2
		B3LYP-D3(BJ)	239.2	259.3
		M06-2X	219.8	238.2
	1-Bromopropane	MP2	208.7	231.5
		M06-2X	212.6	231.7
		M06-2X	210.8	230.4
IC2	Gas	B3LYP	47.2	62.3
		B3LYP-D3(BJ)	31.2	50.3
		M06-2X	34.4	51.5
	1-Bromopropane	MP2	18.3	37.4
		M06-2X	29.4	54.3
		M06-2X	28.5	55.5
TS3	Gas	B3LYP	133.4	148.2
		B3LYP-D3(BJ)	111.2	127.9
		M06-2X	101.1	136.4
	1-Bromopropane	MP2	87.4	120.8
		M06-2X	99.3	140.0
		M06-2X	98.9	141.2
PC	Gas	B3LYP	4.8	31.8
		B3LYP-D3(BJ)	19.9	16.4
		M06-2X	-2.6	1.6
	1-Bromopropane	MP2	-38.3	-11.8
		M06-2X	-22.0	6.7
		M06-2X	-21.7	8.6

^a Single-point calculations were performed with 6-311+G(2df,2p) on geometries obtained using gas-phase B3LYP/6-31G(d) optimizations.

Table B–8. Lesion site Watson-Crick hydrogen bond occupancies^a, distances (in Å), and angles (in °) for the natural DNA, the dA-Ap AAG duplex, and the dG-Ap AAG duplex.

Residue Interaction	Atoms Interacting	Occupancy (%)			Distance (Ang.)			Angle (deg.)		
		Natural	dA-Ap	dG-Ap	Natural	dA-Ap	dG-Ap	Natural	dA-Ap	dG-Ap
G6–C25	N2—H...O2	99.8	65.6	92.7	2.9	2.9	2.9	162.5	159.4	161.4
	N3—H...N1	100.0	63.5	87.2	3.0	3.0	2.9	165.4	161.2	163.6
	O6—H...N4	99.0	53.1	79.7	2.9	2.9	3.0	163.3	161.2	161.6
A7–T24	N1...H—N3	99.7	0.0	74.5	3.0	–	2.9	164.3	–	159.6
	N6—H...O4	96.2	0.0	68.6	3.0	–	3.0	163.7	–	159.7
A8–T23	N1...H—N3	99.8	0.0	0.0	2.9	–	–	164.3	–	–
	N6—H...O4	94.8	0.0	0.0	3.0	–	–	163.7	–	–
G9–C22	N2—H...O2	99.7	97.3	0.0	2.9	2.9	–	161.4	162.4	–
	N3—H...N1	99.9	96.0	0.0	2.9	3.0	–	165.0	164.2	–
	O6—H...N4	97.7	94.1	0.0	3.0	2.9	–	162.1	161.4	–
A10–T21	N1...H—N3	99.8	99.6	82.5	2.9	3.0	2.9	164.2	163.8	162.4
	N6—H...O4	93.6	95.8	77.3	3.0	3.0	3.0	161.6	163.1	159.4

^a Hydrogen-bond occupancies were determined using a cut off of 3.4 Å and 120°

Table B–9. Lesion site Watson-Crick hydrogen bond occupancies^a, distances (in Å), and angles (in °) for natural DNA and the dA-Ap AAG duplex.

Residue Interaction	Atoms Interacting	Occupancy (%)		Distance (Ang.)		Angle (deg.)	
		Natural	dA-Ap	Natural	dA-Ap	Natural	dA-Ap
G6–C25	N2—H...O2	99.8	47.1	2.9	3.0	162.5	159.2
	N3—H...N1	99.8	48.0	3.0	3.0	165.4	156.1
	O6—H...N4	98.7	33.7	2.9	2.9	163.3	160.0
A7–T24	N1...H—N3	99.7	0.0	3.0	–	164.3	–
	N6—H...O4	97.4	0.0	3.0	–	163.5	–
A8–T23	N1...H—N3	99.7	0.0	3.0	–	164.8	–
	N6—H...O4	97.4	0.0	3.0	–	163.1	–
G9–C22	N2—H...O2	99.8	64.7	2.9	3.0	162.4	162.5
	N3—H...N1	99.8	73.1	3.0	2.9	163.8	159.6
	O6—H...N4	97.7	60.9	3.0	2.9	162.9	161.3
A10–T21	N1...H—N3	99.8	92.4	2.9	2.9	164.4	164.0
	N6—H...O4	93.8	86.5	3.0	3.0	162.3	162.7

^a Hydrogen-bond occupancies were determined using a cut off of 3.4 Å and 120°

Table B-11. Atom types and partial charges for dA-Ap.

Fragment	ATOM Name	Atom Type	Partial Charge
1	O5'	OS	-0.4922
	P	P	1.2135
	OP1	O2	-0.7914
	OP2	O2	-0.7914
	O3'	OS	-0.5398
	C5	CB	-0.0075
	C4	CB	0.6306
	C6	CA	0.2913
	N9	N*	-0.1513
	N3	NC	-0.8259
	N7	NB	-0.6362
	N6	N2	-0.0415
	H61	H	0.2413
	N1	NC	-0.6774
	C1'	CT	0.095
	H1'	H2	0.1671
	C8	CK	0.2213
	H8	H5	0.174
	C2	CQ	0.6339
	H2	H5	0.0279
	C2'	CT	-0.0751
	H2'	HC	0.0551
	H2''	HC	0.0551
	O4'	OS	-0.3973
C3'	CT	0.1029	
H3'	H1	0.0952	
C4'	CT	0.1877	
H4'	H1	0.1081	
C5'	CT	-0.0378	
H5'	H1	0.0824	
H5''	H1	0.0824	
2	O5'	OS	-0.4922
	P	P	1.2135
	OP1	O2	-0.7914
	OP2	O2	-0.7914
	O3'	OS	-0.5398
	C5'	CT	0.0049
	H5'	H1	0.0783
	H5''	H1	0.0783
	C4'	CT	0.0571
	H4'	H1	0.0918
	O4'	OS	-0.2359
	C1'	CT	-0.2359
	H1'	H2	0.2144
	C3'	CT	0.203
	H3'	H1	0.0558
	C2'	CT	0.0569
	H2'	HC	0.0163
H2''	HC	0.0163	

Table B-12. Atom types and partial charges for dG-Ap.

Fragment	ATOM Name	Atom Type	Partial Charge
1	O5'	OS	-0.5097
	P	P	1.2158
	OP1	O2	-0.7907
	OP2	O2	-0.7907
	O3'	OS	-0.5616
	C5'	CT	0.0097
	H5'	H1	0.0675
	H5''	H1	0.0675
	C4'	CT	0.1689
	H4'	H1	0.0975
	O4'	OS	-0.3957
	C1'	CT	0.155
	H1'	H2	0.1221
	N9	N*	-0.029
	C8	CK	0.134
	H8	H5	0.1785
	N7	NB	-0.5709
	C5	CB	0.1271
	C6	C	0.5787
	O6	O	-0.5621
	N1	NA	-0.5058
	H1	H	0.3069
	C2	CA	0.4844
	N2	N2	-0.2813
	H2	H	0.3353
	N3	NC	-0.4918
	C4	CB	0.1884
C3'	CT	0.1896	
H3'	H1	0.0696	
C2'	CT	-0.1182	
H2'	HC	0.0555	
H2''	HC	0.0555	
2	O5'	OS	-0.5097
	P	P	1.2158
	OP1	O2	-0.7907
	OP2	O2	-0.7907
	O3'	OS	-0.5616
	C5'	CT	0.0485
	H5'	H1	0.053
	H5''	H1	0.053
	C4'	CT	0.193
	H4'	H1	0.0761
	O4'	OS	-0.2965
	C1'	CT	-0.3467
	H1'	H2	0.2675
	C3'	CT	0.2783
	H3'	H1	0.0026
	C2'	CT	0.0011
	H2'	HC	0.0535
H2''	HC	0.0535	

References

1. Blanchet, C.; Pasi, M.; Zakrzewska, K.; Lavery, R., CURVES+ Web Server For Analyzing and Visualizing the Helical, Backbone and Groove Parameters of Nucleic Acid Structures. *Nucleic Acids Res.* **2011**, *39* (suppl_2), W68-W73.