Title
Structural and Thermodynamic Features of Polycyclic Aromatic Hydrocarbon - DNA Adducts in the NRAS(Q61) Sequence Context

Author
Urwin, Derek

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Structural and Thermodynamic Features of Polycyclic Aromatic Hydrocarbon - DNA Adducts in the NRAS(Q61) Sequence Context

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Chemistry

by

Derek J. Urwin

2022

ABSTRACT OF THE DISSERTATION

Structural and Thermodynamic Features of
Polycyclic Aromatic Hydrocarbon - DNA Adducts in the NRAS(Q61) Sequence Context

by

Derek J. Urwin
Doctor of Philosophy in Chemistry
University of California, Los Angeles, 2022
Professor Anastassia N. Alexandrova, Co-Chair
Professor William M. Gelbart, Co-Chair

The relative genotoxicity of different polycyclic aromatic hydrocarbons (PAHs) is widely thought to be a function of the structural and thermodynamic features of their corresponding PAH-DNA adducts. As a result, accurate parameters for molecular mechanics force fields are crucial to the study of such systems via molecular dynamics (MD). While transferability of parameters among structurally similar molecular systems is frequently a goal when parameterizing novel residues for the CHARMM force field, we will show that planar bay region and non-planar fjord region PAH-DNA adduct systems require distinct dihedral terms to accurately model the torsional potential energy surface of the adduct covalent bond that links a PAH-diol-epoxide and adenine, despite identical atomic connectivity. We then examine the use of the Truncated Singular Value Decomposition and Tikhonov Regularization in standard form to address ill-posed least squares problems $\mathbf{Ax} = \mathbf{b}$ that frequently arise in molecular mechanics force field parameter optimization. Utilizing the Discrete Picard Condition and/or a well-defined gap in the singular value spectrum when $\mathbf{A}$ has a well-determined numerical rank, we are able to systematically determine truncation and in

turn regularization parameters that are correspondingly used to produce truncated and regularized solutions to the ill-posed least squares problem at hand. These solutions in turn result in optimized force field dihedral terms that accurately parameterize the torsional energy landscape. As the solutions produced by this approach are unique, it has the advantage of avoiding the multiple iterations and guess and check work often required to optimize molecular mechanics force field parameters. With optimized parameters for bay and fjord region PAH-DNA adduct systems developed, we conduct alchemical free energy perturbation calculations over closed thermodynamic cycles in order to gauge the relative genotoxicities of several IARC Group 2A/B and 3 PAHs in the NRAS(Q61) DNA sequence context. These calculations reveal that the fjord region PAHs examined in this work as well as other IARC Group 2A/B and 3 PAHs exhibit greater relative binding affinity as compared to the IARC Group 1 known human carcinogen B[a]P. These PAHs are also less likely to form productive PAH-DNA protein binding complexes required in the recognition step of global genomic - nucleotide excision repair, indicating that they are more likely to persist and induce mutations in subsequent DNA replication cycles. Further examination reveals that the intercalated conformation and structural differences among PAH-DNA adducts have an impact on stabilizing van der Waals interactions and hydrogen bonding between nucleobase pairs in NRAS(Q61) that are generally associated with trends in relative binding free energies.

The dissertation of Derek J. Urwin is approved.

Jose Alfonso Rodriguez

Christopher R. Anderson

William M. Gelbart, Committee Co-Chair

Anastassia N. Alexandrova, Committee Co-Chair

University of California, Los Angeles

2022

To . . .

my wife Erin who has been my foundation since the day I met her,

my children Aidan and Everett who remind me how much I love being a father everyday,

my parents Ross and Kimy who always supported and encouraged my studies,

my late brother Isaac, for whom I did this.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor Professor Anastassia Alexandrova for giving a very tired looking thirty-something year old fire fighter the opportunity to study with the finest students and professors the academic world has to offer. Your encouragement has meant the world to me. Thank you to my committee members Professors Chris Anderson, William Gelbart, and Jose Rodriguez for your kindness and guidance that have taught me countless scientific and life lessons that I will put to work for all of my remaining years. Thank you to my brother and sister LA County and Miami-Dade fire fighters - especially at 1s, 3s, and 170s - for always having my back and carrying my family through the tough times of my brother's illness and passing, I wouldn't have made it without you.

Chapter 3 of this thesis is adapted with permission from the publication "Regularization of least squares problems in CHARMM parameter optimization by truncated singular value decompositions." Urwin, D. J., and Alexandrova, A. N. The Journal of Chemical Physics (2021) 154(18), 184101, doi.org/10.1063/5.0045982

Figure 4.11 was included with permission from the publication "3dna: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures," X.J. Lu and W. K. Olson, Nature protocols (2008), vol. 3, no. 7, pp. 1213–1227, doi.org/10.1093/nar/gkg680

Chapters 4 and 5 of this thesis are adapted with permission from a manuscript in preparation "Free energies of binding and formation of the productive complex of PAH-DNA adducts in the NRAS(Q61) sequence context"

VITA

2003	B.S. Applied Mathematics, UCLA
2007-2009	Fire Fighter, Miami-Dade Fire Rescue
2010-Present	Fire Fighter & Engineer, Los Angeles County Fire Department
2017	M.S. Chemistry, UCLA
2015–2021	Teaching Assistant, UCLA Department of Chemistry

PUBLICATIONS

Regularization of least squares problems in CHARMM parameter optimization by truncated singular value decompositions. Urwin, D. J., and Alexandrova, A. N. The Journal of Chemical Physics (2021) 154(18), 184101, doi.org/10.1063/5.0045982.

CHAPTER 1

Introduction

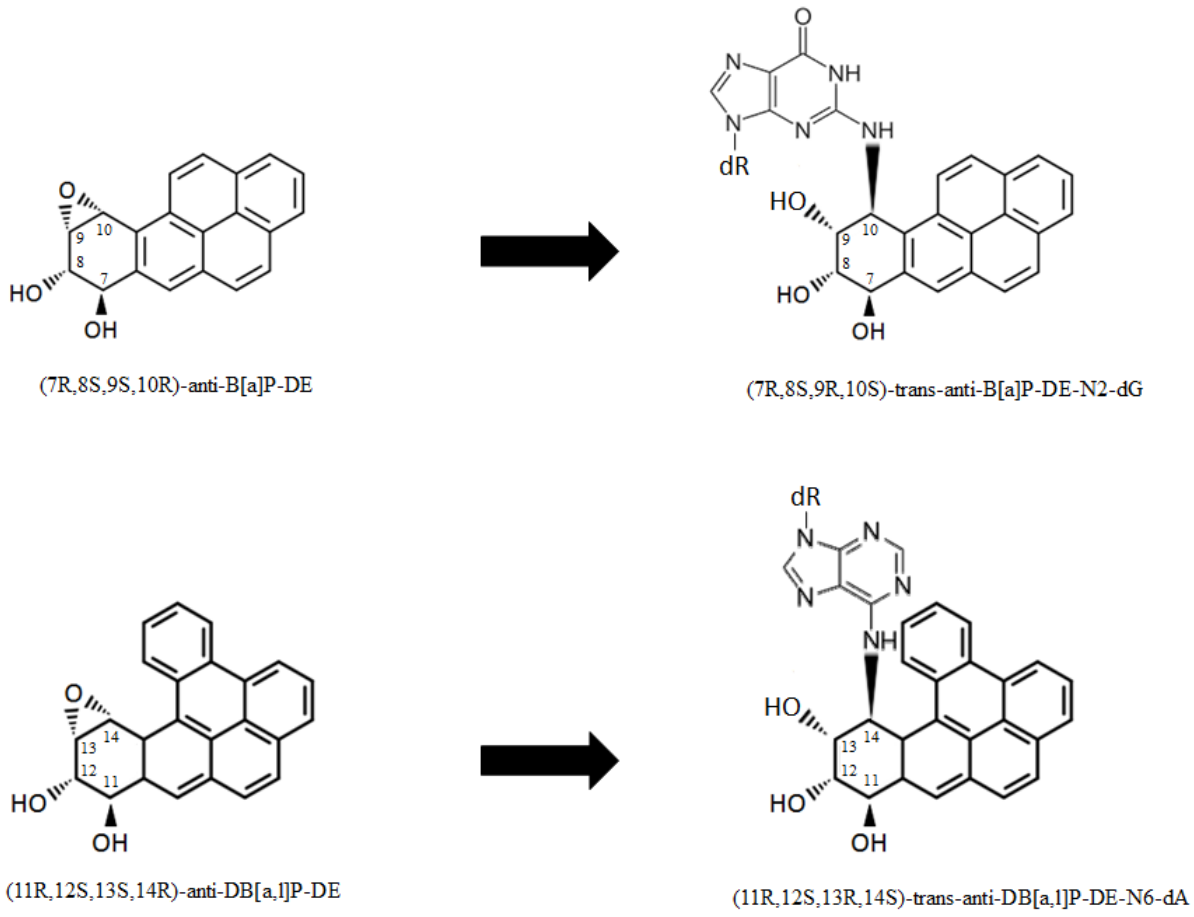
Polycyclic aromatic hydrocarbons (PAHs) are a large class of compounds produced by processes that involve the incomplete combustion of organic substances, many of which are classified as known, probable, or possible human carcinogens (Group 1, 2A, or 2B) by the International Agency for Research on Cancer (IARC).[2] Human exposure to PAHs via inhalation, dermal absorption, and ingestion is pervasive[3–5] as these compounds are produced by ubiquitous processes that range from grilling food to automotive exhaust to catastrophic wildfires that burn in the Western United States and around the world. Environmental and occupational exposures to PAHs is known to be associated with an elevated incidence of cancer in affected populations and occupational exposures in certain fields of work such as fire fighting can be extreme.[6–13] Many PAHs are genotoxic, mutagenic, and ultimately carcinogenic owing to the formation of covalent PAH-DNA adducts at mutational hotspots in the genome.[12–18] Because the process of carcinogenesis may be initiated by the clonal expansion of a single cell through the heritable abrogation of cellular processes that regulate cell division,[13] elucidating the molecular mechanisms at the root of genotoxicity in this large class of compounds is essential to developing carcinogenic risk factor assessments for the occupational and public health communities.

In human cells, bay region PAHs such as benzo[a]pyrene (B[a]P) and fjord region PAHs such as dibenzo[a,l]pyrene (DB[a,l]P) are enantioselectively and diastereoselectively metabolized to PAH-diol-epoxides (PAH-DEs) by cytochrome P450s (CYPs) and epoxide-hydrolases (EHs) that operate in the lipid bi-layer of the endoplasmic reticulum.[13] These enzymes are heavily expressed in the liver as well as most extra-hepatic tissues including the lungs, with the metabolic process

yielding the major products: (7R,8S,9S,10R)-B[a]P-DE and (11R,12S,13S,14R)-DB[a,l]P-DE, respectively (Figure 1.1 - left). Stereoisomers of these compounds are only produced in small quantities.[13,17] Covalent PAH-DNA adducts are preferentially formed via trans opening of the epoxide ring, with the bay region (7R,8S,9S,10R)-B[a]P-DE preferentially binding with the exocyclic amino group of guanine to form (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N2-dG adducts and the fjord region (11R,12S,13S,14R)-DB[a,l]P-DE preferentially binding with the exocyclic amino group of adenine to form (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA adducts (Figure 1.1 - right). Adducts formed by cis opening of the epoxide ring are only formed in small quantities.[13,17] Note also that many planar PAHs with a large ratio of surface area to depth are activating ligands of the cytosolic aryl-hydrocarbon receptor (AhR), which is a transcription factor that mediates expression of cytochrome P450, thus increasing expression of the enzyme primarily responsible for metabolism of PAHs into genotoxic PAH-DEs.[12,13,17]

When PAH-DNA adducts evade genomic repair mechanisms, they are likely to induce nucleotide misincorporation in the complementary DNA strand during replication, thus leading to mutations.[12,13,17,19,20] Studies both in-vitro and in-vivo have shown that bay region PAH-DEs that preferentially form covalent PAH-DNA adducts with guanine tend to induce dG$\rightarrow$dT transversions while fjord region PAH-DEs that preferentially form covalent PAH-DNA adducts with adenine tend to induce dA$\rightarrow$dT transversions.[13,21–26] Since DNA lesions in the template strand of DNA are efficiently repaired by transcription coupled nucleotide excision repair (TC-NER), PAH-DNA adducts in the non-template (coding) strand of DNA, are of primary interest when it comes to studying PAH induced mutagenesis.[12,13] PAH-DNA adducts in the non-template strand are primarily repaired by global genomic nucleotide excision repair (GG-NER), initiation of which is dependent upon the GG-NER recognition step which is characterized by the protein XPC-RAD23B forming a productive complex with a PAH-DNA adduct system and subsequently recruiting repair factors that complete the dual excision GG-NER repair process.[12,13,15,16,27,28]

While humans are known to be exposed to both B[a]P and DB[a,l]P from products of incomplete combustion such as cigarette smoke, it has been noted that DB[a,l]P is the most tumorigenic



(1.1) Top: Bay region (7R,8S,9S,10R)-B[a]P-DE and resulting (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N2-dG adduct
Bottom: Fjord region (11R,12S,13S,14R)-DB[a,l]P-DE and resulting (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA adduct

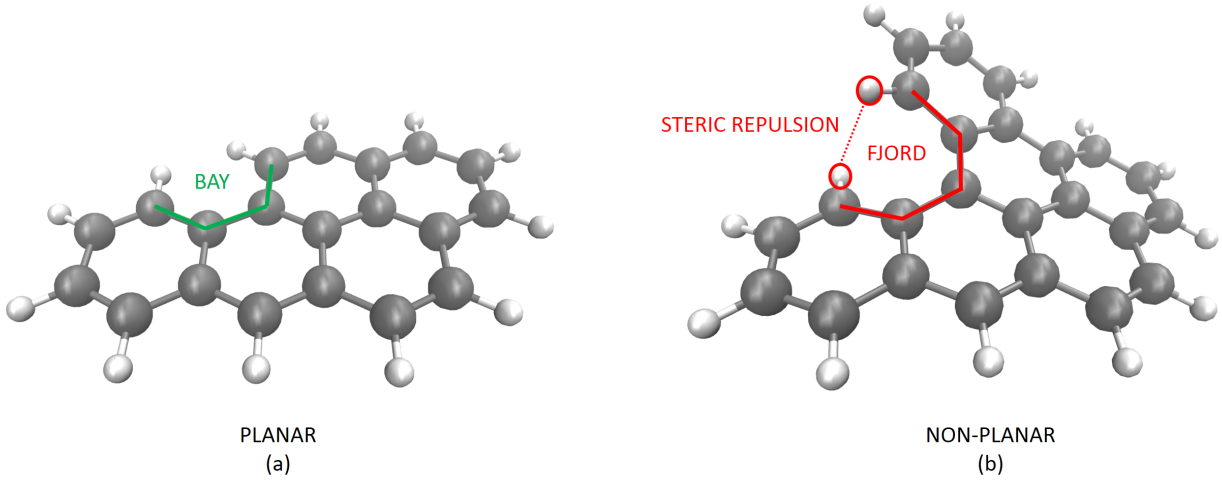
compound known to date, with tumorigenicity estimated to be approximately 100-fold that of B[a]P.[12,13] Despite this, B[a]P is classified as an IARC Group 1 known human carcinogen while DB[a,l]P is classified as a Group 2A probable human carcinogen.[2] Furthermore, B[a]P appears on the list of 16 EPA priority PAHs utilized for environmental risk factor assessments while DB[a,l]P and other PAHs known to be more genotoxic and mutagenic than B[a]P do not.[2,12–16,29,30] As a result, risk factor assessments that do not include the most genotoxic and mutagenic PAHs paint an incomplete picture of the severity of PAH exposures. Because PAHs are a massive class of compounds, and those with four to seven rings have been shown to be the most genotoxic,[13,17] examining the genotoxicities of PAHs other than B[a]P is crucial to developing a more thorough

understanding of the severity of toxic exposures associated with products of incomplete combustion and the downstream biological impact.

The process of examining the genotoxic and carcinogenic potential of a particular compound either in-vitro or in-vivo is both slow and expensive, with results developed on the order of years.[31] For example, mutagenesis assays utilizing transgenic rodents are widely accepted as an effective approach that includes the biology of an entire organism, but requires extensive infrastructure over several generations of animals to complete.[31] As a result, this field of study can benefit from in-silico examination of the relative genotoxicity of a collection of structurally diverse and largely unstudied PAHs. Such in-silico studies can then inform in-vivo and in-vitro research efforts in an approach similar to lead optimization in computational drug design.[32]

The relative genotoxicity of PAH-DNA adducts can be studied and understood as a function of their structural and thermodynamic features.[12,13,17] For example, NMR solution structures and molecular dynamics simulations have shown that in the 5'-d(...CA*C...)-3' 5'-d(...GTG...)-3' sequence context, both the (11S,12R,13S,14R)-trans-anti-DB[a,l]P-DE-N6-dA* and (7S,8R,9S,10R)-trans-anti-B[a]P-DE-N6-dA* adducts assume an intercalated conformation from the major groove without neighboring nucleobase displacement (note these particular adducts result from enantiomers of the major PAH-DEs described above).[12,33–36] However, the flexible and non-planar fjord region DB[a,l]P-DNA adduct system results in thermodynamic stabilization of the DNA duplex characterized by an 11°C increase of the DNA duplex melting point while the rigid and planar bay region B[a]P-DNA adduct system results in destabilization of the DNA duplex characterized by a 13°C decrease of the DNA duplex melting point. Correspondingly, the (11S,12R,13S,14R)-trans-anti-DB[a,l]P-DE-N6-dA* system is known to be almost totally resistant to GG-NER while the (7S,8R,9S,10R)-trans-anti-B[a]P-DE-N6-dA* system is characterized by a mild GG-NER response.[12] The non-planar structure of fjord region PAHs such as DB[a,l]P results from the steric repulsion between hydrogens on opposite ends of the fjord region (Figure 1.2), allowing fjord PAHs to intercalate in energetically favorable conformations that minimize distortions of the DNA duplex and result in stabilizing van der Waals interactions between the PAH and neighboring nu-

cleobases (i.e. enhanced π-stacking).[12,13,16,17,37]



(1.2) (a) Planar bay region benzo[a]pyrene (B[a]P), (b) fjord region dibenzo[a,l]pyrene (DB[a,l]P) where the non-planar structure is caused by steric repulsion between hydrogens on opposite ends of the fjord region

An effective way to study the structural and thermodynamic properties of large numbers of PAH-DNA adduct systems, to then characterize the relative genotoxicity of different PAHs, is via molecular dynamics (MD) simulations. However, PAH-nucleotide systems of interest are not standard residues in commonly utilized molecular mechanics (MM) force fields such as AMBER and CHARMM. Thus before proceeding with MD simulations of entire PAH-DNA adduct systems, it is necessary to develop custom residues and accurate force field parameters to effectively study such systems. Depending on the MM force field utilized, novel residues are often parameterized via tools such as the Generalized AMBER Force Field[38] and ANTECHAMBER,[39] the CHARMM General Force Field (CGenFF) and Paramchem.com,[40–42] and SwissParam.[43] CGenFF for example assigns CHARMM atom types, partial charges, and bonded parameters (bond, angle, and dihedral) by analogy from previously parameterized residues in CGenFF along with penalty scores that indicate the validity of partial charge and parameter assignments. Penalty scores of 50 or greater, require full parameter development and optimization from first principles. In these cases,

the VMD-Force Field Tool Kit (VMD-ffTK)[44] and most recently FFParam[45] are extremely useful packages that can be used to further optimize parameters for a novel residue.

While these are robust tools that work well for most of the force field parameters needed to model PAH-DNA adduct systems in the CHARMM force field, we will begin by showing that bay and fjord region PAH-DNA adduct systems require distinct dihedral terms to accurately model the torsional potential energy landscapes of the PAH-DNA adduct covalent bond between, despite identical atomic connectivity. This is as opposed to the standard approach of optimizing one set of dihedral parameters to be used interchangeably among systems with identical atomic connectivity that might differ structurally. We will then examine the use of the Truncated Singular Value Decomposition and Tikhonov Regularization in standard form to address ill-posed least squares problems $\mathbf{Ax} = \mathbf{b}$ that frequently arise in molecular mechanics force field parameter optimization. Utilizing the Discrete Picard Condition and/or a well-defined gap in the singular value spectrum when $\mathbf{A}$ has a well-determined numerical rank, we will show that truncation and in turn regularization parameters can be determined systematically to produce solutions to the ill-posed least squares problem that are largely insensitive to perturbations of the parameterization target data. These solutions in turn result in optimized force field dihedral terms that accurately parameterize the torsional energy landscape. As the solutions produced by this approach are unique, it has the advantage of avoiding the multiple iterations and guess and check work often required to optimize molecular mechanics force field parameters utilizing standard approaches.

This work was largely focused on developing the parameterization tools described above, which are needed to accurately model PAH-DNA adduct systems in MD simulations. Adenine model systems were utilized in this development process and because highly genotoxic PAHs such as DB[a,l]P tend to form covalent PAH-DNA adducts with adenine, examination of relative PAH genotoxicities will focus on adenine PAH-DNA adducts while guanine PAH-DNA adducts will be the subject of future work. In particular, fjord region PAH-DEs such as DB[a,l]P-DE have a propensity to form PAH-DNA adducts at mutational hotspots in the human genome such as the central adenine of codon 61 in the NRAS proto-oncogene [henceforth NRAS(Q61)]

that subsequently induce dA→dT transversions.[12,13,17,20] NRAS mutations are found in 27.7% of human melanomas and 88.1% of these mutations are found in NRAS(Q61), which normally codes for glutamine with nucleotide sequence CAA.[46] Single nucleotide polymorphisms (SNPs) of NRAS(Q61) abrogate the catalytic activity of the NRAS enzyme, locking it in an active GTP-bound conformation.[18] In particular, NRAS(Q61L) variants that code for leucine rather than glutamine exhibit elevated mitogen-activated protein kinase (MAPK) signaling, resulting in over-representation of the protein kinase CK2α which is associated with cellular proliferation in primary human melanocytes.[18] Studies examining the efficiency of global genomic nucleotide excision repair (GG-NER) of PAH-DNA adducts at the central adenine of NRAS(Q61) (henceforth dA*) in human HeLA cell extracts found that fjord region (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA* adducts were almost entirely repair resistant while stereochemically analogous bay region (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA* adducts were repaired with high efficiency.[13,17] As a result, adducts such as the (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA* adduct in NRAS(Q61) that evade GG-NER and induce dA*→dT transversions that correspond to a CAA→CTA SNP that codes for leucine are a possible source of NRAS(Q61L) mutations in human cancers.

The genotoxicity of a given PAH in a given sequence context is dependent upon the likelihood of the PAH-DE forming a PAH-DNA adduct and the likelihood of that PAH-DNA adduct evading genomic repair mechanisms such as GG-NER. The contrast in GG-NER repair efficiency observed between the (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA* and (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA* adducts described above is known to be associated with the distinct structural and thermodynamic features of each system.[12,13,17] In order to gauge the relative genotoxicity at dA* in the NRAS(Q61) sequence context of IARC Group 2A, 2B, and 3 PAHs as compared to B[a]P, we will first quantify the relative differences in free energy of binding for these PAH-DNA adduct systems. This will allow us to identify those PAH-DEs that are most likely to form covalent PAH-DNA adducts. We will then quantify the relative differences in free energy of formation of the corresponding productive RAD4-RAD23 : PAH-DNA adduct binding complex (henceforth

productive complex) where RAD4-RAD23 is the yeast ortholog of human XPC-RAD23.[15,16,27,28] This will allow us to identify those PAH-DNA adducts that are less likely to be repaired by the GG-NER machinery. Those PAHs that are the most likely to form PAH-DNA adducts and the least likely to form the productive complex are likely to be the most genotoxic in the NRAS(Q61) sequence context and are the most likely to persist and induce mutations in subsequent DNA replication cycles. We will then examine the associated structural features of these PAH-DNA adduct systems.

CHAPTER 2

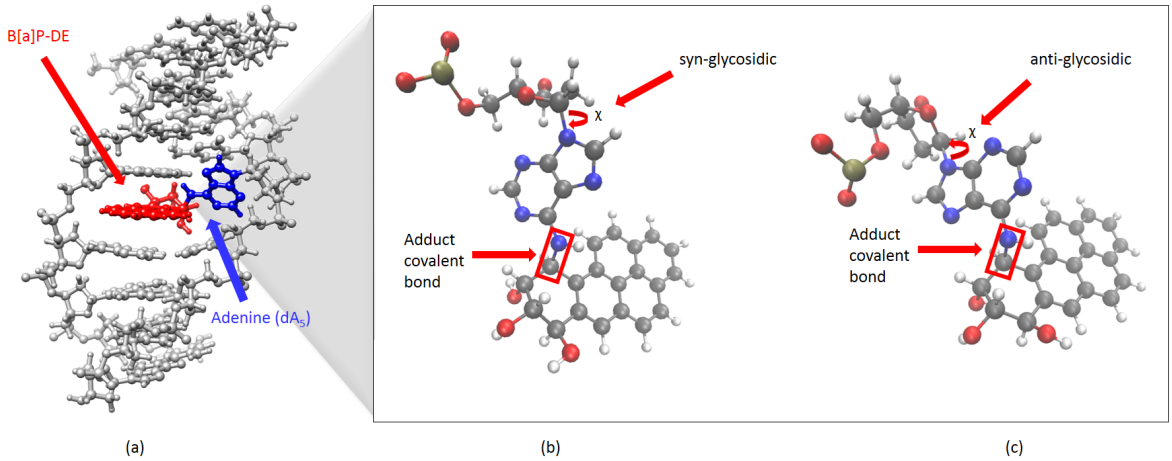
Parameterization of Bay and Fjord Region PAH-DNA Adducts

2.1 Methods

2.1.1 Quantum Mechanical Scans of the Adduct Covalent Bond's Torsional Potential Energy Surface

When parameterizing novel residues for the CHARMM force field, it is standard practice to identify portions of the molecule for which there are existing CHARMM residues and parameters, and then parameterize the sections of the novel residue that link the existing residues.[40] Because they are typically optimized last, the dihedral parameters that characterize freely rotating covalent bonds that link ring systems, such as a PAH-DNA adduct covalent bond, are essentially a correction to non-bonded interactions.[47] As a result, such dihedral parameters are known to perform poorly when used transferably among different CHARMM residues and they are typically optimized by fitting to QM target data derived from model systems.[40,47,48] In the interest of using the most computationally tractable model system, it is natural to first examine a model system consisting of the simplest PAH, naphthalene, in hopes that it will produce a working set of parameters that can be used in other PAH-DNA adduct systems that have identical atomic connectivity in the adduct covalent bond. However, in the work to follow, it will become evident that naphthalene does not suffice as a model system for parameterizing the adduct covalent bond in structurally different bay and fjord region PAH-DNA adduct systems, and each will require distinct dihedral terms to accurately parameterize the adduct covalent bond.

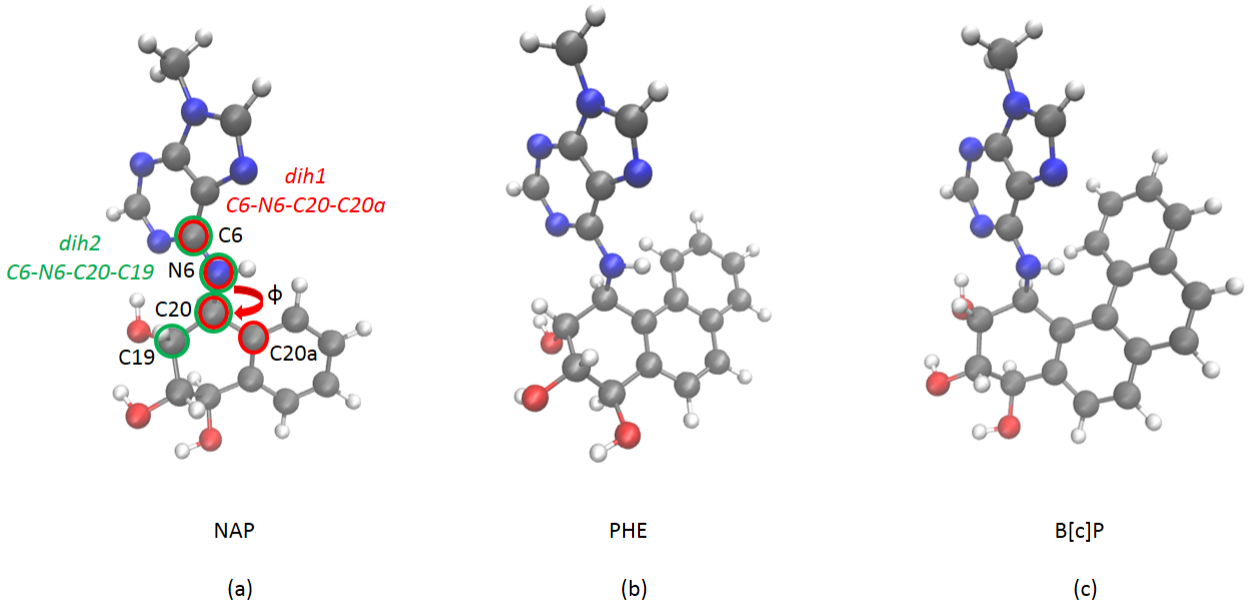
In order to examine how differing structural features impact the adduct covalent bond's tor-



(2.1) (a) (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA$^*_5$ adduct in the 5'-d(GGTCA$^*_5$CGAG)-3' 5'-d(CTCGGGACC)-3' DNA duplex, intercalated without neighboring nucleobase displacement (PDB: 1DXA[1]), (b) (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA* adduct in syn-glycosidic base-sugar conformation, (c) (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA* adduct in anti-glycosidic base-sugar conformation

sional potential energy surface (PES), we developed three model systems based on the NMR solution structure of a B[a]P-DE-N6-dA adduct system formed by trans opening of the epoxide ring in (7R,8S,9S,10R)-B[a]P-DE and binding with the N6 nitrogen of dA$^*_5$'s exocyclic amino group in the 5'-d(GGTCA$^*_5$CGAG)-3' 5'-d(CTCGGGACC)-3' DNA duplex (PDB: 1DXA,[1] 1BPS[49]). The resulting (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA$^*_5$ adduct assumes an intercalated configuration without neighboring nucleobase displacement (Figure 2.1a) with adenine in either syn(major) or anti(minor) conformations of the dA$^*_5$ base-sugar glycosidic torsion angle χ (Figure 2.1b and 2.1c).

The three model systems are composed of 9-methyl-adenine (dA with the N9 nitrogen capped by a methyl group) and a stereochemically analogous covalent PAH adduct [where PAH=naphthalene (NAP), phenanthrene (PHE), or benzo[c]phenanthrene (B[c]P)] replacing the (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA* adduct described above (Figure 2.2a, 2.2b, and 2.2c). This results in (-R,-S,-R,-S)-trans-anti-PAH-DE-N6-dA* model systems that are representative of typical PAH



(2.2) Syn-glycosidic model systems: (a) NAP, dihedral parameter dih_1: C6-N6-C20-C20a circled in red, dihedral parameter dih_2: C6-N6-C20-C19 circled in green, (b) PHE, (c) B[c]P

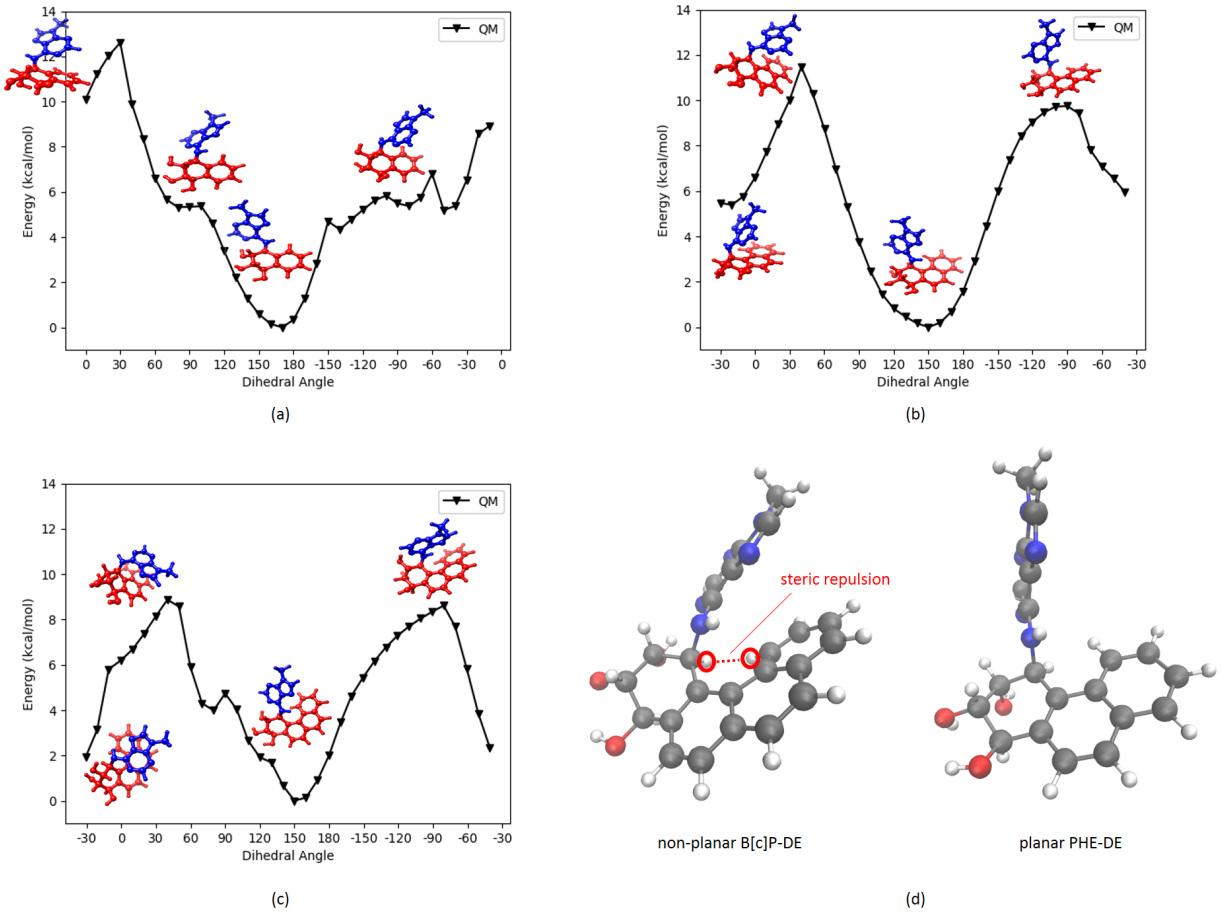
structures of interest such as bay (PHE) and fjord (B[c]P) region PAHs.

The adduct covalent bond's dihedral angle ϕ was defined about the atoms C6-N6-C20-C20a (Figure 2.2a) where C6 and N6 are standard CHARMM atom names from the ADE residue in the CHARMM-Nucleic Acid (NA) force field[50,51] while C20 and C20a are atoms in the PAH portion of the model systems. Note that carbons do not follow standard IUPAC PAH numbering in our model systems to avoid atom names overlapping with those in the CHARMM-NA force field. In the major syn-glycosidic conformation, relaxed QM PES scans of the dihedral angle ϕ were performed in 10° increments for $\phi \in$(-180°,180°] for each of the three model systems (Figure 2.3). Calculations were performed utilizing the Gaussian 16[52] software package at the MP2/6-31G(d) level of theory, which is known to produce accurate torsional potential energy surfaces for the purpose of parameterizing dihedrals in oligonucleotides.[40] Note that these QM PES scans were conducted in-vacuo as the non-base displaced intercalating configuration of the (7R,8S,9R,10S)-trans-

anti-B[a]P-DE-N6-dA$_5^*$ adduct upon which the model systems are based results in the PAH and modified dA$_5^*$ nucleobase residing within the hydrophobic core of the DNA double-helix. Several PAH-DNA adduct systems of toxicological interest assume such configurations.[12,13,16,17,27,37,53] The adduct covalent bond in the NMR solution structure from which the model systems are derived has a dihedral angle of $\phi = 141°$ and the absolute minimum of the PES for each model system occurs at a similar value. Hence, all three plots in Figure 2.3 are centered at their respective absolute minima as this region of the PES will be of primary interest for MM force field parameterization.

The NAP model system, resulted in a PES with an absolute minimum at $\phi = 170°$ and an absolute maximum of approximately 12.6 kcal/mol at $\phi = 30°$. Local minima with values less than 6.0 kcal/mol occur across relatively flat regions of the PES at $\phi = 80°, -140°, -80°$, and $-50°$ (Figure 2.3a). The bay region PHE model system's PES has its absolute minimum $\phi = 150°$ and climbs smoothly to an absolute maximum of approximately 11.5 kcal/mol at $\phi = 40°$ and a local maximum of approximately 9.7 kcal/mol at $\phi = -90°$. A local minimum of approximately 5.4 kcal/mol occurs at $\phi = -20°$ between the absolute and local maxima (Figure 2.3b). The fjord region B[c]P model system's PES is distinguished from the NAP and PHE model systems by maxima that do not exceed 8.9 kcal/mol. The absolute minimum is located at $\phi = 150°$ and the PES climbs smoothly to a local maximum at $\phi = -80°$ while there is a local minimum at $\phi = 80°$ between the absolute minimum and the absolute maximum at $\phi = 40°$ (Figure 2.3c). Natural Bond Orbital[54] analysis of optimized geometries from these QM PES scans revealed that the absolute maxima for each model system at or near $\phi = 40°$ correspond to inversion of the dA* N6 nitrogen in a manner analogous to the two state ammonia system. The model systems were then evaluated for multi-configurational character by conducting single point Complete Active Space (CASSCF) calculations on samples of the MP2 optimized geometries. These showed that MP2/6-31G(d) was a satisfactory level of theory for treating these systems.

These differing potential energy surfaces arise because of the varied aromatic ring structures among the NAP, PHE and B[c]P model systems. In particular, the non-planar fjord region B[c]P model system assumes lower energy conformations as compared to the rigid planar NAP and PHE



(2.3) QM PES scans of the adduct covalent bond dihedral angle ϕ at MP2/6-31G(d) level of theory for syn-glycosidic model systems: (a) NAP, (b) PHE, (c) B[c]P, (d) steric repulsion between hydrogens on opposite ends of the non-planar fjord region B[c]P model system facilitate lower energy conformations than the planar bay region PHE model system.

model systems (Figure 2.3d). As mentioned above, dihedral parameters are typically the last to be optimized in a novel residue and are essentially a correction to non-bonded interactions in the CHARMM force field.[47] As a result, the MM dihedral parameters associated with the adduct covalent bond in a given system will largely characterize and fine tune the corresponding torsional potential energy surface. This in turn influences sampling during MD simulations, with lower energy regions of the torsional potential energy surface being more accessible. At first glance, it may seem that dihedral parameters associated with the adduct covalent bond would be of little importance in a much larger system composed of a DNA duplex. However, in MM force field parameterization, the dihedral parameters associated with freely rotating covalent bonds that link ring systems are crucial to accurate conformational sampling.[47] Furthermore, because PAH genotoxicity is widely thought to be a function of PAH-DNA adduct conformation,[12,13,16,17,27,37,53] effective/accurate sampling in MD simulations will require an accurate torsional potential energy landscape for the freely rotating adduct covalent bond. In turn, this will require custom dihedral terms for the adduct covalent bond in each of the three model systems. We will show this by applying standard approaches to parameterization of novel residues for the CHARMM force field and upon examination of the resulting MM potential energy surfaces, demonstrate that further parameter optimization is warranted.

2.1.2 Molecular Mechanical Parameterization of Model Systems Using Standard Approaches

Our initial parameterization approach followed the standard practice of identifying and linking existing CHARMM residues. In the case of our model systems, both adenine (dA) and NAP are standard residues in the CHARMM-NA and CGenFF 4.1[40–42] force fields (residues: ADE and NAFT respectively), while PHE and B[c]P are not standard residues in the CHARMM force fields. Atom types and partial charges in the dA portion of the NAP model system were set to match atom types and partial charges from the ADE residue in the CHARMM-NA force field. Atom types and partial charges in the aromatic segment of the NAP portion of the model system were set to match those from the NAFT residue in CGenFF. Those from the aliphatic segment of the NAP portion of

the model system were obtained by utilizing low penalty CGenFF/Paramchem.com (ver 2.2)[40–42] generated atom types and partial charges from a stereochemically analogous naphthalene-triol custom residue and adjusted to match atom types and partial charges from analogous hydroxylated CGenFF residues where appropriate. Atom types and partial charges were assigned similarly in the PHE and B[c]P model systems and additional aromatic rings were assigned atom types and partial charges analogous to those in the aromatic segments of the anthracene (ANTR) CGenFF residue.

The covalent bond between the N6 nitrogen in the dA segment and the C20 carbon in the NAP segment (Figure 2.2a) is formed following the procedure described by MacKerell[55] where by the H62 hydrogen from the CHARMM-NA ADE residue is deleted and its partial charge shifted onto the N6 nitrogen. Similarly, a hydrogen is deleted from the C20 carbon of the naphthalene-triol residue and its partial charge shifted onto C20. Finally a bond between N6 and C20 is added to the corresponding custom residue's topology. An advantage of this standard parameterization approach is a custom residue, the majority of whose atom types, partial charges, and associated bonded parameters come from either the CHARMM-NA force field or previously parameterized CGenFF residues. The only exceptions are bonded parameters involved in the adduct covalent bond or the hydroxyl oxygens.

For these parameters, CGenFF/Paramchem.com (ver 2.2)[40–42] was utilized to obtain a set of atom types, partial charges, and bonded parameters, along with corresponding penalty scores for the entire NAP model system. CGenFF does not assign atom types from the CHARMM-NA force field, hence we mapped CGenFF atom types from the dA portion of the model system to their CHARMM-NA analogs in order to assign bonded parameters for the adduct covalent bond (e.g. CGenFF CG2R64 = CHARMM-NA CN2 and CGenFF NG311 = CHARMM-NA NN1). All other parameter assignments described above were retained. Note that the CGenFF NG311 atom type was evaluated for use in parameterizing the N6 nitrogen in dA's exocyclic amino group in each of the three model systems. However, bonded parameter sets assigned by CGenFF and further optimized in ffTK did not fit QM target data for the adduct covalent bond as well as those

described below utilizing the CHARMM-NA NN1 atom type (and thus retaining the corresponding CHARMM-NA parameters).

In order to validate or further optimize the bonded parameters that link the dA and NAP portions in the model system, we proceeded in accordance with guidance provided in CGenFF where by parameters with penalties from 0-10 were accepted as sound, and those with penalties greater than 10 were accepted after validation, manually adjusted to match other analogous CGenFF parameters, or optimized from first principles using ffTK and QM calculations in Gaussian16.

Following this guidance, all bond and angle parameters for the adduct covalent bond assigned by CGenFF were retained. Note that an attempt was made to further optimize the bond and angle parameters in each of the three model systems utilizing ffTK and corresponding QM Hessian calculations and these parameters were then evaluated for transferability among the three model systems. It was found however that the CGenFF assigned bond and angle parameters, all of which had low penalty scores, ultimately resulted in the best fit to QM target data for the adduct covalent bond. This outcome has the added benefit of maximizing the number of transferable parameters among our model systems.

Dihedral parameters involving $sp2$ and $sp3$ hybridized atom types in or adjacent to the aliphatic ring of the NAP portion were manually assigned dihedral parameters from the analogous segments of cyclohexene based CGenFF residues (e.g. MECH and TMCH). The remaining dihedral parameters assigned by CGenFF with penalty scores less than 10, and those involving hydrogens, were retained after evaluating the analogous CGenFF residues from which they were assigned. This left six dihedral parameters with penalty scores of 26 or greater, each including the N6 nitrogen in the adduct covalent bond and/or a hydroxyl oxygen. All partial charges and bonded parameters other than these six dihedrals are used transferably between the three model systems.

In order to optimize these remaining high penalty dihedral parameters, corresponding QM PES scans in addition to those for the adduct covalent bond described above were conducted in each model system at the MP2/6-31G(d) level of theory in Gaussian16.[52] For each model system, the six high penalty dihedral parameters were then optimized utilizing the standard approach in the

VMD-ffTK[44,48] via multiple rounds of multiple iteration Monte Carlo Simulated Annealing and multiple rounds of Downhill Simplex using CGenFF multiplicities, force constants, and phases as initial input.

In order to maximize the number of transferable dihedral parameters among the model systems, the two ffTK optimized dihedral parameters involving hydroxyl oxygens were evaluated for transferability among the model systems. For example, the two ffTK optimized dihedral parameters involving hydroxyl oxygens from the NAP model system were applied to the PHE and B[c]P model systems and the remaining four dihedral parameters involving the N6 nitrogen in the adduct covalent bond were then re-optimized in ffTK for the PHE and B[c]P model systems. After multiple iterations evaluating transferability, it was found that the two ffTK optimized dihedral parameters involving hydroxyl oxygens from the PHE model system transferred well to the B[c]P model system, ultimately resulting in the best fit to QM target data for the adduct covalent bond's torsional PES in the PHE and B[c]P model systems.

Of note, among the six high penalty dihedral parameters that were further optimized, the two highest penalty dihedral parameters were those that parameterize the freely rotating adduct covalent bond, defined by atoms C6-N6-C20-C20a (dih_1) and C6-N6-C20-C19 (dih_2) (Figure 2.2a), with CGenFF penalty scores of 75 and 46.5 respectively (Table 2.1). This highlights the effectiveness of CGenFF penalty scoring, and as these two parameters are of paramount importance to accurate conformational sampling in MD simulations of PAH-DNA adduct systems, they are the primary focus of this phase of our parameterization efforts. The ffTK optimized dihedral terms for dih_1 and dih_2 are listed in Table 2.2 for each of the three model systems

	Atom Names	Atom Types	n	k_n(kcal/mol)	δ_n	Penalty Score
dih_1	C6-N6-C20-C20a	CN2-NN1-CG311-CG2R61	1	2.5	180.00°	75
			2	1.5	0.00°	
			3	0.5	0.00°	
dih_2	C6-N6-C20-C19	CN2-NN1-CG311-CG311	1	2.5	180.00°	46.5
			2	1.5	0.00°	
			3	0.5	0.00°	

Table (2.1) CGenFF dihedral terms for the adduct covalent bond assigned by analogy.

		NAP		PHE		B[c]P	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	1.131	180.00°	2.520	180.00°	2.969	180.00°
	2	1.412	180.00°	1.442	180.00°	2.755	180.00°
	3	0.677	180.00°	1.308	180.00°	0.031	0.00°
dih_2	1	1.508	0.00°	1.562	0.00°	3.500	0.00°
	2	0.825	180.00°	0.377	180.00°	0.669	0.00°
	3	0.095	0.00°	1.094	0.00°	0.056	0.00°

Table (2.2) ffTK optimized dihedral terms for the adduct covalent bond

Relaxed MM PES scans of the adduct covalent bond's dihedral angle ϕ analogous to the relaxed QM PES scans in Figure 2.3 were conducted utilizing unmodified CGenFF as well as ffTK optimized dihedral terms in order to examine the efficacy of the parameterization approach described above and to highlight the effectiveness of CGenFF penalty scoring. The relaxed MM PES scans were conducted by taking the optimized geometries from the relaxed QM PES scan and fixing atoms C6-N6-C20-C20a (dih_1, Figure 2.2a). Each structure was then subjected to 1000 steps of Conjugate-Gradient minimization in NAMD,[56,57] taking the lowest energy structure from each minimization. With the relative energies of each model system's QM PES as target data, errors are listed for the MM PES fit using CGenFF and ffTK dihedral terms in Table 2.3. In all three systems, the RMSE exceeds the 1.0 kcal/mol target for chemical accuracy. Whether or not a set of dihedral terms is satisfactory for modeling a given system is typically judged by these errors along with the overall quality of the MM PES fit to the shape of the QM PES.[40,47] These are shown in Figure 2.4 and described below.

For the NAP model system, CGenFF dihedral terms result in a MM PES with spurious maxima at $\phi = 50°$ and $\phi = -140°$ and spurious minima at $\phi = 40°$, $\phi = -50°$, and $\phi = -70°$ resulting in a poor overall fit (Figure 2.4a - left). The set of ffTK dihedral terms for the NAP model system result in a MM PES that largely approximates the shape of the QM PES (Figure 2.4a - right).

For the PHE model system, CGenFF dihedral terms result in a MM PES that fails to approximate local minima and maxima on the QM PES at $\phi = -20°$ and $\phi = -90°$ respectively. The absolute maximum is overestimated by approximately 8 kcal/mol and shifted from $\phi = 40°$ to $\phi = 30°$ (Figure 2.4b - left). The set of ffTK dihedral terms for the PHE model system result in a MM PES with a spurious inflection point $\phi = -120°$ that does not approximate the local maximum at $\phi = -90°$ and it is shifted an average ± 1.5 kcal/mol from $\phi = -110°$ to $\phi = 30°$ (Figure 2.4b - right).

For the B[c]P model system, CGenFF dihedral terms result in a MM PES with a spurious minima at $\phi = -60°$ and $\phi = 130°$ and spurious maxima at $\phi = -170°$ and $\phi = -140°$ resulting in a poor overall fit. The absolute maximum on the QM PES is overestimated by approximately 9

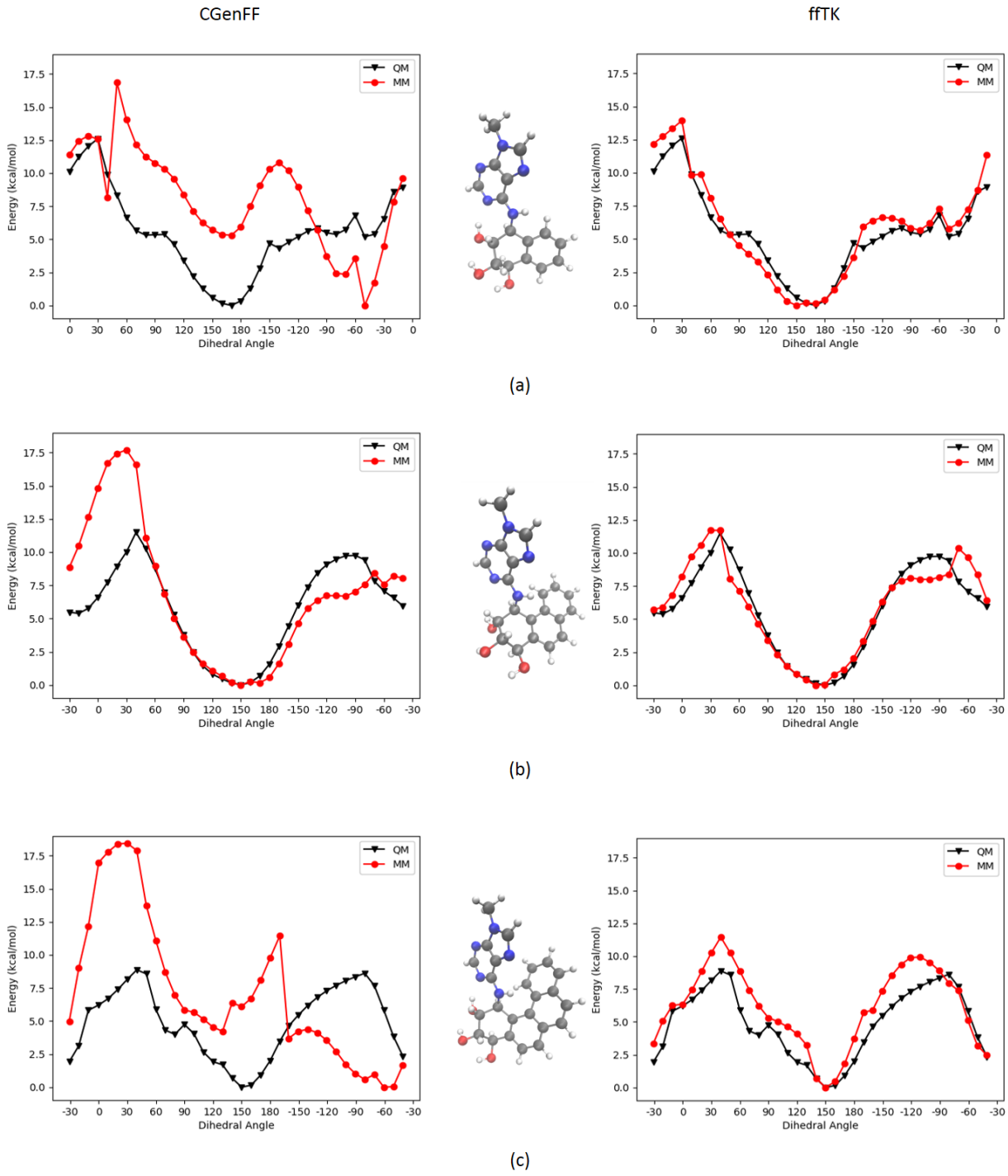
	CGenFF Dihedral Terms			ffTK Dihedral Terms		
	NAP	PHE	B[c]P	NAP	PHE	B[c]P
max abs error	8.5106	8.9684	11.1084	2.4170	2.5971	3.1182
RMSE	4.5752	3.5165	6.0769	1.0757	1.1863	1.6701

Table (2.3) Error Data: Adduct covalent bond dihedral angle (ϕ) MM PES fit to QM PES - ffTK optimized dihedral terms

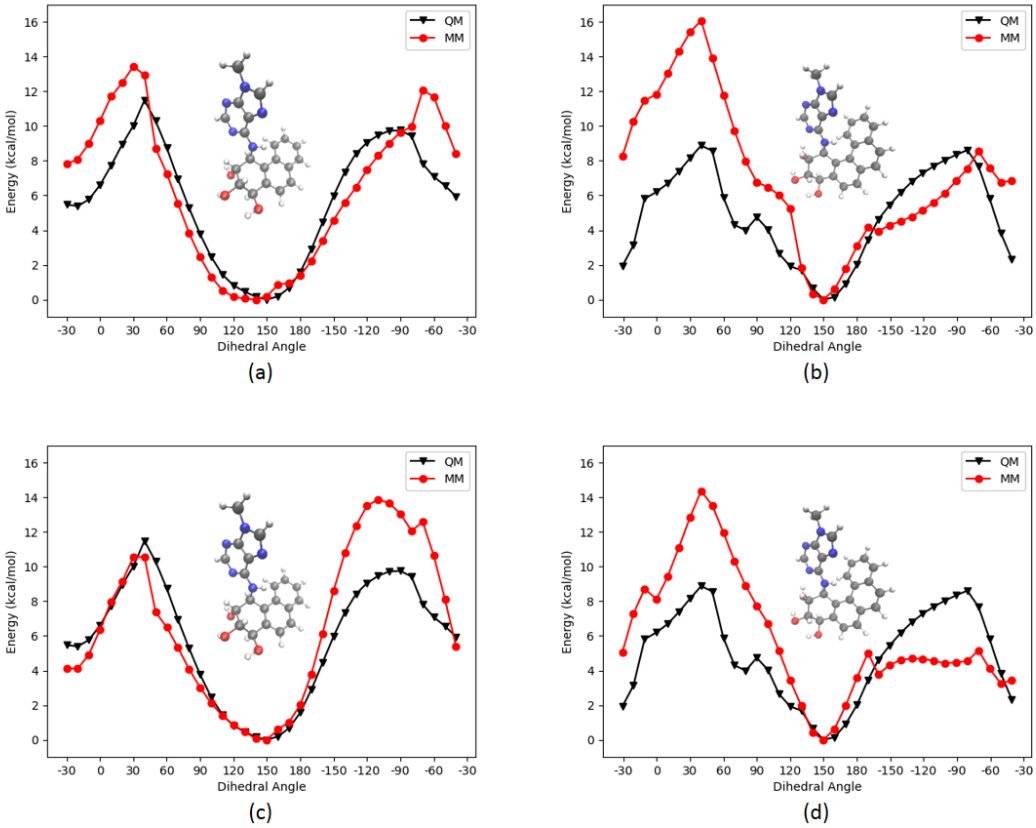
kcal/mol and shifted from $\phi = 40°$ to $\phi = 30°$ (Figure 2.4c - left). The set of ffTK dihedral terms for the B[c]P model system result in a MM PES that overestimates the absolute maximum at $\phi = 40°$ by approximately 3 kcal/mol and shifts the local maximum on the QM PES from $\phi = -80°$ to $\phi = -110°$. The MM PES is shifted an average of +2.2 kcal/mol from $\phi = -170°$ to $\phi = -110°$ and an average of +2.0 kcal/mol from $\phi = 30°$ to $\phi = 120°$ (Figure 2.4c - right).

The ffTK optimized dihedral parameters from each model system were also evaluated for transferability among the model systems. Dihedral terms from the NAP model system were applied to MM PES scans of the adduct covalent bond dihedral angle (ϕ) in the PHE and B[c]P model systems resulting in a poor overall fit in both cases with RMSE_{PHE}=2.1405 kcal/mol (Figure 2.5a) and $\text{RMSE}_{B[c]P}$=3.9011 kcal/mol (Figure 2.5b). Additionally, dihedral terms derived from the PHE model system were applied to a MM PES scan of the adduct covalent bond dihedral angle (ϕ) in the B[c]P model system and vice-versa. Both again resulted in a poor overall fit with RMSE_{PHE}=2.1917 kcal/mol (Figure 2.5c) and $\text{RMSE}_{B[c]P}$=3.1026 kcal/mol (Figure 2.5d).

Conformational sampling in MD simulations will show a preference for lower energy conformations associated with lower energy regions of the corresponding MM PES, hence it is desirable to achieve an accurate fit to the QM PES in those regions (i.e. 5 kcal/mol or less relative to the absolute minimum).[40] Additionally, a good overall fit should be achieved in those regions below 12 kcal/mol relative to the the absolute minimum.[47] Of particular importance is the location of



(2.4) MM PES fit (red circles) to QM PES (black triangles) for the adduct covalent bond dihedral angle (ϕ) for model systems: (a) NAP, (b) PHE, (c) B[c]P. Left column: with unmodified CGenFF dihedral parameters assigned by analogy. Right column: with ffTK optimized dihedral parameters.



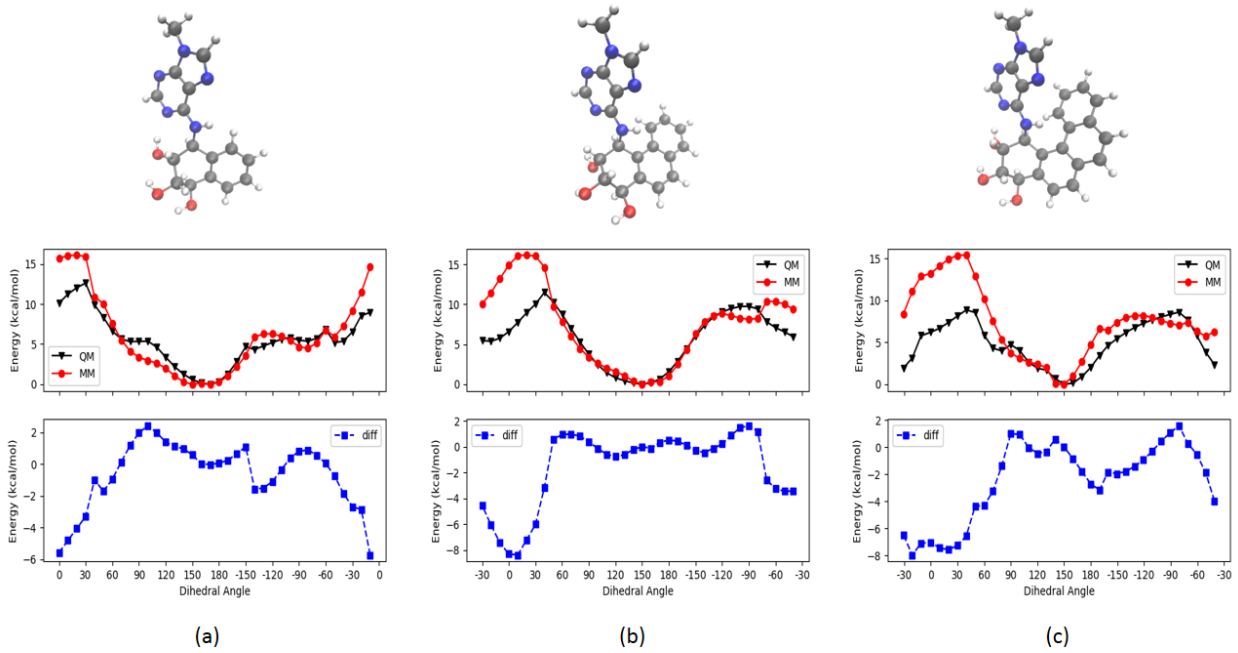
(2.5) MM PES fit (red circles) to QM PES (black triangles) for the adduct covalent bond dihedral angle (ϕ) with ffTK optimized dihedral parameters derived from: (a) NAP model system applied to the PHE model system, (b) NAP model system applied to the B[c]P model system, (c) B[c]P model system applied to the PHE model system, (d) PHE model system applied to the B[c]P model system.

minima and maxima, as well as their respective depths and heights, as these features will largely influence the systems ability to achieve conformational changes and/or the range of the geometric parameters sampled during MD simulations. The accuracy of the adduct covalent bond's torsional MM PES is thus a crucial feature as PAH-DNA adduct genotoxicity is widely thought to be a function of geometric conformation as described above.[12,13,16,17,27,37,53] Using either unmodified CGenFF or ffTK optimized dihedral terms for dih_1 and dih_2, the RMSE for each model system's MM PES fit to the corresponding QM target data examined above exceeds the 1.00 kcal/mol target for chemical accuracy. Additionally, we have shown that NAP does not suffice as a model system to parameterize bay and fjord region PAH model systems. Furthermore, we have shown that dihedral parameters for the adduct covalent bond derived from a bay region PAH model system do not perform well in a fjord region PAH model system and vice-versa.

2.1.3 Least Squares Optimization of Dihedral Parameters

It is clear that bay and fjord region PAH-DNA adduct systems require custom dihedral terms beyond what is achieved with standard parameterization approaches consisting of parameter assignments by analogy, Monte Carlo Simulated Annealing, and Downhill Simplex. To proceed, we will examine the application of least squares fitting to QM target data to optimize dihedral terms. We will begin by optimizing only the dih_1 dihedral parameter, which has the highest CGenFF penalty score of 75, while retaining the remaining CGenFF and ffTK optimized parameters described above. This will allow us to examine the efficacy of least squares fitting in dihedral parameter optimization on a linear system that is not ill-conditioned. In Chapter 4 we will examine approaches for simultaneously optimizing multiple dihedrals to avoid over-fitting and approaches to regularize ill-conditioned systems of equations.

For each of the three model systems, a relaxed MM PES scan of the adduct covalent bond's dihedral angle ϕ analogous to those described above is conducted with the force constants for dih_1 set to zero. Following the approach described by Guvench and MacKerell[48] and where θ_i is the i^{th} discrete scan point, the energies $E_i^{MM_{k_{dih1}=0}}$ of the resulting MM PES are then subtracted from the



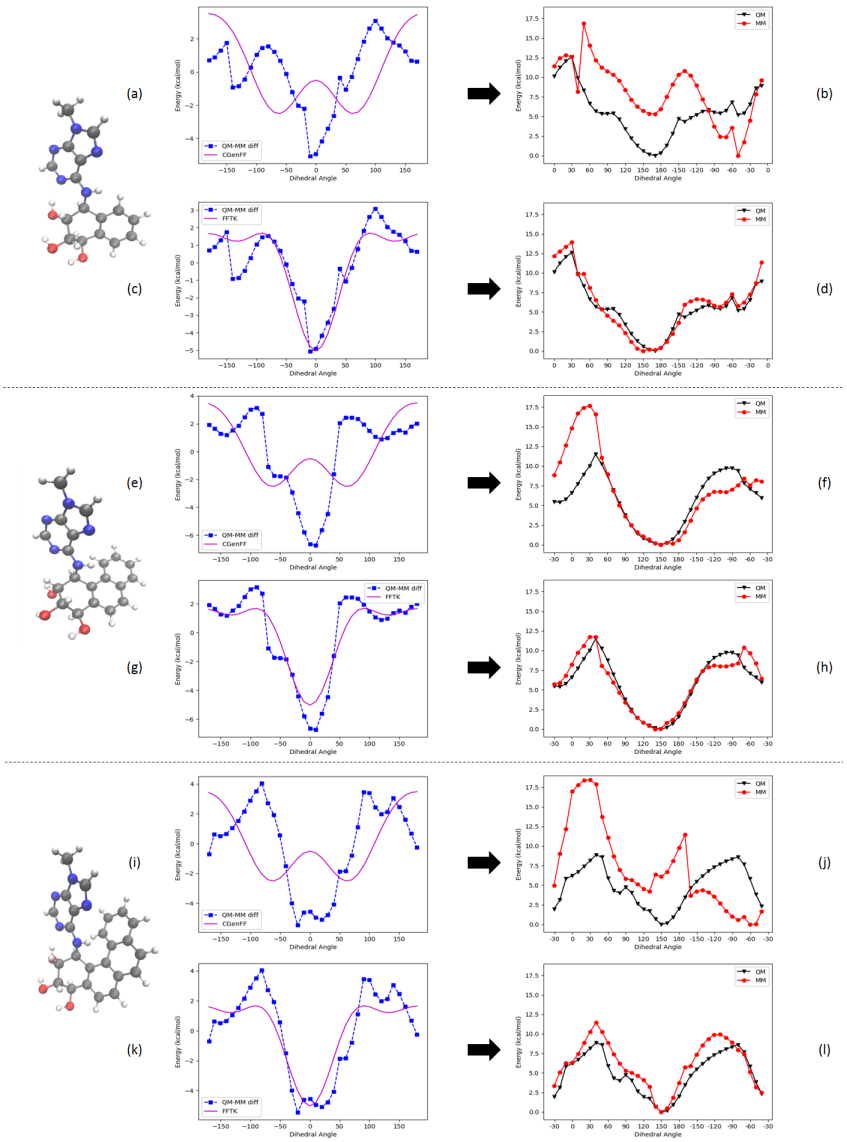
(2.6) Dihedral difference potentials (blue squares) derived from the adduct covalent bond dihedral angle (ϕ) QM PES (black triangles) and MM PES (red circles) with dih_1 force constants set to zero for model systems: (a) NAP, (b) PHE, (c) B[c]P.

energies E_{i}^{QM} of the QM target data described above. The resulting discrete difference potential $E_{diff} = \{E_1, ..., E_m\}$ where $E_{\mathrm{i}} = E_{\mathrm{i}}^{QM} - E_{\mathrm{i}}^{MM_{k_{dih1}=0}}$ then reveals the form of the dihedral potential needed to correct the fit of the MM PES to the QM PES in each of our three model systems (Figure 2.6). The dihedral force constants for dih_1 can then be further optimized to fit this difference potential.

In the case of our model systems, the difficulty in developing dihedral parameters that accurately model QM target data arises due to symmetric dihedral potentials that fail to accurately model the asymmetric dihedral difference potentials plotted in Figure 2.6. The CHARMM class I additive molecular mechanics force field uses a phased cosine series to model the potential energy of a given dihedral parameter ϕ:

$$E_{dihedral_\phi} = \sum_{\mathrm{j}\in M} k_\mathrm{j}\left[1+\cos(\mathrm{j}\phi-\delta_\mathrm{j})\right] \tag{2.1}$$

where $M \subseteq \{1,2,3,4,5,6\}$ and $\delta_\mathrm{j} = 0°$ or $180°$ by convention. Dihedral force constants are the coefficients k_j in (2.1) and they are typically optimized by fitting data from dihedral difference potentials.[47,48] However, restricting the phase constants to $\delta_\mathrm{j} = 0°$ or $180°$, as in the unmodified CGenFF and ffTK optimized dihedral terms for the dih_1 parameter, results in a linear combination of even functions, which will only ever produce an even function. The resulting symmetric dihedral potential is generally a poor fit to the asymmetric difference potentials that arise from our model systems (Figure 2.7 - left column). This in turn results in the non-optimal fit of the MM PES to QM target data (Figure 2.7 - right column).



(2.7) Left column: Unmodified CGenFF and ffTK optimized dih_1 dihedral potentials (magenta line) fit to target difference potential (blue squares). Right column: Resulting MM PES (red circles) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ. Model systems: (a-d) (+)-trans-NAP-DE-N6-dA, (e-h) (+)-trans-PHE-DE-N6-dA, (i-l) (+)-trans-B[c]P-DE-N6-dA. Note that difference potential plots in the left column are centered at 0° in order to illustrate the asymmetric target difference potentials. The resulting MM PES plots in the right column are centered at their respective absolute minima.

Symmetric dihedral potentials are a required feature of a generalized force field, otherwise stereoisomers would result in different torsional energies and symmetric molecules would have asymmetric dihedral potentials, which is unphysical.[40,47,48] Indeed, CGenFF was designed to be a generalized force field for modeling symmetric molecules and stereoisomers of asymmetric molecules, precluding the use of asymmetric dihedrals in general.[40] However, it has been noted[47,58,59] that when developing force field parameters for asymmetric molecules where the chirality is always the same, asymmetric dihedral potentials can be utilized. Because dihedral parameters are largely a correction to 1-4 non-bonded interactions, dihedrals that link ring systems (such as our adduct covalent bond) are typically optimized last in a novel residue and are often unsatisfactory when used transferably between different molecular systems.[40,47] As a result, custom parameters are often substituted into force fields for such dihedrals in order to improve the fit to QM target data.[59]

Thus to accurately fit the asymmetric dihedral difference potentials in our model systems, it is necessary to use the complete basis $\{1, \cos(\mathrm{j}\phi), \sin(\mathrm{j}\phi) \,|\, \mathrm{j} = 1,2,3...\}$. Following the approach described by Hopkins and Roitberg[59] and also examined by Vanommeslaeghe et. al.,[47] the dihedral force constants for $dih1$ are optimized for each model system by calculating the coefficients a_j and b_j that achieve a least squares fit of the truncated Fourier series (2.2) to the corresponding dihedral difference potential and then transforming to the desired force constants k_j in (2.1). However, in this work, the variable phase constants δ_j for the phased cosine series in (2.1) are instead calculated by consistent use of the principle value of the argument of the corresponding point in the complex plane rather than by taking the arctan of a quotient. The necessity for this is explained below.

$$E_{dihedral_\phi} = \sum_{\mathrm{j}} [a_\mathrm{j} \cos(\mathrm{j}\phi) + b_\mathrm{j} \sin(\mathrm{j}\phi)] \quad where \quad \mathrm{j} \leq 6 \tag{2.2}$$

Using this approach to further optimize the dih_1 parameter for each model system will result in an improved fit of the corresponding MM PES to QM target data, which is imperative to accurate conformational sampling in MD simulations of PAH-DNA adduct systems as described above. For emphasis, it is well understood that the inclusion of dihedral multiplicities that are not appropriate

to the symmetry of a given molecule would in general be unphysical,[47,58] and we do not assert that any and all dihedral parameters should be optimized by fitting general Fourier series. The custom dihedral terms for each model system are only intended for use in structurally (i.e. bay vs. fjord) and stereochemically analogous PAH-DNA adduct systems as a correction to the force fields for these systems.

Letting $\theta_1, ..., \theta_m$ be the discrete scan angles utilized in the QM PES scans of the adduct covalent bond dihedral angle ϕ, and letting n be the number of terms in the truncated Fourier series (2.2), we seek a solution vector $\mathbf{x}$ to the matrix equation $\mathbf{Ax} = \mathbf{b}$ that minimizes the 2-norm of the residual $\|\mathbf{r}\|_2^2 = \|\mathbf{b} - \mathbf{Ax}\|_2^2$ where:

$$A_{\mathrm{i},2\mathrm{j}-1} = \cos(\mathrm{j}\theta_\mathrm{i}) - \overline{C_\mathrm{j}} \quad \text{for} \quad \mathrm{i} = 1...m \quad \text{and} \quad \mathrm{j} = 1...n \tag{2.3}$$

$$A_{\mathrm{i},2\mathrm{j}} = \sin(\mathrm{j}\theta_\mathrm{i}) - \overline{S_\mathrm{j}} \quad \text{for} \quad \mathrm{i} = 1...m \quad \text{and} \quad \mathrm{j} = 1...n \tag{2.4}$$

$$\overline{C_\mathrm{j}} = \frac{1}{m}\sum_{\mathrm{i}=1}^{m} \cos(\mathrm{j}\theta_\mathrm{i}) \quad \mathrm{j} = 1...n \tag{2.5}$$

$$\overline{S_\mathrm{j}} = \frac{1}{m}\sum_{\mathrm{i}=1}^{m} \sin(\mathrm{j}\theta_\mathrm{i}) \quad \mathrm{j} = 1...n \tag{2.6}$$

$$b_\mathrm{i} = E_\mathrm{i} - \overline{E} \quad \text{for} \quad \mathrm{i} = 1...m \tag{2.7}$$

$$\overline{E} = \frac{1}{m}\sum_{\mathrm{i}=1}^{m} E_\mathrm{i} \tag{2.8}$$

The quantities $\overline{C_\mathrm{j}}$, $\overline{S_\mathrm{j}}$, and $\overline{E}$ are constants that are used to shift the respective data sets so that their averages are zero.[47,48] In general terms, the residual $\|\mathbf{r}\|_2^2$ is a differentiable function of $\mathbf{x}$, hence a solution vector that minimizes $\|\mathbf{r}\|_2^2$ corresponds to $\mathbf{x}$ such that $\nabla\|\mathbf{r}\|_2^2 = 0$.[60,61] The matrix equation $\mathbf{Ax} = \mathbf{b}$ when expanded takes the form:

$$\begin{bmatrix} \cos(1\theta_1)-\overline{C_1} & \sin(1\theta_1)-\overline{S_1} & \ldots & \cos(n\theta_1)-\overline{C_n} & \sin(n\theta_1)-\overline{S_n} \\ \cos(1\theta_2)-\overline{C_1} & \sin(1\theta_2)-\overline{S_1} & \ldots & \cos(n\theta_2)-\overline{C_n} & \sin(n\theta_2)-\overline{S_n} \\ \cos(1\theta_3)-\overline{C_1} & \sin(1\theta_3)-\overline{S_1} & \ldots & \cos(n\theta_3)-\overline{C_n} & \sin(n\theta_3)-\overline{S_n} \\ \vdots & \vdots & & \vdots & \vdots \\ \cos(1\theta_m)-\overline{C_1} & \sin(1\theta_m)-\overline{S_1} & \ldots & \cos(n\theta_m)-\overline{C_n} & \sin(n\theta_m)-\overline{S_n} \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \\ \vdots \\ a_n \\ b_n \end{bmatrix} = \begin{bmatrix} E_1-\overline{E} \\ E_2-\overline{E} \\ E_3-\overline{E} \\ \vdots \\ E_m-\overline{E} \end{bmatrix}$$

The residual vector $\mathbf{r}$ then has elements of the form:

$$r_{\mathrm{i}} = \left(E_{\mathrm{i}}-\overline{E}\right) - \sum_{\mathrm{j}=1}^{n}\left[a_{\mathrm{j}}A_{\mathrm{i},2\mathrm{j}-1} + b_{\mathrm{j}}A_{\mathrm{i},2\mathrm{j}}\right] \tag{2.9}$$

$$r_{\mathrm{i}} = \left(E_{\mathrm{i}}^{QM} - E_{\mathrm{i}}^{MM_{k_{dih1}=0}} - \overline{E}\right) - \sum_{\mathrm{j}=1}^{n}\left[a_{\mathrm{j}}\left(\cos(\mathrm{j}\theta_{\mathrm{i}})-\overline{C_{\mathrm{j}}}\right) + b_{\mathrm{j}}\left(\sin(\mathrm{j}\theta_{\mathrm{i}})-\overline{S_{\mathrm{j}}}\right)\right] \tag{2.10}$$

Before converting a_{j} and b_{j} to the CHARMM requisite k_{j} and δ_{j} in Equation (2.1), we note that for a given θ_{i}, up to a constant the MM energy of the system $E_{\mathrm{i}}^{MM_{k_{dih1}=opt}}$ resulting from the least squares optimized force constants for dih_1 is:

$$E_{\mathrm{i}}^{MM_{k_{dih1}=opt}} = E_{\mathrm{i}}^{MM_{k_{dih1}=0}} + \sum_{\mathrm{j}=1}^{n}\left[a_{\mathrm{j}}\cos(\mathrm{j}\theta_{\mathrm{i}}) + b_{\mathrm{j}}\sin(\mathrm{j}\theta_{\mathrm{i}})\right] \tag{2.11}$$

We have then that:

$$r_{\mathrm{i}} = E_{\mathrm{i}}^{QM} - E_{\mathrm{i}}^{MM_{k_{dih1}=opt}} - \tilde{E} \quad where \quad \tilde{E} = \overline{E} - \sum_{\mathrm{j}=1}^{n}\left(a_{\mathrm{j}}\overline{C_{\mathrm{j}}} + b_{\mathrm{j}}\overline{S_{\mathrm{j}}}\right) \tag{2.12}$$

and so

$$\|\mathbf{r}\|_2^2 = \sum_{\mathrm{j}=1}^{n}\left(E_{\mathrm{i}}^{QM} - E_{\mathrm{i}}^{MM_{k_{dih1}=opt}} - \tilde{E}\right)^2 \tag{2.13}$$

We have then that the a_{j} and b_{j} that minimize $\|\mathbf{r}\|_2^2$ correspondingly minimize:

$$RMSE = \sqrt{\frac{\sum_{\mathrm{i}=1}^{m}\left(E_{\mathrm{i}}^{QM} - E_{\mathrm{i}}^{MM_{k_{dih1}=opt}}\right)^2}{m}} \tag{2.14}$$

which in turn optimizes the force field to fit the target QM PES.[47,48]

Note that because we conducted equispaced PES scans over the interval (-180°,180°] that include 0°, the constants $\overline{C_j}$ and $\overline{S_j}$ are zero. Furthermore the columns of **A** are orthogonal. This can be seen by considering the m x n complex matrix that corresponds to the m x $2n$ real matrix **A** (note that below $i = \sqrt{-1}$ where as "i" is an index above):

$$\begin{bmatrix} \cos(1\theta_1)-\overline{C_1} & \sin(1\theta_1)-\overline{S_1} & \ldots & \cos(n\theta_1)-\overline{C_n} & \sin(n\theta_1)-\overline{S_n} \\ \cos(1\theta_2)-\overline{C_1} & \sin(1\theta_2)-\overline{S_1} & \ldots & \cos(n\theta_2)-\overline{C_n} & \sin(n\theta_2)-\overline{S_n} \\ \cos(1\theta_3)-\overline{C_1} & \sin(1\theta_3)-\overline{S_1} & \ldots & \cos(n\theta_3)-\overline{C_n} & \sin(n\theta_3)-\overline{S_n} \\ \vdots & \vdots & & \vdots & \vdots \\ \cos(1\theta_m)-\overline{C_1} & \sin(1\theta_m)-\overline{S_1} & \ldots & \cos(n\theta_m)-\overline{C_n} & \sin(n\theta_m)-\overline{S_n} \end{bmatrix} \leftrightarrow \begin{bmatrix} e^{i1\theta_1} & \ldots & e^{in\theta_1} \\ e^{i1\theta_2} & \ldots & e^{in\theta_2} \\ e^{i1\theta_3} & \ldots & e^{in\theta_3} \\ \vdots & & \vdots \\ e^{i1\theta_m} & \ldots & e^{in\theta_m} \end{bmatrix}$$

Observing that $\theta_\alpha = \{-\pi + \frac{\alpha * 2\pi}{m} | \alpha = 1...m\}$ (i.e. $2\pi/m$ is the scan step size where m is even) we have:

$$\sum_{\alpha=1}^{m} e^{ik\theta_\alpha} e^{-il\theta_\alpha} = \begin{cases} 0 \text{ if } k \neq l \\ m \text{ if } k = l \end{cases} \tag{2.15}$$

and so the columns of **A** are orthogonal and **A** has full rank. As a result, we can simply solve the Normal Equations: $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ where $\mathbf{A}^*$ is the conjugate transpose.[60,61] Since **A** is orthogonal, we have $\mathbf{A}^*\mathbf{A} = c\mathbf{I}$ where $c = m/2$ is a constant and **I** is the identity. We then have $\mathbf{x} = \frac{1}{c}\mathbf{A}^*\mathbf{b}$. Note that this is effectively forming the Discrete Fourier Transform. We will examine regularization approaches in the next chapter for the more general case where the matrix **A** may not be orthogonal or may be ill-conditioned. The coefficients a_j and b_j of the truncated Fourier series (2.2) obtained from the least squares solution above are transformed into the CHARMM requisite force constants k_j of the phased cosine series in Equation (2.1) by simple application of the Pythagorean Theorem:

$$k_j = \sqrt{a_j^2 + b_j^2}. \tag{2.16}$$

Note however, the variable phase constants δ_j in Equation (2.1) must be determined using the principle value of the argument of the point (a_j, b_j) in the complex plane:

$$\delta_j = \mathrm{Arg}(a_j + ib_j) \in (-\pi, \pi] \tag{2.17}$$

and not by using $\arctan\left(\frac{b_j}{a_j}\right)$. This difficulty arises because $\tan(\delta_j)$ is not one-to-one on the entire real line. In order to construct the inverse function $\delta_j = \arctan\left(\frac{b_j}{a_j}\right)$ one must select a subinterval of the real line (or branch of the graph) on which $\tan(\delta_j)$ is one-to-one. By convention, this subinterval is chosen to be $(-\frac{\pi}{2}, \frac{\pi}{2})$. Consequently, if $\arctan\left(\frac{b_j}{a_j}\right)$ is used to calculate the corresponding phase constants, δ_j will be forced to take values in the range $(-\frac{\pi}{2}, \frac{\pi}{2})$ and require manual inspection and adjustment to obtain the desired value. Instead, the principle value of the argument should be implemented using the function atan2 which is included in the majority of programming language math libraries (i.e. $\delta_j = \mathrm{atan2}(b_j, a_j)$). We note that while this mathematical detail is not mentioned in previous literature that examines this approach, the atan2 function is implemented in the lsfitpar.c source code.[47]

2.2 Results and Discussion

2.2.1 Optimized Dihedral Parameters

Applying the least squares fitting approach described above to optimize the dih_1 dihedral parameter for each of our model systems, we developed three and six term phased cosine series labeled LS_3 & LS_6 respectively (Table 2.4), resulting in markedly improved fits to the difference potential of each model system (Figure 2.8 - left column).

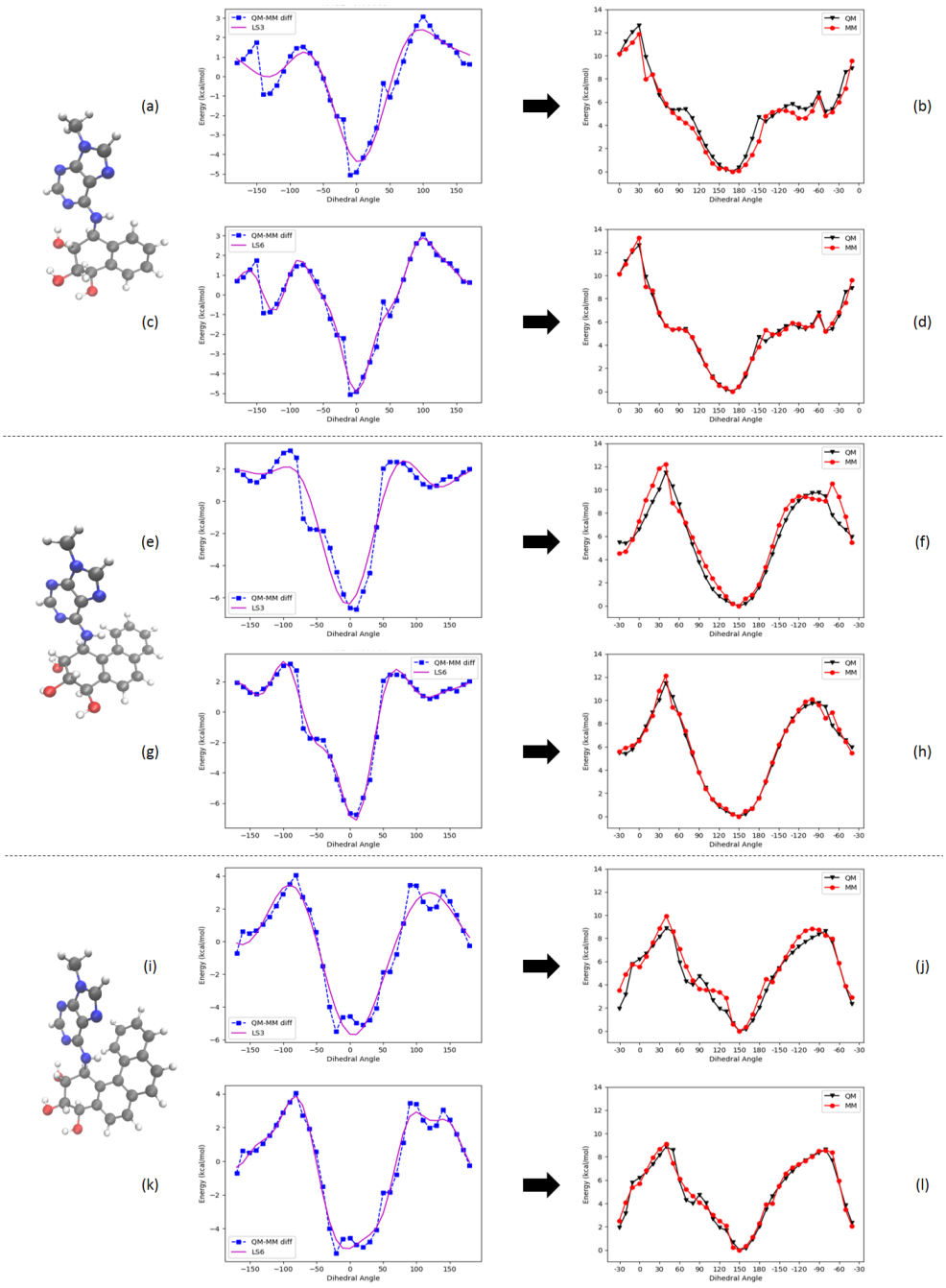
The relaxed MM PES scans of the adduct covalent bond dihedral angle (ϕ) were repeated for each of the model systems with both the LS_3 & LS_6 least squares optimized dih_1 dihedral terms, resulting in markedly improved fits of their respective MM potential energy surfaces to the corresponding QM target data (Figure 2.8 - right column). As listed in Table 2.5, the LS_3 parameter set resulted in a RMSE less than 1.0 kcal/mol and the LS_6 parameter set resulted in a RMSE less than 0.5 kcal/mol in all three model systems.

	NAP		PHE		B[c]P	
n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
1	1.9559	165.2912°	2.9911	176.4400°	2.7834	-170.7915 °
2	1.8223	-161.1383°	2.3374	166.0241°	2.8772	-159.3861 °
3	0.7635	-168.5205°	1.2210	176.7705°	0.4028	120.6574 °
1	1.9559	165.2912°	2.9910	176.4403°	2.7829	-170.7839 °
2	1.8223	-161.1383°	2.3374	166.0239°	2.8758	-159.3766 °
3	0.7635	-168.5206°	1.2210	176.7686°	0.4018	120.6328 °
4	0.4266	84.1696°	0.8175	-90.7949°	0.3840	46.0026 °
5	0.2917	-137.1010°	0.6996	-109.5474°	0.5031	38.3541 °
6	0.4287	171.3839°	0.1776	-179.0679°	0.2011	-164.2161 °

Table (2.4) LS_3 & LS_6 optimized dihedral terms for the adduct covalent bond dihedral parameter *dih*1

	LS_3 Dihedral Terms			LS_6 Dihedral Terms		
	NAP	PHE	B[c]P	NAP	PHE	B[c]P
max abs error	2.0603	2.7300	1.7531	0.9682	1.1329	1.1382
RMSE	0.7755	0.9769	0.7886	0.3806	0.4036	0.4751

Table (2.5) Error Data: Adduct covalent bond dihedral angle (ϕ) MM PES fit to QM PES - LS_3 & LS_6 optimized dihedral terms



(2.8) Left column: LS_3 & LS_6 dih_1 optimized dihedral potentials (magenta line) fit to target difference potential (blue squares). Right column: Resulting MM PES (red circles) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ. Model systems: (a-d) NAP, (e-h) PHE, (i-l) B[c]P. Note that difference potential plots in the left column are centered at 0° in order to illustrate the asymmetric target difference potentials. The resulting MM PES plots in the right column are centered at their respective absolute minima.

2.2.2 Molecular Dynamics Simulations

In order to compare the utility of the three term ffTK and three term LS_3 optimized dihedral terms in a simple test MD simulation of a PAH-DNA adduct system, we turn again to the syn-glycosidic (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA^*_5 adduct in the 5'-d(GGTCA^*_5CGAG)-3' 5'-d(CTCGGGACC)-3' DNA duplex with a mismatched dG_{14} opposite dA^*_5 (Figure 2.1 PDB: 1DXA[1]). In order to model the dynamics of this PAH-DNA adduct system, we developed a custom residue analogous to our model systems described above consisting of the entire adenine residue (ADE) from the CHARMM-NA force field and a bay region B[a]P adduct constructed analogously to our model systems above, resulting in a (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA custom residue that is compatible with the rest of the CHARMM-NA force field. We also developed an analogous fjord region (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA custom residue, minimizing its structure at the MP2/6-31G(d) level of theory, and placing it in the same 5'-d(GGTCA^*_5CGAG)-3' 5'-d(CTCGGGACC)-3' DNA duplex, applying 1000 steps of conjugate gradient minimization in NAMD prior to conducting MD to relieve steric clashes created between the DNA duplex and the additional aromatic ring in the (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA^*_5 custom residue. We emphasize that these MD simulations are a simple test case examining the usability of least squares optimized dihedral terms, and they are not meant to derive structural or thermodynamic conclusions. Such simulations are the subject of Chapters 4 and 5 in this work.

Molecular dynamics simulations were conducted in NAMD[56,57] utilizing the CHARMM36-NA and CGenFF 4.1 force fields along with custom topology/parameter files for the custom B[a]P and DB[a,l]P residues and accompanying parameters. Eighteen Na^+ counter ions were placed along the 5'-d(GGTCA^*_5CGAG)-3' 5'-d(CTCGGGACC)-3' DNA backbone using the CIonize VMD[62] plugin. The system was then solvated in a 60 Å TIP3P[63] explicit water box with the addition of Na^+ and Cl^- ions using the Solvate and Autoionize plugins in VMD to achieve a 120mM NaCl solution ensuring that the solvent extends at least 10 Å beyond the solute. The system was relaxed over 1000 steps of conjugate-gradient minimization with harmonic constraints applied to

the DNA duplex. This was followed by 1 ns of NVT simulation and 5 ns of NPT simulation with harmonic constraints still in place in order to avoid the possibility of the system volume changing too rapidly at the outset of NPT MD. Harmonic restraints were then released and an extra bond added between the dG-N1 and dC-N3 nitrogens in the terminal dG-dC nucleobase pairs in order to avoid end fraying. Production NPT MD simulations were run for 200 ns with periodic boundary conditions at 300 K and 1 atm utilizing Langevin dynamics and the Langevin piston.[64] Electrostatic interactions were treated utilizing the Particle Mesh Ewald[65] method with a cutoff of 12 Å. Lennard-Jones interactions were treated by activating the switching function at 10 Å. RigidBonds was set to all[66,67] in order to utilize a 2 fs time step.

Note that dihedral parameters for phosphate linkages in the DNA backbone that were modified for the CHARMM36-NA force field were reverted to their CHARMM27-NA values as MD simulations of our PAH-DNA adduct systems became unstable at approximately 100 ns during test runs, with several nucleobases and the PAH-DE extruding out of the DNA double helix when utilizing the CHARMM36-NA values. As described by Minhas et al.,[68] these dihedral parameters were modified in CHARMM36-NA in order to improve BI/BII conformational sampling over CHARMM27-NA. However, this causes increased flexibility of the DNA backbone that was found to result in instability of DNA on the microsecond time scale.[68] In the case of our PAH-DNA adduct systems, the presence of the bulky PAH-DNA adduct resulted in unstable trajectories on a shorter time scale.

During 200 ns of MD utilizing ffTK optimized dihedral terms from the bay region PHE model system for the dih_1 dihedral parameter (Table 2.2), the syn-glycosidic B[a]P-DNA adduct system exhibits brief and reversible disruptions of hydrogen bonding in the $dC_4 : dG_{15}$ and $dC_6 : dG_{13}$ base pairs that neighbor the modified dA^*_5 (Figure 2.9 - left - blue and red plots). The syn-glycosidic dA^*_5 then rotates to an anti-glycosidic conformation at approximately 160 ns (Figure 2.9 - left - magenta plot) and persists in this conformation. This results in rupturing of the non-standard $dA_5 : dG_{14}$ base pair that is observed in the starting NMR solution structure[1] (Figure 2.9 - left - green plot) and shifts the B[a]P adduct in the $dG_{13} \mid dG_{14}$ intercalation pocket (Figure 2.9 - left - cyan plot).

It should be noted that previous work has shown that syn-glycosidic (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA$^*_5$ is the major conformer in the 5'-d(GGTCA$^*_5$CGAG)-3' 5'-d(CTCGGGACC)-3' DNA duplex while the anti-glycosidic conformation is a minor conformer.[1,49]

During 200 ns of MD utilizing LS_3 optimized dihedral terms from the bay region PHE model system for the dih_1 dihedral parameter (Table 2.4), the syn-glycosidic B[a]P-DNA adduct system again exhibits brief and reversible disruptions of hydrogen bonding in the dC$_4$: dG$_{15}$ and dC$_6$: dG$_{13}$ base pairs (Figure 2.9 - right - blue and red plots). In this case however, the (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA$^*_5$ maintains its syn-glycosidic conformation with the exception of a brief and reversible rotation to anti-glycosidic at approximately 130 ns (Figure 2.9 - right - magenta plot). Hydrogen bonding in the non-standard dA$_5$: dG$_{14}$ base pair persists for the duration of the simulation with reversible disruptions occurring during the simulation (Figure 2.9 - right - green plot). The B[a]P adduct remains in the dG$_{13}$|dG$_{14}$ intercalation pocket with reversible shifts deeper into the intercalation pocket occurring during the simulation (Figure 2.9 - right - cyan plot).

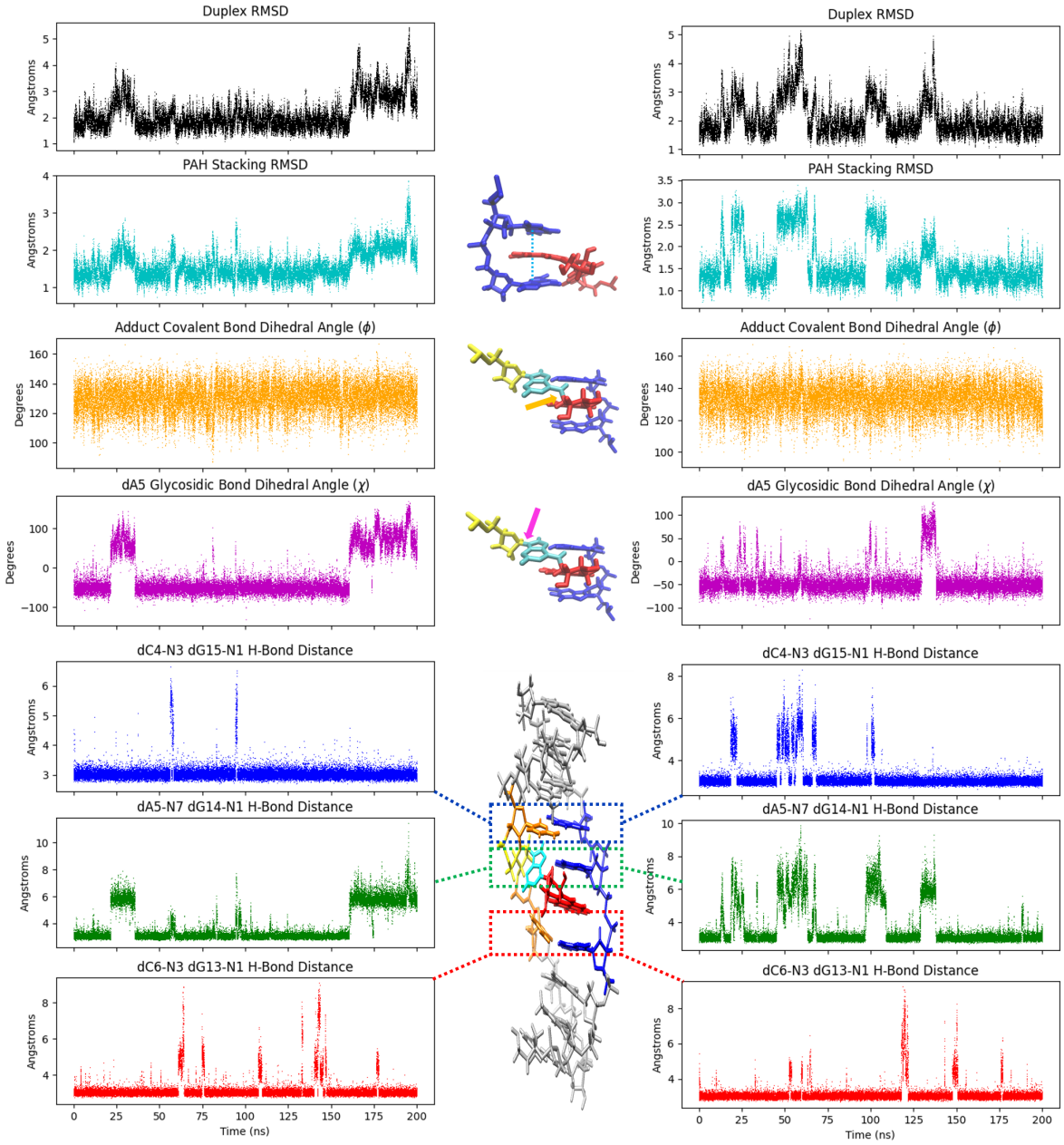
During 200 ns of MD both the ffTK optimized (Table 2.2) and LS$_3$ optimized (Table 2.4) dihedral terms from the fjord region B[c]P model system for the dih_1 dihedral parameter resulted in very similar stable trajectories for the syn-glycosidic DB[a,l]P-DNA adduct system. No disruption of hydrogen bonding was observed in the canonical dC$_4$: dG$_{15}$ and dC$_6$: dG$_{13}$ base pairs nor in the non-standard dA$_5$: dG$_{14}$ base pair (Figure 2.10 - blue, green, and red plots). The (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA$^*_5$ residue maintains its syn-glycosidic conformation for the duration of the trajectory (Figure 2.10 - magenta plots) and the DB[a,l]P adduct remains in the dG$_{13}$ | dG$_{14}$ intercalation pocket for the duration of the trajectory (Figure 2.10 - cyan plots).

Root mean square deviations (RMSD) and standard deviations from the starting NMR solution structure for the DNA duplex and the PAH in the dG$_{13}$ | dG$_{14}$ intercalation pocket for the MD simulations described above are listed in Table 2.6. Note that structural effects such as hydrogen bond disruption and the resulting thermodynamic destabilization of the DNA duplex are thought to be a hallmark feature of intercalated B[a]P-DNA adduct systems and their susceptibility to GG-NER.

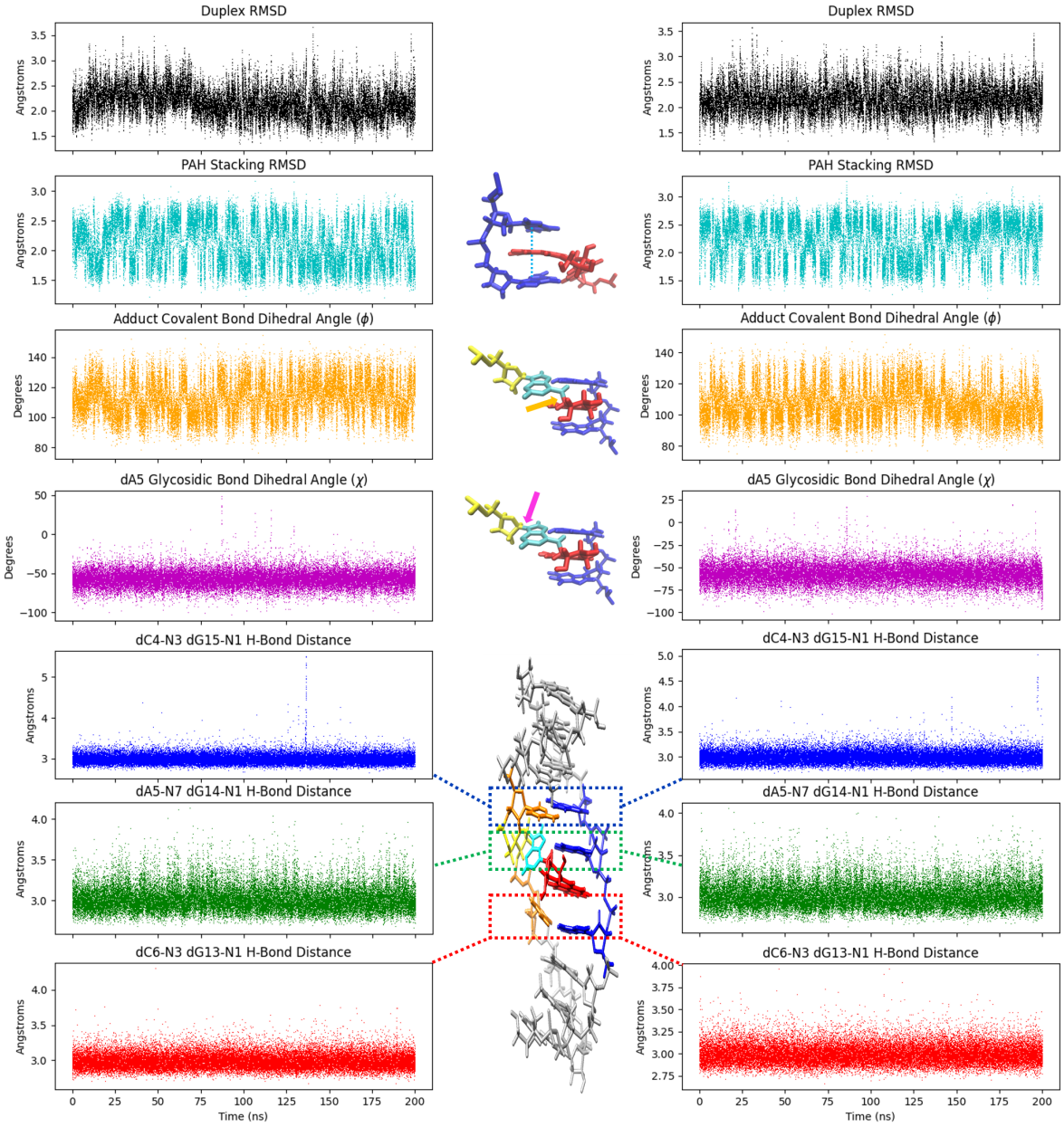
	B[a]P-DE		DB[a,l]P-DE	
	ffTK	LS_3	ffTK	LS_3
DNA Duplex	2.1024±0.6032Å	2.0636±0.6178Å	2.1615±0.2869Å	2.1526±0.2667Å
PAH-DE Intercalation	1.5629±0.3721Å	1.6224±0.5295Å	2.0965±0.3602Å	2.2278±0.3562Å

Table (2.6) RMSD and standard deviation from the starting NMR solution structure for the DNA duplex and the PAH-DE in the dG_{13} | dG_{14} intercalation pocket utilizing ffTK and LS_3 optimized *dih*1 dihedral terms

Disruption of hydrogen bonding with the nucleobase in the unmodified complementary strand opposite the adducted nucleobase (as in the non-standard $dA_5 : dG_{14}$ base pair), and hydrogen bond disruption in neighboring base pairs (as in the $dC_4 : dG_{15}$ and $dC_6 : dG_{13}$ base pairs), is thought to facilitate formation of a productive complex with the XPC-RAD23B protein in the recognition step of the GG-NER pathway. Conversely, the lack of hydrogen bond disruption between base pairs in the DB[a,l]P-DNA adduct system is consistent with the near total resistance to GG-NER in such systems.[12,15,27,53]



(2.9) MD trajectories for syn-glycosidic (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA$^*_5$ in the 5'-d(GGTCA$^*_5$CGAG)-3' 5'-d(CTCGGGACC)-3' DNA duplex with ffTK (left) and LS$_3$ (right) dihedral terms for the adduct covalent bond dihedral parameter *dih*1. Note the glycosidic bond dihedral angle trajectory utilizes C2' in place of O4' to facilitate plotting.



(2.10) MD trajectories for syn-glycosidic (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA$^{*}_{5}$ in the 5'-d(GGTCA$^{*}_{5}$CGAG)-3' 5'-d(CTCGGGACC)-3' DNA duplex with ffTK (left) and LS$_3$ (right) dihedral terms for the adduct covalent bond dihedral parameter *dih*1. Note the glycosidic bond dihedral angle trajectory utilizes C2' in place of O4' to facilitate plotting.

2.3 Conclusion

Studying the structural and thermodynamic features of PAH-DNA adduct systems via molecular dynamics requires accurate parameterization of the covalent bond that links a PAH-DE and a purine. QM PES scans in stereochemically identical NAP, PHE, and B[c]P model systems indicate that the torsional potential energy landscape of the adduct covalent bond is dependent upon the entire PAH structure (i.e. bay vs fjord) and not just upon atomic connectivity. This indicates that dihedral parameters for MM force fields that parameterize the adduct covalent bond are not likely to be transferable among structurally different PAH-DNA adduct systems despite identical atomic connectivity. In MM force field parameterization, accurate dihedral parameters associated with freely rotating single bonds that link ring systems, such as the PAH-DNA adduct covalent bond, are critical for accurate conformational sampling in MD simulations. Standard parameterization approaches for the CHARMM force field yield parameters that result in MM potential energy surfaces for the adduct covalent bond dihedral angle that are a poor overall fit to their respective sets of QM target data and the parameters do not transfer between structurally different bay and fjord region PAH model systems.

In order to improve the fit to QM target data, CGenFF is used to effectively identify high penalty dihedral parameters that should not be assigned by analogy for the adduct covalent bond. These dihedral parameters are optimized by utilizing CGenFF assigned multiplicities, force constants, and phase constants as initial input for ffTK optimization. The dih_1 dihedral terms for the adduct covalent bond are then further optimized by least squares to fit the dihedral difference potential that results from a MM PES scan of the adduct covalent bond dihedral angle, obtained by setting the force constants for dih_1 to zero. Because the PES scan is conducted over an equispaced partition of the interval (-180°,180°] that includes 0°the resulting matrix $\mathbf{A}$ in the corresponding least squares system is orthogonal and hence is of full rank, allowing force constants to be obtained by simply solving the Normal Equations.

CHAPTER 3

Regularization of Least Squares Problems in CHARMM Parameter Optimization by Truncated Singular Value Decompositions

3.1 Introduction

As described in the previous chapter, parameterization of novel residues for use with molecular mechanics (MM) force fields frequently requires optimization of a subset of parameters that cannot be accurately assigned by analogy.[38–45,55] Optimization of such parameters by least squares fitting of force field terms to quantum mechanical (QM) target data is an effective approach to what is often a challenging and tedious task.[47,48,59,69] Broadly, where we typically require $m > n$, the elements of $\mathbf{A} \in \mathbb{R}^{m \times n}$ are composed of the functional form of the force field, the elements of $\mathbf{x} \in \mathbb{R}^n$ are the unknown force field terms to be optimized, and the elements of $\mathbf{b} \in \mathbb{R}^m$ consist of the QM target data. We then seek a solution $\mathbf{x}_0$ to the matrix equation $\mathbf{Ax} = \mathbf{b}$ that minimizes the 2-norm of the residual $\|\mathbf{r}_0\|_2^2 = \|\mathbf{Ax}_0 - \mathbf{b}\|_2^2$. Noting that in general $\|\mathbf{r}\|_2^2$ is a differentiable function of $\mathbf{x}$, the least squares solution is that for which $\nabla\|\mathbf{r}\|_2^2 = 0$.[60,61]

There exist several numerical approaches to solving least squares inverse problems, but when applied to force field parameter optimization utilizing QM target data, such inverse problems are frequently ill-posed as a result of the matrix $\mathbf{A}$ being ill-conditioned, whereby small perturbations to $\mathbf{A}$ or $\mathbf{b}$ result in very large perturbations of the solution $\mathbf{x}_0$. This in turn can result in unphysical force field terms when the ill-posedness of the underlying least squares problem is not addressed

or not recognized.[47,60,61,69–72]

A well established numerical approach to ill-posed least squares problems is Tikhonov Regularization in standard form[71–74] whereby the ill-conditioned matrix $\mathbf{A}$ is augmented by $\lambda \mathbf{I}_n$ where λ is known as the regularization parameter. This results in a least squares problem of full rank:

$$\min \left\| \begin{bmatrix} \mathbf{A} \\ \lambda \mathbf{I}_n \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\| \tag{3.1}$$

with a unique regularized solution $\mathbf{x}_\lambda$.[71,72] This is similar to the force field parameter optimization approach described by Vanommeslaeghe and MacKerell[47] which specifies bias factors as parameters for regularization in non-standard form. Note that regularized least squares problems not in standard form can be transformed into standard form as described by Elden[70,71] and we will thus work with the standard form (3.1) for simplicity.

Another well established numerical approach to ill-posed least squares problems is the Truncated Singular Value Decomposition (TSVD)[71,72] whereby the ill-conditioned matrix $\mathbf{A}$ is decomposed into a product of matrices: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and $\mathbf{\Sigma}$ is diagonal. The resulting truncated solution $\mathbf{x}_k$ is determined by identifying and discarding small singular values that result in unsatisfactory solutions (i.e. truncating the singular value spectrum). This is similar to the force field parameter optimization approach described by Dasgupta et. al.[69] which specifies a critical condition number as a parameter that drives truncation of the singular value spectrum.

While effectively implemented, these previous approaches to ill-posed least squares problems in MM force field parameter optimization specify a range of regularization and truncation parameters based on the user's experience.[47,69] However, Hansen has shown previously that where an ill-posed least squares problem satisfies the Discrete Picard Condition described below, both the regularization parameter λ and the truncation parameter k can be determined systematically if not rigorously.[71,72] The resulting regularized solution $\mathbf{x}_\lambda$ and the truncated solution $\mathbf{x}_k$ will be similar where the Discrete Picard Condition is satisfied and furthermore, the truncation parameter k can

be used to estimate an effective regularization parameter λ.

Application of the TSVD and Discrete Picard Condition as regularization tools for ill-posed least squares problems was developed rigorously in the Numerical Linear Algebra community. Here we will show how these mathematical tools can be applied to MM force field parameterization in order to study a wide range of chemical problems of interest. While previously developed, the mathematics behind this approach is essential to its application to chemical systems, hence in the sections to follow we will restate Hansen's key results,[71,72] abridging some details and elaborating on others for those interested in force field parameterization. We then demonstrate an effective application to optimization of dihedral parameters for PAH-DNA adducts in the CHARMM force field. These systems pose unique challenges as the torsional potential energy surface (PES) of the freely rotating single bond linking the purine in DNA and the PAH adduct (henceforth adduct covalent bond) is asymmetric and highly dependent upon the PAH structure (i.e. bay vs. fjord) despite identical atomic connectivity as described in the previous chapter. Because the genotoxicity and hence carcinogenic potential of PAH-DNA adducts is a function of geometric conformation, accurate parameterization of the adduct covalent bond is essential to accurate conformational sampling in molecular dynamics simulations of such systems.[12,13,17,27,33–37,53] We note however that this approach is applicable to most all ill-posed least squares problems that arise in force field optimization where the Discrete Picard Condition described below is satisfied, not merely dihedral parameter optimization.

3.2 Ill-Posed Least Squares Problems

3.2.1 Filtering Small Singular Values

The source of ill-posed least squares problems is well illustrated in terms of the singular value decomposition of the matrix **A** in the unconstrained linear least squares problem:

$$\min\|\mathbf{Ax}-\mathbf{b}\|_2 \quad \mathbf{A}\in\mathbb{R}^{m\,x\,n} \quad m>n. \tag{3.2}$$

The matrices $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ are symmetric positive semi-definite and hence each has orthogonal eigenvectors and they share positive eigenvalues. As a result the economy SVD of $\mathbf{A}$ has the form:

$$\mathbf{A}=\mathbf{U\Sigma V}^T \tag{3.3}$$

where $\mathbf{U}=[\mathbf{u}_1,...,\mathbf{u}_n]\in\mathbb{R}^{m\,x\,n}$ with orthonormal column vectors $\{\mathbf{u}_i\}$, $\mathbf{V}=[\mathbf{v}_1,...,\mathbf{v}_n]\in\mathbb{R}^{n\,x\,n}$ with orthonormal column vectors $\{\mathbf{v}_i\}$, and $\mathbf{\Sigma}\in\mathbb{R}^{n\,x\,n}$ is a diagonal matrix with $\mathbf{\Sigma}=\text{diag}[\sigma_1,\sigma_2,...,\sigma_n]$.[60,61,71,72] Where $\text{rank}(\mathbf{A})=r<n$ we have:

$$\sigma_1\geq...\geq\sigma_r>\sigma_{r+1}=...=\sigma_n=0. \tag{3.4}$$

Where we assume $\mathbf{A}$ to have full rank equal to n, we have:

$$\sigma_1\geq\sigma_2\geq...\geq\sigma_n>0 \tag{3.5}$$

and the condition number of the matrix $\mathbf{A}$ is defined as $C=\sigma_1/\sigma_n$, where a large condition number indicates the presence of small elements in the singular value spectrum of $\mathbf{A}$. In terms of the SVD, the matrix equation $\mathbf{Ax}=\mathbf{b}$ has the least squares solution:

$$\mathbf{x}_0=\mathbf{A}^+\mathbf{b}=\mathbf{V\Sigma}^+\mathbf{U}^T\mathbf{b} \tag{3.6}$$

where:

$$\mathbf{\Sigma}^+=\text{diag}\left[\frac{1}{\sigma_1},...,\frac{1}{\sigma_n}\right] \tag{3.7}$$

and the solution can be written as:[71,72]

$$\mathbf{x}_0 = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i. \tag{3.8}$$

From this we see that if $\mathbf{A}$ has very small singular values σ_i, these will cause the elements of the solution $\mathbf{x}_0$ to become large. Consequently, small perturbations in $\mathbf{A}$ and/or $\mathbf{b}$ may result in large perturbations of the solution $\mathbf{x}_0$. Such ill-conditioned matrices are characterized by large condition numbers and are often the source of ill-posed least squares problems in force field parameter optimization. These problems can be addressed by regularization methods that filter out small singular values that have a large impact on the solution. Such methods yield an approximate solution to the ill-posed least squares problem by solving a well-posed problem derived from the original ill-posed problem.[71,72]

The TSVD addresses ill-posed least squares problems by truncating the sum in (3.8) at a truncation parameter $k < n$ thus eliminating the impact of small singular values on the solution:

$$\mathbf{x}_k = \mathbf{A}_k^+ \mathbf{b} = \mathbf{V} \Sigma_k^+ \mathbf{U}^T \mathbf{b} \tag{3.9}$$

where:

$$\Sigma_k^+ = \text{diag}\left[\frac{1}{\sigma_1}, ..., \frac{1}{\sigma_k}, 0, ..., 0\right] \tag{3.10}$$

and similar to (3.8), the truncated solution can be written as:[71,72]

$$\mathbf{x}_k = \sum_{i=1}^{k} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i. \tag{3.11}$$

Tikhonov Regularization in standard form addresses ill-posed least squares problems by examining the quadratically constrained least squares problem (3.1), which has the unique solution:

$$\mathbf{x}_\lambda = \text{argmin}\{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2 \|\mathbf{x}\|_2^2\} \tag{3.12}$$

which can be written in terms of the SVD of $\mathbf{A}$ as:

$$\mathbf{x}_\lambda = \mathbf{A}_\lambda^I \mathbf{b} = \left[\mathbf{A}^T\mathbf{A} + \lambda^2 \mathbf{I}_n\right]^{-1} \mathbf{A}^T\mathbf{b} = \mathbf{V}\boldsymbol{\Sigma}_\lambda^+ \mathbf{U}^T\mathbf{b} \tag{3.13}$$

where:

$$\boldsymbol{\Sigma}_\lambda^+ = \text{diag}\left[\frac{\sigma_1}{\sigma_1^2 + \lambda^2}, ..., \frac{\sigma_n}{\sigma_n^2 + \lambda^2}\right] \tag{3.14}$$

and similar to (3.8) and (3.11) the regularized solution can be written in the form:[71,72]

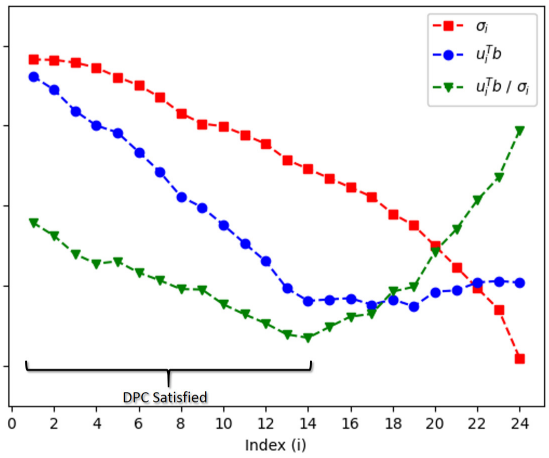
$$\mathbf{x}_\lambda = \sum_{\text{i}=1}^{n} \frac{\sigma_\text{i}^2}{\sigma_\text{i}^2 + \lambda^2} \frac{\mathbf{u}_\text{i}^T\mathbf{b}}{\sigma_\text{i}} \mathbf{v}_\text{i}. \tag{3.15}$$

From this it is apparent that for $\sigma_n \leq \lambda \leq \sigma_1$, the term $\sigma_\text{i}^2/(\sigma_\text{i}^2 + \lambda^2)$ filters out the impact of singular values that are smaller than the regularization parameter λ.

From (3.8), (3.11), and (3.15) it is apparent that the regularized solution $\mathbf{x}_\lambda$ and the truncated solution $\mathbf{x}_k$ will be similar when $\lambda \approx \sigma_k$ as the filter factor $\sigma_\text{i}^2/(\sigma_\text{i}^2 + \lambda^2)$ in (3.15) will dampen the impact of singular values smaller than σ_k on the regularized solution. Indeed Hansen has shown that setting $\lambda \approx (\sigma_k^3\, \sigma_{k+1})^{\frac{1}{4}}$ minimizes the difference between the regularized and truncated solutions while $\lambda \approx (\sigma_k\, \sigma_{k+1})^{\frac{1}{2}}$ minimizes the difference between the corresponding residuals [see Thrm 5.2 Ref 71 for details]. Additionally, the truncated solution $\mathbf{x}_k$ can be calculated as efficiently as the regularized solution $\mathbf{x}_\lambda$. Hence in most cases, the TSVD can be used as a tool to determine the regularization parameter λ or can be used to calculate a regularized solution on its own.[71,72] In the sections to follow, we will examine Hansen's approach to determining the regularization parameter λ and the truncation parameter k in order to obtain satisfactory solutions.

3.2.2 The Discrete Picard Condition

Hansen formulated the Discrete Picard Condition (DPC) to establish a set of conditions under which Tikhonov Regularization in standard form and the TSVD converge to satisfactory solutions of the ill-posed least squares problem at hand. This was motivated by the well established Picard



(3.1) Hypothetical illustration of the Discrete Picard Condition satisfied for $i = 1, ..., 14$. Red squares: singular value spectrum $\{\sigma_i\}$. Blue circles: terms $\{|\mathbf{u}_i^T \mathbf{b}|\}$. Green triangles: coefficients $\{\frac{|\mathbf{u}_i^T \mathbf{b}|}{\sigma_i}\}$.

Condition for Fredholm integral equations of the first kind utilizing the corresponding singular value expansion.[72]

In defining the DPC, it is necessary to examine the coefficient term $\mathbf{u}_i^T \mathbf{b}/\sigma_i$ that appears in the three solutions (3.8), (3.11), and (3.15) described above. Where $\mathbf{A}$ has very small singular values, and the σ_i decay toward zero faster than the corresponding $\mathbf{u}_i^T \mathbf{b}$, our regularization approaches may not be effective at filtering out the impact of small singular values. To quantify this, we can examine the decay of the terms $\mathbf{u}_i^T \mathbf{b}$ relative to the singular values by considering the relationship:

$$\mathbf{u}_i^T \mathbf{b} = \sigma_i^\alpha \quad i = 1, ..., n \tag{3.16}$$

for some $\alpha \geq 0$. Where $\alpha > 1$ and when $\sigma_i < 1$, we see from $\mathbf{u}_i^T \mathbf{b}/\sigma_i = \sigma_i^\alpha/\sigma_i$ that the terms $\mathbf{u}_i^T \mathbf{b}$ decay faster than the corresponding singular values σ_i and where $0 \leq \alpha \leq 1$ the opposite holds. From this Hansen formulates the **Discrete Picard Condition (DPC)**:[72]

In the matrix equation $\mathbf{Ax} = \mathbf{b}$, the unperturbed right hand side $\mathbf{b}$ satisfies the DPC if, for every non-zero singular value, the terms $|\mathbf{u}_i^T \mathbf{b}|$ decay to zero faster on average (not necessarily monotonically) than the singular values σ_i (Figure 3.1).

Hansen has shown that when the DPC is satisfied, error bounds on the regularized and truncated solutions $\mathbf{x}_\lambda$ and $\mathbf{x}_k$ relative to the solution $\mathbf{x}_0$ can be established [Thrm 3.1 Ref 72]:

$$\frac{\|\mathbf{x}_0 - \mathbf{x}_k\|_2}{\|\mathbf{x}_0\|_2} \leq \begin{cases} \sqrt{n} & \text{if} \quad 0 \leq \alpha \leq 1 \\ (\frac{\sigma_{k+1}}{\sigma_1})^{\alpha-1}\sqrt{n} & \text{if} \quad 1 \leq \alpha \end{cases} \tag{3.17}$$

$$\frac{\|\mathbf{x}_0 - \mathbf{x}_\lambda\|_2}{\|\mathbf{x}_0\|_2} \leq \begin{cases} \sqrt{n} & \text{if} \quad 0 \leq \alpha \leq 1 \\ (\frac{\lambda}{\sigma_1})^{\alpha-1}\sqrt{n} & \text{if} \quad 1 \leq \alpha < 3 \\ (\frac{\lambda}{\sigma_1})^{2}\sqrt{n} & \text{if} \quad 3 \leq \alpha \end{cases} \tag{3.18}$$

These indicate that when the DPC is satisfied, and for small σ_k and λ relative to σ_1, the regularized and truncated solutions $\mathbf{x}_\lambda$ and $\mathbf{x}_k$ approximate the solution $\mathbf{x}_0$ and the error bounds improve with faster decay of the terms $\mathbf{u}_1^T\mathbf{b}$ relative to the singular values (i.e. for larger $\alpha > 1$). Note that if there are errors present such as a perturbation $\mathbf{b}+\mathbf{e}$ to the right hand side of the matrix equation, the DPC must be satisfied for the *unperturbed* right hand side for the regularized and truncated solutions to approximate $\mathbf{x}_0$. Additionally, Hansen has shown that when the DPC is satisfied and $\sigma_{k+1} << \sigma_1$, we can choose $\lambda \in [\sigma_{k+1}, \sigma_k]$ for which the regularized and truncated solutions are similar. As above, for larger $\alpha > 1$ the regularized and truncated solutions become yet closer [see Thrm 3.2 Ref 72 for details].

3.2.3 Perturbation Theory

Errors in least squares problems are often isolated to the right hand side of the matrix equation $\mathbf{Ax} = \mathbf{b}$.[71,72] Such is largely the case when using QM target data to optimize force field parameters where the matrix $\mathbf{A}$ consists of the mathematical terms of the MM force field at specified geometries of the molecular system being parameterized, and the right hand side consists of the corresponding QM energies. Computational errors that arise from QM calculations at a given level of theory then result in perturbations $\mathbf{b}+\mathbf{e}$ of the right hand side. Although errors may occur in

the mathematical terms in the elements of the matrix $\mathbf{A}$, we seek to follow Hansen's treatment of Tikhonov Regularization and the TSVD and consider only perturbations $\mathbf{b}+\mathbf{e}$ of the right hand side going forward. In order to proceed, we define several quantities:

$$\mathbf{b}_0 = \mathbf{A}\mathbf{x}_0 \qquad \mathbf{b}_k = \mathbf{A}\mathbf{x}_k \qquad \mathbf{b}_\lambda = \mathbf{A}\mathbf{x}_\lambda, \tag{3.19}$$

$$\begin{aligned} \mathbf{x}_0^{(e)} &= \sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\ \mathbf{x}_k^{(e)} &= \sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\ \mathbf{x}_\lambda^{(e)} &= \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i}, \end{aligned} \tag{3.20}$$

$$\begin{aligned} \tilde{\mathbf{x}}_0 &= \sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T(\mathbf{b}+\mathbf{e})}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\ &= \mathbf{x}_0 + \mathbf{x}_0^{(e)} \end{aligned} \tag{3.21}$$

$$\begin{aligned} \tilde{\mathbf{x}}_k &= \sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T(\mathbf{b}+\mathbf{e})}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\ &= \mathbf{x}_k + \mathbf{x}_k^{(e)}. \end{aligned} \tag{3.22}$$

$$\begin{aligned} \tilde{\mathbf{x}}_\lambda &= \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T(\mathbf{b}+\mathbf{e})}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\ &= \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\ &= \mathbf{x}_\lambda + \mathbf{x}_\lambda^{(e)} \end{aligned} \tag{3.23}$$

Note that the solutions (3.21), (3.22), and (3.23) resulting from the perturbed right hand side $\mathbf{b}+\mathbf{e}$ are the analogs of the solutions (3.8), (3.11), and (3.15) resulting from the unperturbed right hand side.

Hansen has shown [Thrm 4.1 Ref 72] that for $\lambda \in [\sigma_n, \sigma_1]$, the regularization and truncation parameters λ and k can be chosen such that the corresponding solutions $\tilde{\mathbf{x}}_k$ and $\tilde{\mathbf{x}}_\lambda$ are not largely

impacted by the perturbation to the right hand side of the matrix equation, as seen in the following error bounds:

$$\frac{\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|_2}{\|\mathbf{x}_k\|_2} \leq \frac{\sigma_1}{\sigma_k} \frac{\|\mathbf{e}\|_2}{\|\mathbf{b}_k\|_2} \tag{3.24}$$

$$\frac{\|\mathbf{x}_\lambda - \tilde{\mathbf{x}}_\lambda\|_2}{\|\mathbf{x}_\lambda\|_2} \leq \frac{\sigma_1}{2\lambda} \frac{\|\mathbf{e}\|_2}{\|\mathbf{b}_\lambda\|_2}. \tag{3.25}$$

Note that when $\lambda \approx \sigma_k$ the error bounds (3.24) and (3.25) will be similar.

Where the DPC is satisfied, there is a balance to be struck between the error bounds (3.17),(3.18) and the perturbation bounds (3.24),(3.25) when one selects the regularization and truncation parameters. Because (3.17) and (3.18) respectively contain the terms σ_{k+1}/σ_1 and λ/σ_1, the truncated and regularized error bounds will shrink for smaller λ and correspondingly larger k (i.e. smaller σ_k and σ_{k+1}), but the perturbation bounds will grow since (3.24) and (3.25) respectively contain the terms σ_1/σ_k and $\sigma_1/(2\lambda)$, resulting in $\tilde{\mathbf{x}}_\lambda$ and $\tilde{\mathbf{x}}_k$ being more sensitive to perturbations. Where larger λ and smaller k result in smaller perturbation bounds, the error bounds become larger depending upon the rate of decay of the terms $\mathbf{u}_i^T\mathbf{b}$ relative to the singular values (i.e. depending on the value of α).

3.3 Determining Regularization and Truncation Parameters

3.3.1 Analysis of Regularized and Truncated Solutions

It is a standard practice to examine the least squares solutions produced by a given numerical method by plotting the norm of said solutions against the norm of the corresponding residuals.[71,72,75] When examining our regularized and truncated solutions corresponding to the perturbed right hand side $\mathbf{b}+\mathbf{e}$, we will observe a distinct corner in the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ as a function of the regularization parameter λ and in the plot of $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ as a discrete function of the truncation parameter k, that demarcates regions from which λ and k should be selected.

As noted by Hansen, the discussion to follow is not strictly rigorous, but demonstrates a working application of the results outlined thus far. Additional details can be found in Hansen's works on the TSVD and regularization.[71,72]

To illustrate this cornering behavior, the components of the truncated and regularized residuals $\mathbf{r}_k$ and $\mathbf{r}_\lambda$ from the column space of $\mathbf{A}$, corresponding to the unperturbed right hand side are defined as:

$$\mathbf{r}_k = \mathbf{b}_0 - \mathbf{A}\mathbf{x}_k = \mathbf{A}\mathbf{x}_0 - \mathbf{A}\mathbf{x}_k = \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{i=k+1}^{n} \mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} \tag{3.26}$$

$$\mathbf{r}_\lambda = \mathbf{b}_0 - \mathbf{A}\mathbf{x}_\lambda = \mathbf{A}\mathbf{x}_0 - \mathbf{A}\mathbf{x}_\lambda = \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} \tag{3.27}$$

The truncated and regularized residuals $\tilde{\mathbf{r}}_k$ and $\tilde{\mathbf{r}}_\lambda$ corresponding to the perturbed right hand side are defined as:

$$\begin{aligned}
\tilde{\mathbf{r}}_k &= \mathbf{A}\tilde{\mathbf{x}}_0 - \mathbf{A}\tilde{\mathbf{x}}_k \\
&= \mathbf{A}(\mathbf{x}_0 + \mathbf{x}_0^{(e)}) - \mathbf{A}(\mathbf{x}_k + \mathbf{x}_k^{(e)}) \\
&= \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\
&= \sum_{i=k+1}^{n} \mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} + \sum_{i=k+1}^{n} \mathbf{u}_\mathrm{i}^T\mathbf{e}\mathbf{u}_\mathrm{i} \\
&= \mathbf{r}_k + \mathbf{r}_k^{(e)}
\end{aligned} \tag{3.28}$$

$$\begin{aligned}
\tilde{\mathbf{r}}_\lambda &= \mathbf{A}\tilde{\mathbf{x}}_0 - \mathbf{A}\tilde{\mathbf{x}}_\lambda \\
&= \mathbf{A}(\mathbf{x}_0 + \mathbf{x}_0^{(e)}) - \mathbf{A}(\mathbf{x}_\lambda + \mathbf{x}_\lambda^{(e)}) \\
&= \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\
&= \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} + \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{u}_\mathrm{i}^T\mathbf{e}\mathbf{u}_\mathrm{i} \\
&= \mathbf{r}_\lambda + \mathbf{r}_\lambda^{(e)}
\end{aligned} \tag{3.29}$$

For illustrative purposes, we begin by independently examining the curves $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$ and $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$ as functions of the regularization parameter λ.

Examining $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$, the norm of the solution to the unperturbed problem plotted against the norm of the corresponding residual, it is known that $\|\mathbf{x}_\lambda\|_2$ is a decreasing function of $\|\mathbf{r}_\lambda\|_2$ and we have that as $\lambda \to 0$ the filter factor $\sigma_i^2/(\sigma_i^2+\lambda^2) \to 1$ resulting in $\mathbf{x}_\lambda \to \mathbf{x}_0$ and thus $\mathbf{r}_\lambda \to 0$.[72] Hence, for values of λ much smaller than the smallest singular value σ_n, we can make the approximations: $\sigma_i^2/(\sigma_i^2+\lambda^2) \approx 1$ and $\lambda^2/(\sigma_i^2+\lambda^2) \approx \lambda^2/\sigma_i^2$, resulting in $\mathbf{x}_\lambda \approx \mathbf{x}_0$ and:

$$\mathbf{r}_\lambda = \sum_{i=1}^{n} \frac{\lambda^2}{\sigma_i^2+\lambda^2}\mathbf{u}_i^T\mathbf{b}\mathbf{u}_i \approx \sum_{i=1}^{n} \frac{\lambda^2}{\sigma_i^2}\mathbf{u}_i^T\mathbf{b}\mathbf{u}_i. \tag{3.30}$$

Hence we have $\|\mathbf{x}_\lambda\|_2 \approx \|\mathbf{x}_0\|_2$ and since $\mathbf{b}_0 = \sum_{i=1}^{n} \mathbf{u}_i^T\mathbf{b}\mathbf{u}_i$ we have:

$$\begin{aligned} \|\mathbf{r}_\lambda\|_2 \approx \lambda^2 \sqrt{\sum_{i=1}^{n}\left(\frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i^2}\right)^2} &\le \lambda^2 \sqrt{\sum_{i=1}^{n}\left(\frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_n^2}\right)^2} \\ &\le \left(\frac{\lambda}{\sigma_n}\right)^2 \|\mathbf{b}_0\|_2. \end{aligned} \tag{3.31}$$

Thus for these small values of the regularization parameter λ, we have that $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2) \approx (\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_0\|_2)$ and the curve of the norm of the solution to the unperturbed problem plotted against the norm of the corresponding residual traces a nearly horizontal line for small values of $\|\mathbf{r}_\lambda\|_2$. As λ becomes larger, the regularization filter factor $\sigma_i^2/(\sigma_i^2+\lambda^2) < 1$ resulting in the norm of the solution $\|\mathbf{x}_\lambda\|_2$ becoming smaller than $\|\mathbf{x}_0\|_2$ and the corresponding residual $\|\mathbf{r}_\lambda\|_2$ becoming larger. Noting that as the regularization parameter $\lambda \to \infty$ the filter factor $\sigma_i^2/(\sigma_i^2+\lambda^2) \to 0$, resulting in $\mathbf{x}_\lambda \to 0$ and $\mathbf{r}_\lambda \to \mathbf{b}_0$, we have that the curve $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$ veers downwards toward the horizontal axis and the point $\|\mathbf{b}_0\|_2$.[72]

Examining $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$, the norm of the perturbation of the solution plotted against the norm of the perturbation of the residual we assume that for each i the terms $\mathbf{u}_i^T\mathbf{e}$ seen in $\mathbf{x}_0^{(e)}$, $\mathbf{x}_\lambda^{(e)}$, and $\mathbf{r}_\lambda^{(e)}$ (3.20) are all of approximately the same magnitude ε_0 (i.e. the DPC is not satisfied for these terms). As above, we note that as the regularization parameter $\lambda \to 0$, $\mathbf{x}_\lambda^{(e)} \to \mathbf{x}_0^{(e)}$ and

$\mathbf{r}_\lambda^{(e)} \to 0$, and again for very small $\lambda << \sigma_n$ we have $\mathbf{x}_\lambda^{(e)} \approx \mathbf{x}_0^{(e)}$. With the additional assumption that $|\mathbf{u}_\mathrm{i}^T \mathbf{e}| \approx \varepsilon_0$ we have that:

$$\begin{aligned}
\mathbf{x}_0^{(e)} &\approx \varepsilon_0 \sum_{\mathrm{i}=1}^{n} \frac{1}{\sigma_\mathrm{i}} \mathbf{v}_\mathrm{i} \leq \varepsilon_0 \sum_{\mathrm{i}=1}^{n} \frac{1}{\sigma_n} \mathbf{v}_\mathrm{i}, \\
\mathbf{x}_\lambda^{(e)} &\approx \varepsilon_0 \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}}{\sigma_\mathrm{i}^2 + \lambda^2} \mathbf{v}_\mathrm{i}, \\
\mathbf{r}_\lambda^{(e)} &\approx \varepsilon_0 \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2 + \lambda^2} \mathbf{u}_\mathrm{i}.
\end{aligned} \tag{3.32}$$

We have then that $\|\mathbf{x}_\lambda^{(e)}\|_2 \approx \|\mathbf{x}_0^{(e)}\|_2$ where $\varepsilon_0/\sigma_n \leq \|\mathbf{x}_0^{(e)}\|_2 \leq \sqrt{n}\,\varepsilon_0/\sigma_n$. Hence for these small λ we have that $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2) \approx (\|\mathbf{r}_\lambda^{(e)}\|_2, \frac{\sqrt{n}\,\varepsilon_0}{\sigma_n})$ and the curve of the norm of the perturbation of the solution plotted against the norm of the perturbation of the residual traces a nearly horizontal line for small values of $\|\mathbf{r}_\lambda^{(e)}\|_2$. As λ becomes larger than the smallest singular value σ_n we have that $\mathbf{x}_\lambda^{(e)}$ in (3.32) is dominated by the terms for which $\lambda \approx \sigma_\mathrm{i}$ where we can make the approximation: $\sigma_\mathrm{i}/(\sigma_\mathrm{i}^2 + \lambda^2) \approx 1/(2\lambda)$. Supposing there are p such terms, we have that $\|\mathbf{x}_\lambda^{(e)}\|_2 \approx p\varepsilon_0/(2\lambda)$ and hence as $\lambda \to \infty$ we have that $\|\mathbf{x}_\lambda^{(e)}\|_2 \to 0$. Since we also have that $\varepsilon_0 \leq \|\mathbf{r}_\lambda^{(e)}\|_2 \leq \sqrt{n}\,\varepsilon_0$, the curve $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$ decreases rapidly toward the horizontal axis and toward the point $\sqrt{n}\,\varepsilon_0$.[72]

Note that Hansen has shown where the DPC is satisfied and where k is large, we can choose the regularization parameter $\lambda \in [\sigma_{k+1}, \sigma_k]$ such that the analogous plots of the norm of the truncated solution to the unperturbed problem against the norm of the corresponding residual along with the accompanying perturbations: $(\|\mathbf{r}_k\|_2, \|\mathbf{x}_k\|_2)$ and $(\|\mathbf{r}_k^{(e)}\|_2, \|\mathbf{x}_k^{(e)}\|_2)$, closely approximate those of the regularized solution and residual and accompanying perturbations: $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$ and $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$, with deviations occurring where the DPC is not satisfied.[72] Hence, the features discussed above for regularized curves are also observed for the truncated plots.

3.3.2 Regularization and Truncation Parameters Based on the L-Curve

We can organize the results outlined above into the following collection of conditions for the perturbed right hand side $\mathbf{b} + \mathbf{e}$ of the matrix equation [Assumption 5.1 Ref 72]:

1. The unperturbed right hand side $\mathbf{b}$ satisfies the DPC

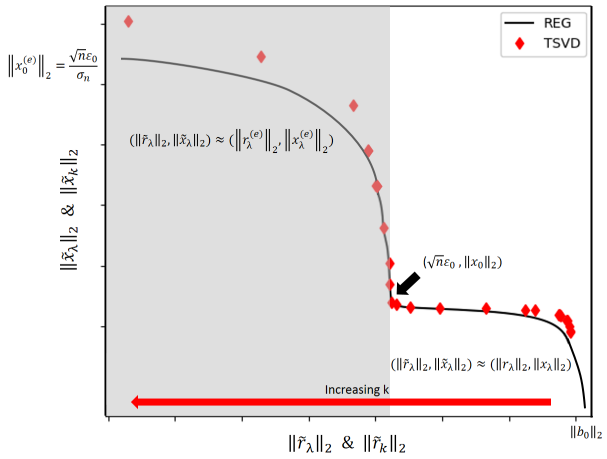
2. $\|\mathbf{e}\|_2 < \|\mathbf{b}_0\|_2$ where $\mathbf{b}_0 = \mathbf{A}\mathbf{x}_0$

3. The perturbation $\mathbf{e}$ is a random vector of zero mean and covariance matrix $\varepsilon_0^2 I$

As we have seen above, the first and second assumptions are required for $\tilde{\mathbf{x}}_k$ and $\tilde{\mathbf{x}}_\lambda$ to produce reasonable approximations of $\mathbf{x}_0$. The third assumption ensures that the errors in the perturbation are uncorrelated and results in the DPC not being satisfied for the perturbation $\mathbf{e}$.

We now examine the the norm of the solution to the perturbed problem plotted against the norm of the corresponding residual: $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ as a function of the regularization parameter λ, applying the analysis utilized to examine the graphs of $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$ and $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$ above and recalling that $\tilde{\mathbf{r}}_\lambda = \mathbf{r}_\lambda + \mathbf{r}_\lambda^{(e)}$ and $\tilde{\mathbf{x}}_\lambda = \mathbf{x}_\lambda + \mathbf{x}_\lambda^{(e)}$. Again, $\|\tilde{\mathbf{x}}_\lambda\|_2$ is a decreasing function of $\|\tilde{\mathbf{r}}_\lambda\|_2$. Where λ is small resulting in the perturbation $\mathbf{x}_\lambda^{(e)}$ dominating the solution $\tilde{\mathbf{x}}_\lambda$ and the DPC not being satisfied, the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ resembles that of the perturbation $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$, running nearly horizontal at $\|\tilde{\mathbf{x}}_\lambda\|_2 \approx \|\mathbf{x}_0^{(e)}\|_2 \approx \sqrt{n}\,\varepsilon_0/\sigma_n$ for correspondingly small values of of the residual $\|\tilde{\mathbf{r}}_\lambda\|_2$, followed by a rapid decrease toward the horizontal axis at the point $\sqrt{n}\,\varepsilon_0$. As the regularization parameter λ grows, the solution to the unperturbed problem $\mathbf{x}_\lambda$ begins to dominate the solution $\tilde{\mathbf{x}}_\lambda$ and the DPC is satisfied, resulting in the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ of the perturbed problem resembling that of unperturbed problem $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$, again running nearly horizontal at $\|\tilde{\mathbf{x}}_\lambda\|_2 \approx \|\mathbf{x}_0\|_2$, then gradually curving toward the horizontal axis at the point $\|\mathbf{b}_0\|_2$ as λ grows large relative to σ_n. As above, the analogous plot of the norm of the truncated solution to the perturbed problem and corresponding residual norm: $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ closely approximates the curve of the Tikhonov regularized solution $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ where the DPC is satisfied.[72]

We can thus observe a corner in the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (which we will jointly refer to as the L-curve) near the point $(\sqrt{n}\,\varepsilon_0, \|\mathbf{x}_0\|_2)$ where the regularized and truncated solutions $\tilde{\mathbf{x}}_\lambda$ and $\tilde{\mathbf{x}}_k$ to the perturbed problem are dominated by the perturbations $\mathbf{x}_\lambda^{(e)}$ and $\mathbf{x}_k^{(e)}$ to the left of the corner and dominated by the solutions $\mathbf{x}_\lambda$ and $\mathbf{x}_k$ to the unperturbed problem to the right of the corner (Figure 3.2). As described by Hansen, the regularized and truncated solutions



(3.2) Hypothetical illustration of a corner in the L-curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds) as functions of λ and k. In the shaded region to the left of the corner, $\mathbf{x}_\lambda^{(e)}$ and $\mathbf{x}_k^{(e)}$ dominate the solution, resulting in $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2) \approx (\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$. In the unshaded region to the right of the corner, $\mathbf{x}_\lambda$ and $\mathbf{x}_k$ dominate the solution, resulting in $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2) \approx (\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$. Regularization and truncation parameters and the corresponding solutions should be selected from the unshaded region and where the DPC is satisfied.

are similar and best approximate $\mathbf{x}_0$ to the right of this corner, and the largest possible value of the truncation parameter k for which the DPC is satisfied for the *perturbed* terms $\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})$ should be chosen. Additionally, the singular values should not be truncated between multiple or nearly multiple (i.e. repeated) singular values. We then have for $\lambda \in [\mathbf{r}_{k+1}, \mathbf{r}_k]$ as described above, the regularized and truncated solutions will be reasonable solutions,[76–78] satisfying: $\|\tilde{\mathbf{x}}_\lambda\|_2 \approx \|\tilde{\mathbf{x}}_k\|_2 \approx \|\mathbf{x}_0\|_2$ and $\|\tilde{\mathbf{r}}_\lambda\|_2 \approx \|\tilde{\mathbf{r}}_k\|_2 \approx \|\mathbf{e}\|_2$ with $\tilde{\mathbf{x}}_\lambda, \tilde{\mathbf{x}}_k \to \mathbf{x}_0$ as $\mathbf{e} \to \mathbf{0}$.[72]

3.3.3 Regularization and Truncation Parameters Based on Numerical Rank

While the analyses above cover selection of regularization and truncation parameters in general, a special and convenient case arises for matrices $\mathbf{A}$ that have well-determined numerical rank. The rank of a matrix $\mathbf{A}$ is the dimension of its column space (i.e. the number of linearly independent column vectors in $\mathbf{A}$), and is revealed by the number of non-zero singular values in the singular value spectrum of $\mathbf{A}$. With ill-conditioned matrices such as those that often occur in MM force field parameter optimization, it is uncommon to find identically zero singular values, but it is very

common to encounter numerically small singular values as discussed above.[71]

When considering the singular value spectrum $\sigma_1 > ... > \sigma_k > \sigma_{k+1} > ... > \sigma_n$ we can examine the relative gap $\omega_k = \sigma_{k+1}/\sigma_k$ between neighboring singular values. We can then define ill-conditioned matrices with well-determined numerical rank k as those that have a large, well-defined gap in the singular value spectrum between σ_k and σ_{k+1}, such that the singular values $\sigma_{k+1}, ..., \sigma_n$ are effectively zero in numerical applications.[71] This is characterized by a relative gap ω_k that is markedly smaller than the other relative gaps in the singular value spectrum. As shown by Hansen, such a well-defined gap can be used to select the truncation parameter k (where the DPC should also be satisfied for the first k singular values) without having to examine the L-curve $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$, yielding the same results as those discussed above. The regularization parameter can then be determined, where λ should be chosen as close to σ_k as possible following the analyses above.[71,72] In the case of ill-conditioned matrices $\mathbf{A}$ where the singular value spectrum decays without a well-defined gap, $\mathbf{A}$ is considered to have ill-determined numerical rank, and we instead have to examine the L-curve $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ as described above.

Although the term "well-defined gap" does not strike one as a rigorous definition, we will see in applications to optimization of dihedral force field parameters below that the relative gap ω_k can differ by several fold as compared to the average relative gap in the system's singular value spectrum, demarcating a numerically well-defined gap and corresponding numerical rank that allows for specification of the truncation parameter k. We refer the reader to Hansen's work on the TSVD and standard texts on numerical linear algebra for additional details on numerical rank and the accompanying perturbation theory.[60,61,71,72]

Note that selection of the truncation parameter either by identifying the corner in the L-curve $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ or by identifying a well-defined gap in the singular value spectrum *results in* the condition number $C = \sigma_1/\sigma_k$ of the matrix $\mathbf{A}$ as a function of the truncation parameter k. If instead the condition number is specified as a parameter that dictates the singular values that are to be discarded when solving ill-posed least squares problems by the TSVD, one runs the risk of the solution $\tilde{\mathbf{x}}_k$ falling in the region for which the DPC is not satisfied for the given problem, thus

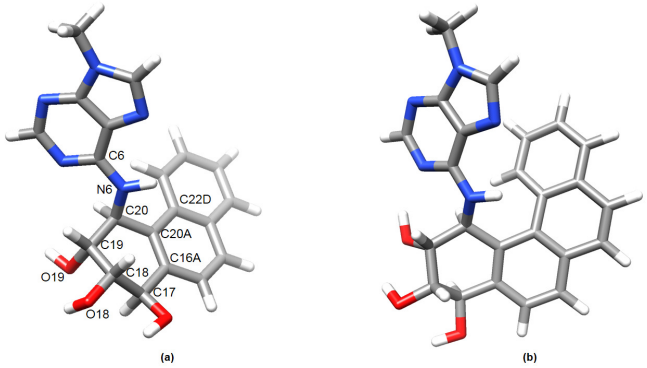
being influenced by the perturbation $\mathbf{x}_k^{(e)}$.

3.4 Dihedral Parameterization of Syn-Glycosidic Model Systems Utilizing Variable Phase

3.4.1 Model Systems

The results outlined above motivate a useful and practical application to ill-posed least squares problems that arise in MM force field parameter optimization. Here we apply the TSVD approach to select truncation and regularization parameters and simultaneously optimize multiple dihedral force field terms for the syn-glycosidic bay region PHE and fjord region B[c]P PAH-DNA adduct model systems described in the previous chapter. We begin with the dihedral parameters dih_1 and dih_2 [Figure 3.3(a): C6-N6-C20-C20a and C6-N6-C20-C19 respectively] that characterize the torsional energy landscape of the adduct covalent bond. As described in the previous chapter, the remainder of the model systems were parameterized using low penalty CGenFF / ParamChem.com (ver 2.2)[40–42] analogy assignments as well as VMD-ffTK[44,62] optimized parameters for those that resulted in high CGenFF penalties. Note that CGenFF / ParamChem.com assigned dihedral parameters for dih_1 and dih_2 were the highest penalty parameters in our model systems (75 and 46.5 respectively) highlighting the need for focused optimization of these parameters as well as the effectiveness of CGenFF / ParamChem.com penalty scoring.

As described in the previous chapter, relaxed QM torsion scans of the adduct covalent bond driven in 10° increments by $\phi_{dih_1} \in$ (-180°,180°] were conducted at the MP2/6-31G(d) level of theory utilizing the Gaussian 16[52] software package for both the PHE and B[c]P model systems (respectively Figure 3.4(c)(d) and Figure 3.5(c)(d) black triangles). Note this results in $m = 36$ discrete scan points $\{(\phi_{dih_1,\mathrm{i}}, \phi_{dih_2,\mathrm{i}}) | \mathrm{i} = 1,...,m\}$ where we plot the respective PESs using the driving geometric parameter ϕ_{dih_1}. An analogous relaxed MM PES scan was conducted with the dihedral force constants for dih_1 and dih_2 set to zero utilizing NAMD[56] and conjugate gradient



(3.3) Model systems:(a) bay region PHE model system, dihedral parameter dih_1: C6-N6-C20-C20a and dihedral parameter dih_2: C6-N6-C20-C19 (b) fjord region B[c]P model system

minimization. Where $\{E_{\mathrm{i}}^{QM}|\,\mathrm{i}=1,...,m\}$ and $\{E_{\mathrm{i}}^{MM_{k_{dih1},k_{dih2}=0}}|\,\mathrm{i}=1,...,m\}$ are respectively the QM and MM energies resulting from the corresponding relaxed PES scans, the discrete difference potential $E^{diff}=\{E_{\mathrm{i}}|\,\mathrm{i}=1,...,m\}$ where $E_{\mathrm{i}}=E_{\mathrm{i}}^{QM}-E_{\mathrm{i}}^{MM_{k_{dih1},k_{dih2}=0}}$ elucidates the form of the dihedral potential that the sum of the dih_1 and dih_2 dihedral force field terms must fit in order for the complete MM PES to accurately model the QM PES.

3.4.2 Inverse Problem with Well-Determined Numerical Rank

In the previous chapter, we have shown the efficacy of utilizing asymmetric dihedral potentials to parameterize dih_1 in our model systems, hence we simultaneously optimize dihedral terms for dih_1 and dih_2 by respectively calculating the coefficients a_{j_1}, b_{j_1} and a_{j_2}, b_{j_2} that achieve a least squares fit of the truncated Fourier series:

$$\begin{aligned} E_{\phi_{dih_1}}+E_{\phi_{dih_2}} = &\sum_{\mathrm{j}_1\in M_1}[a_{\mathrm{j}_1}\cos(\mathrm{j}_1\phi_{dih_1})+b_{\mathrm{j}_1}\sin(\mathrm{j}_1\phi_{dih_1})]+ \\ &\sum_{\mathrm{j}_2\in M_2}[a_{\mathrm{j}_2}\cos(\mathrm{j}_2\phi_{dih_2})+b_{\mathrm{j}_2}\sin(\mathrm{j}_2\phi_{dih_2})] \end{aligned} \tag{3.33}$$

where $M_1, M_2 \subseteq \{1,2,3,4,5,6\}$ are the multiplicities of the dihedral terms. Optimized dihedral

terms are then transformed into the CHARMM requisite dihedral format:

$$E_{\phi_{dih_1}} + E_{\phi_{dih_2}} = \sum_{\mathrm{j}_1 \in M_1} k_{\mathrm{j}_1} \left[1 + \cos(\mathrm{j}_1 \phi_{dih_1} - \delta_{\mathrm{j}_1})\right] + \sum_{\mathrm{j}_2 \in M_2} k_{\mathrm{j}_2} \left[1 + \cos(\mathrm{j}_2 \phi_{dih_2} - \delta_{\mathrm{j}_2})\right] \tag{3.34}$$

using:

$$k_l = \sqrt{a_l^2 + b_l^2} \tag{3.35}$$

$$\delta_l = \mathrm{Arg}(a_l + \boldsymbol{i} b_l) \in (-\pi, \pi] \tag{3.36}$$

where $l = \mathrm{j}_1 \,\mathrm{or}\, \mathrm{j}_2$. Note above that $\boldsymbol{i} = \sqrt{-1}$ where as "i" is an index.

Where $\{(\phi_{dih_1,\mathrm{i}}, \phi_{dih_2,\mathrm{i}}) | \, \mathrm{i} = 1, ..., m\}$ are the PES scan points described above, n_1 and n_2 are the largest multiplicities of the dih_1 and dih_2 dihedral terms respectively (we presume $\mathrm{j}_1 = 1, ..., n_1$ and $\mathrm{j}_2 = 1, ..., n_2$ for simplicity), and where we treat the right hand side of the matrix equation as a perturbation in order to apply the results outlined in the previous sections; the resulting matrix equation $\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{e}$ where $\mathbf{A} \in \mathbb{R}^{m \,\mathrm{x}\, 2(n_1+n_2)}$ and $\mathbf{b} + \mathbf{e} \in \mathbb{R}^m$ have elements of the form:

$$A_{\mathrm{i},2\mathrm{j}-1} = \cos(\mathrm{j}\phi_{dih_1,\mathrm{i}}) - \frac{1}{m}\sum_{\mathrm{i}=1}^{m} \cos(\mathrm{j}\phi_{dih_1,\mathrm{i}}) \tag{3.37}$$

$$A_{\mathrm{i},2\mathrm{j}} = \sin(\mathrm{j}\phi_{dih_1,\mathrm{i}}) - \frac{1}{m}\sum_{\mathrm{i}=1}^{m} \sin(\mathrm{j}\phi_{dih_1,\mathrm{i}}) \tag{3.38}$$

for $\mathrm{i} = 1, ..., m$ and $\mathrm{j} = 1, ..., n_1$

$$A_{\mathrm{i},2\mathrm{j}-1} = \cos((\mathrm{j} - n_1)\phi_{dih_2,\mathrm{i}}) - \frac{1}{m}\sum_{\mathrm{i}=1}^{m} \cos((\mathrm{j} - n_1)\phi_{dih_2,\mathrm{i}}) \tag{3.39}$$

$$A_{\mathrm{i},2\mathrm{j}} = \sin((\mathrm{j} - n_1)\phi_{dih_2,\mathrm{i}}) - \frac{1}{m}\sum_{\mathrm{i}=1}^{m} \sin((\mathrm{j} - n_1)\phi_{dih_2,\mathrm{i}}) \tag{3.40}$$

for $\mathrm{i} = 1, ..., m$ and $\mathrm{j} = n_1 + 1, ..., n_1 + n_2$

$$(b_\mathrm{i} + e_\mathrm{i}) = E_\mathrm{i} - \frac{1}{m}\sum_{\mathrm{i}=1}^{m} E_\mathrm{i} \tag{3.41}$$

for $\mathrm{i} = 1, ..., m$.

Note that the respective data sets are shifted so that their averages are zero and that the elements of **A** can be adjusted as needed to suit the desired multiplicities of the dihedral terms being optimized. The unknown vector $\mathbf{x} \in \mathbb{R}^{2(n_1+n_2)}$ has elements consisting of the unknown Fourier coefficients from (3.33) in the form:

$$\mathbf{x}_{2\mathrm{j}-1} = a_{\mathrm{j}_1} \quad \text{and} \quad \mathbf{x}_{2\mathrm{j}} = b_{\mathrm{j}_1} \tag{3.42}$$

for $\mathrm{j} = 1, ..., n_1$ and where $\mathrm{j}_1 = \mathrm{j}$ and,

$$\mathbf{x}_{2\mathrm{j}-1} = a_{\mathrm{j}_2} \quad \text{and} \quad \mathbf{x}_{2\mathrm{j}} = b_{\mathrm{j}_2} \tag{3.43}$$

for $\mathrm{j} = n_1 + 1, ..., n_1 + n_2$ and where $\mathrm{j}_2 = \mathrm{j} - n_1$.

We obtain optimized Fourier coefficients for (3.33) and in turn optimized dihedral force and phase constants for (3.34) from the least squares solution to the inverse problem.

It is well understood that it is an established best practice to utilize even functions with multiplicities appropriate to the symmetry of the molecular system at hand in order to optimize parameters that are transferable among systems with similar atomic connectivity.[47,55,58] However, where we seek to optimize custom dihedral terms for bay and fjord region PAH-DNA adduct systems that are only meant for use in stereochemically and structurally analogous systems, and where we seek to demonstrate the efficacy of the TSVD approach, dih_1 and dih_2 are each parameterized by a six term series with variable phase. In each case, the singular values σ_i and terms $|\mathbf{u}_\mathrm{i}^T(\mathbf{b}+\mathbf{e})|/\sigma_\mathrm{i}$ were examined for regions over which the DPC is satisfied and for well-defined gaps in the singular value spectrum (Figures 3.4(a) and 3.5(a)). In both cases a well defined gap in the singular value spectrum is observed between σ_{12} and σ_{13}, coinciding with the indices over which the DPC is satisfied in practice. Note also that the singluar values $\sigma_1, ..., \sigma_{12}$ are nearly multiple and the singular value spectrum should not be truncated between nearly multiple singular values. Relative gaps of $\omega_{k=12}(PHE) = 0.1059$ and $\omega_{k=12}(B[c]P) = 0.1182$ are observed where the average relative gaps in each system's singular value spectrum are: $\overline{\omega}(PHE) = 0.7788$ and $\overline{\omega}(B[c]P) = 0.8009$. Additionally, where we treat the right hand side of the matrix equation as described above, we observe

a corner in the log scale graph of the L-curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ and $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ that indicates the truncation parameter should be $k = 12$.

Utilizing these observations, we obtain the TSVD solutions $\tilde{\mathbf{x}}_{k=12}$(PHE) and $\tilde{\mathbf{x}}_{k=12}$(B[c]P) and using Hansen's estimate $\lambda = (\sigma_k \sigma_{k+1})^{\frac{1}{2}}$ we obtain regularized solutions $\tilde{\mathbf{x}}_{\lambda=1.8936}$(PHE) and $\tilde{\mathbf{x}}_{\lambda=1.8800}$(B[c]P). The resulting (and very similar) CHARMM compatible dihedral terms are listed in Tables 3.1 and 3.2.

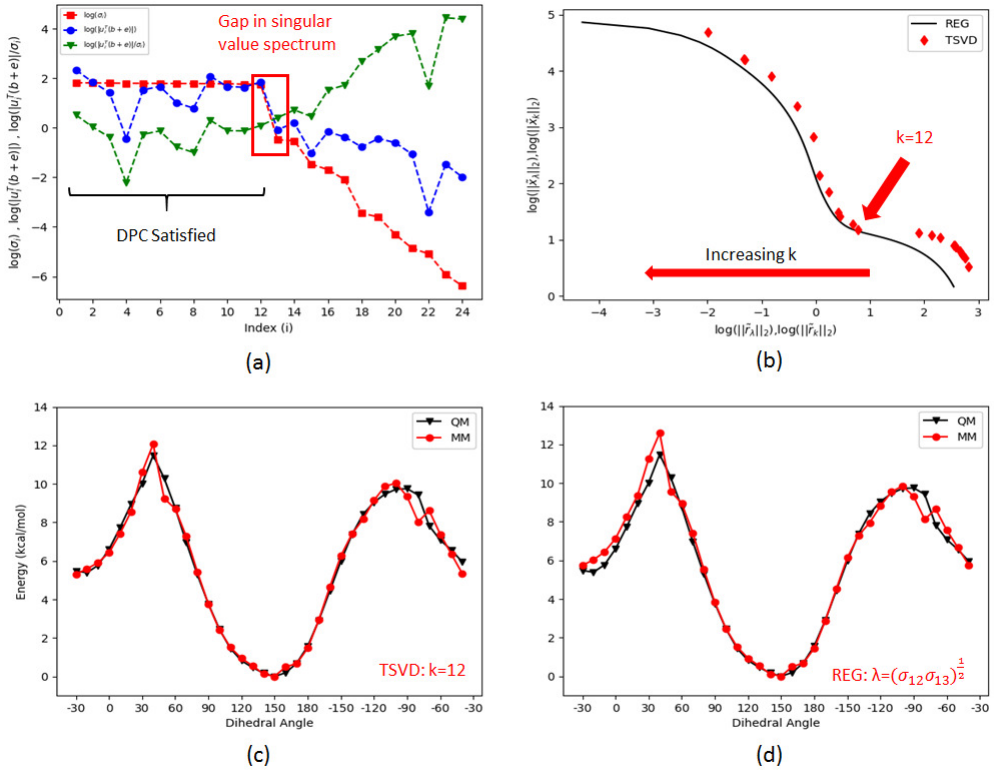
Relaxed MM scans of the adduct covalent bond were repeated for the PHE and B[c]P model systems utilizing the TSVD (Figure 3.4(c) and 3.5(c)) and Tikhonov Regularization (Figure 3.4(d) and 3.5(d)) optimized dihedral terms. In all cases the MM PES achieved an accurate fit to the target QM PES with the resulting RMSEs less than the 1.0 kcal/mol threshold for chemical accuracy (Table 3.3) and demonstrating the effectiveness of this parameterization approach.

PHE					
		$\tilde{\mathbf{x}}_{k=12}$		$\tilde{\mathbf{x}}_{\lambda=1.8936}$	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	1.9836	-163.3974°	1.8087	-163.7064°
	2	1.0361	175.6410°	0.9515	175.5172°
	3	0.1689	-105.5212°	0.1628	-111.3787°
	4	0.4336	-94.13046°	0.4042	-95.8530°
	5	0.3374	-109.4845°	0.3012	-118.7172°
	6	0.0619	-157.9412°	0.1356	98.1378°
dih_2	1	1.9822	-38.1046°	1.8015	-38.1292°
	2	1.0697	69.3318°	0.9610	69.6258°
	3	0.2503	-134.1464°	0.2305	-134.2927°
	4	0.3125	39.4302°	0.2482	37.1320°
	5	0.3365	159.8988°	0.2981	169.9803°
	6	0.1428	-110.1871°	0.2644	-84.9764°

Table (3.1) TSVD and Tikhonov Regularization optimized dihedral terms for the PHE model system.

B[c]P					
		$\tilde{\mathbf{x}}_{k=12}$		$\tilde{\mathbf{x}}_{\lambda=1.8800}$	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	2.8736	-145.8997°	2.6145	-145.5963°
	2	1.5336	-172.9189°	1.3922	-172.3537°
	3	0.2343	126.6249°	0.2158	125.8856°
	4	0.2127	79.1159°	0.2373	94.0870°
	5	0.2334	50.8627°	0.1727	52.2543°
	6	0.2297	172.6102°	0.1487	159.7990°
dih_2	1	2.8802	-20.6135°	2.6183	-20.7331°
	2	1.4288	76.6190°	1.2897	75.8955°
	3	0.1162	108.7005°	0.0857	99.3821°
	4	0.2065	162.5793°	0.2541	143.4855°
	5	0.3242	-64.0855°	0.3336	-63.7747°
	6	0.0994	-29.2053°	0.1220	-56.0529°

Table (3.2) TSVD and Tikhonov Regularization optimized dihedral terms for the B[c]P model system.



(3.4) PHE model system:

(a) Well defined gap in the singular value spectrum between σ_{12} and σ_{13} [red squares: $\{\sigma_i\}$] and in practice, the DPC satisfied for i $= 1,...,12$ resulting in $k = 12$. Note also that the truncation parameter should not be set between (nearly) multiple values. blue circles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|\}$ and green triangles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|/\sigma_i\}$

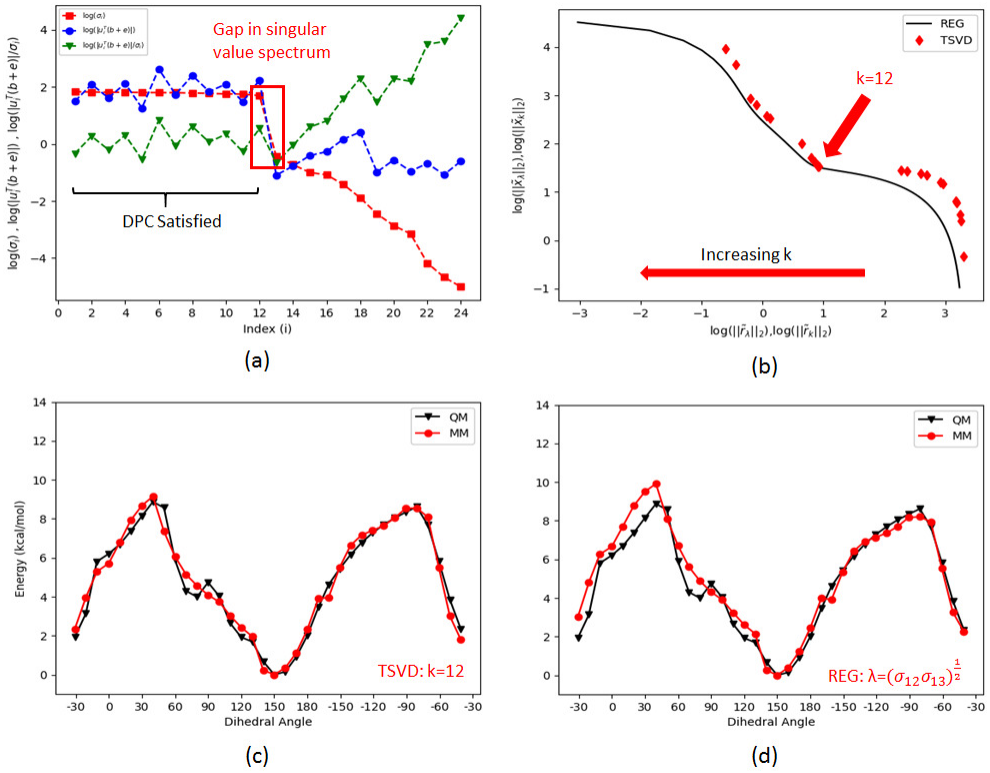
(b) Corner in the log scale L-curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds)

(c) MM PES (red circles) with TSVD optimized dihedral terms ($k = 12$) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ_{dih_1}

(d) MM PES (red circles) with Tikhonov Regularization optimized dihedral terms ($\lambda = (\sigma_{12}\,\sigma_{13})^{\frac{1}{2}} = 1.8936$) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ_{dih_1}

	PHE		B[c]P	
	$\tilde{\mathbf{x}}_{k=12}$	$\tilde{\mathbf{x}}_{\lambda=1.8936}$	$\tilde{\mathbf{x}}_{k=12}$	$\tilde{\mathbf{x}}_{\lambda=1.8800}$
max abs error	1.3889	1.3003	1.2162	1.6645
RMSE	0.4102	0.4885	0.4817	0.6923

Table (3.3) Error Data (kcal/mol): Adduct covalent bond dihedral angle ϕ, MM PES fit to QM PES



(3.5) B[c]P model system:

(a) Well defined gap in the singular value spectrum between σ_{12} and σ_{13} [red squares: $\{\sigma_i\}$] and in practice, the DPC satisfied for i $= 1,...,12$ resulting in $k = 12$. Note also that the truncation parameter should not be set between (nearly) multiple values. blue circles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|\}$ and green triangles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|/\sigma_i\}$

(b) Corner in the log scale L-curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds)

(c) MM PES (red circles) with TSVD optimized dihedral terms ($k = 12$) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ_{dih_1}

(d) MM PES (red circles) with Tikhonov Regularization optimized dihedral terms ($\lambda = (\sigma_{12}\,\sigma_{13})^{\frac{1}{2}} = 1.8800$) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ_{dih_1}

3.4.3 Sensitivity of Solutions to Perturbation and the Impact of Small Singular Values

As described above, truncation and regularization parameters should be selected where the DPC is satisfied and to the *right* of the corner in the corresponding L-curve so that the resulting solutions $\tilde{\mathbf{x}}_k$ and $\tilde{\mathbf{x}}_\lambda$ are not impacted by small singular values and are dominated by $\mathbf{x}_k$ and $\mathbf{x}_\lambda$ respectively. In order to further illustrate the effectiveness of the TSVD as a regularization tool in force field parameter optimization, we examine the impact of incorrectly selected truncation and regularization parameters where the DPC is not satisfied and whose resulting solutions fall to the *left* of the corner in the L-curve. In this region, solutions are impacted by small singular values and are instead dominated by the terms $\mathbf{x}_k^{(e)}$ (3.22) and $\mathbf{x}_\lambda^{(e)}$ (3.23) that result from perturbations.

Random perturbations on the order of 0.01 kcal/mol were applied to each element of the discrete difference potential $\mathbf{E}^{diff}$ in both the PHE and B[c]P model systems. We then examined the truncated solutions (i.e. the resulting dihedral force constants) to the unperturbed and perturbed inverse problems for $k = 12, 18, 20$. The same was done for the corresponding Tikhonov regularized solutions $\tilde{\mathbf{x}}_\lambda$ where $\lambda = (\sigma_k \sigma_{k+1})^{\frac{1}{2}}$. Results for the TSVD and Tikhonov solutions are respectively listed in Tables 3.4 and 3.5 for the PHE model system and Tables 3.6 and 3.7 for the B[c]P model system.

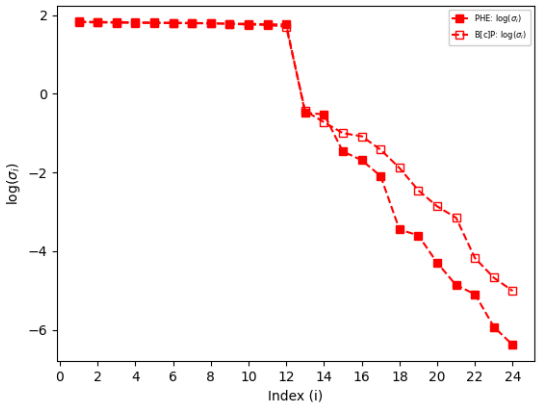
In both model systems, the correct truncated solutions $\tilde{\mathbf{x}}_{k=12}$ and corresponding Tikhonov regularized solutions $\tilde{\mathbf{x}}_\lambda$ (PHE: $\lambda = 1.8936$ and B[c]P: $\lambda = 1.8800$) are largely insensitive to the applied perturbation, with the norm of the difference vector $\|\Delta\tilde{\mathbf{x}}\|_2$ between solutions to the unperturbed and perturbed inverse problems less than or equal to 0.0200 in all cases. This is in contrast to the truncated solutions $\tilde{\mathbf{x}}_{k=18}$ and $\tilde{\mathbf{x}}_{k=20}$ which are more sensitive to perturbations, with $\|\Delta\tilde{\mathbf{x}}\|_2$ increasing with larger k in both model systems (see Tables 3.4-3.7). Additionally the norm of the solutions $\|\tilde{\mathbf{x}}_{k=18}\|_2$ and $\|\tilde{\mathbf{x}}_{k=20}\|_2$ are seen to increase with k as a result of small singular values impacting the solutions. Analogous results are observed for the corresponding Tikhonov regularized solutions $\tilde{\mathbf{x}}_\lambda$ where $\lambda = (\sigma_k \sigma_{k+1})^{\frac{1}{2}}$ and $k = 12, 18, 20$.

From the data in Tables 3.4-3.7, it is apparent that the solutions in the PHE model system

are more sensitive to perturbation and the solution norms more susceptible to inflation with larger truncation parameter k than those in the B[c]P model system. In fact, the B[c]P system is arguably insensitive to perturbation as compared to the PHE system. This can be understood in terms of their respective singular value spectra as seen in Figure 3.6, where the PHE system has smaller singular values (solid red squares) for all $k > 14$. Additionally, in each system, the condition numbers $C_k = \sigma_1/\sigma_k$:

$$C_{k=18}(\text{PHE}) = 194.2396 \quad C_{k=18}(\text{B[c]P}) = 41.1578$$
$$C_{k=20}(\text{PHE}) = 452.5983 \quad C_{k=20}(\text{B[c]P}) = 109.2179$$

indicate that the PHE model system is ill-conditioned to a greater extent than the B[c]P model system and thus more sensitive to perturbation.



(3.6) Singular value spectra for the PHE (solid red squares) and B[c]P (hollow red squares) model systems demonstrating more rapid decay in the PHE model system.

		$\tilde{\mathbf{x}}_{k=12}$				$\tilde{\mathbf{x}}_{k=18}$				$\tilde{\mathbf{x}}_{k=20}$			
		UNPERTURBED		PERTURBED		UNPERTURBED		PERTURBED		UNPERTURBED		PERTURBED	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	1.9836	-163.3974°	1.9857	-163.5037 °	4.5291	174.1027 °	4.1943	175.0157 °	4.1233	158.8797 °	3.8776	159.6725 °
	2	1.0361	175.6410°	1.0333	175.6682 °	6.8878	165.5519 °	6.1975	165.7034 °	8.9274	-89.0745 °	9.8450	-87.1626 °
	3	0.1689	-105.5212°	0.1796	-107.9636 °	4.9212	128.2440 °	4.2283	131.6030 °	18.8473	-100.7282 °	20.9049	-103.1965 °
	4	0.4336	-94.1305°	0.4377	-93.6219 °	6.9269	-48.8337 °	5.8326	-52.2305 °	26.0770	-38.6520 °	25.1327	-36.3397 °
	5	0.3374	-109.4845°	0.3324	-111.4711 °	3.3271	-0.4347 °	3.2266	1.1515 °	9.3805	48.9880 °	9.5120	52.0630 °
	6	0.0619	-157.9412°	0.0596	-163.1498 °	3.3728	94.0064 °	3.3039	96.0202 °	5.4188	125.0950 °	5.3861	127.8041 °
dih_2	1	1.9822	-38.1046°	1.9844	-38.2153 °	1.0673	53.2948 °	1.0043	37.9587 °	3.2897	45.7948 °	3.3304	41.0820 °
	2	1.0697	69.3318°	1.0668	69.3347 °	5.2018	-121.6995 °	4.4230	-122.4575 °	8.2496	-11.4904 °	9.2250	-8.7458 °
	3	0.2503	-134.1464°	0.2619	-133.6433 °	5.5562	-34.8725 °	4.7508	-32.7047 °	17.8446	95.0837 °	20.0402	93.0311 °
	4	0.3125	39.4302°	0.3180	40.4796 °	6.0436	-82.1540 °	5.0021	-85.3869 °	26.7898	-76.2883 °	25.9308	-74.4828 °
	5	0.3365	159.8988°	0.3332	157.8033 °	3.9371	97.6314 °	3.7627	98.2402 °	9.7453	128.4088 °	9.6548	132.1679 °
	6	0.1428	-110.1871°	0.1407	-112.4694 °	3.1921	-61.8583 °	3.1638	-58.5558 °	4.1183	-23.4598 °	4.2050	-20.6666 °
		$\|\tilde{\mathbf{x}}_k\|_2$ = 3.2726		$\|\tilde{\mathbf{x}}_k\|_2$ = 3.2749		$\|\tilde{\mathbf{x}}_k\|_2$ = 16.8489		$\|\tilde{\mathbf{x}}_k\|_2$ = 14.8848		$\|\tilde{\mathbf{x}}_k\|_2$ = 49.7617		$\|\tilde{\mathbf{x}}_k\|_2$ = 50.8057	
		$\|\Delta\tilde{\mathbf{x}}\|_2 = 0.0191$				$\|\Delta\tilde{\mathbf{x}}\|_2 = 2.1578$				$\|\Delta\tilde{\mathbf{x}}\|_2 = 3.5459$			

Table (3.4) PHE model system: TSVD optimized dihedral terms demonstrating sensitivity to perturbation and inflation of solution norms as a function of the truncation parameter $k = 12, 18, 20$.

		$\tilde{\mathbf{x}}_{\lambda=1.8936}$				$\tilde{\mathbf{x}}_{\lambda=0.02947}$				$\tilde{\mathbf{x}}_{\lambda=0.01023}$			
		UNPERTURBED		PERTURBED		UNPERTURBED		PERTURBED		UNPERTURBED		PERTURBED	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	1.8087	-163.7064°	1.8105	-163.7956 °	3.1305	150.3020 °	2.9309	154.7461 °	7.7152	97.7078 °	6.1263	101.8785 °
	2	0.9515	175.5172°	0.9481	175.5477 °	3.5593	-154.2105 °	3.2661	-153.5830 °	12.2113	-131.2689 °	10.8047	-133.3013 °
	3	0.1628	-111.3787°	0.1722	-113.1177 °	3.8697	-68.9391 °	4.3120	-76.5660 °	13.9181	-71.8563 °	15.5114	-78.4202 °
	4	0.4042	-95.8530°	0.4056	-95.3938 °	9.7343	-59.3063 °	8.9094	-59.0581 °	18.9148	-42.3965 °	18.0299	-39.7684 °
	5	0.3012	-118.7172°	0.2950	-120.8344 °	4.0433	12.2131 °	3.9654	16.1683 °	6.7979	41.8796 °	6.8594	46.7146 °
	6	0.1356	98.1378°	0.1388	99.2256 °	3.8580	100.7959 °	3.7637	103.2530 °	4.7480	117.5766 °	4.6218	120.6032 °
dih_2	1	1.8015	-38.1292°	1.8038	-38.2531 °	2.6326	17.9451 °	2.4385	14.3078 °	9.2467	22.0590 °	7.6900	20.7989 °
	2	0.9610	69.6258°	0.9595	69.6212 °	2.2110	-71.6421 °	1.8521	-67.0099 °	11.0438	-58.8617 °	9.5184	-60.2472 °
	3	0.2305	-134.2927°	0.2418	-133.8513 °	3.1579	132.9252 °	3.6397	123.8504 °	13.0112	125.5775 °	14.6892	118.7312 °
	4	0.2482	37.1320°	0.2553	38.5918 °	9.4473	-93.2413 °	8.6665	-92.9398 °	19.2346	-78.7125 °	18.4524	-76.3663 °
	5	0.2981	169.9803°	0.2966	167.6763 °	4.7397	97.0668 °	4.5690	100.3576 °	7.5178	119.5891 °	7.3241	124.4312 °
	6	0.2644	-84.9764°	0.2568	-85.4520 °	3.3882	-50.0110 °	3.3572	-46.9664 °	3.7302	-30.9668 °	3.7289	-27.5632 °
		$\|\tilde{\mathbf{x}}_k\|_2$ = 2.9864		$\|\tilde{\mathbf{x}}_k\|_2$ = 2.9881		$\|\tilde{\mathbf{x}}_k\|_2$ = 17.5619		$\|\tilde{\mathbf{x}}_k\|_2$ = 16.6432		$\|\tilde{\mathbf{x}}_k\|_2$ = 40.5714		$\|\tilde{\mathbf{x}}_k\|_2$ = 39.5211	
		$\|\Delta\tilde{\mathbf{x}}\|_2 = 0.0200$				$\|\Delta\tilde{\mathbf{x}}\|_2 = 1.4336$				$\|\Delta\tilde{\mathbf{x}}\|_2 = 4.0075$			

Table (3.5) PHE model system: Tikhonov regularization optimized dihedral terms demonstrating sensitivity to perturbation and inflation of solution norms as a function of the regularization parameter $\lambda = (\sigma_k \sigma_{k+1})^{\frac{1}{2}}$ where $k = 12, 18, 20$.

		$\tilde{\mathbf{x}}_{k=12}$				$\tilde{\mathbf{x}}_{k=18}$				$\tilde{\mathbf{x}}_{k=20}$			
		UNPERTURBED		PERTURBED		UNPERTURBED		PERTURBED		UNPERTURBED		PERTURBED	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	2.8736	-145.8996 °	2.8749	-145.9935 °	3.2897	-126.7633 °	3.2783	-127.2927 °	4.1625	-127.3966 °	4.0680	-129.0660 °
	2	1.5336	-172.9189 °	1.5315	-172.8940 °	2.3844	-121.1935 °	2.3461	-122.3223 °	0.4154	140.8178 °	0.7416	169.8488 °
	3	0.2343	126.6249 °	0.2345	129.8358 °	1.9092	100.1278 °	1.9040	101.1249 °	8.0944	119.1925 °	7.4995	120.5342 °
	4	0.2127	79.1159 °	0.2101	78.1791 °	7.3480	136.7974 °	7.1422	136.8871 °	6.7676	160.2066 °	6.5635	157.2829 °
	5	0.2334	50.8627 °	0.2352	54.3912 °	1.7836	-85.5908 °	1.8140	-79.4467 °	2.1474	-125.6733 °	2.0030	-118.3994 °
	6	0.2297	172.6102 °	0.2302	171.3138 °	1.0237	148.8379 °	1.1069	149.2706 °	0.2591	107.9973 °	0.4148	132.4549 °
dih_2	1	2.8802	-20.6135 °	2.8813	-20.7036 °	2.2393	-38.4325 °	2.2528	-38.4459 °	1.6502	-69.3508 °	1.5958	-63.7310 °
	2	1.4288	76.6190 °	1.4267	76.6312 °	2.3755	19.7827 °	2.3255	20.4648 °	2.5647	87.4535 °	2.1023	86.6096 °
	3	0.1162	108.7005 °	0.1087	114.3647 °	2.2926	-61.3749 °	2.2933	-61.0758 °	8.4717	-43.2008 °	7.8867	-42.1631 °
	4	0.2065	162.5793 °	0.2087	161.6634 °	7.4297	104.5112 °	7.2181	104.7043 °	7.2445	125.7236 °	7.0088	123.1769 °
	5	0.3242	-64.0855 °	0.3202	-61.9006 °	1.9402	-4.0382 °	1.9663	2.0463 °	2.6795	-46.2743 °	2.4726	-39.8343 °
	6	0.0994	-29.2053 °	0.0962	-26.3620 °	1.3957	1.3460 °	1.4587	2.5513 °	0.2387	-20.9919 °	0.4419	-3.1034 °
		$\|\tilde{\mathbf{x}}_k\|_2$ = 4.6179		$\|\tilde{\mathbf{x}}_k\|_2$ = 4.6177		$\|\tilde{\mathbf{x}}_k\|_2$ = 12.4579		$\|\tilde{\mathbf{x}}_k\|_2$ = 12.2164		$\|\tilde{\mathbf{x}}_k\|_2$ = 16.5617		$\|\tilde{\mathbf{x}}_k\|_2$ = 15.6590	
		$\|\Delta\tilde{\mathbf{x}}\|_2 = 0.0105$				$\|\Delta\tilde{\mathbf{x}}\|_2 = 0.3223$				$\|\Delta\tilde{\mathbf{x}}\|_2 = 1.1201$			

Table (3.6) B[c]P model system: TSVD optimized dihedral terms demonstrating sensitivity to perturbation and inflation of solution norms as a function of the truncation parameter $k = 12, 18, 20$.

		$\tilde{\mathbf{x}}_{\lambda=1.8800}$				$\tilde{\mathbf{x}}_{\lambda=0.1142}$				$\tilde{\mathbf{x}}_{\lambda=0.04948}$			
		UNPERTURBED		PERTURBED		UNPERTURBED		PERTURBED		UNPERTURBED		PERTURBED	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	2.6145	-145.5963 °	2.6157	-145.7004 °	3.4313	-119.5957 °	3.4217	-121.7972 °	5.4720	-93.7359 °	5.3127	-97.7906 °
	2	1.3922	-172.3537 °	1.3909	-172.3748 °	1.6199	-147.1732 °	1.7438	-148.2804 °	1.1758	-147.1666 °	1.5614	-145.7696 °
	3	0.2158	125.8856 °	0.2161	129.1298 °	2.6530	126.2226 °	2.5570	126.9020 °	5.7400	128.6639 °	5.4398	129.0827 °
	4	0.2373	94.0870 °	0.2298	92.8775 °	4.8379	139.4046 °	4.6895	139.0099 °	5.7125	139.9923 °	5.5537	138.7863 °
	5	0.1727	52.2543 °	0.1779	54.4603 °	1.5106	-121.6332 °	1.4155	-114.9485 °	2.1222	-122.1002 °	1.9784	-116.0373 °
	6	0.1487	159.7990 °	0.1591	157.3011 °	0.6869	109.9340 °	0.7721	115.6206 °	0.6220	121.3522 °	0.7494	126.5423 °
dih_2	1	2.6183	-20.7331 °	2.6190	-20.8147 °	2.5806	-49.5735 °	2.4991	-47.9476 °	4.5526	-87.1235 °	4.1521	-87.3711 °
	2	1.2897	75.8955 °	1.2871	75.9721 °	1.4986	49.0753 °	1.4058	45.6093 °	1.6617	62.9923 °	1.4260	51.1453 °
	3	0.0857	99.3821 °	0.0772	104.9061 °	2.8021	-38.5210 °	2.7081	-38.4191 °	6.0206	-35.3236 °	5.7254	-35.2145 °
	4	0.2541	143.4855 °	0.2540	144.0647 °	5.0603	108.3178 °	4.8997	108.1359 °	6.0156	107.3269 °	5.8275	106.3622 °
	5	0.3336	-63.7747 °	0.3244	-61.2201 °	1.9056	-38.9303 °	1.7909	-33.1811 °	2.5888	-40.2021 °	2.4008	-34.8268 °
	6	0.1220	-56.0529 °	0.1226	-51.3962 °	0.8943	-35.2208 °	0.9631	-30.3511 °	0.7816	-26.5348 °	0.9038	-21.4715 °
		$\|\tilde{\mathbf{x}}_k\|_2$ = 4.2006		$\|\tilde{\mathbf{x}}_k\|_2$ = 4.1999		$\|\tilde{\mathbf{x}}_k\|_2$ = 9.7154		$\|\tilde{\mathbf{x}}_k\|_2$ = 9.4645		$\|\tilde{\mathbf{x}}_k\|_2$ = 14.3190		$\|\tilde{\mathbf{x}}_k\|_2$ = 13.7174	
		$\|\Delta\tilde{\mathbf{x}}\|_2 = 0.0190$				$\|\Delta\tilde{\mathbf{x}}\|_2 = 0.3616$				$\|\Delta\tilde{\mathbf{x}}\|_2 = 0.8457$			

Table (3.7) B[c]P model system: Tikhonov regularization optimized dihedral terms demonstrating sensitivity to perturbation and inflation of solution norms as a function of the regularization parameter $\lambda = (\sigma_k \sigma_{k+1})^{\frac{1}{2}}$ where $k = 12, 18, 20$.

3.4.4 Inverse Problem with Ill-Determined Numerical Rank and Optimization of Multiple Dihedral Parameters

In order to examine a case where the inverse problem has ill-determined numerical rank (i.e. there is not a well defined gap in the singular value spectrum), we applied the TSVD approach to simultaneously optimize all dihedrals (excluding those involving hydrogens) with CGenFF penalty scores greater than 9 in the PHE model system (see Table 3.8). In addition to the ϕ_{dih_1} PES utilized above to optimize dih_1 and dih_2, relaxed QM torsion scans of the dihedrals ϕ_{dih_3},...,ϕ_{dih_6} were conducted at the MP2/6-31G(d) level of theory, scanning in $\pm$ 10° increments from the global minimum structure obtained from the ϕ_{dih_1} PES. As the central bond in each of the additional dihedrals ϕ_{dih_3},...,ϕ_{dih_6} is part of the PHE aliphatic ring, an energy cutoff of 10 kcal/mol was utilized for fitting and four scan points on the ϕ_{dih_6} PES that lie greater than 50° from the minimum structure and correspond to a conformational change in the aliphatic ring were excluded. This results in $m = 85$ discrete scan points $\{(\phi_{dih_1,\mathrm{i}}, \phi_{dih_2,\mathrm{i}}, ..., \phi_{dih_6,\mathrm{i}}) | \mathrm{i} = 1, ..., m\}$. As described above, analogous relaxed MM PES scans were conducted with the dihedral force constants for dih_1,..., dih_6 set to zero and the resulting discrete difference potential $E^{diff} = \{E_\mathrm{i} | \mathrm{i} = 1, ..., m\}$ where $E_\mathrm{i} = E_\mathrm{i}^{QM} - E_\mathrm{i}^{MM_{k_{dih1} \cdots k_{dih6}=0}}$ elucidates the form of the dihedral potentials that the sum of the dih_1,..., dih_6 dihedral force field terms must fit in order for the MM PESs to accurately model the their respective QM PESs.

Label	Atom Names	CGenFF Penalty
dih_1	C6-N6-C20-C20A	75
dih_2	C6-N6-C20-C19	46.5
dih_3	N6-C20-C20A-C22D	37.5
dih_4	N6-C20-C19-C18	9.2
dih_5	N6-C20-C19-O19	38.5
dih_6	C16A-C17-C18-O18	26

Table (3.8) Parameters labels, atom names, and CGenFF analogy assignment penalties for simultaneously optimized dihedrals utilizing the TSVD approach in the PHE model system.

As above, where $M_1, ..., M_6 \subseteq \{1, 2, 3, 4, 5, 6\}$ are the multiplicities of the dihedral terms for the

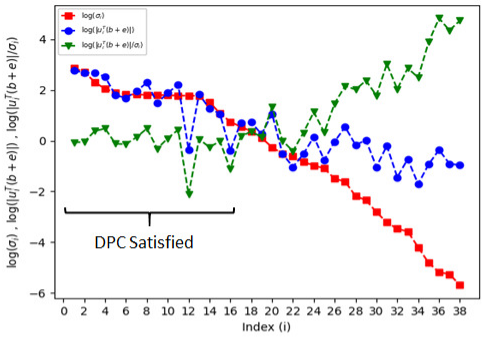
$dih_1, ..., dih_6$ parameters, we seek to optimize the force constants in the sum of dihedral terms:

$$\begin{aligned} E_{\phi_{dih_1}} + ... + E_{\phi_{dih_6}} = & \sum_{j_1 \in M_1} k_{j_1} \left[1 + \cos(j_1 \phi_{dih_1} - \delta_{j_1})\right] + ... \\ & + \sum_{j_6 \in M_6} k_{j_6} \left[1 + \cos(j_6 \phi_{dih_6} - \delta_{j_6})\right] \end{aligned} \tag{3.44}$$

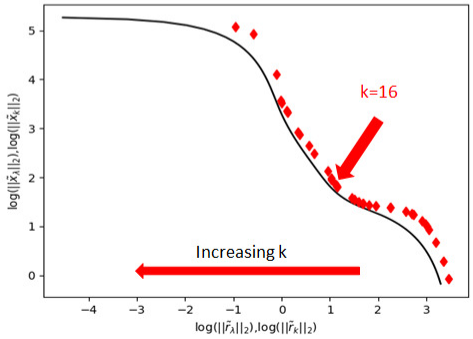
The elements of the matrix $\mathbf{A}$ and vector $\mathbf{b}+\mathbf{e}$ are analogous to those described above with multiplicities listed in Table 3.9 with the exception that dih_5 and dih_6 were parameterized with fixed phase terms (i.e. $\delta = 0^\circ$ or 180°) while dih_1,..., dih_4 are parameterized with variable phase terms. Note that we are aware of the sp2 center on C20A with identical constituent atom types in C22D and C16A and we find that parameterization utilizing the ϕ_{dih_3} PES and variable phase dihedrals suffices for this illustrative example.

We observe in Figure 3.7(a) that there is no obvious gap in the singular value spectrum (red squares), and examination of the relative gaps $\omega_k = \sigma_{k+1}/\sigma_k$ confirms that there is no numerically well defined gap in the singular value spectrum. We then turn our attention to examining the terms $|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|/\sigma_i$ (green triangles) in Figure 3.7(a) where it is apparent that the DPC is satisfied in practice up to $k = 16$. Additionally, the corner in the L-curve in Figure 3.7(b), indicates that $k = 16$ is the appropriate truncation parameter.

Utilizing these observations, we obtain the TSVD solution $\tilde{\mathbf{x}}_{k=16}$ and corresponding Tikhonov regularized solution $\tilde{\mathbf{x}}_{\lambda=1.9089}$ where $\lambda = (\sigma_{16}\sigma_{17})^{\frac{1}{2}}$. The resulting CHARMM compatible dihedral terms are listed in Table 3.9. Relaxed MM scans of ϕ_{dih_1} and ϕ_{dih_3},...,ϕ_{dih_6} were repeated utilizing the TSVD optimized dihedral terms (Figure 3.8(a)-(e) red circles) with each MM PES achieving a good fit to the corresponding QM PES (Figure 3.8(a)-(e) black triangles) with RMSE less than 1.0 kcal/mol in all cases (Table 3.10, top panel). The same was done utilizing the Tikhonov regularized dihedral terms (Figure 3.9(a)-(e)), yielding similar results (Table 3.10, bottom panel). We note that in the cases we have examined, the TSVD approach generally results in slightly better MM PES fits to QM target data, demonstrating the effectiveness of the TSVD as a regularization tool on its own.



(a)

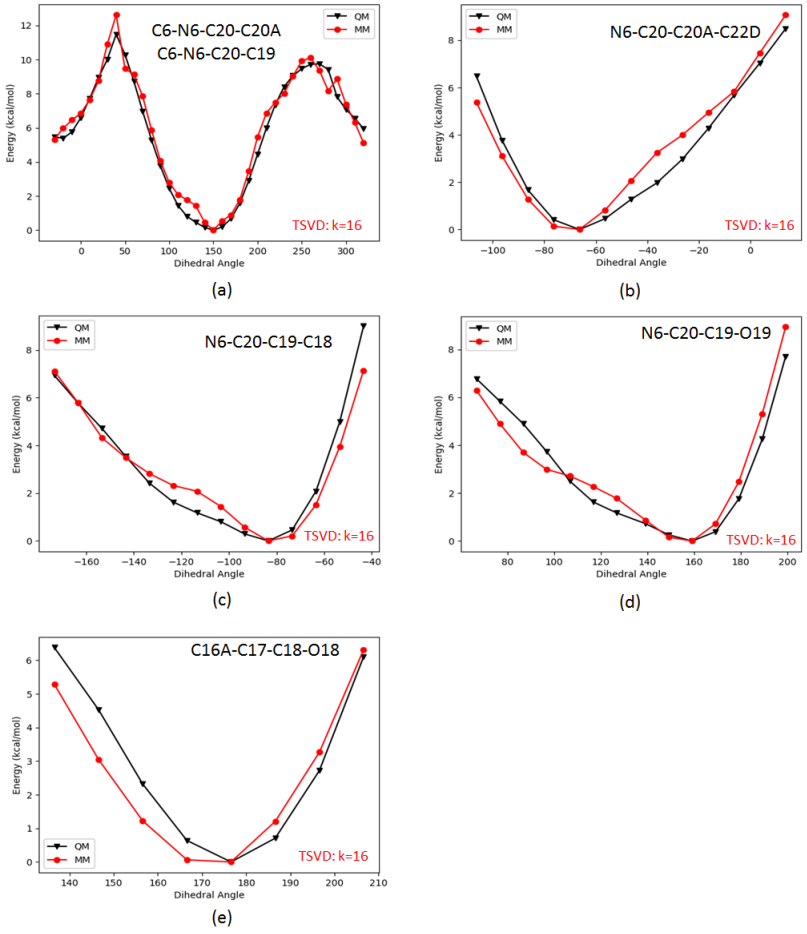


(b)

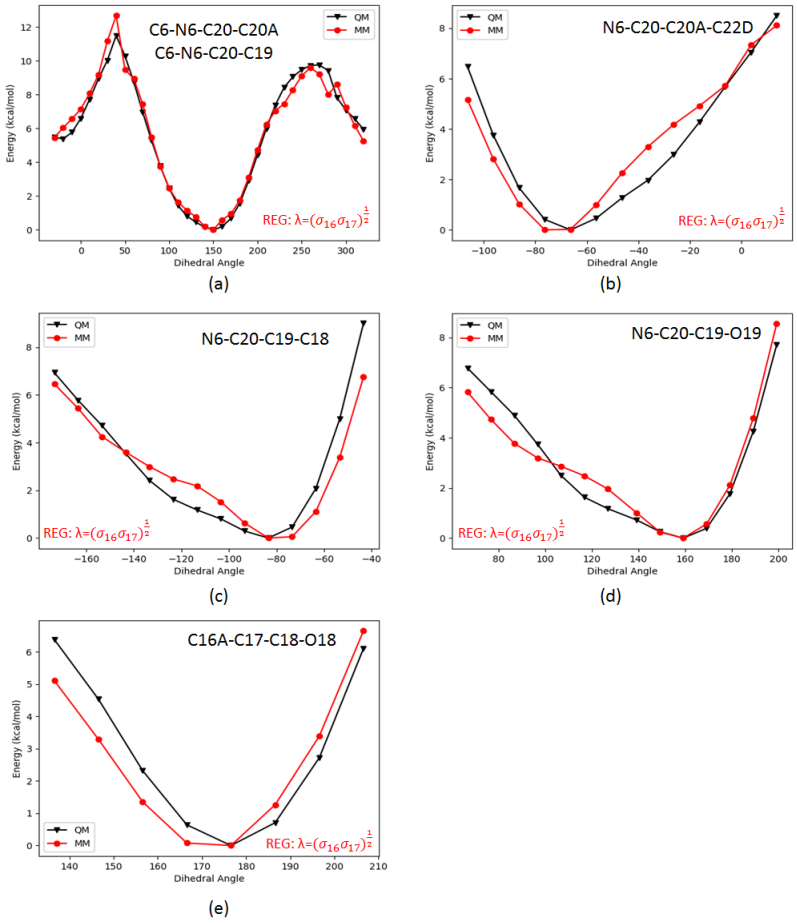
(3.7) Parameter optimization of multiple dihedrals in the syn-glycosidic PHE model system: (a) Singular value spectrum without a well defined gap [red squares: $\{\sigma_i\}$] and in practice, the DPC satisfied for $i = 1,...,16$ resulting in $k = 16$. blue circles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|\}$ and green triangles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|/\sigma_i\}$ (b) Corner in the log scale L-curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds)

PHE					
		$\tilde{\mathbf{x}}_{k=16}$		$\tilde{\mathbf{x}}_{\lambda=1.9089}$	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	2.0685	-175.0417 °	1.8858	-174.4335°
	2	0.9038	155.3065 °	0.8075	152.9721 °
	3	0.3603	-144.0670 °	0.4129	-147.5366°
	4	0.5953	-71.7942 °	0.4924	-62.5803 °
	5	0.2409	-154.2063 °	0.3488	-150.8455°
	6	0.4657	-99.7501 °	0.1543	-73.5062 °
dih_2	1	1.9950	-53.1178 °	1.8044	-53.5979 °
	2	1.1797	60.6358 °	1.0392	63.6460 °
	3	0.1808	-75.6781 °	0.0935	-104.4248°
	4	0.3081	-27.8215 °	0.3180	-15.4688 °
	5	0.8381	-178.2084 °	0.6017	-164.0427°
	6	0.3746	122.4060 °	0.1261	-141.0168°
dih_3	1	0.9650	-134.9080 °	1.1438	-132.9850°
	2	1.2751	179.2733 °	1.3199	-175.2967°
	3	0.7435	128.5063 °	0.3937	149.6747 °
dih_4	1	1.2092	-15.6620 °	1.0770	-17.7731 °
	2	1.2323	-115.6482 °	1.1135	-119.5223°
	3	0.5154	-136.3010 °	0.5451	-164.0336°
dih_5	3	0.3226	180.0000 °	0.7236	180.0000°
dih_6	3	0.4465	0.0000 °	0.8724	0.0000°

Table (3.9) TSVD and Tikhonov Regularization dihedral terms simultaneously optimized for the PHE model system.



(3.8) TSVD parameter optimization of multiple dihedrals in the syn-glycosidic PHE model system ($k = 16$). MM PES (red circles) and target QM PES (black triangles) for: (a) ϕ_{dih_1} & ϕ_{dih_2} (b) ϕ_{dih_3} (c) ϕ_{dih_4} (d) ϕ_{dih_5} (e) ϕ_{dih_6}



(3.9) Tikhonov Regularization parameter optimization of multiple dihedrals in the syn-glycosidic PHE model system ($\lambda = 1.9089$). MM PES (red circles) and target QM PES (black triangles) for: (a) ϕ_{dih_1} & ϕ_{dih_2} (b) ϕ_{dih_3} (c) ϕ_{dih_4} (d) ϕ_{dih_5} (e) ϕ_{dih_6}

	$\tilde{\mathbf{x}}_{k=16}$				
	dih_1 & dih_2	dih_3	dih_4	dih_5	dih_6
max abs error	1.2436	1.2745	1.8532	1.2359	1.4792
RMSE	0.6262	0.6939	0.7088	0.7185	0.8298

	$\tilde{\mathbf{x}}_{\lambda=1.9089}$				
	dih_1 & dih_2	dih_3	dih_4	dih_5	dih_6
max abs error	1.4157	1.3307	2.2479	1.1164	1.2722
RMSE	0.5591	0.7997	0.9264	0.6745	0.8269

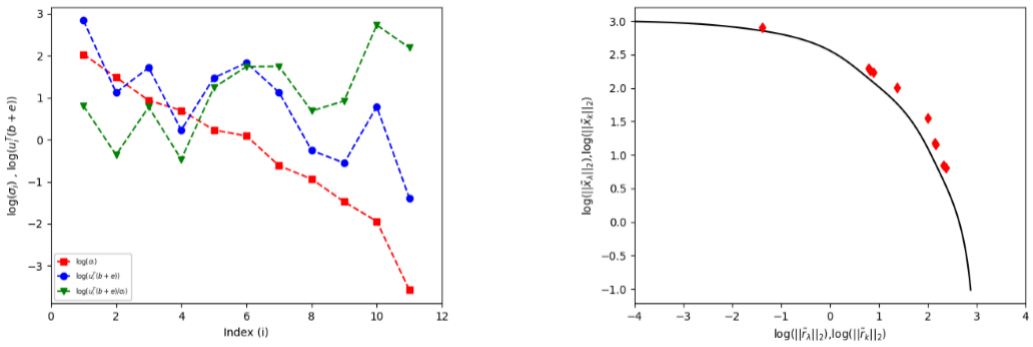
Table (3.10) Error Data (kcal/mol): MM PES fits to QM target data obtained with TSVD ($\tilde{\mathbf{x}}_{k=16}$) and Tikhonov Regularization ($\tilde{\mathbf{x}}_{\lambda=1.9089}$) dihedral terms simultaneously optimized for the syn-glycosidic PHE model system.

3.5 Dihedral Parameterization of Anti-Glycosidic Model Systems Utilizing Fixed Phase

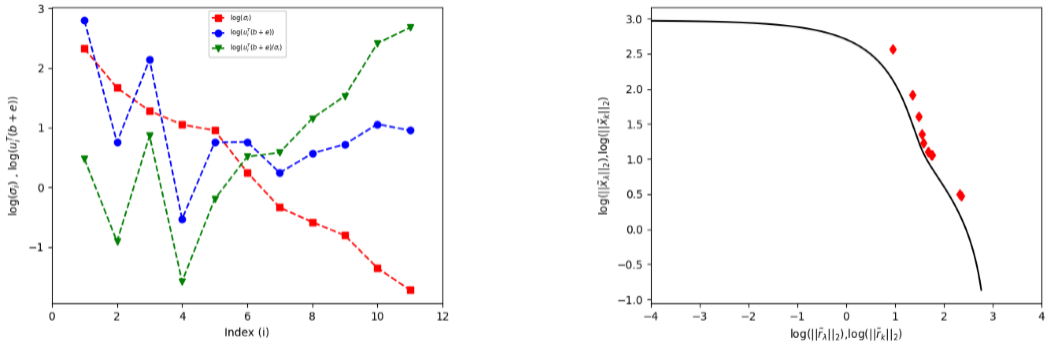
Utilizing the approach to generate QM target data detailed in the sections above, parameters were reoptimized for the anti-glycosidic versions of the PHE and B[c]P model systems utilizing the updated CGenFF 4.4 force field and CGenFF / Paramchem.com (ver 2.4) program[40–42] that were issued during the course of this work. These updates resulted in new parameters assigned by analogy, with only two dihedral parameters having penalty scores above 10: dih_1 (penalty 26) and dih_6 (penalty 13). Additionally, dihedral multiplicities were distinct from those assigned by CGenFF / ParamChem.com (ver 2.2) in some cases. Because the dih_1 - dih_5 dihedrals govern the relative orientation of adenine and the adducted PAH, and because dih_6 has a high penalty score, all six parameters were reoptimized for use with the CGenFF 4.4 force field despite some having low penalty scores. However, we elected to use even dihedral functions (i.e. phase factors of 0.00° or 180.00°) and the multiplicities assigned by CGenFF / ParamChem.com (ver 2.4) (Table 3.11), despite the approach detailed above, due to the lower penalty scores associated with the force field update and to evaluate the efficacy of TSVD parameter optimization with even dihedrals.

Simultaneous optimization of dihedral terms for dih_1 - dih_6 was carried out utilizing a 10 kcal/mol cutoff for fitting QM target data, resulting in systems of equations which are essentially not ill-conditioned with C_{PHE} = 270.57 and $C_{B[c]P}$ = 57.32. As a result, we utilized a combination of the TSVD approach described above and one round of the Downhill Simplex method in the VMD-ffTK. Note this approach is distinct from the standard VMD-ffTK[44,48] approach of using CGenFF assigned dihedral terms as an initial guess for multiple rounds of multiple iteration Monte Carlo Simulated Annealing followed by multiple rounds of Downhill Simplex. In both the PHE and B[c]P systems, there is no obvious gap in the singular value spectrum (red squares in Figures 3.10(a) and 3.11(a) respectively), and examination of the relative gaps $\omega_k = \sigma_{k+1}/\sigma_k$ in the singular value spectrum confirms that there is no numerically well defined gap in the singular value spectrum of either system. Examination of the terms $|\mathbf{u}_\mathrm{i}^T(\mathbf{b}+\mathbf{e})|/\sigma_\mathrm{i}$ (green triangles in Figures

3.10(a) and 3.11(a) respectively) indicate that the DPC is satisfied in practice up to $k = 4$ in both systems. Note that there is no obvious corner in the L-curve for either system (Figures 3.10(b) and 3.11(b) respectively). In each system, the respective TSVD $\tilde{\mathbf{x}}_{k=4}$ solution was utilized as an initial input for one round of Downhill Simplex optimization in the VMD-ffTK yielding the dihedral terms listed in Table 3.11 in two steps. In the PHE model system, all six sets of optimized dihedral terms result in an MM PES that achieves a good fit to the QM PES (Figure 3.12), with RMSEs less than 1.06 kcal/mol across all data points, and RMSEs as low as 0.2248 kcal/mol for data points within 30° of the absolute minimum (Table 3.12). In the B[c]P model system, all six sets of optimized dihedral terms result in an MM PES that achieves a good fit to the QM PES (Figure 3.13), with RMSEs less than 1.26 kcal/mol across all data points, and RMSEs as low as 0.3051 kcal/mol for data points within 30° of the absolute minimum (Table 3.13). Note that we have also included data (Tables 3.12 and 3.13) and plots (Figures 3.12 and 3.13) of the N6-C20-C20A-C16A (dih_7) dihedral PES which is parameterized by the dih_3 parameter in both systems and also achieves a good fit of the MM PES to the QM PES. Together with the work detailed above, this demonstrates the general efficacy of the TSVD parameter optimization approach for these systems as opposed to multiple round / multiple iteration approaches.



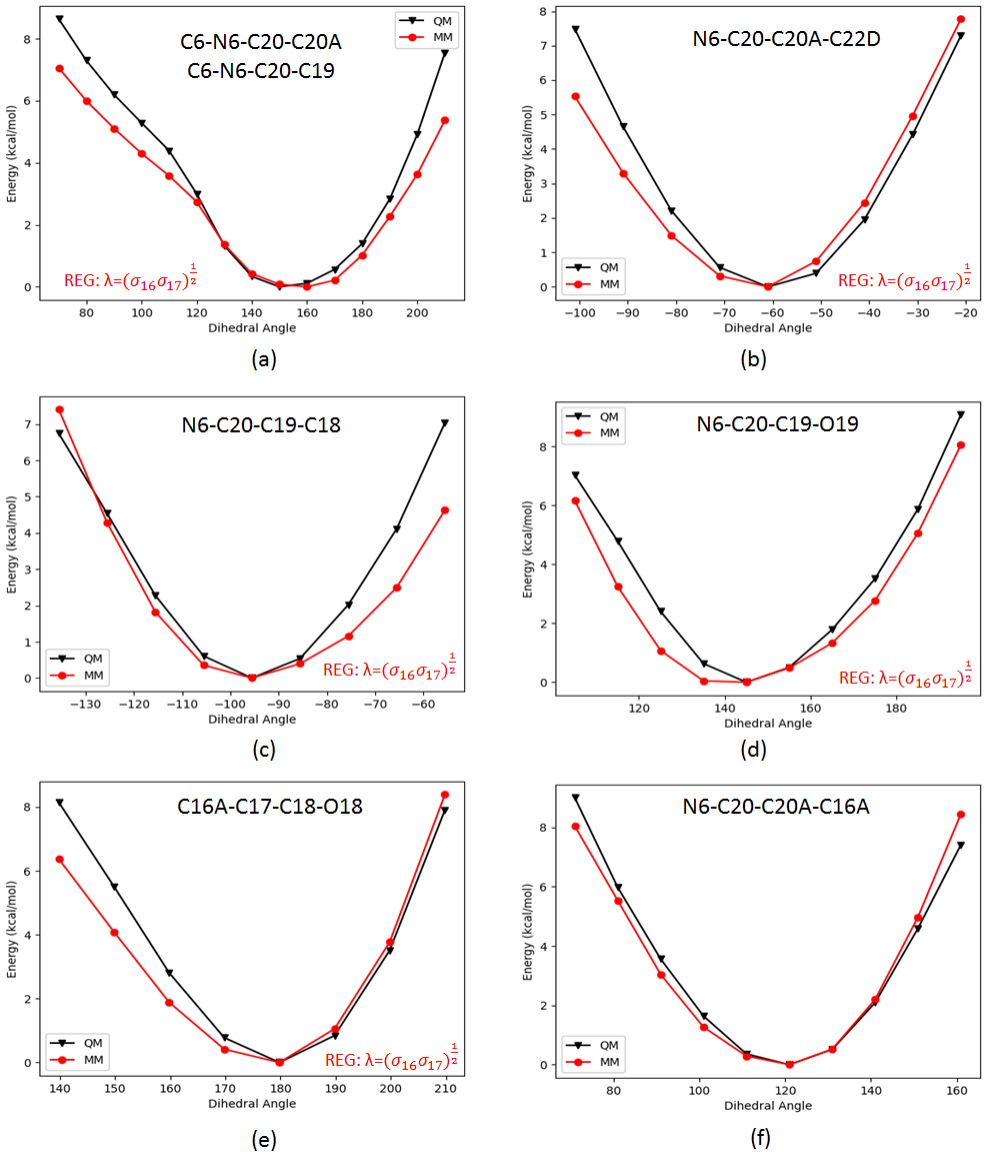
(3.10) Parameter optimization of multiple dihedrals in the anti-glycosidic PHE model system: (a) Singular value spectrum without a well defined gap [red squares: $\{\sigma_i\}$] and in practice, the DPC satisfied for i = 1,...,4 resulting in $k = 4$. blue circles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|\}$ and green triangles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|/\sigma_i\}$ (b) Lack of a corner in the log scale L-curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds)



(3.11) Parameter optimization of multiple dihedrals in the anti-glycosidic B[c]P model system: (a) Singular value spectrum without a well defined gap [red squares: $\{\sigma_i\}$] and in practice, the DPC satisfied for i $= 1,...,4$ resulting in $k = 4$. blue circles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|\}$ and green triangles: $\{|\mathbf{u}_i^T(\mathbf{b}+\mathbf{e})|/\sigma_i\}$ (b) Lack of a corner in the log scale L-curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds)

Dihedral Terms					
		PHE		B[c]P	
		TSVD $\tilde{\mathbf{x}}_{k=4}$ and Downhill Simplex		TSVD $\tilde{\mathbf{x}}_{k=4}$ and Downhill Simplex	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	2.507	180.00 °	3.000	180.00°
	3	0.047	0.00 °	0.396	180.00°
dih_2	1	1.335	180.00 °	0.377	180.00°
	3	0.474	0.00 °	1.530	0.00°
dih_3	2	1.607	180.00 °	0.083	0.00°
dih_4	1	2.191	0.00 °	0.731	0.00°
	3	1.445	180.00 °	0.532	180.00°
dih_5	1	3.000	0.00 °	0.636	0.00°
	3	1.454	180.00 °	2.555	180.00°
dih_6	1	2.999	180.00 °	0.666	0.00°
	3	0.750	0.00 °	1.206	0.00°

Table (3.11) Dihedral terms simultaneously optimized for the anti-glycosidic PHE and B[c]P model systems utilizing TSVD and Downhill Simplex.



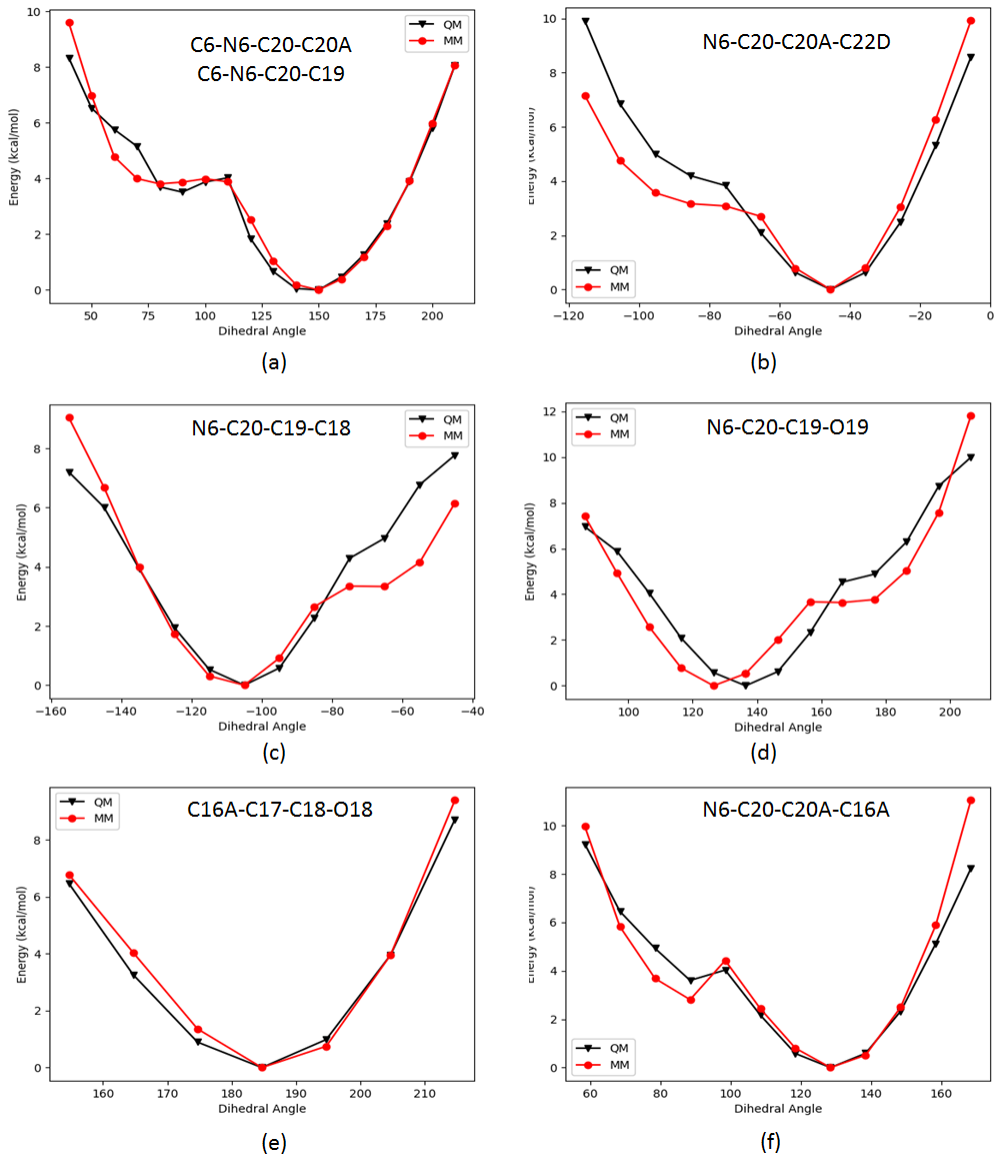
(3.12) TSVD/Downhill Simplex parameter optimization of multiple dihedrals in the anti-glycosidic PHE model system ($k = 4$). MM PES (red circles) and target QM PES (black triangles) for: (a) ϕ_{dih_1} & ϕ_{dih_2} (b) ϕ_{dih_3} (c) ϕ_{dih_4} (d) ϕ_{dih_5} (e) ϕ_{dih_6} (f) ϕ_{dih_7}

	$\tilde{\mathbf{x}}_{k=4}$ and Downhill Simplex					
	dih_1 & dih_2	dih_3	dih_4	dih_5	dih_6	dih_7
max abs error	2.1766	1.9542	2.4081	1.5427	1.7726	1.0370
RMSE	0.9690	0.8890	1.0520	0.8776	0.9015	0.5250
RMSE (w/in 30° of min)	0.2248	0.6612	0.7295	0.8657	0.6927	0.2861

Table (3.12) Error Data (kcal/mol): MM PES fits to QM target data obtained with TSVD ($\tilde{\mathbf{x}}_{k=4}$) and Downhill Simplex dihedral terms simultaneously optimized for the anti-glycosidic PHE model system.

	$\tilde{\mathbf{x}}_{k=4}$ and Downhill Simplex					
	dih_1 & dih_2	dih_3	dih_4	dih_5	dih_6	dih_7
max abs error	1.2636	2.7488	2.6225	1.8121	0.7775	2.8194
RMSE	0.5251	1.2607	1.1994	1.1642	0.4523	0.9982
RMSE (w/in 30° of min)	0.3051	0.5643	0.4230	1.1383	0.4523	0.3570

Table (3.13) Error Data (kcal/mol): MM PES fits to QM target data obtained with TSVD ($\tilde{\mathbf{x}}_{k=4}$) and Downhill Simplex dihedral terms simultaneously optimized for the anti-glycosidic B[c]P model system.



(3.13) TSVD/Downhill Simplex parameter optimization of multiple dihedrals in the anti-glycosidic B[c]P model system ($k = 4$). MM PES (red circles) and target QM PES (black triangles) for: (a) ϕ_{dih_1} & ϕ_{dih_2} (b) ϕ_{dih_3} (c) ϕ_{dih_4} (d) ϕ_{dih_5} (e) ϕ_{dih_6} (f) ϕ_{dih_7}

3.6 Conclusion

We have seen that in molecular mechanics force field parameter optimization, ill-posed least squares problems can be understood in terms of small elements in the singular value spectrum of the matrix **A** that cause standard least squares solutions to blow up, resulting in unusable force field terms. Both the TSVD and Tikhonov Regularization in standard form are effective approaches to ill-posed least squares problems that eliminate or dampen the impact of small singular values on the least squares solution. In order to effectively apply these approaches, truncation and regularization parameters must be selected so that the resulting solutions are not overtly impacted by perturbations in the matrix equation. To this end, we have outlined Hansen's development of the Discrete Picard Condition and accompanying results that allow for systematic determination of the appropriate truncation parameter. This in turn allows for systematic determination of a corresponding regularization parameter, with the resulting truncated and regularized solutions being similar. This approach has been effectively applied to optimization of dihedral parameters in genotoxic PAH-DNA adducts that results in MM PESs that fit target QM PESs with chemical accuracy. As the TSVD and accompanying truncated solutions can be calculated as efficiently as Tikhonov regularized solutions in standard form, and because the truncation parameter can be used to determine the regularization parameter, the TSVD is an effective approach to ill-posed least squares problems that arise in force field parameter optimization.

CHAPTER 4

Free Energies of Binding and Formation of the Productive Complex of PAH-DNA Adducts in the NRAS(Q61) Sequence Context

4.1 Methods

4.1.1 Free Energy Calculations on Closed Thermodynamic Cycles

The Gibbs free energy is defined as :

$$G = H - TS \tag{4.1}$$

where H is the enthalpy, T the temperature, and S the entropy of the system. The enthalpy is defined as $H = U + PV$, where U is the internal energy, P the pressure, and V the system volume.[79] The free energy is the minimum amount of energy required to drive an uphill process that is gradual enough so that the system is constantly in equilibrium with its surroundings. In the absence of an external energy input, systems evolve to their lowest free energy state.[79]

Being that the probability of finding a system in a state X is proportional to its Boltzmann factor:

$$p(X) \propto e^{G(X)/k_B T} \tag{4.2}$$

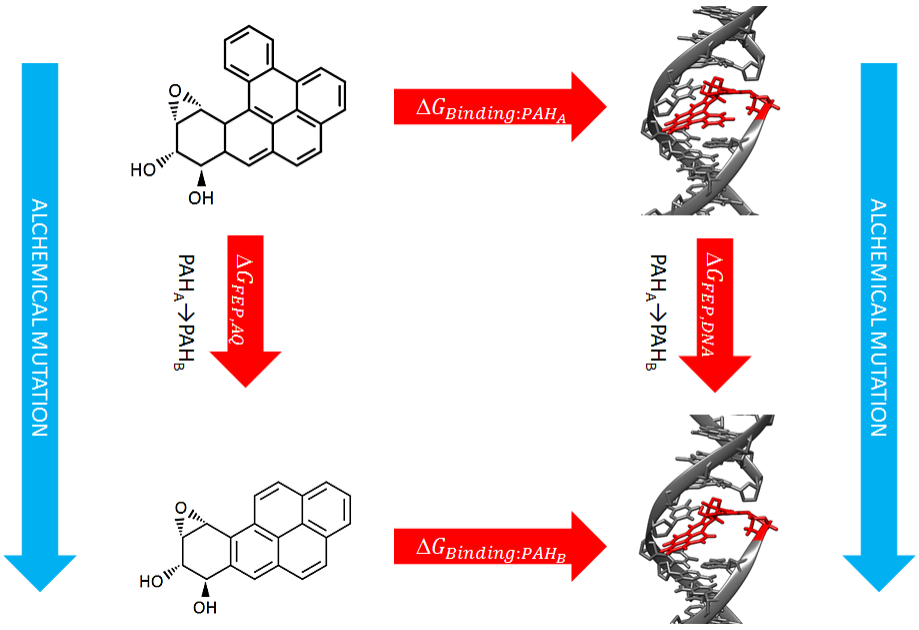
where k_B is the Boltzmann constant, the free energy difference between an initial unbound state A and a final bound state B of a given system can be estimated by examining the ratio of their

probabilities:[79,80]

$$\frac{p(B)}{p(A)} = e^{-\frac{G(B)-G(A)}{k_B T}} = e^{-\frac{\Delta G}{k_B T}}. \tag{4.3}$$

While this is a mathematically convenient expression, calculating the absolute free energies $G(A)$ and $G(B)$ in systems consisting of biological macromolecules is largely intractable due to the size of the systems of interest and the challenges posed by quasi-nonergodicity, where such systems may be formally ergodic but computational simulations of such systems do not properly sample phase space.[79,80] This may result in calculated statistical averages being strongly dependent on initial simulation conditions, yielding inaccurate results. Alternative simulation approaches designed to calculate the change in free energy due to binding ($\Delta G_{Binding}$) such as umbrella sampling are often hampered by the need to sample the entire binding and unbinding processes along a reaction coordinate while utilizing biasing potentials to overcome potential barriers.[79,80] In order to overcome these challenges, we employ the alchemical free energy perturbation (FEP) approach over closed thermodynamic cycles to calculate the relative free energy of binding ($\Delta\Delta G_{Binding}$) and the relative free energy of formation of the productive complex ($\Delta\Delta G_{Repair}$) for a given pair of PAHs. These alchemical FEP calculations have the advantage of being computationally tractable because the alchemical transformation is from PAH_A to PAH_B which requires a much smaller perturbation of the system than approaches such as umbrella sampling.[79,80] In order to illustrate the approach, we utilize DB[a,l]P and B[a]P as an example pair. The implementation of the associated alchemical FEP calculations are discussed in the sections to follow.

To calculate the relative free energy of binding of a (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA* adduct (PAH_A) as compared to a (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA* adduct (PAH_B) in the NRAS(Q61) sequence context, we employ a closed thermodynamic cycle where two alchemical FEP calculations are performed (Figure 4.1). The first is an alchemical transformation from (11R,12S,13S,14R)-DB[a,l]P-DE $\rightarrow$ (7R,8S,9S,10R)-B[a]P-DE in solution that yields $\Delta G_{FEP,AQ:PAH_A \rightarrow PAH_B}$ (Figure 4.1 - left leg). The second is an alchemical transformation from (11R,12S,13R,14S)-trans-anti-DB[a,l]P-DE-N6-dA* $\rightarrow$ (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA* in a NRAS(Q61) centered DNA 11-mer that yields $\Delta G_{FEP,DNA:PAH_A \rightarrow PAH_B}$ (Figure 4.1 - right



(4.1) Closed thermodynamic cycle examining $\Delta\Delta G_{Binding}$

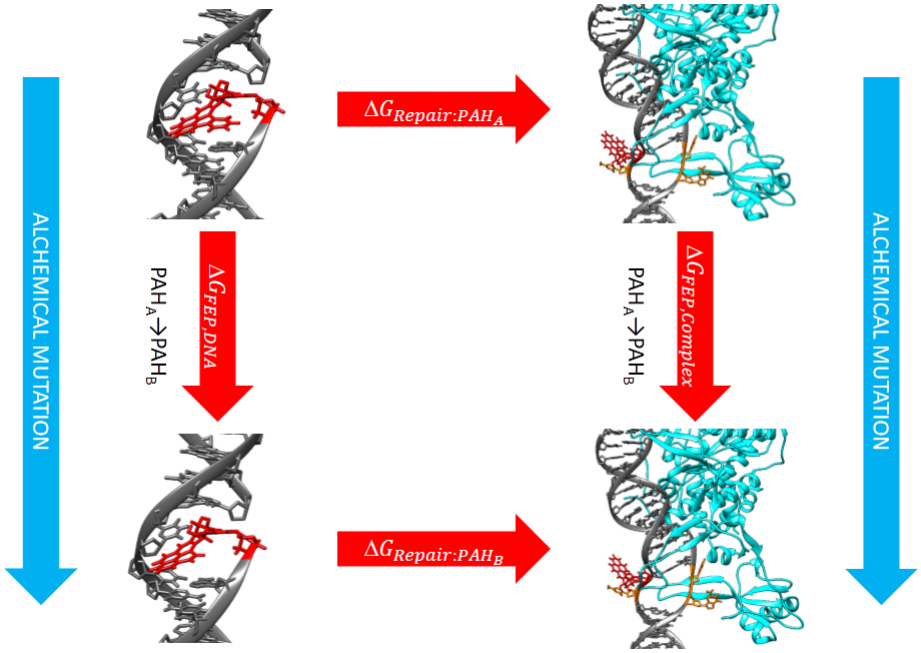
leg). Note that the changes in free energy of DB[a,l]P-DE forming a DB[a,l]P-DNA adduct ($\Delta G_{Binding:PAH_A}$ - Figure 4.1 - top leg) and B[a]P-DE forming a B[a]P-DNA adduct ($\Delta G_{Binding:PAH_B}$ - Figure 4.1 - bottom leg) are not calculated. Because free energy is a state function and thus independent of path, we have that:

$$\Delta G_{Binding:PAH_A} + \Delta G_{FEP,DNA:PAH_A \rightarrow PAH_B} = \Delta G_{FEP,AQ:PAH_A \rightarrow PAH_B} + \Delta G_{Binding:PAH_B} \tag{4.4}$$

and hence that:

$$\begin{aligned} \Delta\Delta G_{Binding} =& \Delta G_{Binding:PAH_B} - \Delta G_{Binding:PAH_A} \\ =& \Delta G_{FEP,DNA:PAH_A \rightarrow PAH_B} - \Delta G_{FEP,AQ:PAH_A \rightarrow PAH_B} \end{aligned} \tag{4.5}$$

Although we have not directly calculated $\Delta G_{Binding:PAH_A}$ and $\Delta G_{Binding:PAH_B}$, we are able to calculate the relative free energy of binding $\Delta\Delta G_{Binding}$ using values obtained from the alchemical FEP calculations in order to determine whether PAH_A or PAH_B is more likely to form a PAH-DNA adduct.



(4.2) Closed thermodynamic cycle examining $\Delta\Delta G_{Repair}$

In order to calculate the relative free energy of a DB[a,l]P-DNA adduct (PAH_A) forming the corresponding productive complex as compared to a B[a]P-DNA adduct (PAH_B), we employ a closed thermodynamic cycle analogous to the one described above. The first is the alchemical mutation described above from DB[a,l]P $\rightarrow$ B[a]P in an NRAS(Q61) centered DNA 11-mer that yields $\Delta G_{FEP,DNA:PAH_A \rightarrow PAH_B}$ (Figure 4.2 - left leg). The second is an alchemical mutation from DB[a,l]P $\rightarrow$ B[a]P in the productive complex that yields $\Delta G_{FEP,Complex:PAH_A \rightarrow PAH_B}$ (Figure 4.2 - right leg). Note that the changes in free energy of a DB[a,l]P-DNA adduct and a B[a]P-DNA adduct binding with RAD4-RAD23 and forming the productive complex ($\Delta G_{Repair:PAH_A}$ - Figure 4.2 - top leg and $\Delta G_{Repair:PAH_B}$ - Figure 4.2 - bottom leg, respectively) are not calculated. From this thermodynamic cycle, we have that:

$$\Delta G_{Repair:PAH_A} + \Delta G_{FEP,Complex:PAH_A \rightarrow PAH_B} = \Delta G_{FEP,DNA:PAH_A \rightarrow PAH_B} + \Delta G_{Repair:PAH_B} \quad (4.6)$$

and hence that:

$$\begin{aligned}\Delta\Delta G_{Repair} =& \Delta G_{Repair:PAH_B} - \Delta G_{Repair:PAH_A} \\ =& \Delta G_{FEP,Complex:PAH_A \rightarrow PAH_B} - \Delta G_{FEP,DNA:PAH_A \rightarrow PAH_B}\end{aligned} \tag{4.7}$$

Although we have not directly calculated $\Delta G_{Repair:PAH_A}$ and $\Delta G_{Repair:PAH_B}$, we are able to calculate the relative free energy of formation of the productive complex $\Delta\Delta G_{Repair}$ using values obtained from the alchemical FEP calculations in order to determine whether the PAH_A-DNA adduct or the PAH_B-DNA adduct is more likely to form the productive complex.

4.1.2 Coupled Hamiltonian - Dual Topology Approach

In order to implement the alchemical FEP calculations described above, we utilize the dual topology and coupled Hamiltonian approach in NAMD.[81]

Simulations meant to simulate physiological conditions are carried out under constant temperature and pressure resulting in an NPT ensemble with a partition function defined as:

$$Q(N,P,T) = \frac{1}{h^{3N}N!}\int\int\int e^{-\beta[\mathbf{H}(\mathbf{p},\mathbf{q})+PV]}dV d\mathbf{p}d\mathbf{q} \tag{4.8}$$

where $\beta = 1/k_BT$ and $\mathbf{H}(\mathbf{p},\mathbf{q}) = K(\mathbf{p}) + U(\mathbf{q})$ is the Hamiltonian of the system consisting of the kinetic energy $K(\mathbf{p})$ as a function of the momentum vector and the potential energy $U(\mathbf{q})$ as a function of the coordinate vector. We then have that the Gibbs free energy is defined as:

$$G = -\frac{1}{\beta}ln[Q(N,P,T)]. \tag{4.9}$$

From this we can determine the difference in free energy between an initial state A and final state B utilizing the free energy perturbation approach of Zwanzig:[79,80]

$$\begin{aligned}\Delta G_{A\rightarrow B} = G_B - G_A = -\frac{1}{\beta}ln\frac{Q_B}{Q_A} =& -\frac{1}{\beta}ln\left[\frac{\int\int e^{-\beta[\mathbf{H}_B(\mathbf{p},\mathbf{q})-\mathbf{H}_A(\mathbf{p},\mathbf{q})+\mathbf{H}_A(\mathbf{p},\mathbf{q})]}d\mathbf{p}d\mathbf{q}}{\int\int e^{-\beta\mathbf{H}_A(\mathbf{p},\mathbf{q})}d\mathbf{p}d\mathbf{q}}\right] \\ =& -\frac{1}{\beta}ln\left[\frac{\int\int e^{-\beta[\mathbf{H}_B(\mathbf{p},\mathbf{q})-\mathbf{H}_A(\mathbf{p},\mathbf{q})]}e^{-\beta\mathbf{H}_A(\mathbf{p},\mathbf{q})}d\mathbf{p}d\mathbf{q}}{\int\int e^{-\beta\mathbf{H}_A(\mathbf{p},\mathbf{q})}d\mathbf{p}d\mathbf{q}}\right] \\ =& -\frac{1}{\beta}ln\left\langle e^{-\beta[\mathbf{H}_B(\mathbf{p},\mathbf{q})-\mathbf{H}_A(\mathbf{p},\mathbf{q})]}\right\rangle_A\end{aligned} \tag{4.10}$$

where $\mathbf{H}_A$ and $\mathbf{H}_B$ are the Hamiltonians for the initial state A and final state B respectively, and the quantity $-\frac{1}{\beta} ln \left\langle e^{-\beta[\mathbf{H}_B(\mathbf{p},\mathbf{q})-\mathbf{H}_A(\mathbf{p},\mathbf{q})]} \right\rangle_A$ is an ensemble average.

Note however that FEP calculations carried out with this approach will only provide accurate estimates of the free energy difference between states if the target state B is sufficiently similar to the reference state A.[79,80] Similarity of target and reference states can be gauged in terms of overlap of important regions in the phase space of each state. Important regions in phase space are those containing configurations with highly probable energies that make the largest contribution to the estimated free energy in Equation 4.10. These important regions must be sufficiently sampled in both the reference and target states to obtain an accurate estimate of the difference in free energy between the two.[79,80] If there is insufficient overlap of the phase space of the reference and target states, configurations generated in the reference state A will be high energy states with low probability when evaluated using the Hamiltonian of the target state $\mathbf{H}_B$ and will thus make a minimal contribution to the ensemble average in Equation 4.10 leading to in accurate results.[79,80]

In order to over come this obstacle, we employ the coupled Hamiltonian-dual topology approach[79–81] where by the initial state A and final state B are defined concurrently, and the path from A to B is divided into a discrete set of N unphysical intermediate states where moieties from the states A and B incrementally fade out or fade in, exploiting the fact that free energy is a state function and thus independent of path. To this end, the parameter $\lambda_i \in [0,1]$ where $i = 0,1,...,N$ is introduced and the coupled Hamiltonian is defined as:[79–81]

$$\mathbf{H}_{\lambda_i} = (1-\lambda_i)\mathbf{H}_A + \lambda_i \mathbf{H}_B \tag{4.11}$$

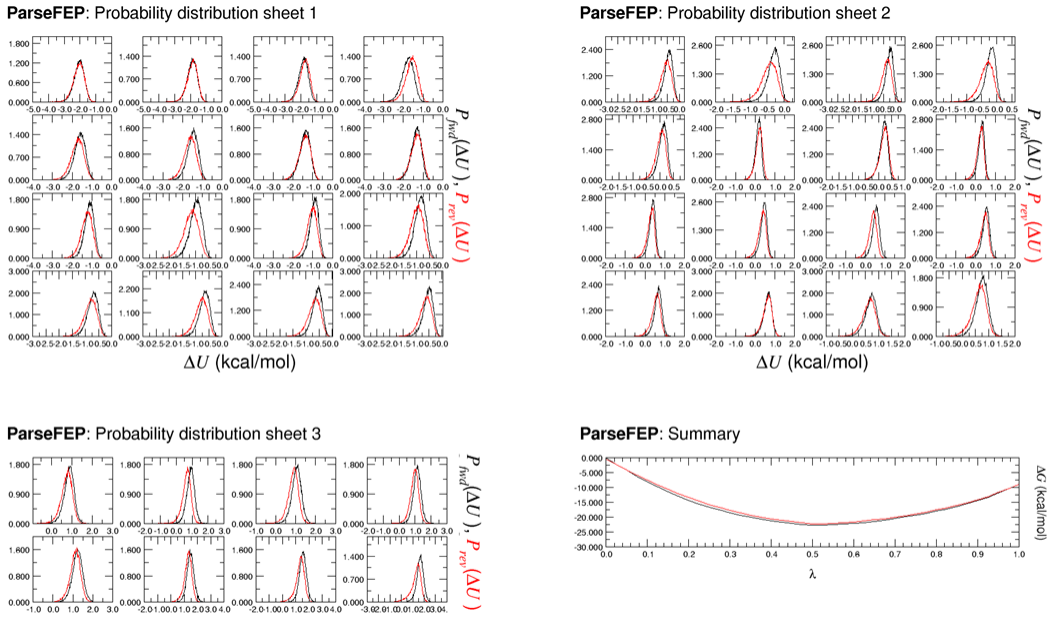
For $\lambda_0 = 0$ we have $\mathbf{H}_{\lambda_0} = \mathbf{H}_A$ and for $\lambda_N = 1$ we have $\mathbf{H}_{\lambda_N} = \mathbf{H}_B$. For $\lambda_i \in (0,1)$, the system topology is in an unphysical intermediate state between A and B, hence the term "alchemical transformation" from the initial state A to the final state B. We have then that the free energy difference in Equation 4.10 can be estimated as the sum of the free energy differences between intermediate states λ_i and λ_{i+1}:

$$\Delta G_{A\to B} = -\frac{1}{\beta} \sum_{i=0}^{N-1} ln \left\langle e^{-\beta[\mathbf{H}_{\lambda_{i+1}}(\mathbf{p},\mathbf{q})-\mathbf{H}_{\lambda_i}(\mathbf{p},\mathbf{q})]} \right\rangle_i \tag{4.12}$$

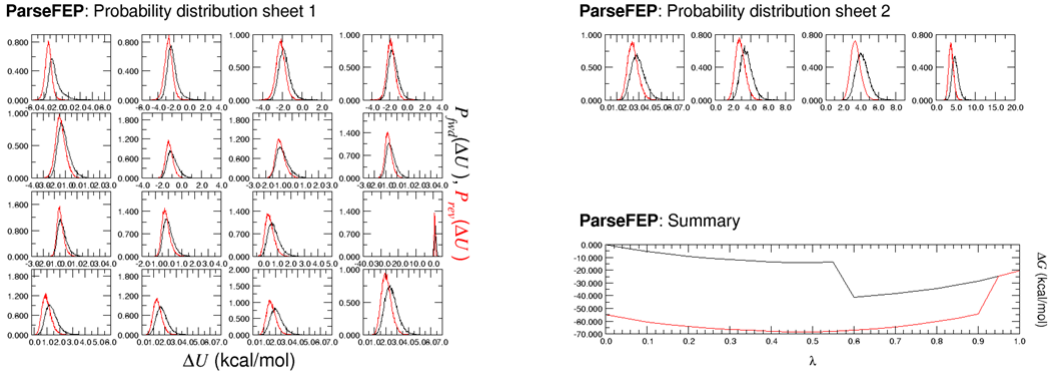
Equilibrium sampling in the reference state λ_i is carried out, and for each configuration, the energy is evaluated using the Hamiltonian $\mathbf{H}_{\lambda_i}$ and then evaluated again using the Hamiltonian $\mathbf{H}_{\lambda_{i+1}}$. For each configuration, the energy difference is evaluated, and a corresponding ensemble average is computed to estimate the free energy.[79–81] This is carried out for both the forward alchemical transformation $A \rightarrow B$ and the backward alchemical transformation $B \rightarrow A$ so that individual states serve as both reference and target states.[79–81]

As described by Liu et al.[82] simulations for which phase space overlap is insufficient, thus resulting in inaccurate free energy calculations, can be identified graphically by examining plots of the probability distributions characterizing the forward and backward alchemical FEP transformations utilizing the ParseFEP[82] VMD plugin. Figure 4.3 depicts plots of the forward ($P_{fwd}(\Delta U)$) and backward ($P_{rev}(\Delta U)$) probability distributions and the change in free energy as a function of λ_i for the alchemical FEP calculation transforming (11R,12S,13R,14S)-trans-DB[a,l]P-DE-N6-dA* $\rightarrow$ (7R,8S,9R,10S)-trans-B[a]P-DE-N6-dA* over 40 λ windows where phase space overlap is sufficient to ensure that low energy configurations of target states are also low energy configurations of reference states and that the FEP calculation has converged. Figure 4.4 depicts plots of the same alchemical FEP calculation conducted over 20 λ windows where phase space overlap is insufficient and the resultant free energy calculation is not converged. Forward transformations are plotted in black and backward transformations are plotted in red.

In the dual topology approach, both the initial state A and final state B are defined concurrently with atoms from A fading out over the discrete set of windows $\{\lambda_i | i = 0, 1, ..., N\}$ while atoms from B fade in. This creates the possibility of so called "end point catastrophes"[79,81,83] near the initial and final states where small inter-atomic distances can manifest as moieties fade in. Because non-bonded interactions are described by Coulomb and Lennard-Jones potentials in the CHARMM molecular mechanics force field, interatomic distances close to zero will cause the corresponding potential to be extremely large, leading to numerical instabilities in the simulation.[79,81,83] This is addressed in NAMD by replacing the standard Coulomb and Lennard-Jones potentials in the force field with a soft-core potential for moieties that are alchemically fading in or out.[81] The soft-core



(4.3) Forward ($P_{fwd}(\Delta U)$) and backward ($P_{rev}(\Delta U)$) probability distributions for the alchemical FEP calculation transforming (11R,12S,13R,14S)-trans-DB[a,l]P-DE-N6-dA* → (7R,8S,9R,10S)-trans-B[a]P-DE-N6-dA* over 40 λ windows depicting sufficient phase space overlap and a converged FEP calculation.



(4.4) Forward ($P_{fwd}(\Delta U)$) and backward ($P_{rev}(\Delta U)$) probability distributions for the alchemical FEP calculation transforming (11R,12S,13R,14S)-trans-DB[a,l]P-DE-N6-dA* → (7R,8S,9R,10S)-trans-B[a]P-DE-N6-dA* over 20 λ windows depicting insufficient phase space overlap and a FEP calculation that has not converged.

potential has the general form:

$$V_{ij}(r_ij) = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r(\alpha_Q(1-\lambda)+r_{ij}^p)^{1/p}} + 4\lambda\varepsilon_{ij}\left(\frac{1}{[\alpha_{LJ}(1-\lambda)+(r_{ij}/\sigma_{ij})^s]^{12/s}} - \frac{1}{[\alpha_{LJ}(1-\lambda)+(r_{ij}/\sigma_{ij})^s]^{6/s}}\right) \quad (4.13)$$

where q_i and q_j are partial atomic charges, ε_0 is the dielectric constant in vacuum, ε_r is the relative dielectric constant, r_{ij} is the interatomic distance, $\alpha_Q, \alpha_{LJ} \in \mathbb{R}$ and $p, s \in \mathbb{Z}$ are constant parameters described by Beutler et al.[83]

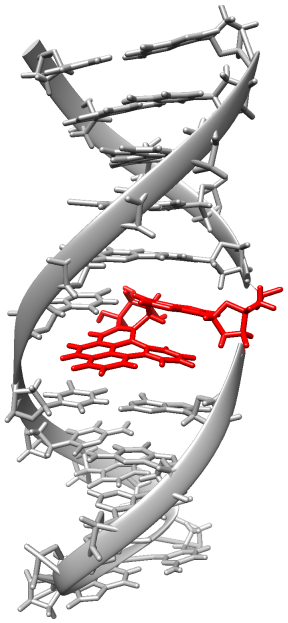
4.1.3 Dual Topology Model Systems

The dual topology PAH-DNA adduct systems examined in this work are based on the x-ray crystal structure of a bay region (1S,2R,3S,4R)-trans-anti-B[a]A-DE-N6-dA^*_6 adduct in a DNA 11-mer centered on NRAS(Q61) and containing NRAS codons 60-62, where dA^*_6 is the central adenine of NRAS(Q61) (PDB: 1DL4[84]):

5'-	C_1	G_2	G_3	A_4	C_5	A^*_6	A_7	G_8	A_9	A_{10}	G_{11}	-3'
3'-	G_{22}	C_{21}	C_{20}	T_{19}	G_{18}	T_{17}	T_{16}	C_{15}	T_{14}	T_{13}	C_{12}	-5'

This adduct is formed by trans opening of the epoxide ring of (1R,2S,3S,4R)-B[a]A-DE, resulting in intercalation from the major groove on the 3' side of dA^*_6 and it is known to be tumorigenic in mouse skin and in newborn mice.[84,85] In order to illustrate how dual topology PAH-DNA adduct systems are built, we again utilize the DB[a,l]P and B[a]P pair as an illustrative example.

The (1S,2R,3S,4R)-trans-anti-B[a]A-DE-N6-dA^*_6 adduct in the 1DL4 model system was modified into a (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA^*_6 adduct utilizing Chimera.[86] Note that the relative carbon numbering in (1R,2S,3S,4R)-B[a]A-DE is reversed from that of (7R,8S,9S,10R)-B[a]P-DE owing to the different root compounds being anthracene and pyrene, respectively, but these and all of the systems we will examine are structurally and stereochemically analogous. As a result, we will hence forth refer to all (-R-S-S-R)-PAH-DEs in solution as PAH-DE systems and all (-R,-S,-R,-S)-trans-anti-PAH-DE-N6-dA^*_6 adducts in DNA as PAH-DNA adduct systems. In order to define a dual topology residue that includes a DB[a,l]P-DNA adduct, coordinates for an additional aromatic ring were simultaneously added to the "l" side of the B[a]P-DNA adduct residue (Figure 4.5). Non-bonded clashes and contacts were identified and relieved via geometric structure editing in Chimera.[86] In order to build the dual topology residue for the analogous

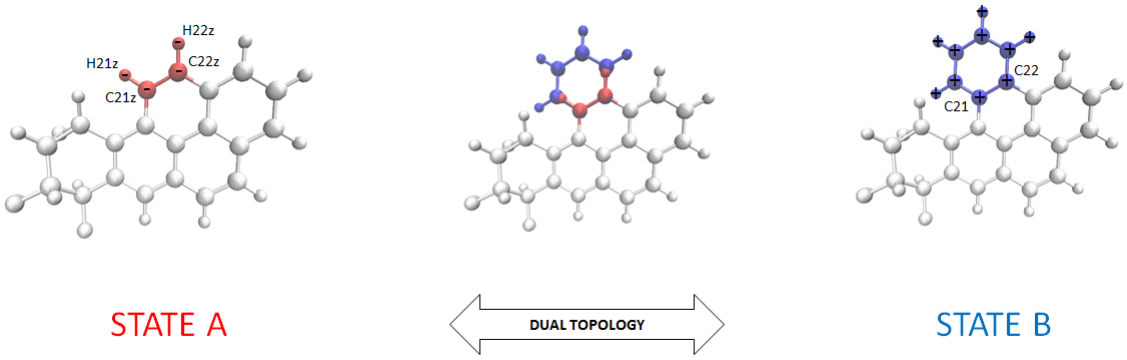


(4.5) PAH-DNA adduct system based on the x-ray crystal structure of a (1S,2R,3S,4R)-trans-anti-B[a]A-DE-N6-dA$^{*}_{6}$ adduct in the NRAS(Q61) sequence context

B[a]P/DB[a,l]P-DE system in solvent, the B[a]P/DB[a,l]P-DNA dual topology residue was removed from the NRAS(Q61) centered DNA 11-mer and edited into a dual topology residue that includes both B[a]P-DE and DB[a,l]P-DE as depicted in the center of Figure 4.6. Note that the dual topology aromatic ring structure of the B[a]P/DB[a,l]P-DNA residue is identical to that of the B[a]P/DB[a,l]P-DE residue.

In both the B[a]P/DB[a,l]P-DNA and the B[a]P/DB[a,l]P-DE dual topology residues, the carbon atoms C21z and C22z on the "l" side of the B[a]P-DE segment have a -0.115 charge that are balanced by hydrogens H21z and H22z with a +0.115 charge (Figure 4.6 - left - State A) where as in the DB[a,l]P-DE segment, the analogous C21 and C22 carbons have a 0.000 charge in order to maintain partial charge balance in the system (Figure 4.6 - right - State B). This approach is based on the partial charges assigned to analogous carbons and hydrogens in the CGenFF residues for naphthalene and anthracene.[40] Coordinates for moieties fading in and those fading out of the dual topology residue are defined simultaneously in the system's PDB file. Atoms from the initial

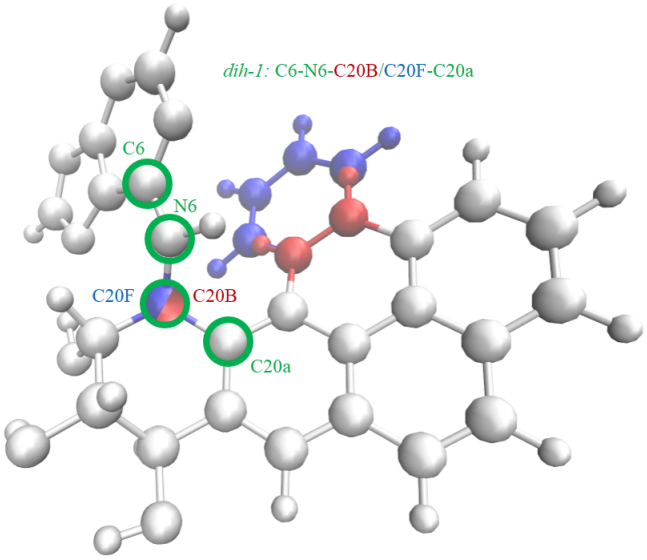
state *A* that fade out are tagged with a -1.00 in the B column of the PDB file while atoms from the final state *B* that fade in are tagged with a 1.00 in the B column of the PDB file. Note that NAMD creates an exclusion list of moieties that are fading in and out, and they do not interact with one another during molecular dynamics.[81] In the case of B[a]P-DE alchemically mutating into DB[a,l]P-DE as depicted in Figure 4.6, moieties in the initial state *A* shaded in red and labeled with a $-$ sign fade out (Figure 4.6 - left - State A) while moieties in the final state *B* shaded in blue and labeled with a $+$ sign fade in (Figure 4.6 - right - State B). Corresponding CHARMM compatible topology files defining atom names, atom types, partial charges, and bonds were developed for the B[a]P/DB[a,l]P-DNA and the B[a]P/DB[a,l]P-DE dual topology residues.



(4.6) Dual topology B[a]P/DB[a,l]P-DE residue: red atoms marked with a (-) in the initial state A fade out while blue atoms marked with a (+) in the final state B fade in.

Note that the alchemical mutation from the B[a]P-DNA adduct to the DB[a,l]P-DNA adduct requires a special case since it is a mutation from a bay PAH to a fjord PAH. As we have shown previously in Chapter 2, bay and fjord region PAH-DNA adducts require distinct sets of terms for dihedral parameters involving the adduct covalent bond between N6 in the exocyclic-amino group of adenine and the C20B/C20F aliphatic carbon in the PAH as depicted in Figure 4.7. In order to implement this alchemical mutation with the appropriate dihedral parameters, we utilize the additional atom types CG311B and CG311F for atoms C20B and C20F respectively. When alchemically transforming from a bay to fjord PAH-DNA adduct, atom C20B fades out while atom C20F fades in. Correspondingly, atoms H20B and H20F are utilized to maintain partial

charge balance during the alchemical mutation. In this way, the CG311B atom type is used to parameterize dihedrals that include the adduct covalent bond in bay systems while the CG311F atom type is used to parameterize the same dihedrals in fjord systems. This allows for distinct sets of dihedral terms to be utilized during alchemical transformations from bay to/from fjord region PAH-DNA adduct systems as shown in Table 4.1.



(4.7) Bay to fjord dual topology B[a]P/DB[a,l]P-DNA residue: atom C20B / C20F with corresponding atom types CG311B / CG311F are used to parameterize systems that alchemically transform between bay and fjord region PAHs

	Atom Names	Atom Types	n	k_n(kcal/mol)	δ_n
Bay	C6-N6-*C20B*-C20a	CN2-NN1-CG311B-CG2R61	1	2.507	180.00°
			3	0.047	0.00°
Fjord	C6-N6-*C20F*-C20a	CN2-NN1-CG311F-CG2R61	1	3.00	180.00°
			3	0.396	180.00°

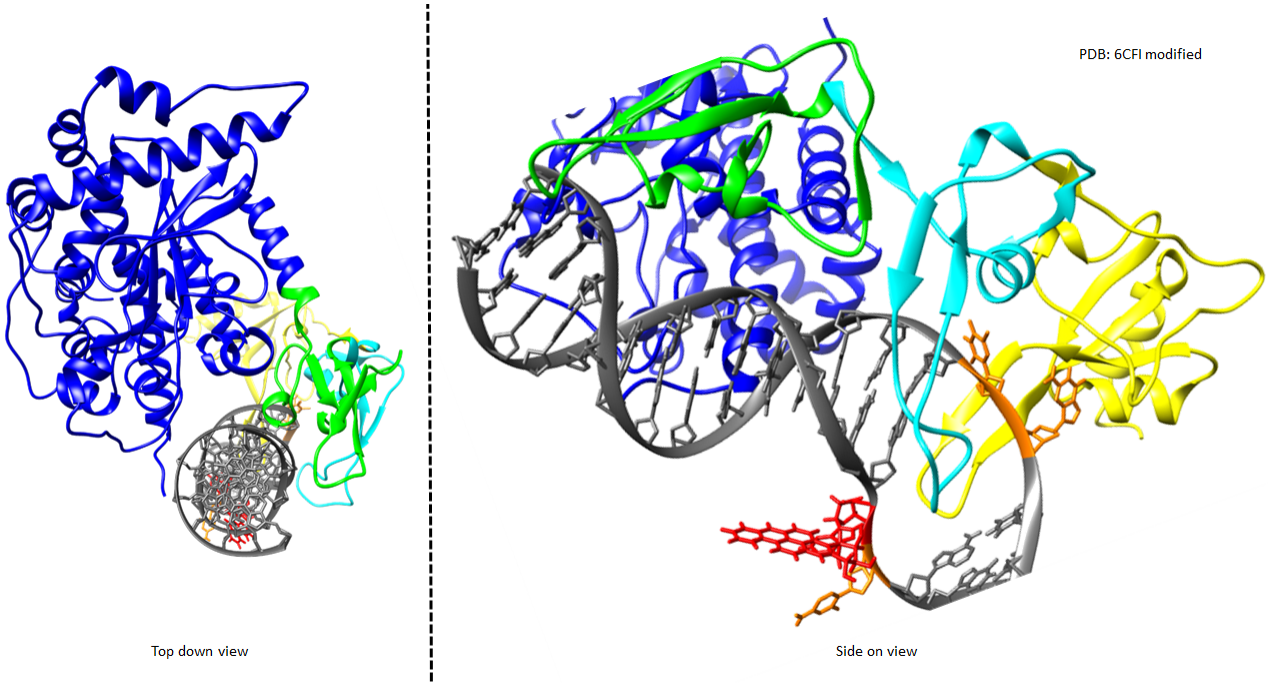
Table (4.1) Distinct dihedral terms for the $dih-1$ parameter in bay and fjord systems

The dual topology productive complex systems are based on the x-ray crystal structure of RAD4-RAD23 bound to a UV induced 6-4 photoproduct in DNA (PDB: 6CFI[87]), using an approach similar to that of Mu et al.[53] The DNA fragment was modified into a 23-mer containing the NRAS(Q61) centered 11-mer with a PAH-DNA adduct (highlighted in blue below), while the rest of the DNA sequence is retained from the 6CFI system:

5'-	T_1	T_2	G_3	C_4	G_5	G_6	A_7	C_8	A^*_9	A_{10}	G_{11}	A_{12}	A_{13}	G_{14}	G_{15}	T_{16}	T_{17}	G_{18}	A_{19}	G_{20}	T_{21}	C_{22}	A_{23}	-3'
3'-	A_{46}	A_{45}	C_{44}	G_{43}	C_{42}	C_{41}	T_{40}	G_{39}	T_{38}	T_{37}	C_{36}	T_{35}	T_{34}	C_{33}	C_{32}	A_{31}	A_{30}	C_{29}	T_{28}	C_{27}	A_{26}	G_{25}	T_{24}	-5'

The structure of the RAD4-RAD23 protein consists of a transglutaminase-homology domain (TGD, blue in Figure 4.8) and three beta hairpin domains (BHD1-3, green, cyan, and yellow respectively in Figure 4.7).[15,28,53,87] The productive complex bound to DNA containing a 6-4 photoproduct is characterized by the TGD and BHD1 domains forming a clamp like structure as seen in the left panel of Figure 4.8. The two nucleotides that comprise the thymine photodimer are extruded from the DNA duplex and exposed to solvent while their partner adenines are also extruded and bound by BHD2 and BHD3.[87] The BHD2 domain contacts the minor groove and the BHD3 domain inserts via the major groove and occupies the space left in the DNA duplex by the extruded nucleotides, as seen in the right panel of Figure 4.8.[15,53,87] Analogously, the productive complex bound to a PAH-DNA adduct system designed for this work consists of the adduct containing dA^*_9 (red in Figure 4.8) and its 5' neighboring dC_8 (orange in Figure 4.8) being extruded from the hydrophobic core of the DNA duplex and exposed to solvent while their partner dG_{39} and dT_{38} nucleobases in the complementary strand (orange in Figure 4.8) are also extruded and bound by the BHD2 and BHD3 domains. The TGD and BHD1 domains of the RAD4-RAD23 protein, along with the nucleotides highlighted in red above are unmodified from the original 6CFI system and are frozen during molecular dynamics. The dual topology B[a]P/DB[a,l]P-DNA residue in the productive complex is identical to that in the PAH-DNA adduct system.

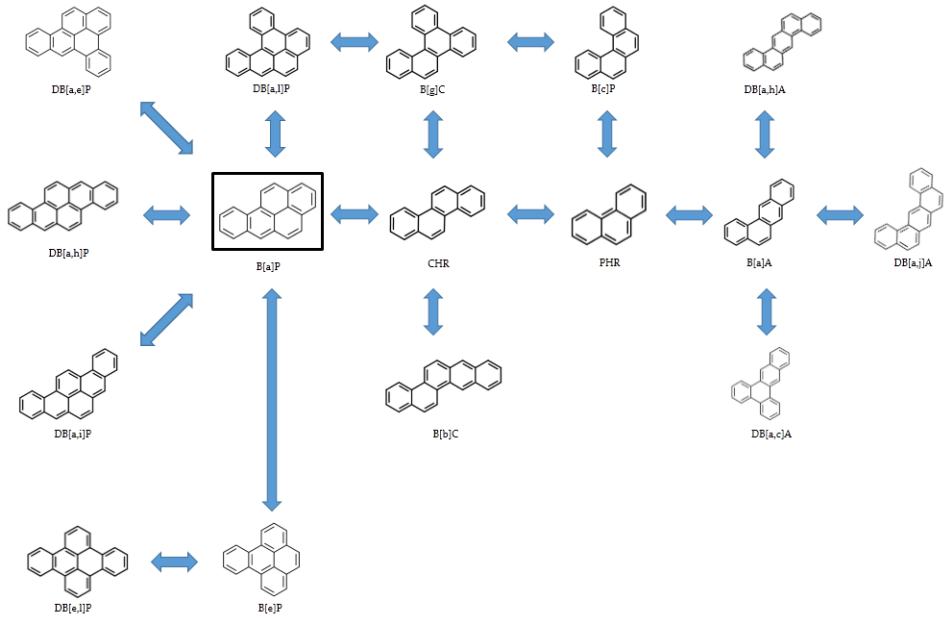
In addition to the B[a]P/DB[a,l]P-DE and B[a]P/DB[a,l]P-DNA dual topology residues described above, analogous dual topology residues were developed for each pair of PAHs connected by double headed arrows depicted in Figure 4.9. These PAHs are also listed in Table 4.2 along with



(4.8) PAH-DNA adduct in the productive complex based on the x-ray crystal structure of RAD4-RAD23 in complex with a UV induced 6-4 photoproduct in DNA

their IARC grouping and full chemical name. This collection of dual-topology residues allows us to examine closed thermodynamic cycles analogous to those depicted in Figures 4.1 and 4.2 for each pair of PAHs. We are then able to quantify relative free energies of binding and relative free energies of formation of the analogous productive complex for each pair of PAHs.

Note that with the exception of the B[a]P/B[e]P pair, each pair of PAHs in Figure 4.9 differs by one aromatic ring. More drastic ring topology transformations are avoided in order to minimize the likelihood of insufficient phase space overlap and thermodynamic cycle closure errors during alchemical FEP calculations. As described by Gapsys et al.,[88] a key assumption in the application of alchemical FEP calculations over closed thermodynamic cycles is that the free energy contribution from moieties alchemically fading in/out are identical in both of the alchemical FEP legs of the thermodynamic cycle. It has been noted that alchemical FEP simulations involving ring-topology transformations may be hampered if conformational distortion between states occurs,



(4.9) Alchemical FEP calculations over closed thermodynamic cycles are carried out for the PAH pairs connected by double headed arrows

and that such ring-topology transformations are the subject of current research.[89,90] Liu et al.[90] showed that when utilizing a *single*-topology approach in relative binding free energy calculations, thermodynamic cycle closure errors can manifest as a result of such conformational distortions in non-aromatic rings. It was noted that such errors should not manifest in *dual*-topology approaches such as that used in this work, because the ring-topology transformations are implemented without utilizing multiply connected dummy atoms interacting with remaining atoms.[89,90] Jiang et al.[89] conducted hybrid single-dual-topology alchemical FEP simulations of a myeloid cell leukemia 1 (MCL1) protein ligand system involving a six-membered aromatic ring extension of the ligand utilizing NAMD and found that the extended ring maintained its topology and that the approach resulted in a calculated relative binding free energy within chemical accuracy as compared to experiment and previous simulations. We also note the importance of addressing thermodynamic cycle closure concerns of the type described by Gapsys et al.[88] regarding the use of softcore potentials and the NAMD alchDecouple 'off' option. In binding free energy calculations of a DNA-binding protein that involve alchemical mutations of the DNA sequence, thermodynamic cy-

cle closure errors manifested due to the retention of non-bonded interactions between nucleobase pairs that are alchemical fading in and out.[88] In conjunction with standard bonded interactions between the moieties that are alchemical fading in/out and regular atoms in the remainder of the system, non-negligible free energy contributions resulting from different conformational ensembles in the DNA-only and DNA-protein systems that do not cancel were observed.[88] Noting that PAH-DNA adducts consist of rigid aromatic rings in the PAH that aromatically stack with neighboring nucleobases, these systems are not likely to undergo conformational distortions that would result in thermodynamic cycle closure errors such as those described above in systems involving alchemically mutating non-aromatic rings or nucleotides in DNA with freely rotating glycosidic bonds. We thus proceed with the alchemical FEP calculations over the closed thermodynamic cycles described above with an eye toward validating this approach for PAH-DNA adduct systems in this and future work.

Polycyclic Aromatic Hydrocarbons					
BAY			FJORD		
Name	Abbreviation	IARC Group	Name	Abbreviation	IARC Group
phenanthrene	PHE	3	benzo[c]phenanthrene	B[c]P	2B
chrysene	CHR	2B	benzo[g]chrysene	B[g]C	3
benzo[a]pyrene	B[a]P	1	dibenzo[a,l]pyrene	DB[a,l]P	2A
benzo[e]pyrene	B[e]P	3			
dibenzo[a,h]pyrene	DB[a,h]P	2B			
dibenzo[a,i]pyrene	DB[a,i]P	2B			
dibenzo[a,e]pyrene	DB[a,e]P	3			
dibenzo[e,l]pyrene	DB[e,l]P	3			
benz[a]anthracene	B[a]A	2B			
dibenz[a,c]anthracene	DB[a,c]A	3			
dibenz[a,h]anthracene	DB[a,h]A	2A			
dibenz[a,j]anthracene	DB[a,j]A	3			
benzo[b]chrysene	B[b]C	3			

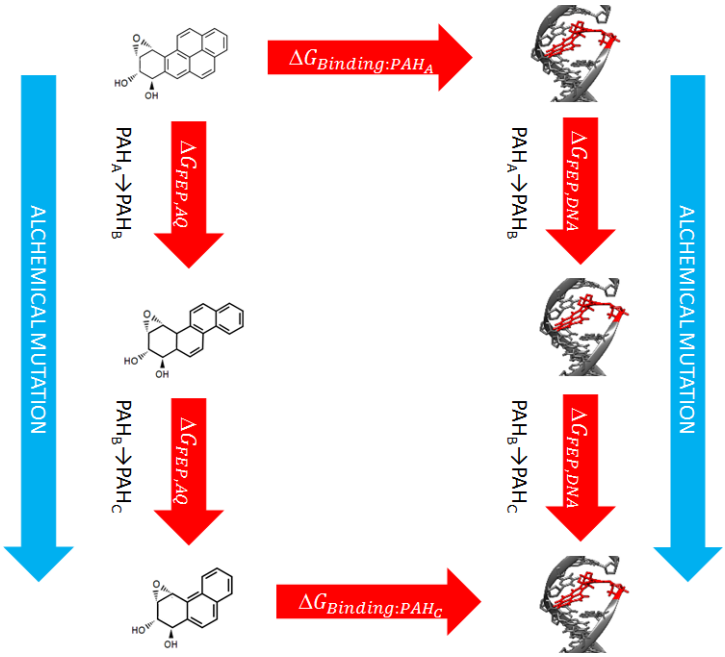
Table (4.2) PAHs examined in this work, their abbreviations, and their IARC Grouping.

In order to examine the relative genotoxicity at the central dA* in NRAS(Q61) of the PAHs examined in this work as compared to the IARC Group 1 known human carcinogen B[a]P, we *estimate* the relative binding free energy as compared to B[a]P of PAHs that differ from B[a]P by

more than one aromatic ring (i.e. PAHs other than those connected to B[a]P by one double headed arrow in Figure 4.7) by concatenating closed thermodynamic cycles as depicted in Figure 4.11 where:

$$\begin{aligned}\Delta\Delta G_{Binding} =& \Delta G_{Binding:PAH_C} - \Delta G_{Binding:PAH_A} \\ =& \left(\Delta G_{FEP,DNA:PAH_A \rightarrow PAH_B} + \Delta G_{FEP,DNA:PAH_B \rightarrow PAH_C}\right) - \\ & \left(\Delta G_{FEP,AQ:PAH_A \rightarrow PAH_B} + \Delta G_{FEP,AQ:PAH_B \rightarrow PAH_C}\right)\end{aligned} \tag{4.14}$$

For example, the relative binding free energy of stereochemically analogous B[a]A as compared to B[a]P is *estimated* by concatenating the closed thermodynamic cycles between B[a]P $\leftrightarrow$ CHR $\leftrightarrow$ PHE $\leftrightarrow$ B[a]A. The relative free energy of formation of the corresponding productive complex is obtained by concatenating the appropriate analogs of these closed thermodynamic cycles. With thermodynamic cycle closures in mind, we emphasize that these are estimates of relative free energy differences and that this approach will be further validated in future work examining various approaches to PAH ring topology transformations.



(4.10) Concatenated closed thermodynamic cycles

4.1.4 Computational Methods

Molecular dynamics simulations were conducted in NAMD 2.14 and GPU accelerated NAMD 3.0[56,57] utilizing the CHARMM36-NA[50,51,91] and CGenFF 4.4[40] force fields along with custom topology/parameter files for the dual-topology PAH-DE and PAH-DNA adduct residues and accompanying parameters. Na^+ counter ions were placed in the PAH-DNA adduct and corresponding productive complex systems using the CIonize VMD plugin.[62] The PAH-DE systems were solvated in a 40 Å TIP3P[63] explicit water box, the PAH-DNA adduct systems were solvated in a 70 Å box, and the productive complex systems were solvated in a [95x120x115]Å box, ensuring that the solvent extends at least 15 Å beyond the solute in each system. Na^+ and Cl^- ions were added to all systems using the Solvate and Autoionize plugins in VMD to achieve a 100mM NaCl solution.

With the dual topology residue in the initial state A, the solvated PAH-DNA adduct and productive complex systems were relaxed over 1000 steps of conjugate-gradient minimization with harmonic constraints applied to nucleobases in the DNA duplex. This was followed by 1 ns of NVT simulation and 5 ns of NPT simulation with harmonic constraints remaining in place to avoid the system volume changing too rapidly at the outset of the NPT equilibration simulation. Harmonic restraints were then released and an extra bond added between the dG-N1 and dC-N3 nitrogens in the terminal dG-dC nucleobase pairs of the PAH-DNA adduct system and between the dA-N1 and dT-N3 nitrogens in the terminal dA-dT base pairs of the productive complex systems in order to avoid end fraying. Production NPT MD simulations were run for 100 ns in the PAH-DNA and productive complex systems with periodic boundary conditions at 300 K and 1 atm utilizing Langevin dynamics and the Langevin piston.[64] Electrostatic interactions were treated utilizing the Particle Mesh Ewald[65] method with a cutoff of 12 Å. Lennard-Jones interactions were treated by activating the switching function at 10 Å. RigidBonds was set to all[66,67] in order to utilize a 2 fs time step. In the PAH-DE systems, NPT production simulations were run for 50 ns with the dual topology residue in the initial state A.

Dihedral parameters for phosphate linkages in the DNA backbone that were modified for the CHARMM36-NA force field were reverted to their CHARMM27-NA values. As described by Minhas et. al.,[68] these dihedral parameters were modified in CHARMM36-NA in order to improve BI/BII conformational sampling over CHARMM27-NA. However, this causes increased flexibility of the DNA backbone that was found to result in instability of DNA on the microsecond time scale[68] which we found manifested in trial MD simulations.

Forward and backward alchemical FEP simulations followed the same MD protocols listed above and were conducted over 20 windows for all PAH-DE, PAH-DNA, and productive complex systems where the initial and final states were either both bay PAHs or both fjord PAHs. When mutating from a bay PAH to/from a fjord PAH, 40 windows were utilized for PAH-DNA and productive complex systems . Each window consisted of 200 ps of equilibration and 800 ps of production. A soft-core potential as described above was used with the NAMD van der Waals radius-shifting coefficient set to 4.0[81] so that the Lennard-Jones potential is shifted from $r^2 \rightarrow r^2 + 4(1-\lambda)$. Electrostatic interactions for moieties fading in were fully decoupled from the simulation for $\lambda \in [0.0, 0.5]$ with coupling of electrostatic interactions linearly increasing from $\lambda \in [0.5, 1.0]$. Electrostatic interactions of moieties fading in are fully coupled to the simulation at $\lambda = 1.0$. Electrostatic interactions for moieties fading out are linearly decoupled for $\lambda \in [0.0, 0.5]$. At $\lambda = 0$ electrostatics are fully coupled to the simulation and then linearly decrease as λ increases until fully decoupled for $\lambda \in [0.5, 1.0]$.[92]

The ParseFEP[82] VMD plugin was utilized to calculate free energy differences and statistical errors resulting from forward and backward alchemical free energy perturbation calculations using the Bennett Acceptance Ratio.[82,93] Sufficient phase space overlap of reference and target states was evaluated graphically following the approach described by Liu et al.[82] utilizing ParseFEP probability distribution plots from the forward and backward alchemical transformations. Enthalpy and entropy estimates were obtained from ParseFEP utilizing the approach described by Liu et al.[82]

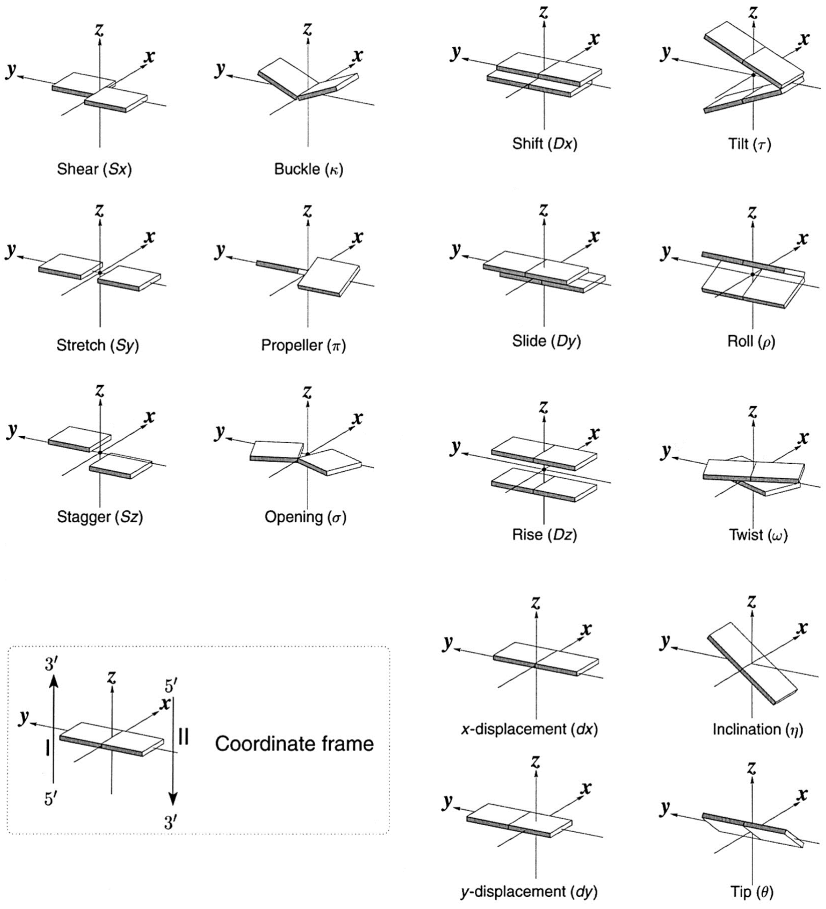
Rigid-body parameters describing the geometry of sequential DNA base steps, canonical DNA base pairs, and refined major and minor groove widths were calculated for each PAH-DNA adduct

system utilizing the x3DNA software package.[94–97] Parameters describing sequential DNA base steps include: shift, tilt, slide, roll, rise, and twist while parameters describing the geometry of canonical DNA base pairs include: shear, buckle, stretch, propeller, stagger, and opening (Figure 4.11, included with permission from the author of Ref[94]). Trajectories over the corresponding 100 ns NPT production run, or an equilibrated subset, were utilized to calculate average values and standard deviations for each rigid-body parameter. Equilibrated subsets of the 100 ns NPT production run were identified as those for which the DNA duplex RMSD, PAH RMSD, dA*$_6$-PAH α' bond angle, dA*$_6$-PAH β' bond angle, and dA*$_6$ glycosidic χ bond angle trajectories steadily fluctuate about their respective averages. Trajectory plots for each system are included in Appendix B. Average rigid-body parameters were then compared to those calculated over an analogous 100 ns NPT production run of an unmodified NRAS(Q61) centered 11-mer (i.e. no PAH-DNA adduct, and henceforth unmodified DNA). Average structures of PAH-DNA adduct systems over the corresponding 100 ns NPT production run, or an equilibrated subset, were calculated using the NAFlex webserver.[98]

The NAMD Energy Plugin in VMD was utilized to calculate average configurational (bond, angle, dihedral, and improper) and average non-bonded (electrostatic and van der Waals) energies of interest in the PAH-DNA adduct systems. Excluding the PAH moiety, the average configurational and non-bonded energies of the three canonical DNA base pairs that comprise NRAS codon 61 ($dC_5 : dG_{18}$, $dA^*_6 : dT_{17}$, and $dA_7 : dT_{16}$) were calculated over the 100 ns NPT production runs of each PAH-DNA adduct system. The strength of stabilizing π-stacking interactions from PAH-DNA adduct intercalation was quantified by calculating average van der Waals interactions over the corresponding 100 ns NPT production run between the aromatic rings of a given PAH and the neighboring nucleobases that form the primary $dT_{16} \mid dT_{17}$ and secondary $dA^*_6 \mid dA_7$ intercalation pockets described in the following sections. Total van der Waals interactions from PAH-DNA adduct intercalation is defined as the sum of van der Waals interactions between the PAH and the dT_{16} and dT_{17} nucleobases in the primary intercalation pocket and between the PAH and the dA^*_6 and dA_7 nucleobases in the secondary intercalation pocket,

i.e. $E_{vdW:Intercalation} = E_{vdW:dT_{16} \mid dT_{17}} + E_{vdW:dA^{*}_{6} \mid dA_{7}}$.

Hydrogen bond occupancies between the three canonical DNA base pairs of NRAS codon-61, were calculated utilizing a Python script measuring electronegative atom distances (e.g. dA_7-N1 : dT_{16}-N3) and donor-hydrogen-acceptor angles for each PAH-DNA adduct system over the corresponding 100 ns NPT production run. Similar to the approach implemented by Cai et al.,[36] tolerance for the donor-hydrogen-acceptor angle was set to > 140°. Four tolerances for electronegative atom distances were examined, {3.0, 3.1, 3.2, 3.3 Å} and it was found that the 3.1 Å tolerance for electronegative atom distances was most suitable for evaluating differences in hydrogen bond occupancy between systems. Percent hydrogen bond occupancies were then compared to those of unmodified DNA calculated over an analogous 100ns NPT production run.

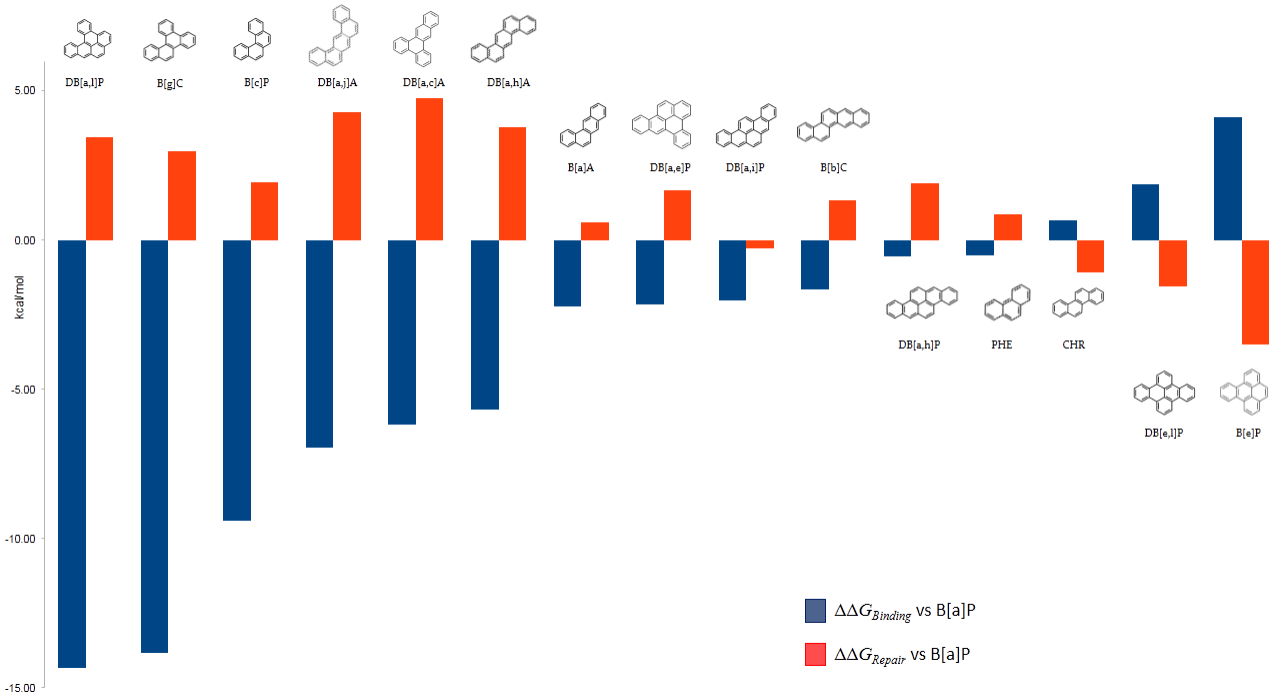


(4.11) DNA rigid-body parameters.

4.2 Results and Discussion

Alchemical FEP calculations were carried out over closed thermodynamic cycles for the PAH pairs connected by double headed arrows in Figure 4.9 as described above. Relative free energies of binding and formation of the productive complex for these pairs along with ParseFEP estimates of the corresponding enthalpy and entropy changes are listed in Tables 6.1 and 6.2 of Appendix A. ParseFEP plots demonstrating sufficient phase space overlap and convergence of alchemical FEP calculations for PAH-DE, PAH-DNA, and productive complex systems are also included in Appendix A. Relative free energies of binding ($\Delta\Delta G_{Binding}$) of (-R,-S,-R,-S)-trans-anti-PAH-DE-N6-dA$^*_6$ adducts as compared to a (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA$^*_6$ adduct were calculated for the PAHs listed in Table 4.2 by concatenating thermodynamic cycles as described above. The relative free energies of binding are plotted in Figure 4.12 (blue bars) and organized from left to right in order of decreasing relative binding affinity. The PAHs are grouped into four categories: strongly preferred ($\Delta\Delta G_{Binding}$ < -5.50 kcal/mol), weakly preferred (-2.50 kcal/mol < $\Delta\Delta G_{Binding}$ < -1.50 kcal/mol), equally preferred (-1.00 kcal/mol < $\Delta\Delta G_{Binding}$ < 1.00 kcal/mol), and non-preferred (1.50 kcal/mol < $\Delta\Delta G_{Binding}$) as compared to B[a]P. Relative free energies of formation of the corresponding productive complexes ($\Delta\Delta G_{Repair}$) as compared to B[a]P were calculated by concatenating analogous thermodynamic cycles and are also plotted in Figure 4.12 (red bars). Associated structural, energetic, and hydrogen bonding characteristics of the NRAS(Q61) DNA 11-mer are briefly outlined below and discussed in greater detail in Chapter 5.

Among the strongly preferred PAHs, the three fjord region systems: DB[a,l]P, B[g]C, and B[c]P respectively exhibit the greatest relative binding affinity as compared to B[a]P. These are followed by the bay region DB[a,j]A, DB[a,c]A, and DB[a,h]A systems respectively, which each have one additional aromatic ring on a B[a]A root compound (see Figure 4.9). Relative free energies of binding among the strongly preferred PAHs range from $\Delta\Delta G_{Binding:\mathrm{DB[a,l]P}}$ = -14.34 kcal/mol to $\Delta\Delta G_{Binding:\mathrm{DB[a,h]A}}$ = -5.68 kcal/mol. With the exception of B[c]P, the strongly preferred PAHs all assume a similar conformational motif, intercalating from the major groove with the aromatic rings



(4.12) Relative free energies of binding ($\Delta\Delta G_{Binding}$) and repair ($\Delta\Delta G_{Repair}$) of (-R,-S,-R,-S)-trans-anti-PAH-DE-N6-dA^*_6 adducts as compared to a (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA^*_6 adduct at the central dA^*_6 in NRAS(Q61)

of the PAH positioned in both a primary intercalation pocket formed by dT_{16} and dT_{17} in the complementary strand (Figure 4.13 green box) and a secondary intercalation pocket formed by dA^*_6 and dA_7 (Figure 4.13 blue box). This conformation positions the aromatic rings of the strongly preferred PAHs in an ideal position for strong stabilizing van der Waals interactions (i.e. π-stacking) with the dT_{16} and dT_{17} nucleobases in the primary intercalation pocket ($E_{vdW:dT_{16}\,|\,dT_{17}}$ in Table 4.3) and with the dA^*_6 and dA_7 nucleobases in the secondary intercalation pocket ($E_{vdW:dA^*_6\,|\,dA_7}$ in Table 4.3).

The conformational motif assumed by these strongly preferred PAHs is distinct from that assumed by the bay region B[a]P, which intercalates from the major groove with it's aromatic rings positioned solely in the primary dT_{16} | dT_{17} intercalation pocket (Figure 4.15). This conformation results in strong van der Waals interactions in the primary intercalation pocket and comparatively weak van der Waals interactions between the aromatic rings of B[a]P and the nucleobases in the

secondary dA^*_6 | dA_7 intercalation pocket (Table 4.3). The strongly preferred fjord region B[c]P, which has the weakest relative binding affinity as compared to B[a]P among the fjord region PAHs, assumes a conformation similar to B[a]P, with it's aromatic rings - including the fjord aromatic ring - positioned solely in the primary intercalation pocket and exhibiting strong van der Waals interactions with dT_{16} and dT_{17} while van der Waals interactions with dA^*_6 and dA_7 in the secondary intercalation pocket are weak (Table 4.3).

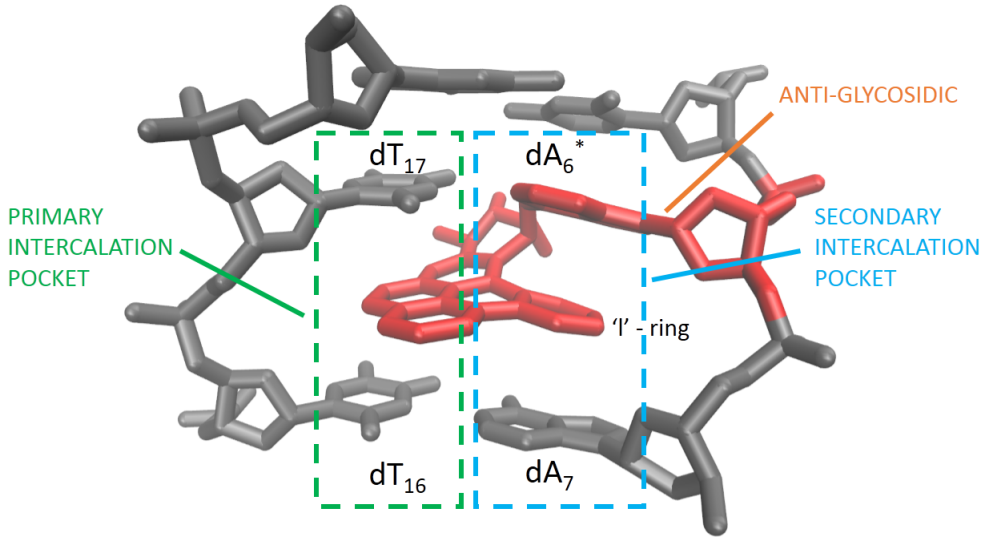
From this it is evident that greater relative binding affinity is generally associated with stronger total van der Waals interactions from PAH-DNA adduct intercalation:

$$E_{vdW:Intercalation} = E_{vdW:dT_{16} \mid dT_{17}} + E_{vdW:dA^*_6 \mid dA_7}$$

These values are listed in Table 4.3 and plotted in Figure 4.14 (green bars) along with the relative free energies of binding (blue bars). Stronger total van der Waals interactions from intercalation are in turn generally associated with the number and orientation of aromatic rings in the PAH that are effectively positioned for non-bonded interactions with the nucleobases in the primary and secondary intercalation pockets.

This is illustrated by noting that each of the fjord region PAHs exhibits greater relative binding affinity than its bay region analog (Table 6.1 of Appendix A). For example, the flexible and non-planar fjord aromatic ring of the DB[a,l]P-DNA adduct is positioned for effective π-stacking in the secondary intercalation pocket while the remaining aromatic rings of the B[a]P root compound are also positioned for effective π in the primary intercalation pocket (Figure 4.13). In contrast, the rigid and planar bay region B[a]P-DNA adduct with one less aromatic ring lacks the structure to effectively π-stack in the secondary intercalation pocket, limiting stabilizing van der Waals interactions to the primary intercalation pocket (Figure 4.15). As a result, total van der Waals interactions from intercalation are stronger in the DB[a,l]P system with $E_{vdW:Intercalation,DB[a,l]P}$ = -27.49 kcal/mol than the B[a]P system with $E_{vdW:Intercalation,B[a]P}$ = -15.73 kcal/mol. Similarly, the flexible and non-planar fjord aromatic ring of the B[c]P-DNA adduct is positioned for more effective π-stacking in the primary intercalation pocket than its rigid and planar bay region PHE-DNA analog with one less aromatic ring. Analogous comparisons can also be drawn between the fjord

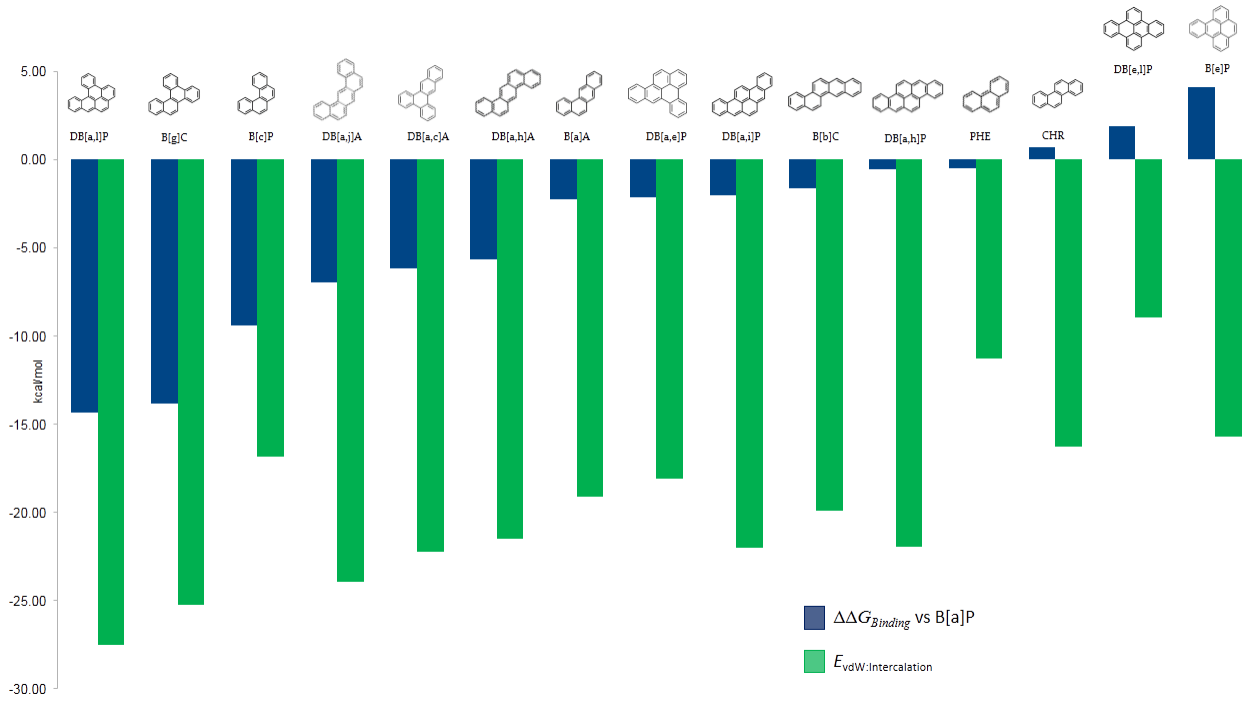
region B[g]C-DNA adduct and its bay region CHR-DNA analog and between the DB[a,j/c/h]A-DNA adducts and their B[a]A root compound, which has one less aromatic ring. These structural features and their impact on stabilizing van der Waals interactions are discussed in detail in Chapter 5.



(4.13) Average structure of the anti-glycosidic conformation of the (11R,12S,13R,14S)-trans-DB[a,l]P-DE-N6-dA^*_6 adduct intercalated from the major groove with its aromatic rings positioned in the primary intercalation pocket formed by dT_{16} and dT_{17} boxed in green and secondary intercalation pocket formed by dA^*_6 and dA_7 boxed in blue.

The strongly preferred systems are followed by the weakly preferred bay region B[a]A, DB[a,e]P, DB[a,i]P, and B[b]C systems, where relative free energies of binding as compared to B[a]P range from $\Delta\Delta G_{Binding}$:B[a]A = -2.23 kcal/mol to $\Delta\Delta G_{Binding}$:B[b]C = -1.65 kcal/mol. These systems assume an intercalated conformation similar to that of the strongly preferred bay region DB[a,j/c/h]A systems described above, but total van der Waals interactions from intercalation are generally weaker. The DB[a,i]P system is an exception to the trend of stabilizing van der Waals interactions decreasing with relative binding affinity, exhibiting strong van der Waals interactions in both the primary and secondary intercalation pockets.

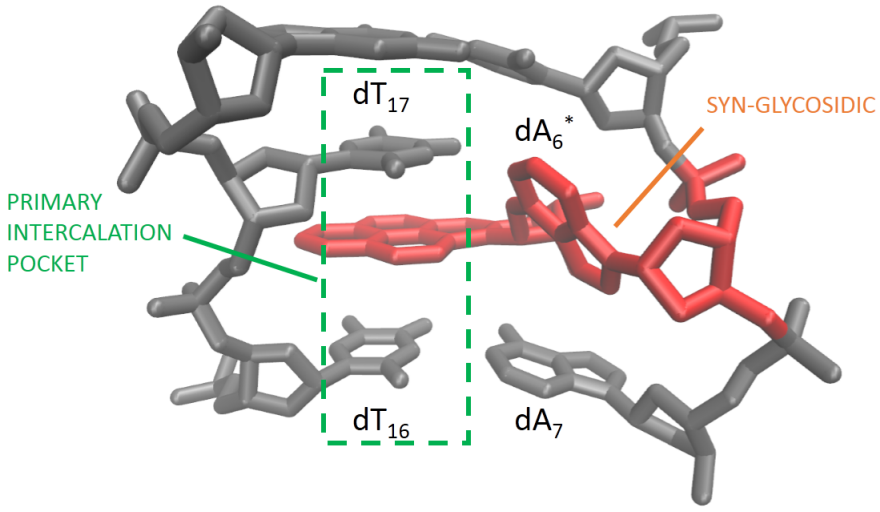
The bay region DB[a,h]P, PHE, and CHR systems are equally preferred as compared to B[a]P with $\Delta\Delta G_{Binding}$ falling between -0.53 kcal/mol and 0.68 kcal/mol. The PHE and CHR systems



(4.14) Relative free energies of binding ($\Delta\Delta G_{Binding}$) of (-R,-S,-R,-S)-trans-anti-PAH-DE-N6-dA^*_6 adducts as compared to a (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA^*_6 adduct at the central dA^*_6 in NRAS(Q61) and total van der Waals interactions from PAH intercalation ($E_{vdW:Intercalation} = E_{vdW:dT_{16}\,|\,dT_{17}} + E_{vdW:dA^*_6\,|\,dA_7}$)

assume a conformational motif similar to that of B[c]P and B[a]P, with their aromatic rings positioned for strong van der Waals interactions in the primary intercalation pocket and weak van der Waals interactions in the secondary intercalation pocket resulting in total van der Waals interactions from intercalation that are weaker than the weakly preferred systems. The DB[a,h]P system is another exception to the trend of stabilizing van der Waals interactions decreasing with relative binding affinity, exhibiting strong van der Waals interactions in both the primary and secondary intercalation pockets.

Finally, the bay region DB[e,l]P and B[e]P systems are not preferred as compared to B[a]P with $\Delta\Delta G_{Binding}$ = 1.88 kcal/mol and 4.11 kcal/mol respectively. The DB[e,l]P system assumes a non-intercalated major groove conformation where by the aromatic rings of DB[e,l]P are not effectively positioned for van der Waals interactions in either the primary or secondary intercalation pocket



(4.15) Average structure of the syn-glycosidic conformation of the (7R,8S,9R,10S)-trans-B[a]P-DE-N6-dA^*_6 adduct intercalated from the major groove with its aromatic rings positioned solely in the primary intercalation pocket formed by dT_{16} and dT_{17} boxed in green.

and van der Waals interactions from intercalation are the weakest of all the systems examined. The B[e]P system assumes a conformation similar to the B[c]P, PHE, CHR, and B[a]P systems with strong van der Waals interactions in the primary intercalation pocket and weak van der Waals interactions in the secondary intercalation pocket.

PAH-DNA Adduct	α'_{avg}	β'_{avg}	χ_{avg}	$E_{vdW:dT_{16}\|dT_{17}}$	$E_{vdW:dA^*_6\|dA_7}$	$E_{vdW:Interc.}$	$\Delta\Delta G_{Binding}$	$\Delta\Delta G_{Repair}$
	degrees	degrees	degrees	kcal/mol	kcal/mol	kcal/mol	kcal/mol	kcal/mol
DB[a,l]P	56.59 (8.43)	84.52 (7.67)	-114.91 (21.66)	-13.38 (1.13)	-14.11 (1.33)	-27.49	-14.34	3.46
B[g]C	56.55 (8.61)	81.25 (8.32)	-108.73 (19.28)	-12.17 (1.42)	-13.07 (1.60)	-25.24	-13.82	2.98
B[c]P	7.34 (12.43)	144.42 (13.36)	-79.4 (21.63)	-12.74 (1.18)	-4.08 (2.13)	-16.82	-9.38	1.96
DB[a,j]A	39.12 (11.42)	112.31 (8.29)	-91.33 (19.83)	-13.2 (0.99)	-10.72 (1.34)	-23.92	-6.96	4.31
DB[a,c]A	45.77 (11.68)	111.69 (11.47)	-94.76 (23.96)	-14.8 (1.21)	-7.42 (1.73)	-22.22	-6.17	4.75
DB[a,h]A	44.46 (11.17)	102.04 (9.67)	-90.63 (19.11)	-13.12 (1.03)	-8.38 (1.62)	-21.50	-5.68	3.80
B[a]A	41.39 (12.85)	107.55 (12.00)	-84.86 (17.79)	-12.49 (1.05)	-6.63 (1.54)	-19.12	-2.23	0.59
DB[a,e]P	116.85 (16.91)	4.68 (21.69)	-140.24 (10.82)	-8.23 (0.95)	-9.88 (1.78)	-18.11	-2.14	1.68
DB[a,i]P	74.78 (42.3)	69.0 (48.95)	-122.28 (25.21)	-13.31 (2.14)	-8.70 (2.75)	-22.01	-2.00	-0.26
B[b]C	49.86 (18.49)	91.76 (18.42)	-99.64 (26.25)	-12.19 (1.39)	-7.70 (1.84)	-19.89	-1.65	1.32
DB[a,h]P	73.51 (34.78)	66.19 (40.42)	-133.2 (20.31)	-11.83 (1.76)	-10.13 (1.99)	-21.96	-0.53	1.91
PHE	16.2 (13.47)	146.05 (13.70)	-64.72 (16.63)	-10.14 (0.99)	-1.16 (1.36)	-11.30	-0.51	0.88
B[a]P	5.40 (11.86)	154.1 (22.27)	-64.32 (9.31)	-13.69 (1.01)	-2.04 (1.32)	-15.73	0.00	0.00
CHR	25.28 (24.09)	127.71 (33.36)	-75.96 (20.93)	-12.32 (1.10)	-3.94 (2.72)	-16.26	0.68	-1.08
DB[e,l]P	-0.14 (9.53) *	-33.89 (9.11)	-61.88 (8.63)	-7.73 (1.21)	-1.22 (1.09)	-8.95	1.88	-1.54
B[e]P	28.53 (20.10)	137.48 (14.00)	-71.42 (23.42)	-12.92 (1.37)	-2.81 (1.25)	-15.73	4.11	-3.49
Unmodified	N/A	N/A	-108.79 (12.06)	-5.62 (0.64)	-7.17 (0.72)	-12.79	N/A	N/A

Table (4.3) Average PAH linkage torsion angles α'_{avg} and β'_{avg} and average glycosidic torsion angle χ_{avg} (standard deviation in parenthesis, see Figure 5.1 for angle definitions); $E_{vdW:dT_{16}\,|\,dT_{17}}$: average van der Waals interactions between the aromatic rings of the PAH and the dT_{16} and dT_{17} nucleobases of the primary intercalation pocket (standard deviation in parenthesis), $E_{vdW:dA^*_6\,|\,dA_7}$: average van der Waals interactions between the aromatic rings of the PAH and the dA^*_6 and dA_7 nucleobases of the secondary intercalation pocket (standard deviation in parenthesis); $E_{vdW:Intercalation} = E_{vdW:dT_{16}\,|\,dT_{17}} + E_{vdW:dA^*_6\,|\,dA_7}$: total van der Waals interactions from PAH intercalation; $\Delta\Delta G_{Binding}$: relative free energy of binding of a (-R,-S,-R,-S)-trans-anti-PAH-DE-N6-dA^*_6 adduct as compared to a (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA^*_6 adduct at the central dA^*_6 in NRAS(Q61); $\Delta\Delta G_{Repair}$: relative free energy of formation of the productive complex of a (-R,-S,-R,-S)-trans-anti-PAH-DE-N6-dA^*_6 adduct as compared to a (7R,8S,9R,10S)-trans-anti-B[a]P-DE-N6-dA^*_6 adduct at the central dA^*_6 in NRAS(Q61)

The six strongly preferred PAHs, are also the least likely to form the corresponding productive complex, exhibiting the most energetically unfavorable relative free energies of formation with $\Delta\Delta G_{Repair}$ ranging from 1.96 kcal/mol to 4.75 kcal/mol (Figure 4.12 red bars). Noting that the analogs of the $dC_5 : dG_{18}$ and $dA^*_6 : dT_{17}$ base pairs are extruded in the productive complex, the likelihood of formation of the productive complex is generally associated with the extent of hydrogen bond disruption in the three base pairs of NRAS(Q61) ($dC_5 : dG_{18}$, $dA^*_6 : dT_{17}$, and $dA_7 : dT_{16}$) as compared to unmodified DNA (Table 4.4). Hydrogen bonding is minimally disrupted in the strongest binding fjord region DB[a,l]P and B[g]C systems, which is consistent with formation of the productive complex being energetically unfavorable in these systems. Meanwhile, mild disruption of the dA^*_6-N1 : dT_{17}-N3 hydrogen bond in the B[c]P system is consistent with formation of the productive complex being less energetically unfavorable than in the DB[a,l]P and B[g]C systems. The DB[a,j/c/h]A systems exhibit mild disruptions of the three hydrogen bonds in the $dC_5 : dG_{18}$ base pair and mild to moderate disruptions of the dA^*_6-N1 : dT_{17}-N3 hydrogen bond. These hydrogen bonding disruptions are countered by enhancement of the dA^*_6-N6 : dT_{17}-O4 and dA_7-N6 : dT_{16}-O4 hydrogen bonds. Together with the strong stabilizing van der Waals interactions from intercalation in these PAH-DNA adduct systems, these hydrogen bonding characteristics are consistent with formation of the productive complex being energetically unfavorable. The structural features associated with these hydrogen bonding characteristics are discussed in further detail in Chapter 5.

The strongly preferred PAHs are more likely to form PAH-DNA adducts at the central dA^* in the NRAS(Q61) sequence context than B[a]P, and are less likely to form the corresponding productive complex. As a result, these PAH-DNA adducts are more likely to evade repair by GG-NER and persist to induce mutations during subsequent DNA replication cycles. This is consistent with studies in human HeLA cell extracts that have shown that DB[a,l]P has a propensity to form GG-NER resistant covalent adducts at the central dA^* of NRAS(Q61) while stereochemically analogous B[a]P adducts were repaired with high efficiency.[12,13,17,20] This also supports the notion that DB[a,l]P is more genotoxic than B[a]P in the NRAS(Q61) sequence context. The B[g]C system

PAH-DNA Adduct	dC_5-N3:dG_{18}-N1	dC_5-N4:dG_{18}-O6	dC_5-O2:dG_{18}-N2	dA^*_6-N1:dT_{17}-N3	dA^*_6-N6:dT_{17}-O4	dA_7-N1:dT_{16}-N3	dA_7-N6:dT_{16}-O4
DB[a,l]P	-7.89	-6.32	-2.52	+2.13	+1.50	+1.24	-1.23
B[g]C	-7.87	-6.76	-1.89	+6.62	-3.95	+1.44	-1.16
B[c]P	-3.02	-1.06	-4.37	**-14.70**	+6.75	-3.35	-0.88
DB[a,j]A	**-14.57**	**-14.30**	-5.65	**-26.90**	+15.36	-5.25	+11.67
DB[a,c]A	**-16.11**	**-12.77**	-7.16	**-16.67**	+4.87	-1.77	+2.60
DB[a,h]A	**-17.87**	**-14.96**	-6.01	**-20.94**	+13.50	-2.41	+10.44
B[a]A	**-12.36**	**-10.28**	-5.19	**-26.15**	+13.88	-1.45	+2.91
DB[a,e]P	-8.16	-4.13	-6.76	**-20.47**	**-73.64**	**-9.41**	+17.69
DB[a,i]P	**-13.71**	-8.18	-7.63	**-30.01**	**-19.16**	-3.85	+6.11
B[b]C	**-19.45**	**-14.74**	**-9.61**	**-25.11**	+8.20	-2.65	+7.91
DB[a,h]P	**-14.23**	-8.12	-7.07	**-18.95**	**-22.88**	-0.68	+5.29
PHE	-0.84	-0.30	-3.46	**-9.94**	**-9.21**	-3.13	**-14.47**
B[a]P	+3.79	+3.94	+0.70	**-12.00**	+0.68	-0.91	**-16.19**
CHR	-6.29	-4.44	-3.38	**-18.98**	+2.07	-8.50	**-12.06**
DB[e,l]P	-5.19	-7.86	-1.14	**-85.59**	**-79.96**	**-9.81**	+2.75
B[e]P	-3.82	-2.91	-4.24	**-19.05**	+2.37	-3.15	**-14.15**

Table (4.4) Differences in hydrogen bond occupancy as compared to unmodified DNA (percentage points) for base pairs in the NRAS(Q61) 3-mer.

exhibits a relative binding affinity and an energetic aversion to formation of the productive complex that is comparable to DB[a,l]P while B[c]P exhibits similar characteristics of a lesser magnitude. As a result, these fjord region PAHs are likely to exhibit greater genotoxicity at the central dA^* in the NRAS(Q61) sequence context than B[a]P. While the bay region DB[a,j/c/h]A systems exhibit weaker relative binding affinities than the fjord region systems, they exhibit a greater energetic aversion to forming the productive complex and are thus likely to be more genotoxic at the central dA^* in the NRAS(Q61) sequence context than B[a]P as well.

Among the weakly preferred and equally preferred PAHs, formation of the productive complex is more energetically favorable and thus more likely than in the strongly preferred systems with $\Delta\Delta G_{Repair}$ ranging from -1.08 kcal/mol to 1.91 kcal/mol. The greater likelihood of formation of the productive complex in these systems is associated with hydrogen bond disruptions that are consistent with extrusion of the $dC_5 : dG_{18}$ and $dA^*_6 : dT_{17}$ base pairs being more energetically favorable than in the strongly preferred PAH-DNA adduct systems. Hydrogen bond disruptions in the B[a]A and B[b]C systems are similar to those seen in the strongly preferred DB[a,j/c/h]A

systems, but enhancement of the dA^*_6-N6 : dT_{17}-O4 and dA_7-N6 : dT_{16}-O4 hydrogen bonds is generally of an equal or lesser magnitude than the DB[a,j/c/h]A systems. The DB[a,e]P-DE system exhibits moderate to severe disruptions of both hydrogen bonds in the dA^*_6 : dT_{17} base pair. The DB[a,i]P and DB[a,h]P systems exhibit mild to moderate disruptions of hydrogen bonding in the dC_5 : dG_{18} and dA^*_6 : dT_{17} base pairs without enhanced hydrogen bonding of the magnitude observed in the strongly preferred DB[a,j/c/h]A systems. The PHE and CHR systems exhibit mild hydrogen bond disruption in the dA^*_6 : dT_{17} and dA_7 : dT_{16} base pairs. In conjunction with stabilizing van der Waals interactions from intercalation that are generally weaker in these systems than in the strongly preferred PAH-DNA adduct systems, these hydrogen bonding characteristics are consistent with formation of the productive complex being energetically favorable as compared to the strongly preferred systems.

From this it is evident that the weakly preferred and equally preferred PAHs are slightly more or just as likely to form PAH-DNA adducts as compared to B[a]P, and are slightly less likely or just as likely to form the productive complex. These systems generally exhibit weaker stabilizing van der Waals interactions from intercalation and hydrogen bonding characteristics that make extrusion of the dC_5 : dG_{18} and dA^*_6 : dT_{17} base pairs more energetically favorable than the strongly preferred systems. As a result, these PAH-DNA adducts are less likely to evade repair by GG-NER and less likely to persist and induce mutations during subsequent DNA replication cycles as compared to the strongly preferred systems and the genotoxicity of these PAHs at the central dA of NRAS(Q61) is likely to be comparable to that of B[a]P, which is efficiently repaired by GG-NER in the NRAS(Q61) sequence context.

Finally, the two non-preferred PAHs, are the most likely to form the productive complex with $\Delta\Delta G_{Repair:\mathrm{DB[e,l]P}}$ = -1.54 kcal/mol and $\Delta\Delta G_{Repair:\mathrm{B[e]P}}$ = -3.49 kcal/mol. In the DB[e,l]P system, hydrogen bonding in the dA^*_6 : dT_{17} base pair is severely disrupted. In the B[e]P system, hydrogen bonding in the dA^*_6 : dT_{17} and dA_7 : dT_{16} base pairs is mildly disrupted. Noting that the DB[e,l]P system exhibits the weakest total van der Waals interactions from intercalation, while the B[e]P system exhibits the third weakest among all of the systems examined, these hydrogen bonding

characteristics are consistent with formation of the productive complex being energetically favorable as compared to the weakly and equally preferred PAHs. These PAHs are less likely to form covalent DNA adducts at the central dA in NRAS(Q61) than B[a]P, and are less likely to evade repair than the readily repaired B[a]P. As a result, these PAH-DNA adducts are unlikely to persist and induce mutations during subsequent DNA replication cycles than B[a]P, and are likely to be less genotoxic at the central dA in the NRAS(Q61) sequence context than B[a]P.

CHAPTER 5

Structural Features of PAH-DNA Adducts in the NRAS(Q61) Sequence Context

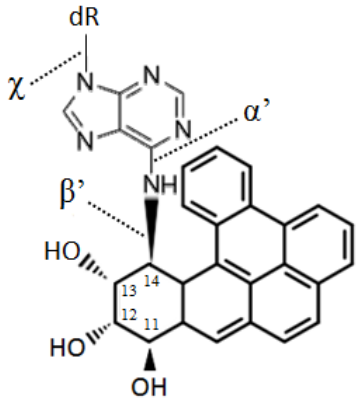
This chapter will expand upon the structural features of each PAH-DNA adduct system that were outlined in Chapter 4 by examining the association between relative binding affinity, PAH structure, intercalated conformation, DNA rigid-body parameters, stabilizing van der Waals interactions from intercalation (π-stacking), hydrogen bond occupancy, and conformational / non-bonded energies of the NRAS(Q61) 3-mer.

5.1 Strongly Preferred PAH-DNA Adducts

5.1.1 Conformational Details and van der Waals Interactions

With the exception of B[c]P, the conformational motif assumed by the strongly preferred PAH-DNA adducts is characterized by dA_6^* assuming an average anti-glycosidic conformation and the average PAH adduct linkage site torsion angles α' and β' assuming associated average values that facilitate positioning of the aromatic rings of the PAH for strong stabilizing van der Waals interactions in both the primary dT_{16} | dT_{17} and the secondary dA_6^* | dA_7 intercalation pockets (see Figure 5.1 for torsion angle definitions). The DB[a,l]P and B[g]C systems assume a conformation where dA_6^* is firmly anti-glycosidic with χ_{avg} = -114.91° and -108.73° respectively, while α'_{avg} = 56.59° and 56.55° and β'_{avg} = 84.52° and 81.25° respectively. The DB[a,j/c/h]A systems assume conformations where dA_6^* fluctuates between syn and anti-glycosidic about the -90° syn/anti-glycosidic

threshold during equilibration with χ_{avg} = -91.33°, -94.76°, and -90.63° respectively, while while α'_{avg} = 39.12°, 45.77°, and 44.46° and β'_{avg} = 112.31°, 111.69°, and 102.04° respectively (Table 4.3).

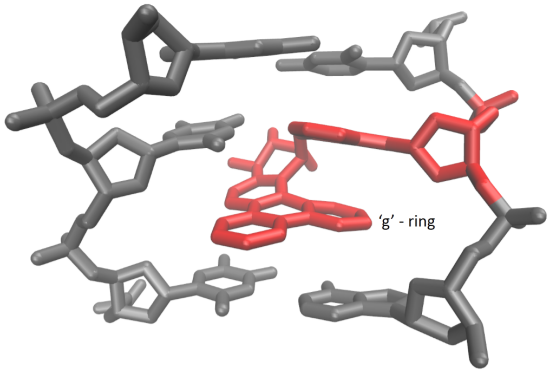


(5.1) PAH linkage torsion angles α'_{avg} and β'_{avg} and average glycosidic torsion angle χ_{avg}

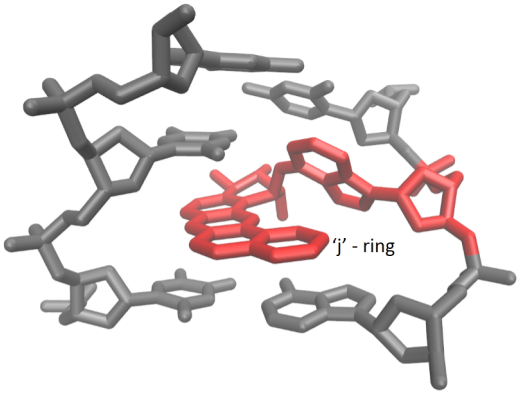
Among the strongly preferred PAHs, average values of $E_{vdW:dT_{16} \mid dT_{17}}$ in the primary intercalation pocket range from -14.80 kcal/mol to -12.17 kcal/mol, as compared to unmodified DNA where van der Waals interactions between dT_{16} and dT_{17} in the absence of an intercalated PAH average -5.62 kcal/mol (Table 4.3), demonstrating a clear stabilizing effect associated with PAH-DNA adduct intercalation. In the two most strongly preferred DB[a,l]P and B[g]C systems, the fjord aromatic rings on the 'l' side of the B[a]P root (Figure 4.13), and on the 'g' side of the CHR root (Figure 5.2) are positioned for particularly strong van der Waals interactions in the secondary intercalation pocket that exceed those in the primary intercalation pocket with $E_{vdW:dA^*_6 \mid dA_7}$ = -14.11 kcal/mol and -13.07 kcal/mol respectively. In unmodified DNA, van der Waals interactions between dA^*_6 and dA_7 in the absence of an intercalated PAH average -7.17 kcal/mol demonstrating an additional stabilizing effect associated with intercalation of these fjord PAHs (Table 4.3). These enhanced van der Waals interactions are a function of the firmly anti-glycosidic conformation assumed by dA^*_6 in these two systems, which results in the plane of the dA^*_6 nucleobase being nearly parallel to the plane of the fjord aromatic ring in each system. This facilitates strong π-stacking between the fjord aromatic ring and both the dA^*_6 and dA_7 nucleobases (Figure 4.13 and 5.2 re-

spectively). Note that this is not observed in the fjord B[c]P system which does not assume the same conformational motif and is discussed further below.

In the DB[a,j]A system, van der Waals interactions in the secondary intercalation pocket do not exceed those of the primary intercalation pocket but are comparatively strong with $E_{vdW:dA^*_6 \mid dA_7}$ = -10.72 kcal/mol (Table 4.3). The additional aromatic ring on the 'j' side of the B[a]A root is situated to have strong van der Waals interactions with dA_7 in the secondary intercalation pocket, but as described above, the modified dA^*_6 assumes an average glycosidic angle of χ_{avg} = -91.33° ± 19.83° resulting in the plane of the modified dA^*_6 being situated on average diagonal to the plane of the additional aromatic ring on the 'j' side of the B[a]A root thus limiting van der Waals interaction in the secondary intercalation pocket (Figure 5.3).



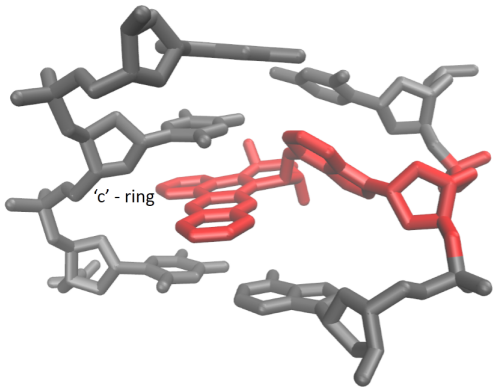
(5.2) Average structure of the B[g]C-DNA adduct



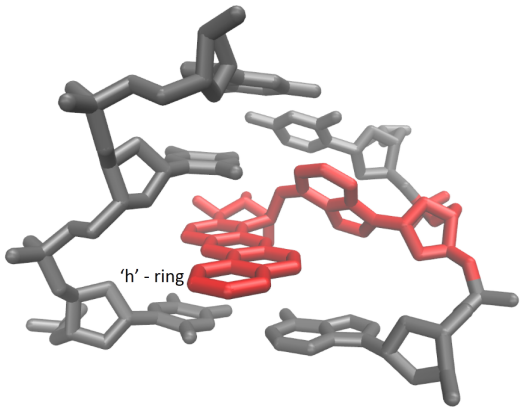
(5.3) Average structure of the DB[a,j]A-DNA adduct

In the DB[a,c]A system, the additional aromatic ring on the 'c' side of the B[a]A root is situated in the primary $dT_{16} \mid dT_{17}$ intercalation pocket (Figure 5.4), enhancing van der Waals interactions as compared to other systems with $E_{vdW:dT_{16} \mid dT_{17}}$ = -14.80 kcal/mol being the strongest van der Waals interaction observed in the primary intercalation pocket among all PAHs examined in this work. In the DB[a,h]A system, the additional aromatic ring on the 'h' side of the B[a]A root is positioned to avoid steric clashes with the sugar phosphate backbone in both strands of the DNA duplex, and does not serve to enhance van der Waals interactions in either the primary or

secondary intercalation pockets (Figure 5.5). As in the DB[a,j]A system, the plane of dA^*_6 is on average oriented diagonal to the plane of the aromatic rings in the DB[a,c]A-DE and DB[a,h]A-DE systems, limiting π-stacking with dA^*_6, and resulting in van der Waals interactions in the secondary intercalation pocket that are comparable to those between dA_6 and dA_7 in unmodified DNA (Table 4.3).



(5.4) Average structure of the DB[a,c]C-DNA adduct

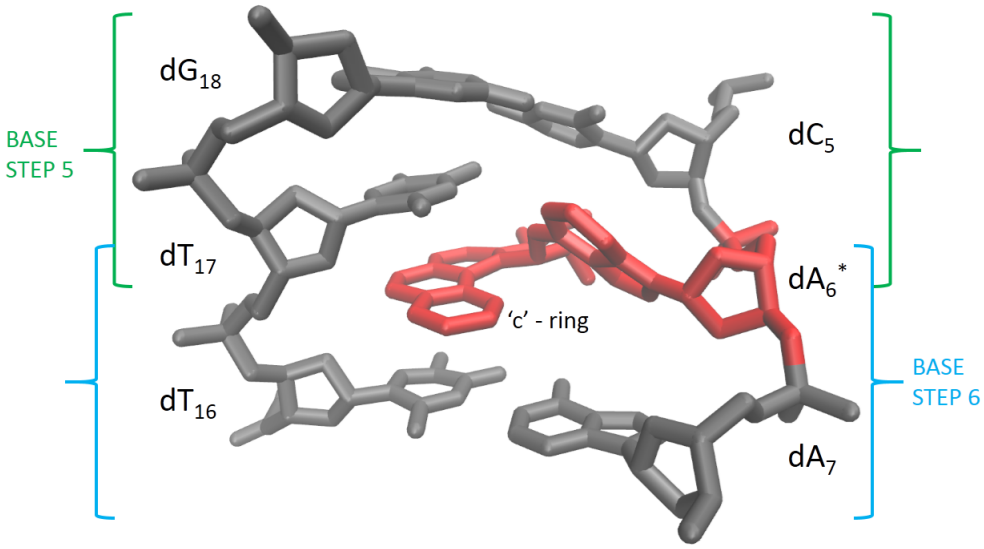


(5.5) Average structure of the DB[a,h]A-DNA adduct

As described above, the conformational motif assumed by the DB[a,l]P, B[g]C, and DB[a,j/c/h]A systems stands in contrast to the intercalated conformation observed in the B[a]P system, which is characterized by dA^*_6 assuming an average syn-glycosidic conformation with χ_{avg} = -64.32° and the adduct linkage site torsion angles assuming associated average values of α'_{avg} = 5.40° and β'_{avg} = 154.10°, resulting in the aromatic rings of B[a]P being positioned solely in the primary intercalation pocket. As a result, van der Waals interactions in the primary intercalation pocket are strong with $E_{vdW:dT_{16}\,|\,dT_{17}}$ = -13.69 kcal/mol while those in the secondary intercalation pocket are comparatively weak with $E_{vdW:dA^*_6\,|\,dA_7}$ = -2.04 kcal/mol (Table 4.3). Among the strongly preferred PAHs, the B[c]P system assumes an intercalated conformation that is similar to B[a]P where dA^*_6 assumes an average syn-glycosidic conformation with $\chi_{avg} = -79.40°$ with average adduct linkage site torsion angles of $\alpha'_{avg} = 7.34°$ and $\beta'_{avg} = 144.42°$ (Table 4.3). This conformation results in B[c]P intercalating from the major groove with its aromatic rings - including the fjord aromatic

ring - situated solely in the primary intercalation pocket (Figure 5.6). This is in contrast to the conformation seen in the other fjord DB[a,l]P and B[g]C systems where the additional fjord aromatic ring is situated in the secondary intercalation pocket, exhibiting strong van der Waals interactions with the dA^*_6 and dA_7 nucleobases. As a result, the B[c]P-DNA adduct's van der Waals interactions in the primary intercalation pocket are strong with $E_{vdW:dT_{16} \mid dT_{17}}$ = -12.74 kcal/mol while those in the secondary intercalation pocket are comparatively weak with $E_{vdW:dA^*_6 \mid dA_7}$ = -4.08 kcal/mol (Table 4.3).

Note that among the fjord region PAHs, relative binding affinities and total van der Waals interactions from intercalation (Figure 4.14) decrease with the decreasing number of aromatic rings in the PAH (5 in DB[a,l]P, 4 in B[g]C, and 3 in B[c]P). The DB[a,l]P system has one more aromatic ring than B[g]C situated in the primary intercalation pocket (Figures 4.13 and 5.2 respectively), resulting in stronger van der Waals interactions with dT_{16} and dT_{17} in the DB[a,l]P system than the B[g]C system. This additional aromatic ring in turn positions the fjord aromatic ring in the DB[a,l]P system for stronger van der Waals interactions with dA^*_6 and dA_7 in the secondary intercalation pocket than the B[g]C system. The B[c]P system exhibits the weakest van der Waals interactions as a result of there being no aromatic rings positioned in the secondary intercalation pocket (Table 4.3).



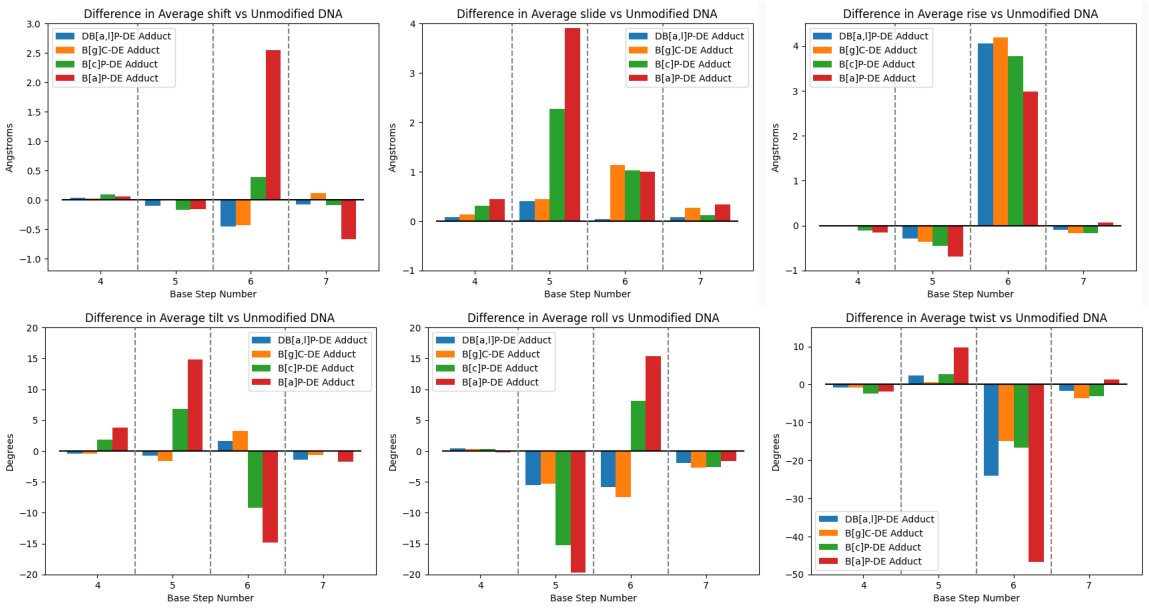
(5.6) Average structure of the syn-glycosidic conformation of the B[c]P-DNA adduct intercalated from the major groove with its aromatic rings positioned solely in the primary intercalation pocket formed by dT_{16} and dT_{17}. Base step 5 formed by the dC_5 : dG_{18} and dA_6^* : dT_{17} base pairs bracketed in green. Base step 6 formed by the dA_6^* : dT_{17} and dA_7 : dT_{16} base pairs bracketed in blue.

5.1.2 Rigid-Body Parameters and Hydrogen Bonding

Analysis of average DNA rigid-body base step parameters comprised of shift, slide, rise, tilt, roll, and twist (Figures 5.7 and 5.8) indicate that structural distortions of the DNA duplex resulting from a covalent PAH-DNA adduct at dA_6^* in NRAS(Q61), as compared to unmodified DNA, are generally limited to the NRAS(Q61) 3-mer consisting of the base steps formed by the $dC_5 : dG_{18}$ and $dA_6^* : dT_{17}$ base pairs (base step 5 in Figure 5.6) and the $dA_6^* : dT_{17}$ and $dA_7 : dT_{16}$ base pairs (base step 6 in Figure 5.6). The total energy (configurational and non-bonded) of the NRAS(Q61) 3-mer is measured in each system without the PAH in order to exclude the configurational and non-bonded energy of the PAH and to compare the total energy of the nucleotide configuration to that of unmodified DNA (Table 5.1). Configurational energy differences between PAH-DNA adduct systems and unmodified DNA manifest in increased dihedral and to a lesser extent increased angle energies, while both electrostatic and van der Waals energies are increased, indicating energetically unfavorable distortions of the NRAS(Q61) 3-mer caused by PAH-DNA adduct intercalation.

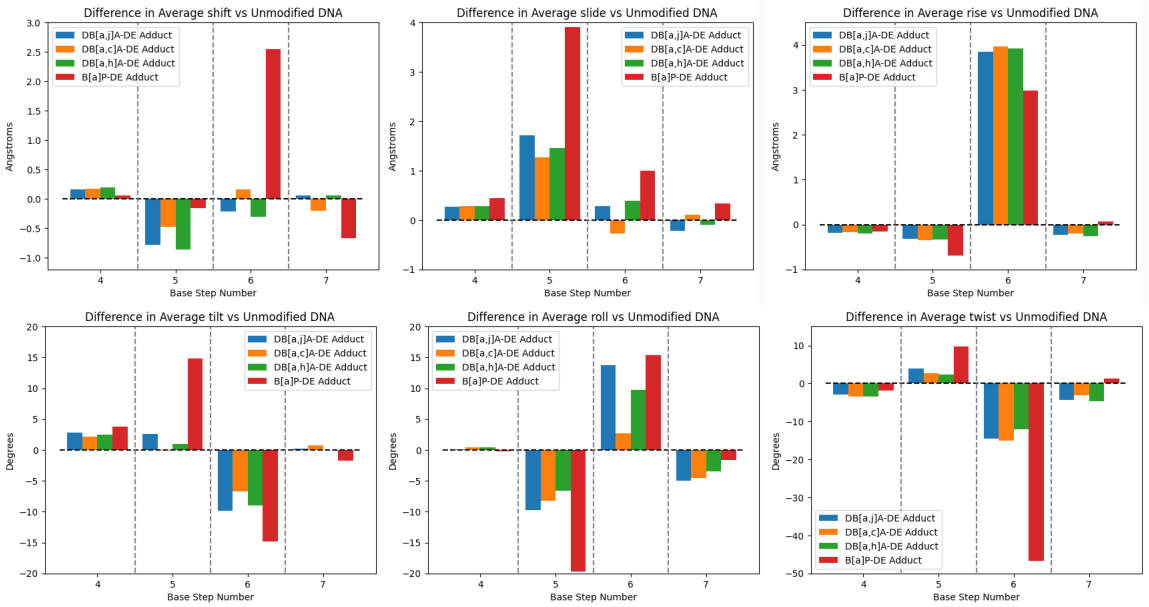
Among the strongly preferred PAHs, structural distortion of the DNA duplex as compared to unmodified DNA is generally characterized by large increases of approximately +3.5Å to +4.0Å in average rise and moderate decreases of approximately -15° to -25° in average twist at base step 6 (Figures 5.7 and 5.8 - blue, orange, and green bars). This is in marked contrast to B[a]P for which the increase in average rise is smaller at +2.98Å and the decrease in average twist is much larger at - 46.76° (Figures 5.7 and 5.8 - red bars). In the strongly preferred PAH-DNA adduct systems, the increase in rise at base step 6 occurs to accommodate the intercalated PAH between the dT_{16} and dT_{17} nucleobases of the primary intercalation pocket and the dA_6^* and dA_7 nucleobases of the secondary intercalation pocket (Figures 4.13 and 5.2 - 5.5). This is as opposed to the B[a]P system whose aromatic rings are positioned solely in the primary intercalation pocket, not requiring the larger increase in rise (Figure 4.15). Note that while the aromatic rings of B[c]P are positioned solely in the primary intercalation pocket, similar to the B[a]P system, B[c]P is non-planar owing to the steric hindrance between the aromatic ring on the 'c' side of the PHE root and the aliphatic ring that is characteristic of fjord region PAHs. This requires a greater degree of rise in base step

6 than B[a]P, but less than that required to accommodate the non-planar fjord region DB[a,l]P and B[g]C systems whose fjord aromatic rings are situated in the secondary intercalation pocket and have a greater depth than the planar B[a]P.



(5.7) Strongly preferred fjord PAH-DNA adducts: distortions in average base step parameters as compared to unmodified DNA.

Decreases in average twist at base step 6 (i.e. unwinding of the DNA double helix) is associated with widening of the major and minor grooves of the DNA duplex to avoid steric clashes between the intercalated PAH and the sugar-phosphate backbones of both DNA strands (see major and minor groove trajectories in Appendix B). The intercalated conformation of the B[a]P system requires a greater degree of unwinding to accommodate the aromatic rings that are positioned solely in the primary intercalation pocket (Figure 4.15), effectively spanning the major groove and resulting in a larger decrease in average twist. The substantial unwinding of the DNA double helix in the B[a]P system is associated with a markedly higher total energy of E_{Total}=-7.31 kcal/mol in the NRAS(Q61) 3-mer as compared to unmodified DNA for which E_{Total}=-121.91 kcal/mol (Table 5.1). The difference in total energy is rooted in weaker non-bonded interactions in the B[a]P system. The strongly preferred PAHs, whose aromatic rings are positioned in the primary and secondary intercalation pockets, effectively extend into the hydrophobic core of the DNA duplex



(5.8) Strongly preferred DB[a,j/c/h]A-DNA adducts: distortions in average base step parameters as compared to unmodified DNA.

and require a smaller decrease in average twist, to accommodate the PAH (Figures 4.13 and 5.2 - 5.5). Although the B[c]P system intercalates from the major groove with its aromatic rings positioned solely in the primary intercalation pocket, B[c]P is the smallest of the strongly preferred PAHs with only three aromatic rings, and these aromatic rings do not extend toward the sugar-phosphate backbone in the complementary strand as they do in the B[a]P system (Figures 4.15 and 5.6 respectively). As a result, B[c]P requires a smaller decrease in average twist to accommodate its aromatic rings. The smaller decrease in average twist among the strongly preferred PAHs is associated with lower total energy configurations of the NRASQ61) 3-mer as compared to B[a]P system with E_{Total} ranging from -52.08 kcal/mol to -42.55 kcal/mol for these systems (Table 5.1).

Note that the DB[a,l]P and B[g]C systems which have the greatest relative binding affinity, generally have fewer and smaller structural distortions (Figure 5.7 blue and orange bars respectively) than the B[a]P system (Figure 5.7 red bars). The B[c]P system exhibits notable distortions in slide, tilt, and roll in base steps 5 and 6 (Figure 5.7 green bars) that are generally not observed in the other fjord region DB[a,l]P and B[g]C systems but that are similar to those seen in the

PAH-DNA Adduct	Bond	Angle	Dihedral	Improper	Electrostatic	van der Waals	Total Conformational	Total Non-Bonded	Total Energy
	kcal/mol	kcal/mol	kcal/mol	kcal/mol	kcal/mol	kcal/mol	kcal/mol	kcal/mol	kcal/mol
Unmodified	56.26 (6.32)	151.73 (9.19)	199.43 (4.39)	3.10 (1.08)	-553.86 (9.15)	21.43 (6.13)	410.51 (11.72)	-532.42 (10.04)	-121.91 (12.79)
DB[a,l]P	57.99 (6.49)	159.59 (9.97)	215.33 (6.37)	2.90 (1.05)	-523.95 (12.09)	42.17 (6.13)	435.82 (11.76)	-481.78 (12.27)	-45.96 (14.80)
B[g]C	57.93 (6.48)	159.02 (9.87)	213.78 (6.06)	2.92 (1.06)	-527.54 (11.88)	41.80 (6.27)	433.66 (11.91)	-485.74 (12.69)	-52.08 (15.65)
B[c]P	57.45 (6.38)	158.79 (9.54)	212.77 (5.47)	2.89 (1.05)	-514.90 (16.25)	38.20 (6.41)	431.90 (11.62)	-476.70 (16.18)	-44.80 (18.40)
DB[a,j]A	57.27 (6.41)	157.82 (9.56)	220.74 (5.51)	2.89 (1.03)	-523.02 (10.63)	38.23 (6.20)	438.71 (11.80)	-484.79 (11.31)	-46.08 (13.58)
DB[a,c]A	57.25 (6.39)	156.13 (9.47)	222.44 (5.44)	2.92 (1.05)	-520.88 (10.89)	39.58 (6.18)	438.75 (11.78)	-481.30 (11.74)	-42.55 (14.14)
DB[a,h]A	57.18 (6.43)	156.68 (9.33)	221.03 (5.16)	2.88 (1.04)	-526.71 (10.28)	38.89 (6.17)	437.77 (11.69)	-487.82 (10.98)	-50.05 (13.29)
B[a]A	57.40 (6.43)	157.94 (9.66)	220.18 (5.46)	2.88 (1.03)	-522.94 (12.31)	37.79 (6.29)	438.39 (11.72)	-485.16 (12.58)	-46.76 (14.68)
DB[a,e]P	57.64 (6.40)	155.95 (9.44)	222.69 (6.51)	3.28 (1.14)	-510.37 (11.03)	41.60 (5.96)	439.55 (11.85)	-468.77 (11.69)	-29.22 (13.57)
DB[a,i]P	57.38 (6.47)	156.30 (9.42)	223.88 (8.58)	2.96 (1.07)	-519.05 (15.58)	41.71 (6.15)	440.52 (12.87)	-477.34 (16.37)	-36.83 (15.20)
B[b]C	57.22 (6.40)	156.22 (9.41)	223.56 (5.84)	2.93 (1.06)	-525.79 (11.66)	39.86 (6.24)	439.93 (11.79)	-485.93 (12.43)	-46.00 (14.92)
DB[a,h]P	57.38 (6.41)	155.98 (9.38)	224.59 (7.39)	2.98 (1.07)	-520.33 (14.12)	42.03 (6.13)	440.93 (12.43)	-478.30 (14.73)	-37.37 (14.85)
PHE	57.67 (6.46)	157.57 (9.33)	215.31 (4.89)	2.89 (1.05)	-511.57 (11.98)	34.65 (6.43)	433.44 (11.72)	-476.92 (12.95)	-43.48 (14.93)
B[a]P	57.89 (6.38)	158.08 (9.47)	219.50 (5.66)	2.86 (1.04)	-482.48 (14.15)	36.85 (6.10)	438.32 (11.77)	-445.63 (15.18)	-7.31 (16.10)
CHR	57.71 (6.43)	158.05 (9.60)	220.23 (5.94)	2.90 (1.06)	-503.91 (23.29)	37.79 (6.45)	438.89 (11.83)	-466.12 (22.72)	-27.24 (23.52)
DB[e,l]P	57.77 (6.33)	155.67 (9.71)	213.31 (6.06)	2.95 (1.06)	-489.66 (11.42)	45.76 (5.99)	429.69 (11.61)	-443.90 (12.51)	-14.21 (14.31)
B[e]P	57.52 (6.50)	158.35 (9.54)	219.14 (5.67)	2.90 (1.04)	-515.48 (12.15)	36.41 (6.66)	437.90 (11.97)	-479.07 (12.56)	-41.17 (14.63)

Table (5.1) Average conformational and non-bonded energies in the NRAS(Q61) 3-mer with PAHs excluded, standard deviations in parenthesis.

B[a]P system, albeit of a lesser magnitude. This is consistent with B[c]P assuming an intercalated conformation that is similar to B[a]P, and distinct from the intercalated conformation assumed by DB[a,l]P and B[g]C. The DB[a,j/c/h]A systems also exhibit distortions similar to the B[a]P system of lesser magnitude in slide at base step 5, tilt at base step 6, and roll at base steps 5 and 6 (Figure 5.8 blue, orange, green, and red bars respectively). These distortions are associated with the average dA^*_6 glycosidic torsion angle assumed by these systems. The B[a]P and B[c]P systems assume average syn-glycosidic conformations with χ_{avg} = -64.32° and χ_{avg} = -79.40°, respectively. The DB[a,j/c/h]A systems assume average anti-glycosidic conformations that are very close to the syn/anti-glycosidic threshold of -90° with standard deviations of approximately 20° (Table 4.3) resulting in conformations that fluctuate between syn and anti-glycosidic. In this conformational motif, the plane of the dA^*_6 nucleobase is on average oriented diagonal to the plane of neighboring nucleobases, requiring the distortions in slide, tilt, and roll in order to accommodate the diagonal orientation of dA^*_6. This is in contrast to the DB[a,l]P and B[g]C systems which assume firmly anti-glycosidic conformations of dA^*_6 and thus do not require the associated distortions in slide, tilt, and roll.

Analysis of average DNA rigid-body base pair parameters comprised of shear, stretch, stagger, buckle, propeller, and opening and associated changes in percent hydrogen bond occupancy as compared to unmodified DNA (Table 4.4) indicate that disruption of base pairing resulting from a covalent PAH-DNA adduct at dA^*_6 in NRAS(Q61), as compared to unmodified DNA, occurs primarily in the $dA^*_6 : dT_{17}$ base pair and to a lesser extent in the $dC_5 : dG_{18}$ base pair. Some disruption of the $dA_7 : dT_{16}$ base pair occurs in the weaker binding PAH systems as described below. In Table 4.4, decreases in hydrogen bond occupancy ranging from 9 to 20 percentage points are high-lighted in yellow, decreases ranging from 20 to 30 percentage points are high-lighted in orange, and decreases greater than 30 percentage points are high-lighted in red. Increases in hydrogen bond occupancy greater than 10 percentage points are high-lighted in green to denote enhanced hydrogen bond occupancy as compared to unmodified DNA.

PAH-DNA Adduct	dC_5-N3:dG_{18}-N1	dC_5-N4:dG_{18}-O6	dC_5-O2:dG_{18}-N2	dA^*_6-N1:dT_{17}-N3	dA^*_6-N6:dT_{17}-O4	dA_7-N1:dT_{16}-N3	dA_7-N6:dT_{16}-O4
DB[a,l]P	3.03(0.14)	3.09(0.29)	2.90(0.14)	2.95(0.13)	2.94(0.21)	2.92(0.11)	3.03(0.24)
B[g]C	3.03(0.14)	3.09(0.28)	2.89(0.14)	2.92(0.11)	2.98(0.25)	2.92(0.11)	3.02(0.23)
B[c]P	3.00(0.12)	3.04(0.24)	2.90(0.15)	**3.05(0.25)**	2.89(0.23)	2.95(0.12)	3.02(0.24)
DB[a,j]A	**3.04(0.14)**	**3.12(0.29)**	2.90(0.14)	**3.09(0.21)**	2.83(0.13)	2.96(0.12)	2.94(0.18)
DB[a,c]A	**3.05(0.19)**	**3.13(0.39)**	2.91(0.15)	**3.05(0.21)**	2.94(0.29)	2.94(0.12)	2.99(0.21)
DB[a,h]A	**3.05(0.16)**	**3.14(0.34)**	2.90(0.15)	**3.08(0.25)**	2.85(0.15)	2.95(0.12)	2.95(0.19)
B[a]A	**3.04(0.15)**	**3.10(0.30)**	2.90(0.15)	**3.12(0.30)**	2.84(0.14)	2.94(0.12)	3.00(0.21)
DB[a,e]P	3.02(0.15)	3.08(0.32)	2.93(0.16)	**3.06(0.18)**	**4.63(0.63)**	**2.99(0.14)**	2.89(0.16)
DB[a,i]P	**3.04(0.17)**	3.11(0.34)	2.91(0.15)	**3.68(1.38)**	**3.89(1.70)**	2.96(0.23)	2.98(0.21)
B[b]C	**3.06(0.16)**	**3.14(0.32)**	**2.92(0.16)**	**3.22(0.53)**	2.91(0.38)	2.95(0.12)	2.97(0.2)
DB[a,h]P	**3.05(0.19)**	3.11(0.35)	2.91(0.17)	**3.21(0.66)**	**3.67(1.43)**	2.93(0.12)	2.98(0.20)
PHE	3.00(0.11)	3.02(0.21)	2.90(0.15)	**3.02(0.20)**	**2.97(0.19)**	2.95(0.12)	**3.11(0.28)**
B[a]P	2.98(0.11)	3.00(0.20)	2.87(0.13)	**3.02(0.14)**	2.96(0.22)	2.94(0.12)	**3.10(0.25)**
CHR	3.01(0.14)	3.06(0.28)	2.89(0.15)	**3.12(0.40)**	2.95(0.27)	3.15(0.82)	**3.27(0.86)**
DB[e,l]P	3.02(0.14)	3.09(0.29)	2.89(0.14)	**6.23(0.85)**	**5.06(1.19)**	**2.96(0.13)**	3.00(0.22)
B[e]P	3.01(0.16)	3.05(0.29)	2.90(0.16)	**3.09(0.31)**	2.94(0.23)	2.95(0.12)	**3.11(0.29)**
Unmodified	3.00(0.13)	3.05(0.26)	2.88(0.14)	2.96(0.12)	2.94(0.19)	2.93(0.11)	3.02(0.23)

Table (5.2) Average hydrogen bond distances (Å) for base pairs in the NRAS(Q61) 3-mer, standard deviation in parenthesis.

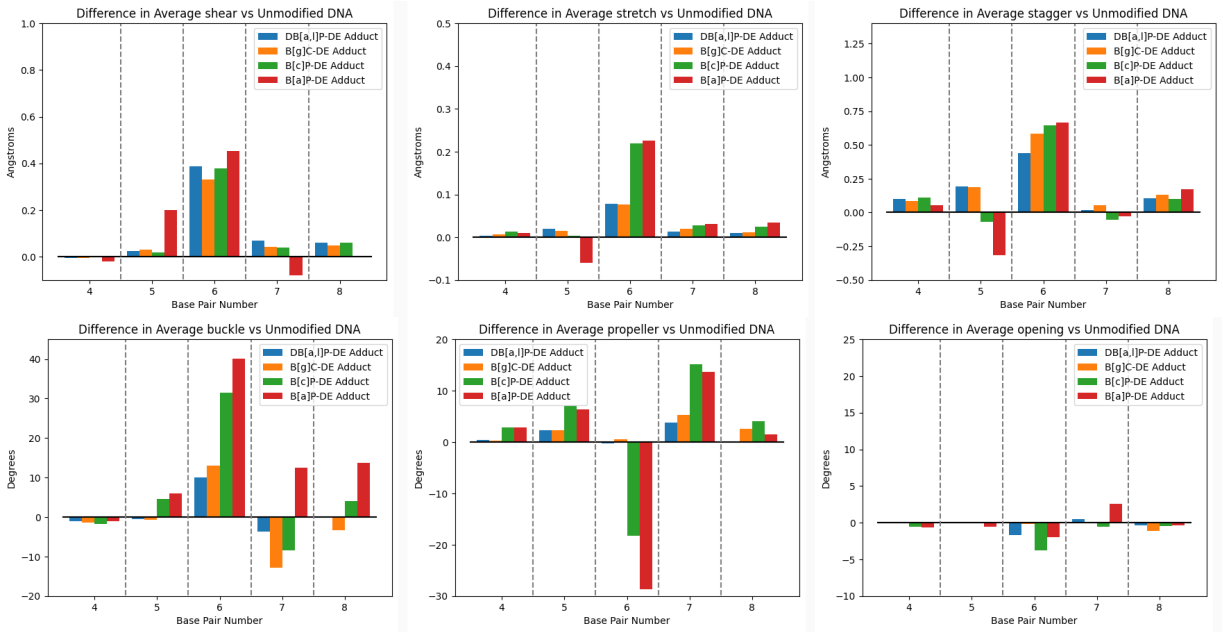
Among the three strongest binding PAHs, hydrogen bonding in the three base pairs of NRAS(Q61) is minimally disrupted in the DB[a,l]P and B[g]C systems, while there is a moderate -14.70 percentage point decrease in the dA^*_6-N1:dT_{17}-N3 hydrogen bond in the B[c]P system (Table 4.4).

Among the next three strongest binding DB[a,j/c/h]A systems, there are decreases ranging from -16.67 to -26.90 percentage points in the dA^*_6-N1:dT_{17}-N3 hydrogen bond occupancy. Analysis of rigid-body base pair parameters in these systems indicate that disruptions in the dA^*_6-N1:dT_{17}-N3 hydrogen bond are associated with increases in average stretch of approximately +0.2Å, increases in average buckle of approximately +30°, and decreases in average propeller of approximately +20° in the dA^*_6 : dT_{17} base pair (base pair 6 in Figures 5.9 and 5.10). These differences in base pair parameters result in increased average distances for the dA^*_6-N1 : dT_{17}-N3 hydrogen bond that range from 3.05Å to 3.09Å with standard deviations ranging from 0.21Å to 0.25Å as compared to unmodified DNA for which the average hydrogen bond distance is 2.96 ± 0.12Å (Table 5.2). Recalling that the threshold for hydrogen bond occupancy is set at an electronegative atom distance less than 3.1Å, this accounts for the corresponding disruptions in hydrogen bond occupancies. The DB[a,l]P and B[g]C systems, which do not exhibit disruptions in the dA^*_6-N1:dT_{17}-N3 hydrogen bond, exhibit much smaller increases in stretch and buckle (Figure 5.9 blue and orange bars respectively) that are less than half those seen in the B[c]P (Figure 5.9 green bars) and DB[a,j/c/h]A systems (Figure 5.10 blue, orange, and green bars respectively). There is essentially no change in propeller in the DB[a,l]P and B[g]C systems due to the firmly anti-glycosidic conformation of dA^*_6 in these two systems while the change in propeller in the other strongly preferred systems is due to the borderline syn/anti-glycosidic conformation of dA^*_6 in these systems described above. Note that increases in shear and stagger seen in base pair 6 do not appear to be associated with disruption of the dA^*_6-N1:dT_{17}-N3 hydrogen bond as these increases occur in the DB[a,l]P and B[g]C systems which do not exhibit hydrogen bond disruption as well as the B[c]P and DB[a,j/c/h]A systems which do exhibit hydrogen bond disruption.

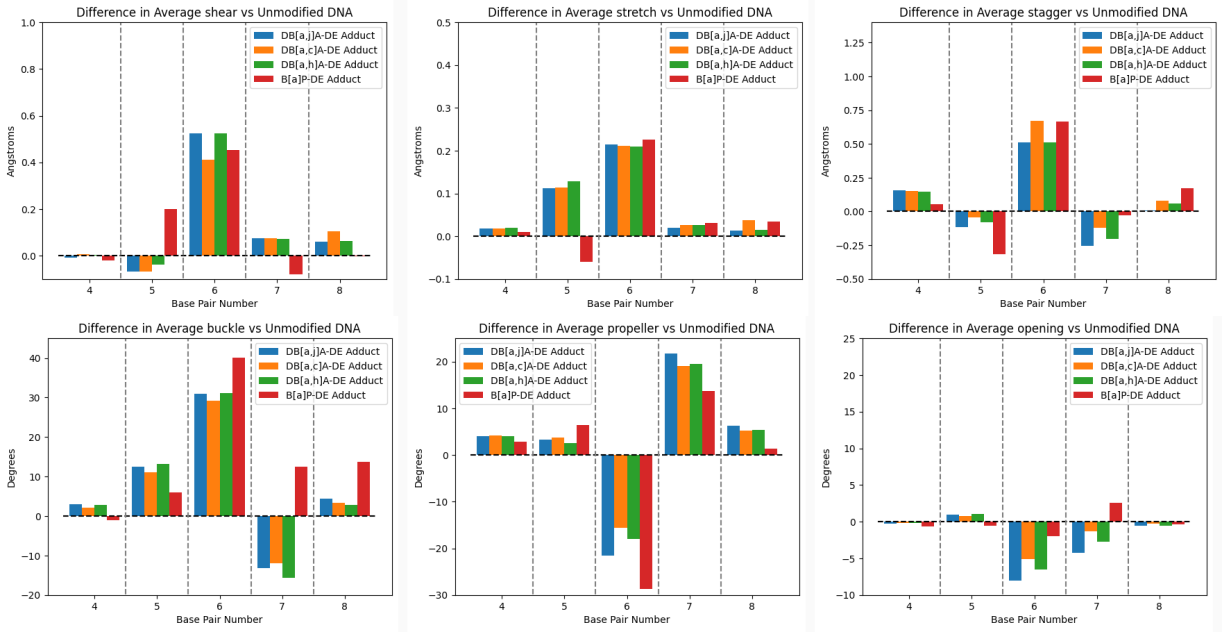
The DB[a,j/c/h]A-DE systems also exhibit decreases ranging from -12.77 to -17.87 percentage points in the dC_5-N3 : dG_{18}-N1 and dC_5-N4 : dG_{18}-O6 hydrogen bond occupancies (Table 4.4). These disruptions are associated with increases in average stretch of approximately 0.1Å and increases in average buckle of approximately 10° in the dC_5 : dG_{18} base pair (base pair 5 in Figure 5.10). These differences in base pair parameters result in increased average distances for the dC_5-

N3 : dG_{18}-N1 and dC_5-N4:dG_{18}-O6 hydrogen bonds ranging from 3.04Å to 3.14Å with standard deviations ranging from 0.14Å to 0.39Å as compared to unmodified DNA for which the average dC_5-N3 : dG_{18}-N1 hydrogen bond distance is 3.00 $\pm$ 0.13Å and the average dC_5-N4:dG_{18}-O6 hydrogen bond distance is 3.05 $\pm$ 0.26Å (Table 5.3).

Note that the dA^*_6-N6 : dT_{17}-O4 and the dA_7-N6 : dT_{16}-O4 hydrogen bond occupancies are increased as compared to unmodified DNA in the DB[a,j/h]A systems (Table 4.4). Increased occupancy of the dA^*_6-N6 : dT_{17}-O4 hydrogen bond is associated with decreased opening in the dA^*_6 : dT_{17} base pair in these systems (base pair 6 in Figure 5.10), resulting in decreased average distances for the dA^*_6-N6 : dT_{17}-O4 hydrogen bond of 2.83 $\pm$ 0.13Å and 2.85 $\pm$ 0.15Å respectively as compared to unmodified DNA for which the average distance is 2.94 $\pm$ 0.19Å. Increased occupancy of the dA_7-N6 : dT_{16}-O4 hydrogen bond is associated with minimal increases in stretch, decreases buckle, increases in propeller, and small decreases in opening of the dA_7 : dT_{16} base pair (base pair 7 in Figure 5.10) that result in marginally decreased average distances for the dA_7-N6 : dT_{16}-O4 hydrogen bond of 2.94 $\pm$ 0.18Å and 2.95 $\pm$ 0.19Å as compared to unmodified DNA for which the average distance is 3.02 $\pm$ 0.19Å (Table 5.3).



(5.9) Strongly preferred fjord PAH-DNA adducts: distortions in average base pair parameters as compared to unmodified DNA

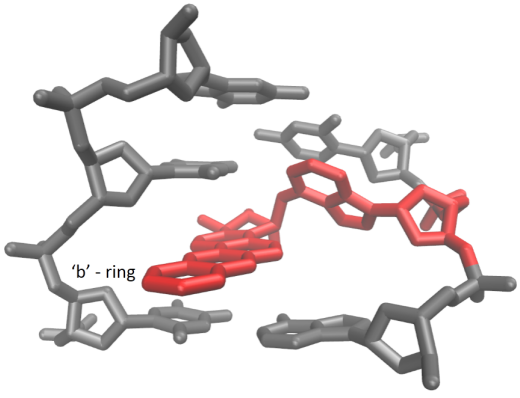


(5.10) Strongly preferred DB[a,j/c/h]A-DNA adducts: distortions in average base pair parameters as compared to unmodified DNA

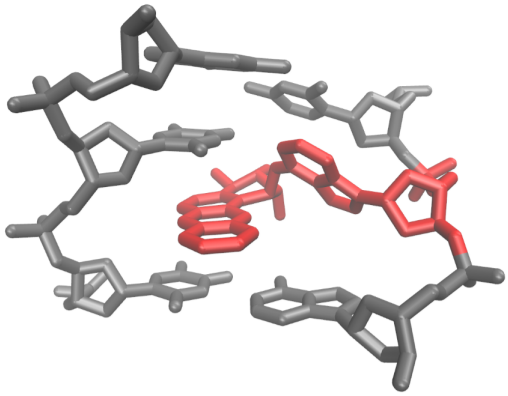
5.2 Weakly Preferred PAH-DNA Adducts

5.2.1 Conformational Details and van der Waals Interactions

The B[b]C and B[a]A systems (Figures 5.11 and 5.12 respectively) assume a conformational motif similar to that described above for the strongly preferred DB[a,j/c/h]A systems with χ_{avg} = -99.64° and -84.86°, α'_{avg} = 49.86° and 41.39°, and β'_{avg} = 91.76° and 107.55° respectively (Table 4.3). The aromatic rings of these PAHs are positioned for strong van der Waals interactions with dT_{16} and dT_{17} in the primary intercalation pocket that are comparable to those observed in the strongly preferred PAHs with $E_{vdW:dT_{16} \mid dT_{17}}$ = -12.19 kcal/mol and -12.49 kcal/mol respectively (Table 4.3). However, both systems lack an aromatic ring positioned for strong van der Waals interactions with dA^*_6 and dA_7 in the secondary intercalation pocket resulting in values of $E_{vdW:dA^*_6 \mid dA_7}$ = -7.70 kcal/mol and -6.63 kcal/mol respectively, that are comparable to those in unmodified DNA between dA_6 and dA_7 in the absence of an intercalated PAH. These weaker stabilizing van der Waals interactions are also a function of the plane of the modified dA^*_6 being on average oriented diagonal to the plane of the aromatic rings of the PAH as described above for the DB[a,j/c/h]A systems.

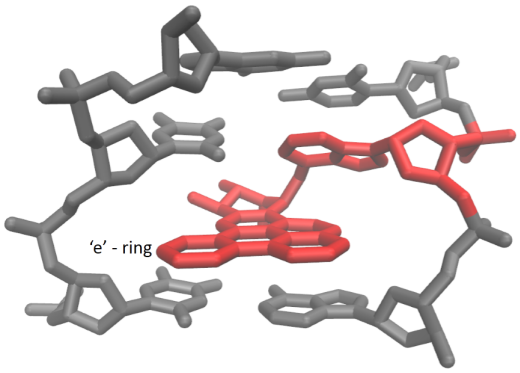


(5.11) Average structure of the B[b]C-DNA adduct

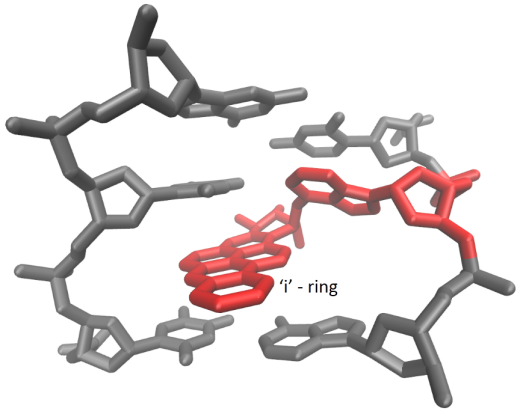


(5.12) Average structure of the B[a]A-DNA adduct

While still intercalating from the major groove with the aromatic rings of the PAH positioned



(5.13) Average structure of the DB[a,e]P-DNA adduct



(5.14) Average structure of the DB[a,i]P-DNA adduct

in the primary and secondary intercalation pockets, the DB[a,e]P and DB[a,i]P systems assume unique conformations that are distinct from those described above in order to accommodate the additional aromatic ring on the 'e' side and 'i' side of the B[a]P root in the respective systems. In the DB[a,e]P system, dA^{*}_{6} assumes an average anti-glycosidic conformation that is firmly anti with χ_{avg} = -140.24° accompanied by average adduct linkage site torsion angles of α'_{avg} = 116.85° and β'_{avg} = 4.68° that position the aromatic rings of DB[a,e]P to avoid steric clashes that would otherwise occur between the additional 'e' aromatic ring and neighboring nucleobases or the DNA backbone in the complementary strand (Figure 5.13). For example, if the DB[a,e]P system were to assume values of α'_{avg} = 55.00° and β'_{avg} = 85.00° similar to those seen in the DB[a,l]P system, the additional 'e' aromatic ring would clash with dT_{16} in the complementary strand. This conformation results in the aromatic rings of the B[a]P root shifting out of the primary dT_{16} | dT_{17} intercalation pocket and partially shifting into the secondary dA^{*}_{6} | dA_{7} intercalation pocket, leaving only the additional 'e' aromatic ring in an ideal position for van der Waals interactions with the dT_{16} and dT_{17} nucleobases. This is reflected in the diminished van der Waals interactions in the primary intercalation pocket of the DB[a,e]P system as compared to other systems, with $E_{vdW:dA_{16} \mid dA_{17}}$ = -8.23 kcal/mol which is 4 to 5 kcal/mol weaker than van der Waals interactions observed in the primary intercalation pocket in the majority of other systems examined (Table 4.3). In the

secondary dA^*_6 | dA_7 intercalation pocket of the DB[a,e]P system, $E_{vdW:dA^*_6 \mid dA_7}$ = -9.88 kcal/mol, which is stronger than most systems and is comparable to the DB[a,j]A system which has an aromatic ring situated in the secondary intercalation pocket. This results from the aromatic rings of the B[a]P root being partially shifted into the secondary intercalation pocket and the plane of the firmly anti-glycosidic dA^*_6 being nearly parallel to the plane of the aromatic rings in DB[a,e]P.

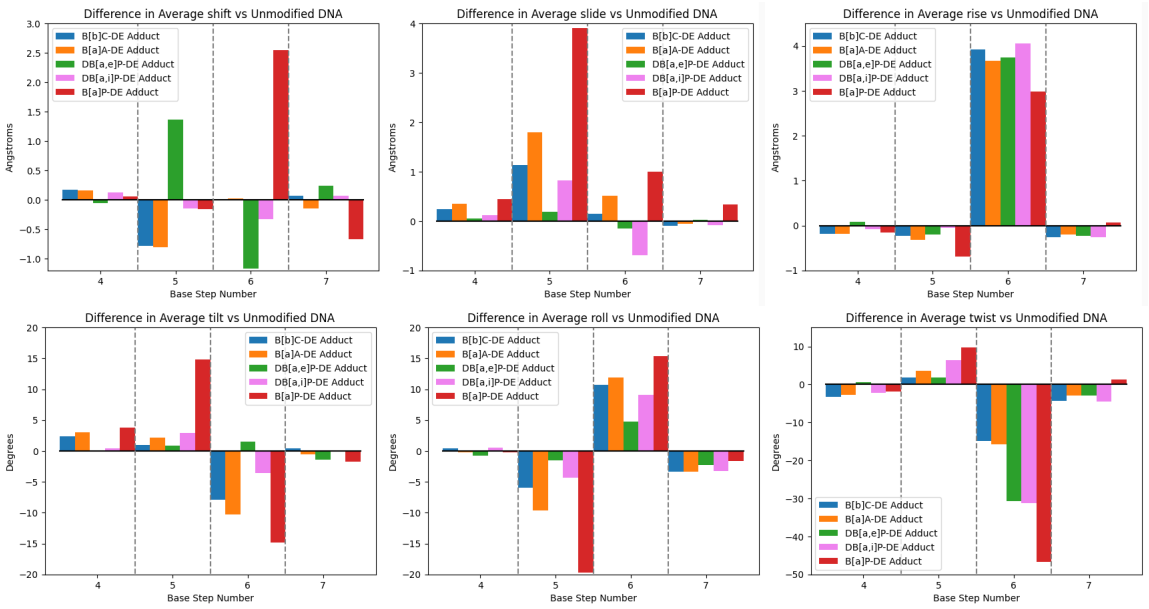
In the DB[a,i]P system, dA^*_6 assumes an average anti-glycosidic conformation with χ_{avg} = -122.28° while the adduct linkage torsion angles take average values of α'_{avg} = 74.78° ± 42.30° and β'_{avg} = 69.00° ± 48.95° (Table 4.3). Note that the large standard deviations in average values for α' and β' and visual examination of the slide, rise, twist, and tilt trajectories for this system (Appendix B) indicate that there is a conformational change in the DNA duplex that lasts approximately 10ns before returning to the equilibrated structure. For most of the trajectory, α' and β' fluctuate about approximate average values of 50° and 100° respectively, which is similar to the average values observed in the strongly preferred systems. Additionally, in the average structure calculated by NAFlex depicted in Figure 5.14, χ = 118.49°, α' = 64.27°, and β' = 80.14° which is very similar to the conformations assumed by the DB[a,l]P and B[g]C systems. This conformation places the 'i' aromatic ring in a position to avoid steric clashes with the sugar phosphate backbone in both strands of the DNA duplex and results in the aromatic rings of DB[a,i]P being ideally situated for strong van der Waals interactions in the primary intercalation pocket. As a result, the van der Waals interactions in the primary intercalation pocket are among the strongest observed with $E_{vdW:dT_{16} \mid dT_{17}}$ = -13.31 kcal/mol. In the secondary intercalation pocket, the aromatic rings of DB[a,i]P are positioned for van der Waals interactions with the dA^*_6 and dA_7 nucleobases that are comparable to those seen in the strongly preferred systems with $E_{vdW:dA^*_6 \mid dA_7:DB[a,i]P\text{-}DE}$ = -8.70 kcal/mol (Table 4.3). This results in $E_{vdW:Intercalation}$ = -22.01 kcal/mol, with the DB[a,i]P system exhibiting stronger total van der Waals interactions from intercalation than several systems with greater relative binding affinity. Further examination of this case will be the subject of future work.

5.2.2 Rigid-Body Parameters and Hydrogen Bonding

Similar to the strongly preferred PAHs, the weakly preferred PAHs exhibit a large increase in average rise of approximately +3.5Å to +4.0Å in base step 6 to accommodate the PAH between the dT_{16} and dT_{17} nucleobases of the primary intercalation pocket and the dA^*_6 and dA_7 nucleobases of the secondary intercalation pocket. There is a decrease in average twist of approximately -15° in the B[b]C and B[a]A systems while the DB[a,e]P and DB[a,i]P systems exhibit a larger decrease of approximately -30° (Figure 5.15). The larger decrease in average twist seen in the DB[a,e]P and DB[a,i]P systems is associated with widening of the major and minor grooves of the DNA duplex to accommodate the larger six ring structure (five aromatic and one aliphatic) of these PAHs in the hydrophobic core of the DNA duplex and is associated with a higher total energy in the NRAS(Q61) 3-mer of E_{Total} = -29.22 kcal/mol and -36.83 kcal/mol respectively (Table 5.1). The smaller five and four ring B[b]C and B[a]A systems which require a lesser decrease in twist to accommodate the PAHs correspondingly have lower total energy configurations of E_{Total} = -46.00 kcal/mol and -46.76 kcal/mol respectively (Table 5.1). As with the strongly preferred PAHs described above, the increase in average rise among the weakly preferred PAHs is greater than the +2.98Å seen in the B[a]P system and the decrease in average twist is less than the -46.76° seen in the B[a]P system. Correspondingly, the weakly preferred PAHs assume lower energy conformations of the NRAS(Q61) 3-mer as compared to the B[a]P system where E_{Total} = -7.31 kcal/mol.

Similar to the DB[a,j/c/h]A systems described above, the B[b]C and B[a]A systems exhibit distortions similar to the B[a]P system of lesser magnitude. There are increases in slide at base step 5, decreases in tilt at base step 6, decreases in roll at base steps 5, and increases in roll at base step 6 (Figure 5.15 blue, orange, and red bars respectively). As in the DB[a,j/c/h]A systems, these distortions are associated with the average dA^*_6 glycosidic torsion angle assumed by these systems where the B[b]C system assumes an average anti-glycosidic conformation with χ_{avg} = -99.64° ± 26.25° while the B[a]A system assumes an average syn-glycosidic conformation with χ_{avg} = -84.86° ± 17.79° resulting in conformations that fluctuate between syn and anti-glycosidic

with the plane of the dA^*_6 nucleobase on average oriented diagonal to the plane of the neighboring nucleobases, requiring the distortions in slide, tilt, and roll to accommodate the diagonal orientation of dA^*_6. The DB[a,e]P system exhibits a notable increase in average shift at base step 5 and a notable decrease at base step 6 that is not observed in other systems, while exhibiting minimal differences in slide, tilt, and roll at base steps 5 and 6 (Figure 5.15 green bars). The DB[a,i]P system meanwhile exhibits minimal differences in shift and tilt at base steps 5 and 6, small differences in slide at base steps 5 and 6, and a moderate increase in roll at base step 6 (Figure 5.15 violet bars).



(5.15) Weakly preferred PAH-DNA adducts: distortions in average base step parameters as compared to unmodified DNA

The B[b]C and B[a]A systems exhibit disruptions in hydrogen bonding and distortions in base pair parameters that are again very similar to those observed in the DB[a,j/c/h]A systems with decreases of -19.45 and -12.36 percentage points in the dC_5-N3 : dG_{18}-N1 hydrogen bond occupancy and decreases of -14.74 and -10.28 percentage points in the dC_5-N4 : dG_{18}-O6 hydrogen bond occupancy respectively (Table 4.4). These disruptions are associated with increases in average stretch of +0.13Å and +0.10Å and increases in average buckle of +12.43° and +12.44° in base pair 5 of the B[b]C and B[a]A systems respectively. There are minimal differences in average propeller in both systems (Figure 5.16 blue and orange bars). These differences in base pair 5

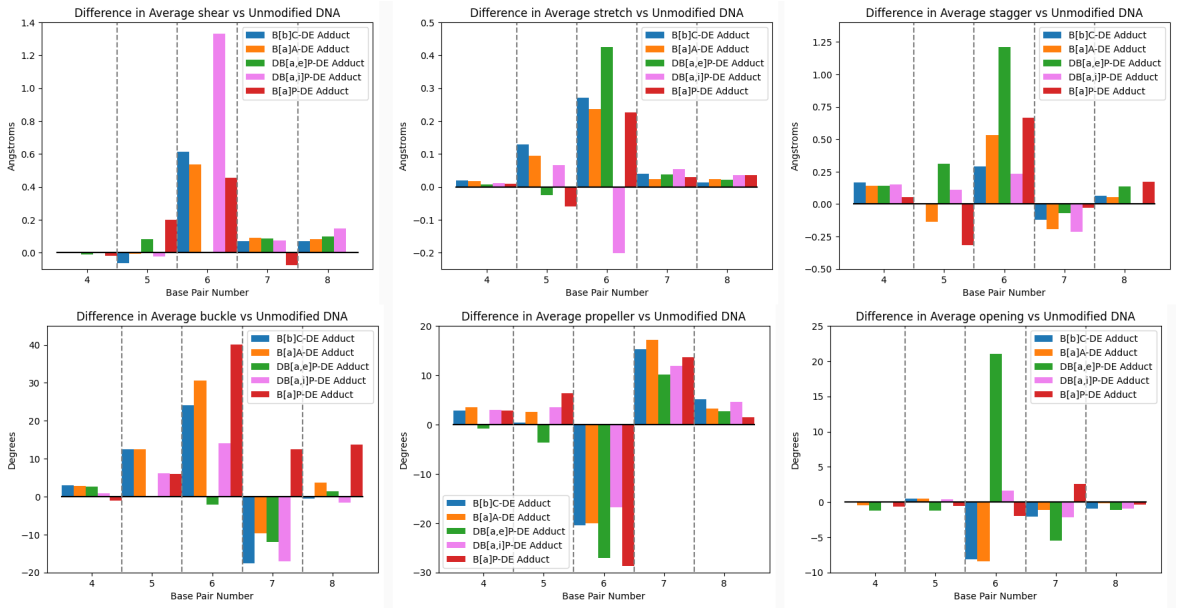
parameters result in increased average distances for the dC_5-N3 : dG_{18}-N1 hydrogen bond of 3.06 $\pm$ 0.16Å and 3.04 $\pm$ 0.15Å as compared to unmodified DNA and increased average distances for the dC_5-N4 : dG_{18}-O6 hydrogen bond of 3.14 $\pm$ 0.32Å and 3.10 $\pm$ 0.30Å for the B[b]C and B[a]A systems respectively (Table 5.2).

The B[b]C and B[a]A systems also exhibit disruptions of -25.11 and -26.15 percentage points the dA^*_6-N1 : dT_{17}-N3 hydrogen bond respectively. These disruptions are associated with distortions in base pair 6 parameters consisting of increases in average stretch of +0.27Å and +0.24Å, increases in average buckle of +24.17° and +30.53°, and decreases in average propeller of -20.39° and -20.04° respectively. These distortions result in increased average distances for the dA^*_6-N1 : dT_{17}-N3 hydrogen bond of 3.22 $\pm$ 0.53Å and 3.12 $\pm$ 0.30Å in the B[b]C and B[a]A systems respectively as compared to unmodified DNA (Table 5.2). Note that decreases in average opening in base pair 6 serve to increase occupancy for the dA^*_6-N6 : dT_{17}-O4 hydrogen bond (Table 5.3) as compared to unmodified DNA by decreasing the average distance of the hydrogen bond in both systems (Table 5.2).

In the DB[a,e]P system, the -20.47 percentage point disruption in the dA^*_6-N1 : dT_{17}-N3 hydrogen bond occupancy is accompanied by a large -73.64 percentage point disruption of the dA^*_6-N6 : dT_{17}-O4 hydrogen bond (Table 4.4). These hydrogen bond disruptions are primarily associated with a large increase in average stretch of +0.42Å, a large decrease in average propeller of -27.06°, and large increase in average opening of +21.04° in base pair 6, while the difference in buckle as compared to unmodified DNA is minimal (Figure 5.16 - green bars). These large differences in stretch, propeller, and opening result in increased average distances of 3.06 $\pm$ 0.18Å for the dA^*_6-N1 : dT_{17}-N3 hydrogen bond and 4.63 $\pm$ 0.63 Å for dA^*_6-N6 : dT_{17}-O4 hydrogen bond as compared to unmodified DNA (Table 5.2). Decreased occupancy of the dA_7-N1 : dT_{16}-N3 hydrogen bond and increased occupancy of the dA_7-N6 : dT_{16}-O4 hydrogen bond is associated with a minimal increase in stretch, a decrease buckle, an increase in propeller, and a decrease in opening of the dA_7 : dT_{16} base pair (Figure 5.16 - base pair 7 - green bars). This results in an increased average distance of 2.99 $\pm$ 0.14Å for the dA_7-N1 : dT_{16}-N3 hydrogen bond as compared to unmodified

DNA and a decreased average distance of 2.89 $\pm$ 0.16Å for the dA$_7$-N6 : dT$_{16}$-O4 hydrogen bond as compared to unmodified DNA (Table 5.2).

In the DB[a,i]P system, the -30.01 percentage point disruption in the dA$^*_6$-N1 : dT$_{17}$-N3 hydrogen bond occupancy is also accompanied by a -19.16 percentage point disruption of the dA$^*_6$-N6 : dT$_{17}$-O4 hydrogen bond (Table 4.4). These hydrogen bond disruptions are primarily associated with an increase in average buckle of +14.15° and a decrease in average propeller of -16.73° in base pair 6, while there is a decrease in average stretch of -0.20Å as compared to unmodified DNA (Figure 5.16 - violet bars). This results in increased average distances of 3.68 $\pm$ 1.38Å and 3.89 $\pm$ 1.70 Å for the dA$^*_6$-N1 : dT$_{17}$-N3 and dA$^*_6$-N6 : dT$_{17}$-O4 hydrogen bonds as compared to unmodified DNA (Table 5.2). There is also a -13.71 percentage point disruption of the dC$_5$-N3 : dG$_{18}$-N1 hydrogen bond occupancy that is associated with a small increase of +0.07Å in stretch, a small increase of +6.22° in buckle, and a small increase in propeller of +3.55° in base step 5. This results in an increased average distance for the dC$_5$-N3 : dG$_{18}$-N1 of 3.04 $\pm$ 0.17Å as compared to unmodified DNA (Table 5.2).

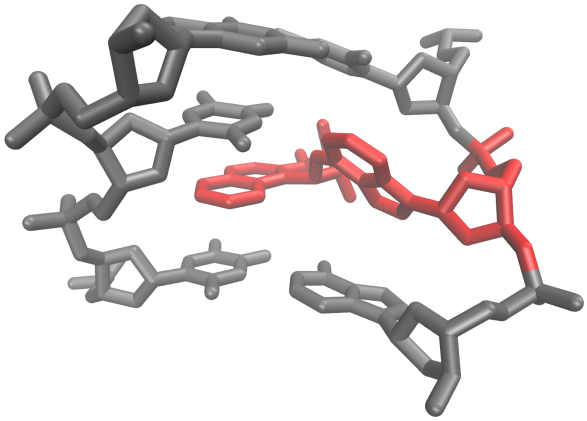


(5.16) Weakly preferred PAH-DNA adducts: distortions in average base pair parameters as compared to unmodified DNA

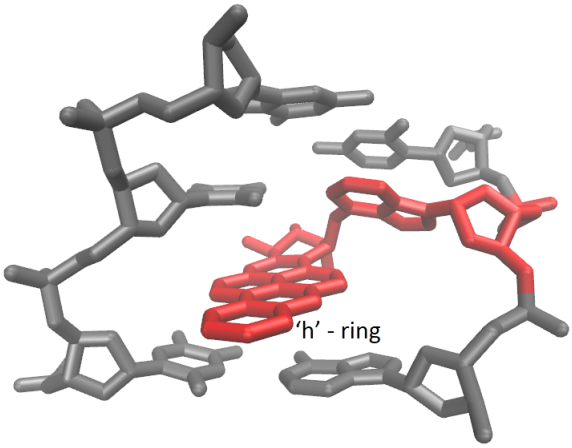
5.3 Equally Preferred PAH-DNA Adducts

5.3.1 Conformational Details and van der Waals Interactions

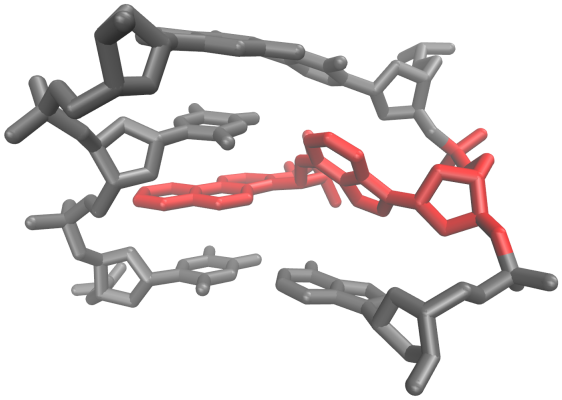
The PHE and CHR systems (Figures 5.17 and 5.19 respectively) assume a conformational motif similar to that of the B[c]P and B[a]P systems where the PAH intercalates from the major groove with its aromatic rings positioned solely in the primary dT_{16} | dT_{17} intercalation pocket. The dA^{*}_{6} nucleobase assumes an average syn-glycosidic conformation with χ_{avg} = -64.72° and -75.96° respectively and the adduct linkage site torsion angles take average values of α'_{avg} = 16.20° and 25.28° and β'_{avg} = 146.05° and 127.71° respectively (Table 4.3). These are similar to the values of χ_{avg}, α'_{avg}, and β'_{avg} observed in the B[c]P and B[a]P systems and they result in the aromatic rings of the PHE and CHR systems being positioned for van der Waals interactions with the dT_{16} and dT_{17} nucleobases in the primary intercalation pocket while van der Waals interactions with the dA^{*}_{6} and dA_{7} nucleobases are comparatively weak. In the PHE system $E_{vdW:dA^{*}_{6} \mid dA_{7}}$ = -10.14 kcal/mol in the primary intercalation pocket while there are minimal van der Waals interactions in the secondary intercalation pocket with $E_{vdW:dT_{16} \mid dT_{17}}$ = -1.16 kcal/mol, resulting in $E_{vdW:Intercalation}$ = -11.30 kcal/mol being the weakest total van der Waals interactions from intercalation observed among the PAHs examined. The PHE system is the smallest of all the PAHs examined with only two aromatic rings that π-stack in the primary intercalation pocket while essentially not interacting with the secondary intercalation pocket. Similarly in the CHR system $E_{vdW:dA^{*}_{6} \mid dA_{7}}$ = -12.32 kcal/mol in the primary intercalation pocket while there are minimal van der Waals interactions in the secondary intercalation pocket with $E_{vdW:dT_{16} \mid dT_{17}}$ = -3.94 kcal/mol (Table 4.3).



(5.17) Average structure of the PHE-DNA adduct



(5.18) Average structure of the DB[a,h]P-DNA adduct



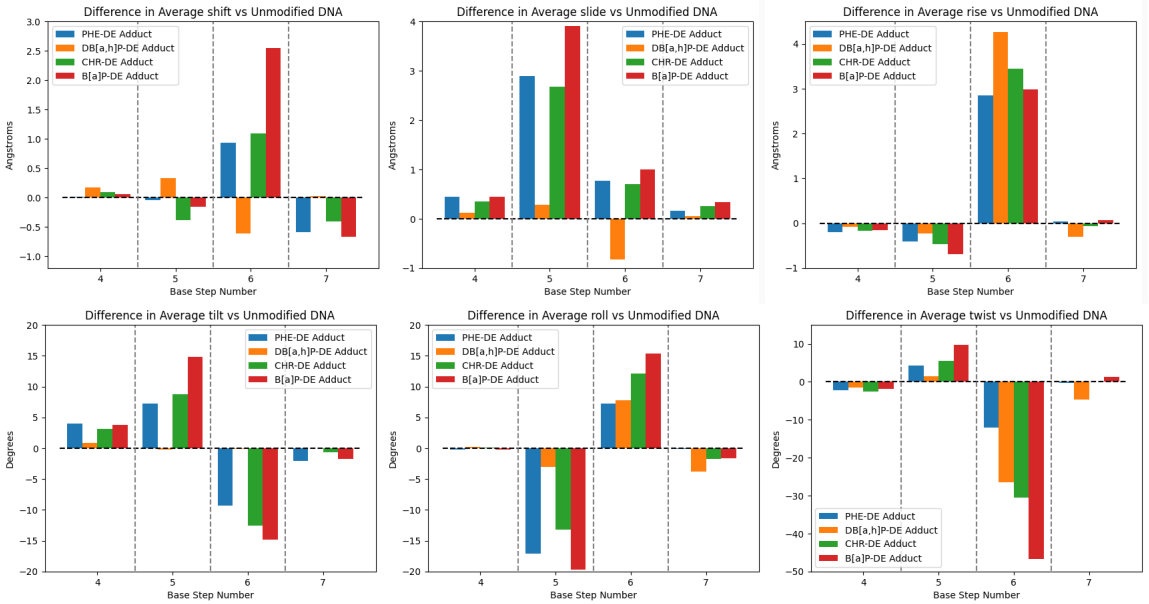
(5.19) Average structure of the CHR-DNA adduct

The DB[a,h]P system assumes a conformation similar to the DB[a,i]P system where dA^*_6 assumes an average anti-glycosidic conformation with χ_{avg} = -133.20°, resulting in the plane of the dA^*_6 nucleobase being oriented diagonal to the plane of the aromatic rings in the DB[a,h]P (Figure 5.18). The adduct linkage torsion angles take average values of α'_{avg} = 73.51° ± 34.78° and β'_{avg} = 66.19° ± 40.42° (Table 4.3). This conformation places the 'h' aromatic ring in a position to avoid steric clashes with the sugar phosphate backbone in both strands of the DNA duplex and results in the aromatic rings of the B[a]P root being situated partly in the primary intercalation pocket and partly in the secondary intercalation pocket. As a result, van der Waals interactions in the primary intercalation pocket are weaker than in most systems with $E_{vdW:dT_{16} \mid dT_{17}}$ = -11.83 kcal/mol while van der Waals interactions in the secondary intercalation pocket are stronger than in most systems with $E_{vdW:dA^*_6 \mid dA_7}$ = -10.13 kcal/mol. As a result $E_{vdW:Intercalation}$ = -21.96 kcal/mol indicating that the DB[a,h]P system has stronger total van der Waals interactions from intercalation than several systems with greater relative binding affinity. Further examination of this case will be the subject of future work. Note also that the large standard deviations in average values for α' and β' and examination of equilibration trajectories (Appendix B) indicate that there are five short yet distinct segments of the trajectory where the DB[a,h]P adduct shifts within the primary and secondary intercalation pockets, temporarily assuming a higher energy intercalated conformation.

5.3.2 Rigid-Body Parameters and Hydrogen Bonding

With three rings (two aromatic and one aliphatic), PHE is the smallest of the PAHs examined in this work and correspondingly exhibits the smallest increase in average rise at +2.86Å and the smallest decrease in average twist at -12.06° in order to accommodate the PHE-DNA adduct. The larger five ring DB[a,h]P and four ring CHR systems exhibit increases in average rise of +4.26Å and +3.43Å and decreases in average twist of -26.41° and -30.60° respectively, which are comparable to those seen in the systems examined above in order to accommodate the PAH-DNA adduct. Correspondingly, the NRAS(Q61) 3-mer in the PHE system has a lower total energy of E_{Total} = -43.48 kcal/mol as compared to the DB[a,h]P and CHR systems where E_{Total} = -37.37 kcal/mol

and -27.24 kcal/mol respectively. The PHE and CHR systems assume a conformational motif analogous to the B[a]P system as described above and correspondingly exhibit distortions similar to the B[a]P system of lesser magnitude. There are increases in shift at base step 6, increases in slide at base steps 5 and 6, increases in tilt at base step 5, decreases in tilt at base step 6, decreases in roll at base step 5, and increases in roll at base step 6 (Figure 5.20 - blue, green, and red bars respectively). The DB[a,h]P system meanwhile exhibits small increases in average shift and slide and a small decrease in average roll at base step 5, small decreases in average shift and slide and an increase in roll at base step 6, and negligible differences in tilt at base steps 5 and 6 (Figure 5.20 - orange bars).



(5.20) Equally preferred PAH-DNA adducts: distortions in average base step parameters as compared to unmodified DNA

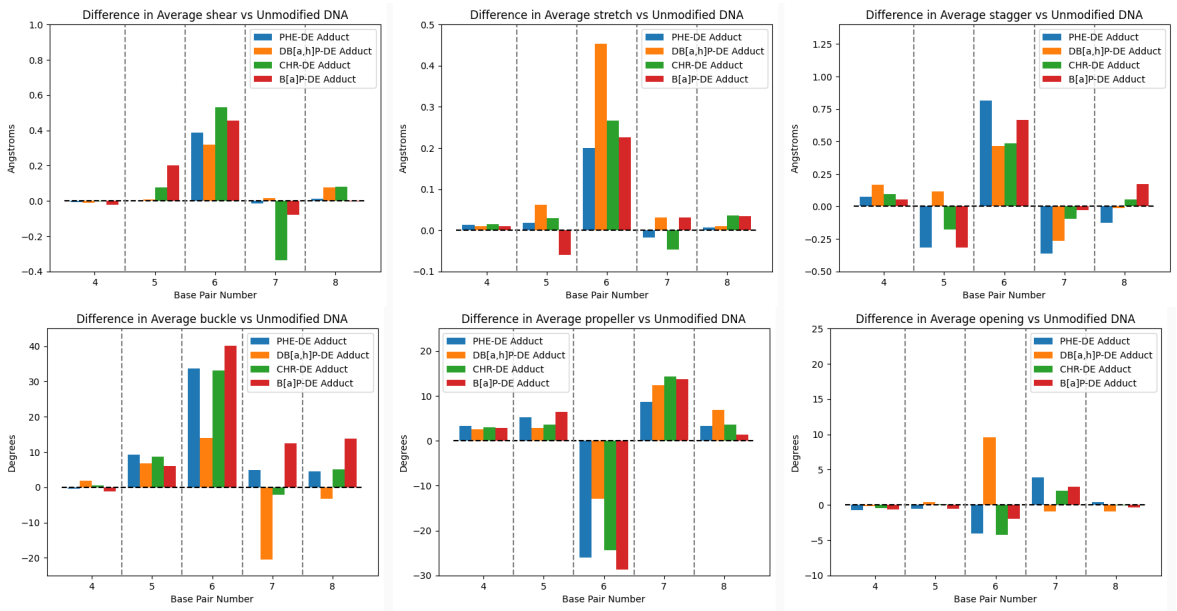
In the PHE system, the dA^*_6-N1 : dT_{17}-N3 and dA^*_6-N6 : dT_{17}-O4 hydrogen bonds exhibit decreases of -9.94 and -9.21 percentage points in occupancy. These are associated with increases of +0.20Å in average stretch and +33.63° in average buckle as well as decreases of -26.02° in average propeller and -4.09° in average opening at base pair 6 (Figure 5.21 - blue bars). Corresponding increases in average hydrogen bond distance are minimal (Table 5.2) and account for the marginal decreases in hydrogen bond occupancy in the dA^*_6 : dT_{17} base pair. There is a decrease of -14.47

percentage points in the dA_7-N6 : dT_{16}-O4 hydrogen bond occupancy associated with increases of +4.86° in buckle, +8.64° in propeller, and +3.93° in opening in base pair 7 (Figure 5.21 - blue bars). This results in an increased average distance of 3.11Å ± 0.28Å for the dA_7-N6 : dT_{16}-O4 hydrogen bond as compared to unmodified DNA (Table 5.2).

In the DB[a,h]P system, the dC_5-N3 : dG_{18}-N1 hydrogen bond occupancy exhibits a decrease of -14.23 percentage points associated with a combination of small differences of +0.06Å in stretch, +6.79° in buckle, and +2.82° in propeller in base step 5 (Figure 5.21 - orange bars), resulting in an increased average distance for the dC_5-N3 : dG_{18}-N1 hydrogen bond of 3.05Å ± 0.19Å as compared to unmodified DNA (Table 5.2). There are also decreases of -18.95 and -22.88 percentage points in the dA^*_6-N1 : dT_{17}-N3 and dA^*_6-N6 : dT_{17}-O4 hydrogen bond occupancies associated with increases of +0.45Å in average stretch and +13.89° in average buckle, a decrease of -12.96° in average propeller, and an increase of +9.61° in average opening at base step 6 (Figure 5.21 - orange bars). This results in an increased average distance of 3.21Å ± 0.66Å for the dA^*_6-N1 : dT_{17}-N3 hydrogen bond. The dA^*_6-N6 : dT_{17}-O4 hydrogen bond exhibits an increased average distance of 3.67Å ± 1.43Å and a decreased average dA^*_6-N6 : dA^*_6-H61 : dT_{17}-O4 hydrogen bond angle of 132.69° ± 55.35° as compared to unmodified DNA for which the average hydrogen bond angle is 161.53° ± 10.90°. The relatively large differences in base pair parameters, the large standard deviations in average hydrogen bond length and angle, and examination of the dA_7 : dT_{16} base pair hydrogen bond trajectories indicate that the base pair intermittently separates during the equilibration simulation with both bases remaining in the hydrophobic core of the DNA duplex, but dT_{16} shifting away from dA_7 and toward the major groove (Appendix B).

In the CHR system, there are decreases of -18.98 and -12.06 percentage points in the dA^*_6-N1 : dT_{17}-N3 and dA_7-N6 : dT_{16}-O4 hydrogen bond occupancies respectively. Disruptions of the dA^*_6-N1 : dT_{17}-N3 hydrogen bond are concomitant with differences of +0.27Å in average stretch, +33.07° in average buckle, -24.35° in average propeller, and -4.27° in average opening in base pair 6 (Figure 5.21 - green bars). This results in an increased average distance of 3.12Å ± 0.40Å for the dA^*_6-N1 : dT_{17}-N3 hydrogen bond as compared to unmodified DNA (Table 5.2). Disruption of the

dA_7-N6 : dT_{16}-O4 hydrogen bond is associated with differences of +14.25°in average propeller and +1.98° in average opening in base pair 7 (Figure 5.21 - green bars) resulting in an increased average hydrogen bond distance of 3.15Å $\pm$ 0.82Å as compared to unmodified DNA (Table 5.2).



(5.21) Equally preferred PAH-DNA adducts: distortions in average base pair parameters as compared to unmodified DNA

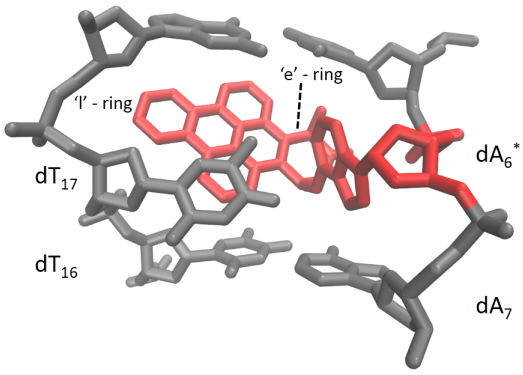
5.4 Non-Preferred PAH-DNA Adducts

5.4.1 Conformational Details and van der Waals Interactions

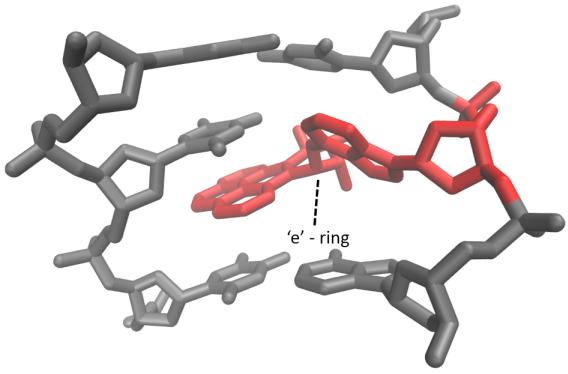
The DB[e,l]P system assumes a major groove conformation where its aromatic rings do not intercalate in either the primary $dT_{16} \mid dT_{17}$ or secondary $dA_6^* \mid dA_7$ intercalation pockets (Figure 5.22). The dA_6^* nucleobase assumes a syn-glycosidic conformation with χ_{avg} = -61.88° while the adduct linkage torsion angles assume average values of α''_{avg} = -0.14° and β'_{avg} = -33.89°, placing the plane of the aromatic rings of DB[e,l]P-DNA adduct nearly perpendicular to the plane of the neighboring $dC_5 : dG_{18}$ and $dA_7 : dT_{16}$ base pairs. Note we have used α'' that measures the N1-C6-N6-C20 dihedral as opposed to α' that measures the C5-C6-N6-C20 dihedral to avoid averaging positive and negative values assumed by α' in the equilibration trajectory of this system. This conformation results in minimal van der Waals interactions in the secondary intercalation pocket with $E_{vdW:dA_6^* \mid dA_7}$ = -1.22 kcal/mol. The dT_{17} nucleobase assumes an unusual syn-glycosidic conformation with a torsion angle of -67.33°, resulting in the plane of the dT_{17} nucleobase being oriented nearly parallel to the aromatic rings of the DB[e,l]P-DNA adduct and allowing for moderate van der Waals interactions. Meanwhile the plane of the dT_{16} nucleobase is nearly perpendicular to the plane of the aromatic rings of the DB[e,l]P-DNA adduct. As a result $E_{vdW:dT_{16} \mid dT_{17}}$ = -7.73 kcal/mol, which is the lowest observed among the PAH-DNA adducts examined, and comparable to van der Waals interactions between dT_{16} and dT_{17} in unmodified DNA. As a result, $E_{vdW:Intercalation}$ = -8.95, indicating that van der Waals interactions do not have an overall stabilizing impact in the DB[e,l]P system (Table 4.3).

The B[e]P system assumes a conformation similar to that assumed by the B[c]P, PHE, CHR, and B[a]P systems, where the B[e]P-DNA adduct intercalates from the major groove with its aromatic rings positioned solely in the primary intercalation pocket (Figure 5.23). The dA_6^* nucleobase assumes an average syn-glycosidic conformation with χ_{avg} = -71.42° and adduct linkage torsion angles of α'_{avg} = 28.53° and β'_{avg} = 137.48° resulting in strong van der Waals interactions in the primary intercalation pocket with $E_{vdW:dT_{16} \mid dT_{17}}$ = -12.92 kcal/mol while van der Waals interactions

in the secondary intercalation pocket are weak with $E_{vdW:dA^{*}_{6} \mid dA_7}$ = -2.81 kcal/mol. As a result, $E_{vdW:Intercalation}$ = -15.73 kcal/mol, which is identical to the B[a]P-DNA adduct system (Table 4.3).



(5.22) Average structure of the DB[e,l]P-DNA adduct

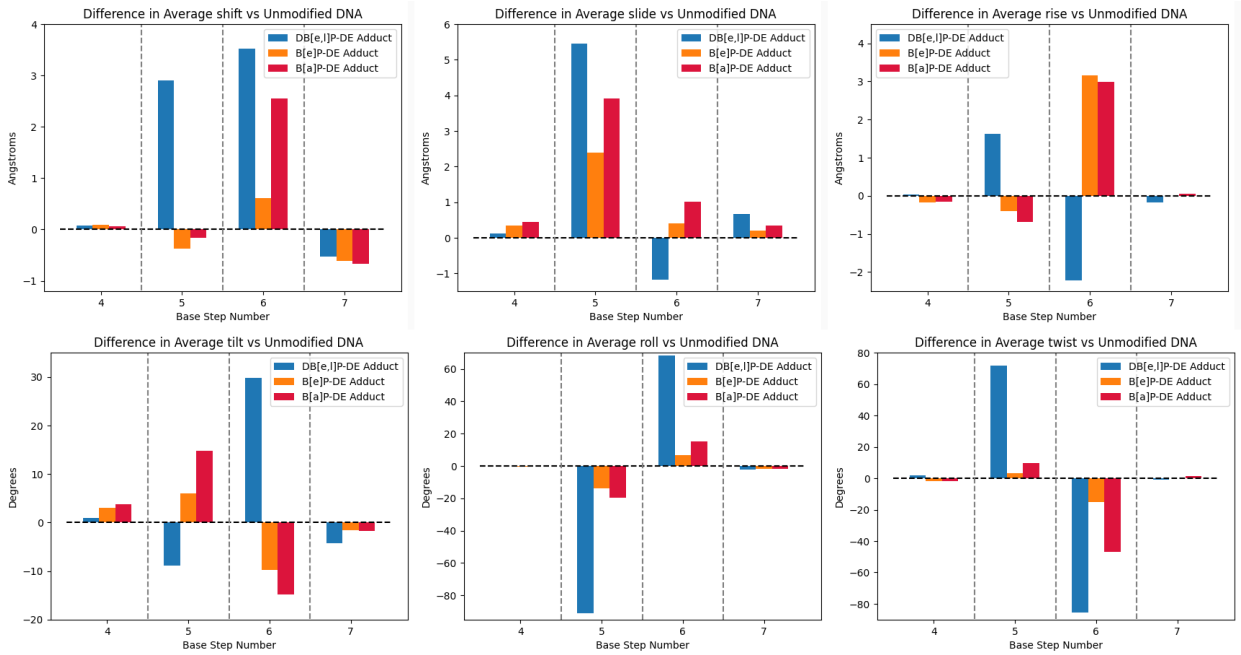


(5.23) Average structure of the B[e]P-DNA adduct

5.4.2 Rigid-Body Parameters and Hydrogen Bonding

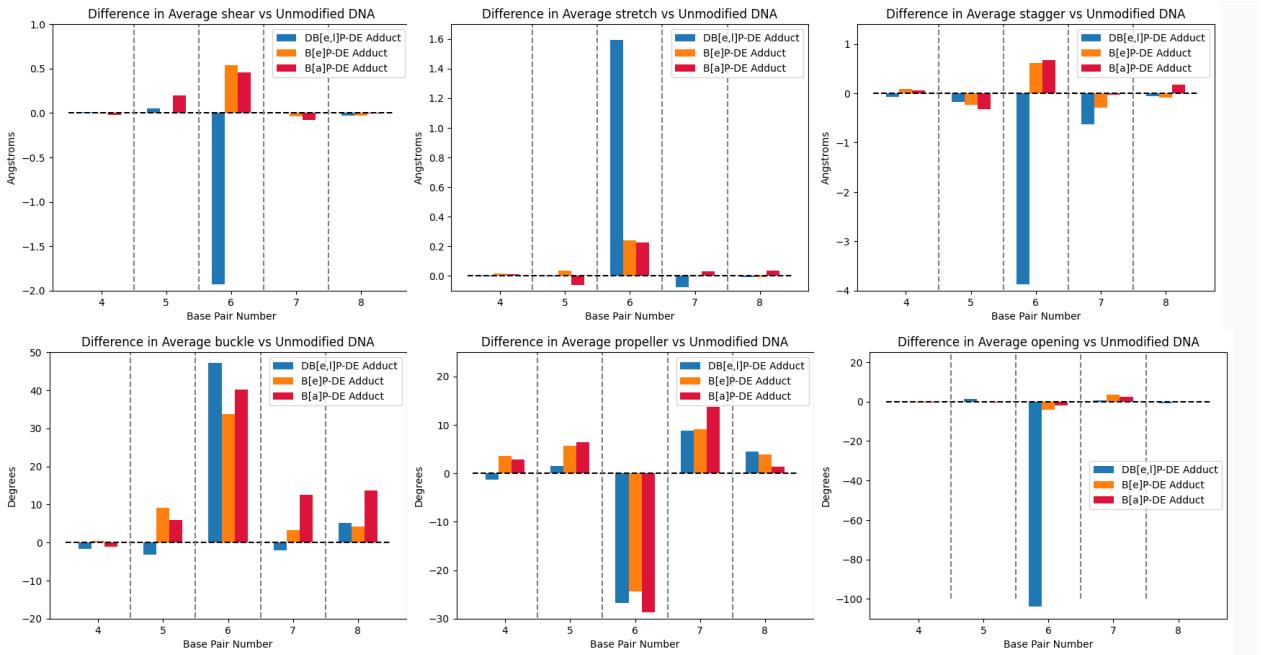
As a result of its non-intercalated major groove conformation, the DB[e,l]P-DNA adduct system is largely distorted in comparison to other PAH-DNA adduct systems examined in this work. The system exhibits a decrease in average rise of -2.23Å at base step 6, and is the only PAH-DNA adduct examined that does not exhibit an increase in average rise. Meanwhile, the decrease in average twist of -85.76° at base step 6 is much larger than that seen in the B[a]P system. At the dA^*_6 | dC_7 base step, average values of remaining rigid-body parameters differ by +3.52Å for shift, -1.19Å for slide, +29.84° for tilt, and +68.49° for roll (Figure 5.24 - blue bars). At the dC_5 | dA^*_6 base step, there are also large differences in shift (+2.91Å), slide (+5.45Å), rise (+1.63Å), tilt (-8.83°), roll (-91.15°), and twist (+71.66°) not observed in other PAH-DNA adduct systems examined. Correspondingly, the NRAS(Q61) 3-mer takes a high total energy value of E_{Total} = -14.21 kcal/mol (Table 5.1).

In the B[e]P-DNA adduct system, the increase in average rise at base step 6 is similar to that seen in B[a]P at +3.15Å while the decrease in average twist at base step 6 is smaller at -15.25°. As in the systems above that assume the same conformational motif, the B[e]P system exhibits distortions shift, slide, tilt, and roll that are similar to the B[a]P system of lesser magnitude (Figure 5.24 - orange bars). The total energy in the NRAS(Q61) 3-mer takes a value of E_{Total} = -41.17 kcal/mol similar to that seen in the PHE and B[c]P systems (Table 5.1). The DB[e,l]P system exhibits disruptions in hydrogen bonding in the dA^*_6:dT_{17} base pair with a decrease of -86.26 percentage points in the dA^*_6-N1:dT_{17}-N3 hydrogen bond occupancy, and a decrease of -80.10 percentage points in the dA^*_6-N6:dT_{17}-O4 hydrogen bond occupancy (Table 4.4). Disruption of hydrogen bonding in the dA^*_6:dT_{17} base pair is accompanied by large differences in stretch of +1.59Å, in buckle of +47.24°, in propeller of -26.76°, and in opening of -103.79° (Figure 5.25 - base pair 6 - orange bars). These differences in base pair parameters result in an increased average distance of 6.23 $\pm$ 0.85Å for the dA^*_6-N1:dT_{17}-N3 hydrogen bond and an increased average distance of 5.06 $\pm$ 1.19Å for the dA^*_6-N6:dT_{17}-O4 hydrogen bond as compared to unmodified DNA (Table 5.2).



(5.24) Non-preferred PAH-DNA adducts: distortions in average base step parameters as compared to unmodified DNA

The B[e]P system exhibits disruptions in hydrogen bonding in the dA^{*}_{6}:dT_{17} and dA_{7}:dT_{16} base pairs with a decrease of -19.05 percentage points in the dA^{*}_{6}-N1:dT_{17}-N3 hydrogen bond occupancy, and a decrease of -14.15 percentage points in the dA_{7}-N6:dT_{16}-O4 hydrogen bond occupancy (Table 4.4). Disruption of hydrogen bonding in the dA^{*}_{6}:dT_{17} base pair is accompanied by differences in stretch of +0.24Å, in buckle of +33.76°, in propeller of -24.42°, and in opening of -4.03° (Figure 5.25 - base pair 6 - orange bars). These differences in base pair parameters result in an increased average distance of 3.09 $\pm$ 0.31Å for the dA^{*}_{6}-N1:dT_{17}-N3 hydrogen bond and an increased average distance of 3.11 $\pm$ 0.29Å for the dA_{7}-N6:dT_{16}-O4 hydrogen bond as compared to unmodified DNA (Table 5.3).



(5.25) Non-preferred PAH-DNA adducts: distortions in average base pair parameters as compared to unmodified DNA

CHAPTER 6

Appendices

6.1 Appendix A

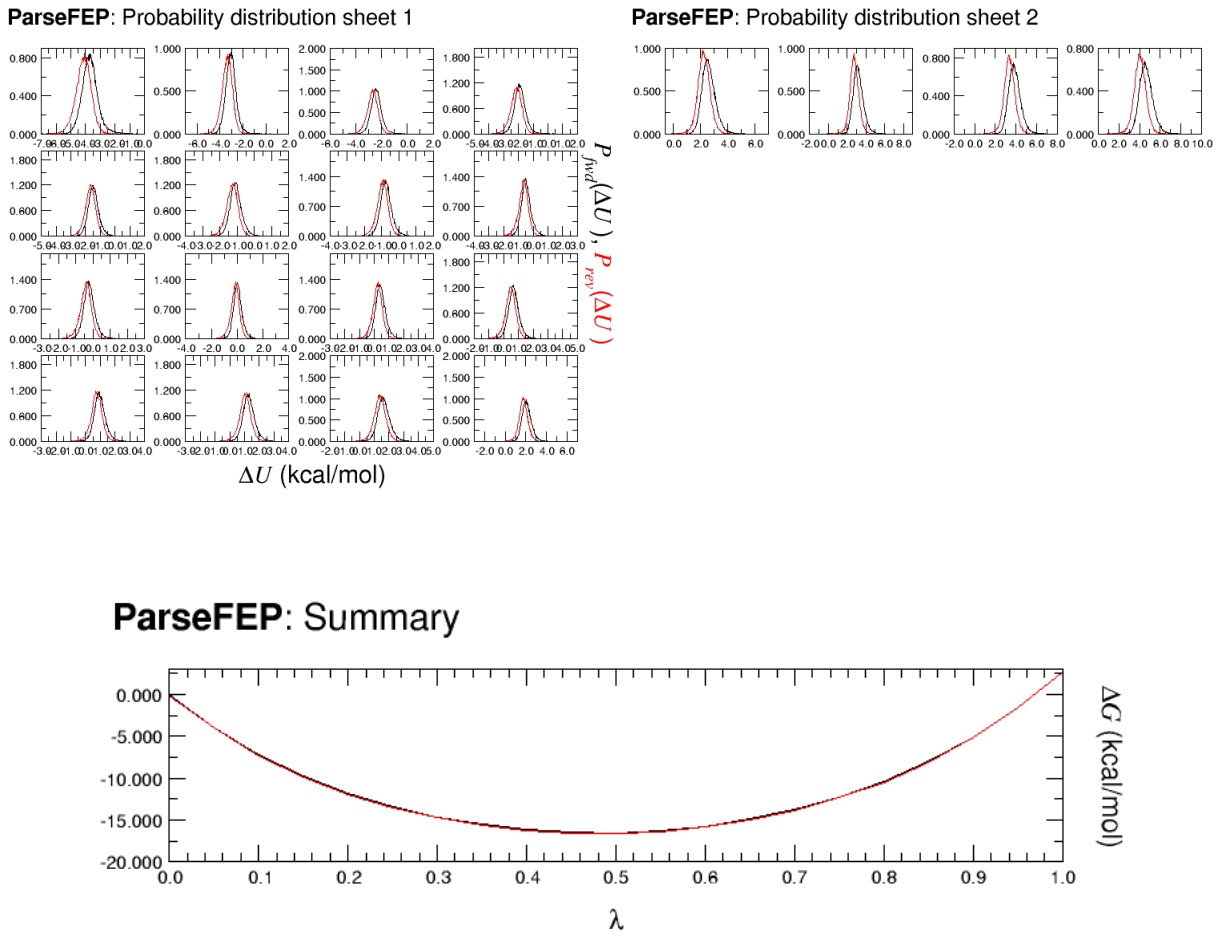
Transformation	ΔG_{AQ}	Error(ΔG_{AQ})	ΔG_{DNA}	Error(ΔG_{DNA})	$\Delta G_{Complex}$	Error($\Delta G_{Complex}$)	$\Delta\Delta G_{Binding}$	$\Delta\Delta G_{Repair}$
B[c]P ↔ B[g]C	15.00	0.02	10.56	0.07	12.08	0.04	-4.44	1.53
B[g]C ↔ DB[a,l]P	0.30	0.01	-0.22	0.04	0.26	0.03	-0.52	0.47
B[a]P ↔ DB[a,l]P	22.69	0.03	8.35	0.05	11.81	0.07	-14.34	3.46
B[a]P ↔ DB[a,e]P	13.74	0.02	11.60	0.08	13.28	0.03	-2.14	1.68
B[a]P ↔ DB[a,i]P	4.42	0.02	2.41	0.04	2.16	0.03	-2.00	-0.26
B[a]P ↔ DB[a,h]P	2.48	0.02	1.95	0.04	3.86	0.07	-0.53	1.91
B[a]P ↔ B[e]P	4.93	0.02	9.04	0.06	5.55	0.02	4.11	-3.49
B[e]P ↔ DB[e,l]P	14.10	0.02	11.86	0.06	13.81	0.05	-2.23	-0.86
B[a]P ↔ CHR	-0.44	0.01	0.25	0.02	-0.83	0.01	0.68	-1.08
CHR ↔ B[g]C	24.83	0.02	7.99	0.05	11.58	0.05	-16.84	3.59
CHR ↔ B[b]C	-3.13	0.02	-5.46	0.04	-3.06	0.02	-2.33	-3.06
CHR ↔ PHE	-6.15	0.02	-7.34	0.08	-5.38	0.03	-1.19	1.96
PHE ↔ B[c]P	13.21	0.02	1.41	0.09	1.78	0.04	-11.81	0.38
PHE ↔ B[a]A	-3.20	0.02	-4.93	0.06	-5.22	0.05	-1.72	-0.29
B[a]A ↔ DB[a,j]A	6.34	0.02	1.60	0.04	5.32	0.02	-4.73	3.72
B[a]A ↔ DB[a,c]A	8.65	0.02	4.71	0.04	8.87	0.04	-3.94	4.16
B[a]A ↔ DB[a,h]A	6.08	0.02	2.63	0.04	5.84	0.03	-3.45	3.21

Table (6.1) Free energy differences (kcal/mol)

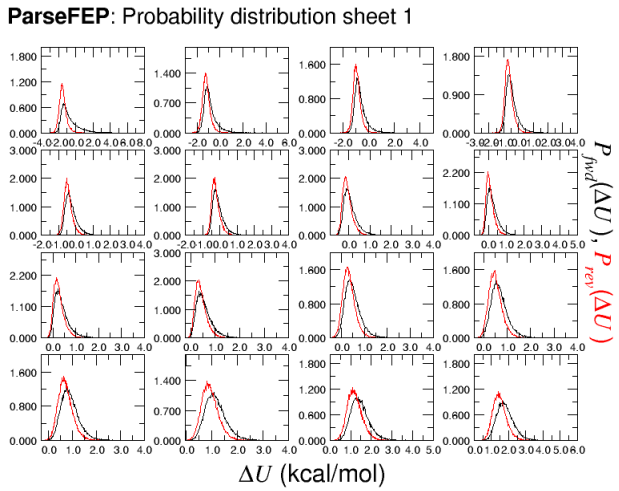
Transformation	ΔU_{AQ}	$T\Delta S_{AQ}$	ΔU_{DNA}	$T\Delta S_{DNA}$	$\Delta U_{Complex}$	$T\Delta S_{Complex}$
B[c]P ↔ B[g]C	5.01	-11.47	-17.71	-29.31	-3.60	9.68
B[g]C ↔ DB[a,l]P	0.16	-1.54	2.63	2.06	-3.36	-5.06
B[a]P ↔ DB[a,l]P	5.58	-15.55	5.29	-3.63	38.82	27.34
B[a]P ↔ DB[a,e]P	14.36	1.93	3.50	-10.00	-3.44	-15.92
B[a]P ↔ DB[a,i]P	-0.60	-3.99	2.56	-0.52	-29.34	-32.33
B[a]P ↔ DB[a,h]P	-1.64	-5.04	8.23	5.35	19.43	15.85
B[a]P ↔ B[e]P	2.49	-0.24	0.03	-7.00	-15.94	-19.16
B[e]P ↔ DB[e,l]P	7.01	-5.75	-6.44	-19.66	9.61	-3.52
B[a]P ↔ CHR	0.16	1.91	-1.96	-0.86	-0.63	0.94
CHR ↔ B[g]C	6.89	-16.40	1.87	-6.10	-27.18	-39.13
CHR ↔ B[b]C	-5.84	-1.96	5.57	10.24	-7.76	-3.90
CHR ↔ PHE	-1.64	3.46	5.53	12.07	1.84	6.60
PHE ↔ B[c]P	1.10	-10.88	6.05	4.78	1.34	-0.96
PHE ↔ B[a]A	-7.96	-3.98	-0.61	-4.52	10.93	15.53
B[a]A ↔ DB[a,j]A	-0.33	-5.63	-2.23	-4.31	23.86	18.62
B[a]A ↔ DB[a,c]A	2.75	-4.66	-17.43	-23.37	13.59	5.31
B[a]A ↔ DB[a,h]A	0.78	-4.25	0.31	-3.06	-6.89	-11.85

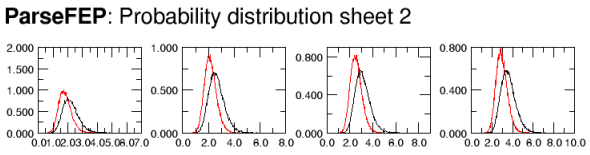
Table (6.2) Enthalpy and entropy estimates (kcal/mol)

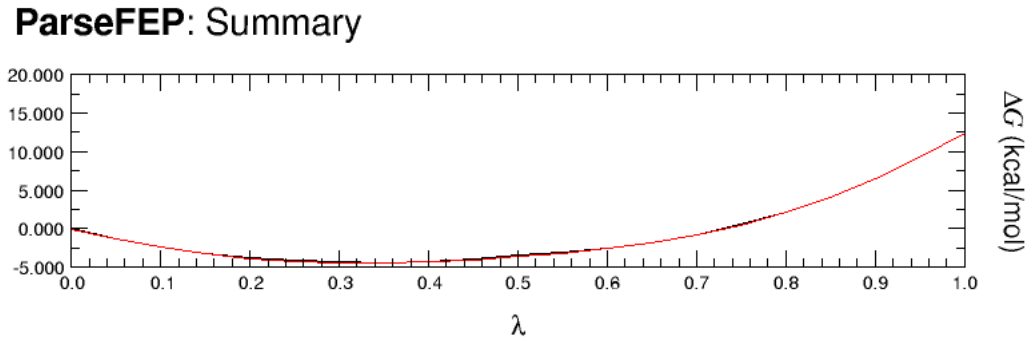
6.1.1 FEP Plots: PAH-DEs in Solution



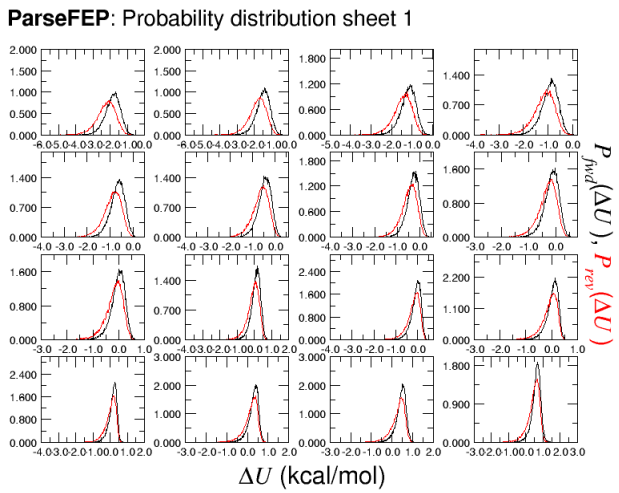
(6.1) B[a]P-DE ↔ B[e]P-DE in solution

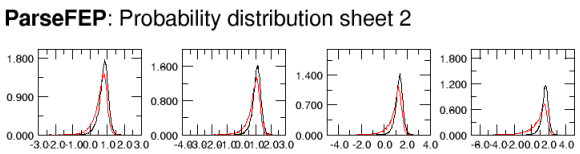


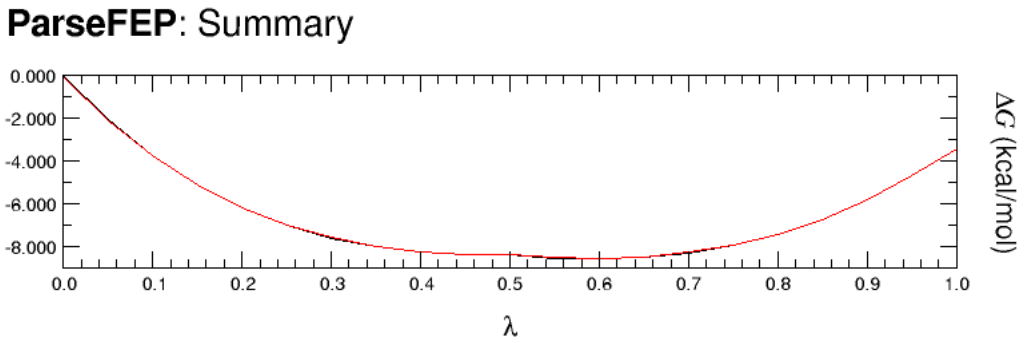




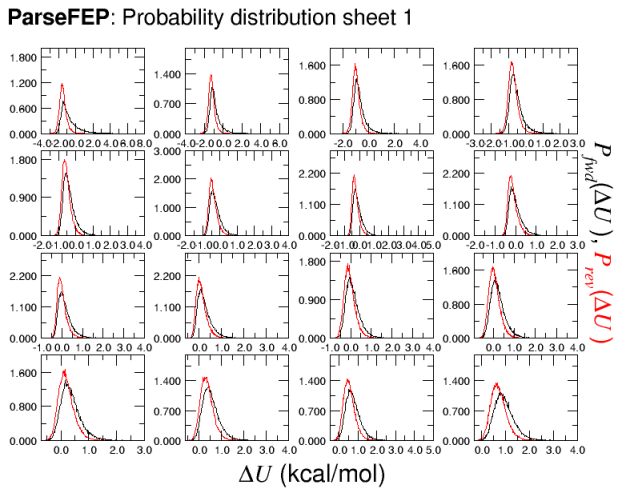
(6.2) B[a]P-DE ↔ DB[a,e]P-DE in solution

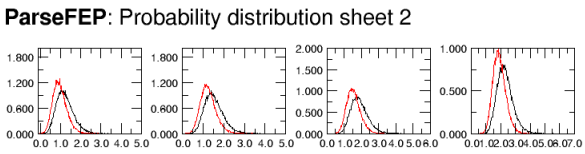


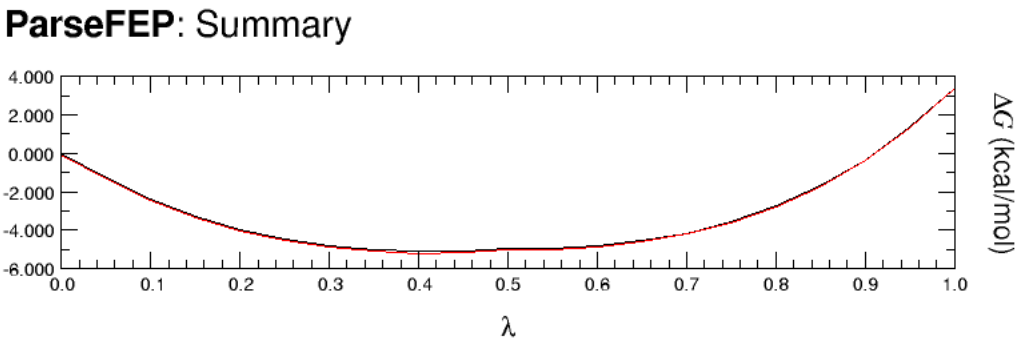




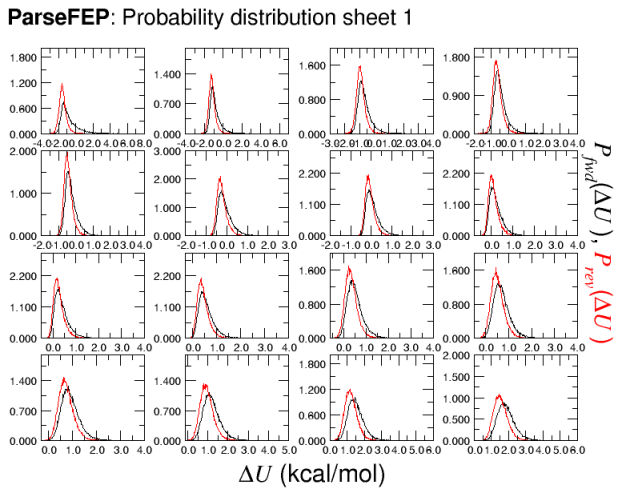
(6.3) B[a]P-DE ↔ DB[a,h]P-DE in solution

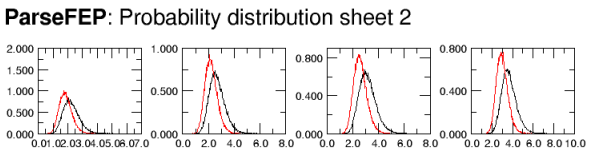


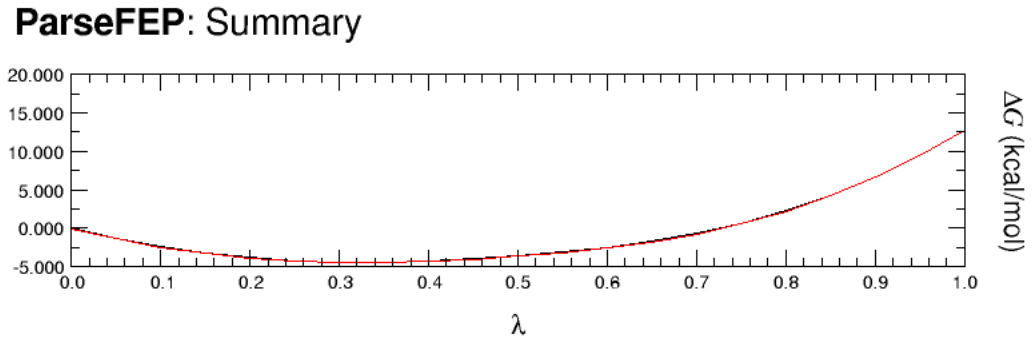




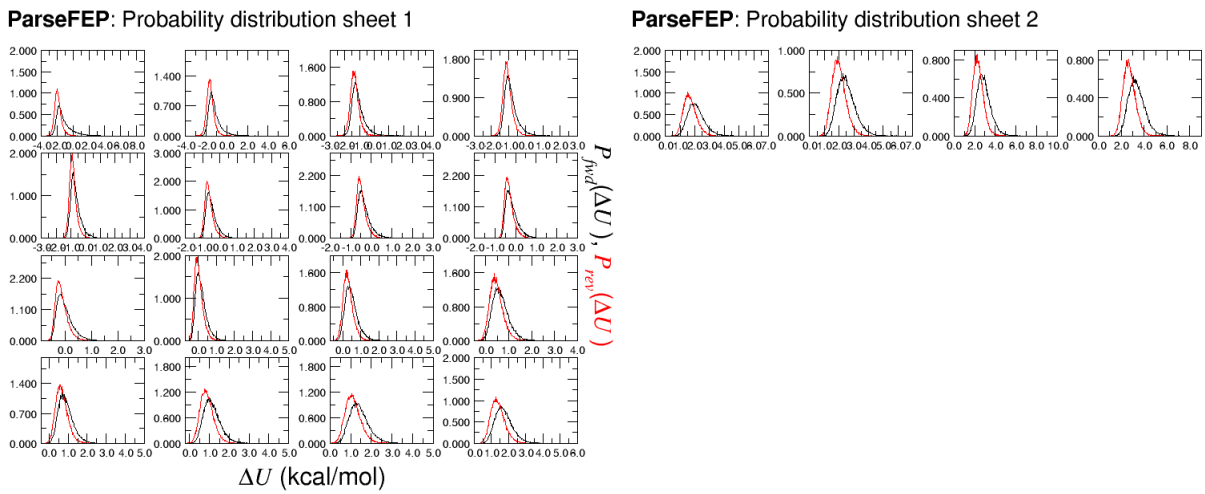
(6.4) B[a]P-DE ↔ DB[a,i]P-DE in solution

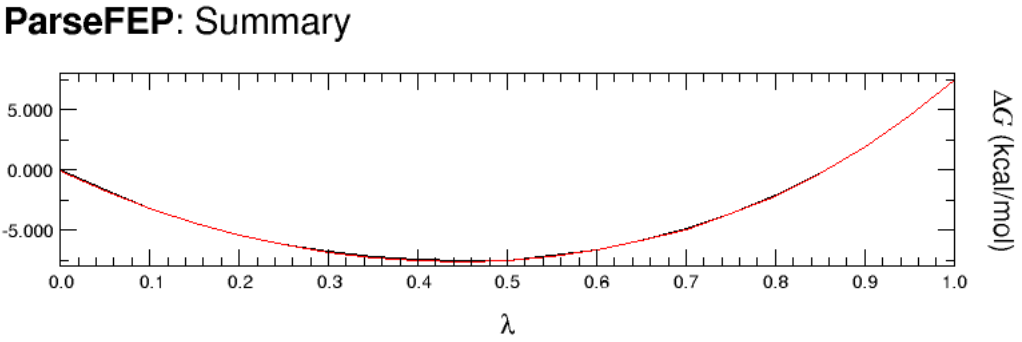




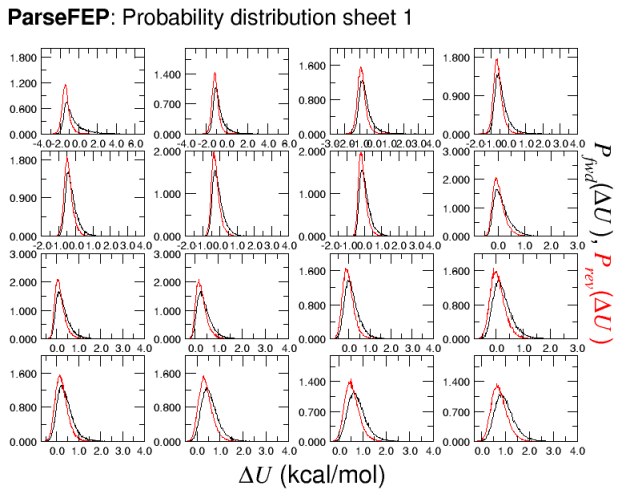


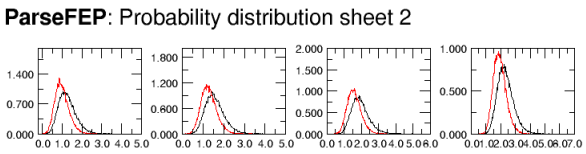
(6.5) B[e]P-DE ↔ DB[e,l]P-DE in solution

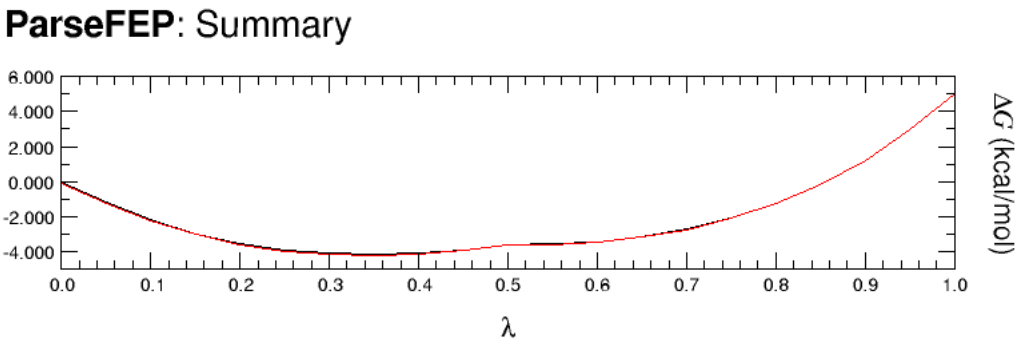




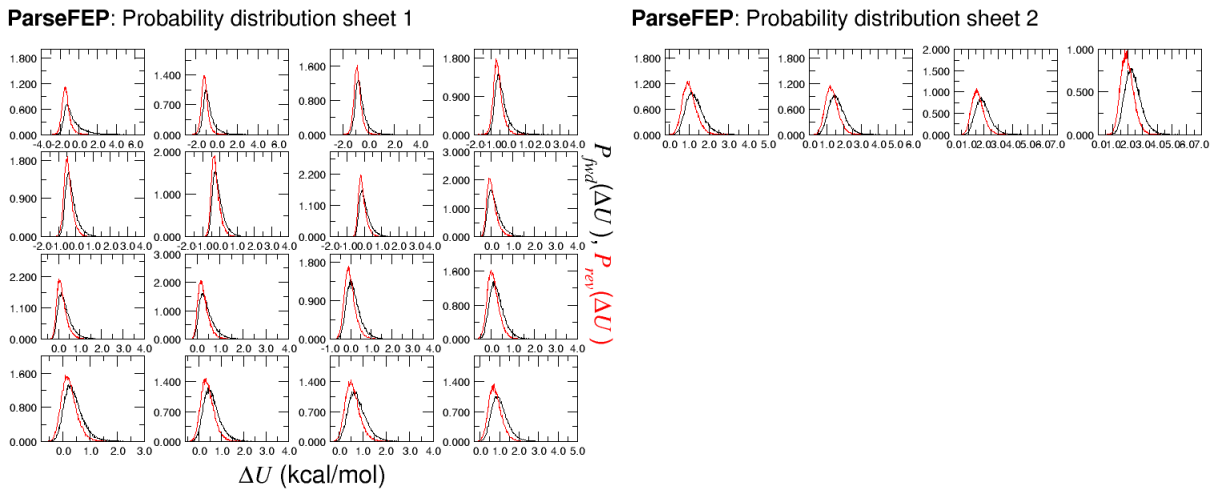
(6.6) B[a]A-DE ↔ DB[a,c]A-DE in solution

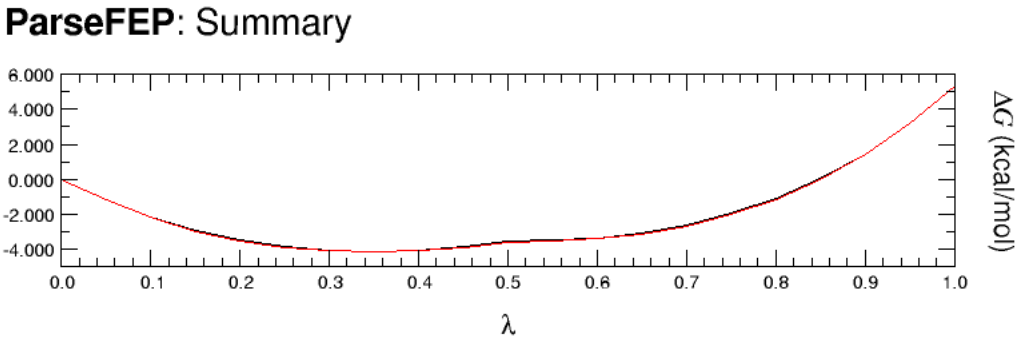




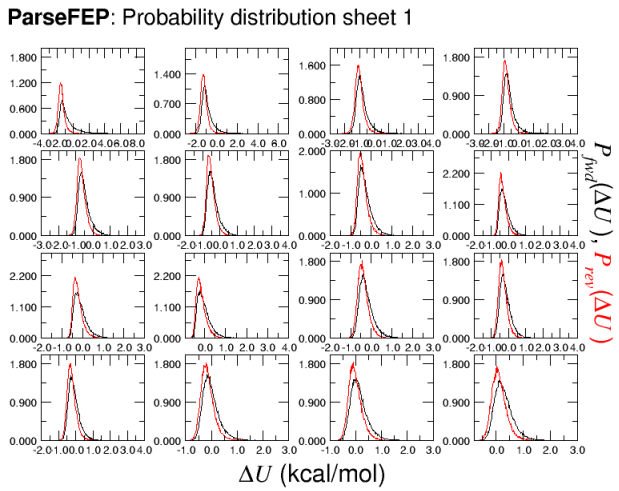


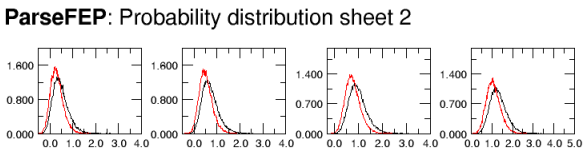
(6.7) B[a]A-DE ↔ DB[a,h]A-DE in solution

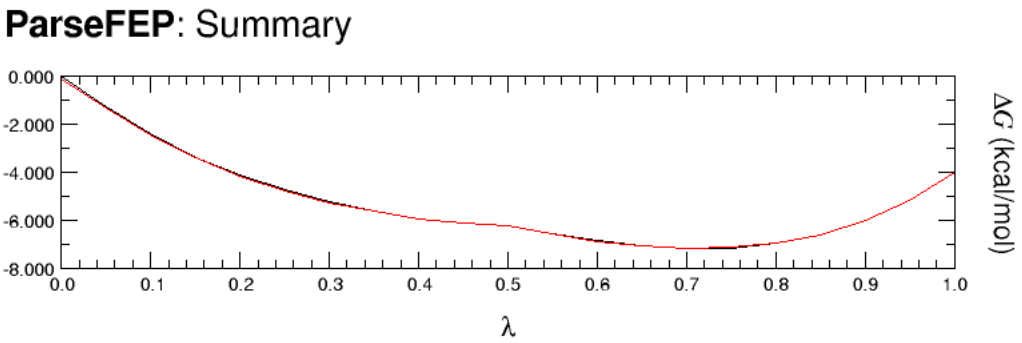




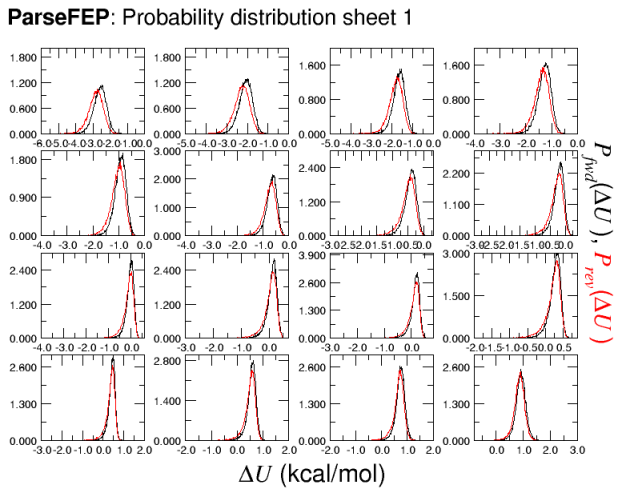
(6.8) B[a]A-DE ↔ DB[a,j]A-DE in solution

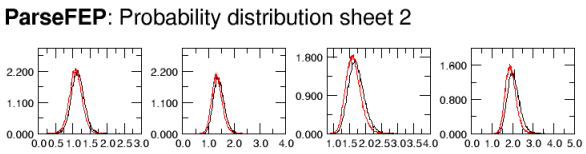


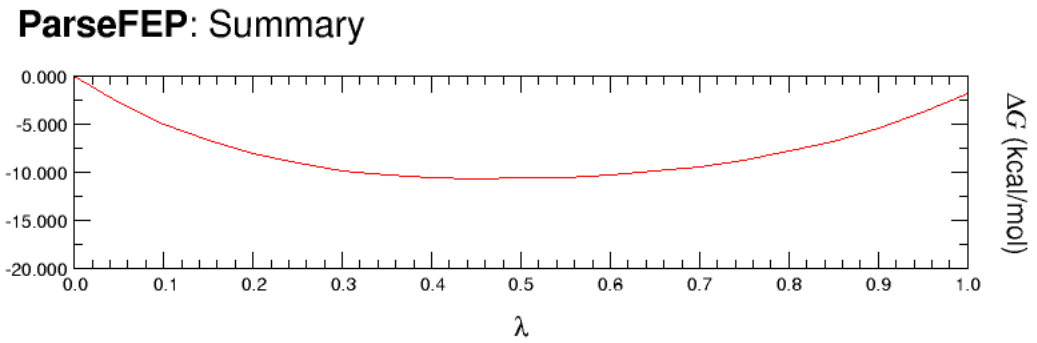




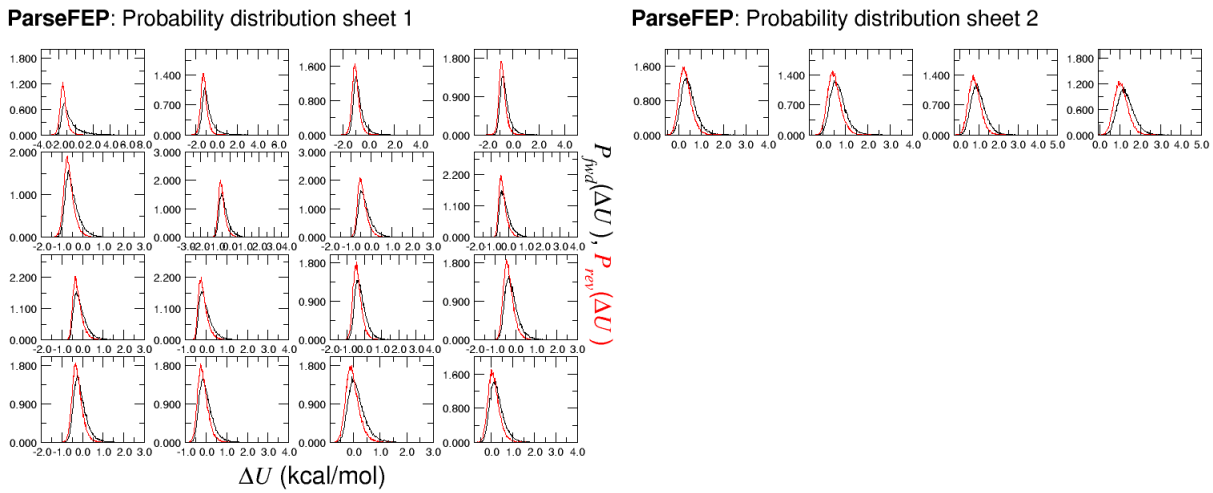
(6.9) PHR-DE ↔ B[a]A-DE in solution

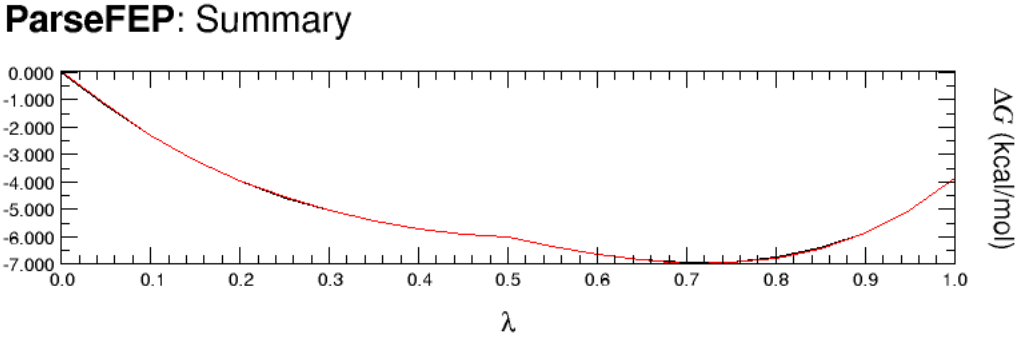




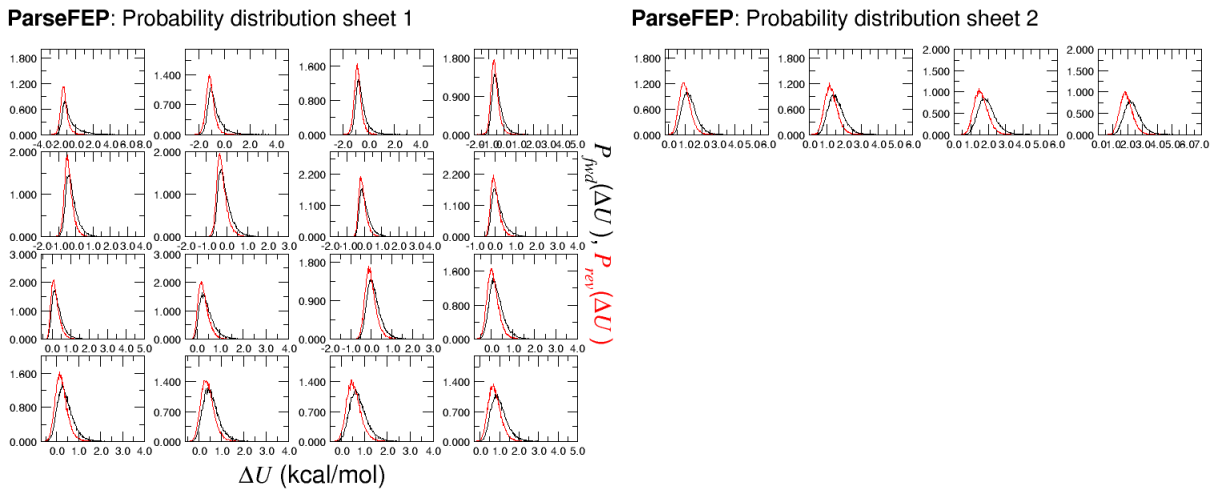


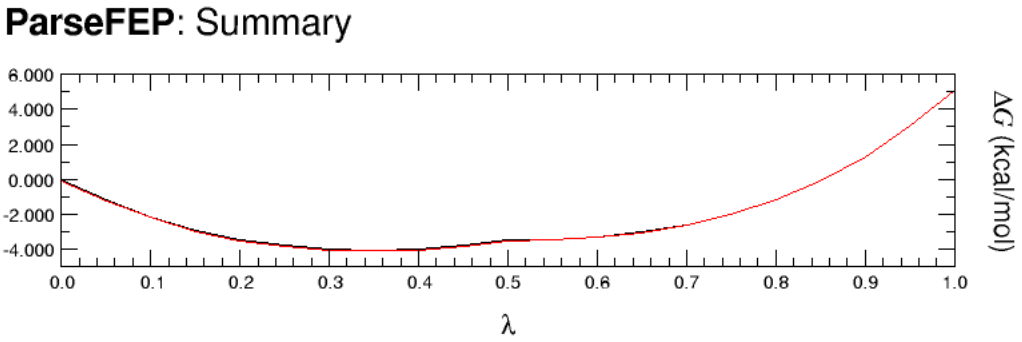
(6.10) B[a]P-DE ↔ CHR-DE in solution



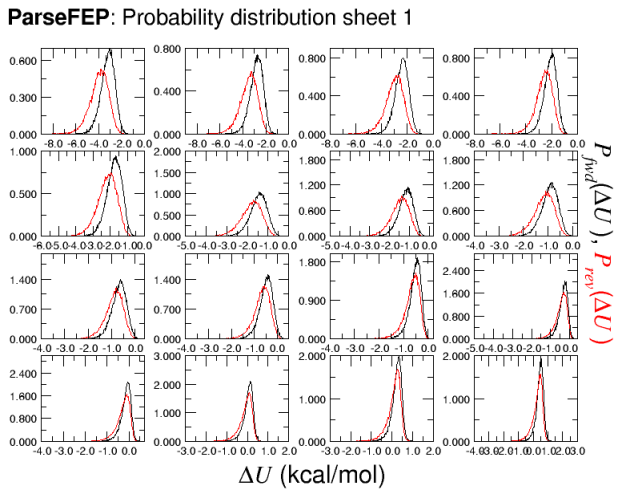


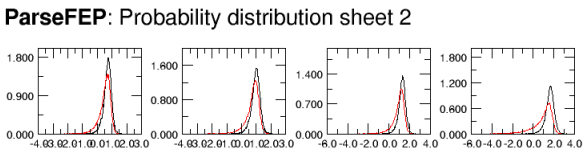
(6.11) CHR-DE ↔ B[b]C-DE in solution

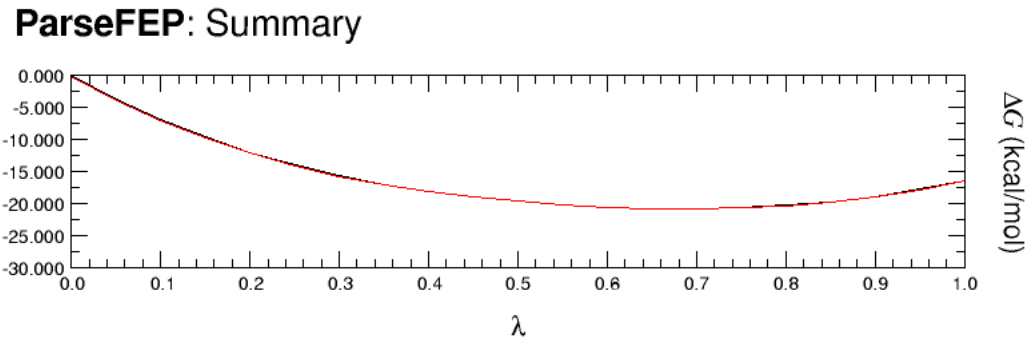




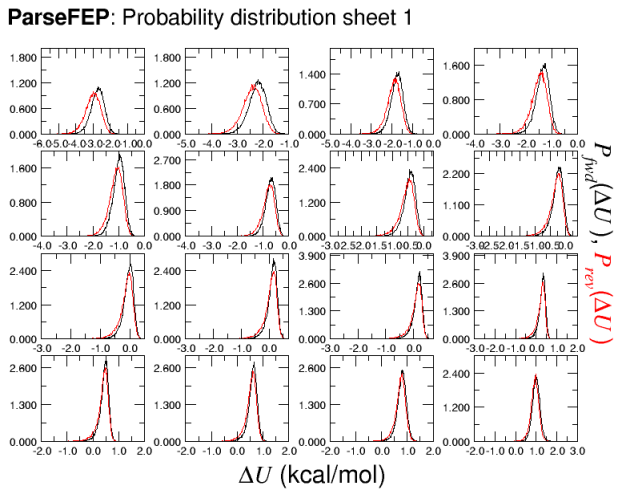
(6.12) PHR-DE ↔ CHR-DE in solution

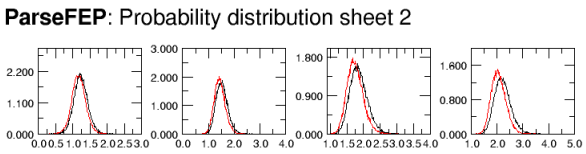


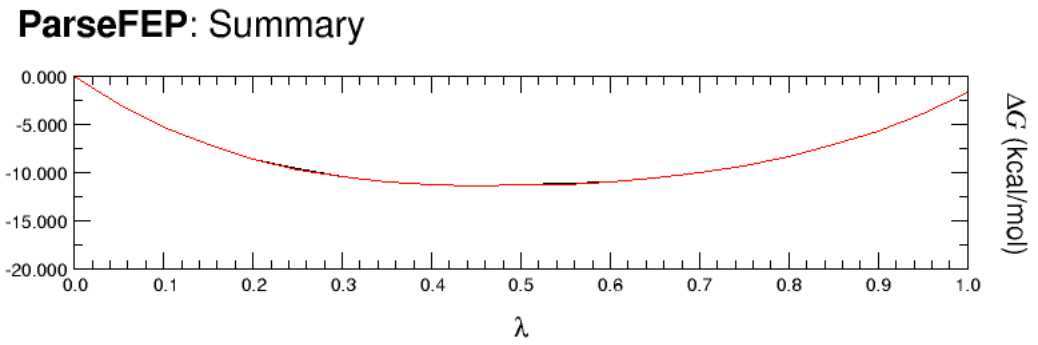




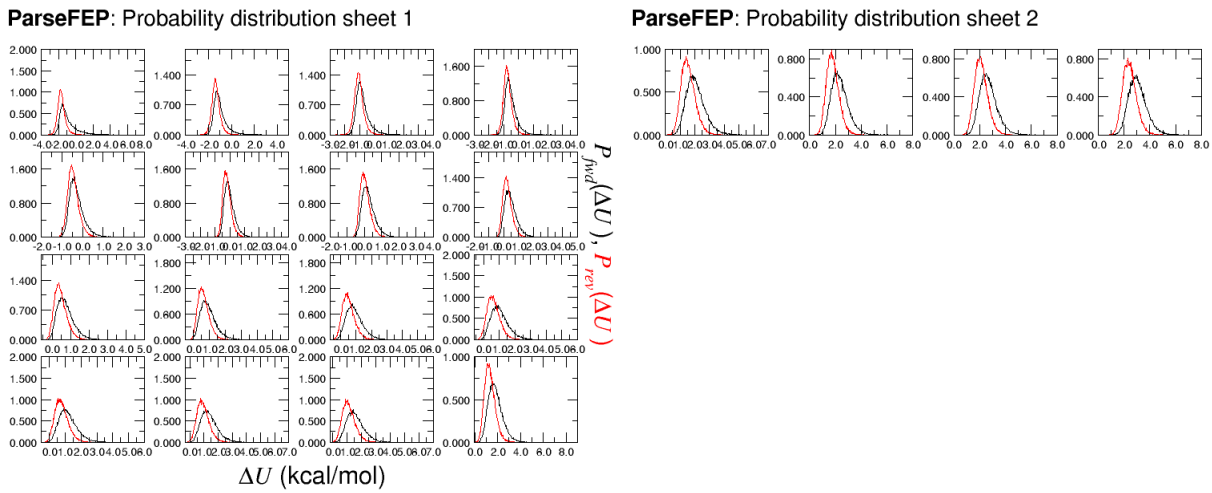
(6.13) B[g]C-DE ↔ B[c]P-DE in solution

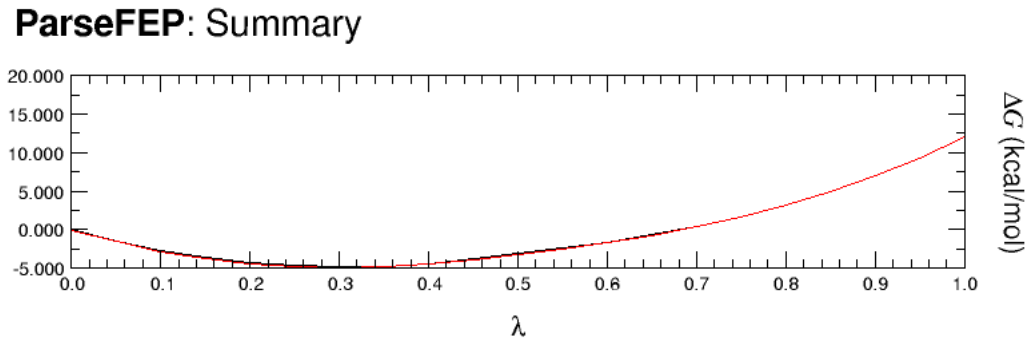




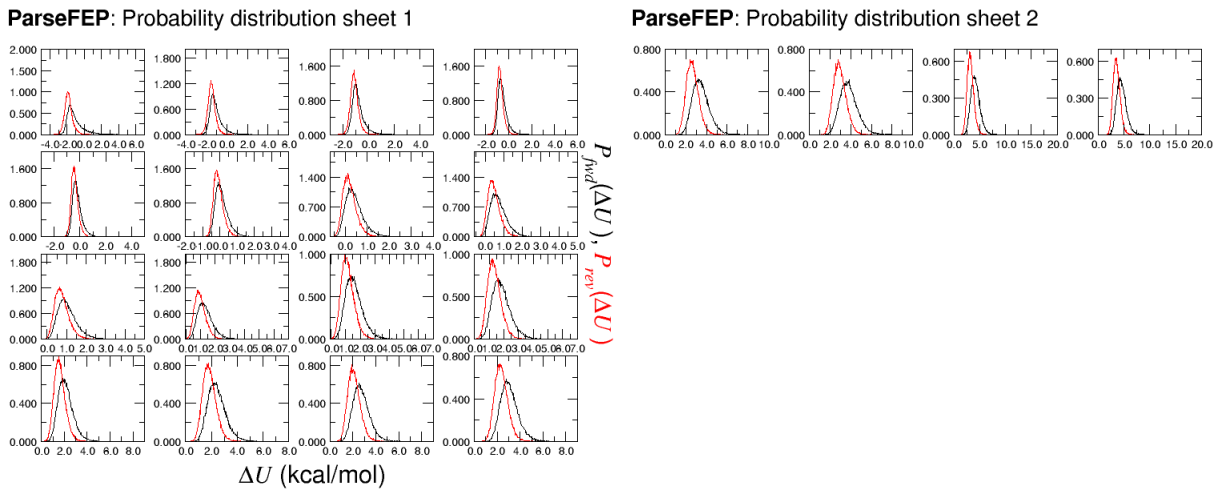


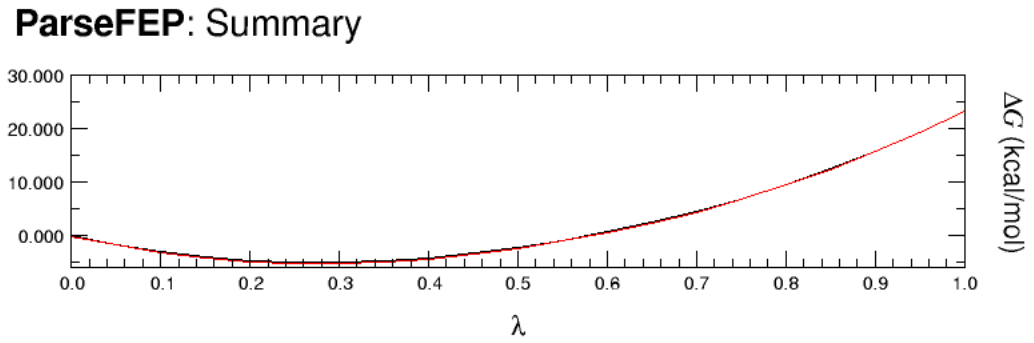
(6.14) DB[a,l]P-DE ↔ B[g]C-DE in solution



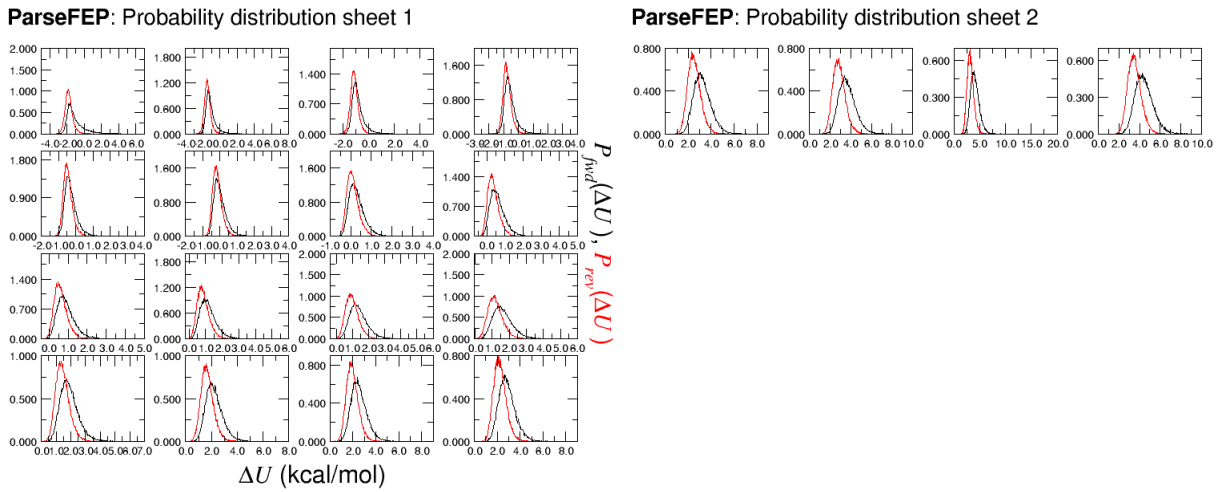


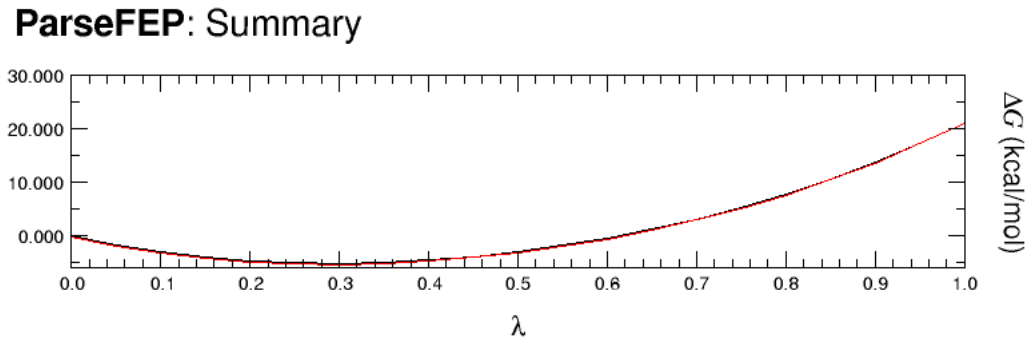
(6.15) B[c]P-DE ↔ PHR-DE in solution





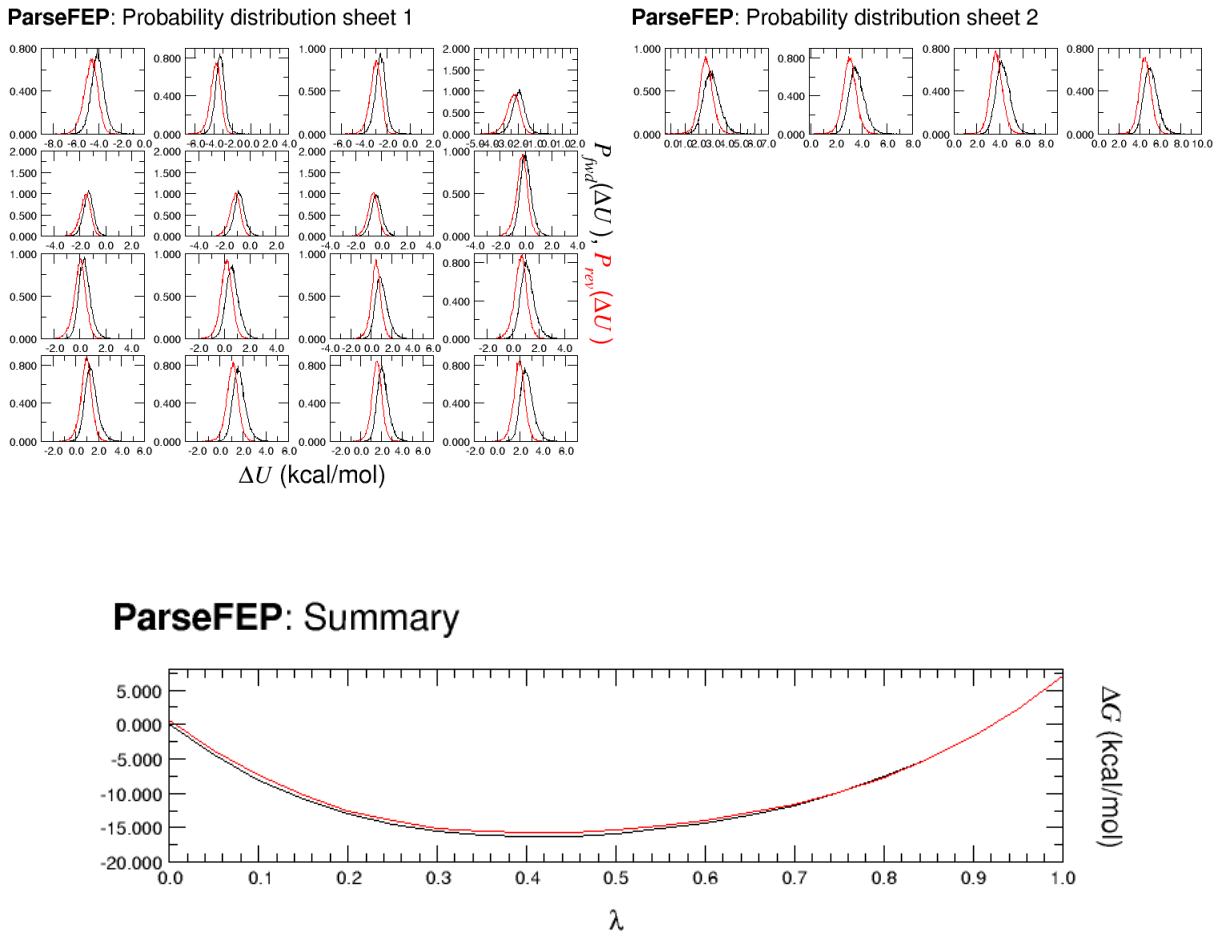
(6.16) B[g]C-DE ↔ CHR-DE in solution



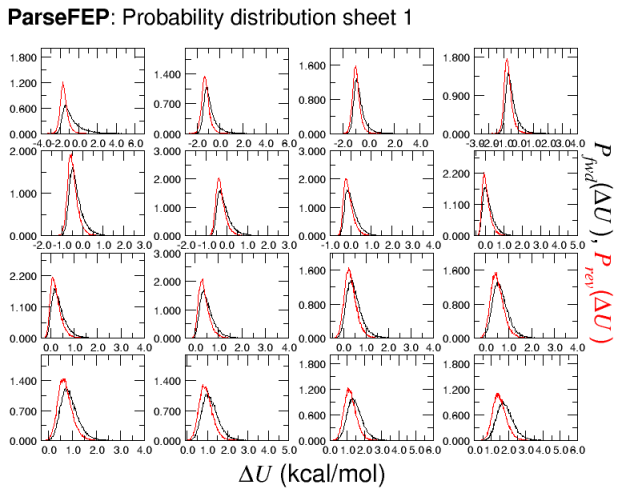


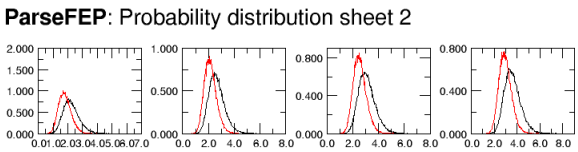
(6.17) DB[a,l]P-DE ↔ B[a]P-DE in solution

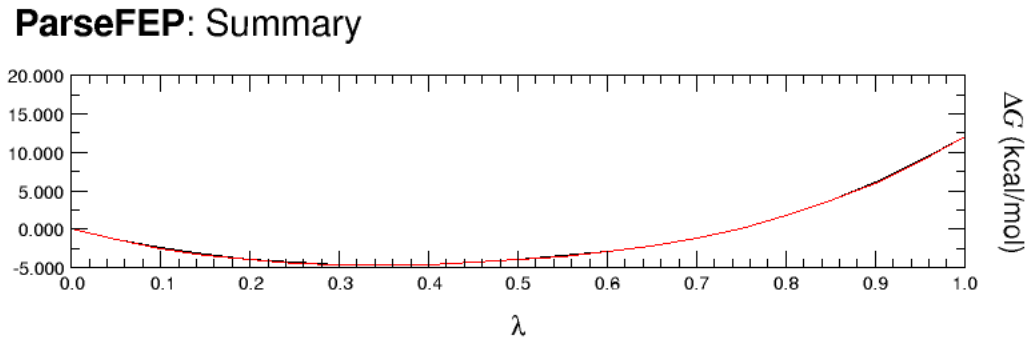
6.1.2 FEP Plots: PAH-DNA Adducts



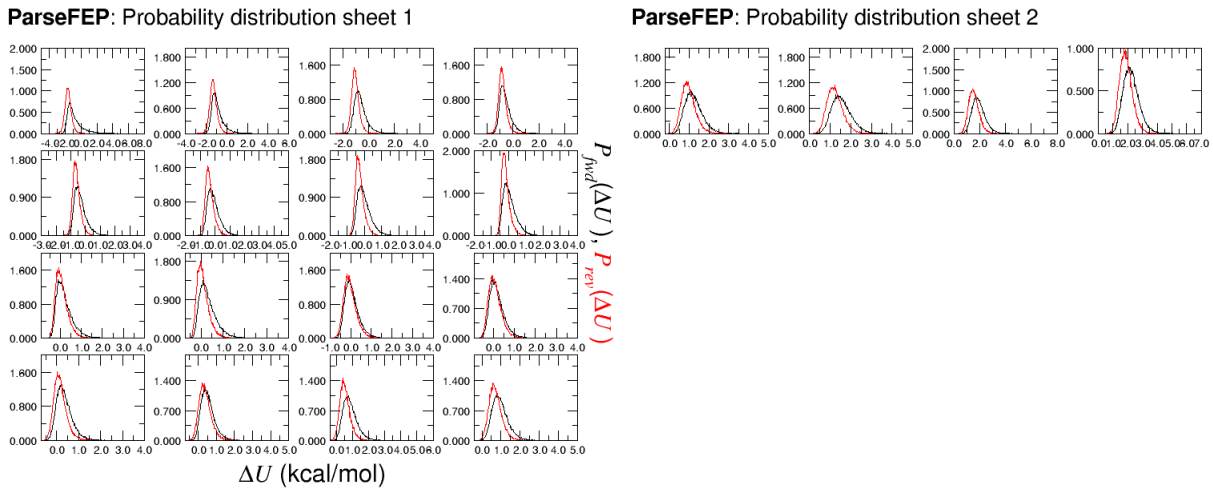
(6.18) B[a]P-DNA ↔ B[e]P-DNA adduct

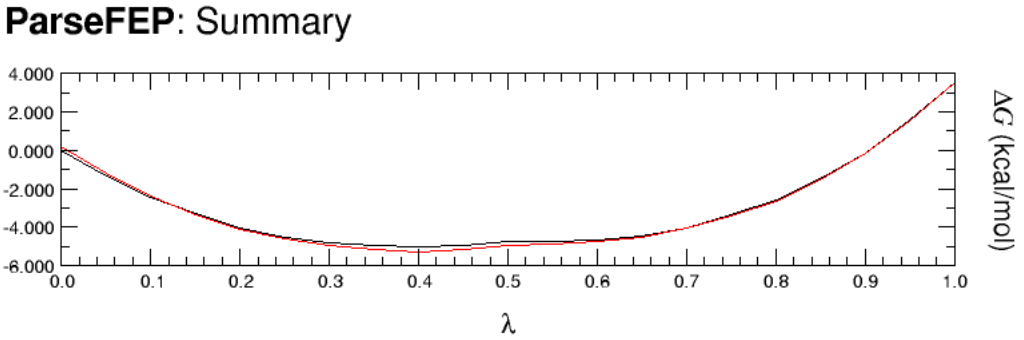




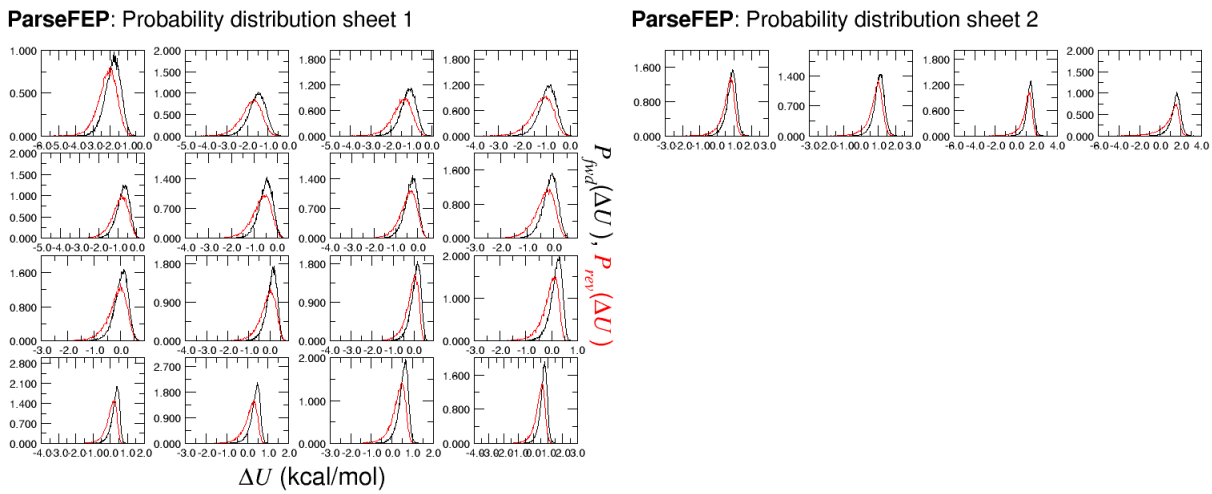


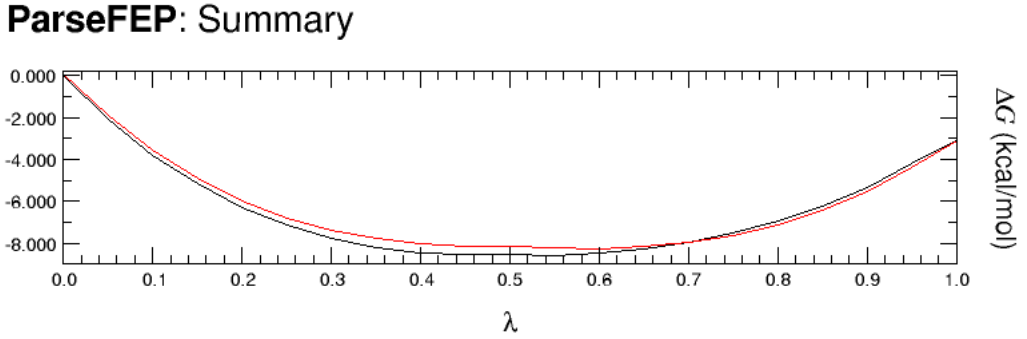
(6.19) B[a]P-DNA ↔ DB[a,e]P-DNA adduct



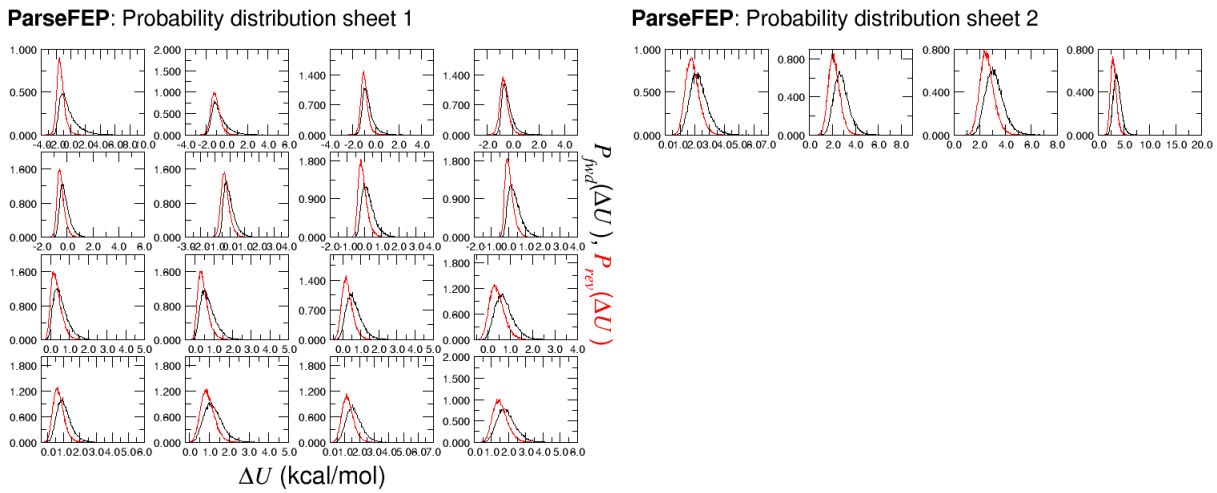


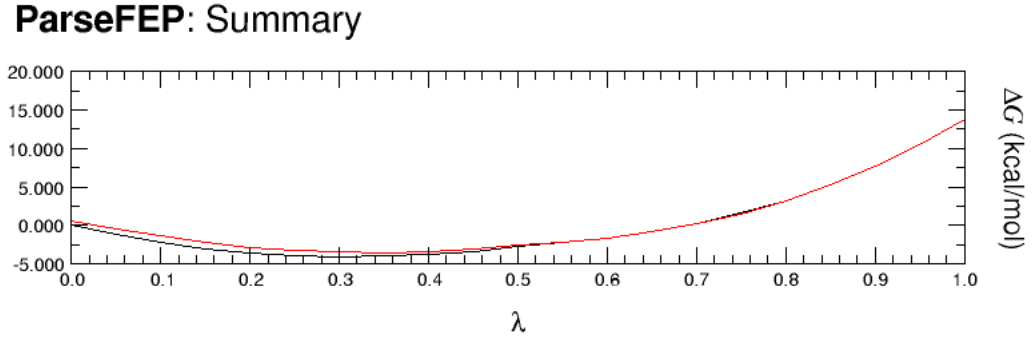
(6.20) B[a]P-DNA ↔ DB[a,h]P-DNA adduct



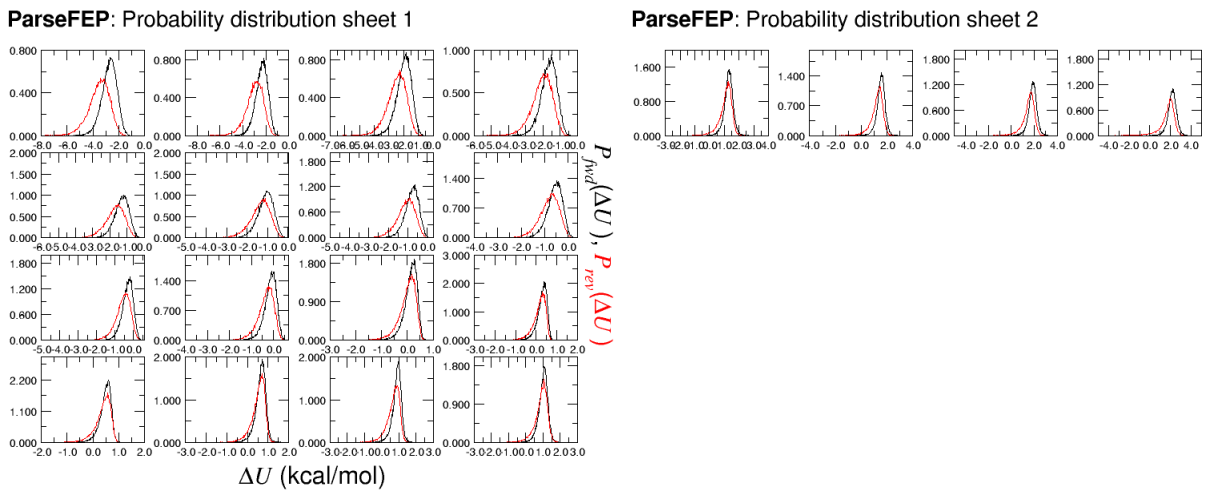


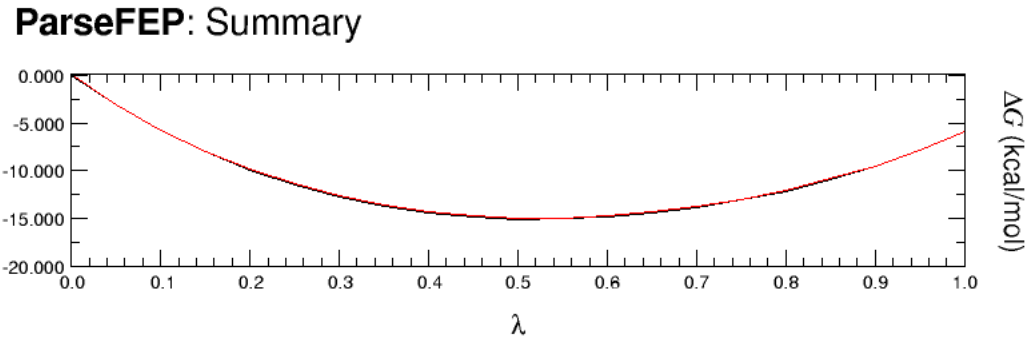
(6.21) DB[a,i]P-DNA ↔ B[a]P-DNA adduct



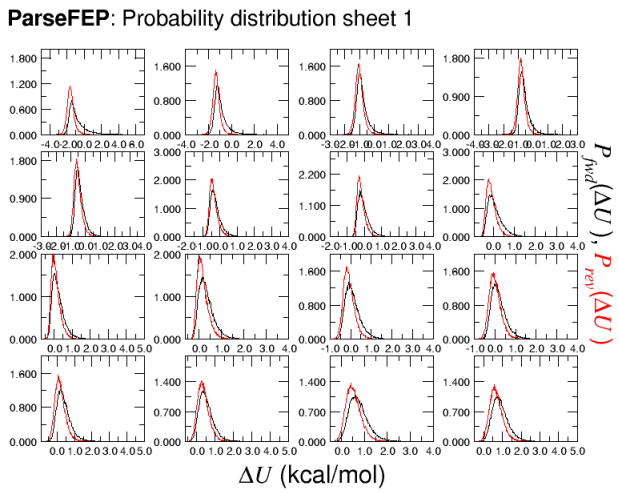


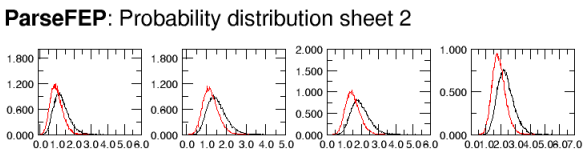
(6.22) B[e]P-DNA ↔ DB[e,l]P-DNA adduct

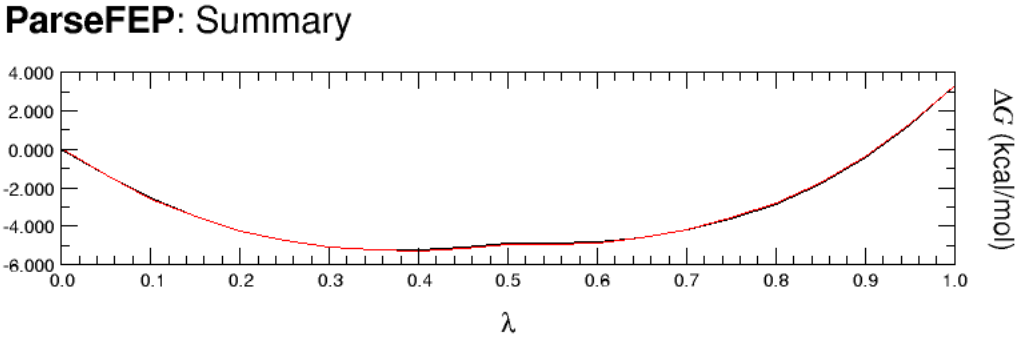




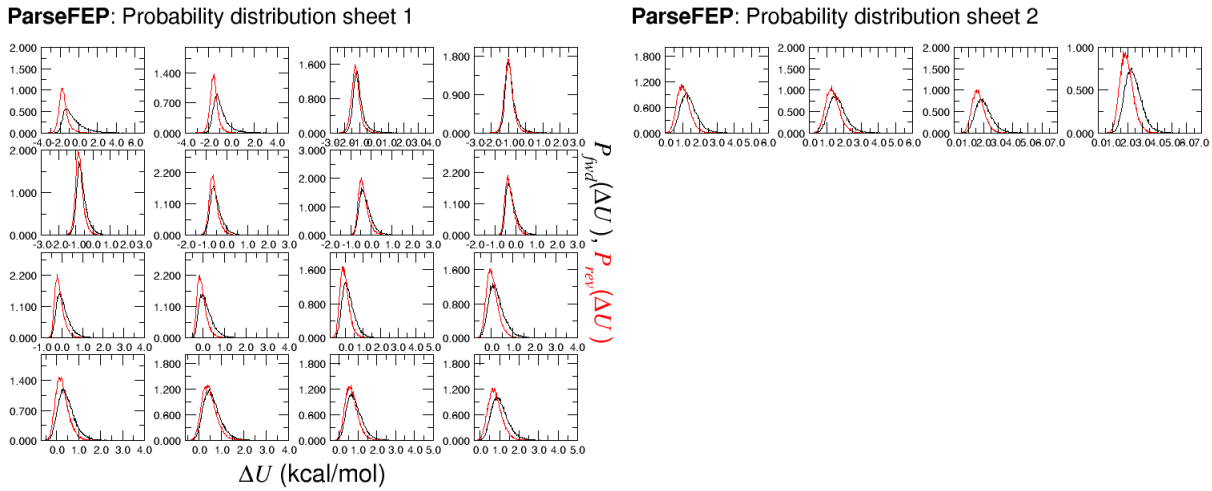
(6.23) DB[a,c]A-DNA ↔ B[a]A-DNA adduct

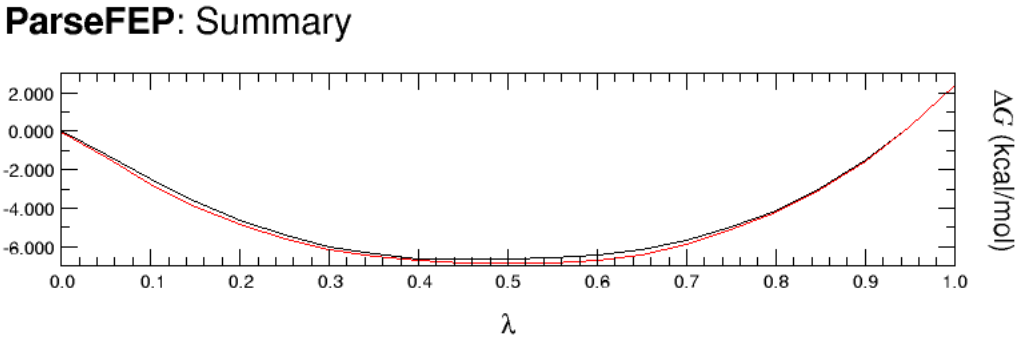




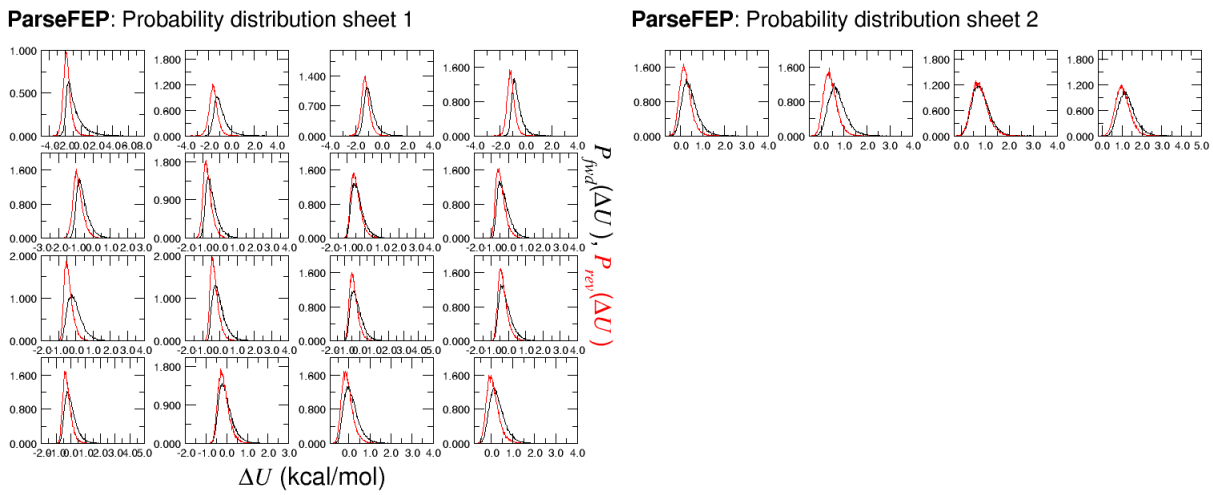


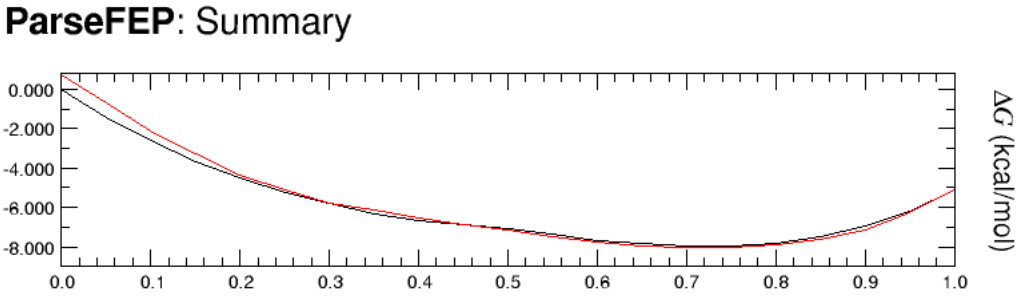
(6.24) B[a]A-DNA ↔ DB[a,h]A-DNA adduct





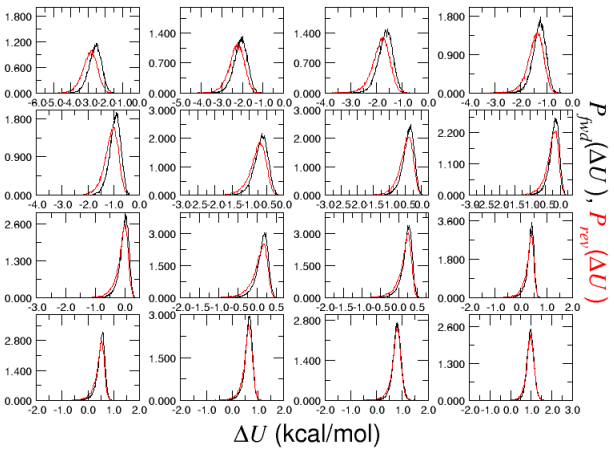
(6.25) B[a]A-DNA ↔ DB[a,j]A-DNA adduct



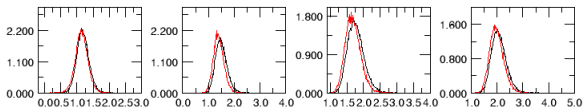


(6.26) B[a]A-DNA ↔ PHR-DNA adduct

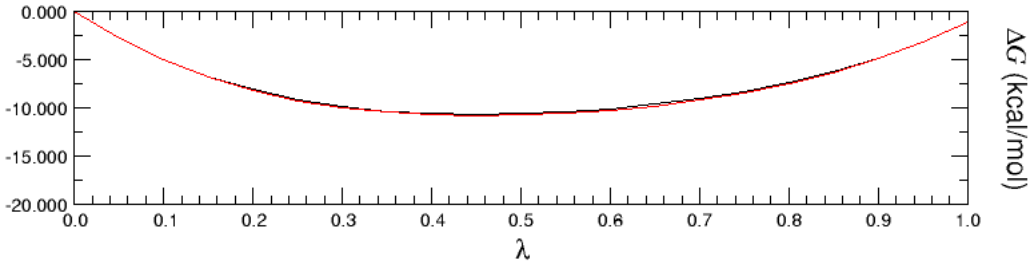
ParseFEP: Probability distribution sheet 1



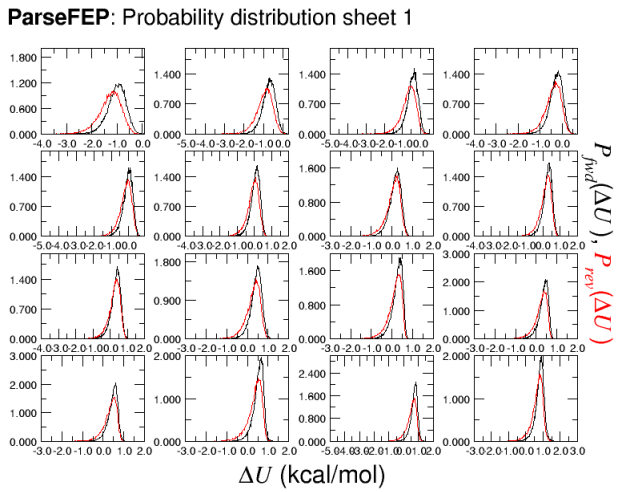
ParseFEP: Probability distribution sheet 2

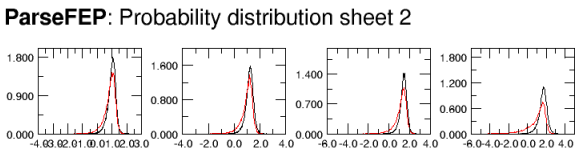


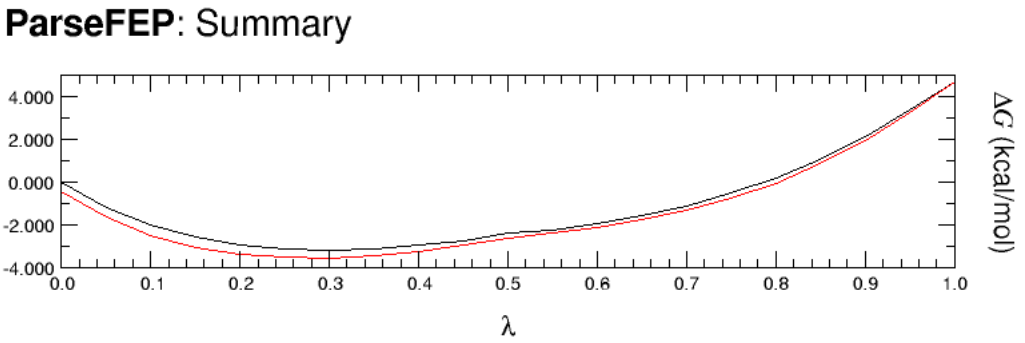
ParseFEP: Summary



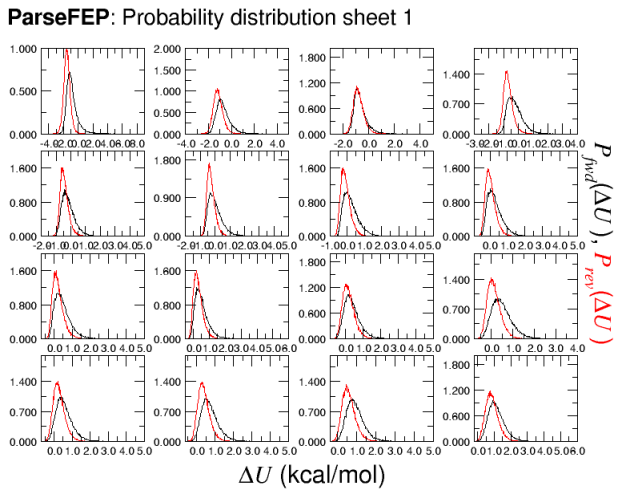
(6.27) CHR-DNA ↔ B[a]P-DNA adduct

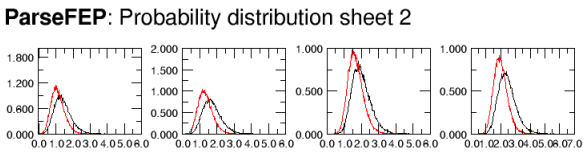


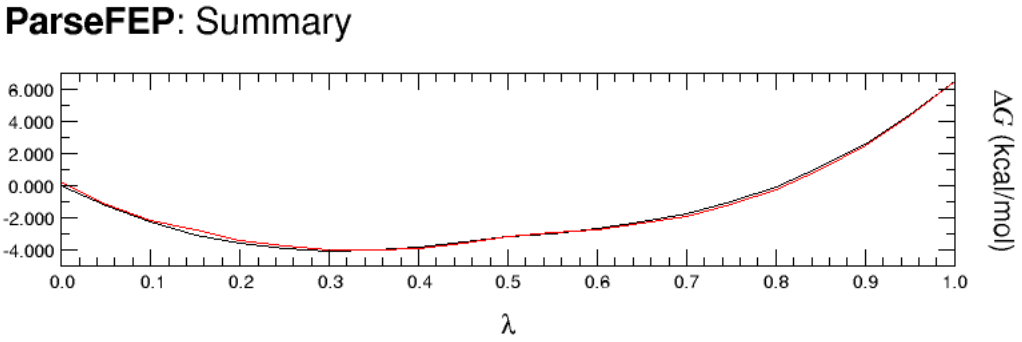




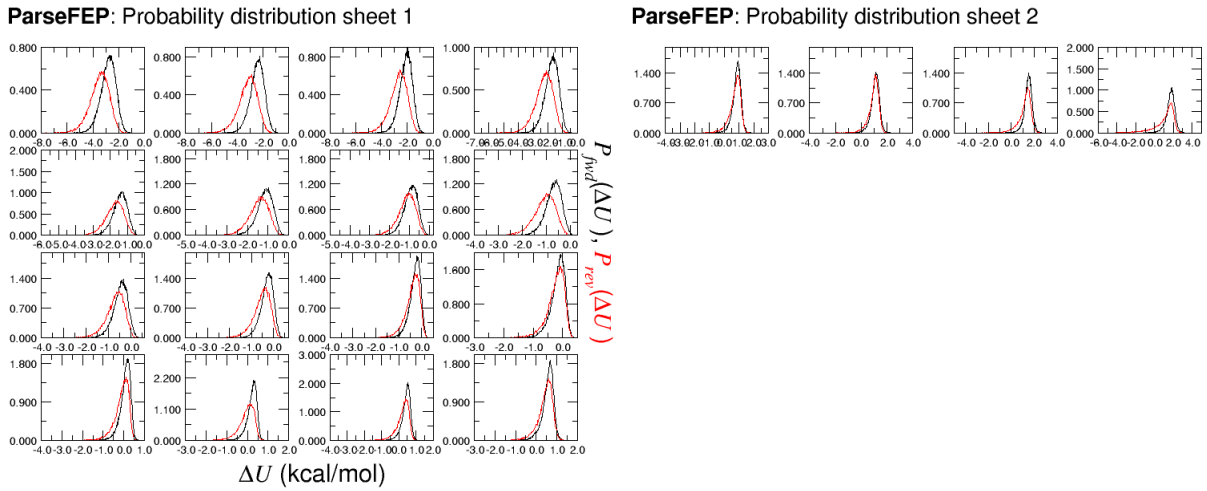
(6.28) CHR-DNA ↔ B[b]C-DNA adduct

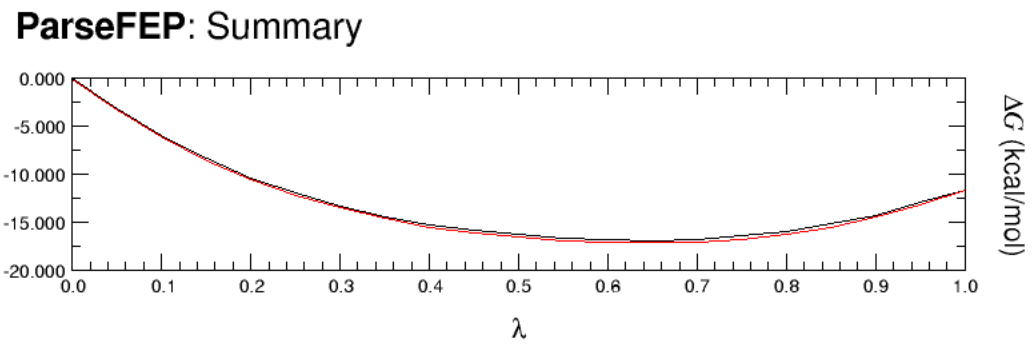




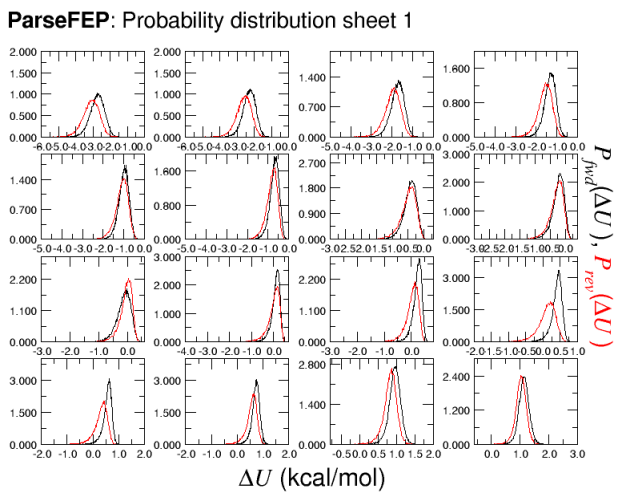


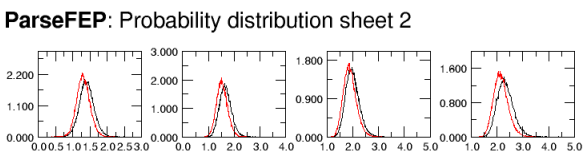
(6.29) PHR-DNA ↔ CHR-DNA adduct

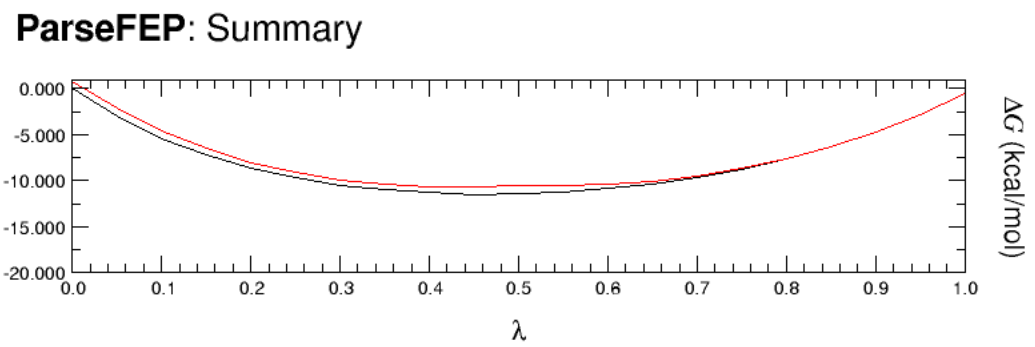




(6.30) B[g]C-DNA ↔ B[c]P-DNA adduct

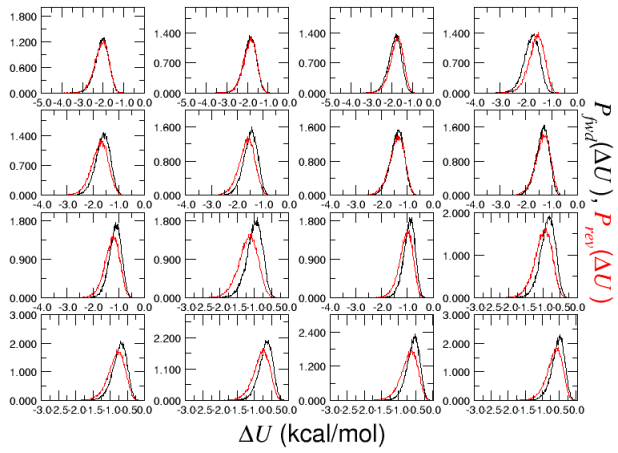




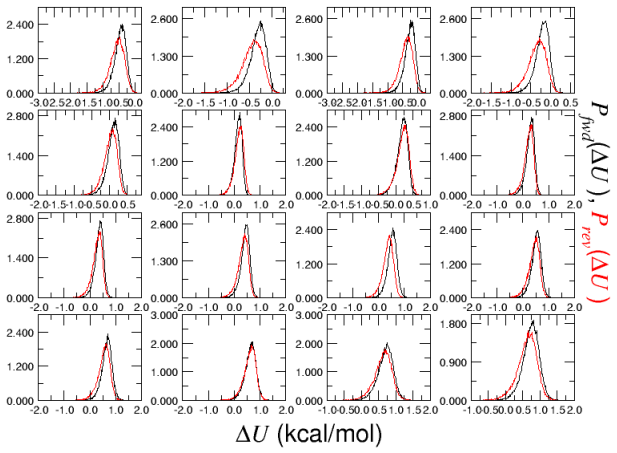


(6.31) DB[a,l]P-DNA ↔ B[g]C-DNA adduct

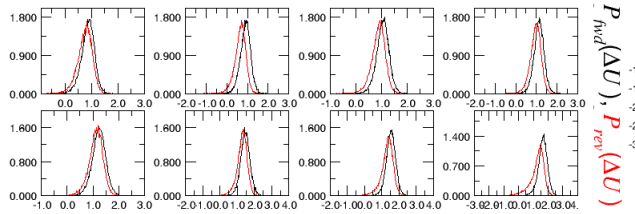
ParseFEP: Probability distribution sheet 1



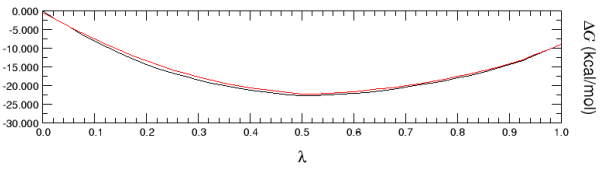
ParseFEP: Probability distribution sheet 2



ParseFEP: Probability distribution sheet 3

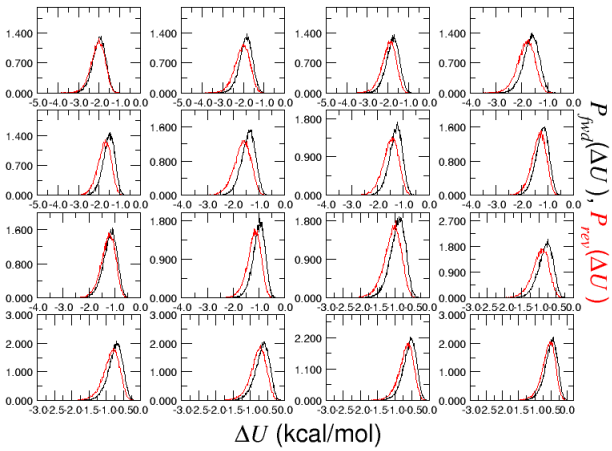


ParseFEP: Summary

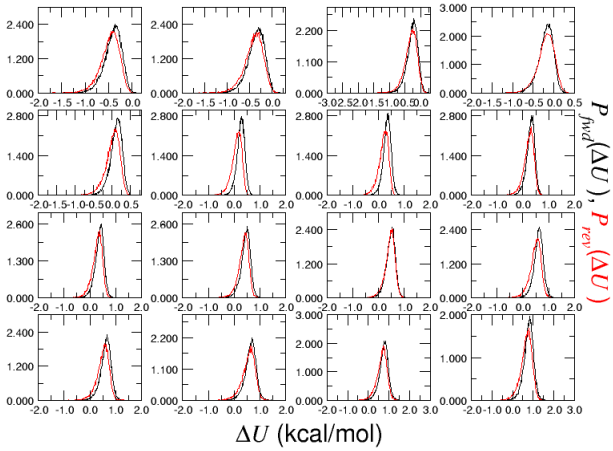


(6.32) DB[a,l]P-DNA ↔ B[a]P-DNA adduct

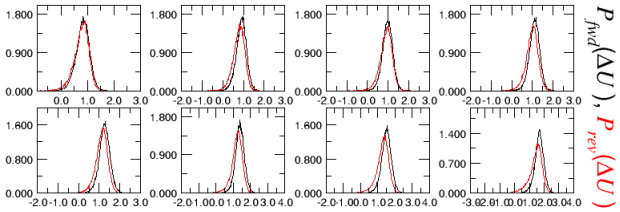
ParseFEP: Probability distribution sheet 1



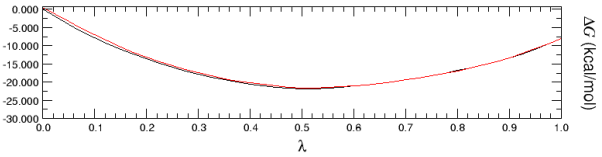
ParseFEP: Probability distribution sheet 2



ParseFEP: Probability distribution sheet 3

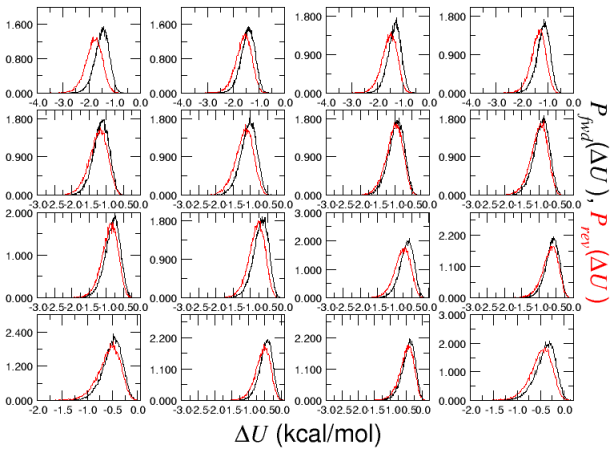


ParseFEP: Summary

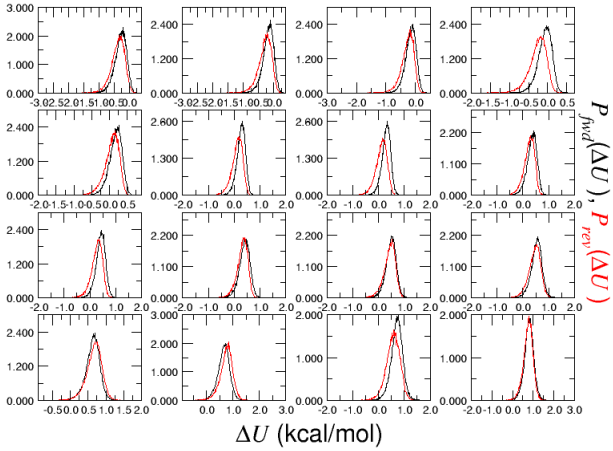


(6.33) B[g]C-DNA ↔ CHR-DNA adduct

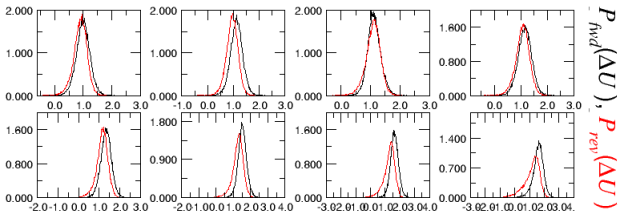
ParseFEP: Probability distribution sheet 1



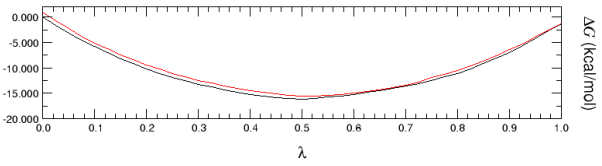
ParseFEP: Probability distribution sheet 2



ParseFEP: Probability distribution sheet 3

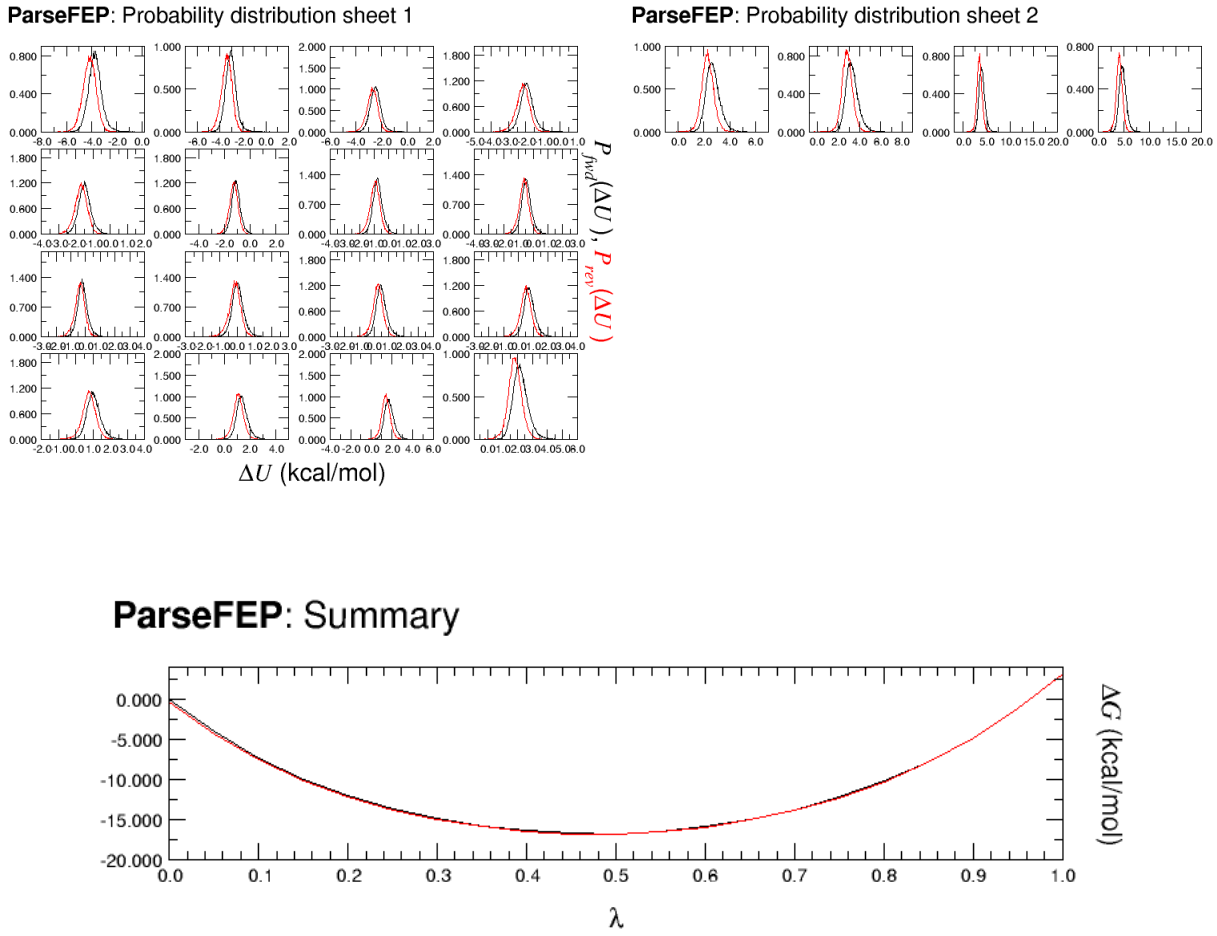


ParseFEP: Summary

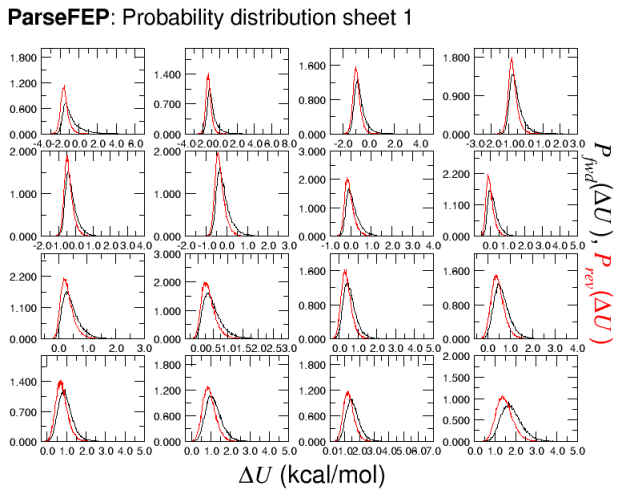


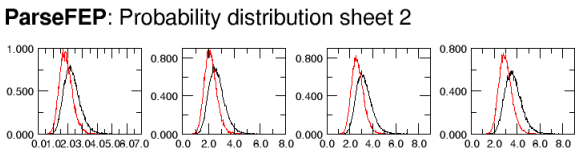
(6.34) B[c]P-DNA ↔ PHR-DNA adduct

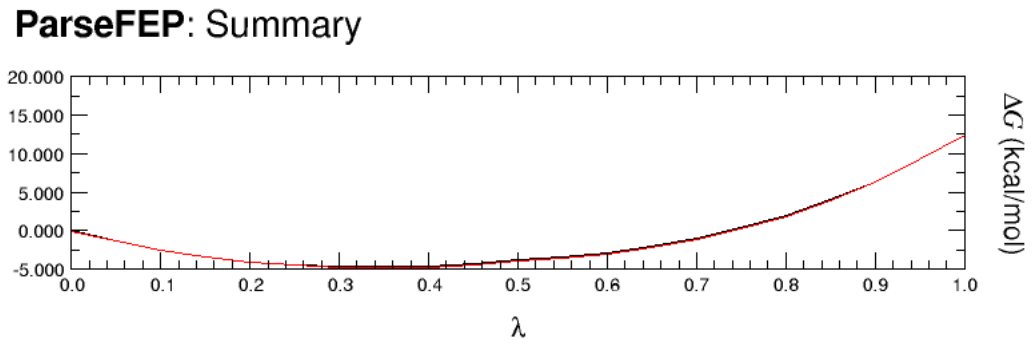
6.1.3 FEP Plots: PAH-DNA Adducts in the Productive Complex



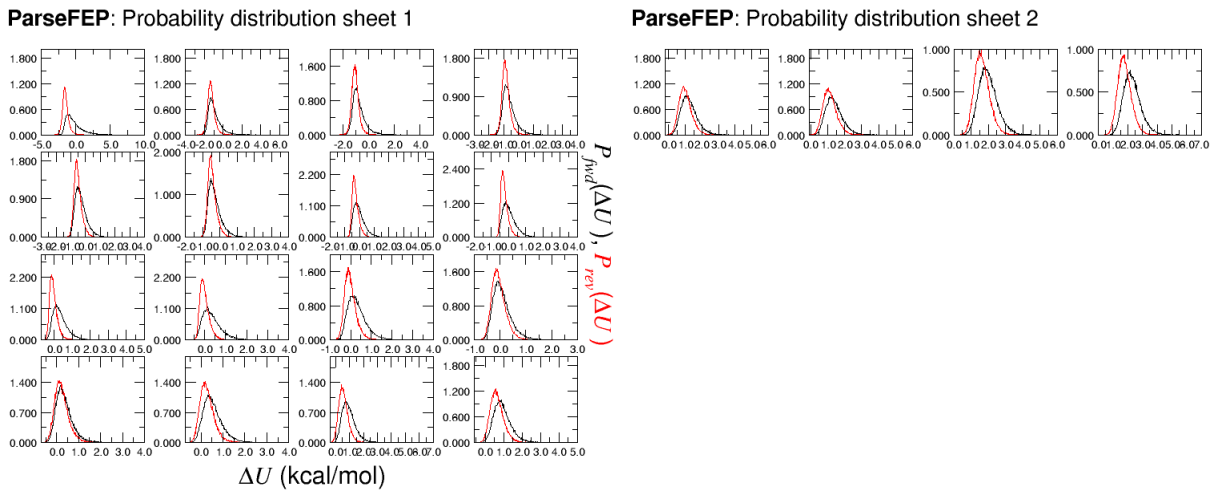
(6.35) B[a]P-DNA ↔ B[e]P-DNA adduct in productive complex

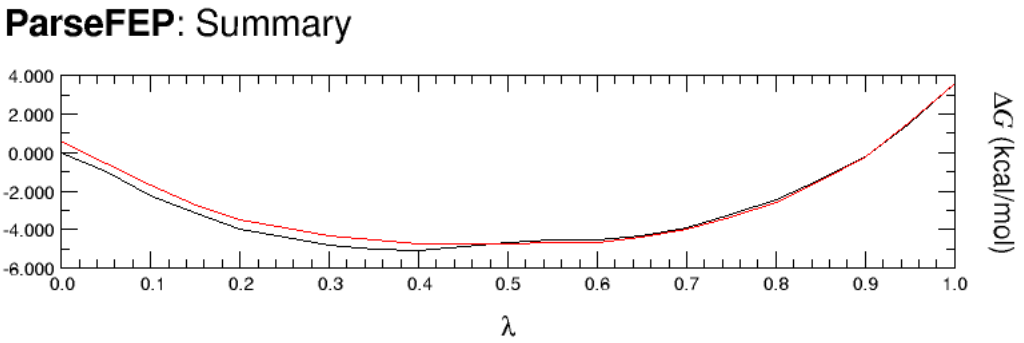




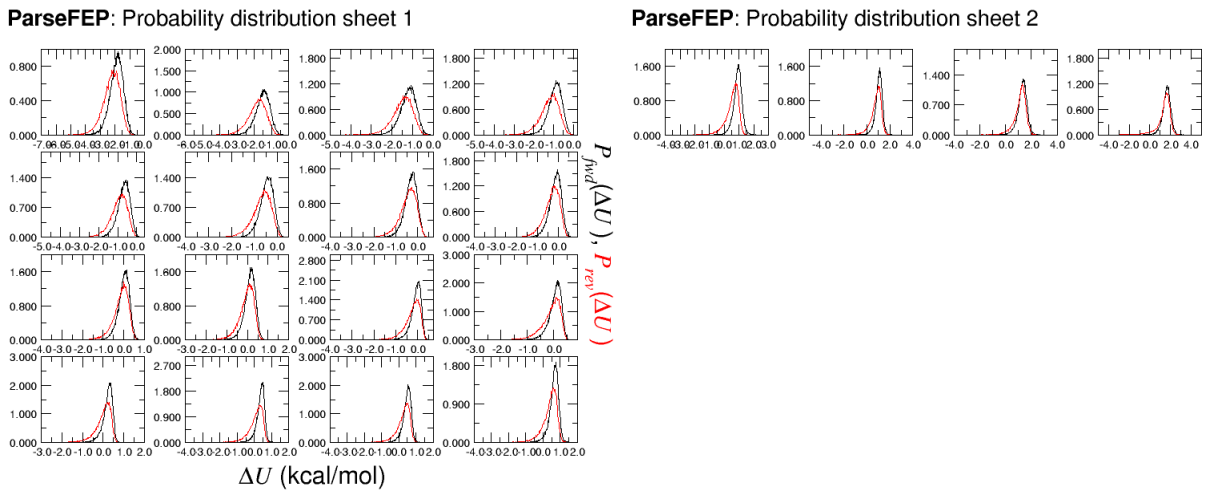


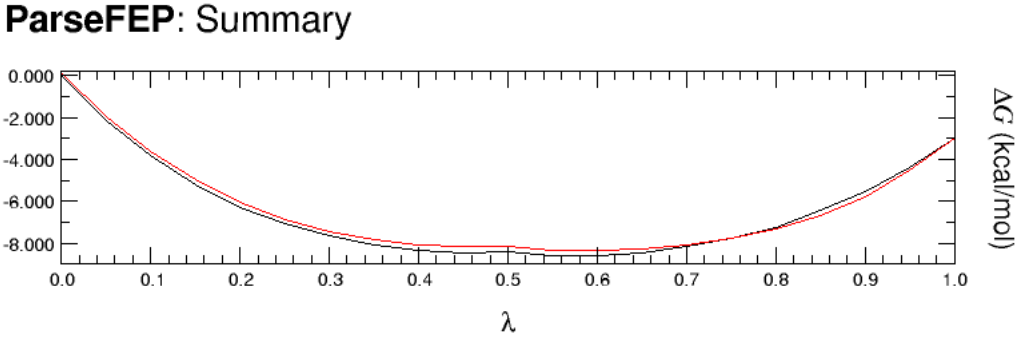
(6.36) B[a]P-DNA ↔ DB[a,e]P-DNA adduct in productive complex



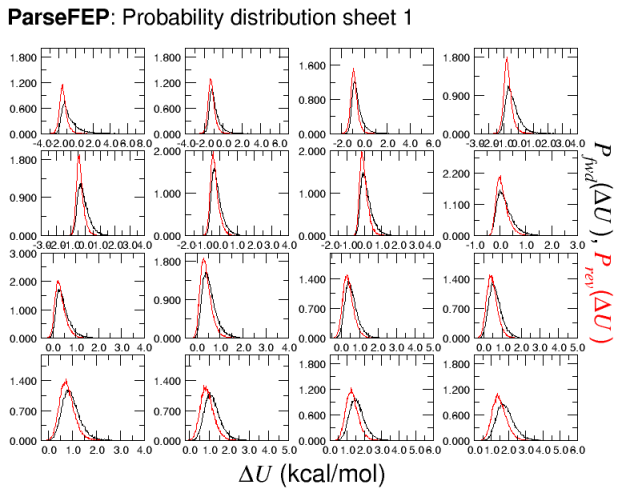


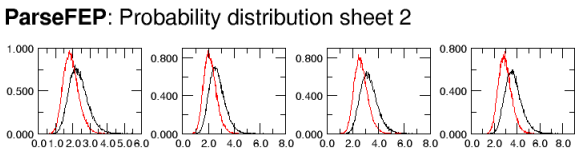
(6.37) B[a]P-DNA ↔ DB[a,h]P-DNA adduct in productive complex

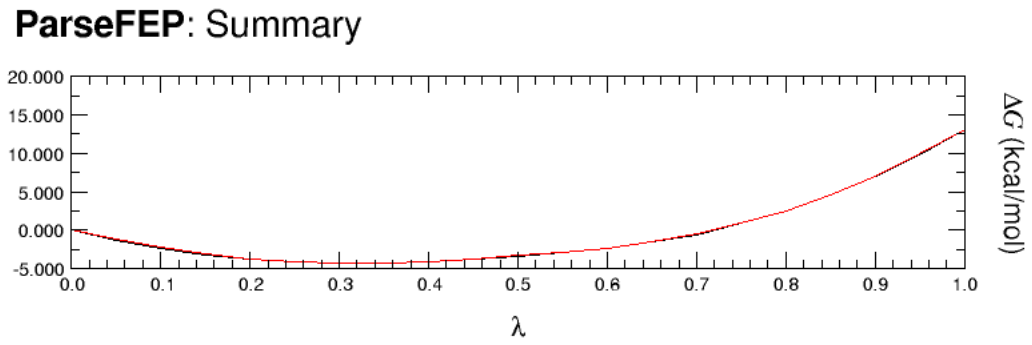




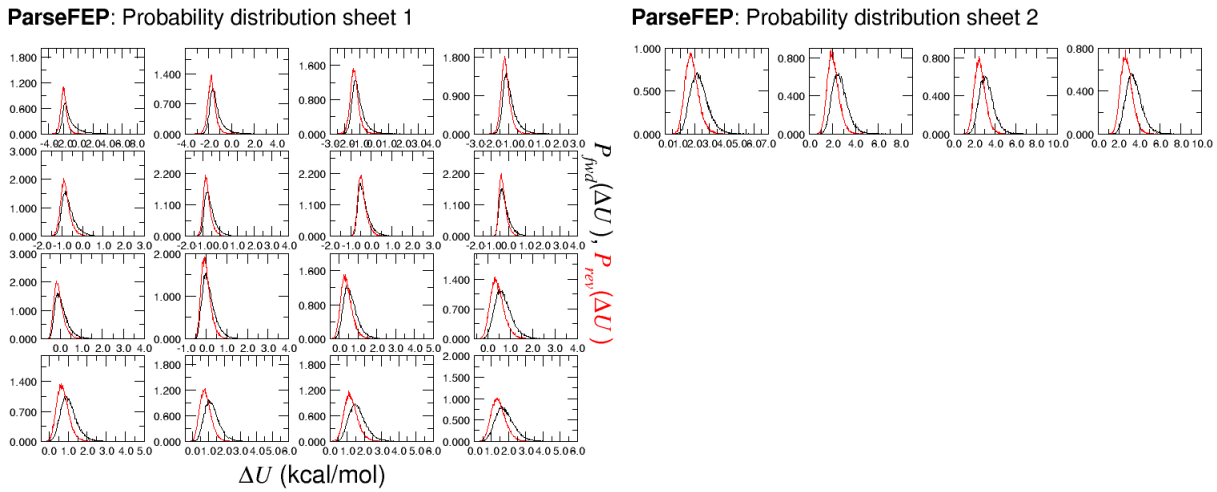
(6.38) DB[a,i]P-DNA ↔ B[a]P-DNA adduct in productive complex

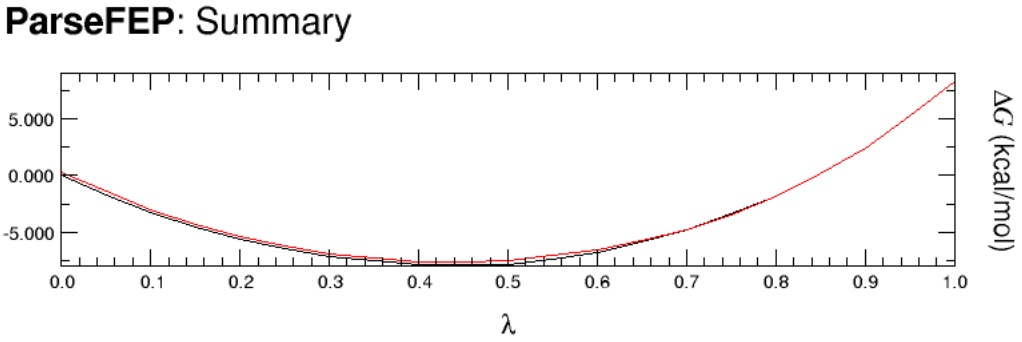




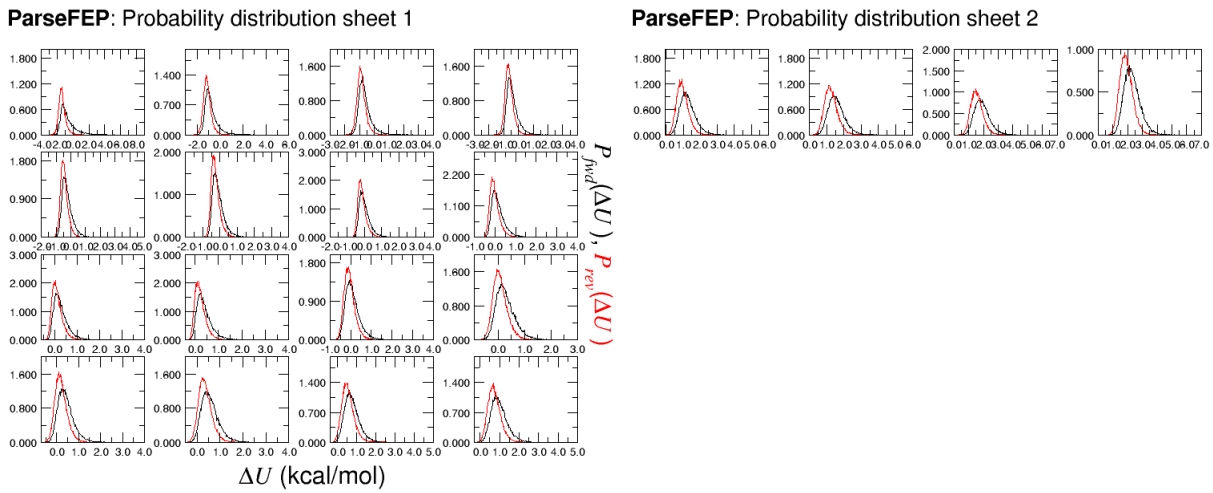


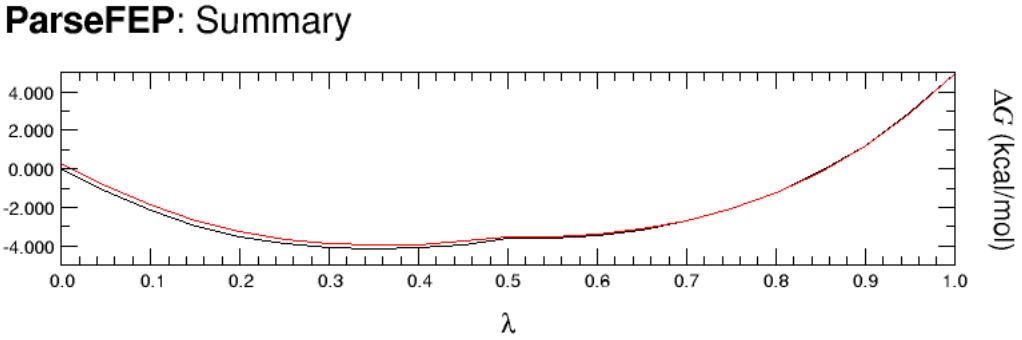
(6.39) B[e]P-DNA ↔ DB[e,l]P-DNA adduct in productive complex





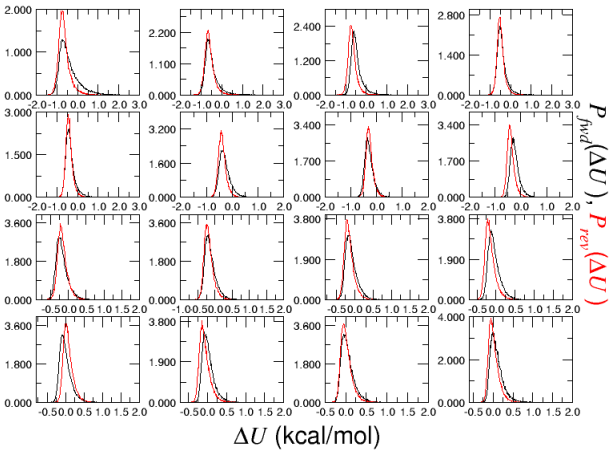
(6.40) B[a]A-DNA ↔ DB[a,c]A-DNA adduct in productive complex



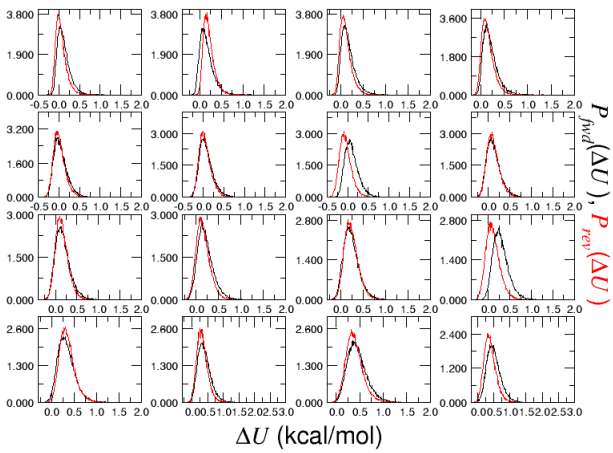


(6.41) B[a]A-DNA ↔ DB[a,h]A-DNA adduct in productive complex

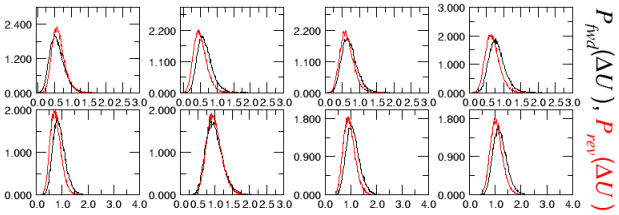
ParseFEP: Probability distribution sheet 1



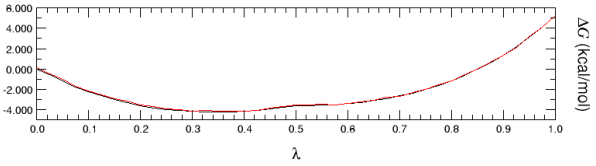
ParseFEP: Probability distribution sheet 2



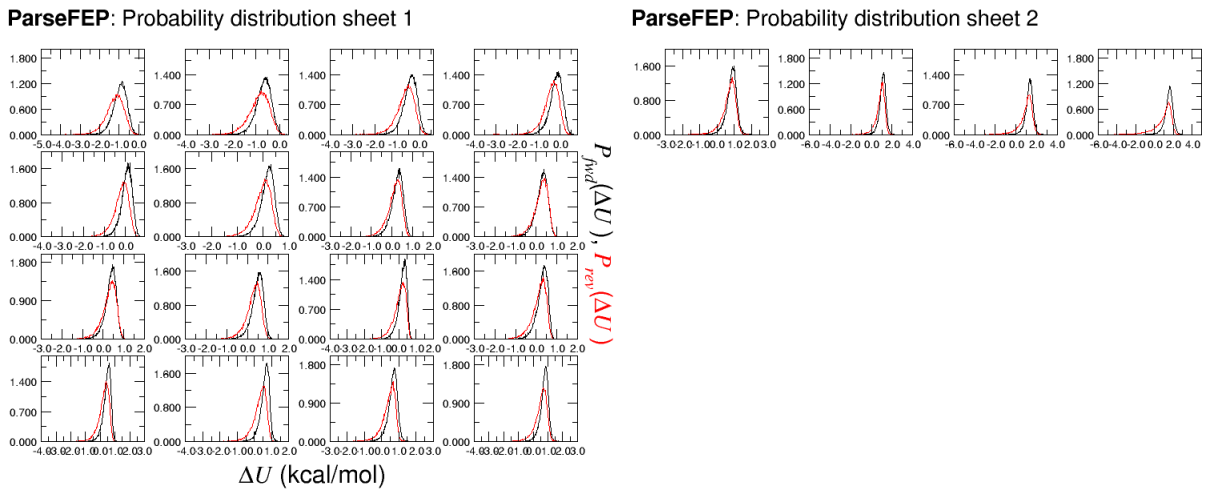
ParseFEP: Probability distribution sheet 3

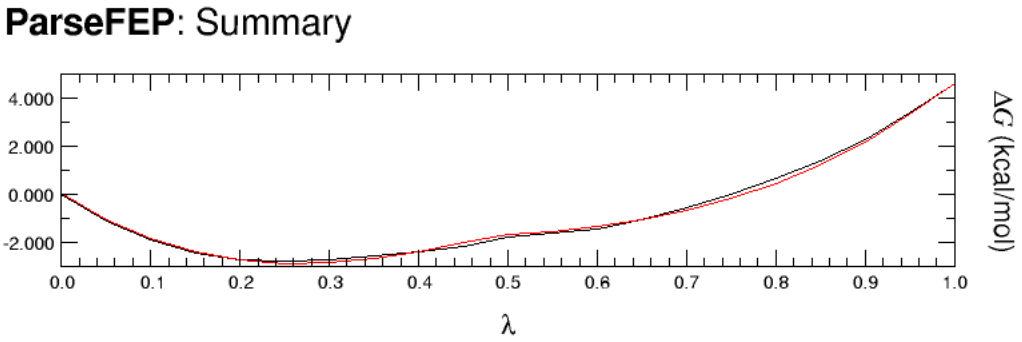


ParseFEP: Summary

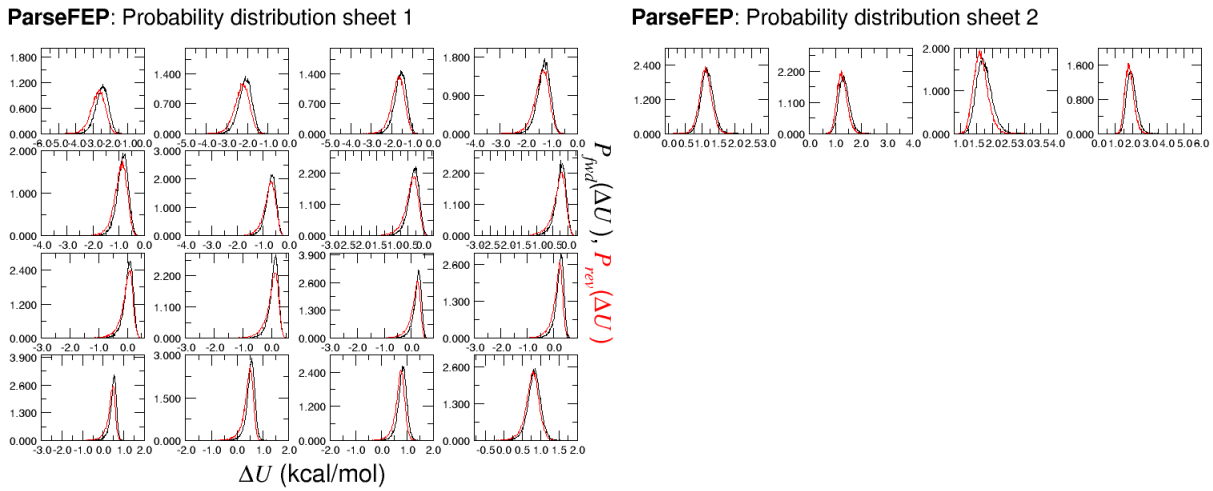


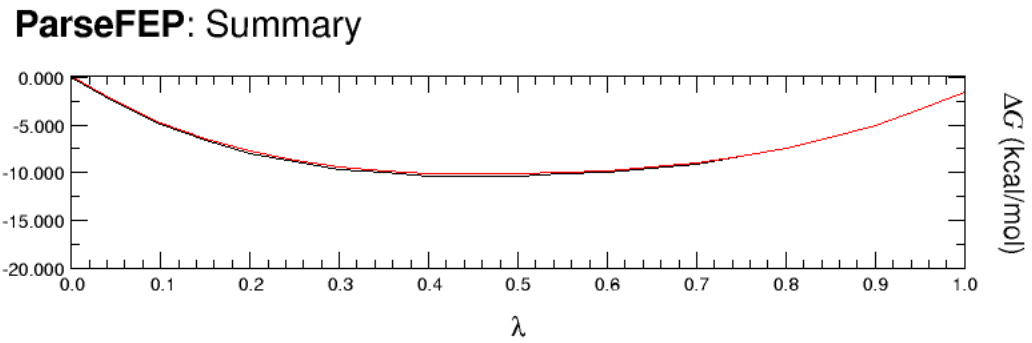
(6.42) B[a]A-DNA ↔ DB[a,j]A-DNA adduct in productive complex



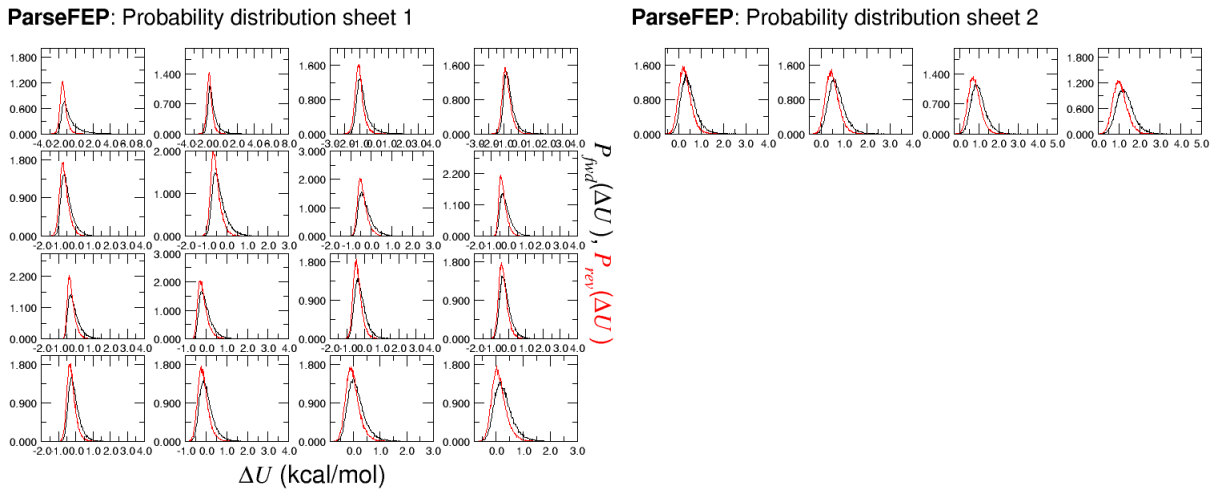


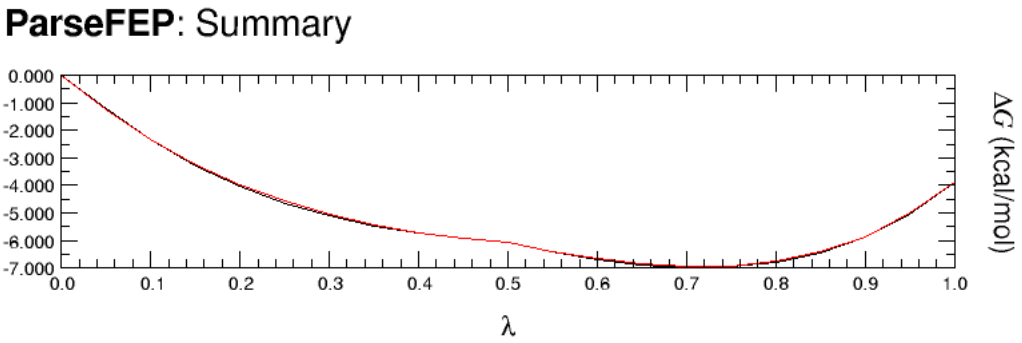
(6.43) PHR-DNA ↔ B[a]A-DNA adduct in productive complex



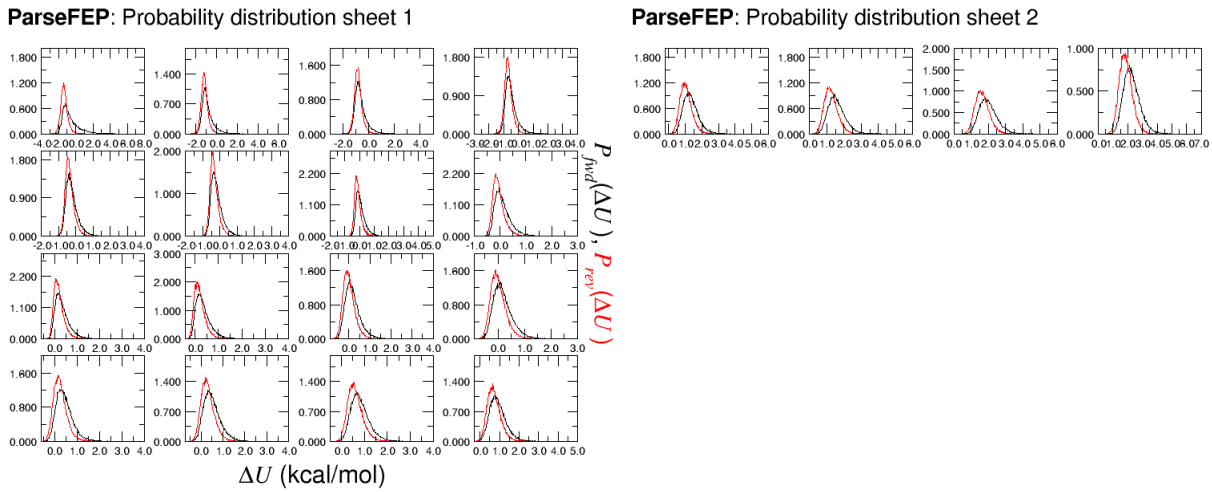


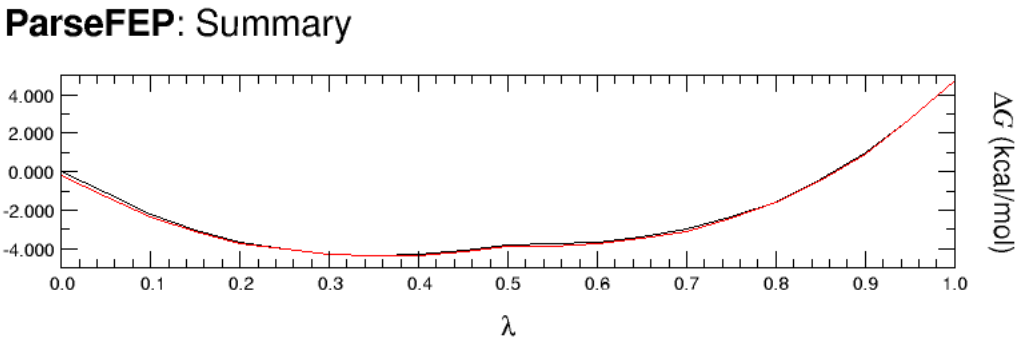
(6.44) B[a]P-DNA ↔ CHR-DNA adduct in productive complex



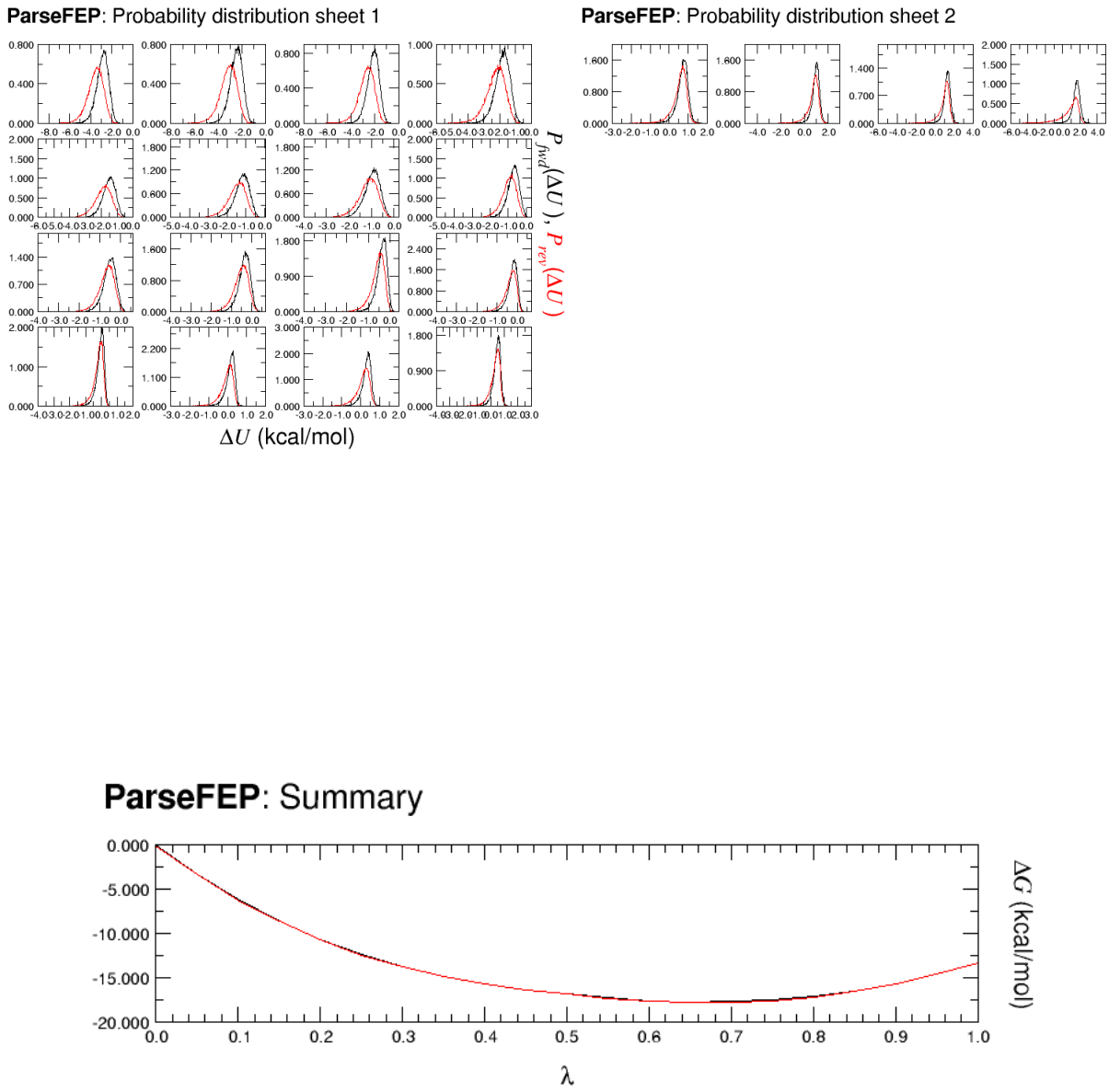


(6.45) CHR-DNA ↔ B[b]C-DNA adduct in productive complex

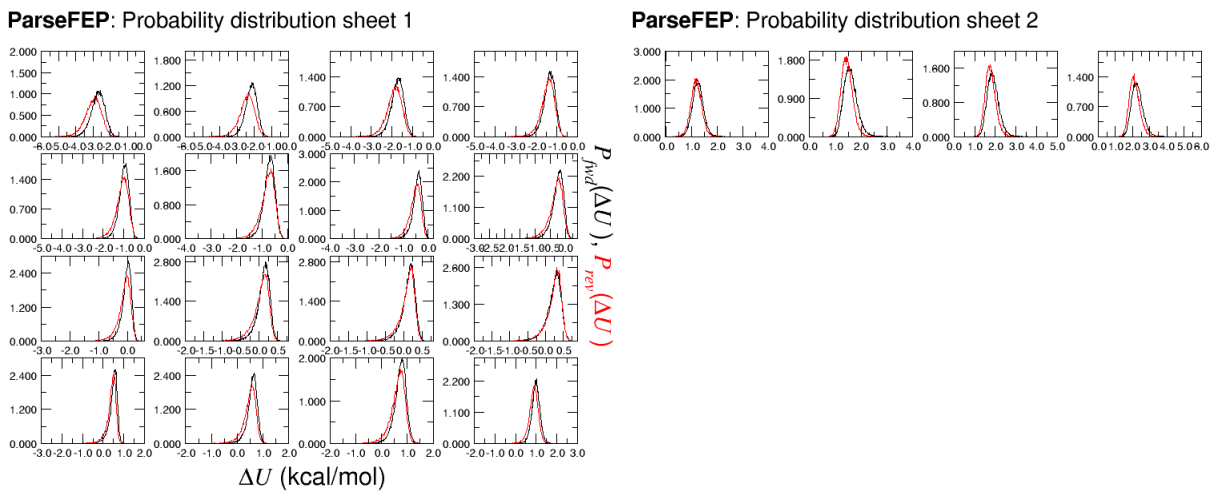


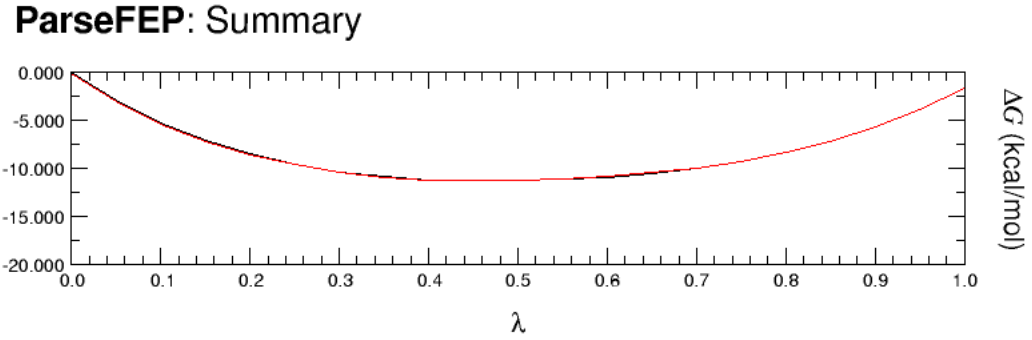


(6.46) PHR-DNA ↔ CHR-DNA adduct in productive complex



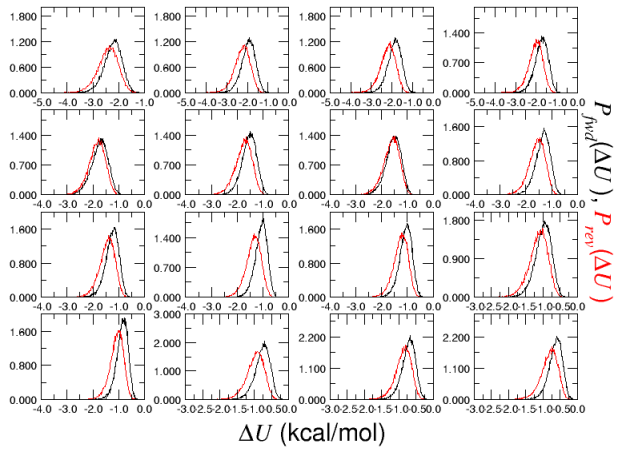
(6.47) B[g]C-DNA ↔ B[c]P-DNA adduct in productive complex



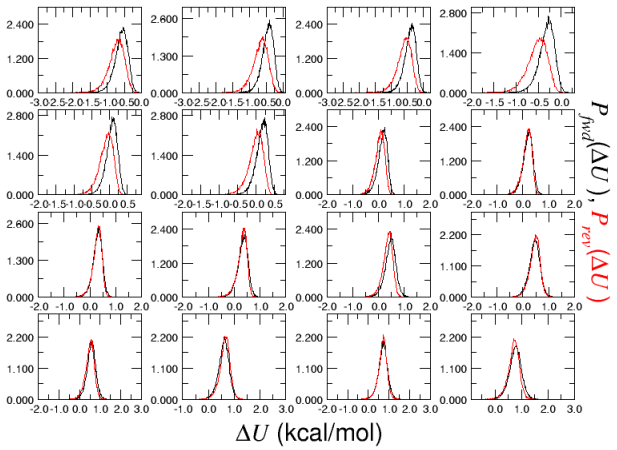


(6.48) DB[a,l]P-DNA ↔ B[g]C-DNA adduct in productive complex

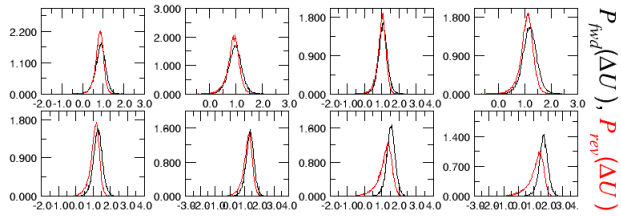
ParseFEP: Probability distribution sheet 1



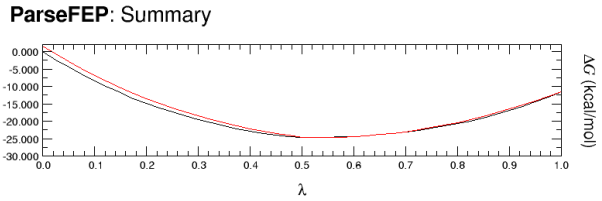
ParseFEP: Probability distribution sheet 2



ParseFEP: Probability distribution sheet 3

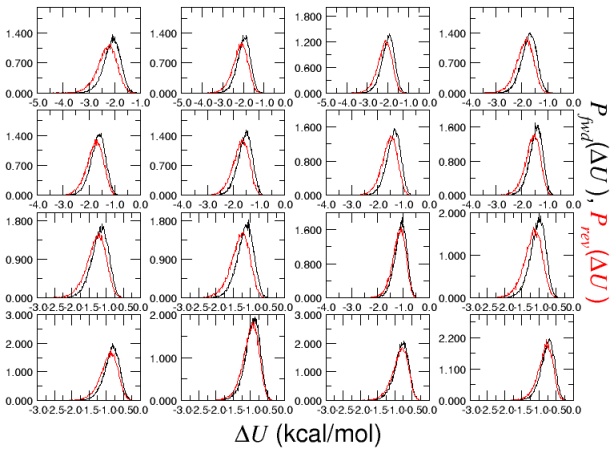


ParseFEP: Summary

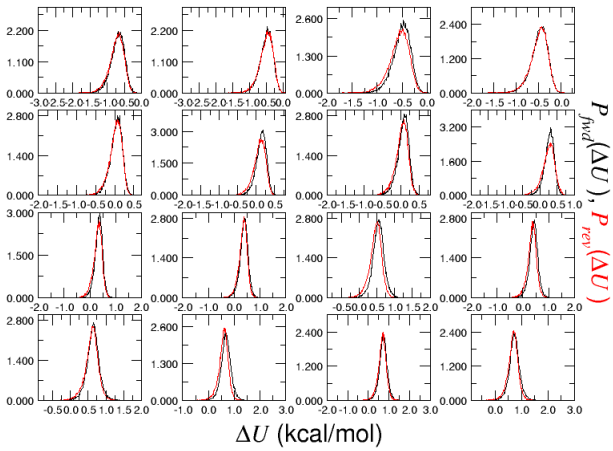


(6.49) DB[a,l]P-DNA ↔ B[a]P-DNA adduct in productive complex

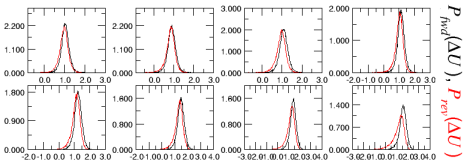
ParseFEP: Probability distribution sheet 1



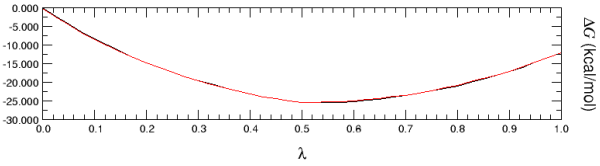
ParseFEP: Probability distribution sheet 2



ParseFEP: Probability distribution sheet 3



ParseFEP: Summary

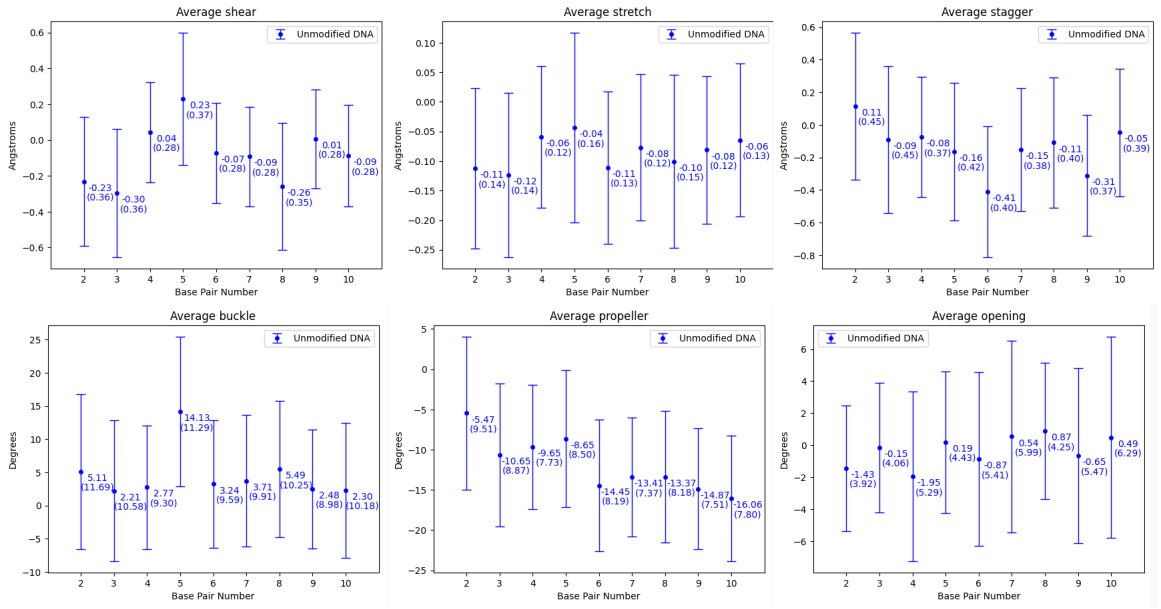


(6.50) B[g]C-DNA ↔ CHR-DNA adduct in productive complex

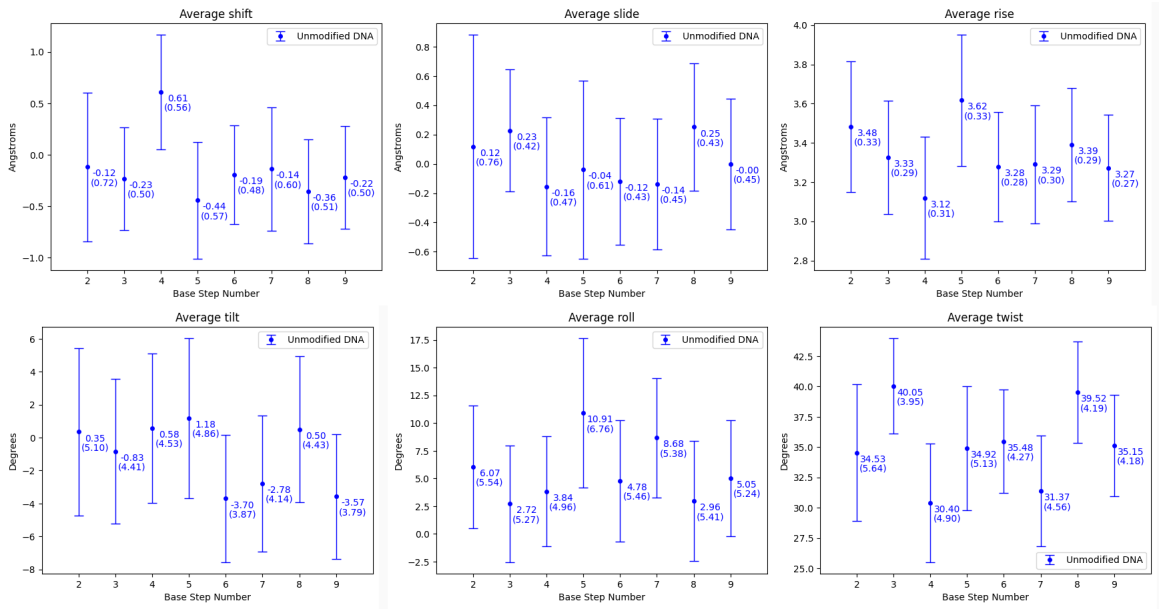
6.2 Appendix B

6.2.1 MD Trajectories and Rigid Body Parameters

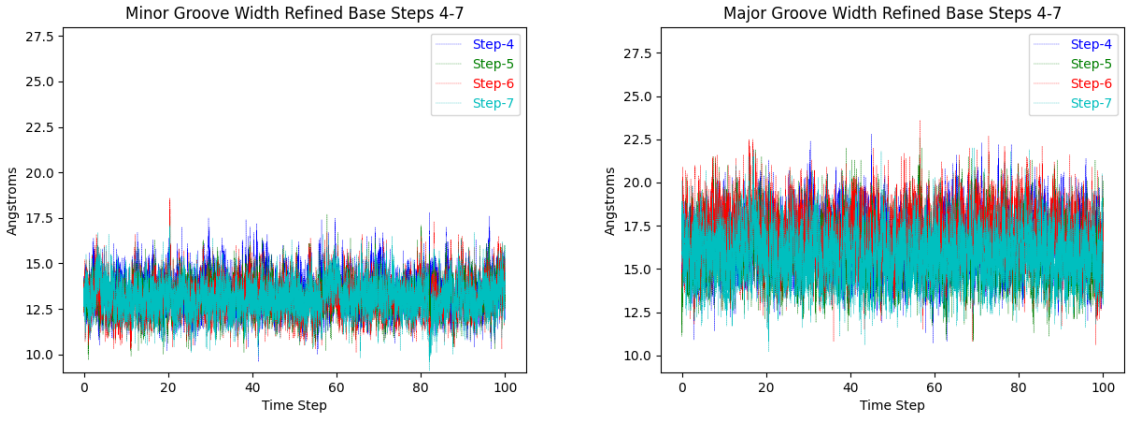
6.2.1.1 Unmodified NRAS(Q61) 11-mer



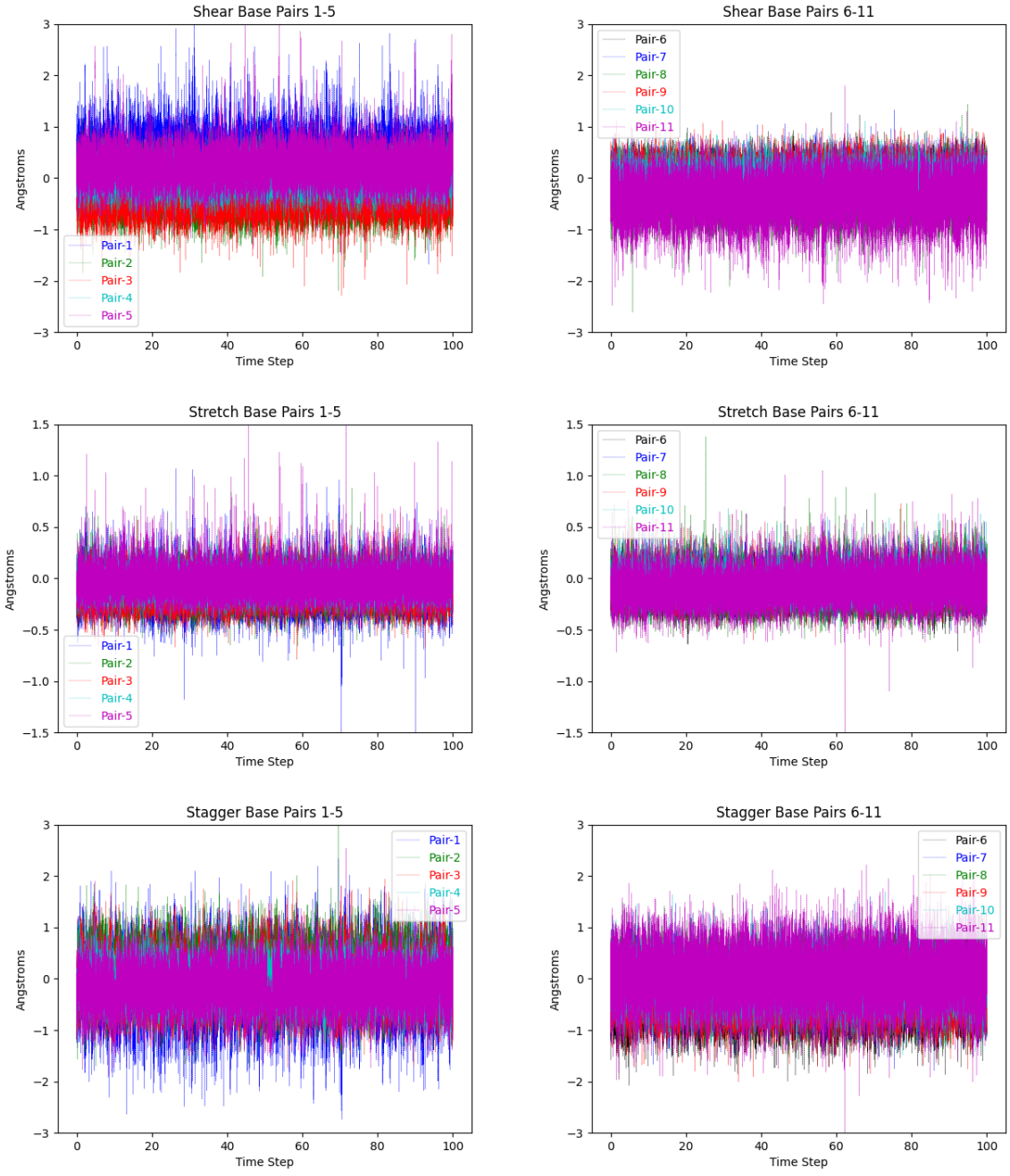
(6.51) Unmodified DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



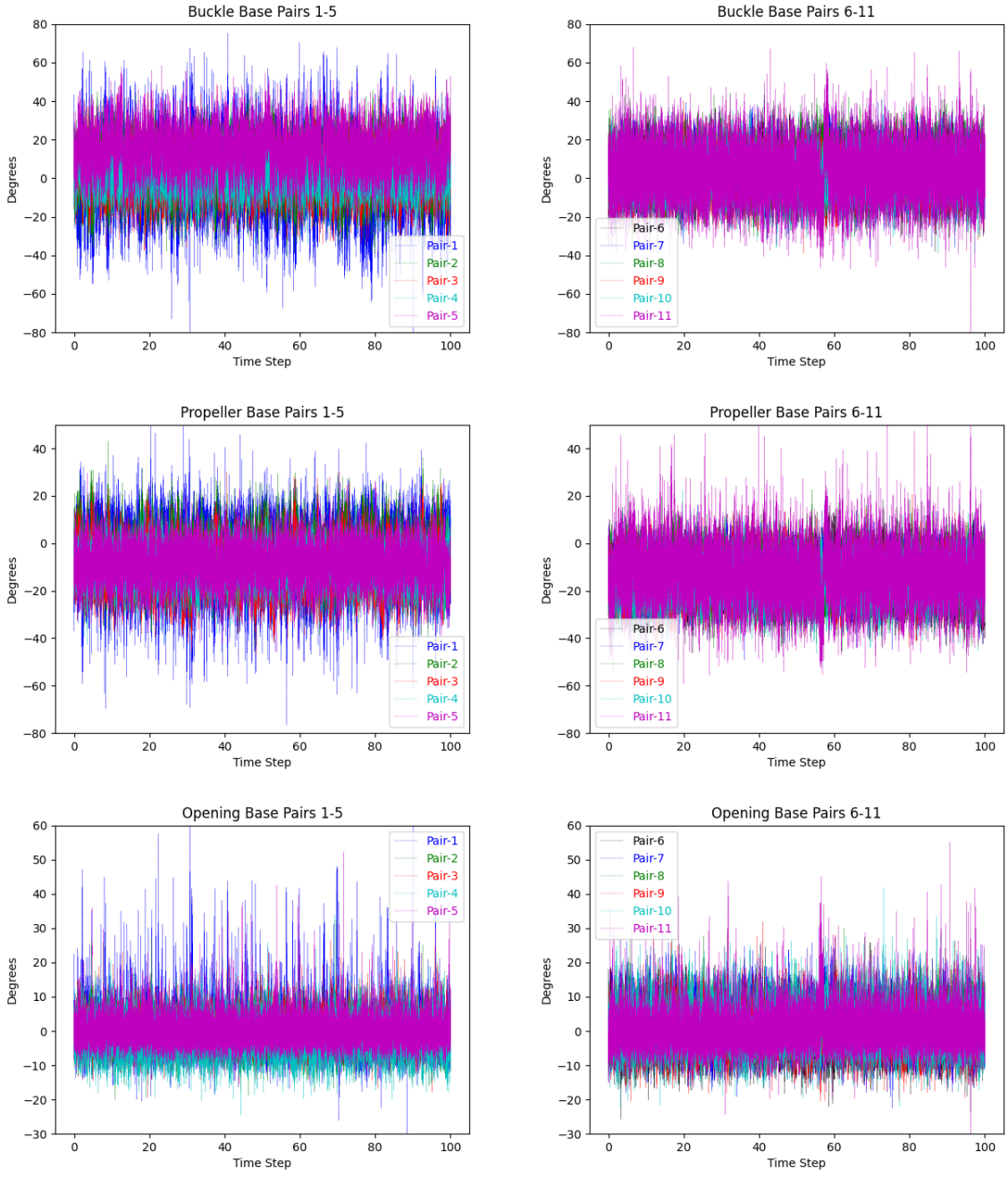
(6.52) Unmodified DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



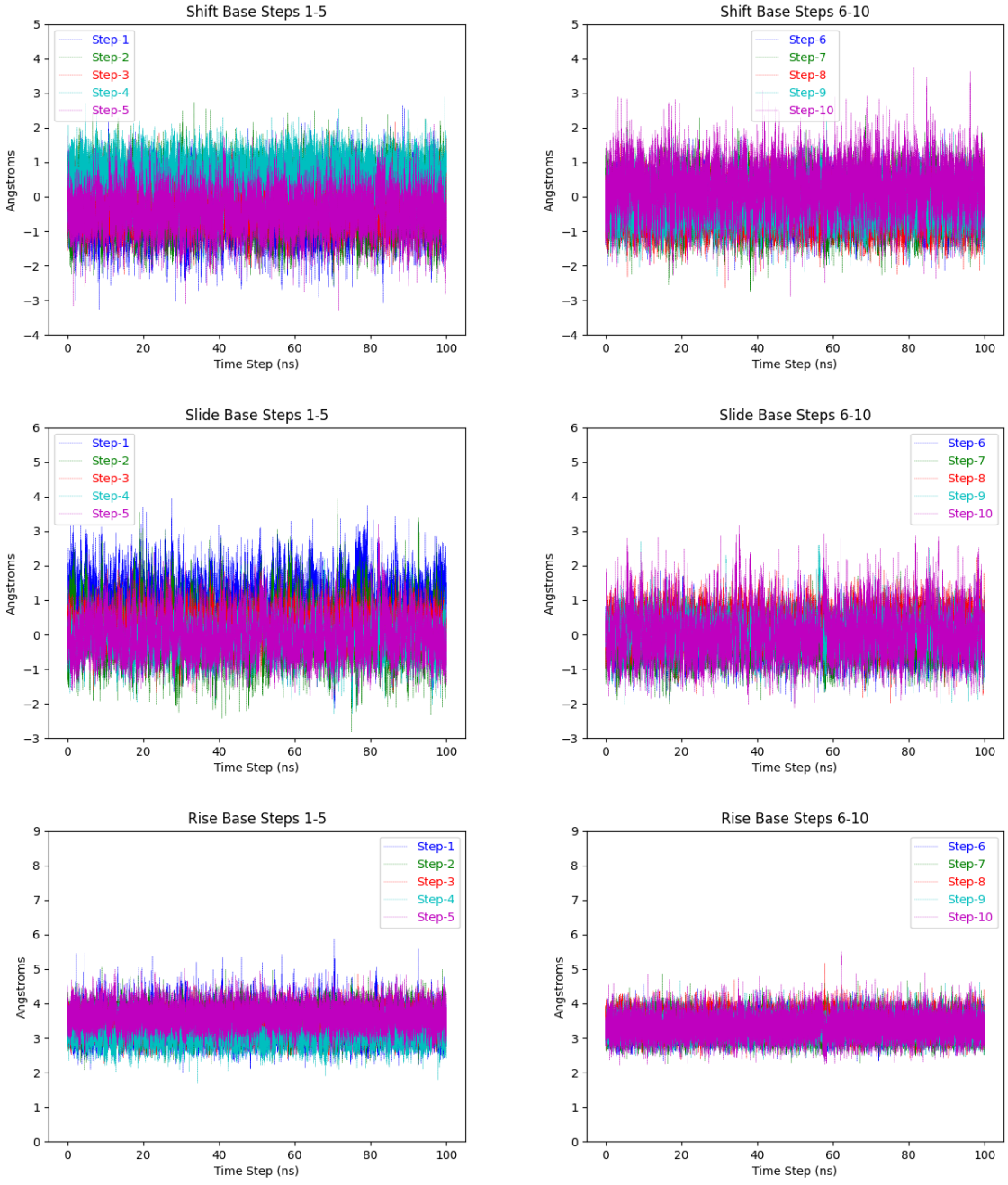
(6.53) Unmodified DNA: Refined major and minor groove trajectories



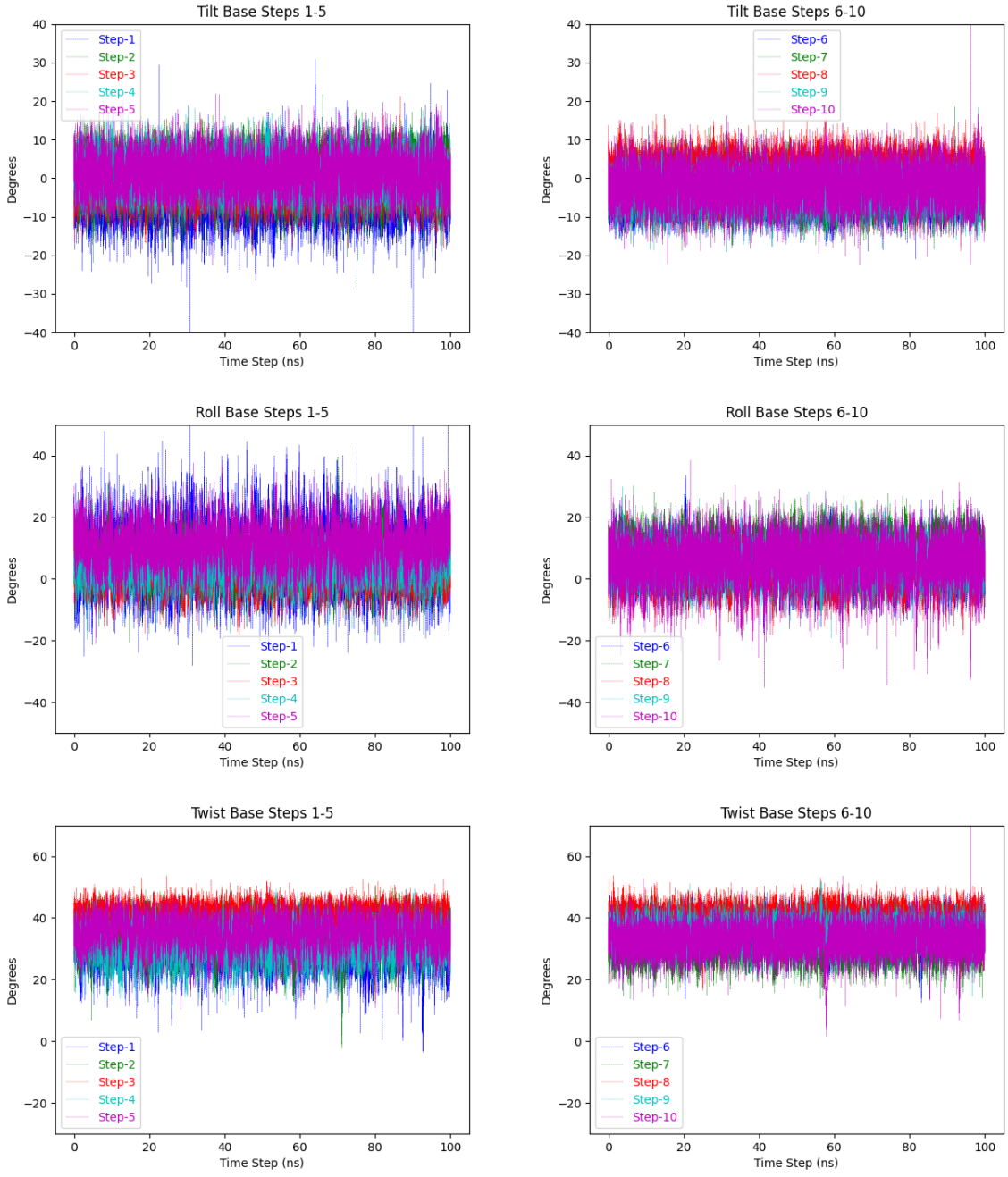
(6.54) Unmodified DNA: Base pair trajectories



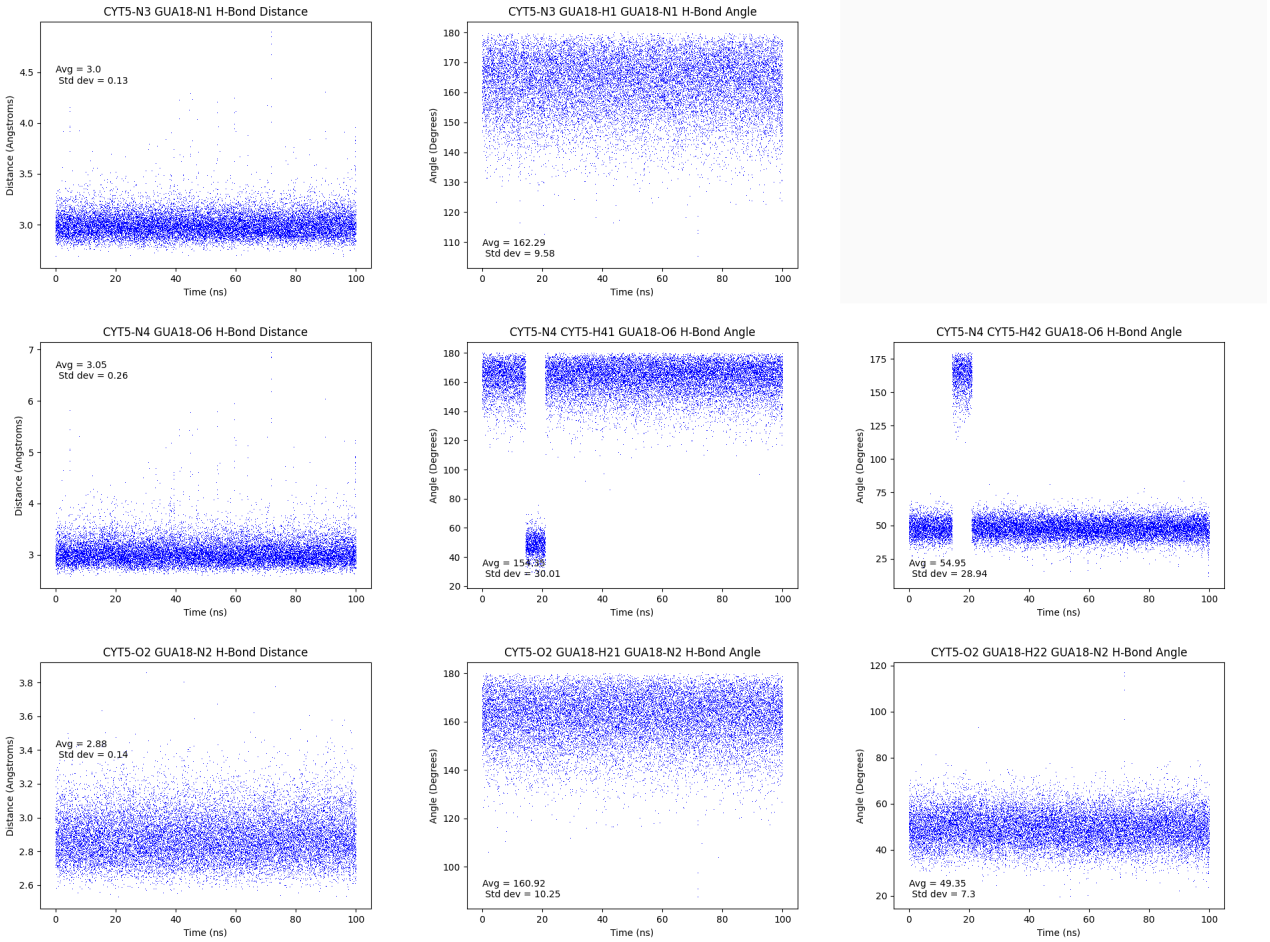
(6.55) Unmodified DNA: Base pair trajectories



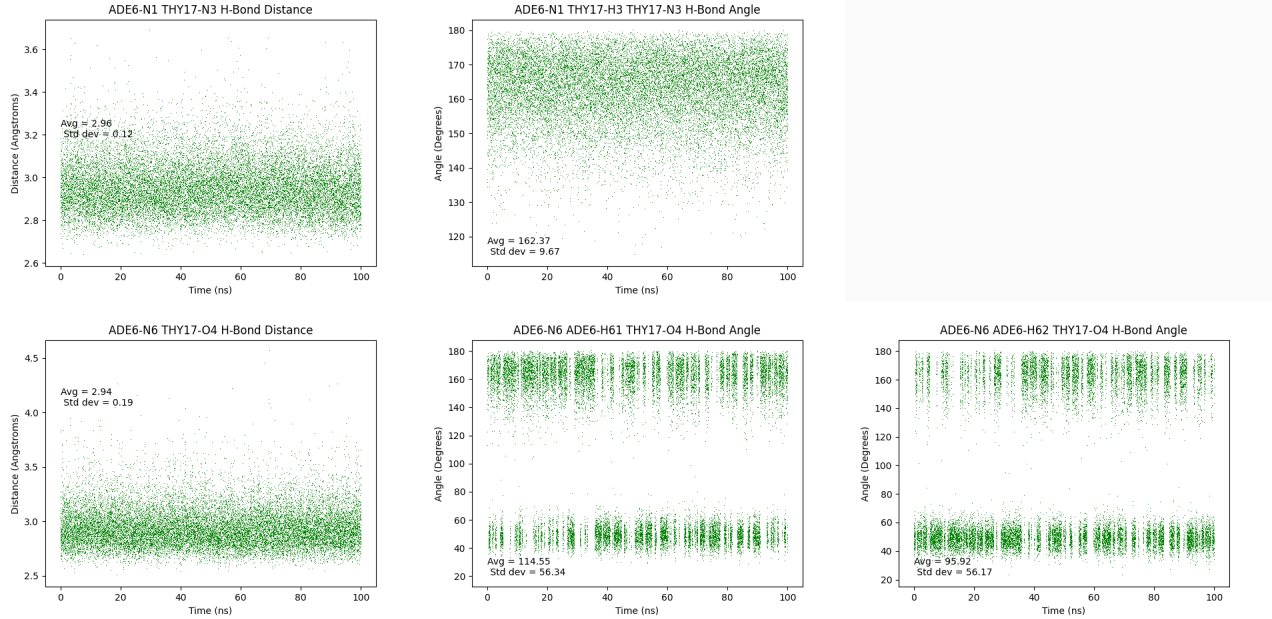
(6.56) Unmodified DNA: Base step trajectories



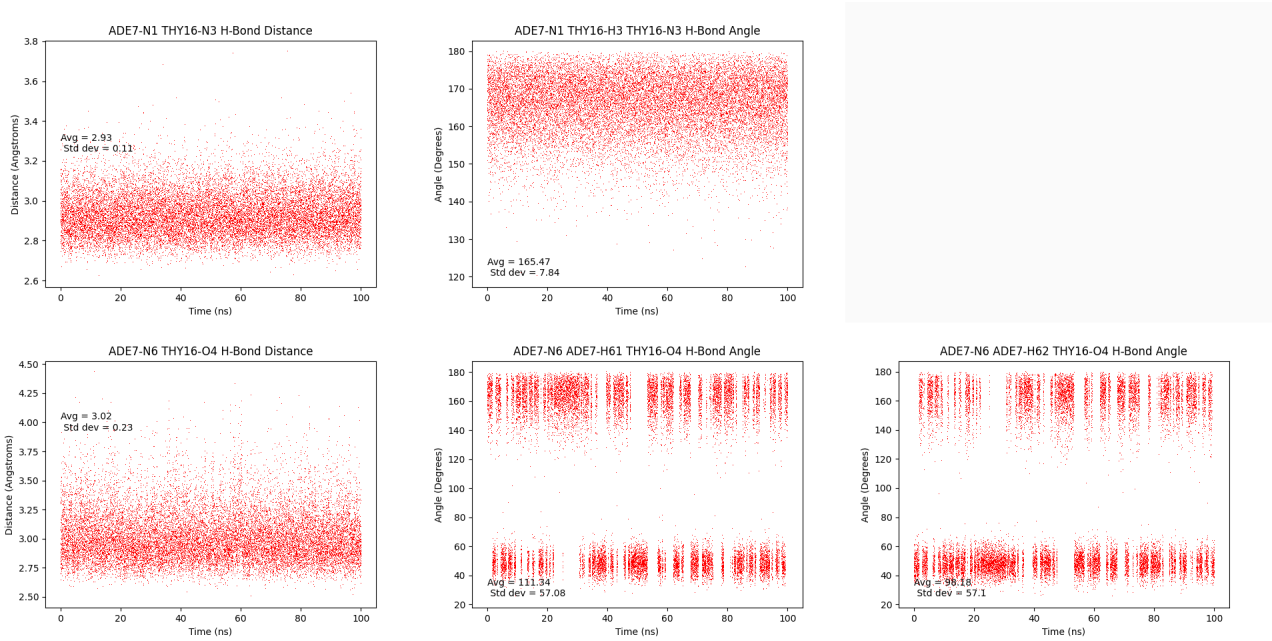
(6.57) Unmodified DNA: Base step trajectories



(6.58) Unmodified DNA: dC_5:dG_{18} hydrogen bond trajectories



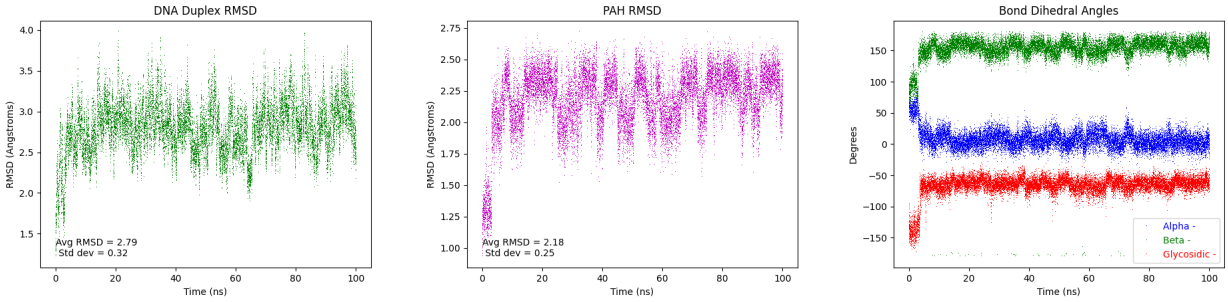
(6.59) Unmodified DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories



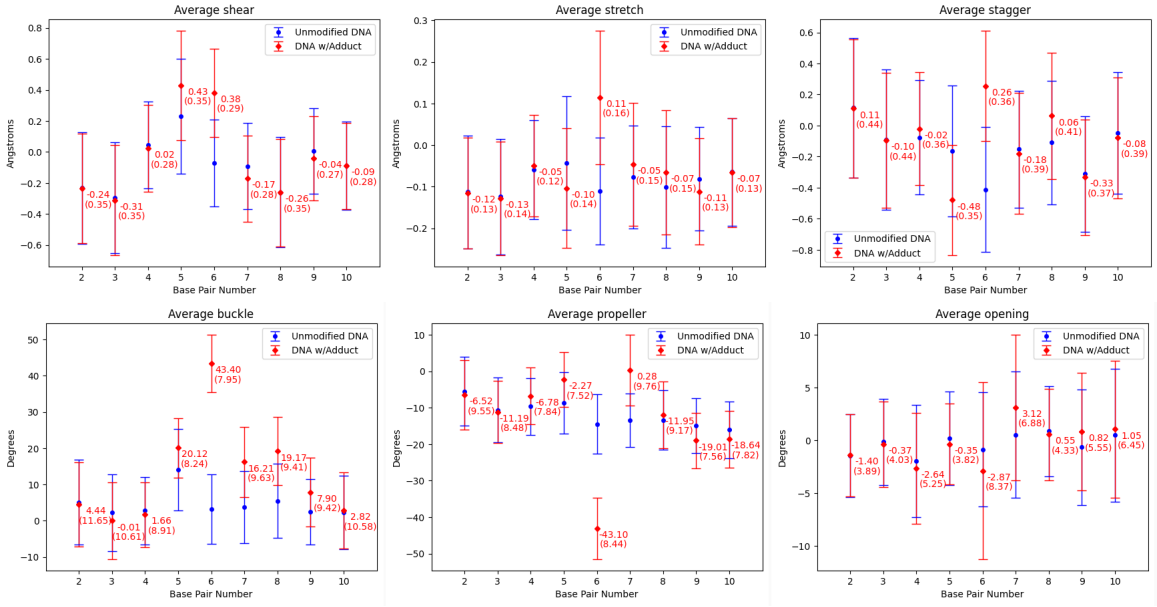
(6.60) Unmodified DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

6.2.1.2 B[a]P-DNA

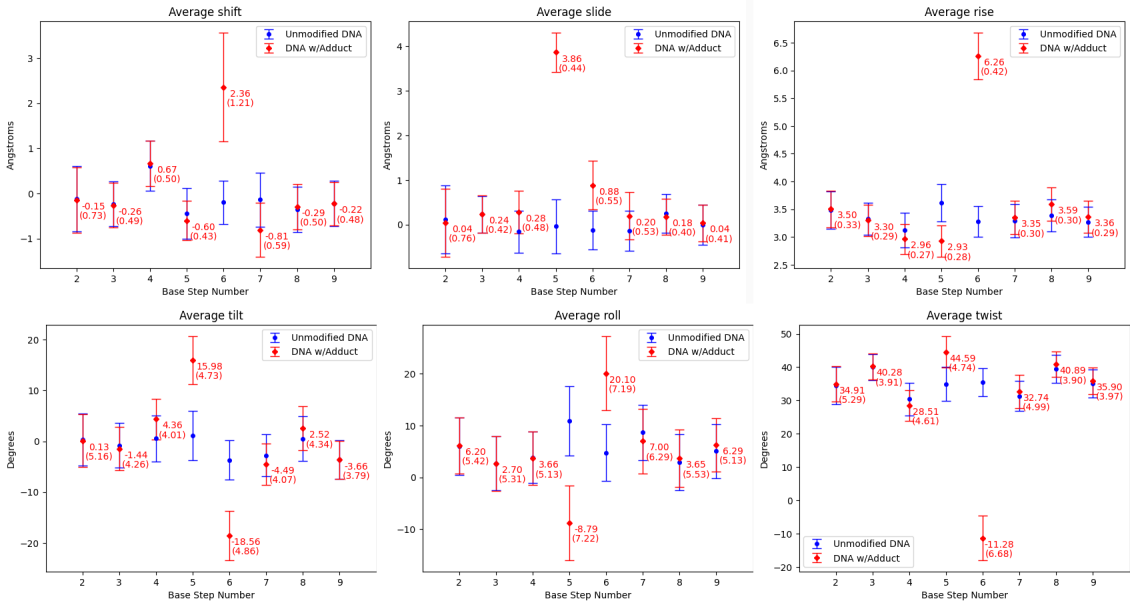
Average base step and base pair rigid-body parameters for the B[a]P-DNA system were calculated over an equilibrated 90 ns subset of the NPT production run.



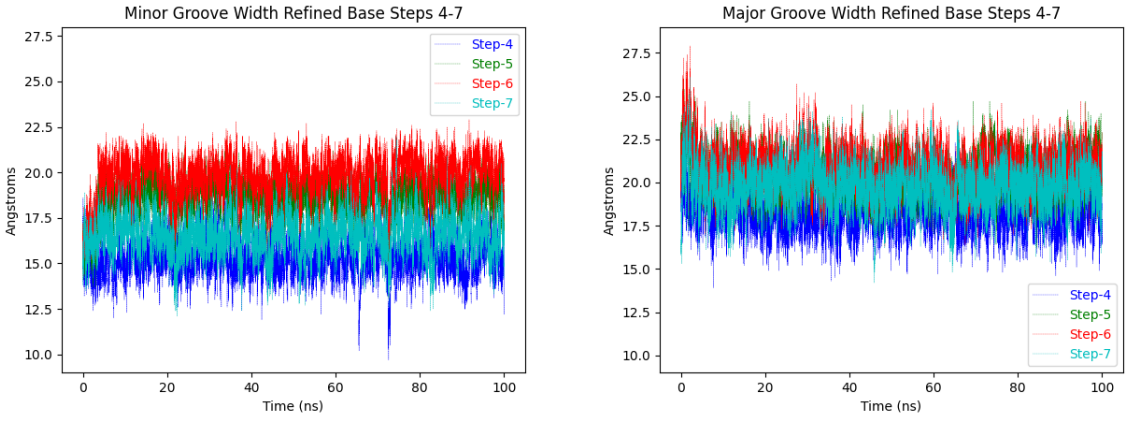
(6.61) B[a]P-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



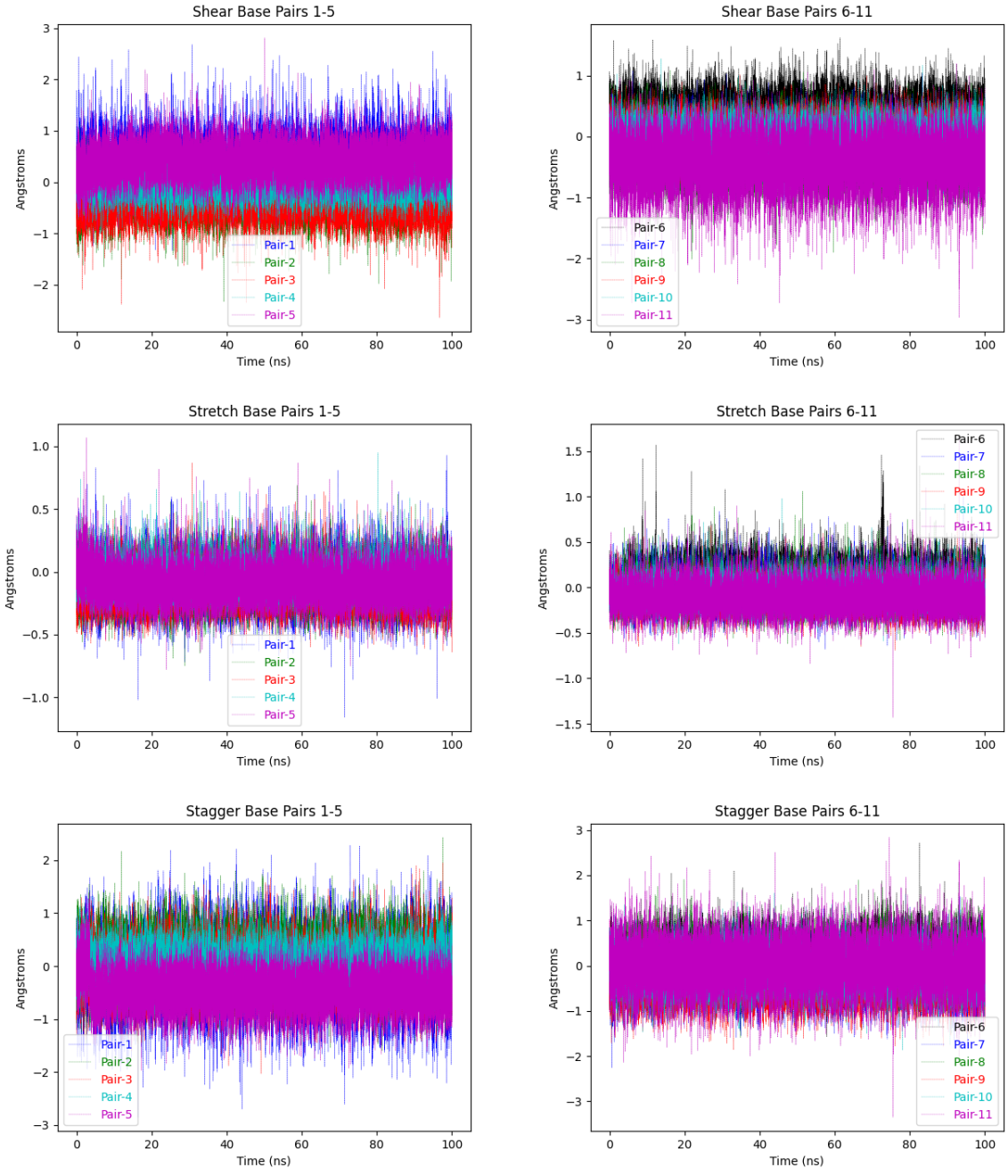
(6.62) B[a]P-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



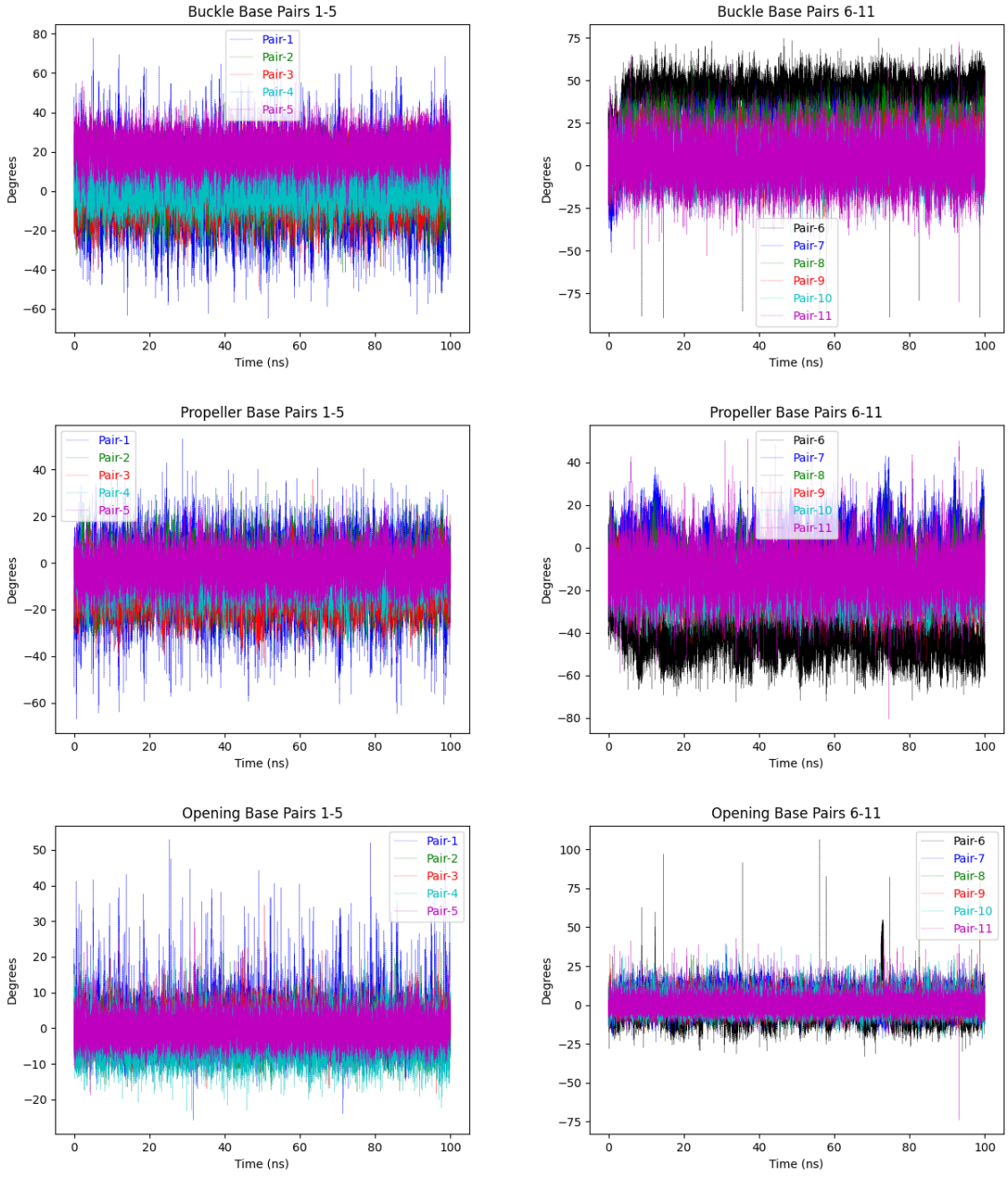
(6.63) B[a]P-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



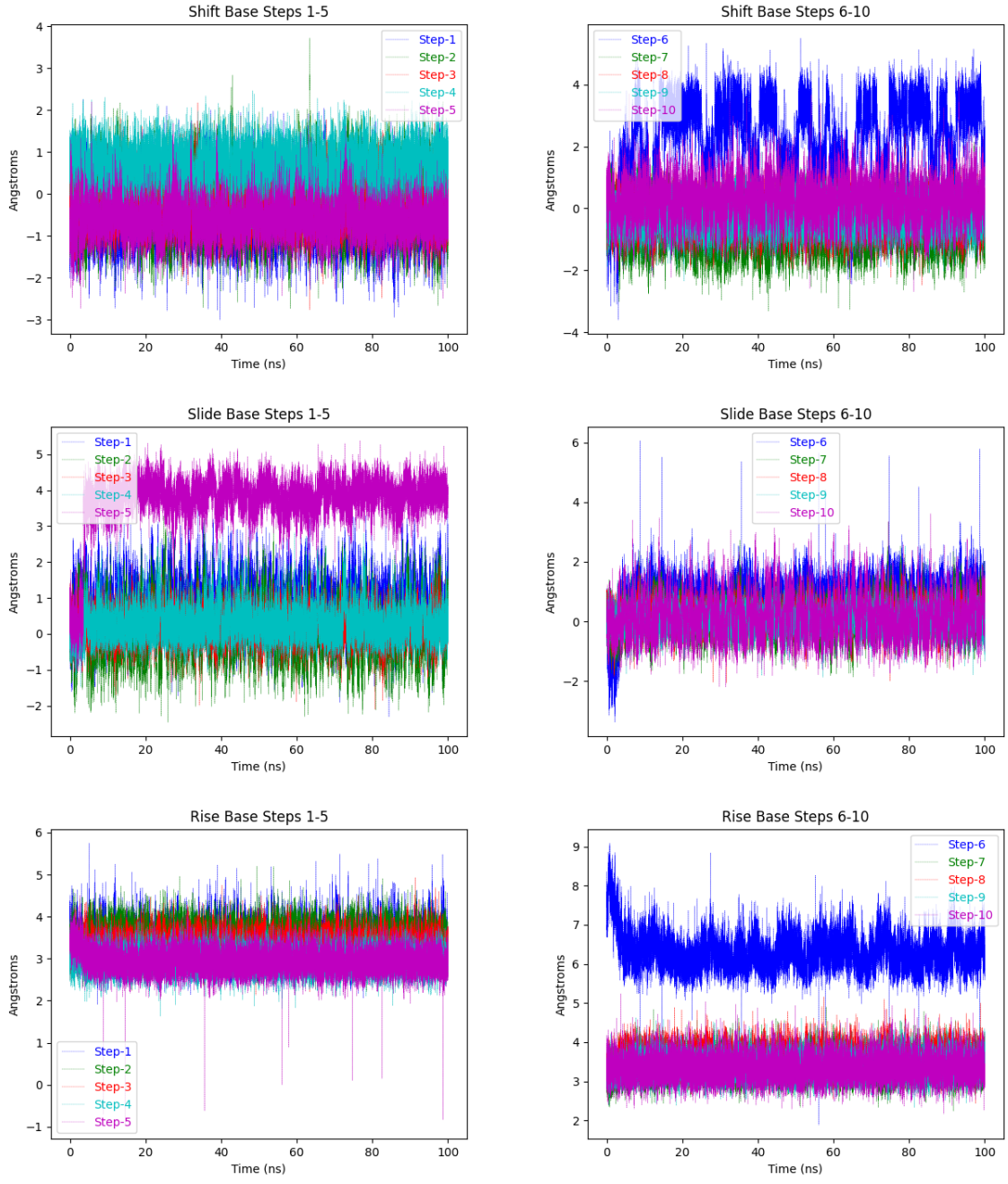
(6.64) B[a]P-DNA: Refined major and minor groove trajectories



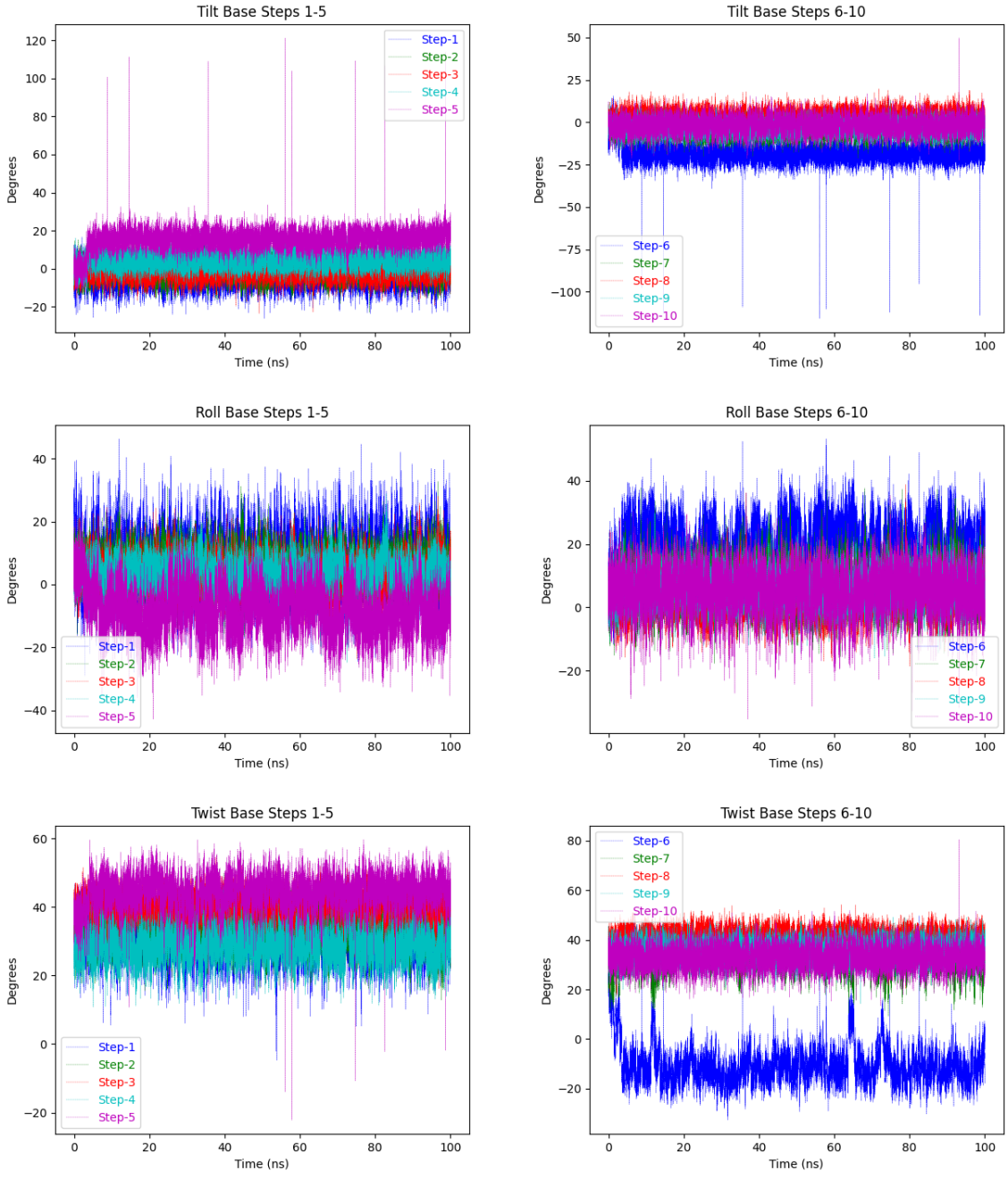
(6.65) B[a]P-DNA: Base pair trajectories



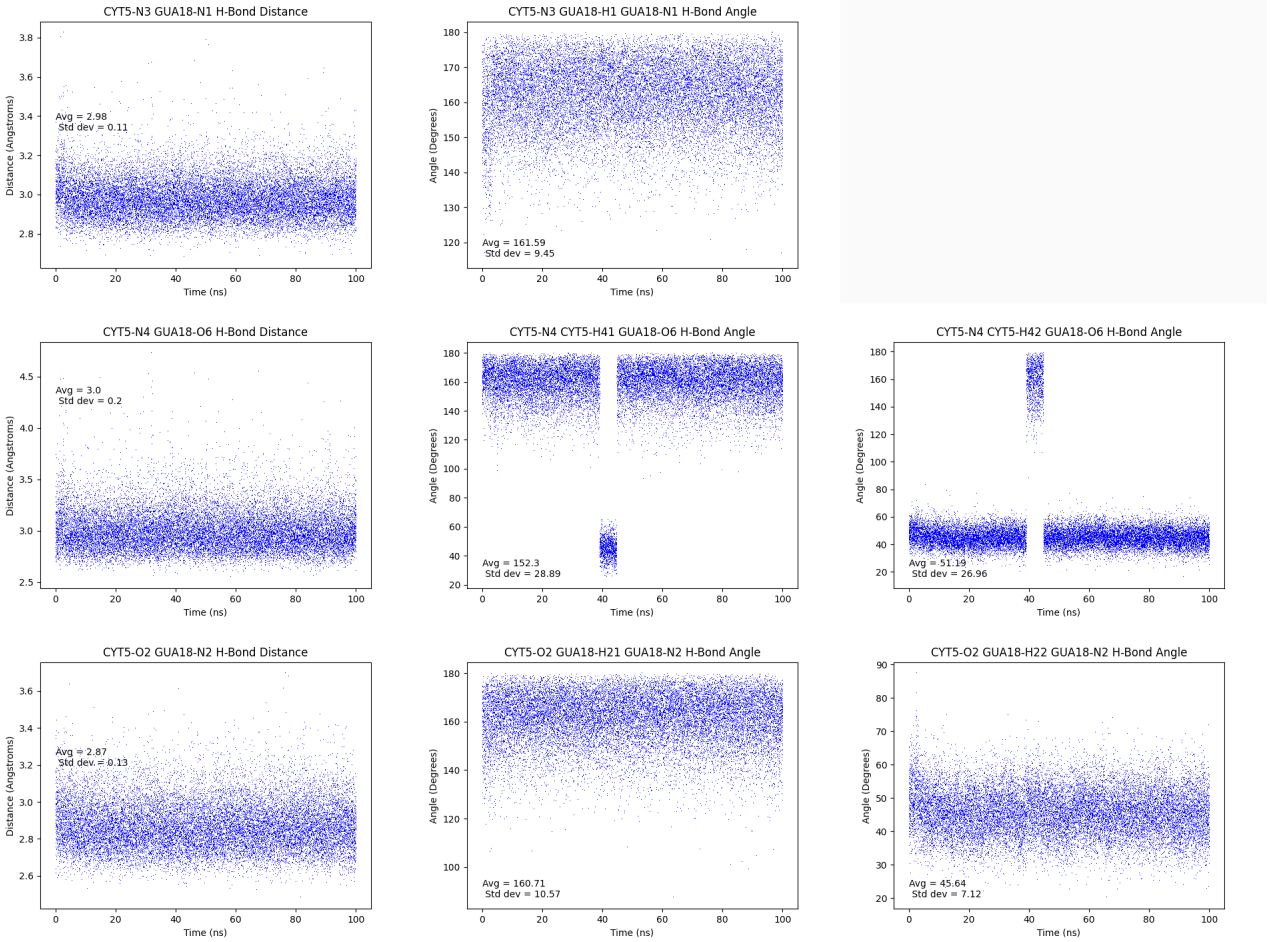
(6.66) B[a]P-DNA: Base pair trajectories



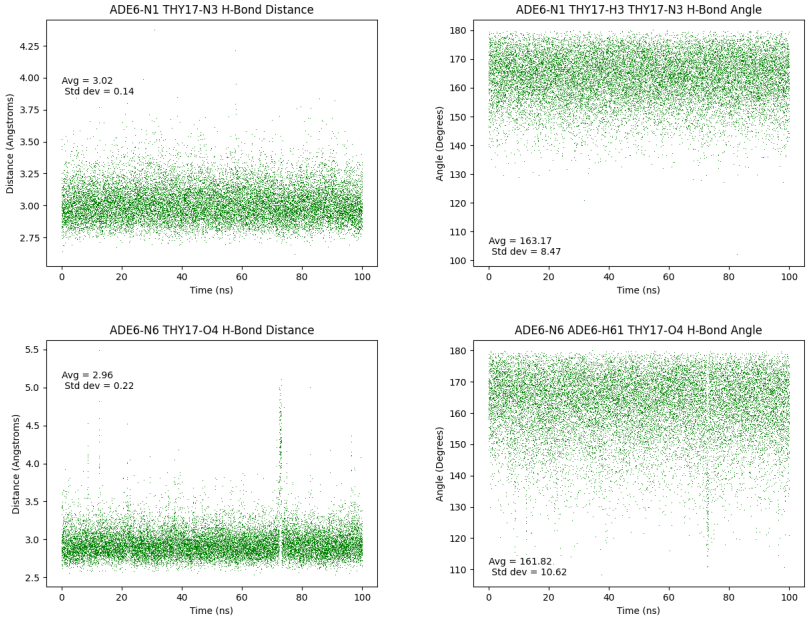
(6.67) B[a]P-DNA: Base step trajectories



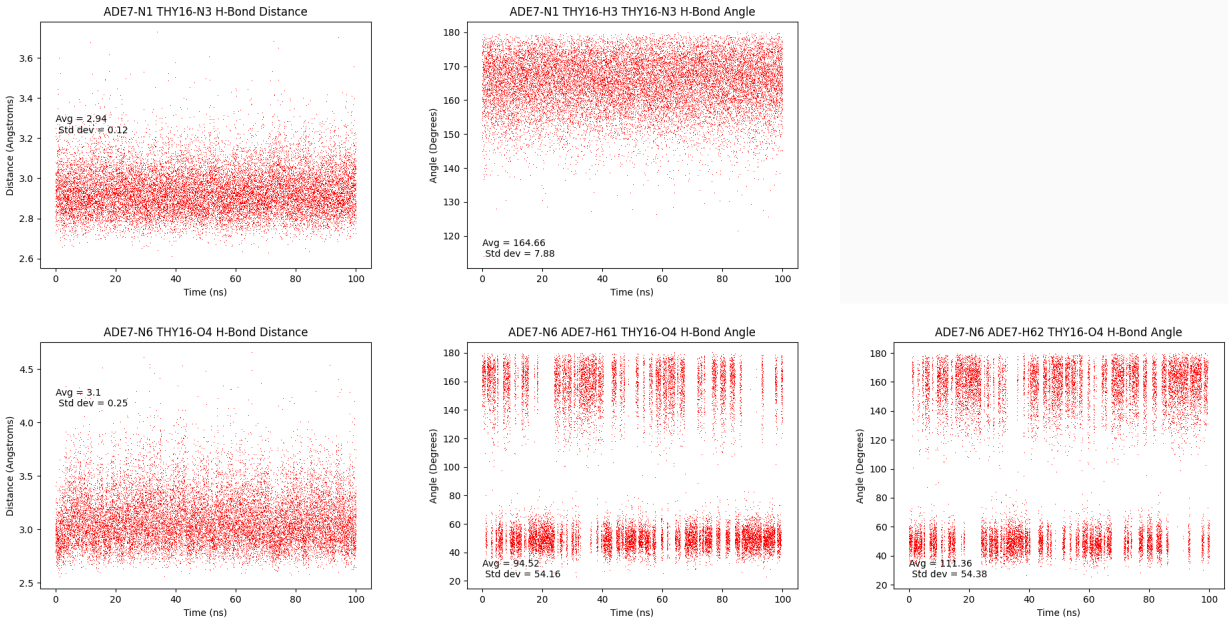
(6.68) B[a]P-DNA: Base step trajectories



(6.69) B[a]P-DNA: dC_5:dG_{18} hydrogen bond trajectories

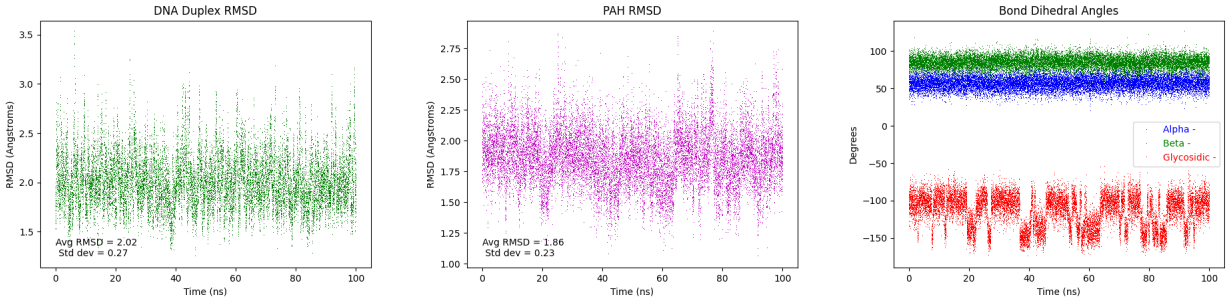


(6.70) B[a]P-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

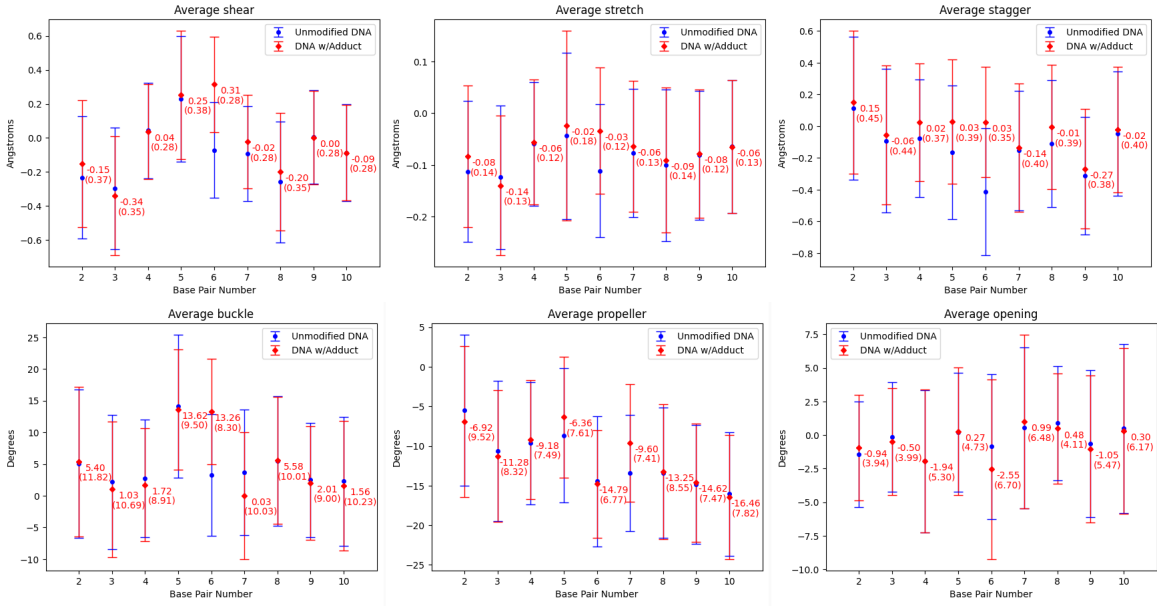


(6.71) B[a]P-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

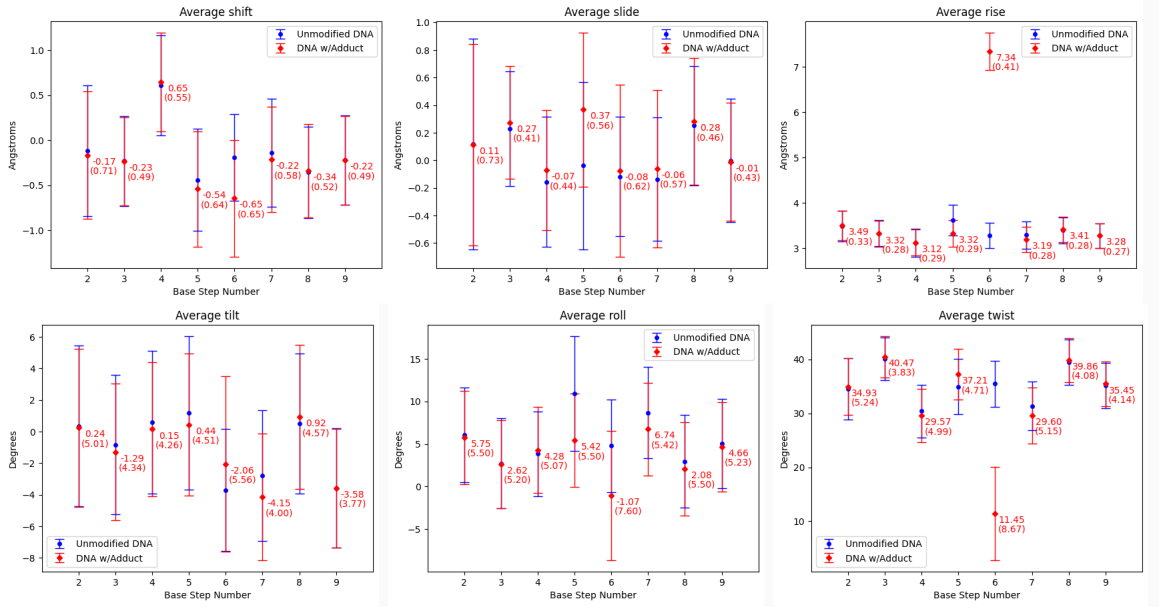
6.2.1.3 DB[a,l]P-DNA



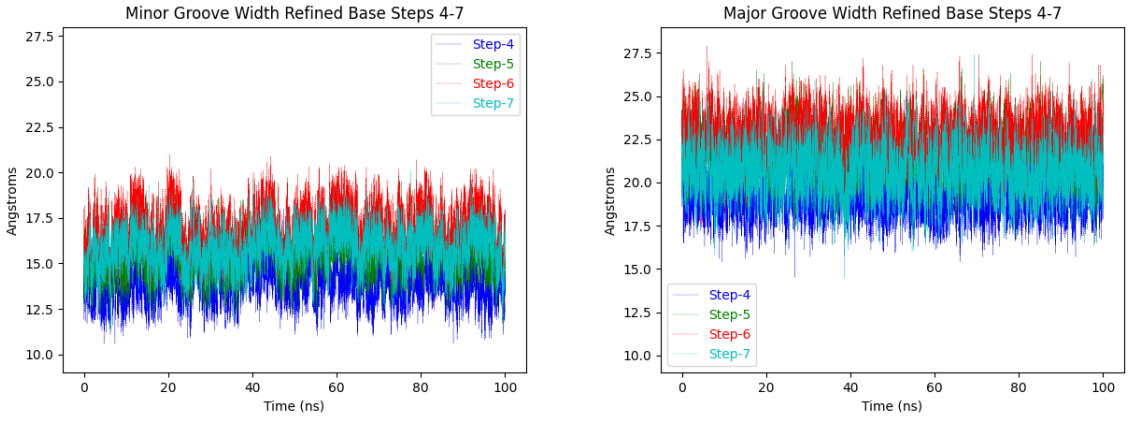
(6.72) DB[a,l]P-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



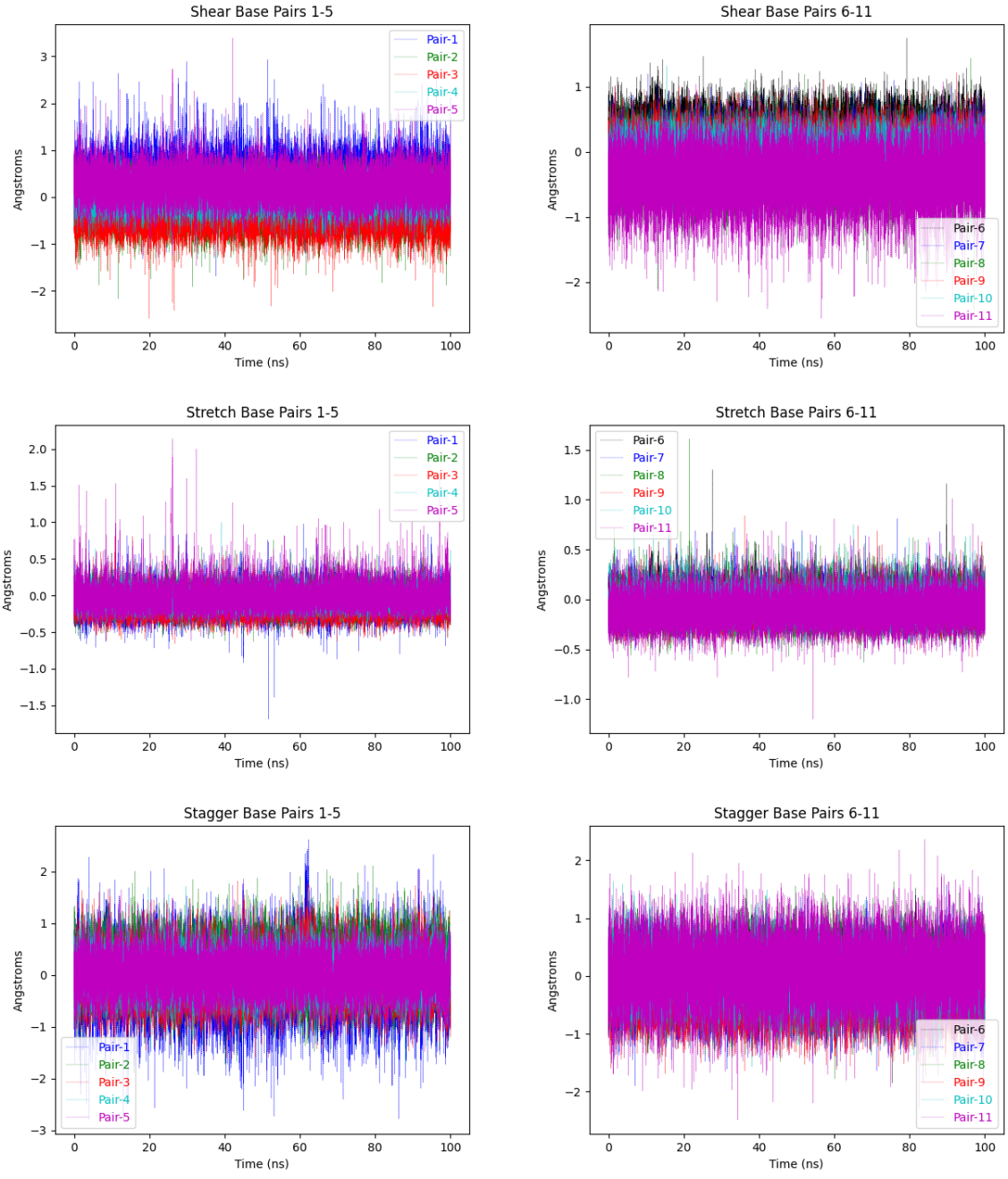
(6.73) DB[a,l]P-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



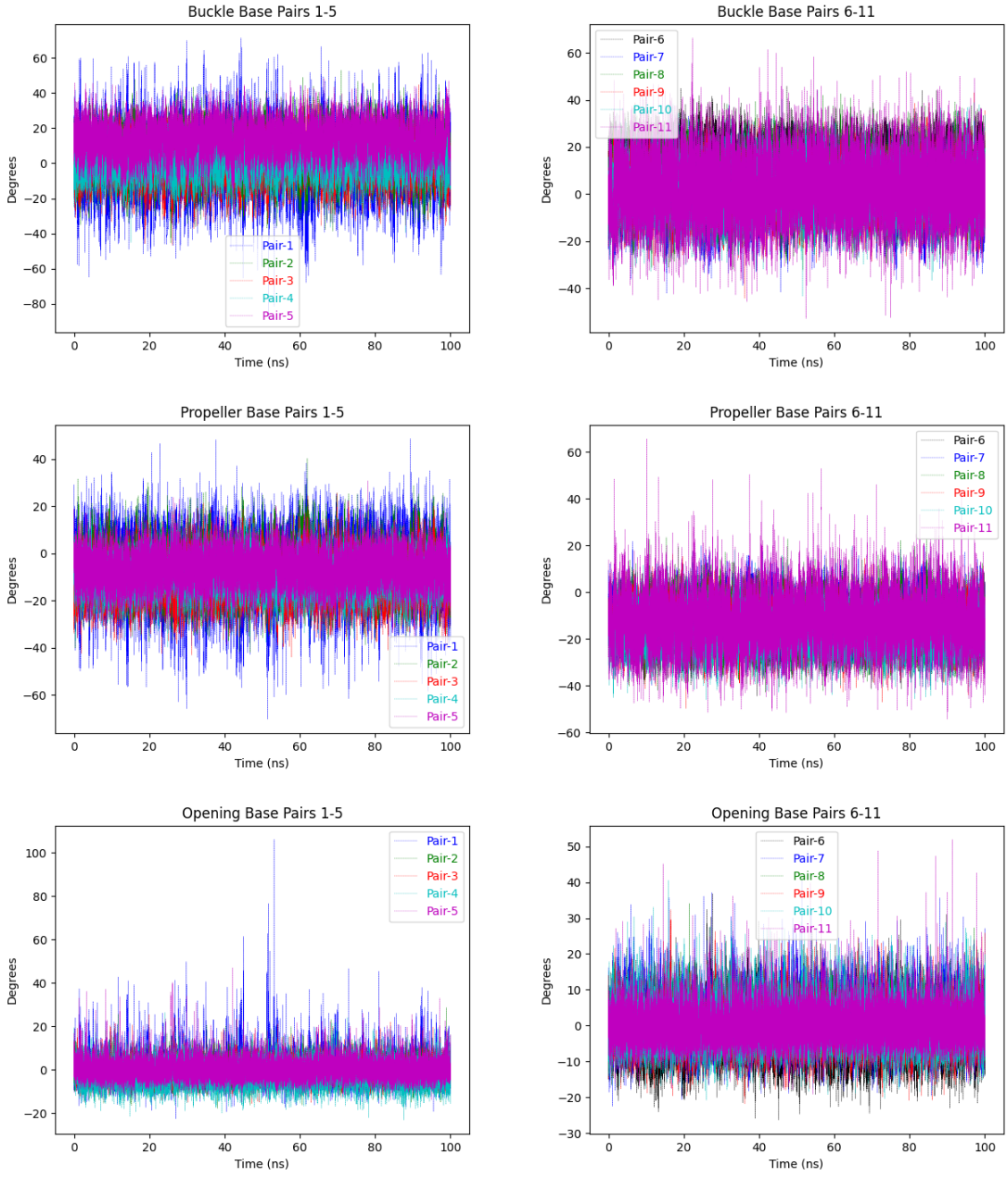
(6.74) DB[a,l]P-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



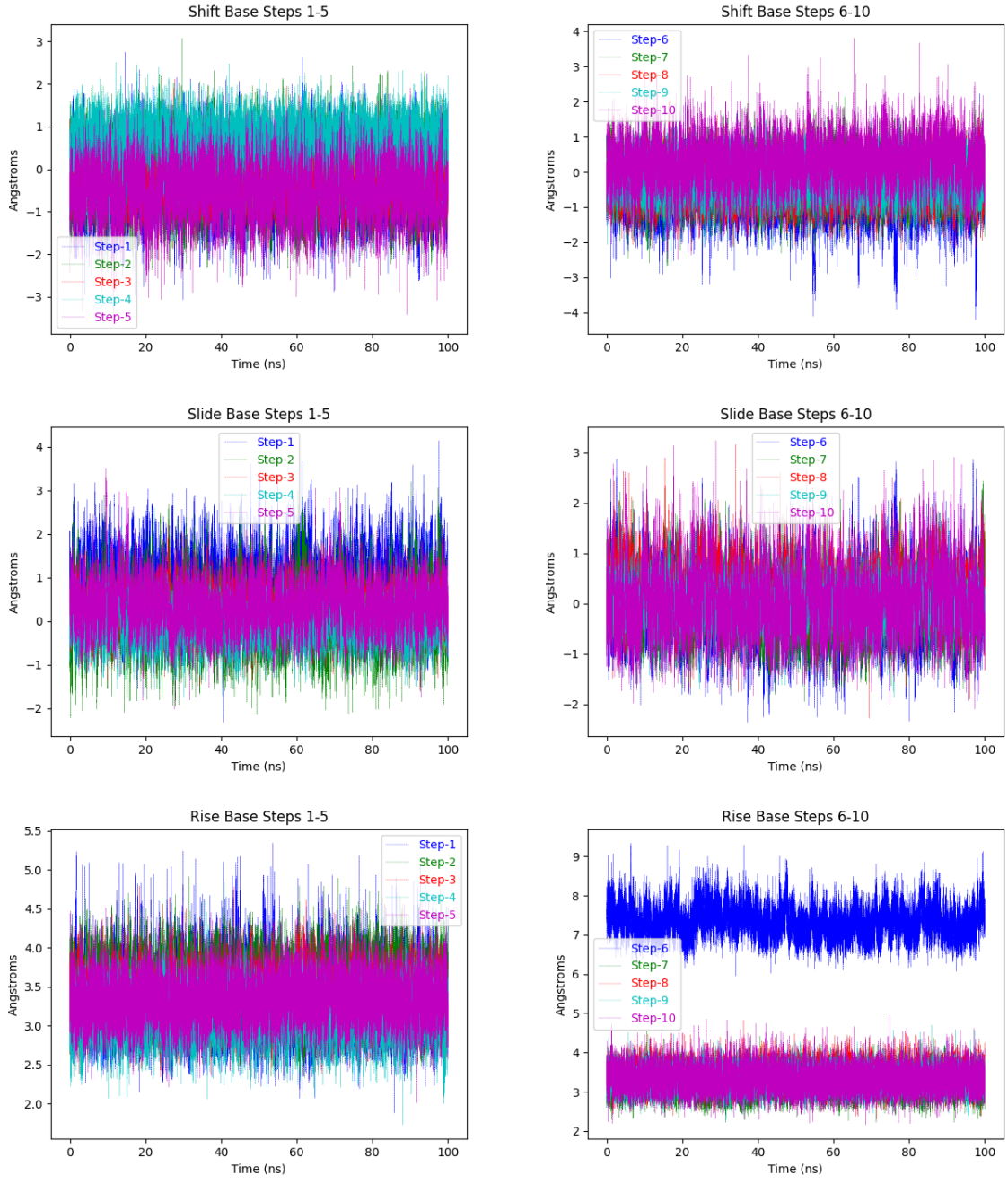
(6.75) DB[a,l]P-DNA: Refined major and minor groove trajectories



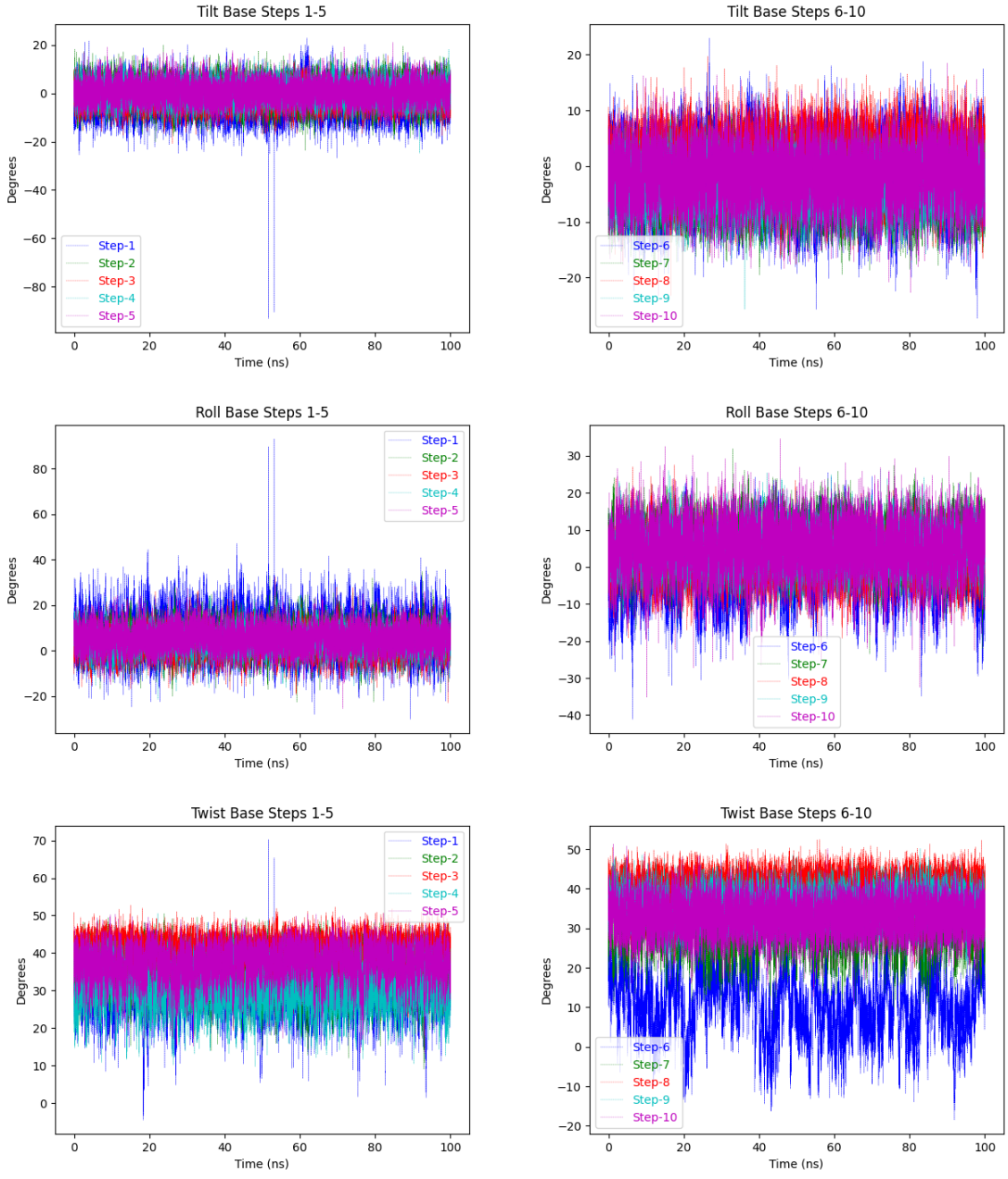
(6.76) DB[a,l]P-DNA: Base pair trajectories



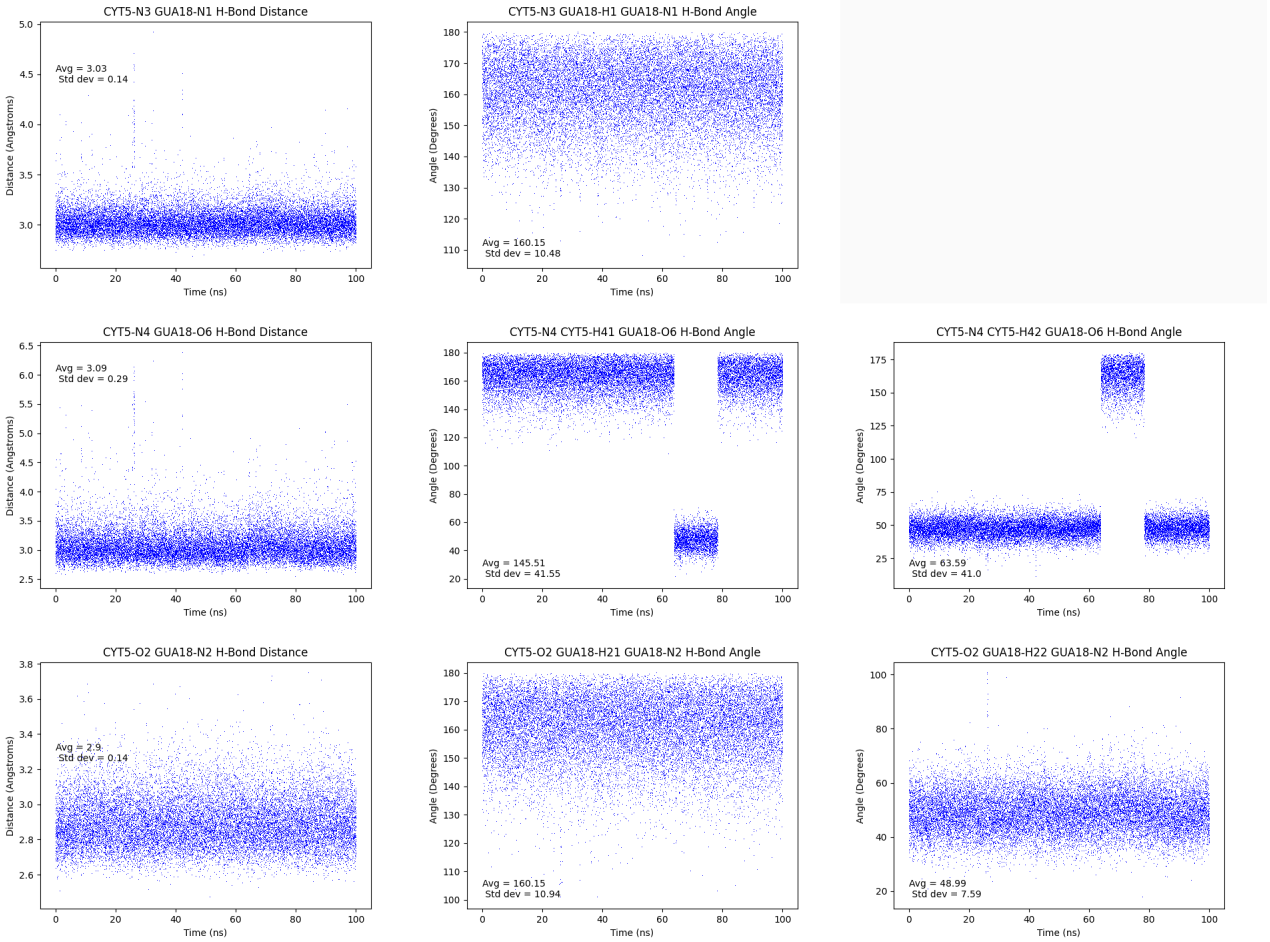
(6.77) DB[a,l]P-DNA: Base pair trajectories



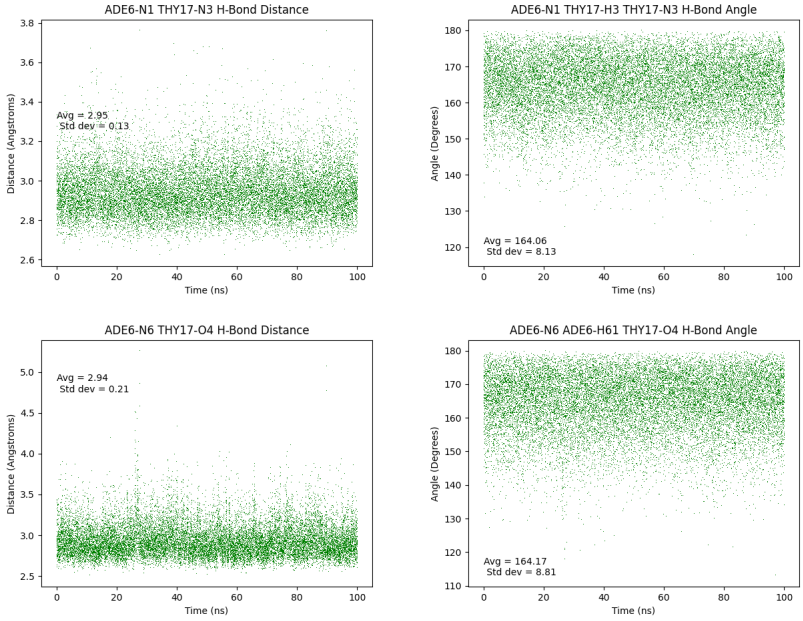
(6.78) DB[a,l]P-DNA: Base step trajectories



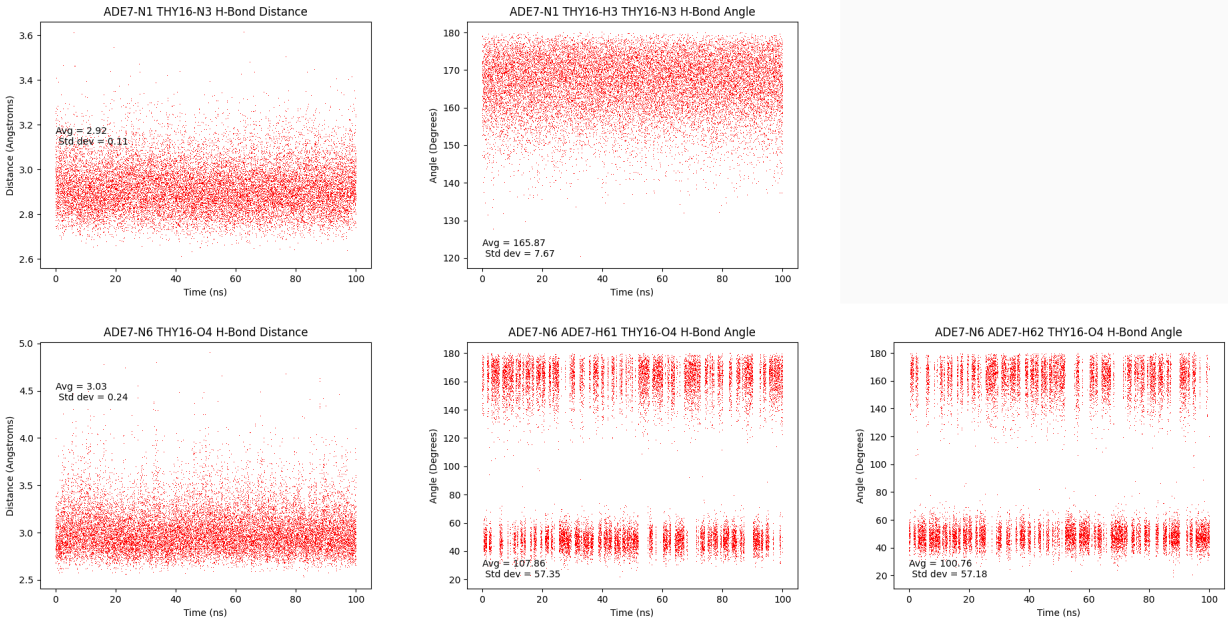
(6.79) DB[a,l]P-DNA: Base step trajectories



(6.80) DB[a,l]P-DNA: dC_5:dG_{18} hydrogen bond trajectories

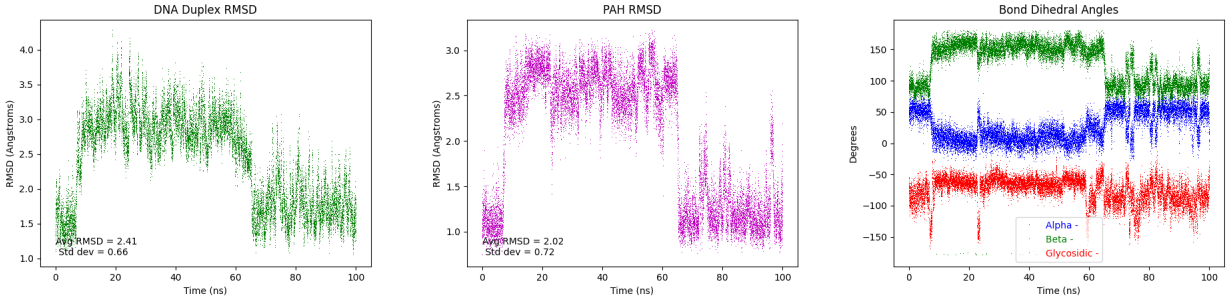


(6.81) DB[a,l]P-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

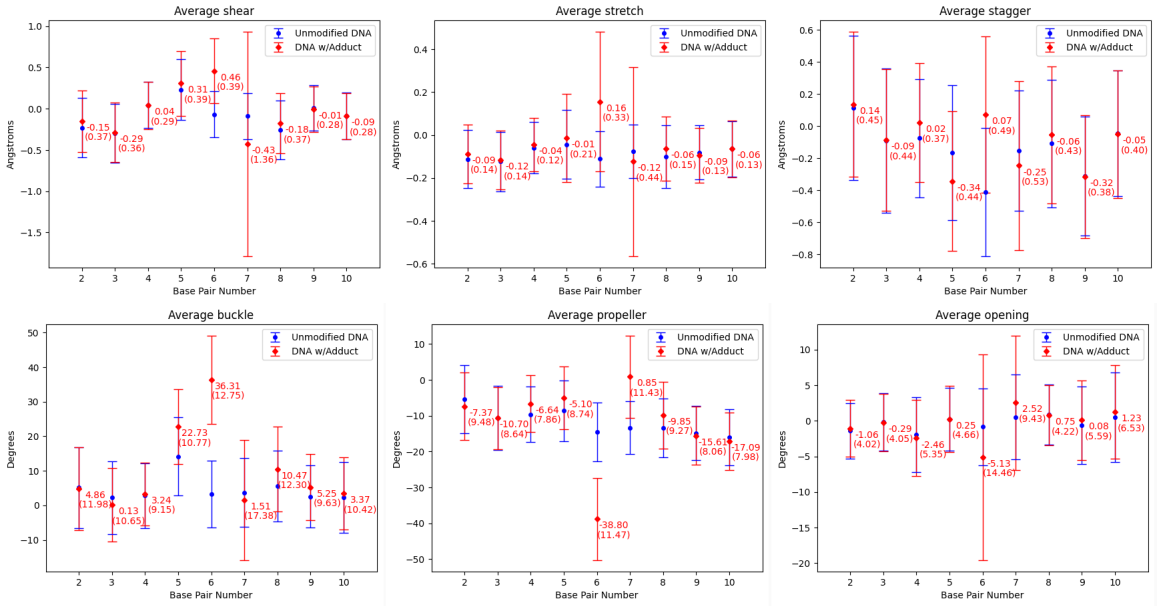


(6.82) DB[a,l]P-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

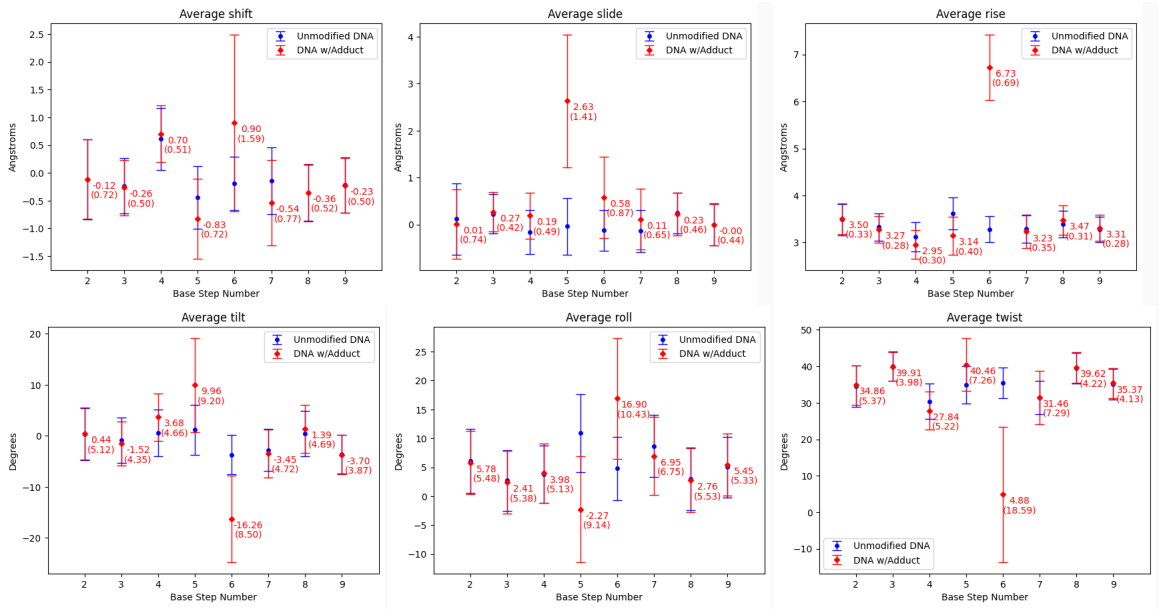
6.2.1.4 CHR-DNA



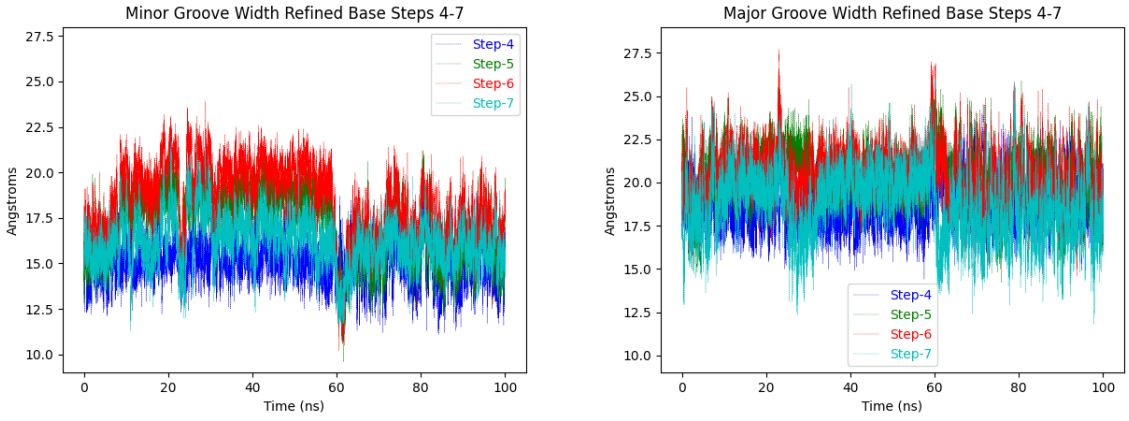
(6.83) CHR-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



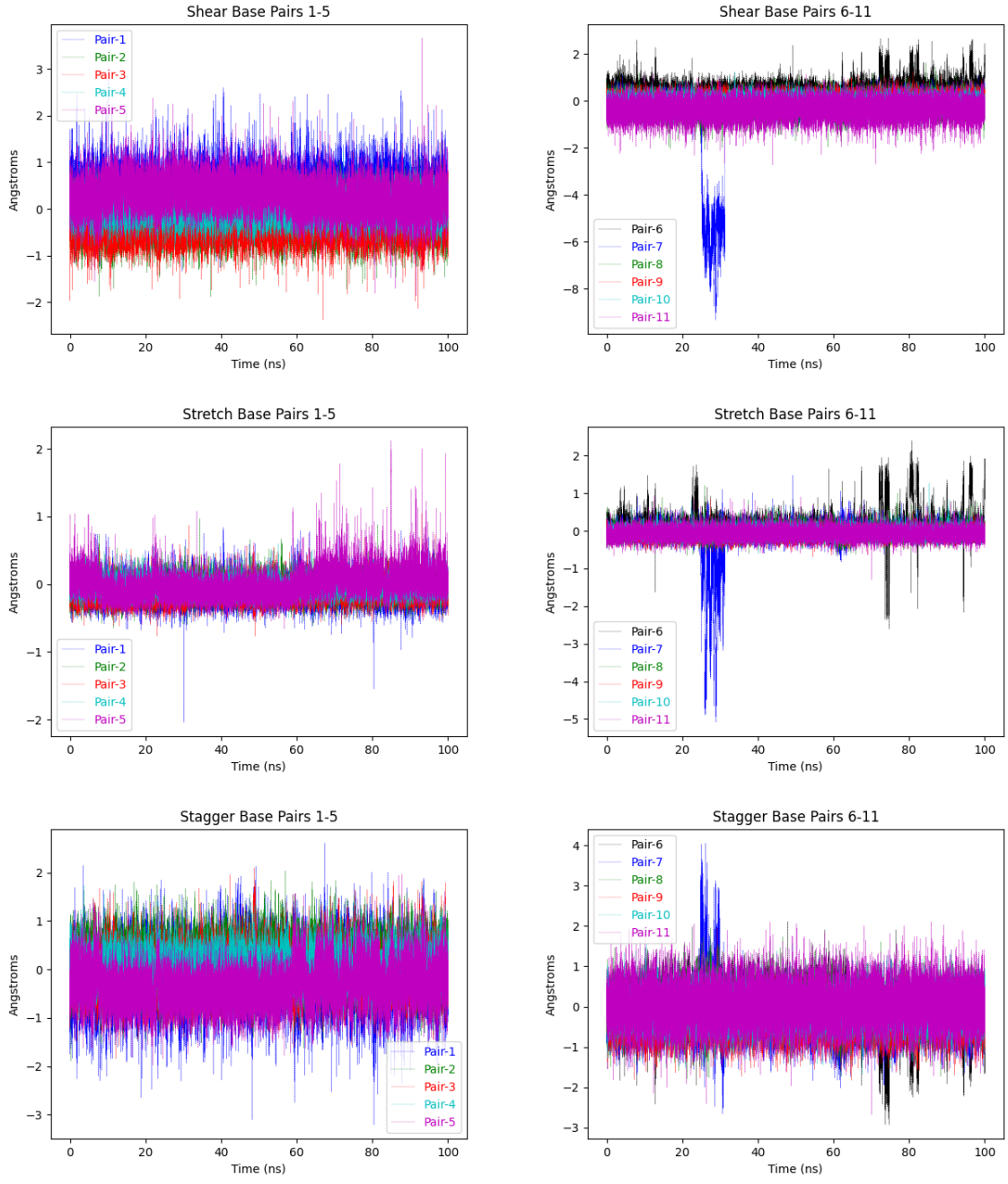
(6.84) CHR-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



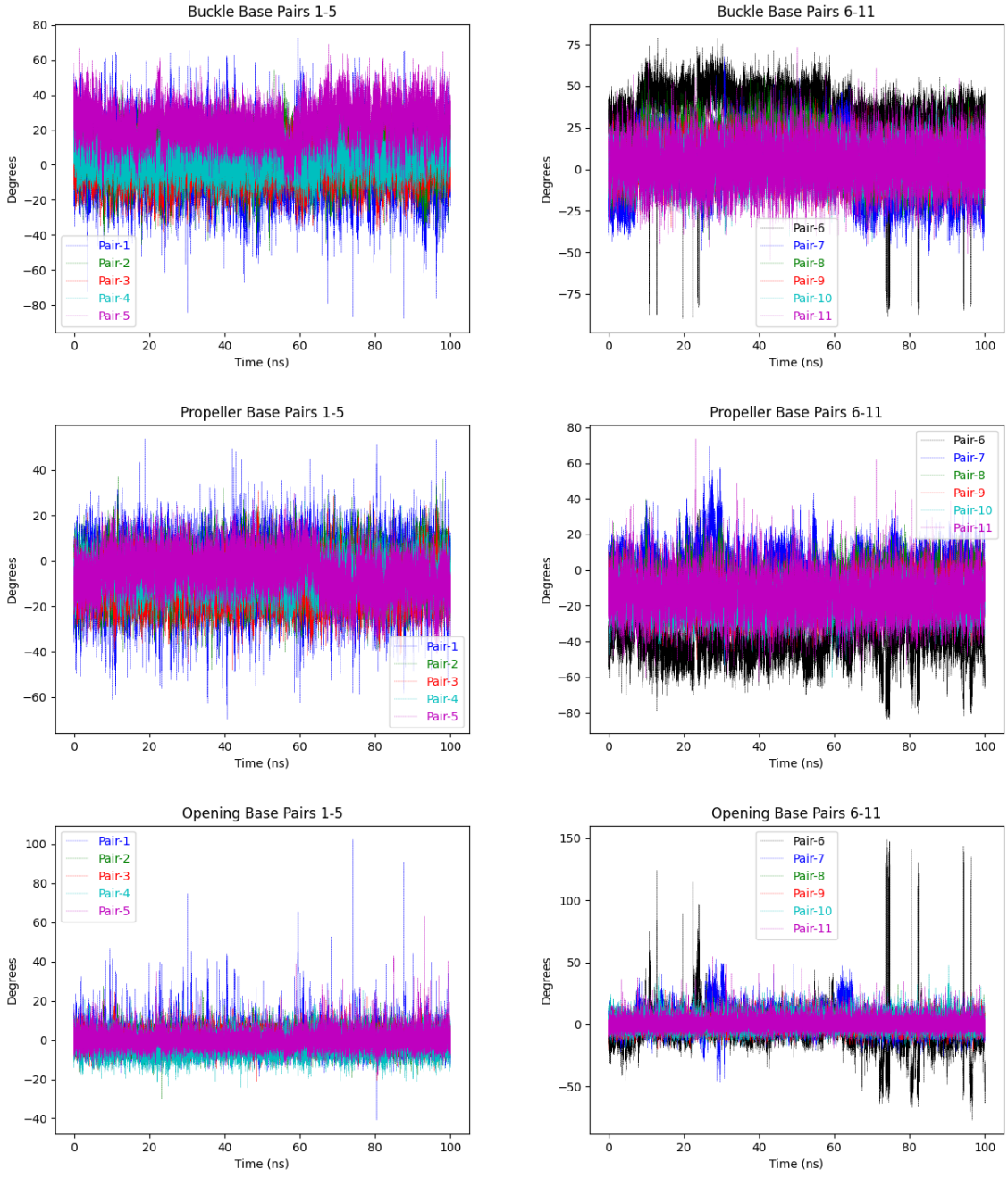
(6.85) CHR-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



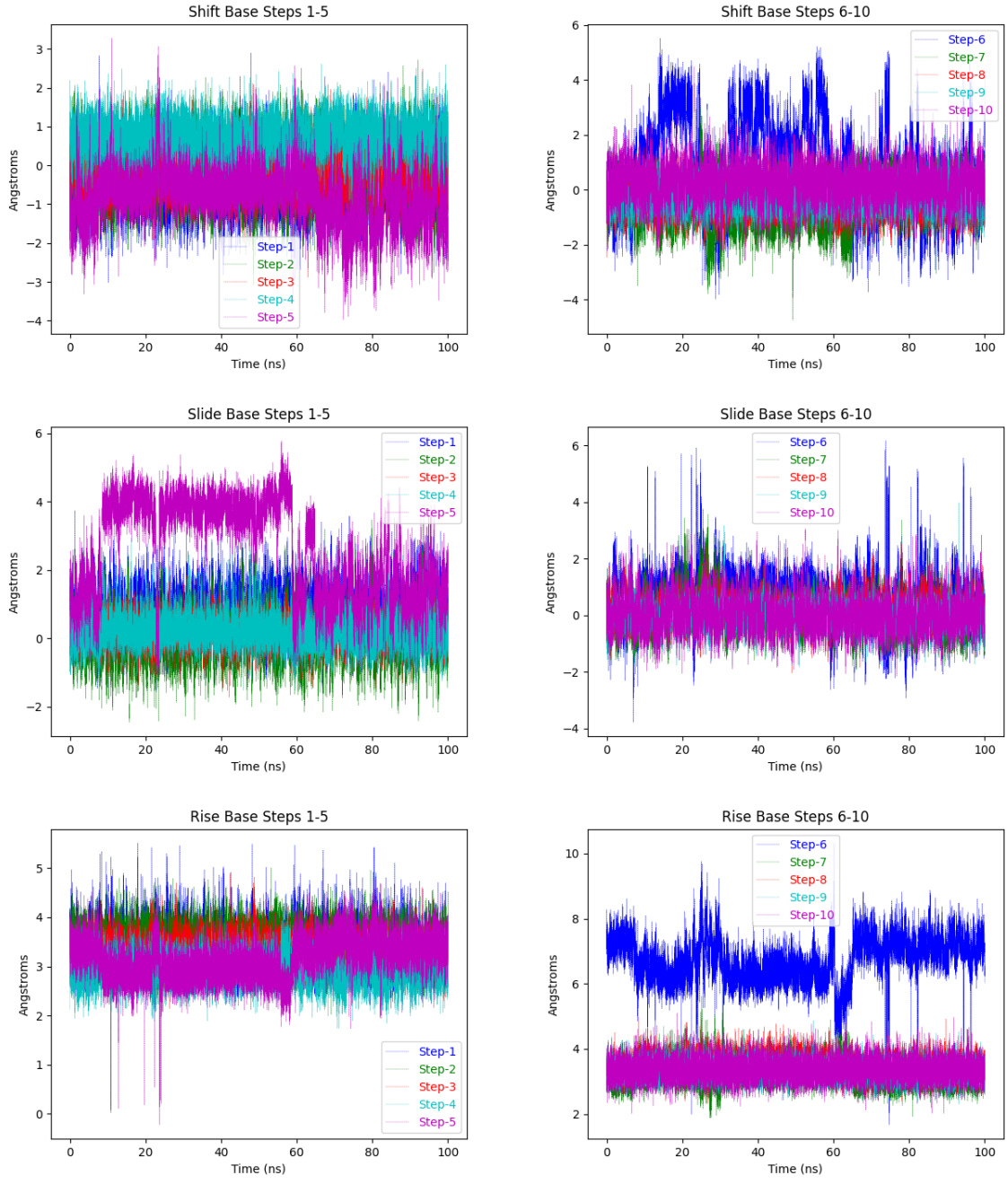
(6.86) CHR-DNA: Refined major and minor groove trajectories



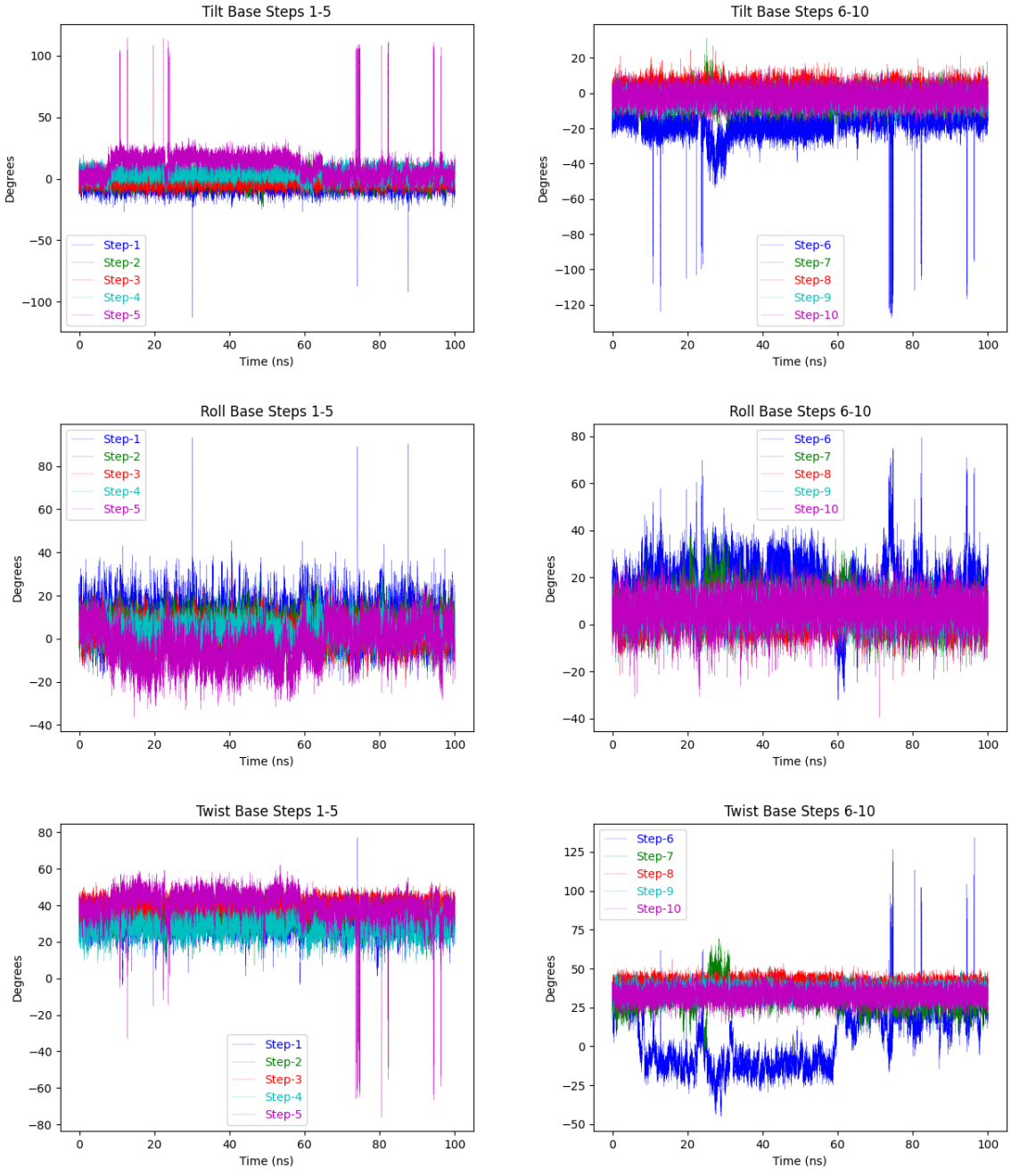
(6.87) CHR-DNA: Base pair trajectories



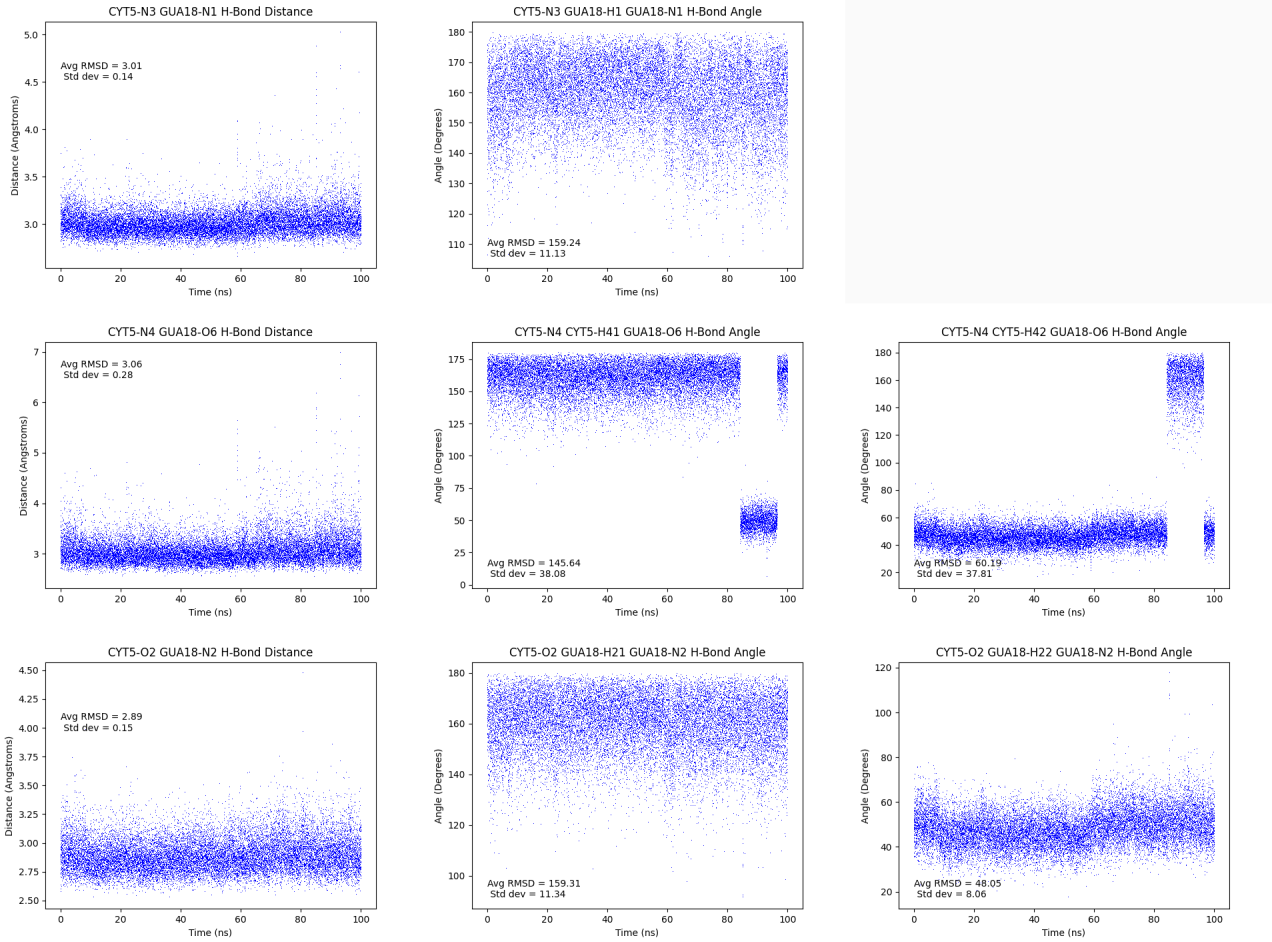
(6.88) CHR-DNA: Base pair trajectories



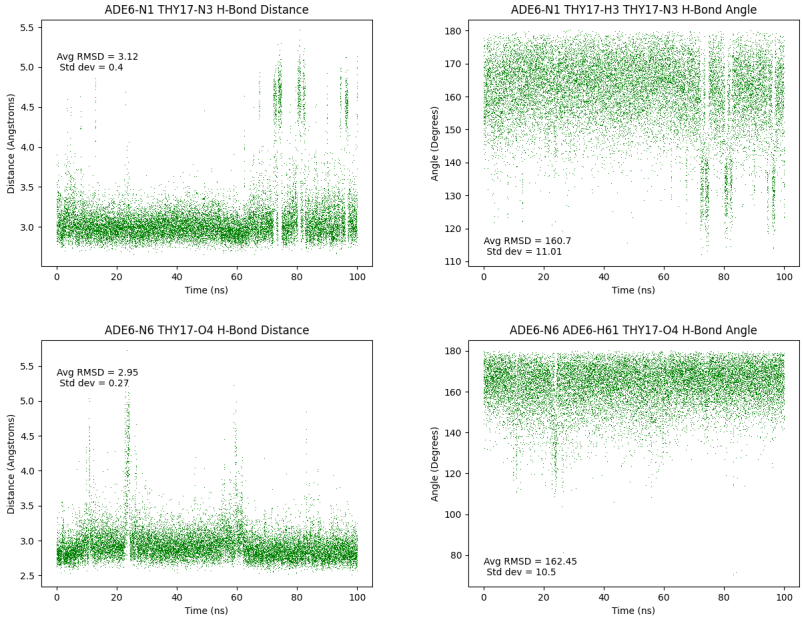
(6.89) CHR-DNA: Base step trajectories



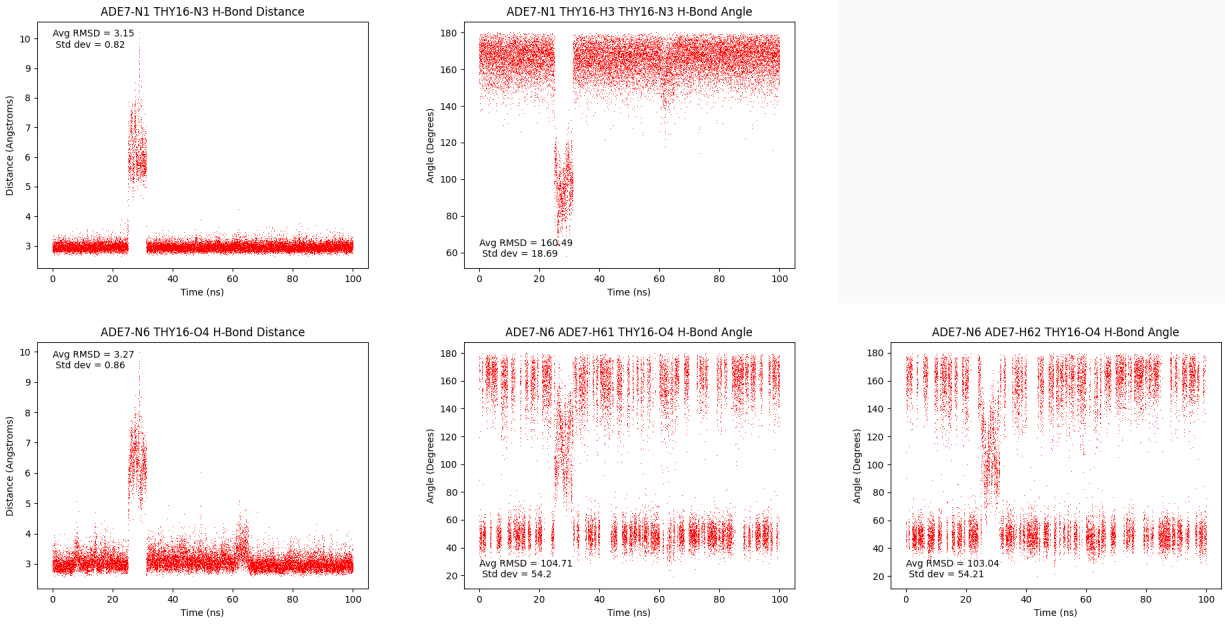
(6.90) CHR-DNA: Base step trajectories



(6.91) CHR-DNA: dC_5:dG_{18} hydrogen bond trajectories

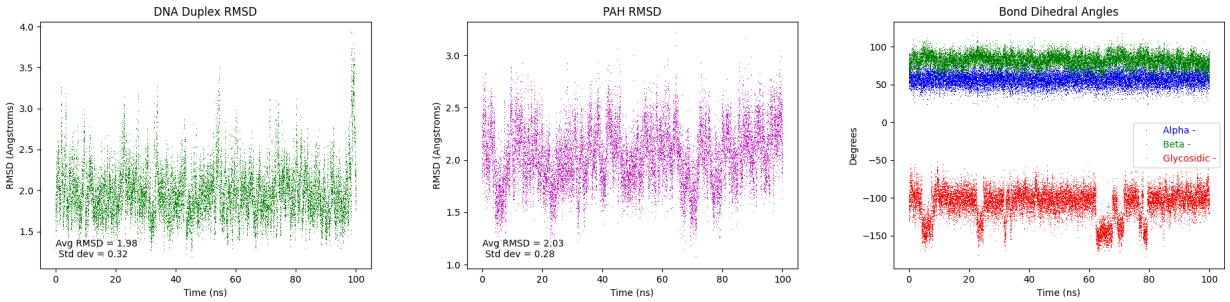


(6.92) CHR-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

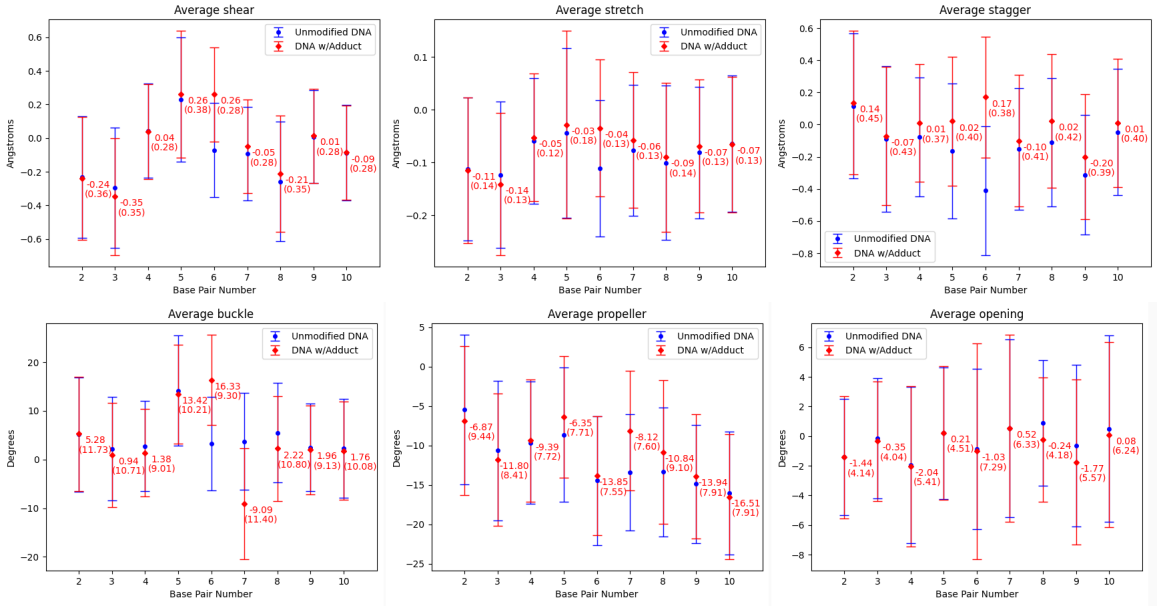


(6.93) CHR-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

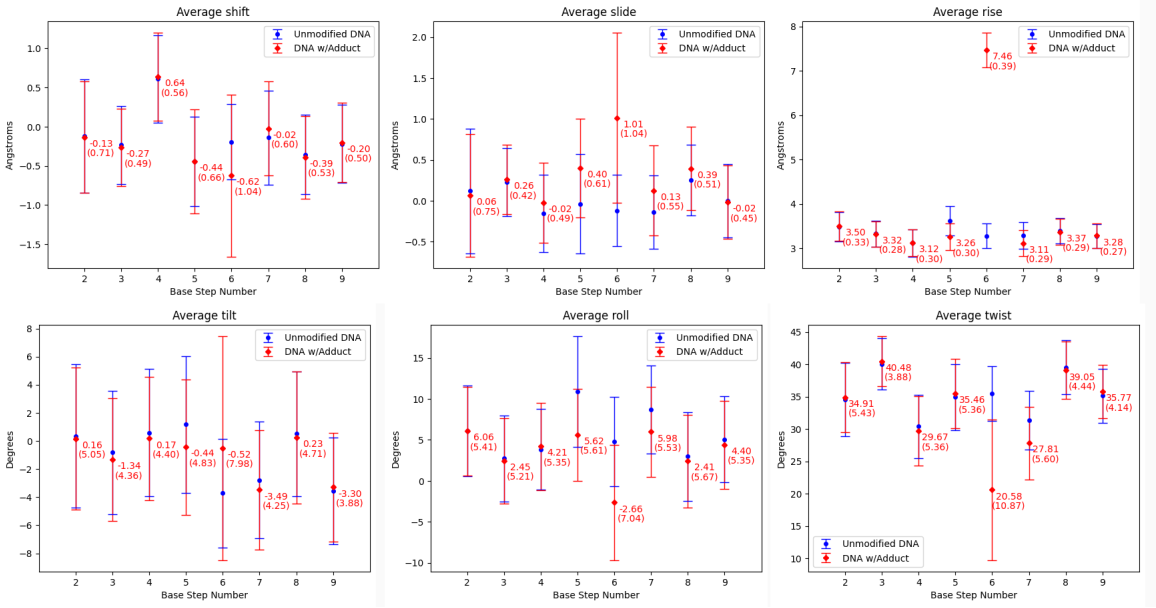
6.2.1.5 B[g]C-DNA



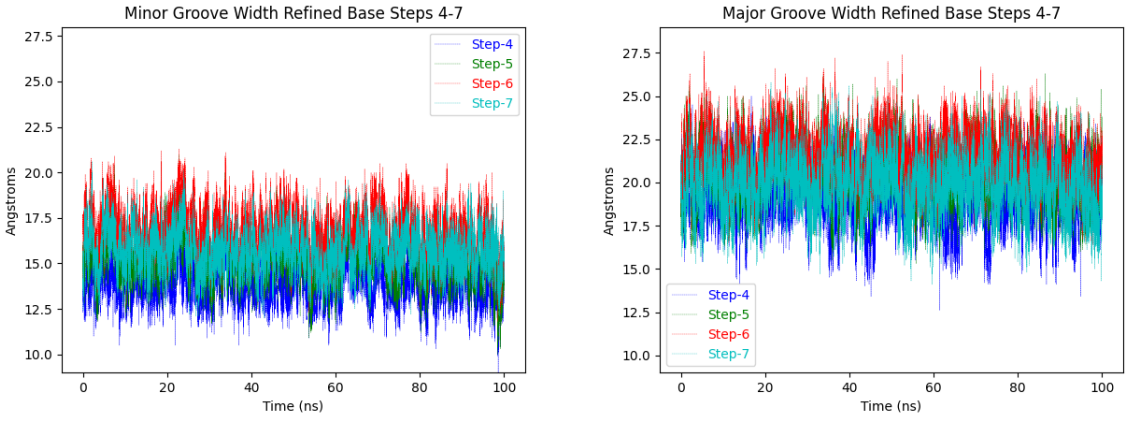
(6.94) B[g]C-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



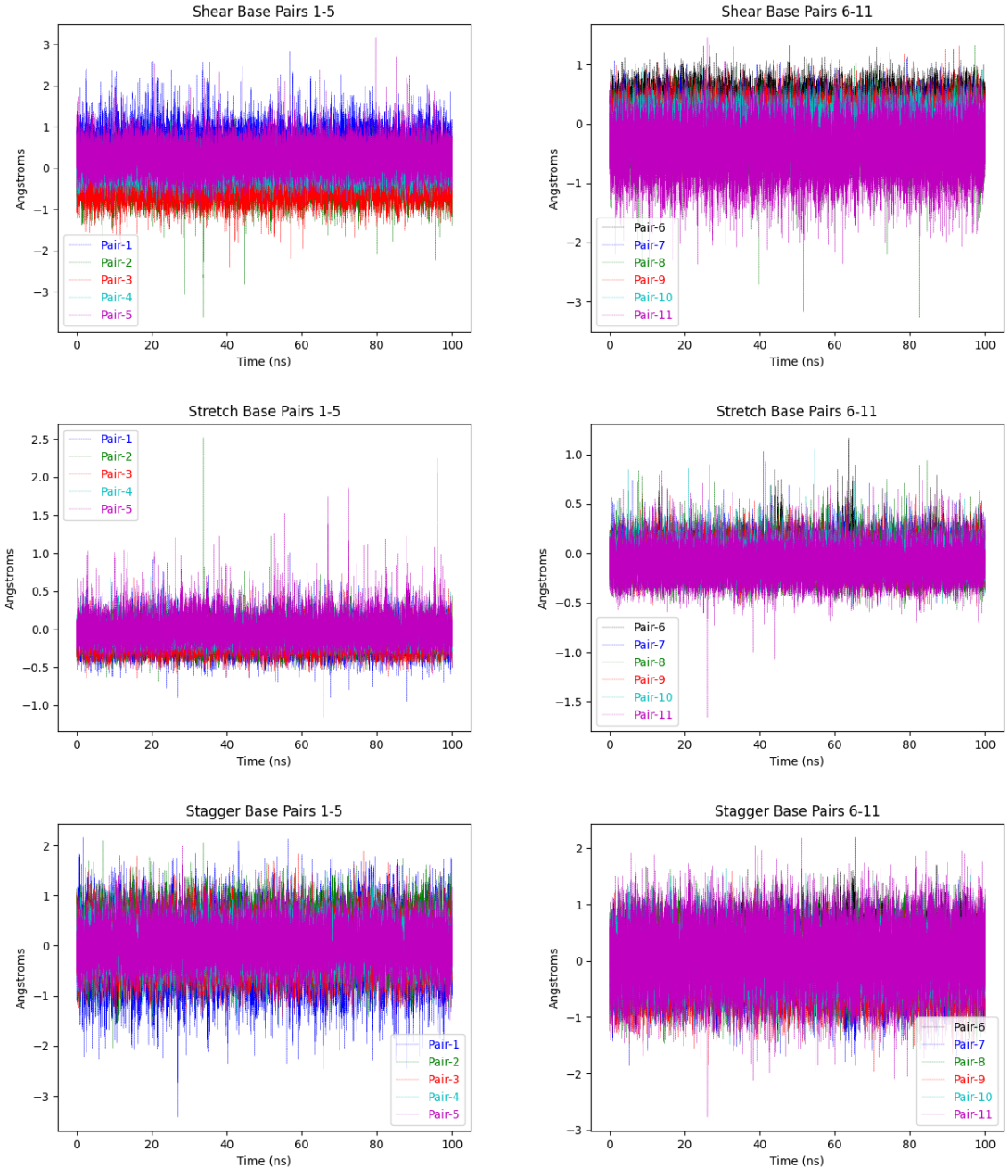
(6.95) B[g]C-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



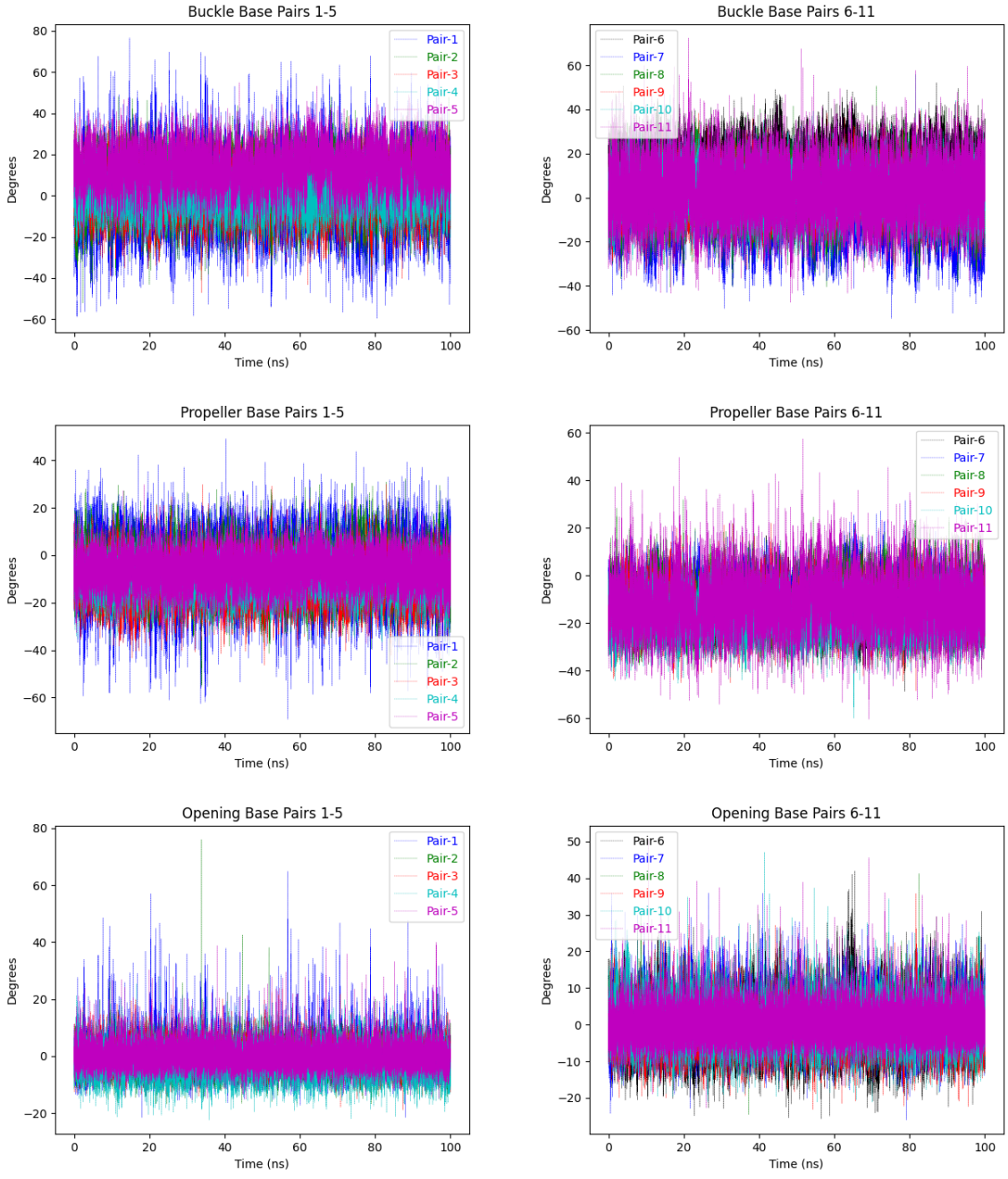
(6.96) B[g]C-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



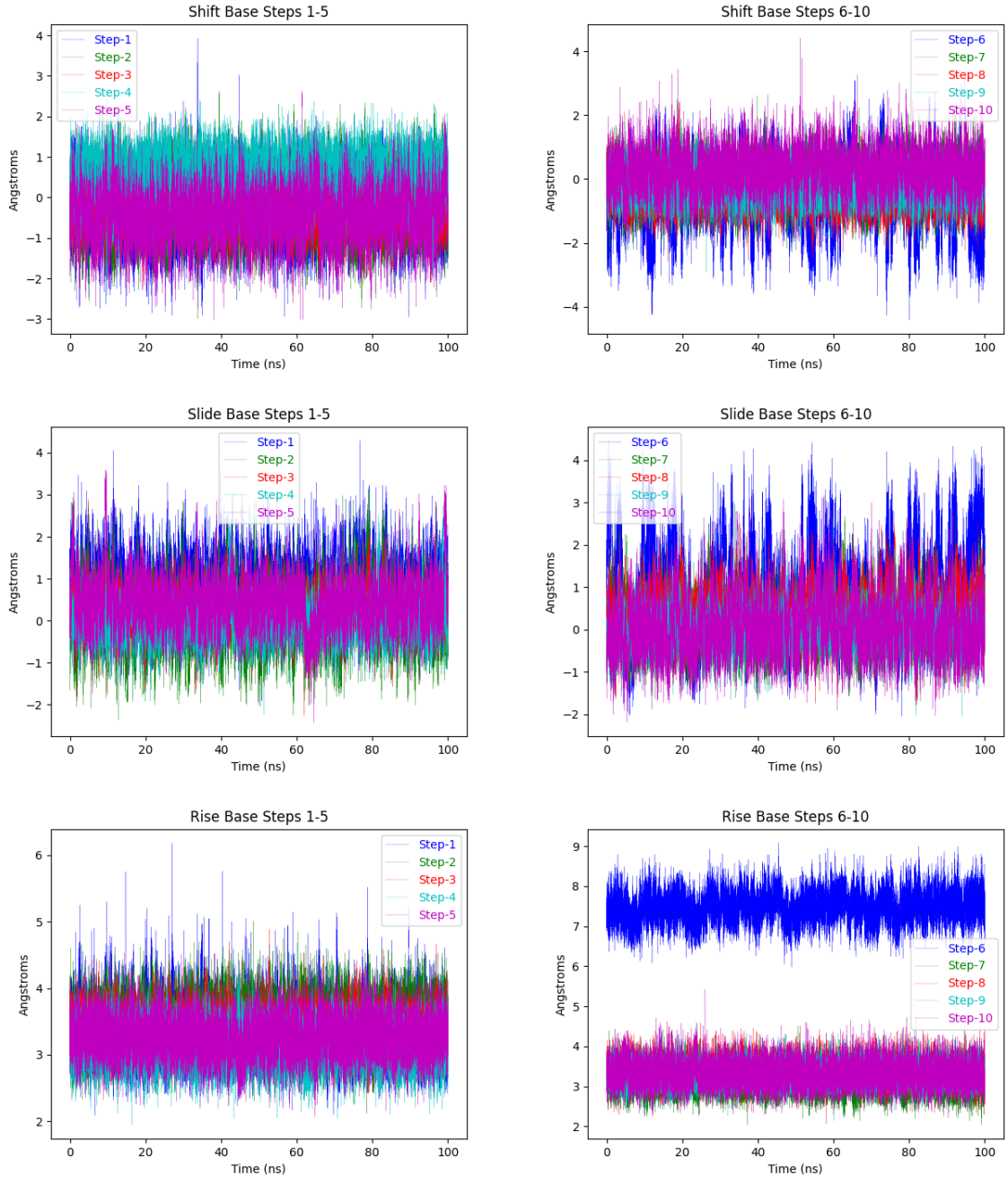
(6.97) B[g]C-DNA: Refined major and minor groove trajectories



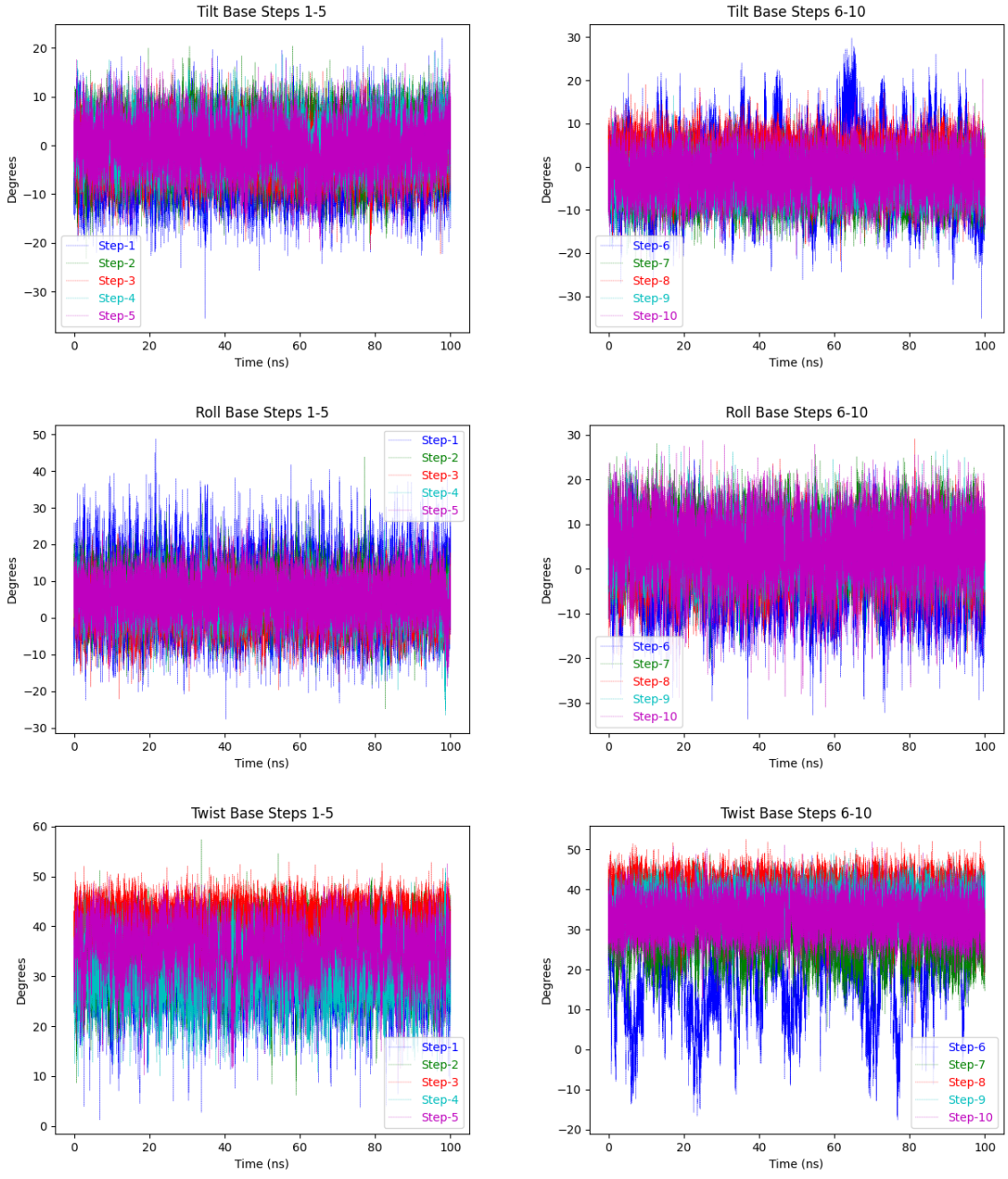
(6.98) B[g]C-DNA: Base pair trajectories



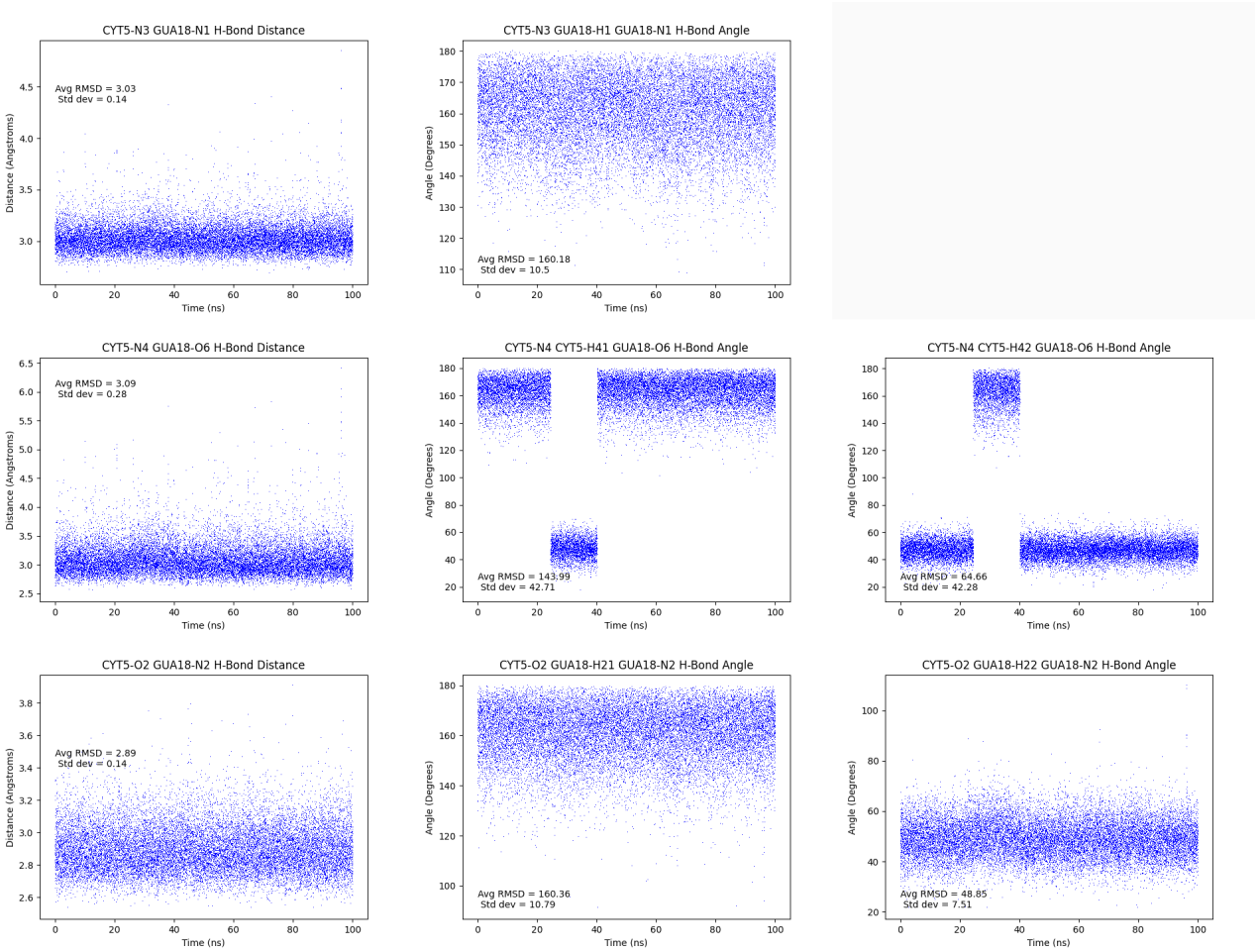
(6.99) B[g]C-DNA: Base pair trajectories



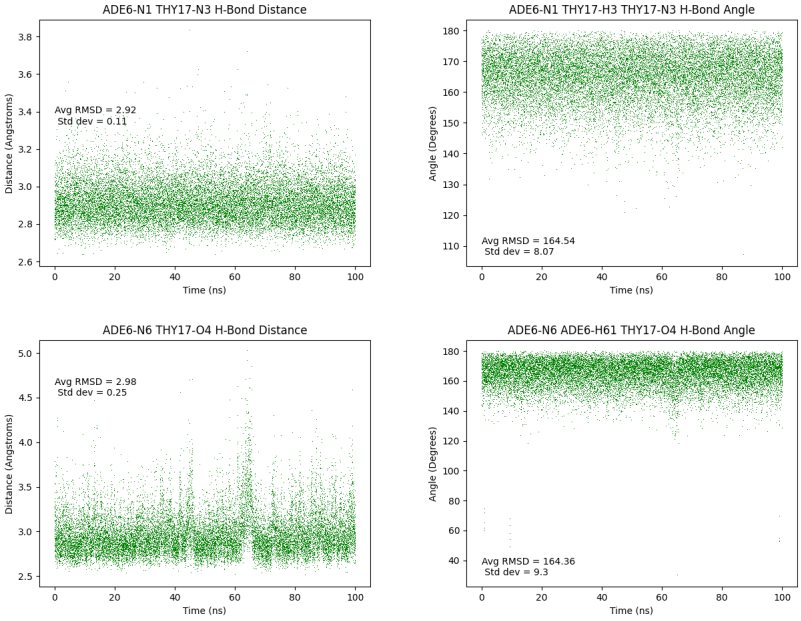
(6.100) B[g]C-DNA: Base step trajectories



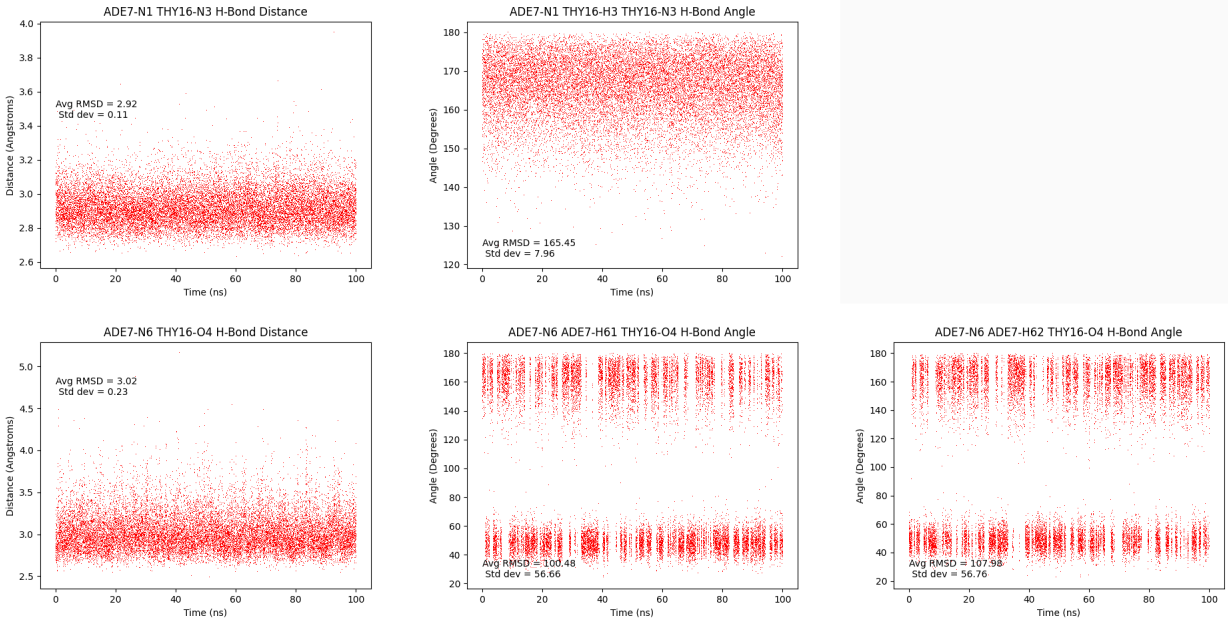
(6.101) B[g]C-DNA: Base step trajectories



(6.102) B[g]C-DNA: dC_5:dG_{18} hydrogen bond trajectories

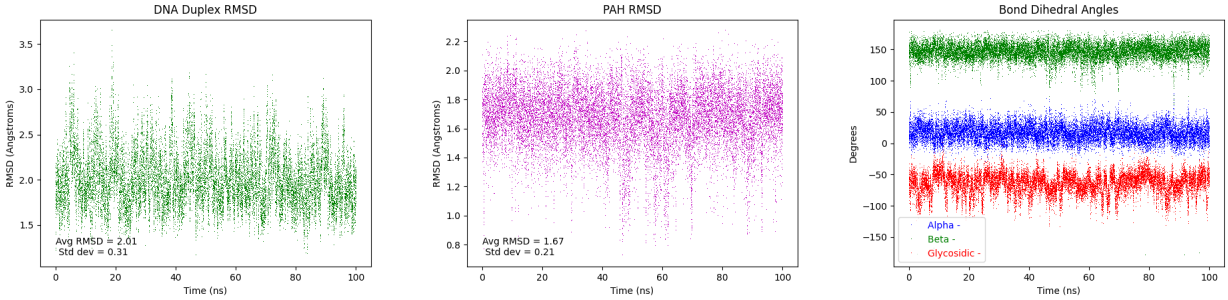


(6.103) B[g]C-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

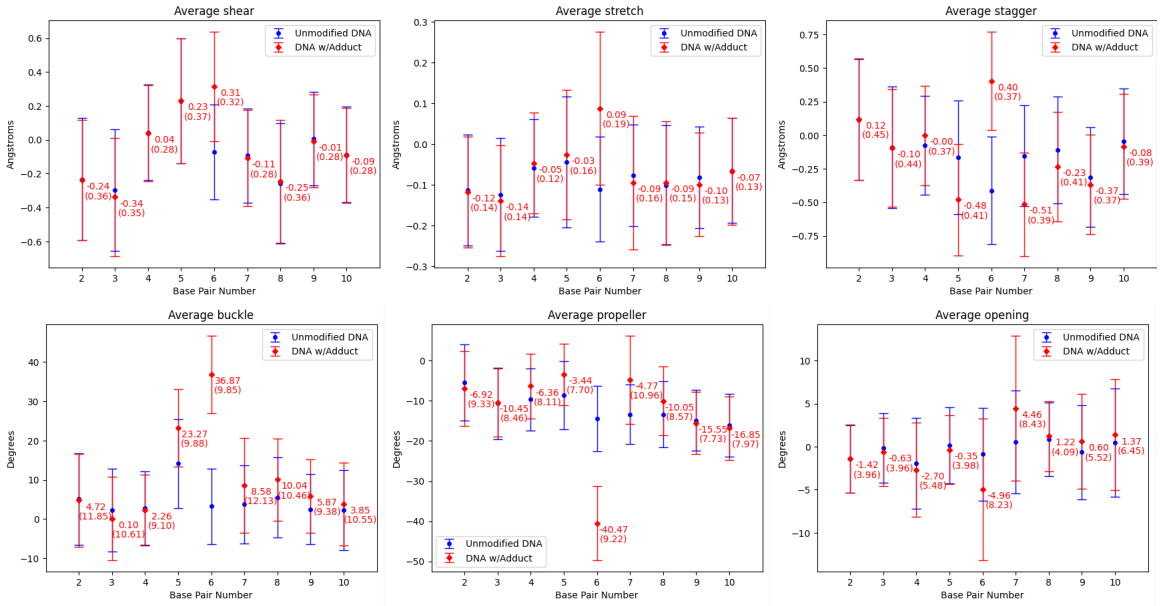


(6.104) B[g]C-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

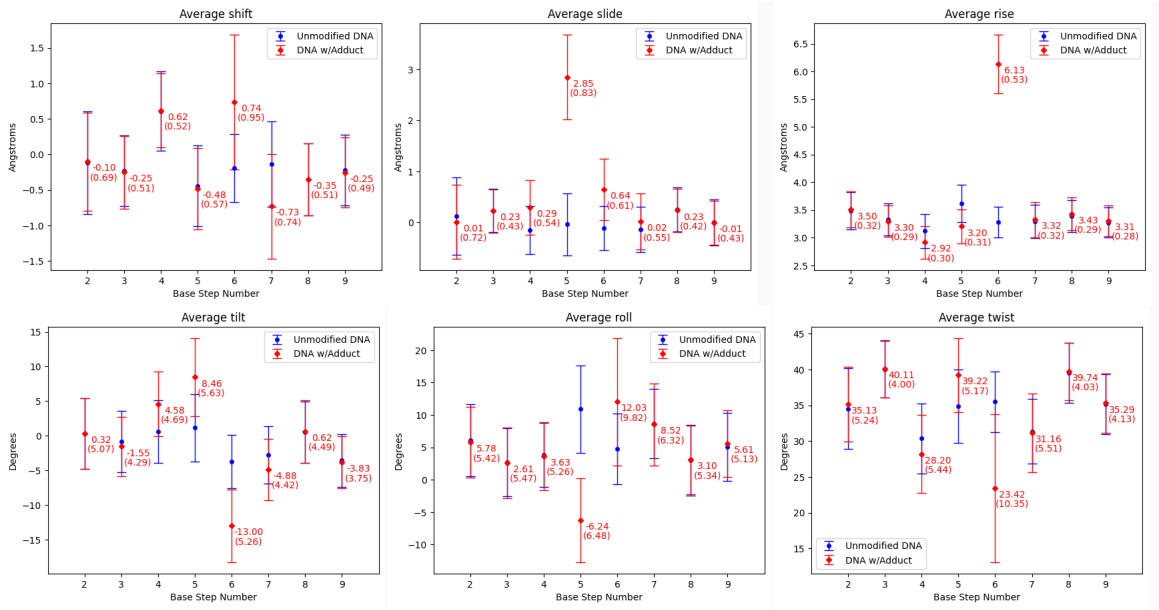
6.2.1.6 PHE-DNA



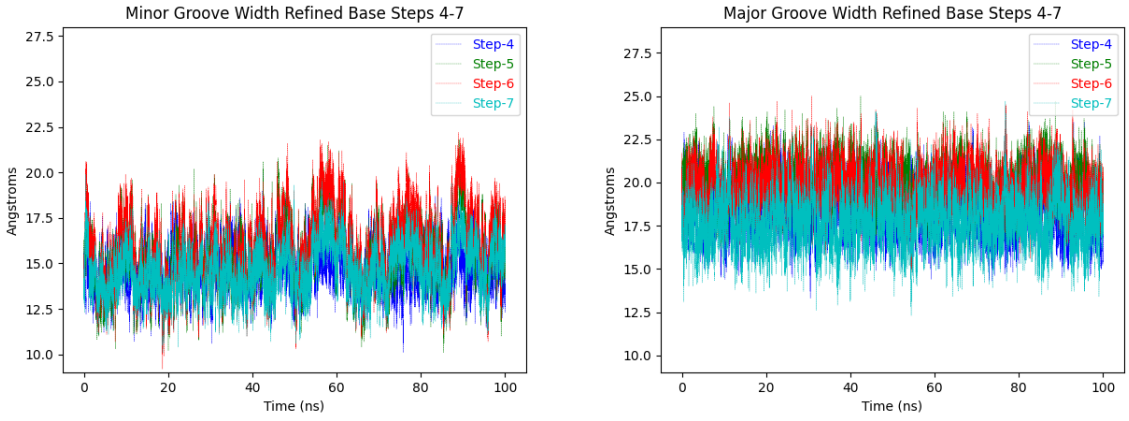
(6.105) PHE-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



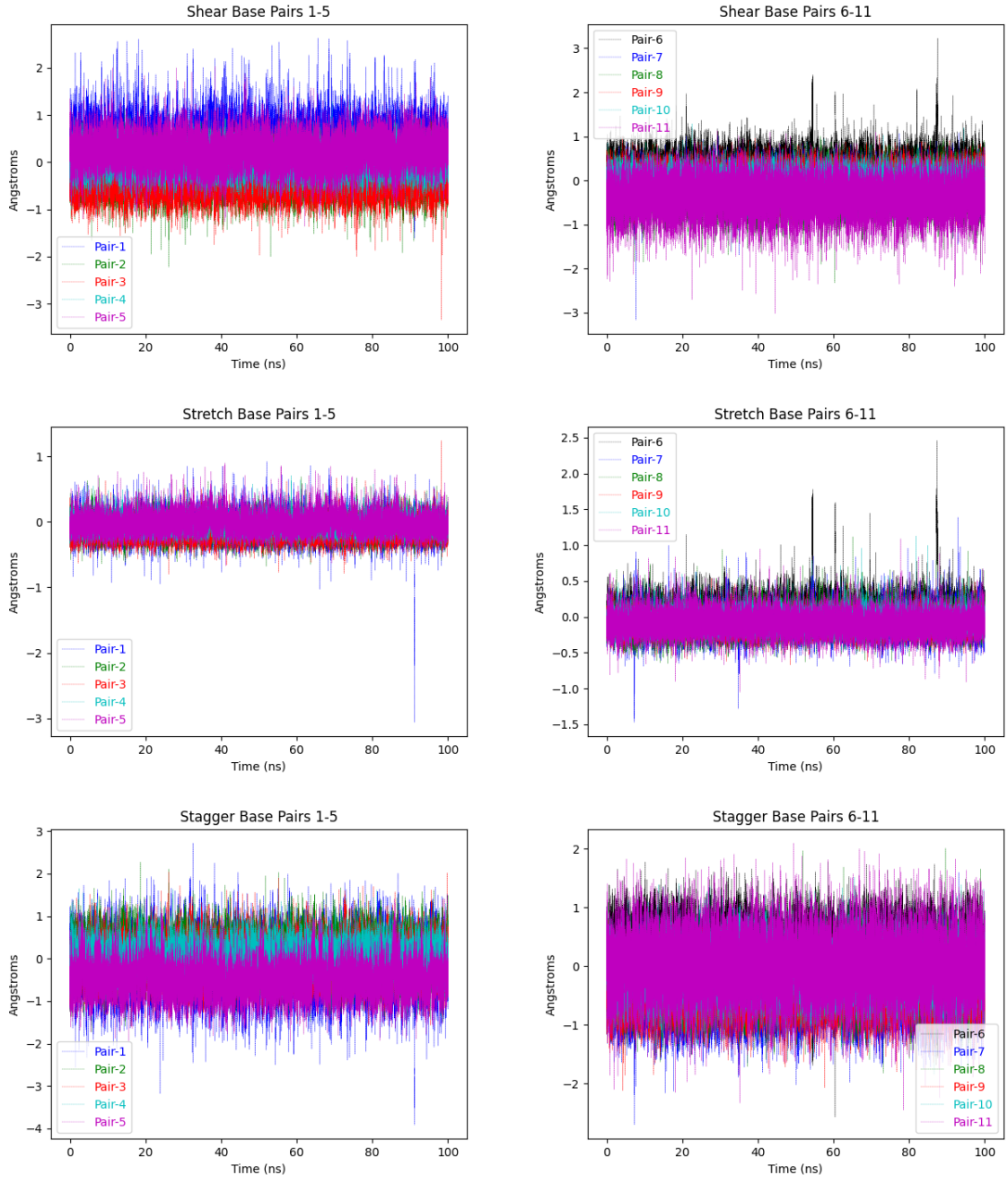
(6.106) PHE-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



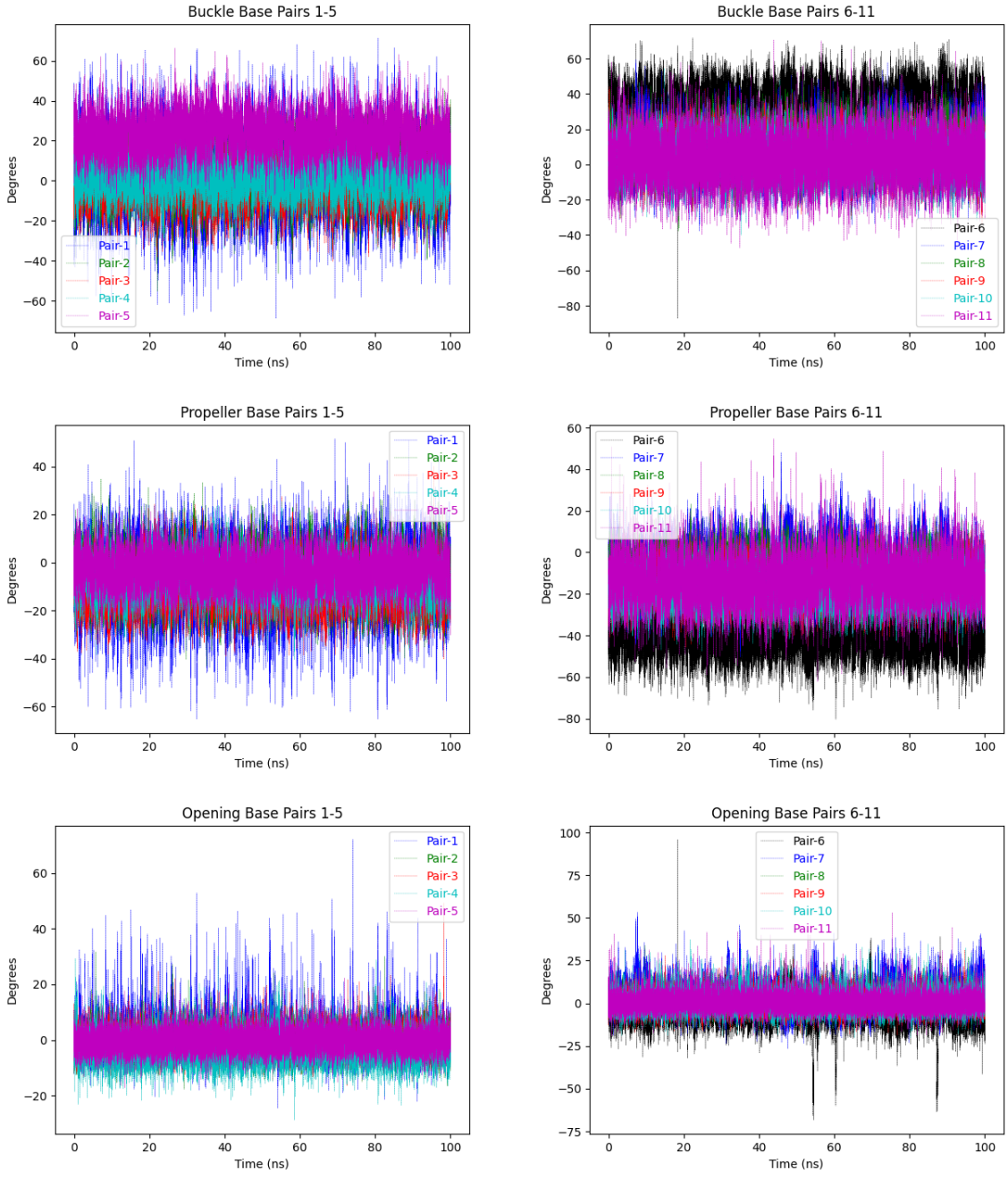
(6.107) PHE-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



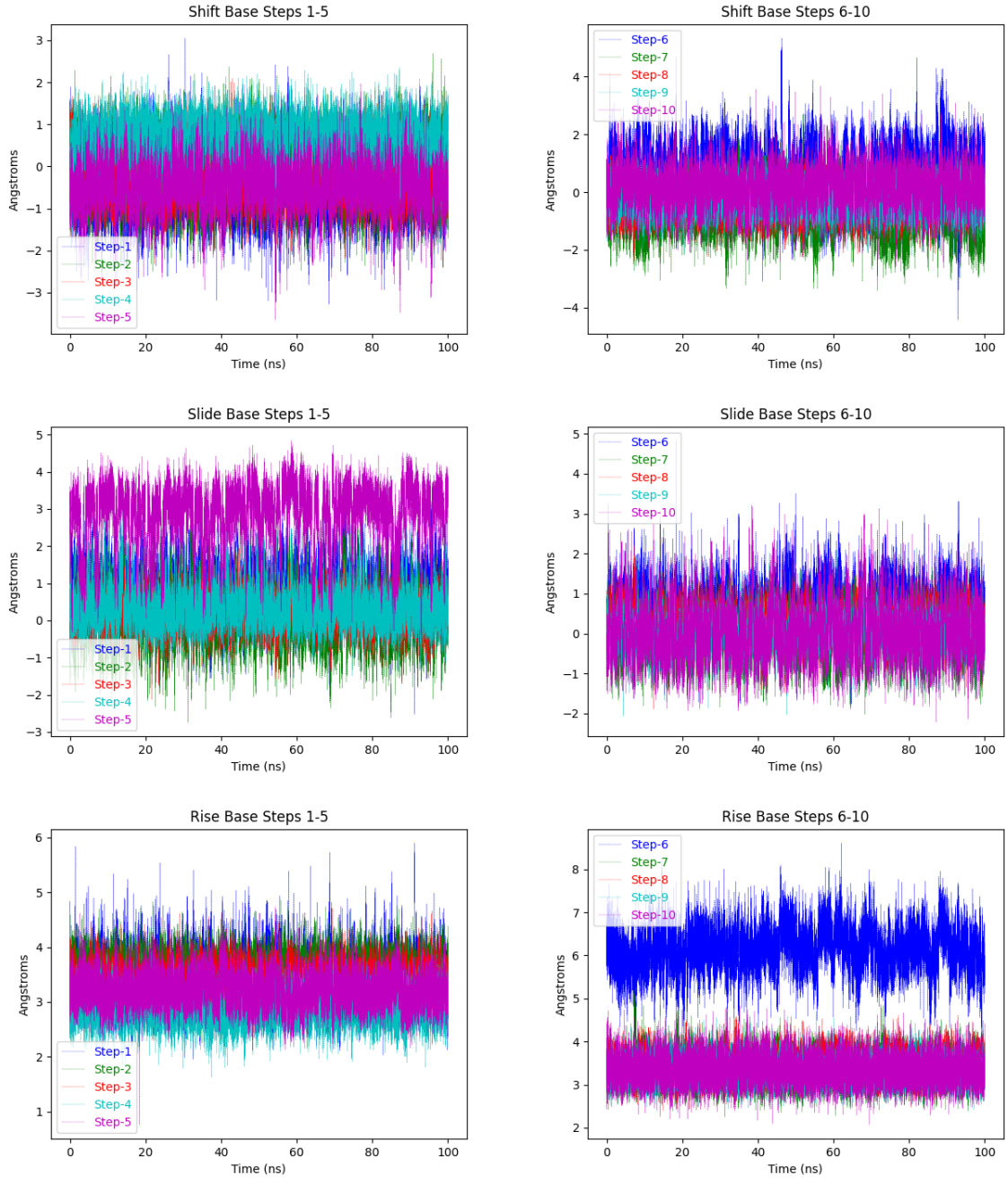
(6.108) PHE-DNA: Refined major and minor groove trajectories



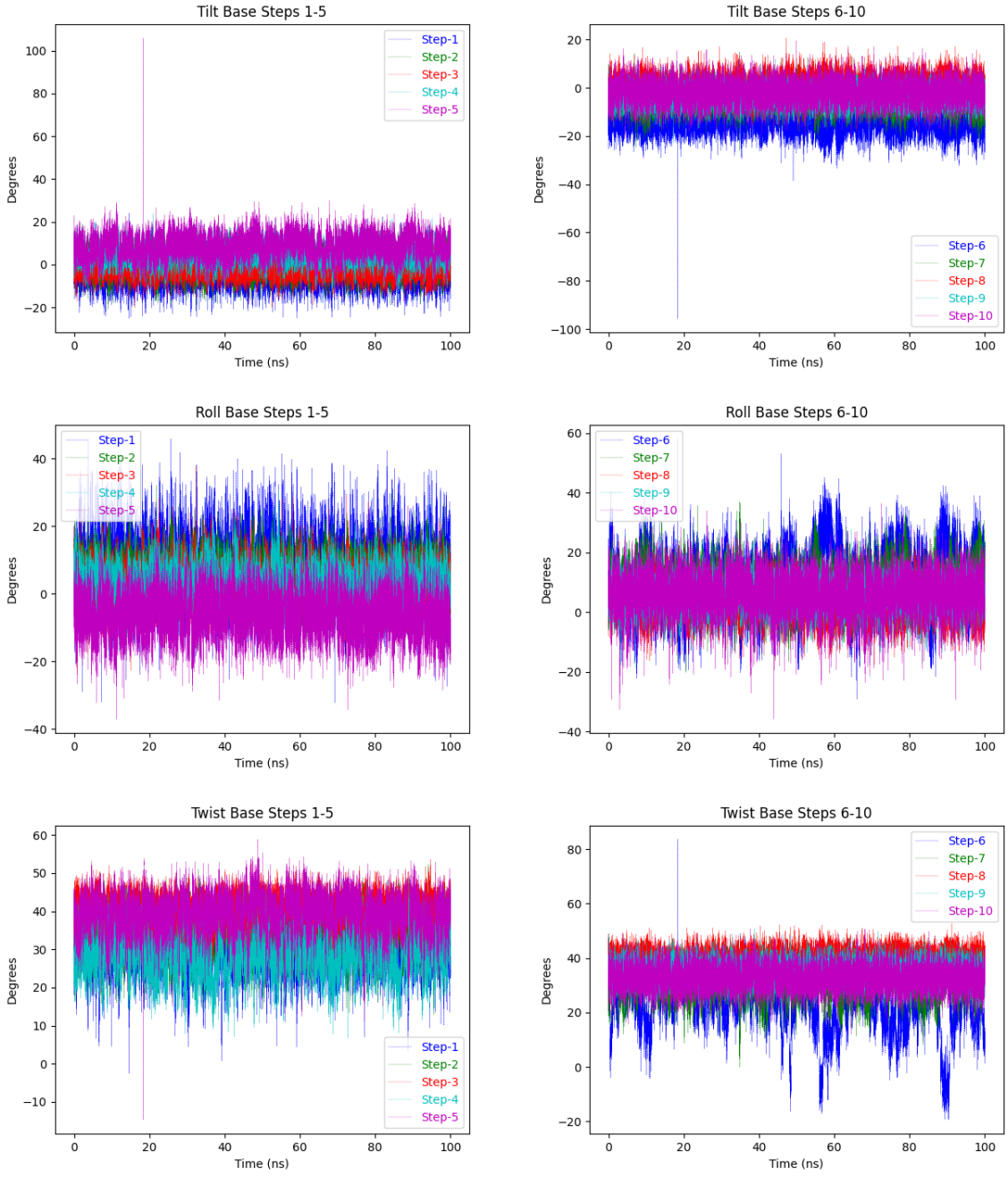
(6.109) PHE-DNA: Base pair trajectories



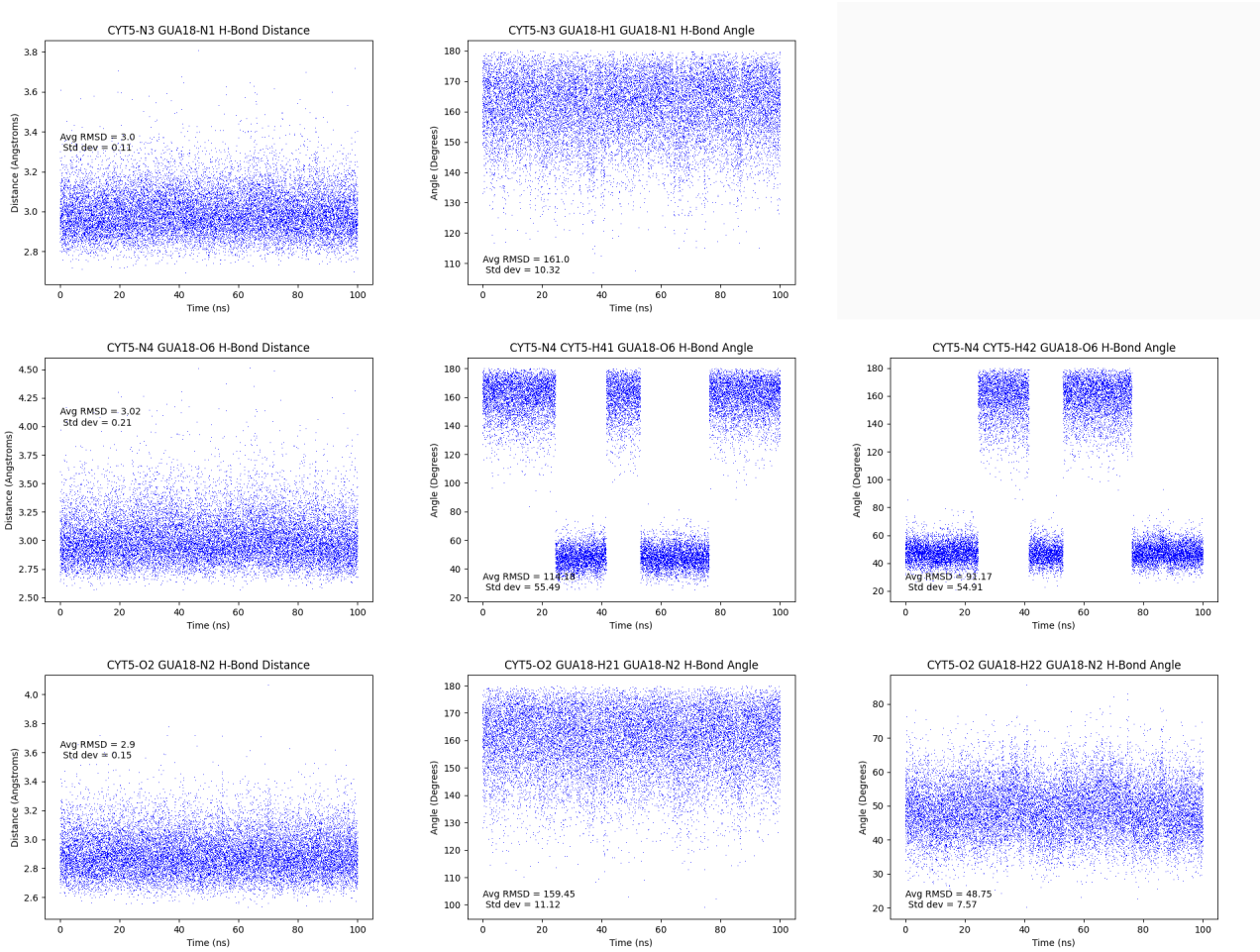
(6.110) PHE-DNA: Base pair trajectories



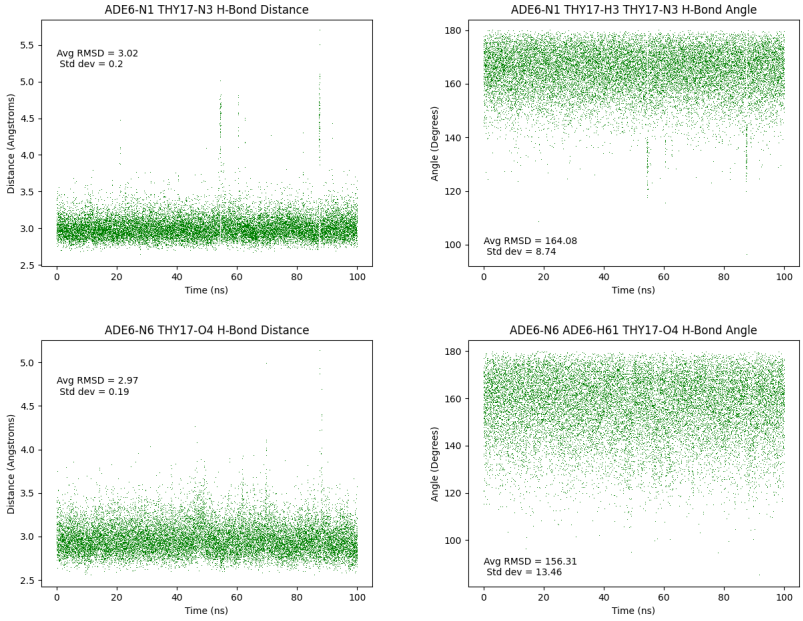
(6.111) PHE-DNA: Base step trajectories



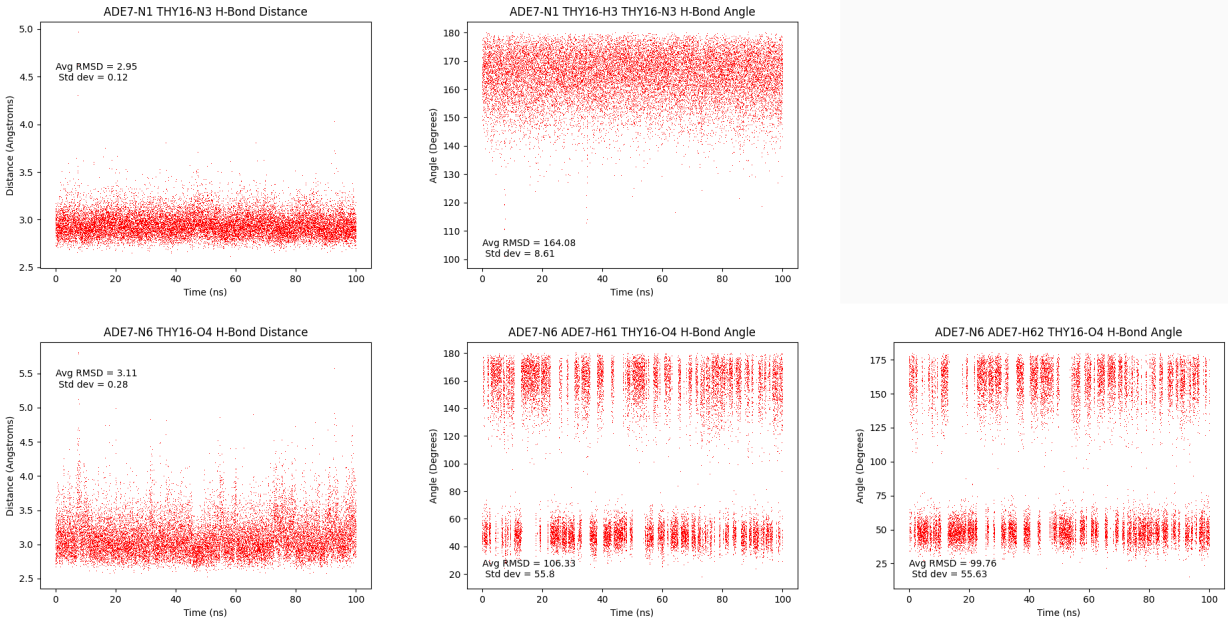
(6.112) PHE-DNA: Base step trajectories



(6.113) PHE-DNA: dC_5:dG_{18} hydrogen bond trajectories

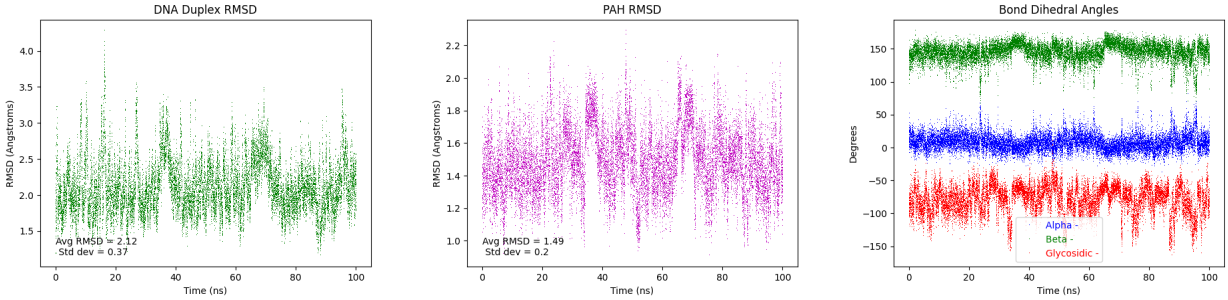


(6.114) PHE-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

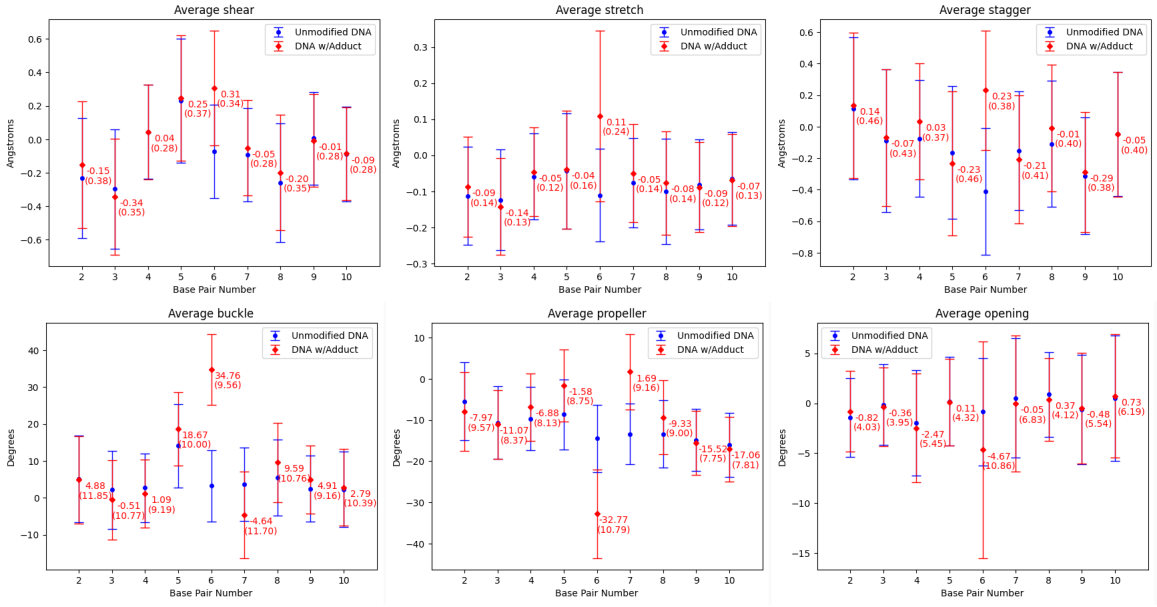


(6.115) PHE-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

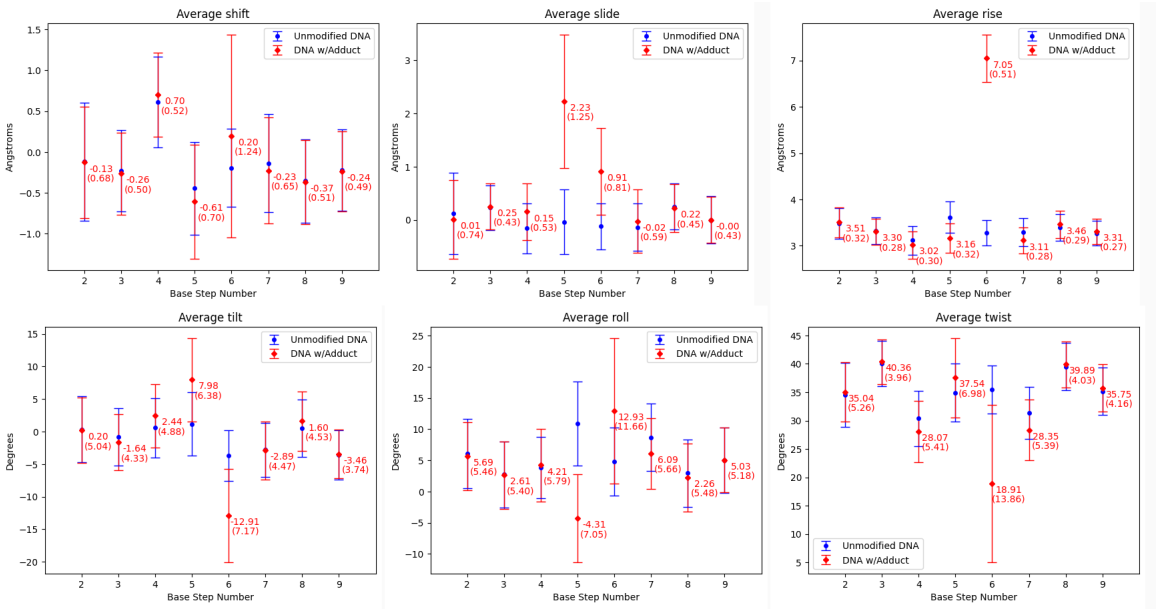
6.2.1.7 B[c]P-DNA



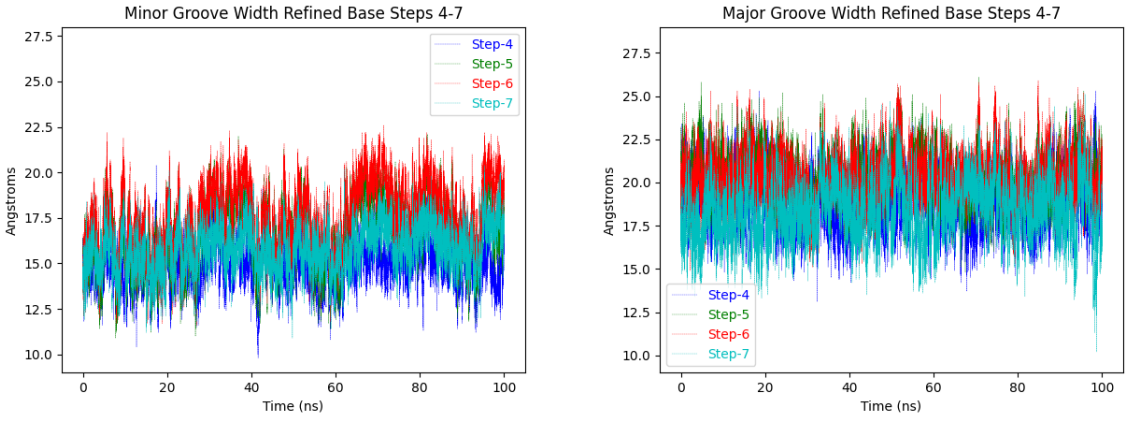
(6.116) B[c]P-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



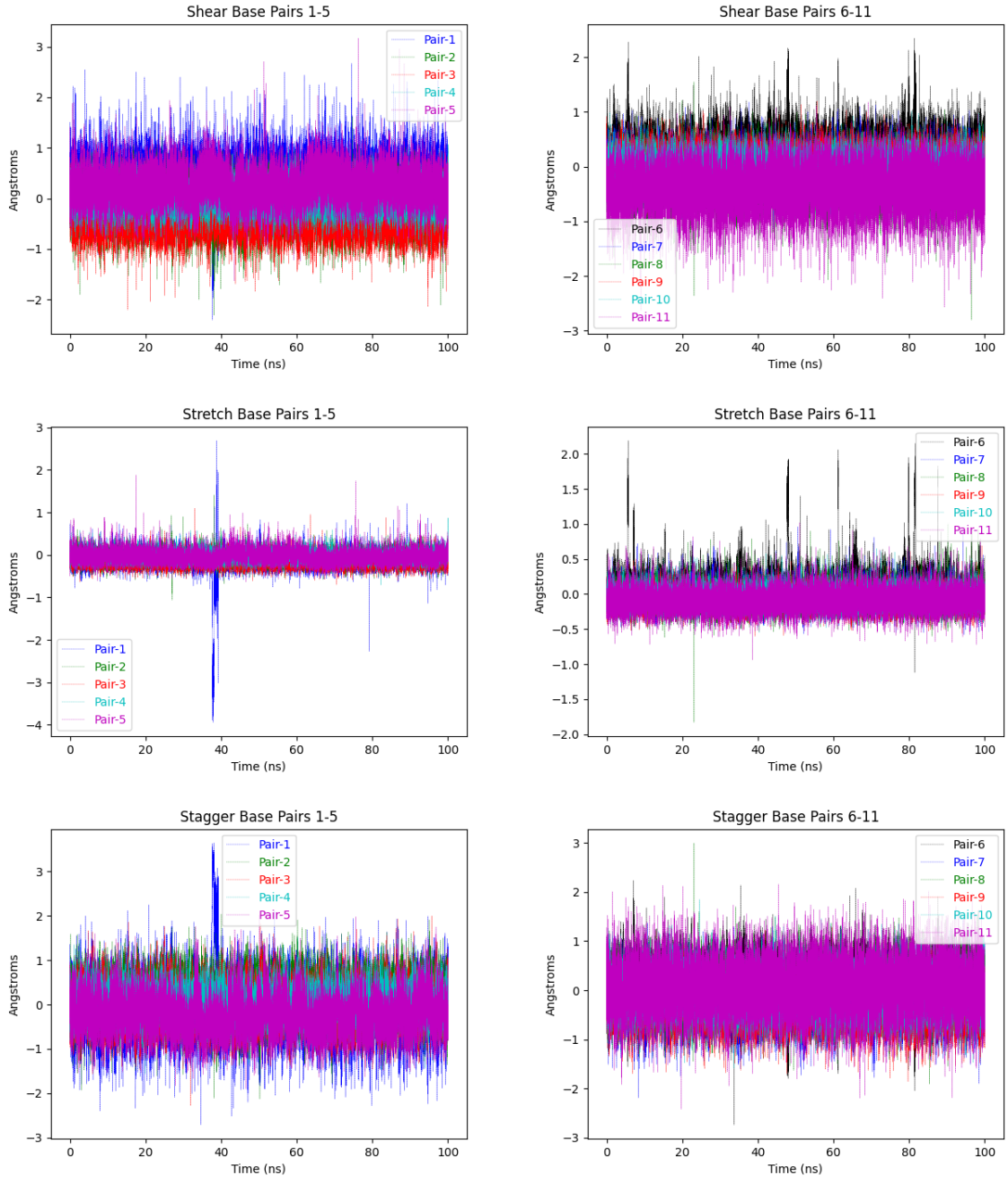
(6.117) B[c]P-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



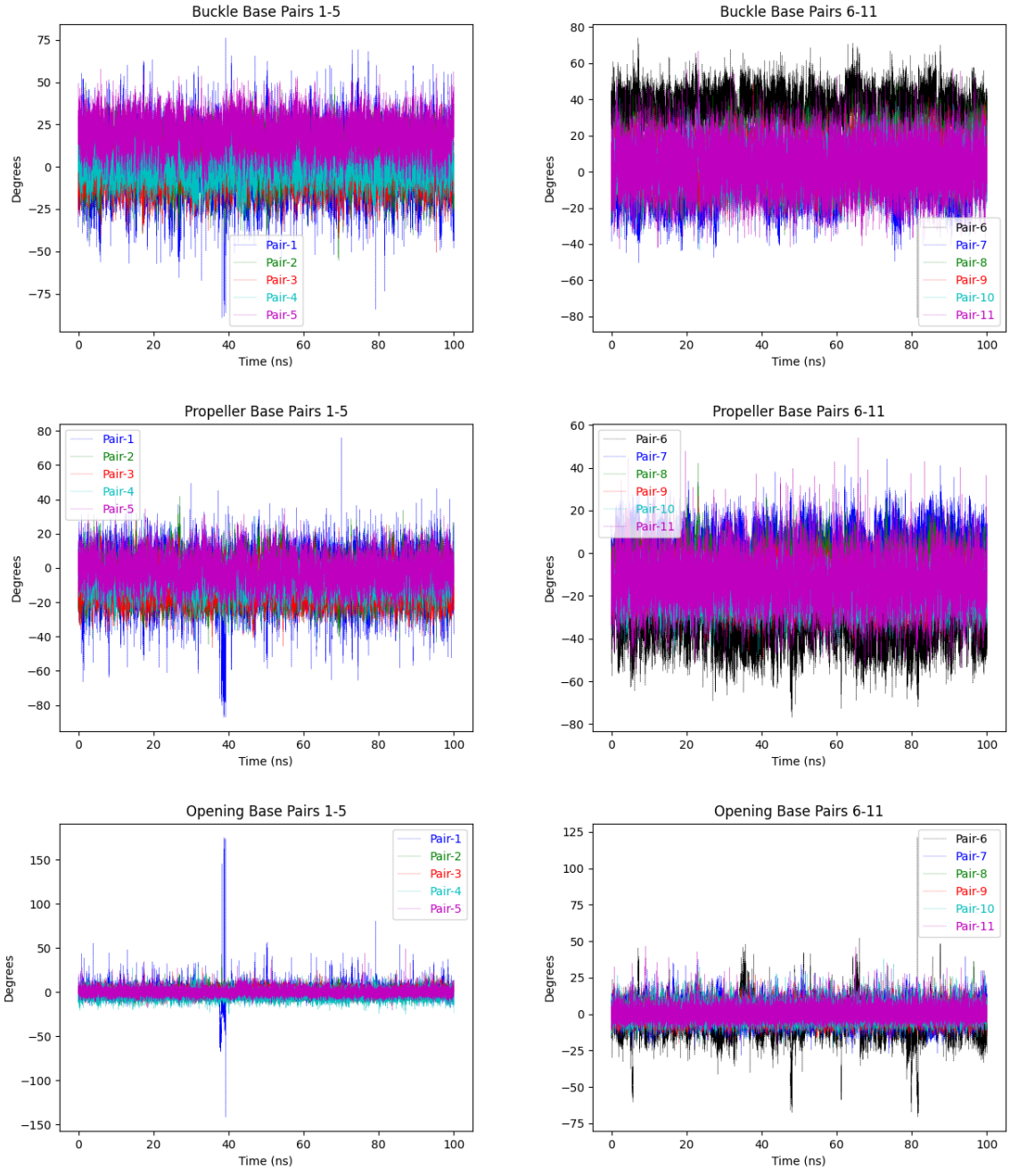
(6.118) B[c]P-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



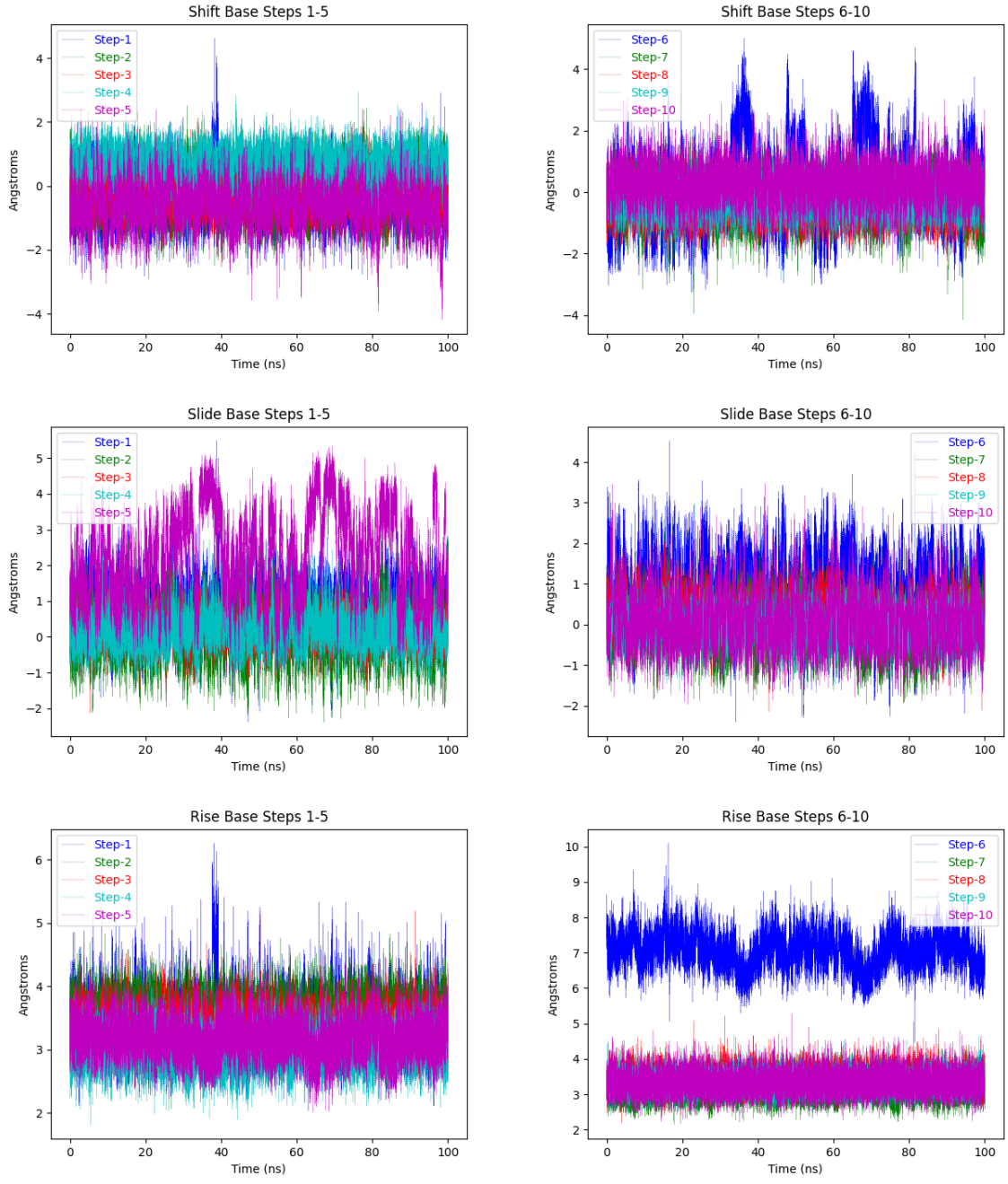
(6.119) B[c]P-DNA: Refined major and minor groove trajectories



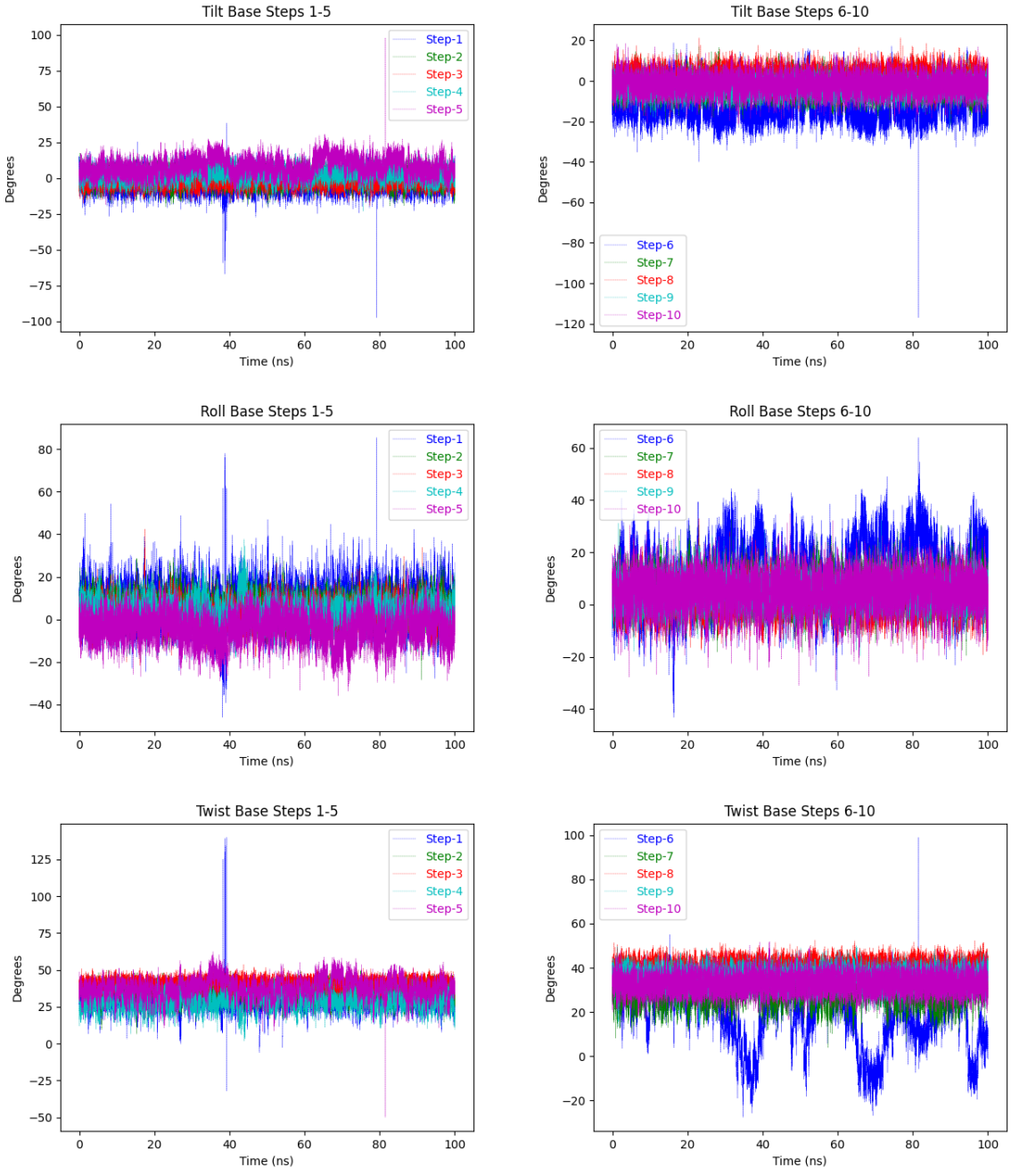
(6.120) B[c]P-DNA: Base pair trajectories



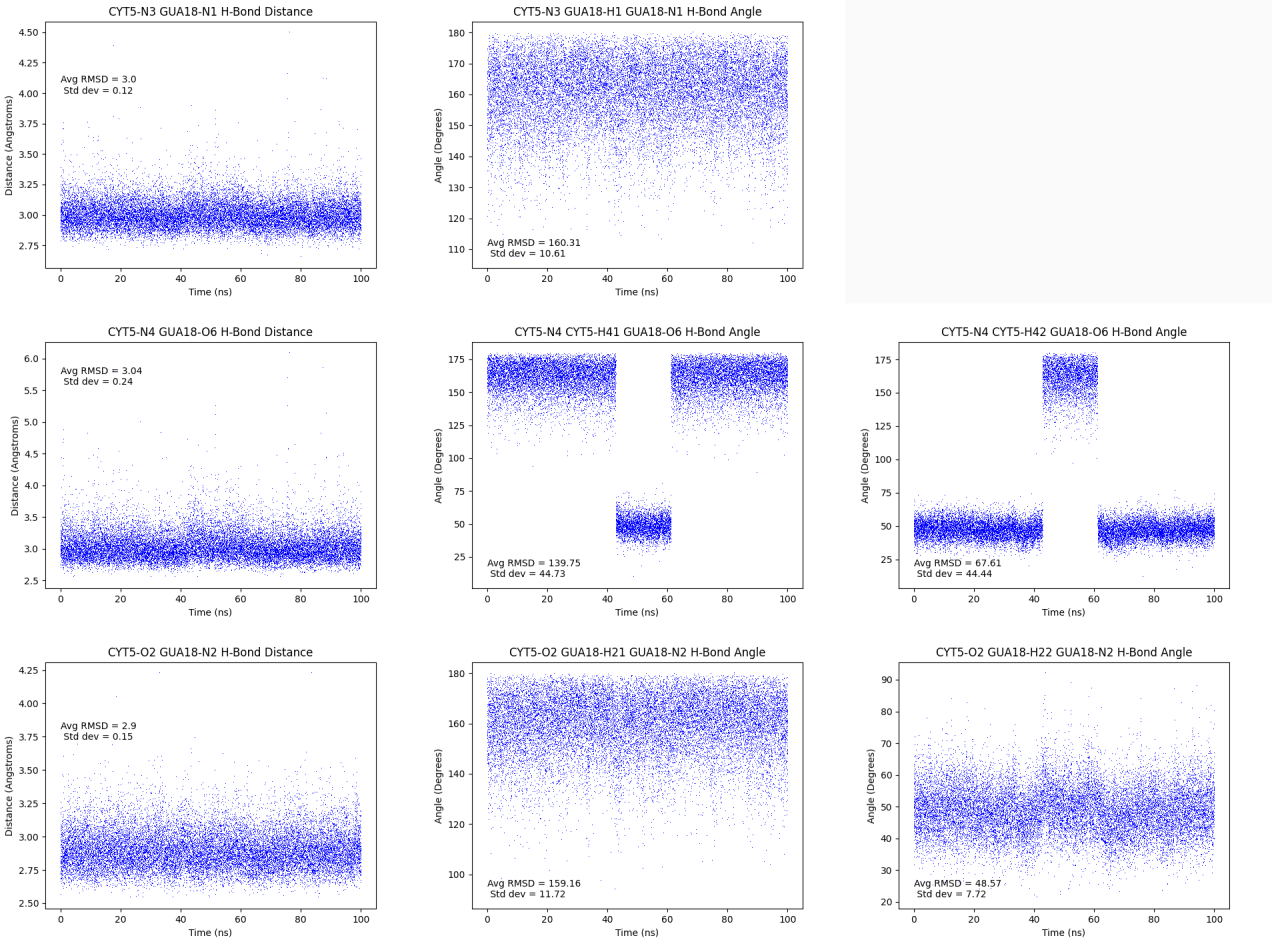
(6.121) B[c]P-DNA: Base pair trajectories



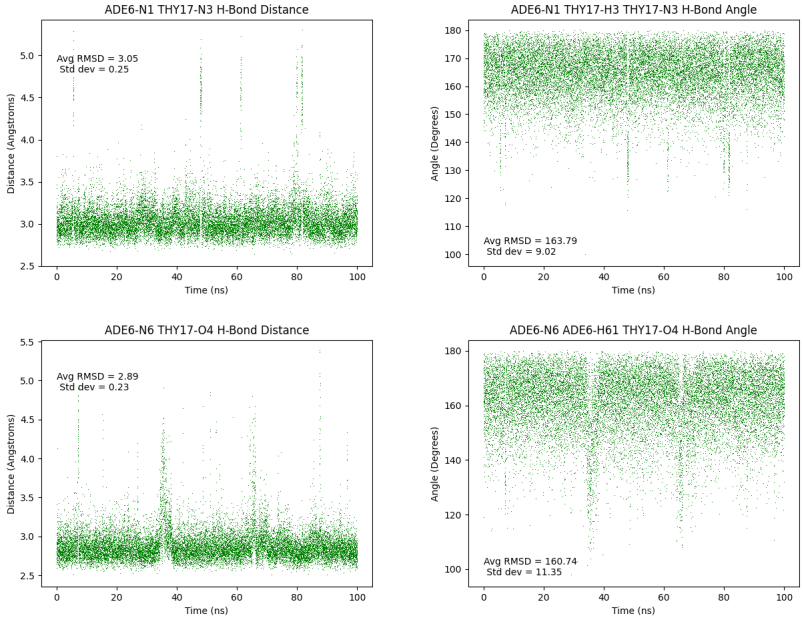
(6.122) B[c]P-DNA: Base step trajectories



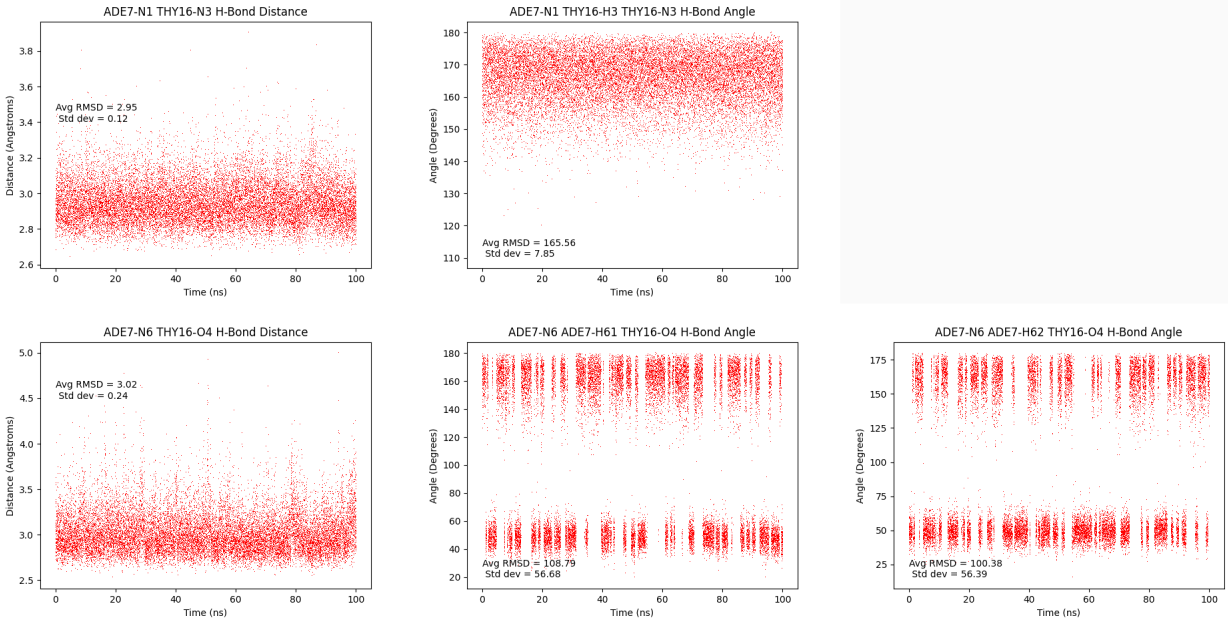
(6.123) B[c]P-DNA: Base step trajectories



(6.124) B[c]P-DNA: dC_5:dG_{18} hydrogen bond trajectories

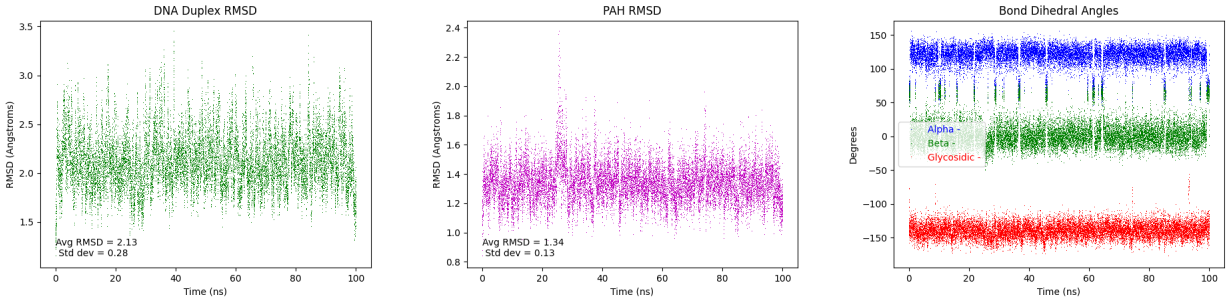


(6.125) B[c]P-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

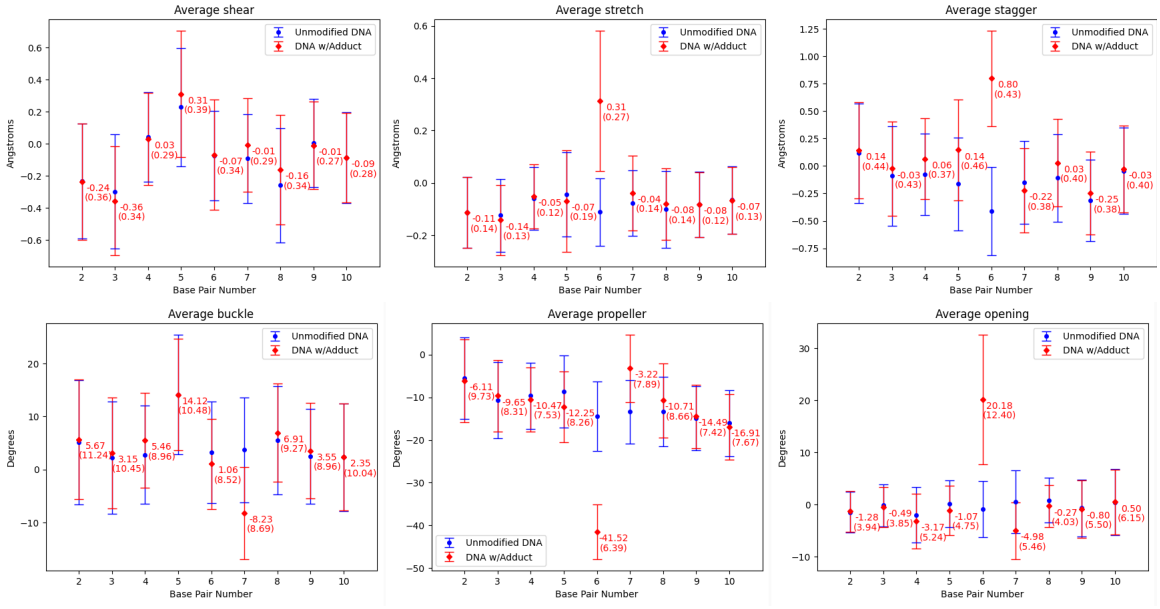


(6.126) B[c]P-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

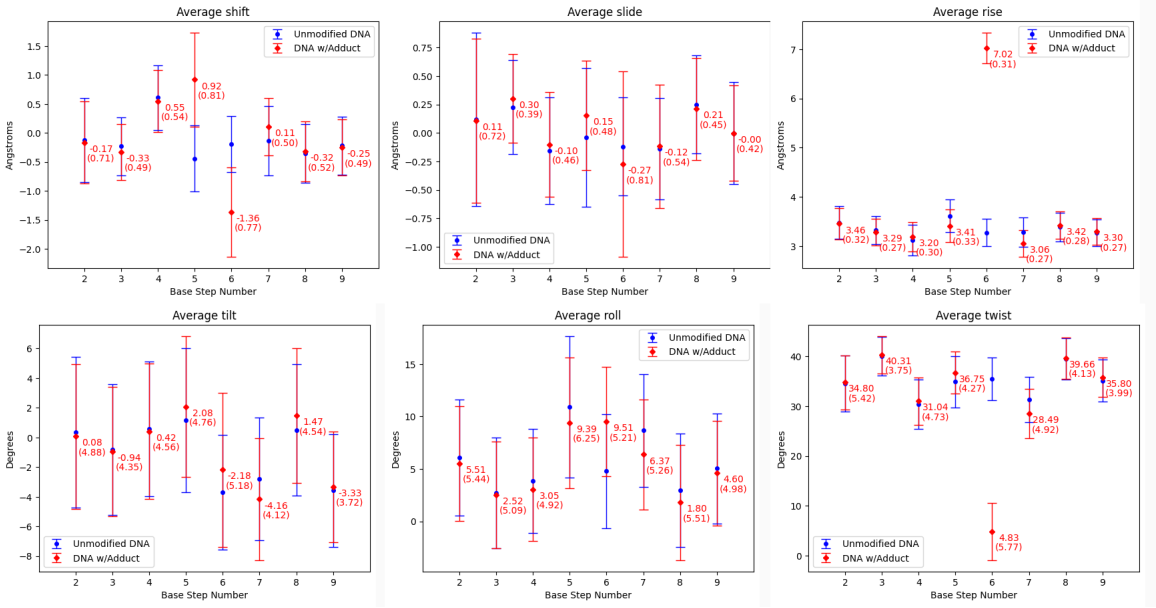
6.2.1.8 DB[a,e]P-DNA



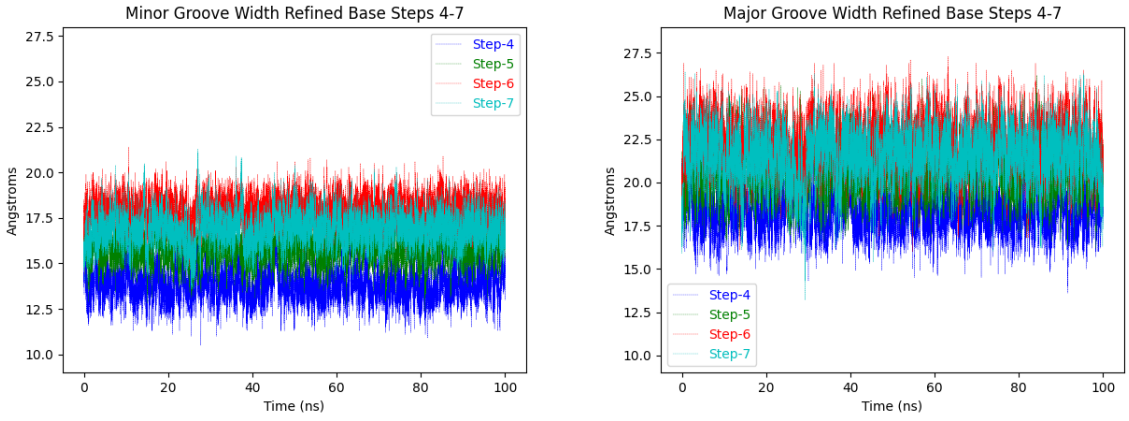
(6.127) DB[a,e]P-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



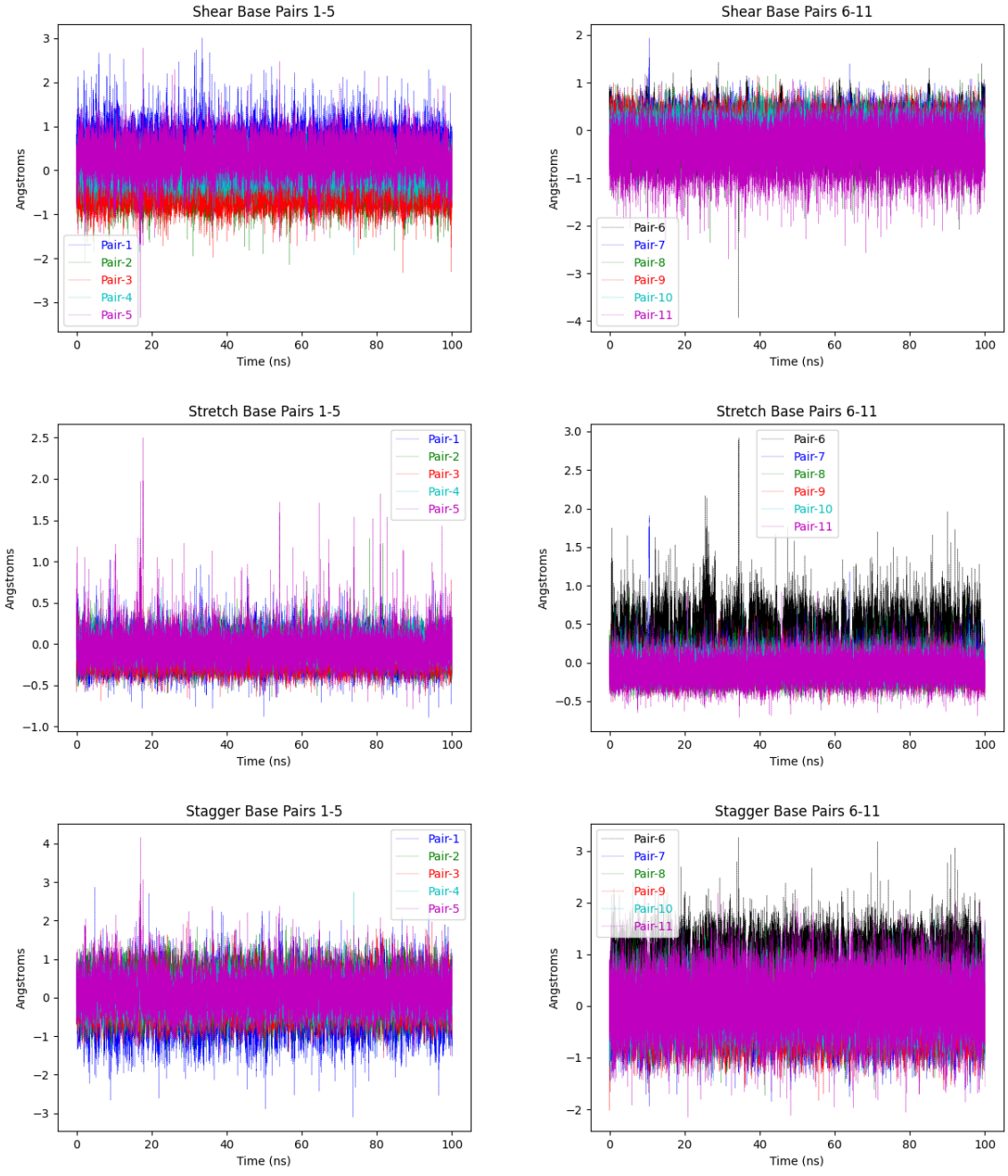
(6.128) DB[a,e]P-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



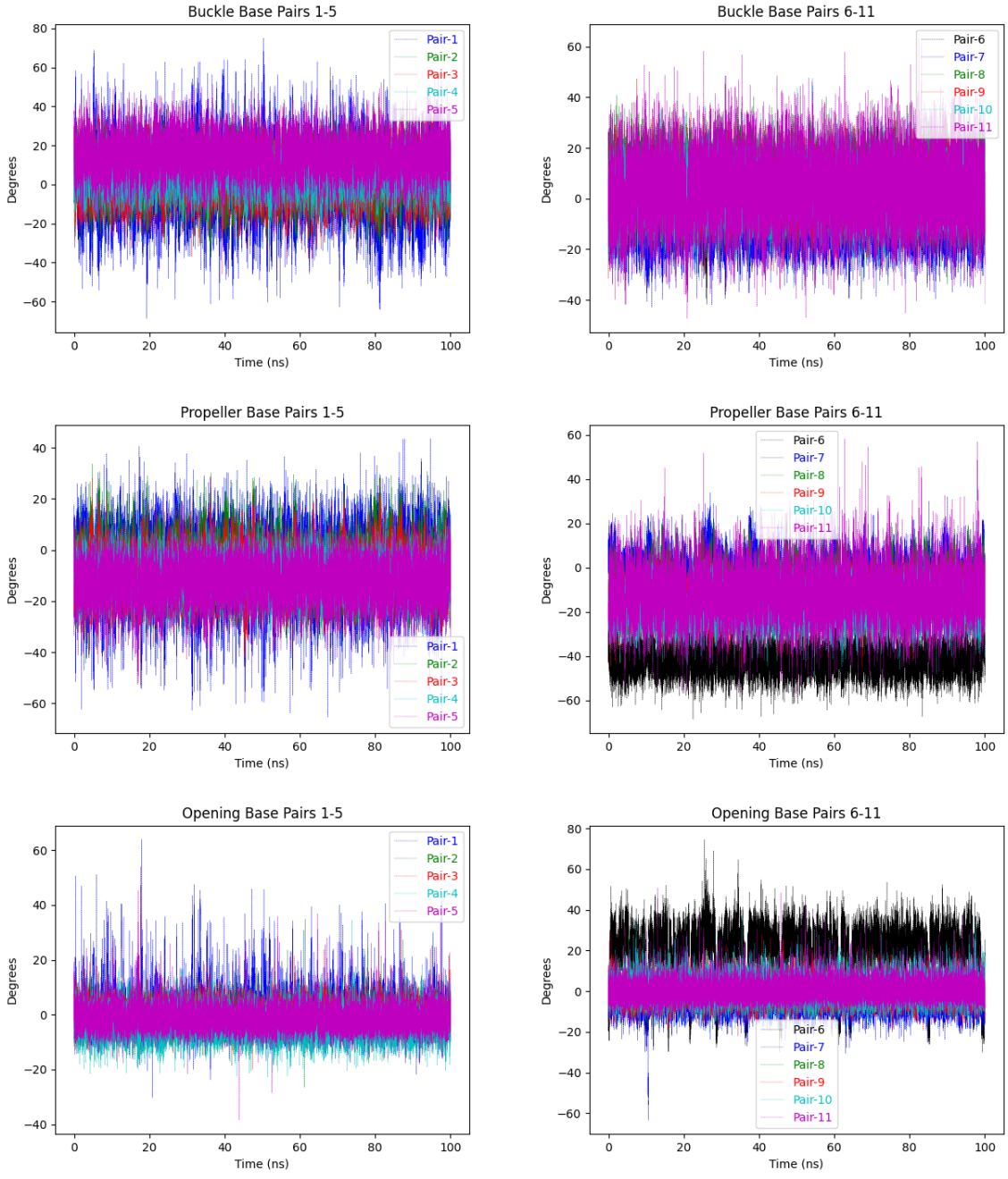
(6.129) DB[a,e]P-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



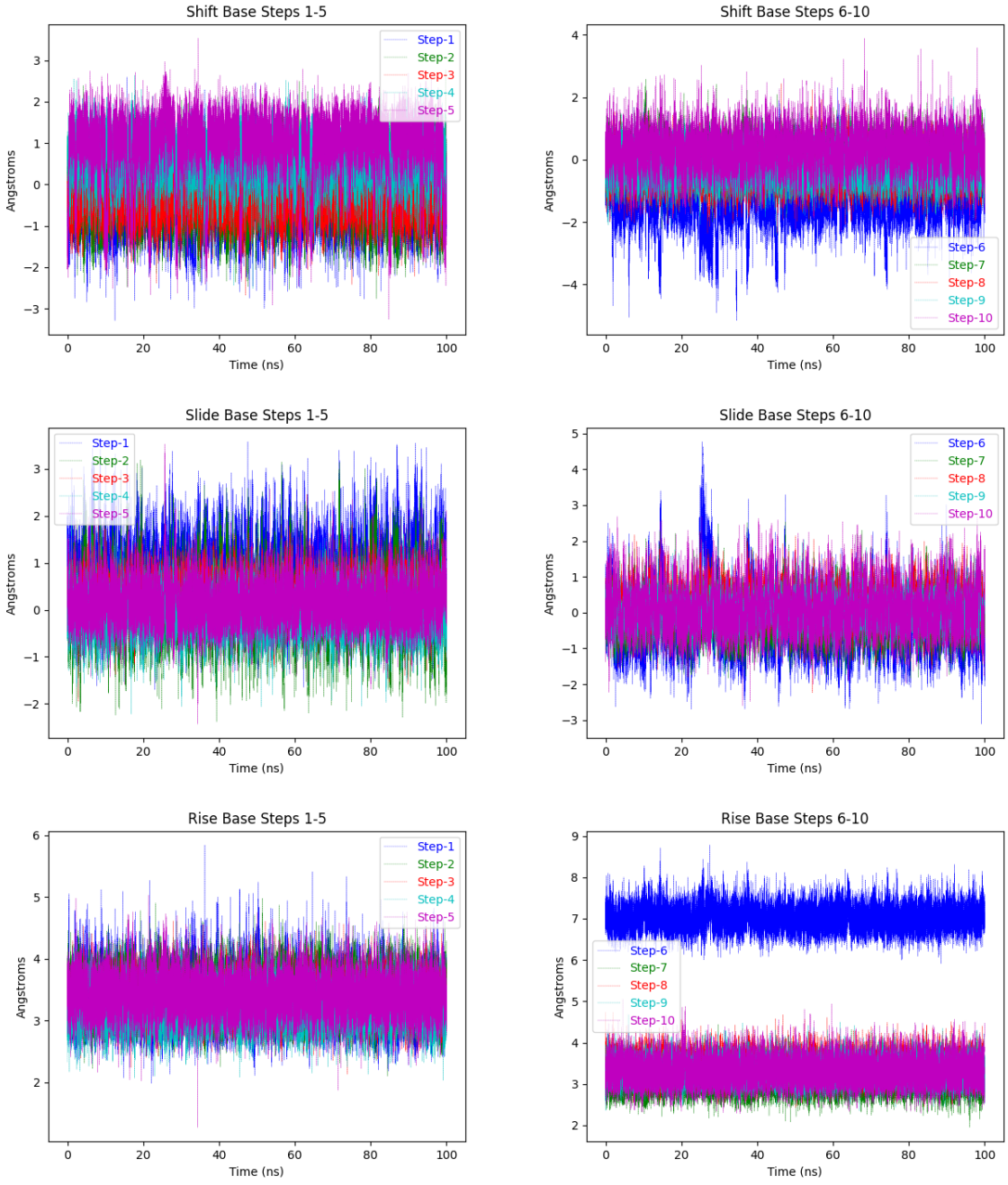
(6.130) DB[a,e]P-DNA: Refined major and minor groove trajectories



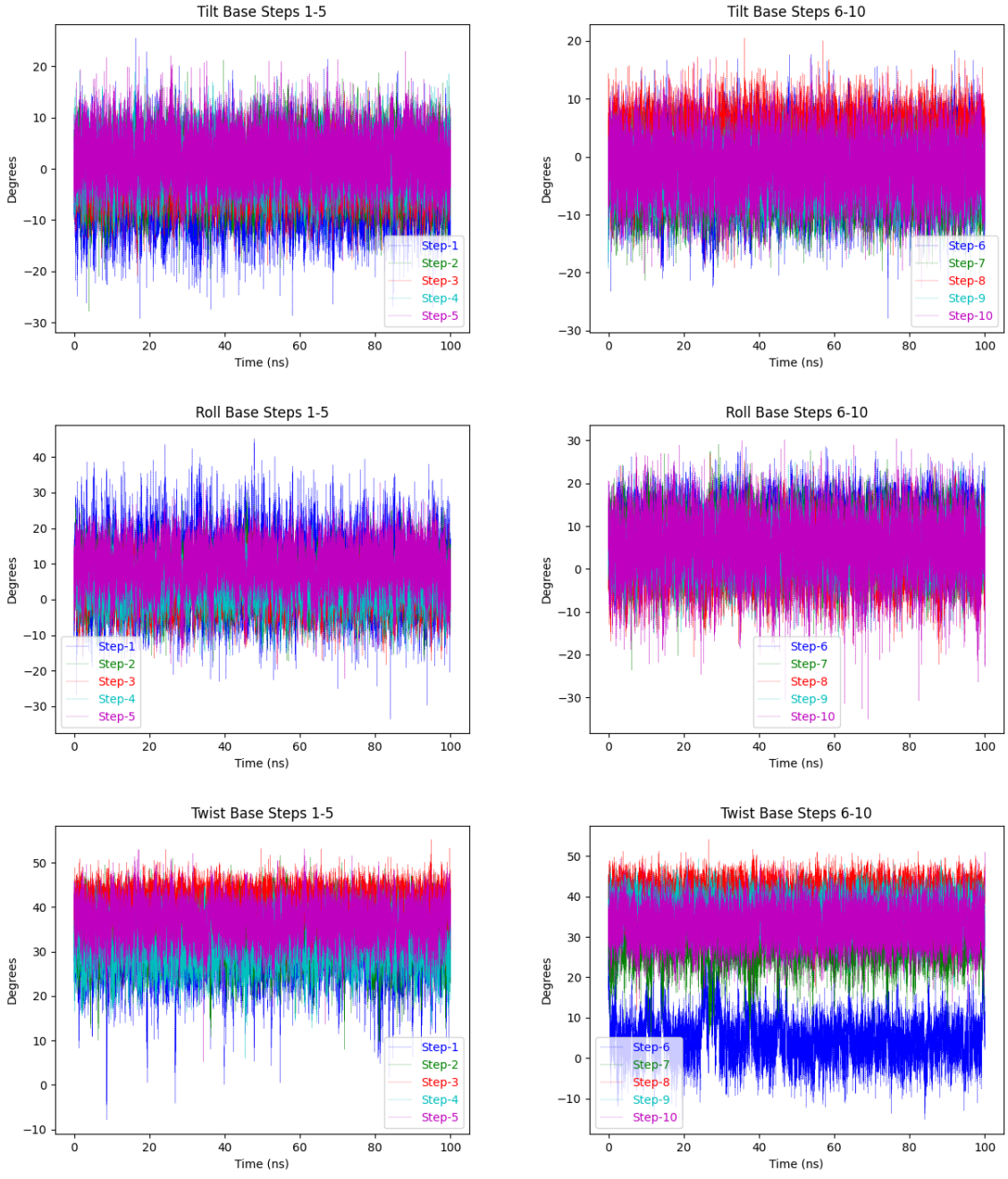
(6.131) DB[a,e]P-DNA: Base pair trajectories



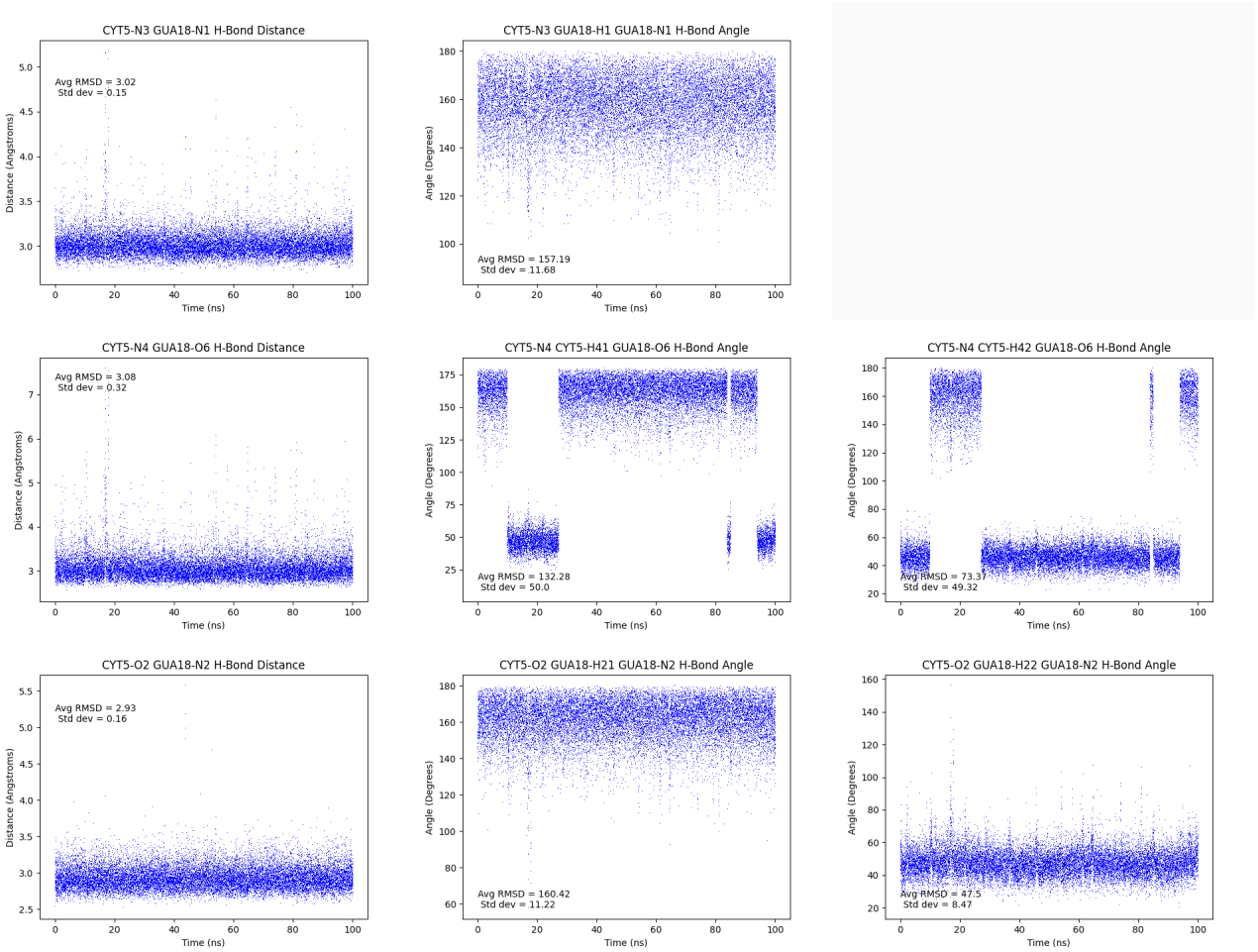
(6.132) DB[a,e]P-DNA: Base pair trajectories



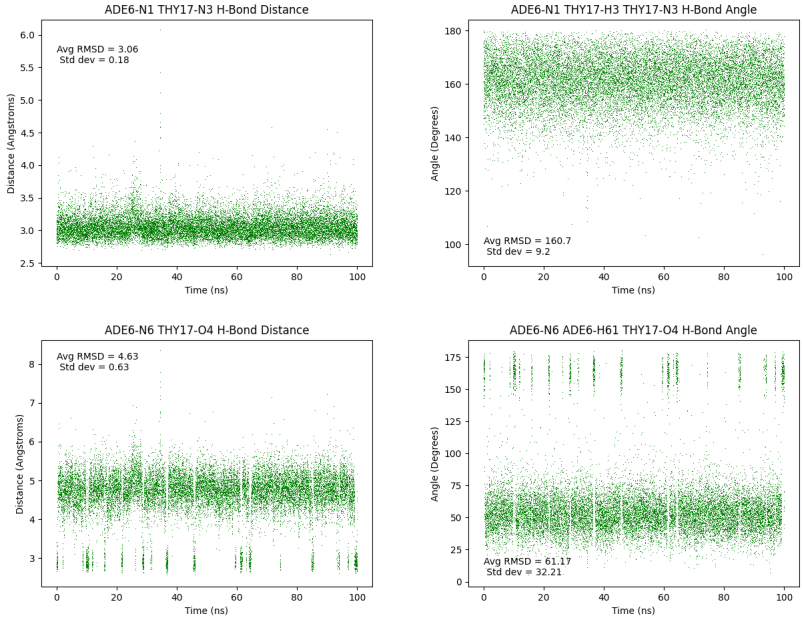
(6.133) DB[a,e]P-DNA: Base step trajectories



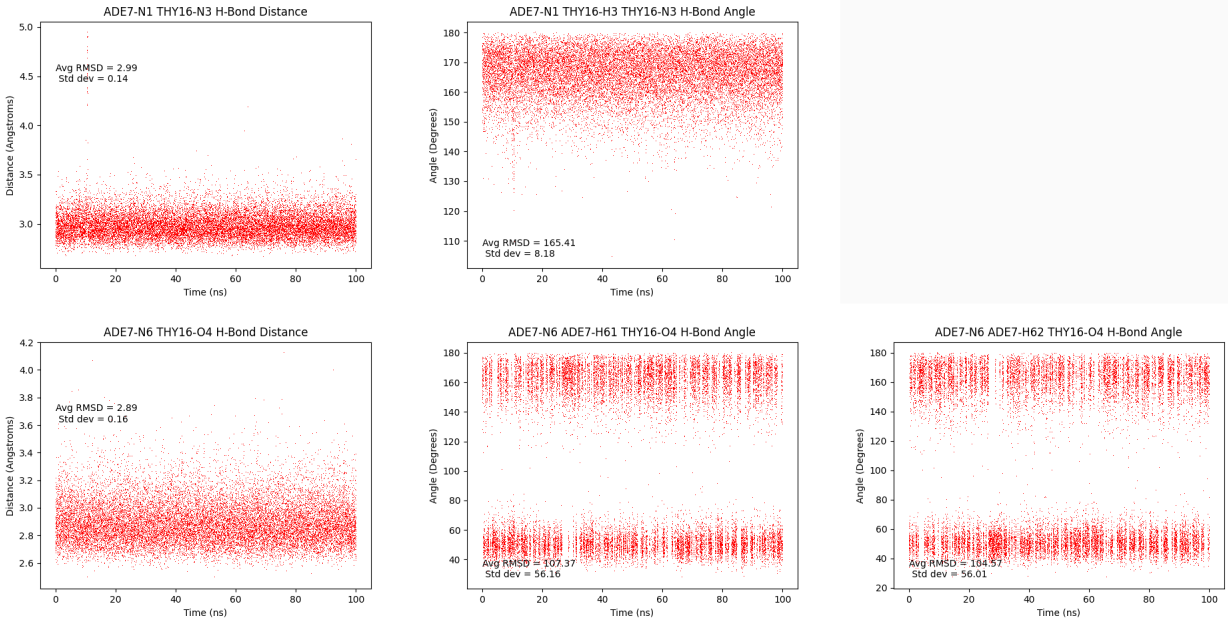
(6.134) DB[a,e]P-DNA: Base step trajectories



(6.135) DB[a,e]P-DNA: dC_5:dG_{18} hydrogen bond trajectories

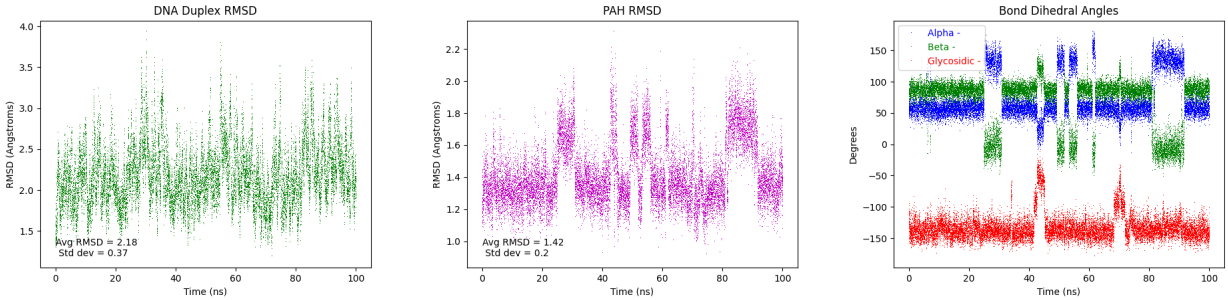


(6.136) DB[a,e]P-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

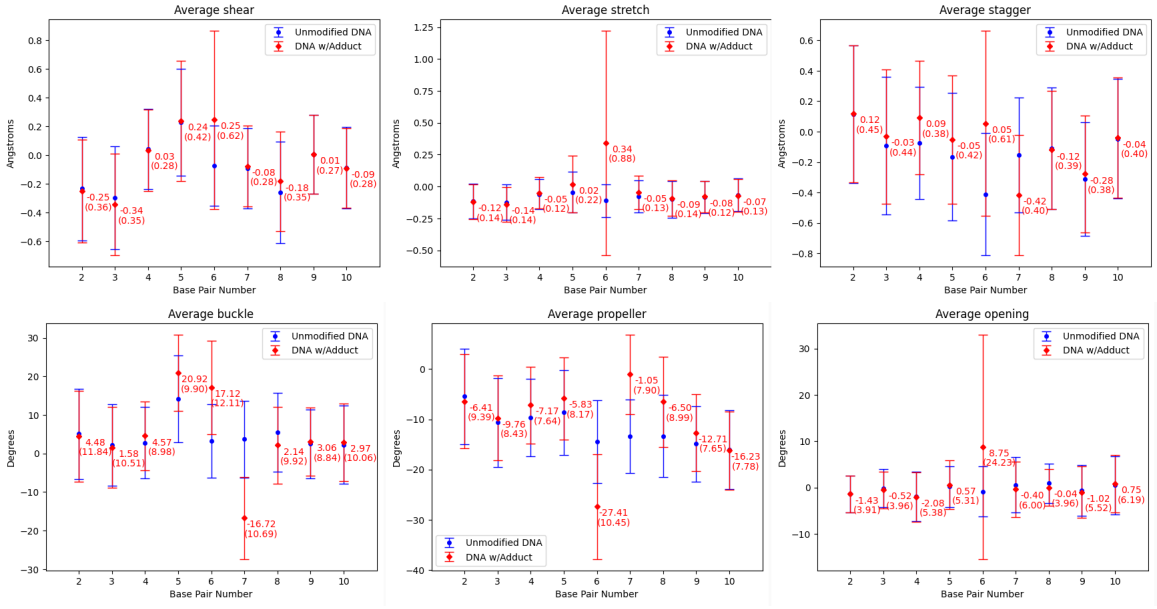


(6.137) DB[a,e]P-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

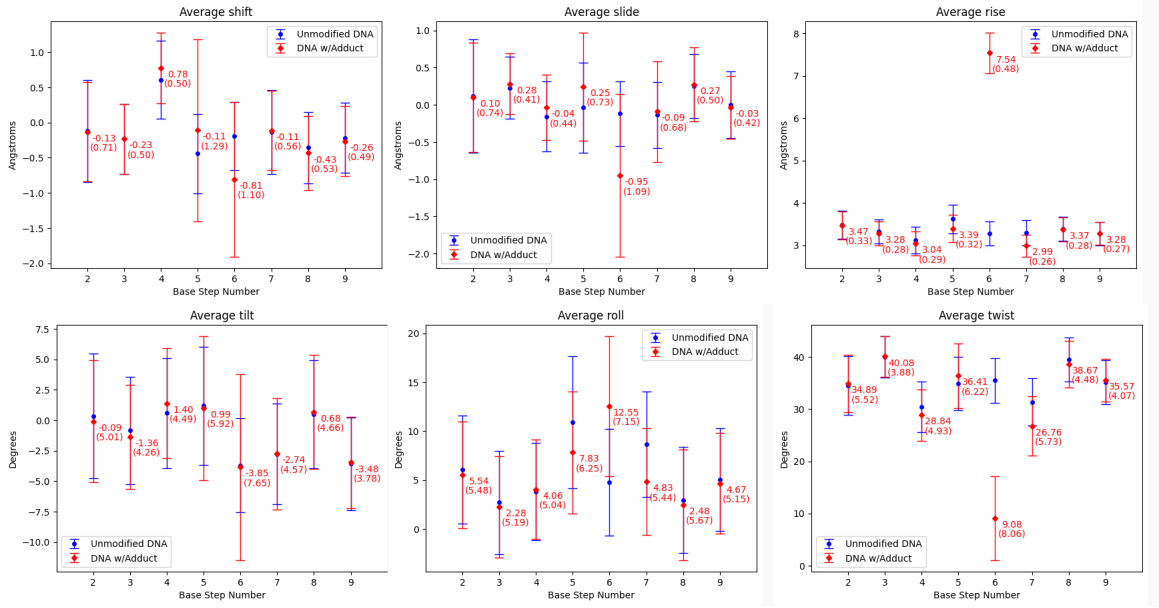
6.2.1.9 DB[a,h]P-DNA



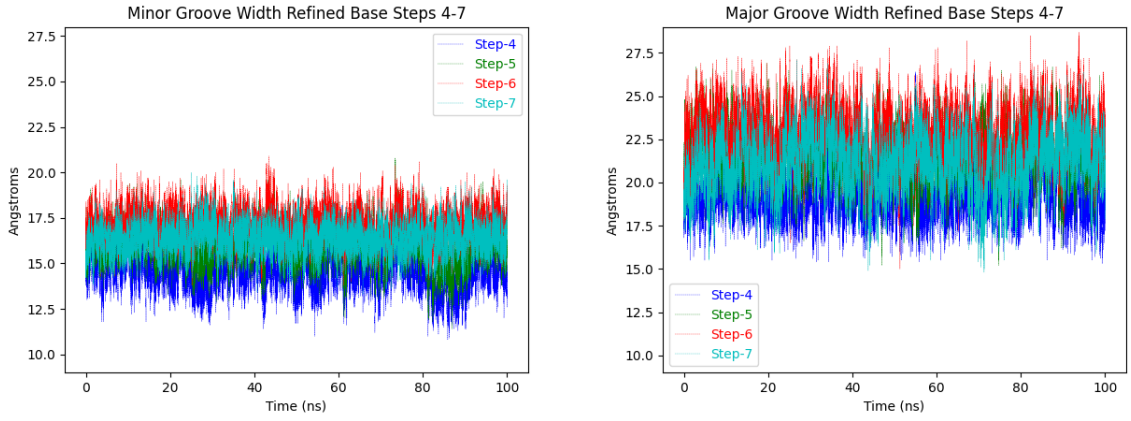
(6.138) DB[a,h]P-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



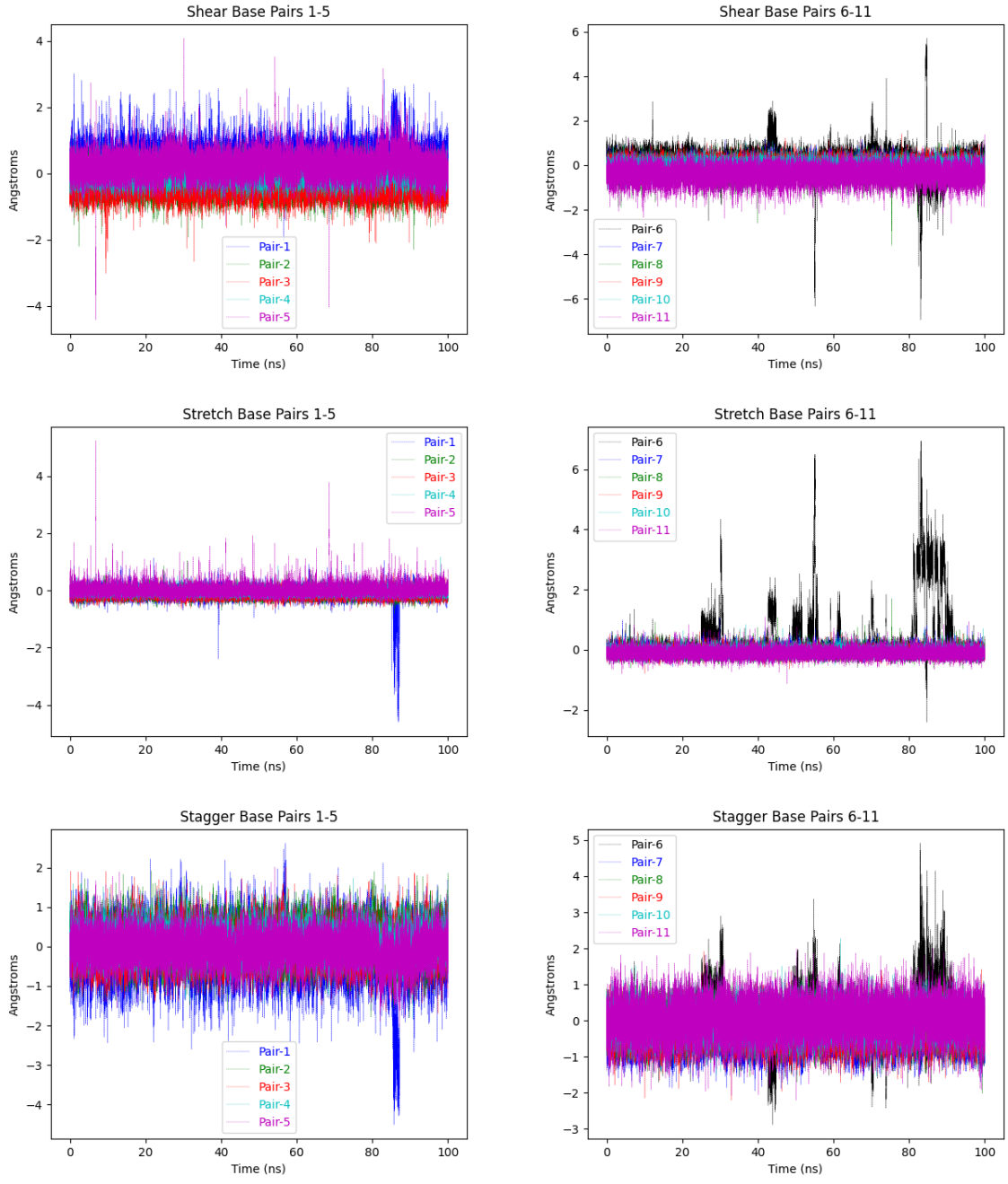
(6.139) DB[a,h]P-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



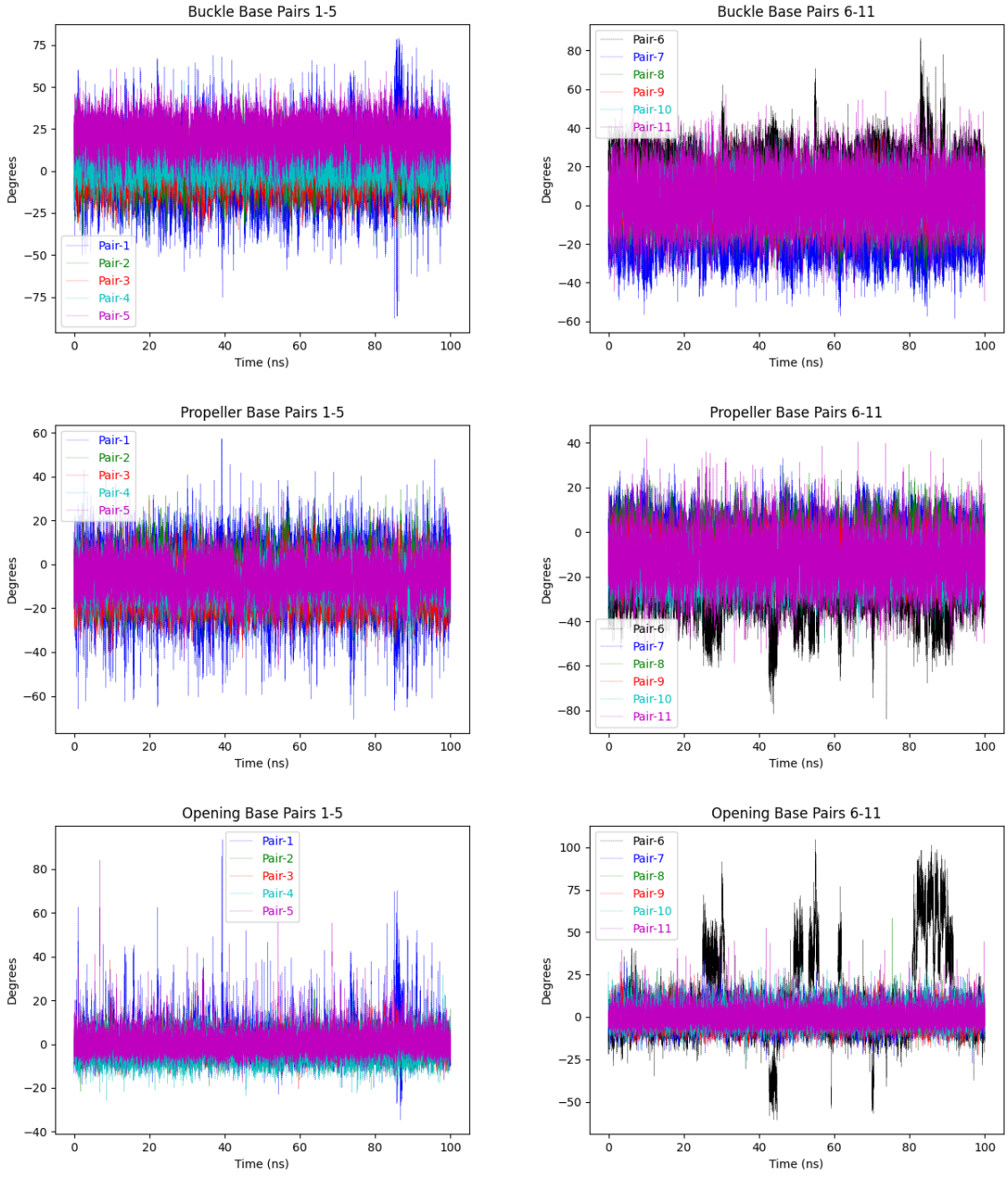
(6.140) DB[a,h]P-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



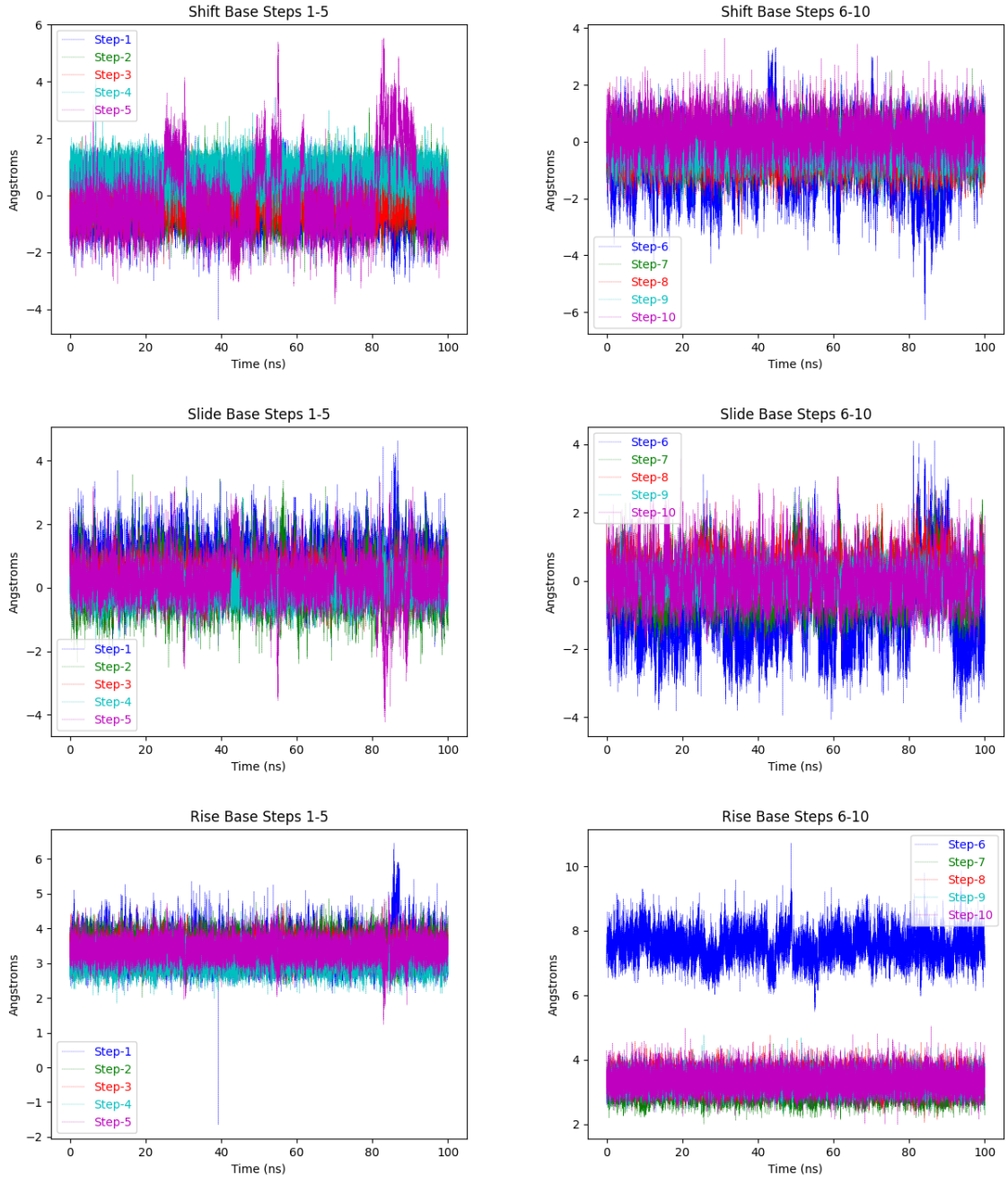
(6.141) DB[a,h]P-DNA: Refined major and minor groove trajectories



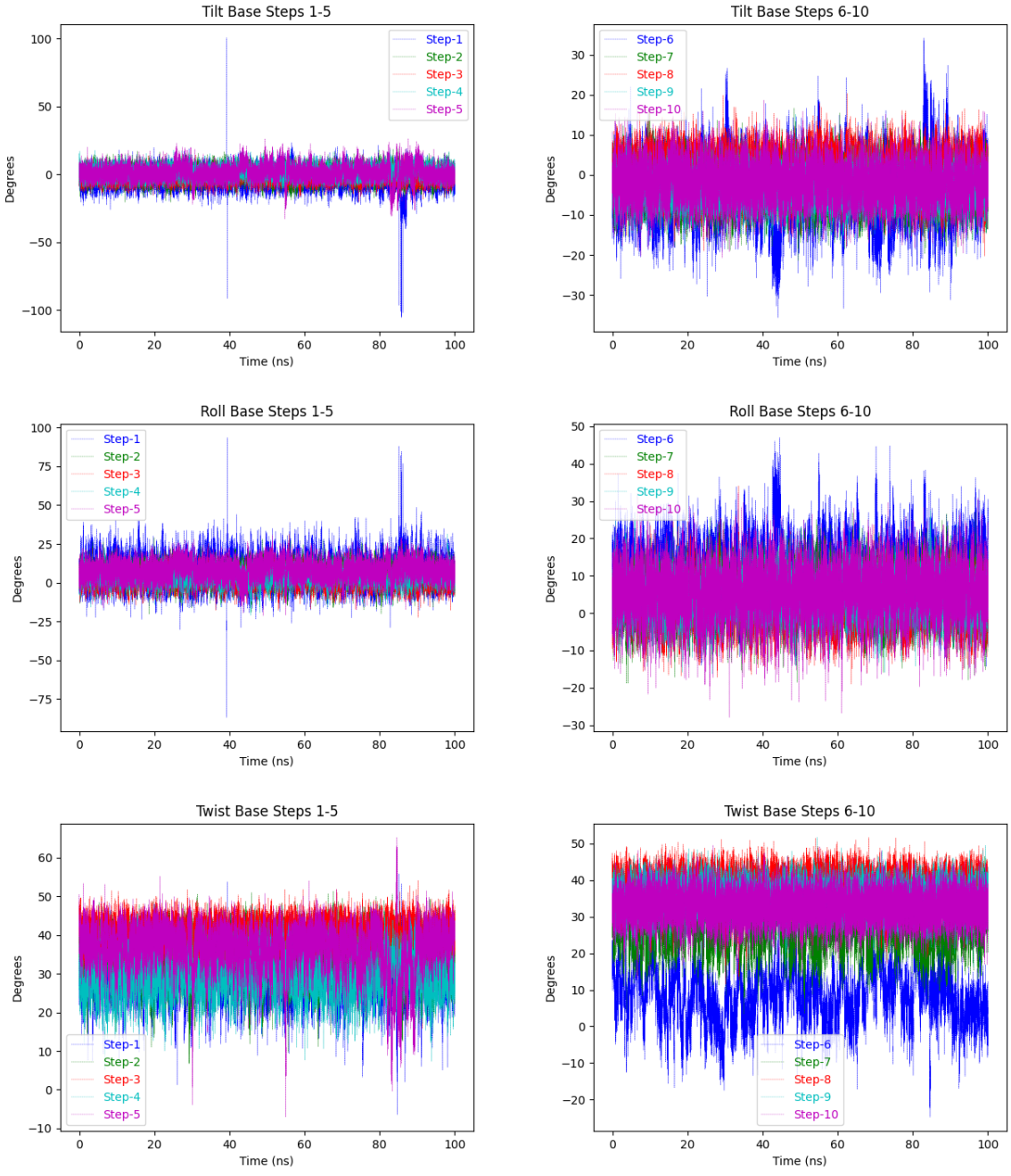
(6.142) DB[a,h]P-DNA: Base pair trajectories



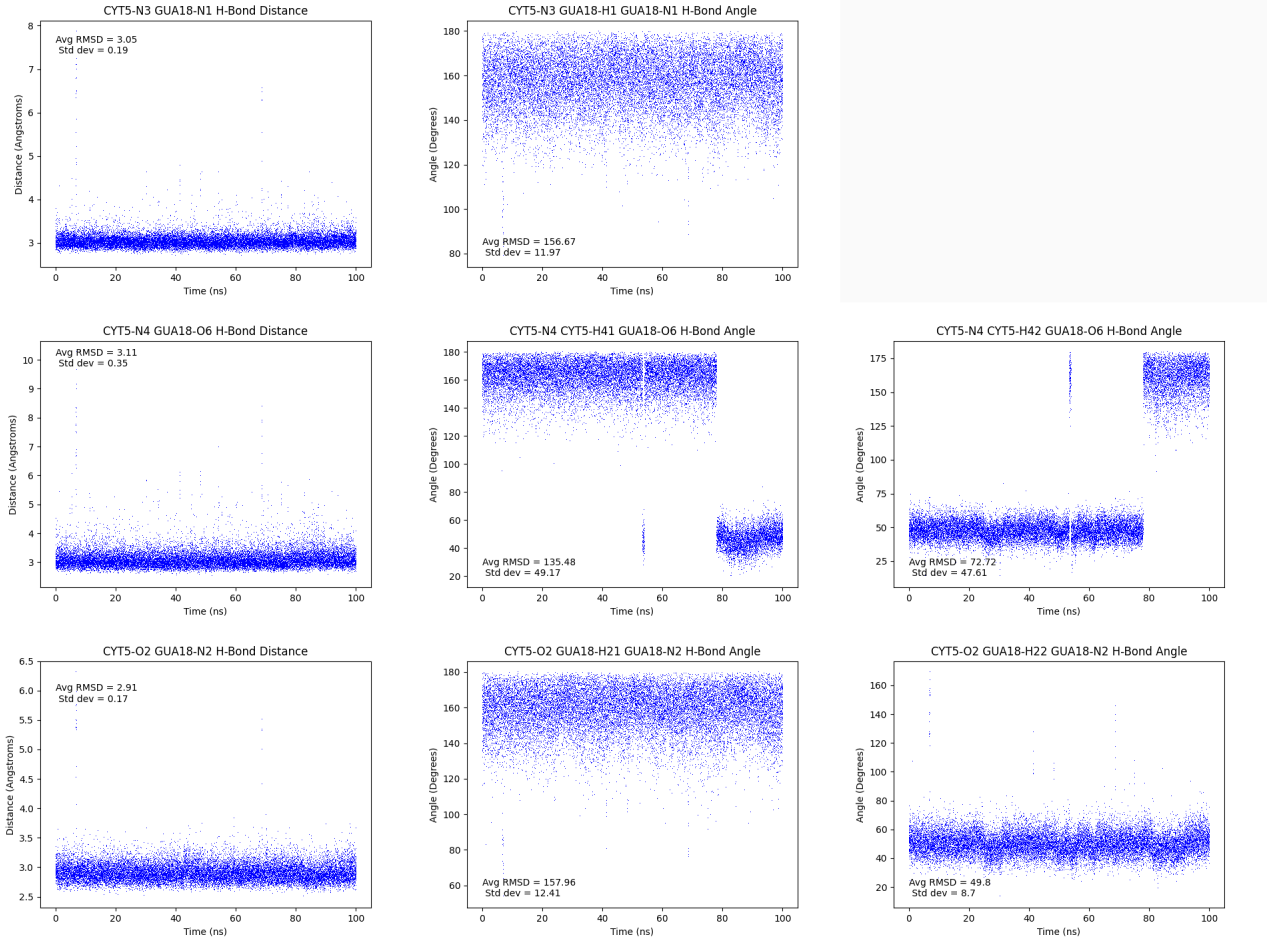
(6.143) DB[a,h]P-DNA: Base pair trajectories



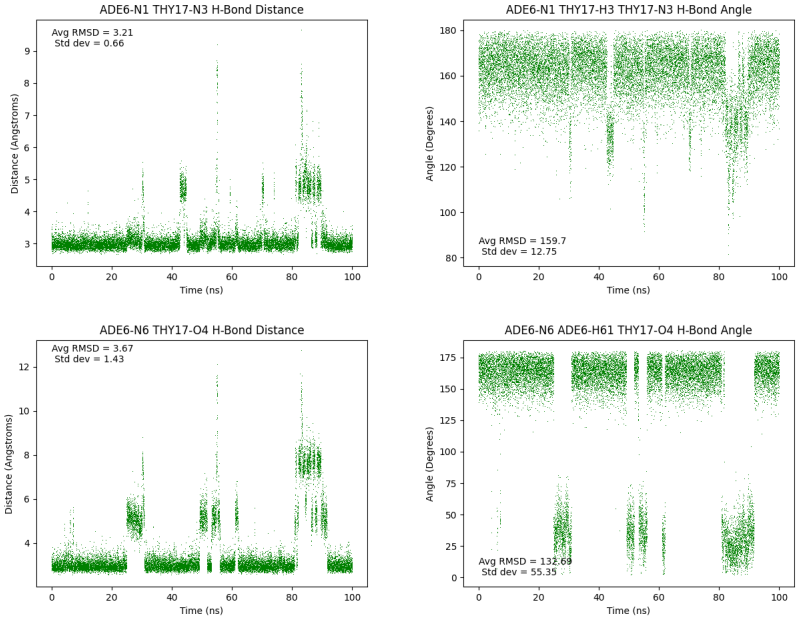
(6.144) DB[a,h]P-DNA: Base step trajectories



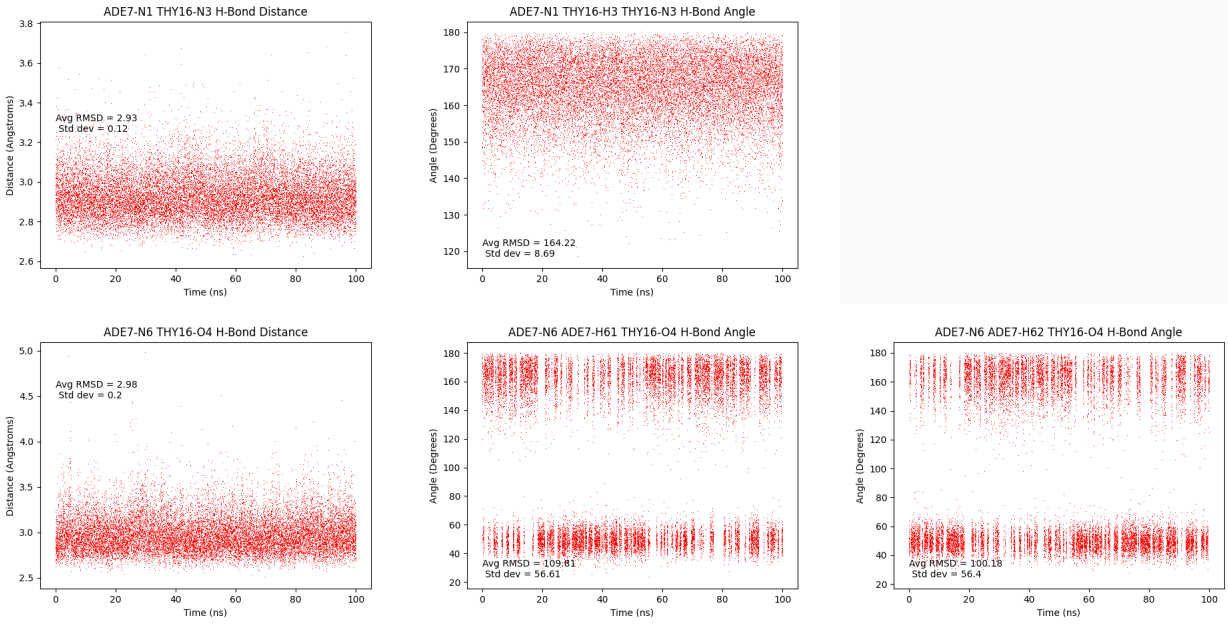
(6.145) DB[a,h]P-DNA: Base step trajectories



(6.146) DB[a,h]P-DNA: dC_5:dG_{18} hydrogen bond trajectories



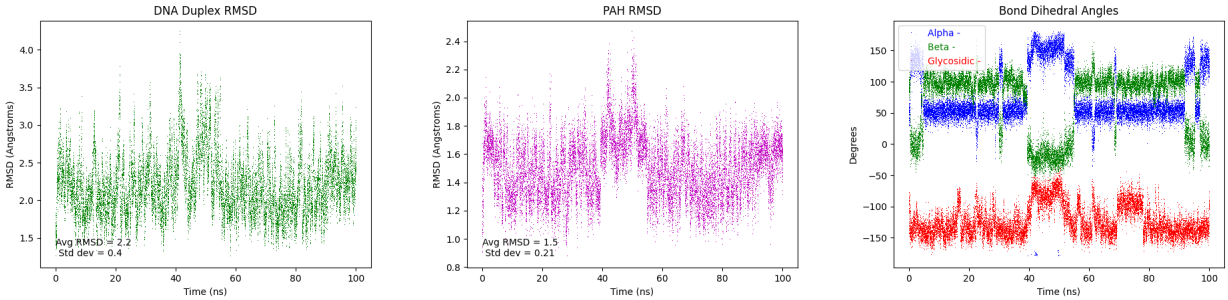
(6.147) DB[a,h]P-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories



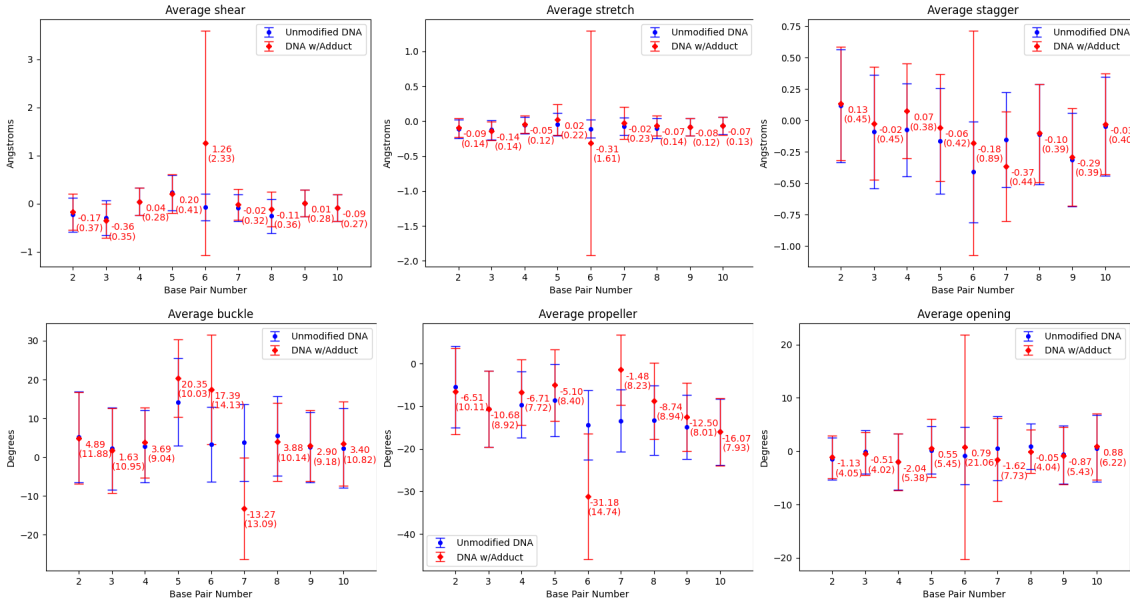
(6.148) DB[a,h]P-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

6.2.1.10 DB[a,i]P-DNA

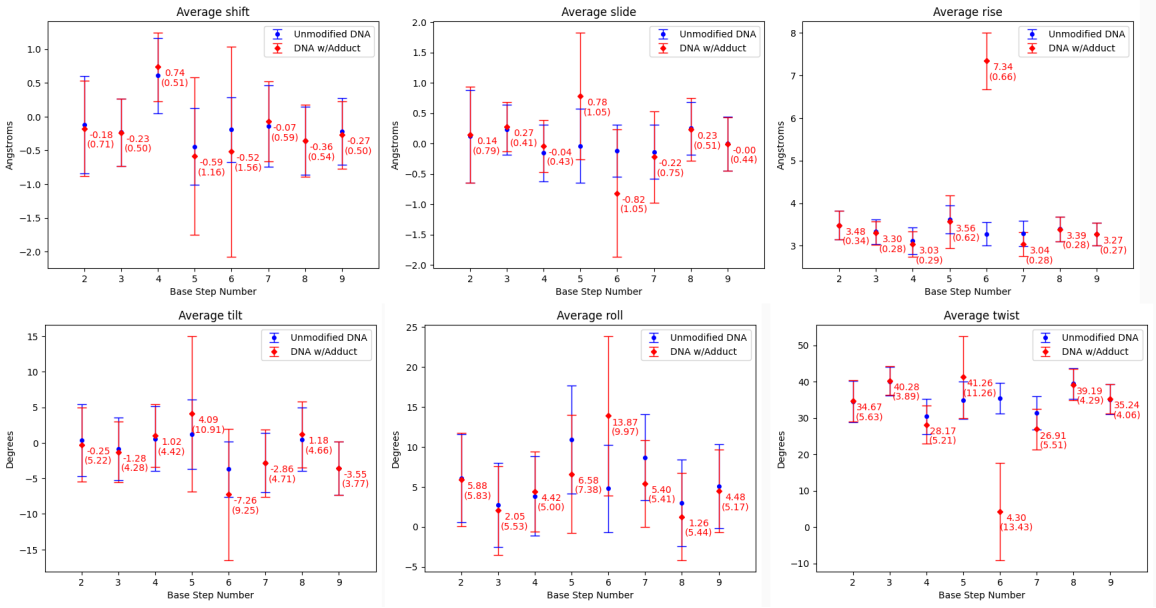
Average base step and base pair rigid-body parameters for the DB[a,i]P-DNA system were calculated over an equilibrated 90 ns subset of the NPT production run.



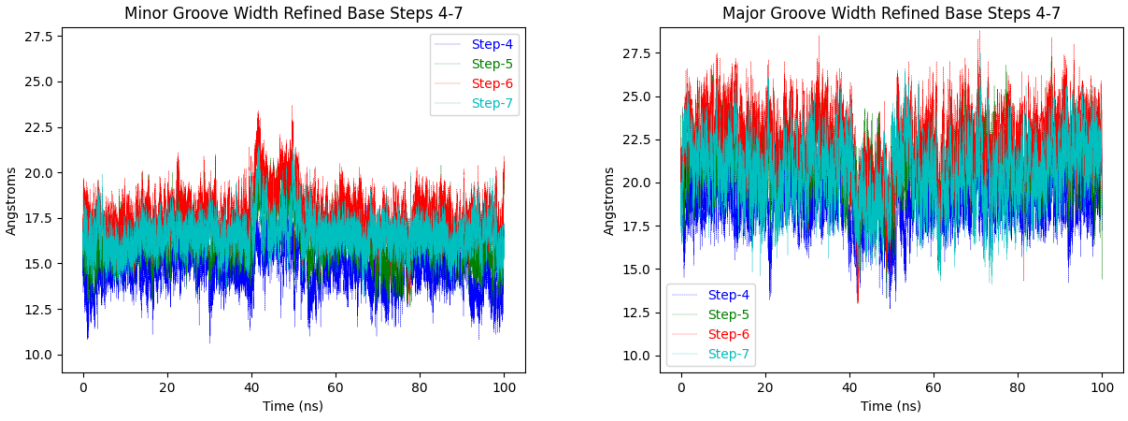
(6.149) DB[a,i]P-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



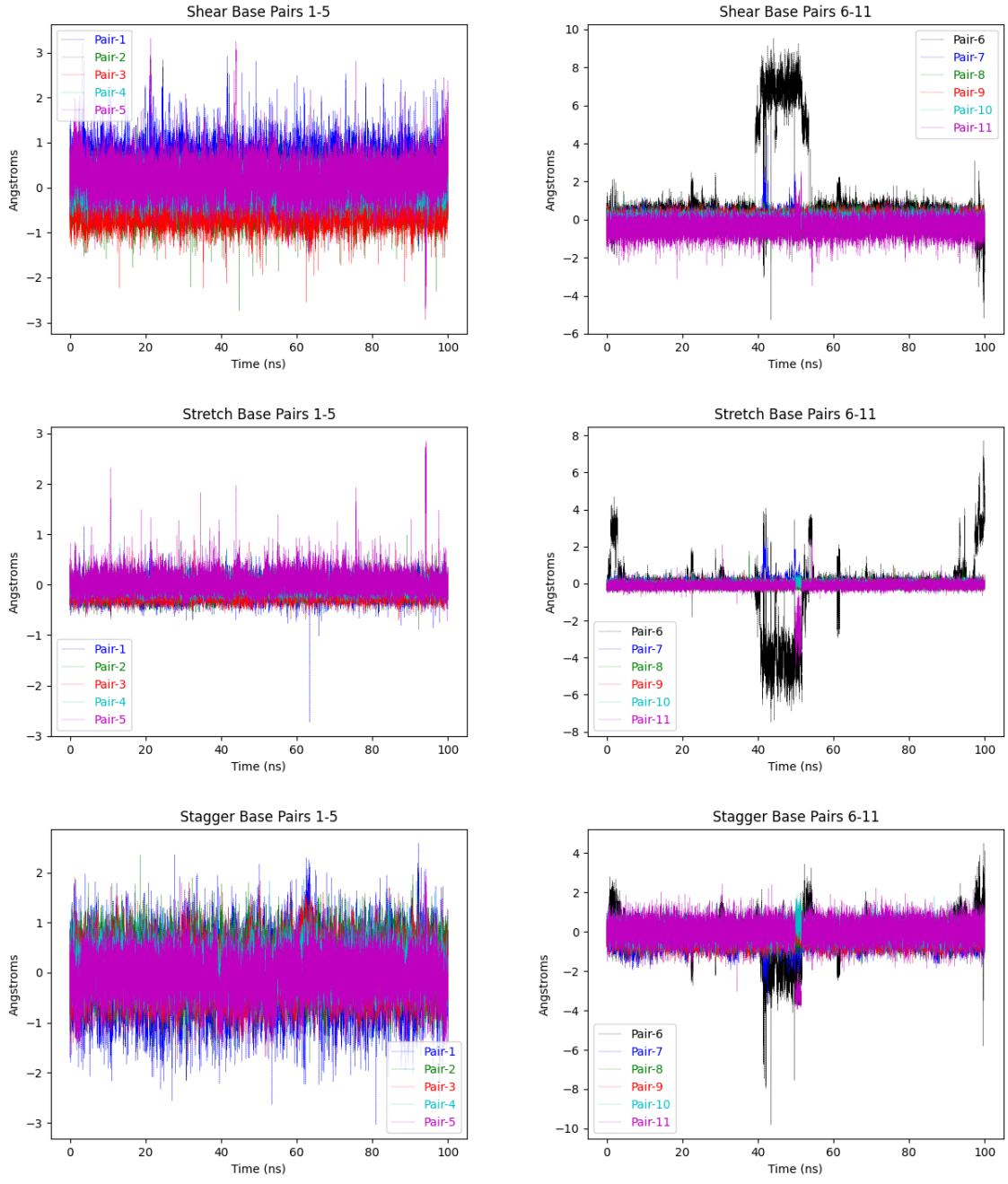
(6.150) DB[a,i]P-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



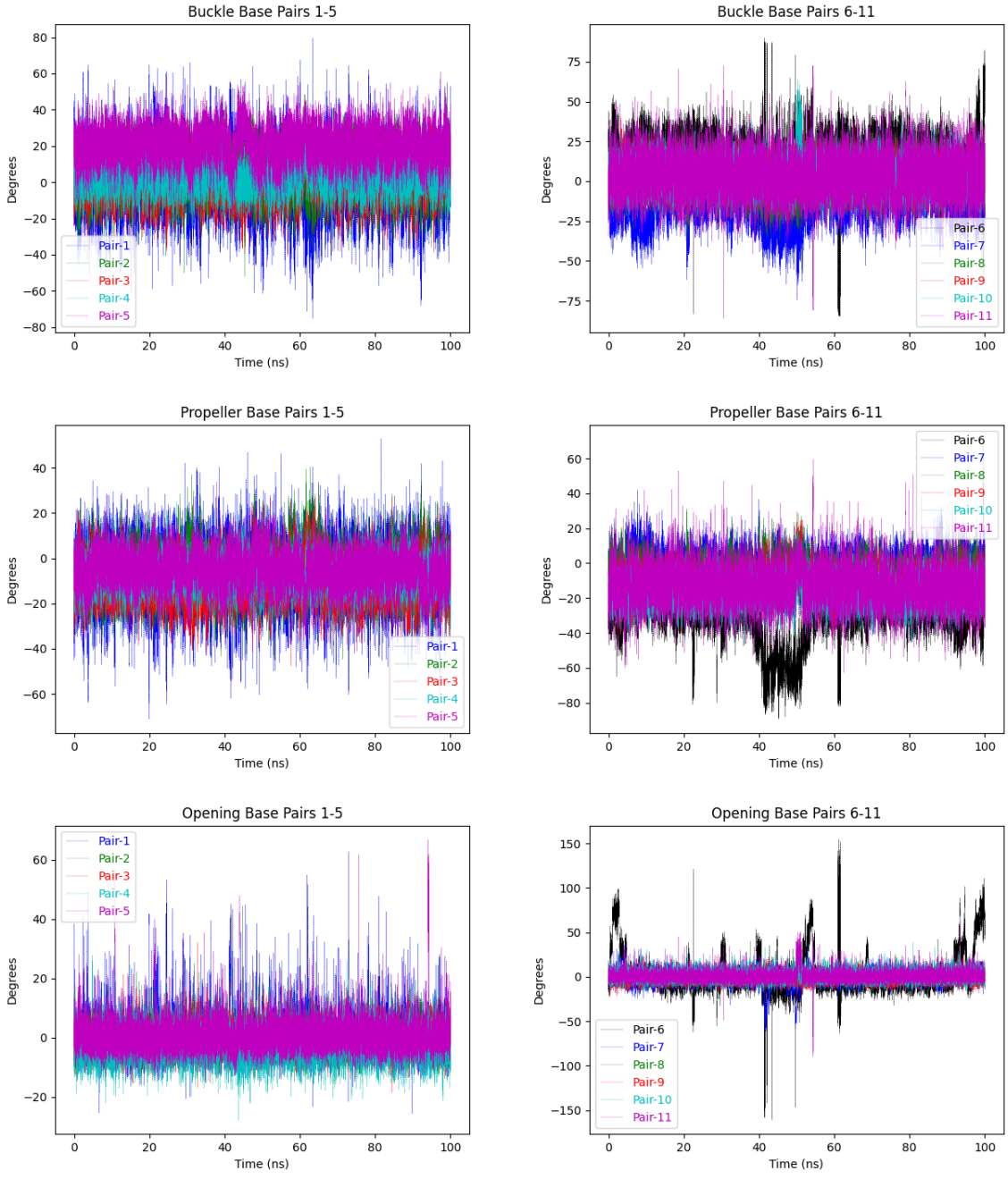
(6.151) DB[a,i]P-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



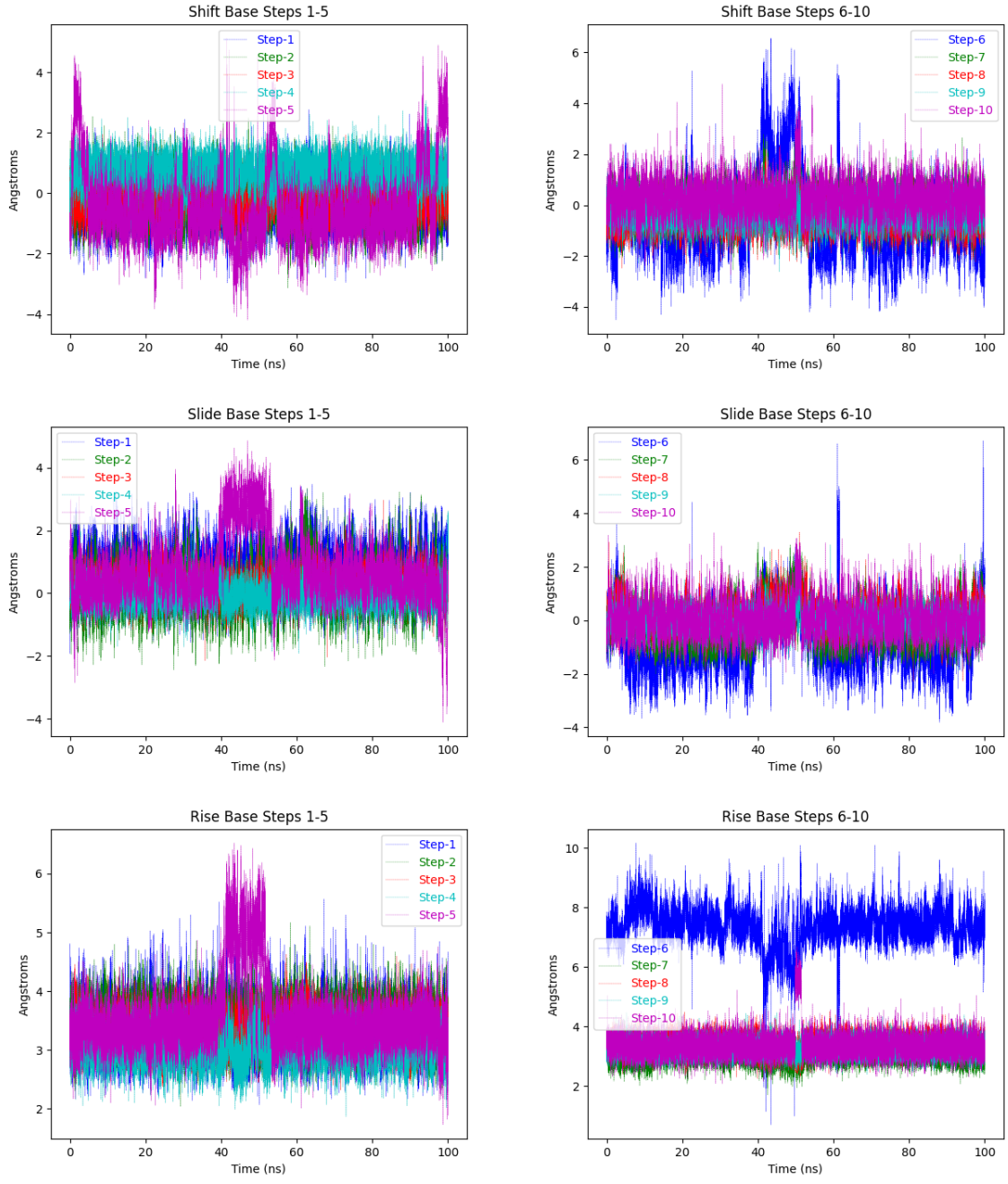
(6.152) DB[a,i]P-DNA: Refined major and minor groove trajectories



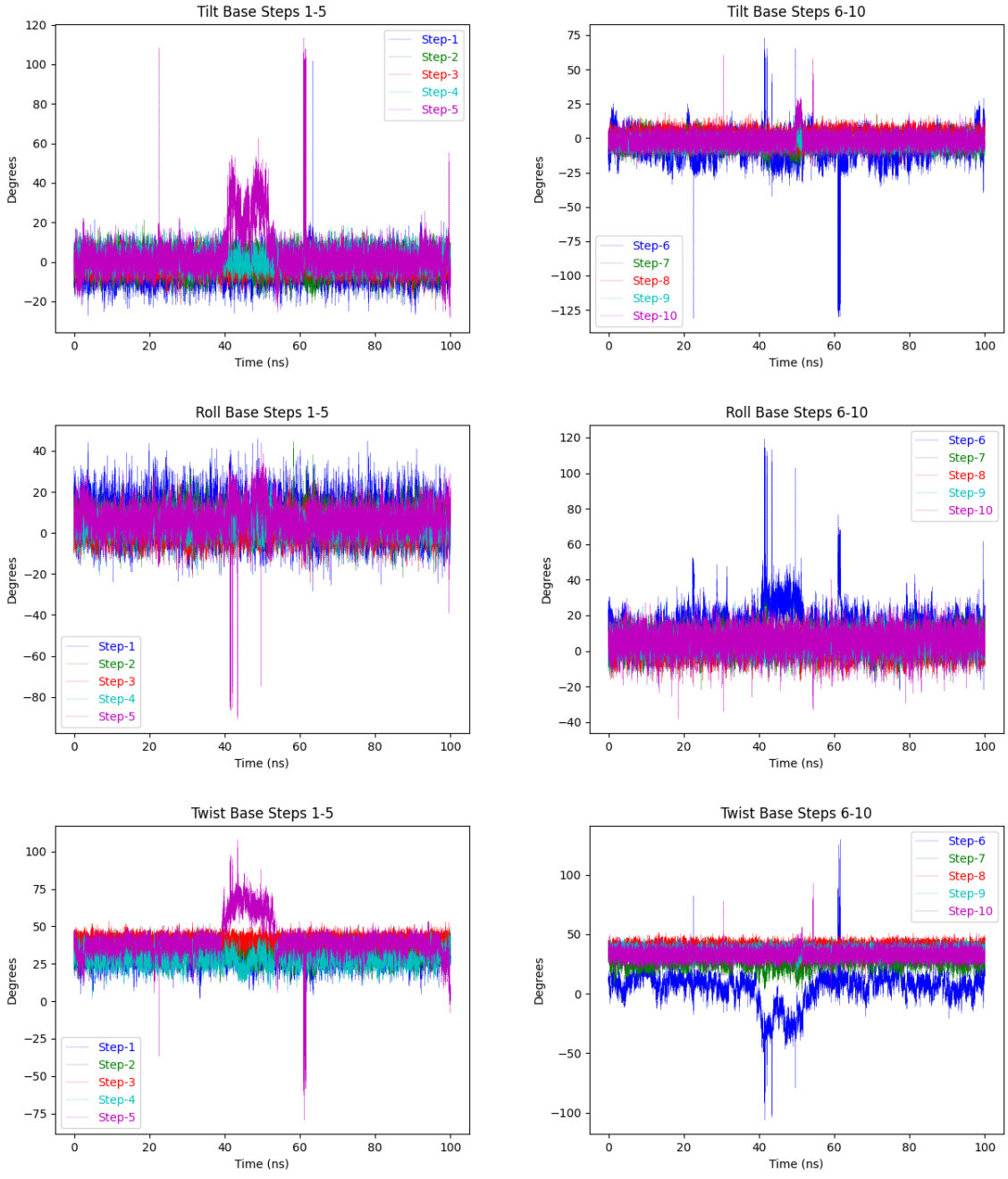
(6.153) DB[a,i]P-DNA: Base pair trajectories



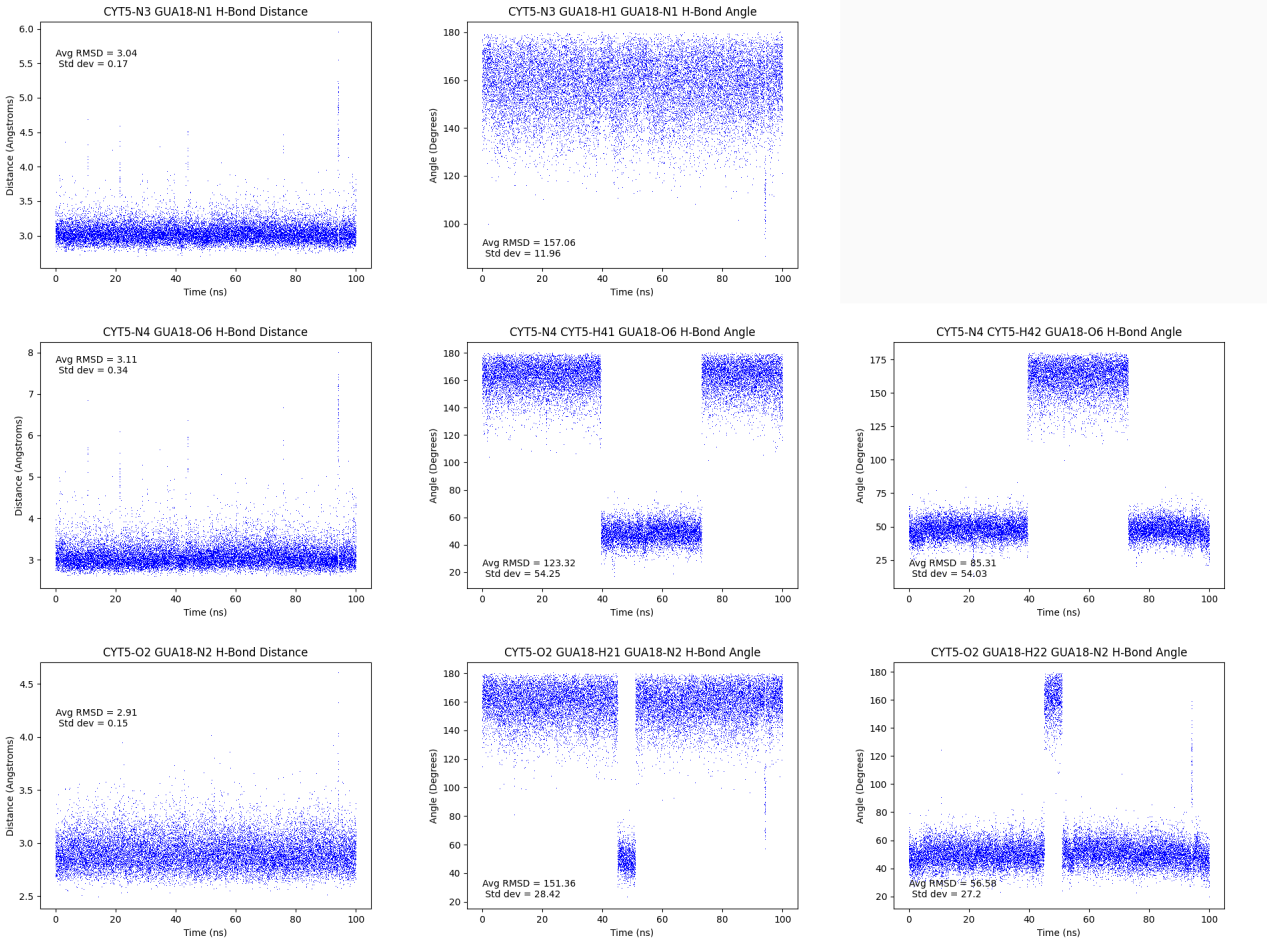
(6.154) DB[a,i]P-DNA: Base pair trajectories



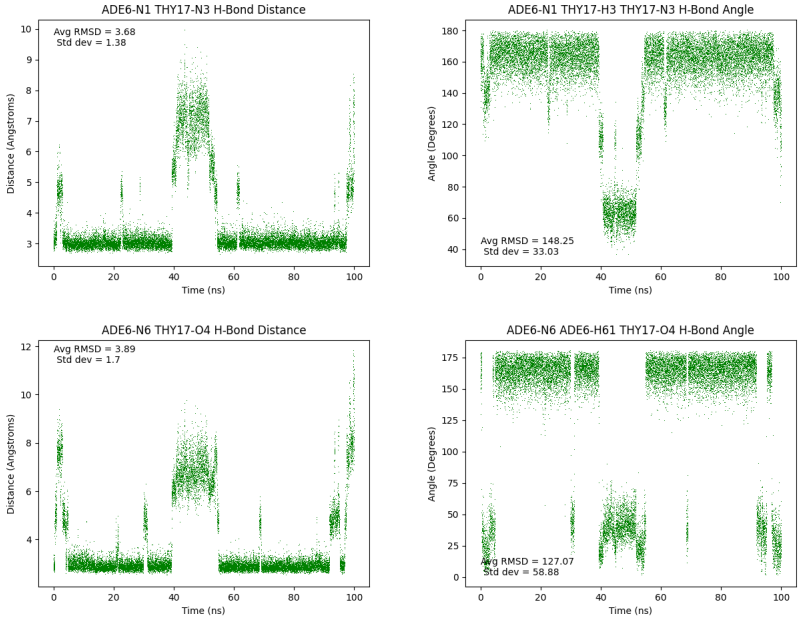
(6.155) DB[a,i]P-DNA: Base step trajectories



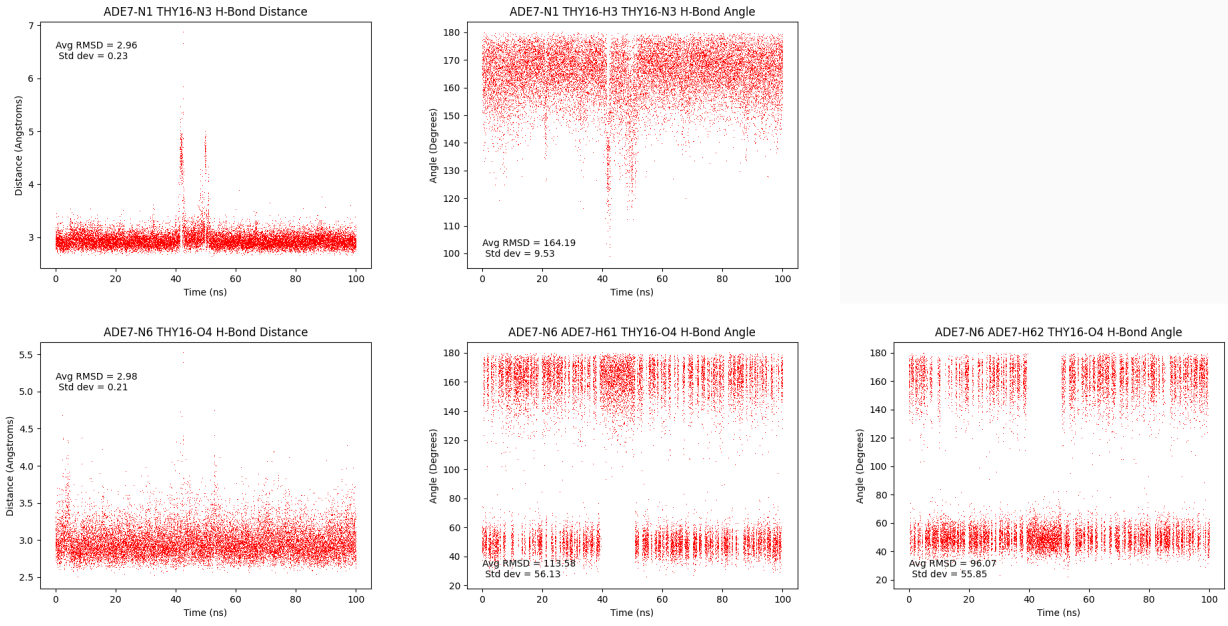
(6.156) DB[a,i]P-DNA: Base step trajectories



(6.157) DB[a,i]P-DNA: dC_5:dG_{18} hydrogen bond trajectories

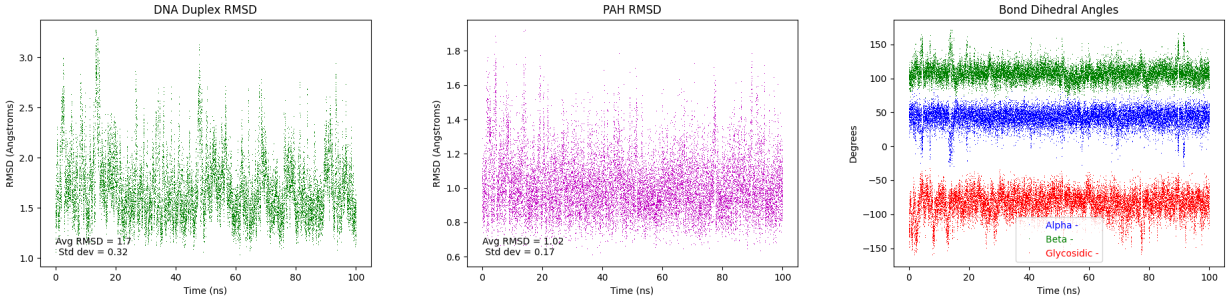


(6.158) DB[a,i]P-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

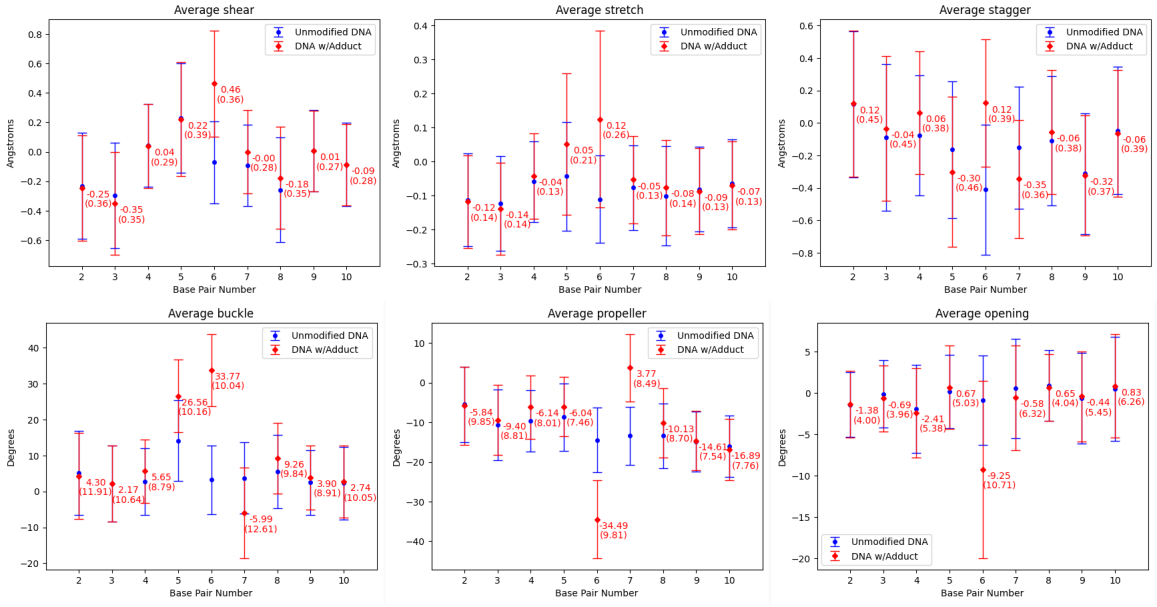


(6.159) DB[a,i]P-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

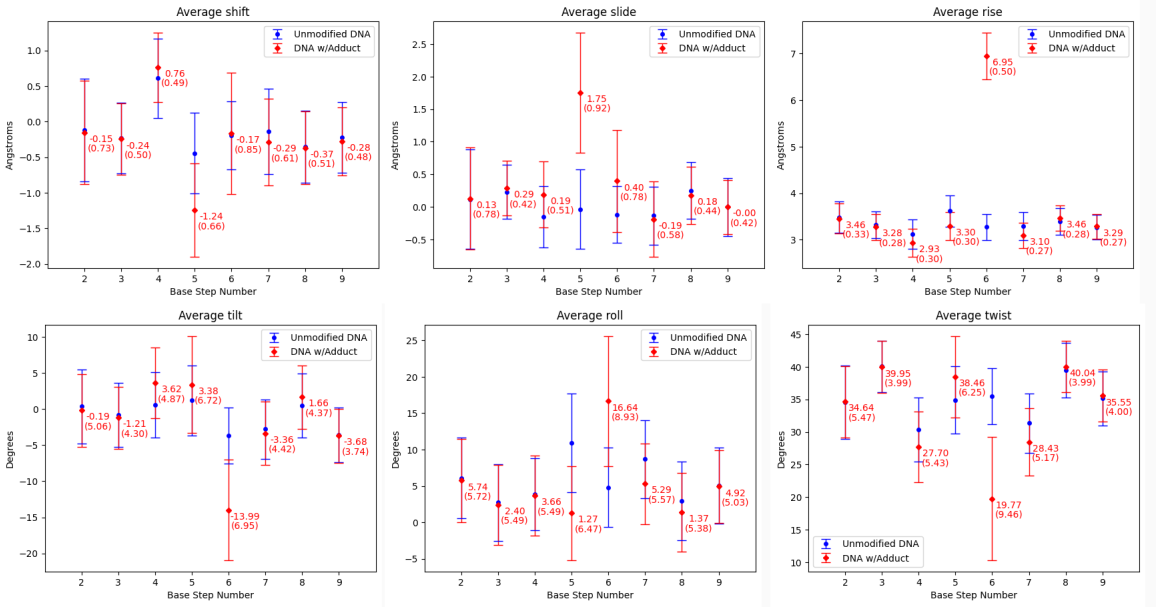
6.2.1.11 B[a]A-DNA



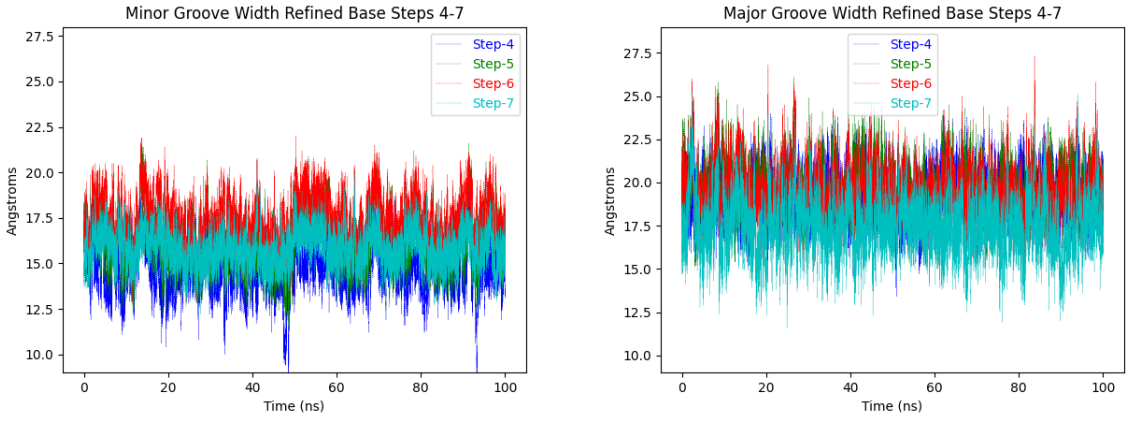
(6.160) B[a]A-DNA duplex RMSD; PAH RMSD; α, β, χ trajectories



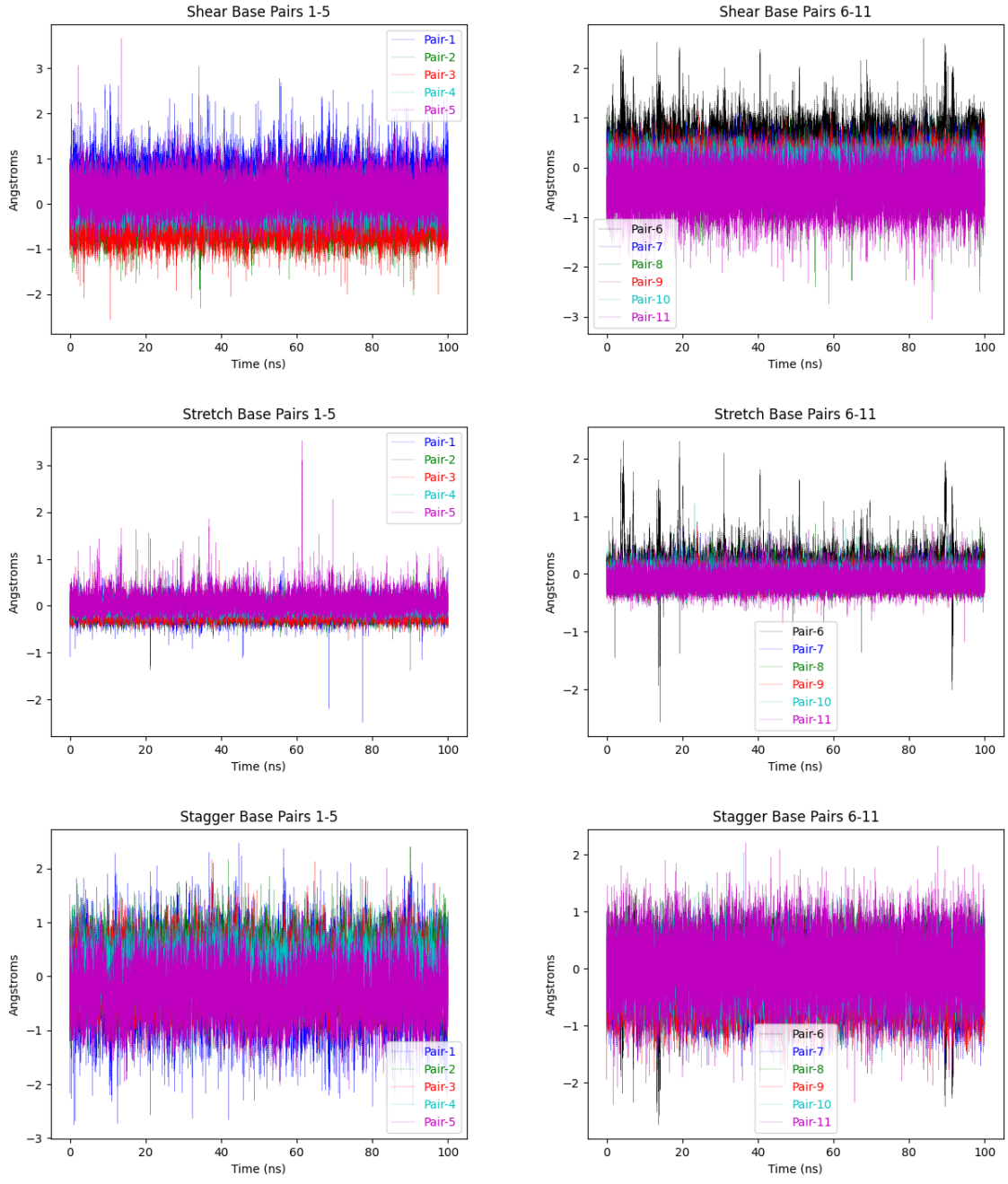
(6.161) B[a]A-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



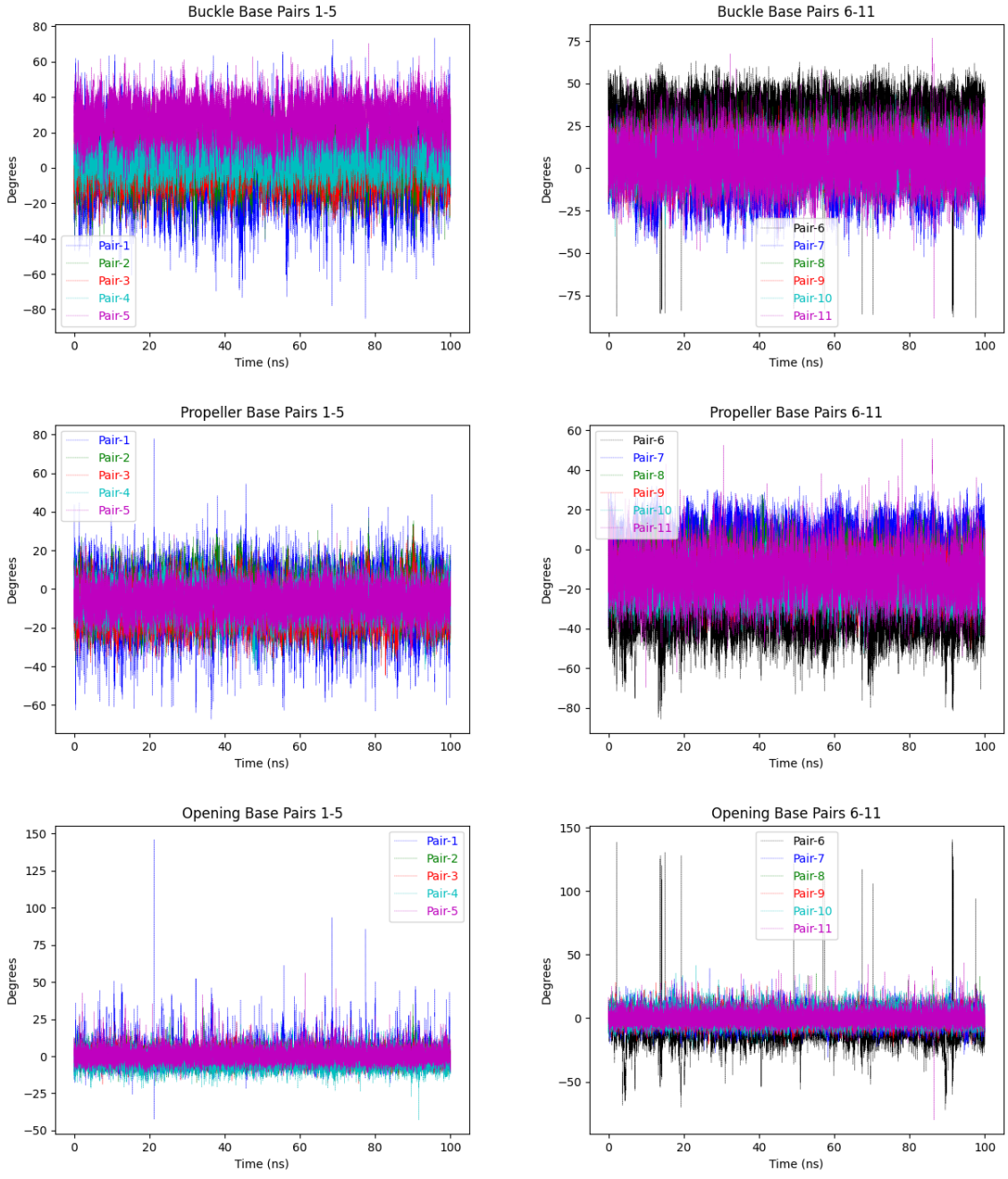
(6.162) B[a]A-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



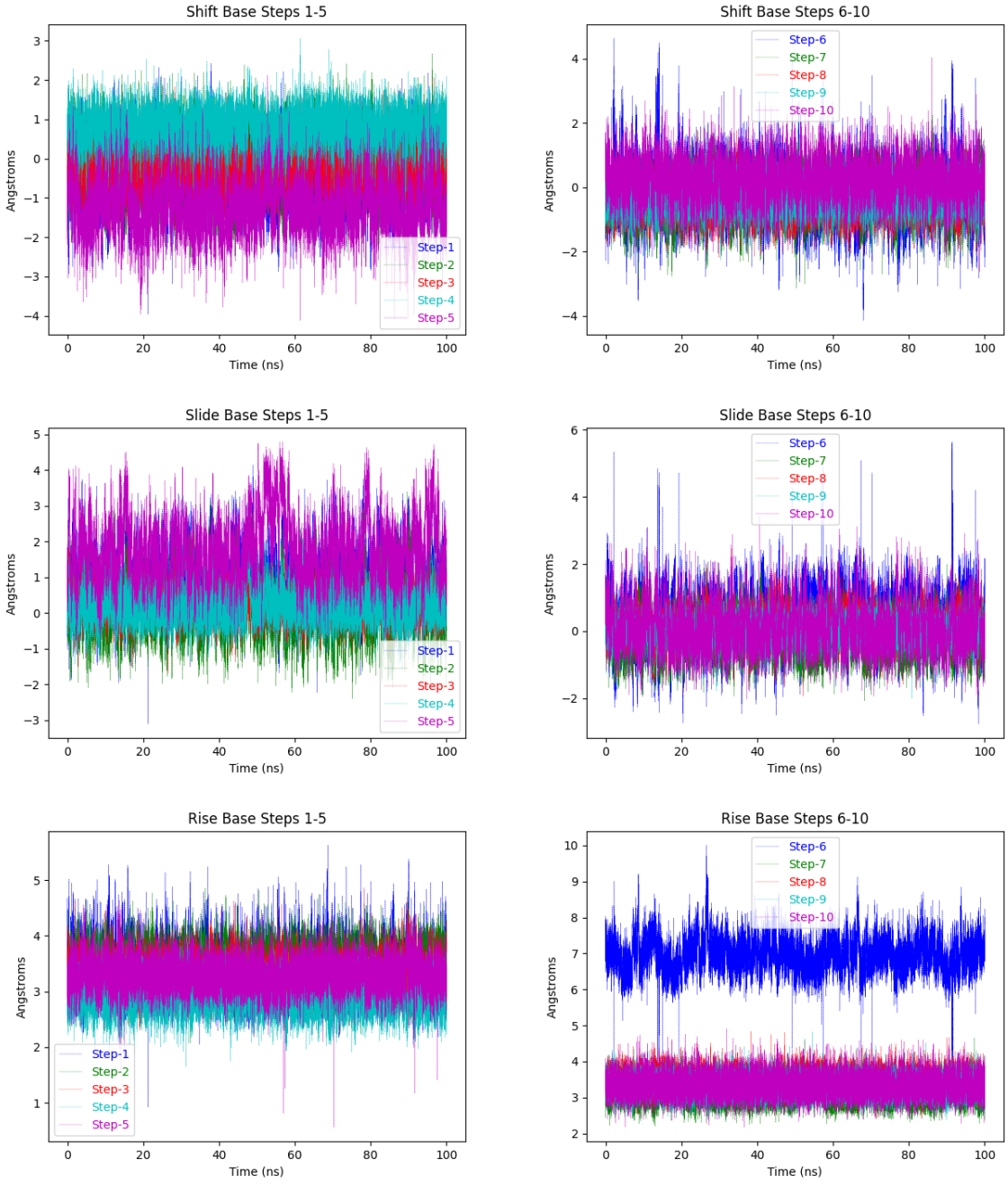
(6.163) B[a]A-DNA: Refined major and minor groove trajectories



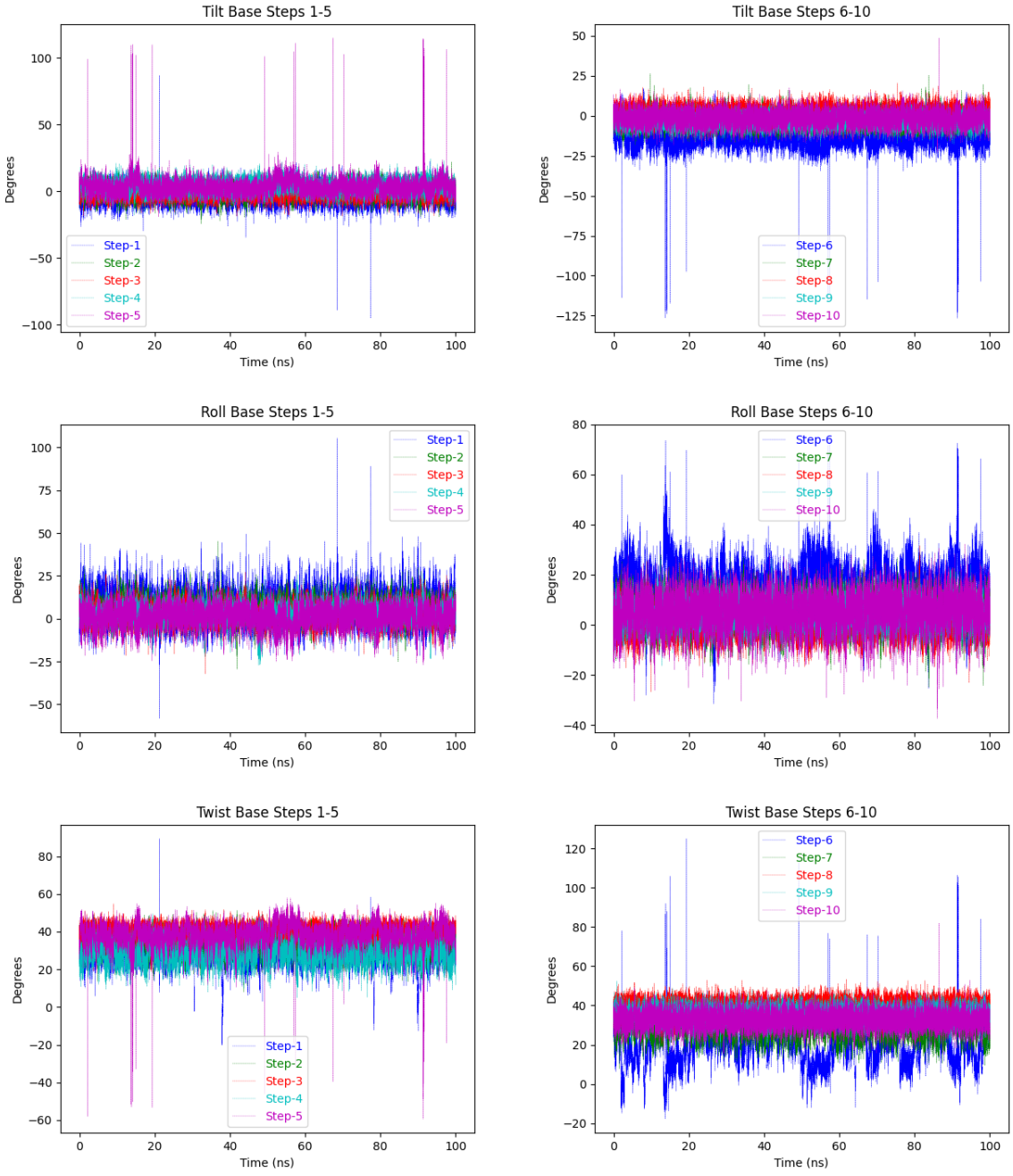
(6.164) B[a]A-DNA: Base pair trajectories



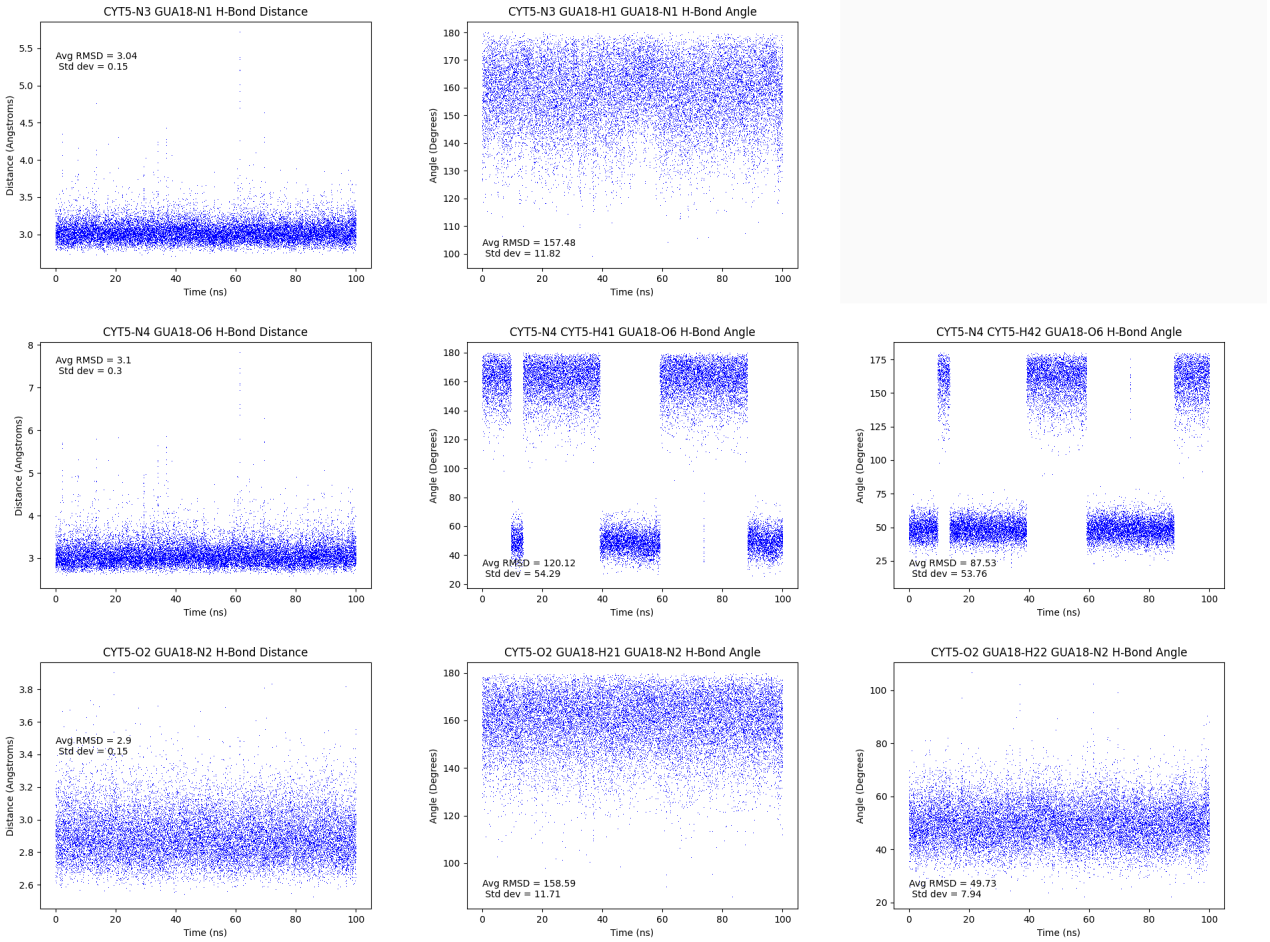
(6.165) B[a]A-DNA: Base pair trajectories



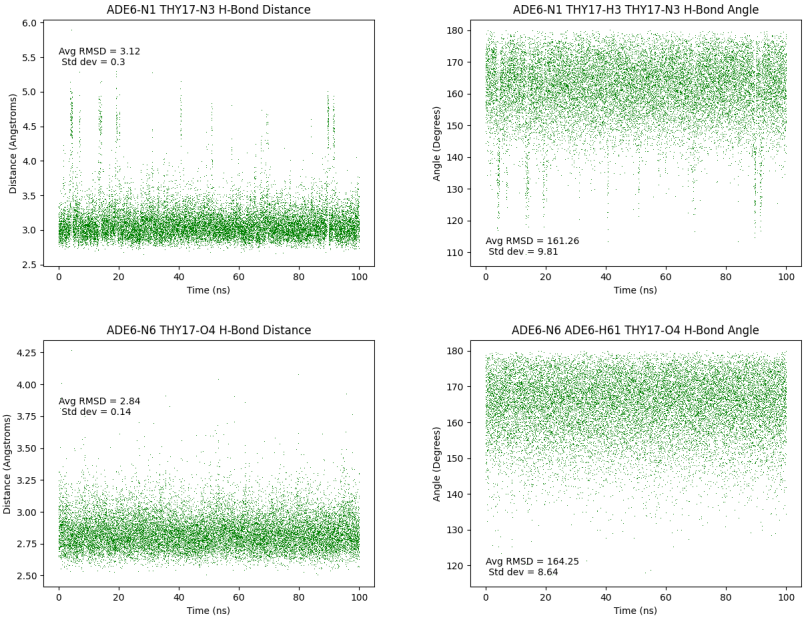
(6.166) B[a]A-DNA: Base step trajectories



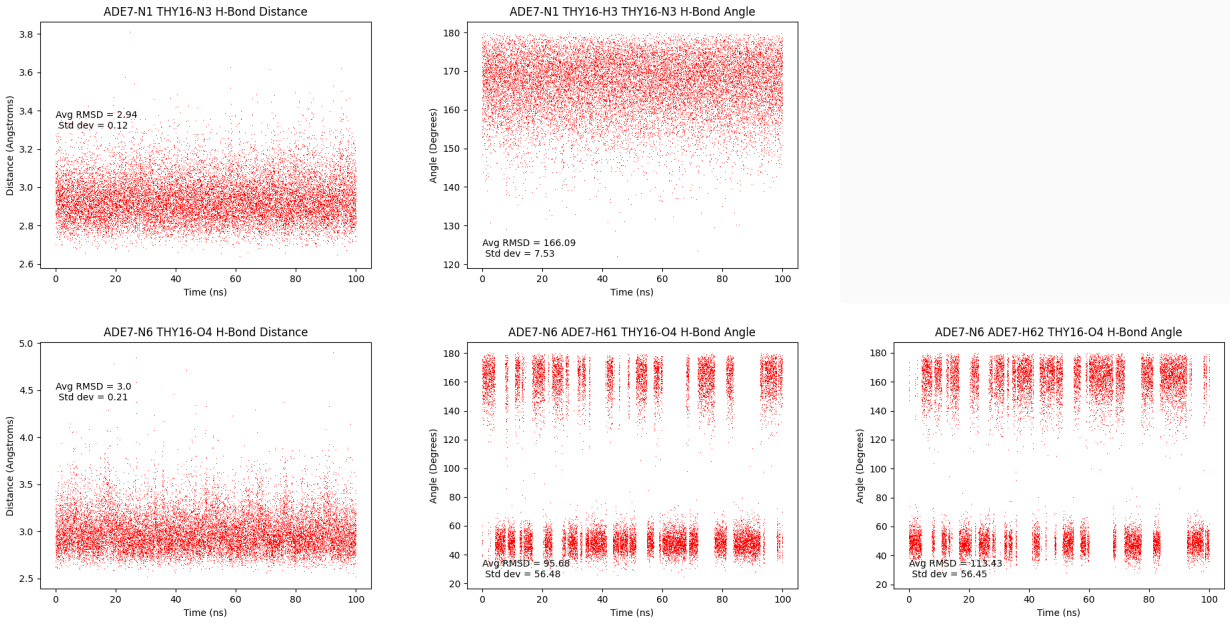
(6.167) B[a]A-DNA: Base step trajectories



(6.168) B[a]A-DNA: dC_5:dG_{18} hydrogen bond trajectories

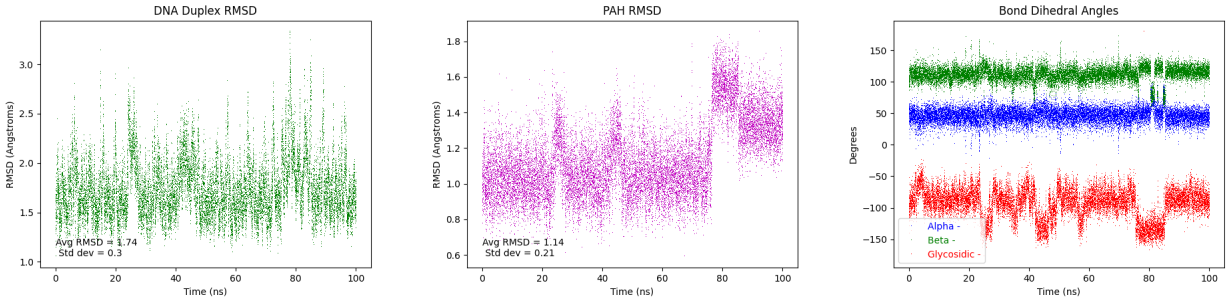


(6.169) B[a]A-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

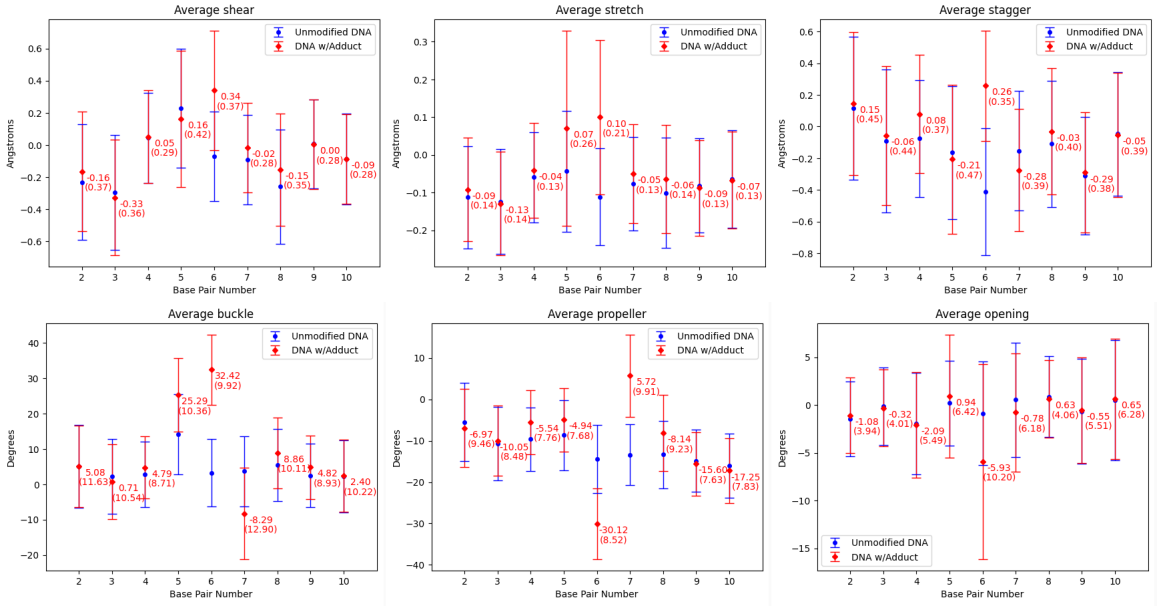


(6.170) B[a]A-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

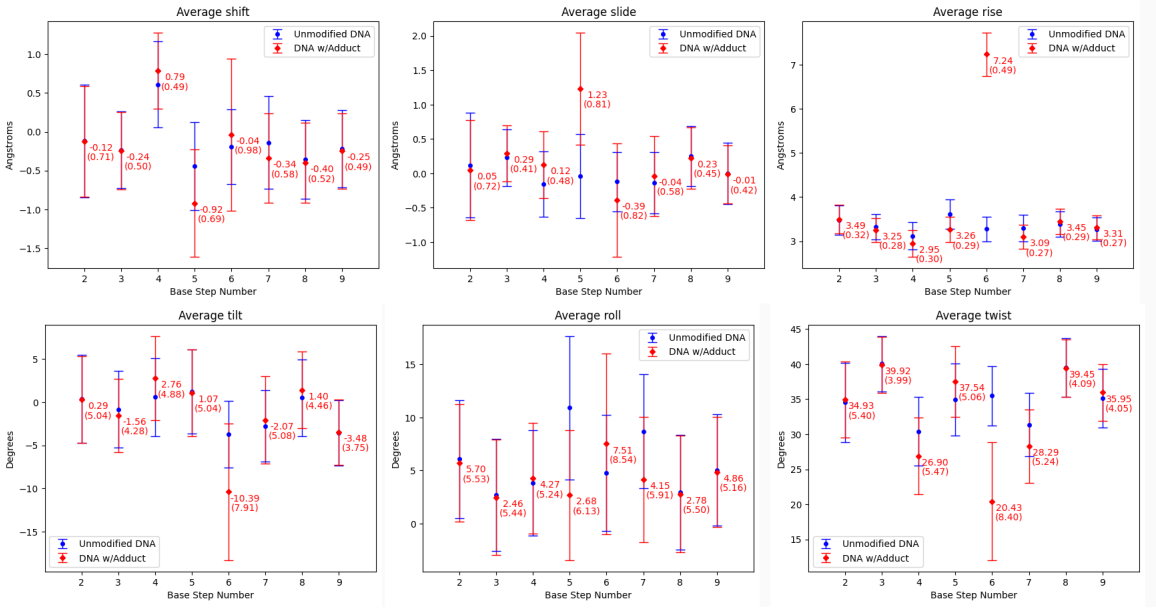
6.2.1.12 DB[a,c]A-DNA



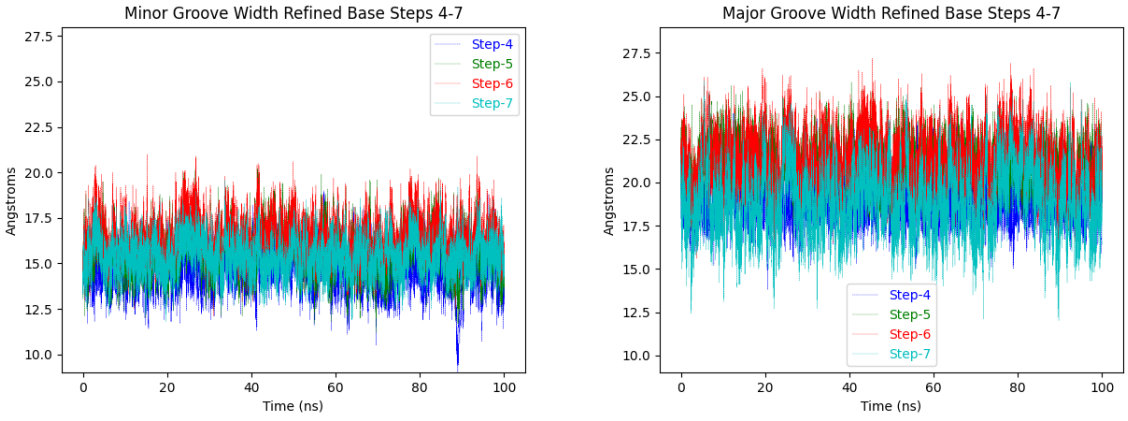
(6.171) DB[a,c]A-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



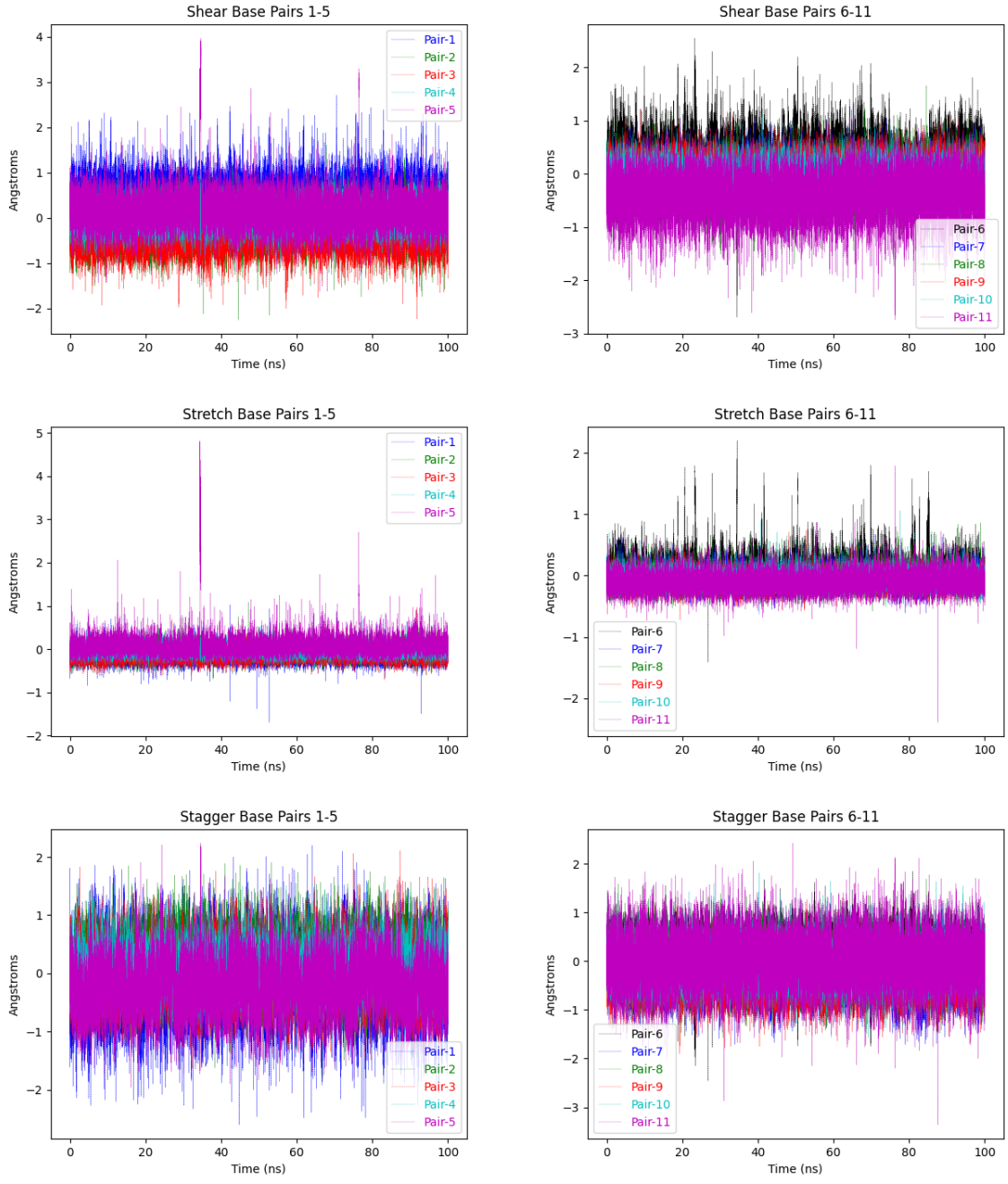
(6.172) DB[a,c]A-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



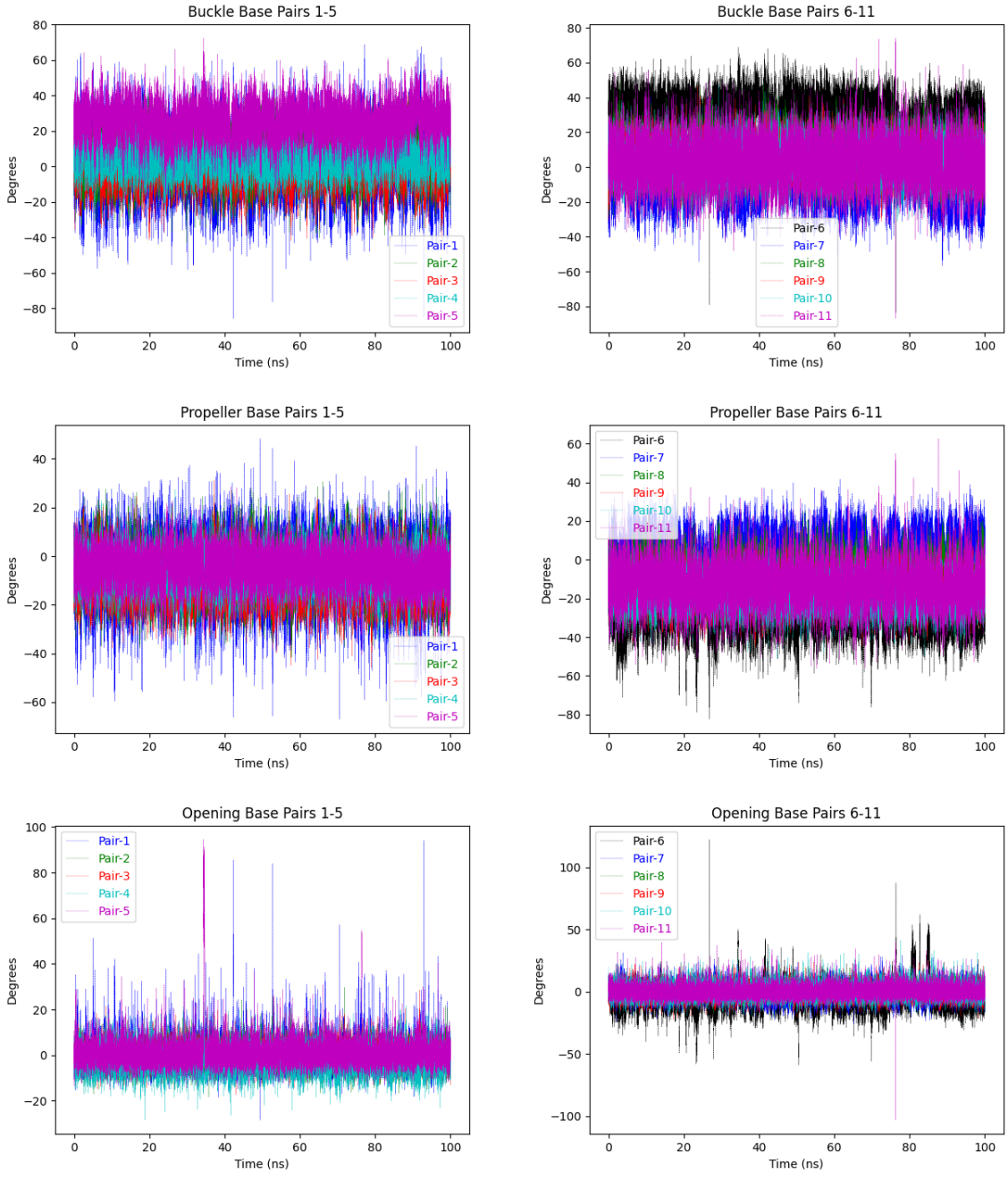
(6.173) DB[a,c]A-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



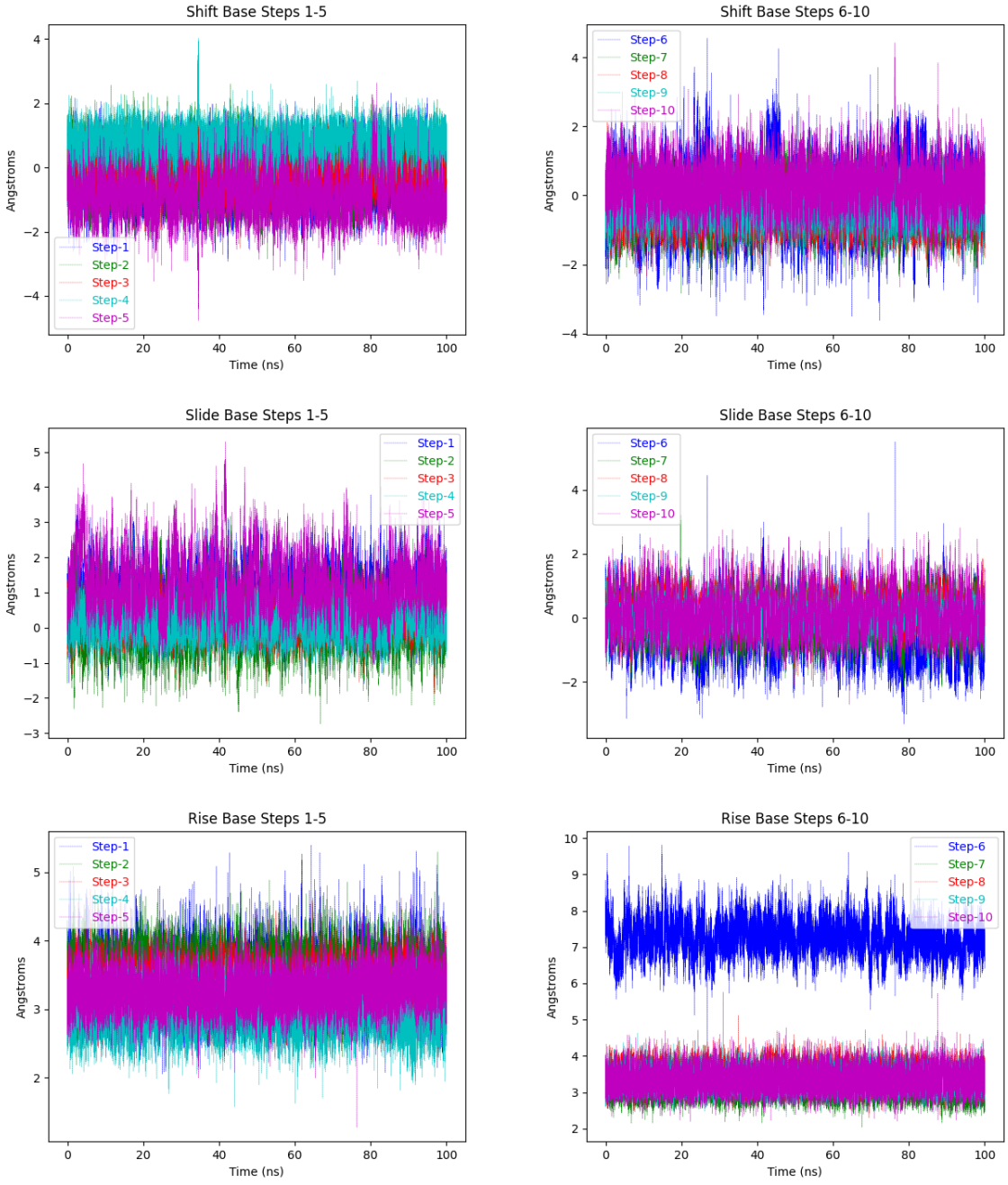
(6.174) DB[a,c]A-DNA: Refined major and minor groove trajectories



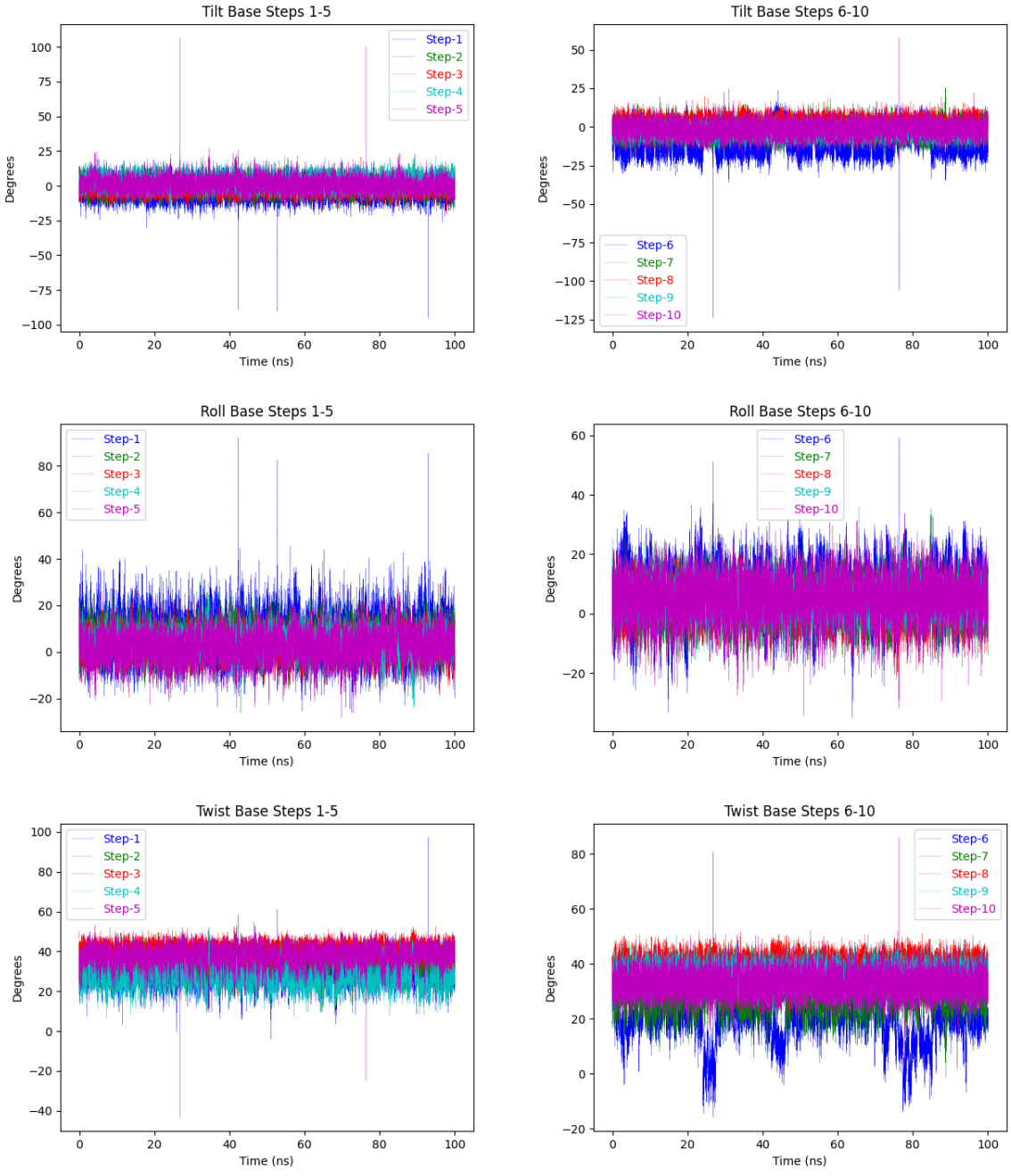
(6.175) DB[a,c]A-DNA: Base pair trajectories



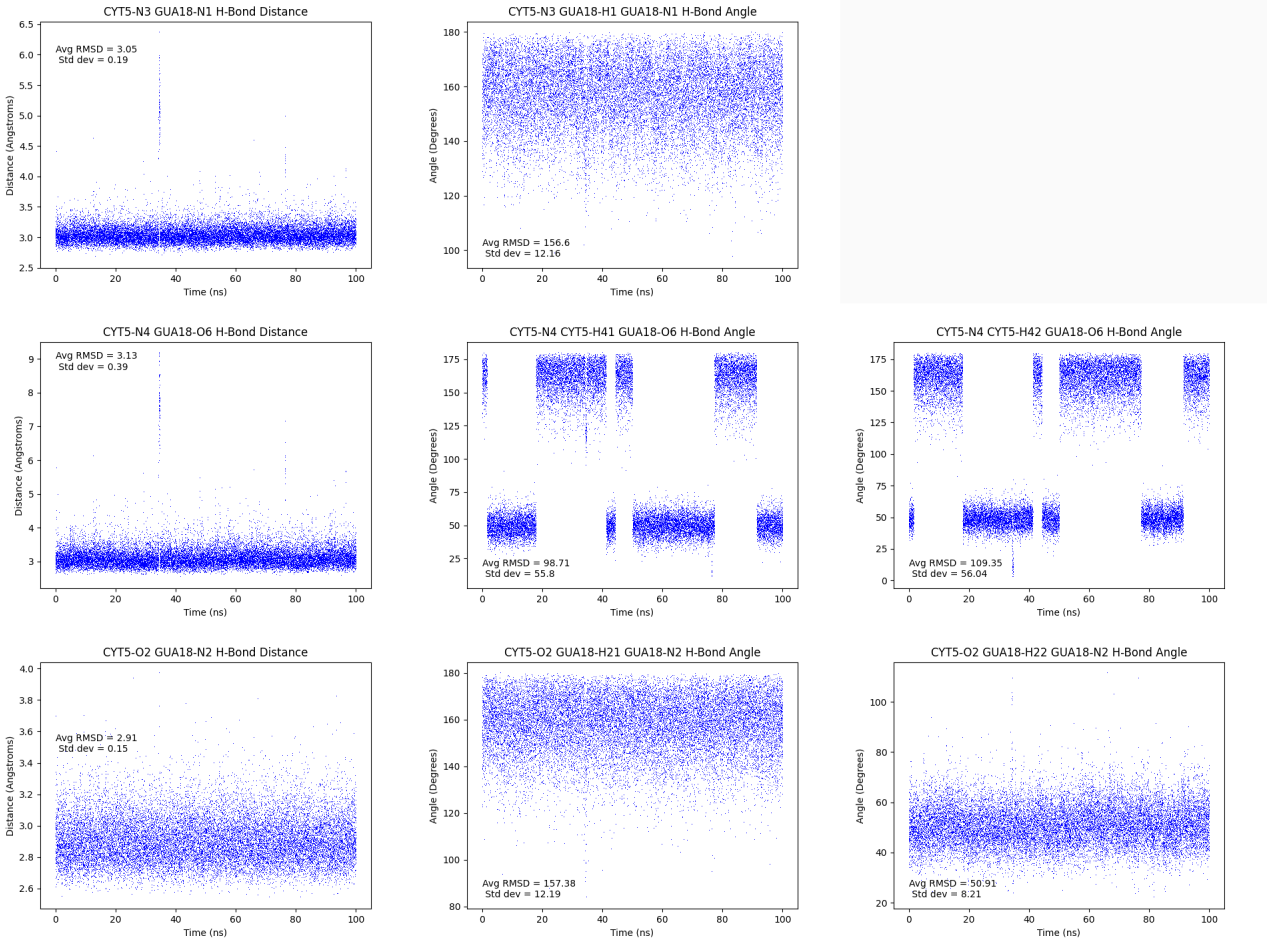
(6.176) DB[a,c]A-DNA: Base pair trajectories



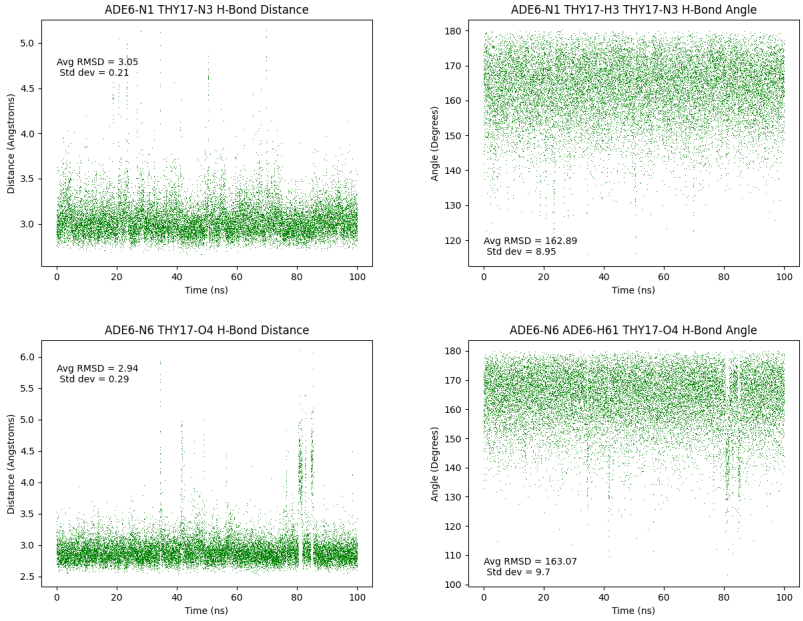
(6.177) DB[a,c]A-DNA: Base step trajectories



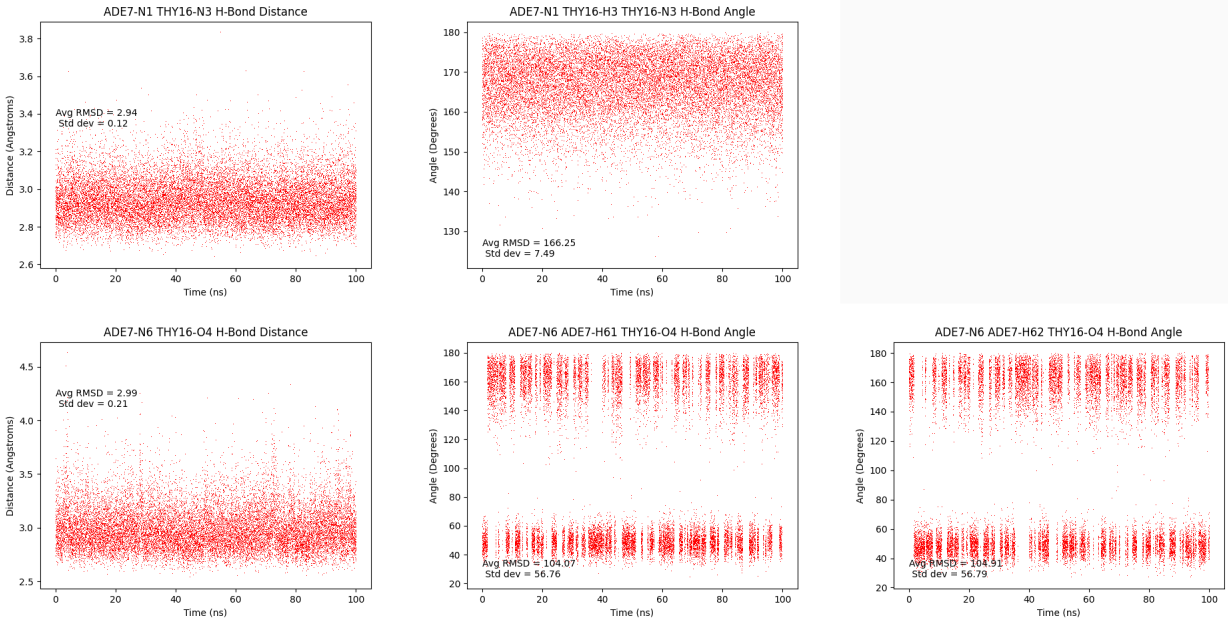
(6.178) DB[a,c]A-DNA: Base step trajectories



(6.179) DB[a,c]A-DNA: dC_5:dG_{18} hydrogen bond trajectories

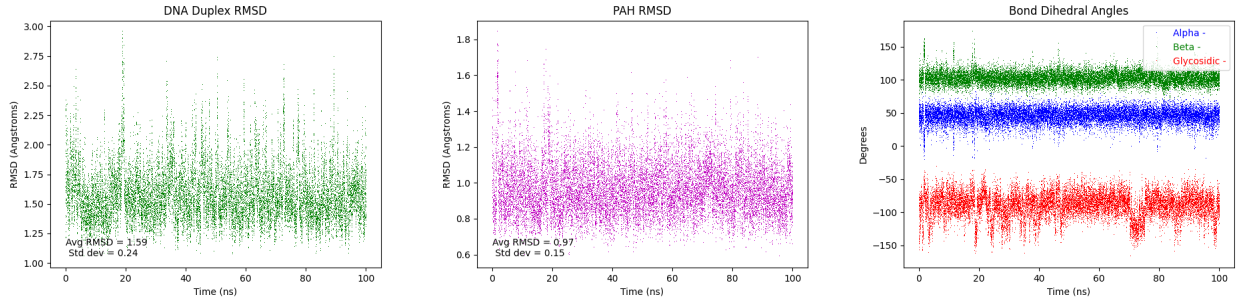


(6.180) DB[a,c]A-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

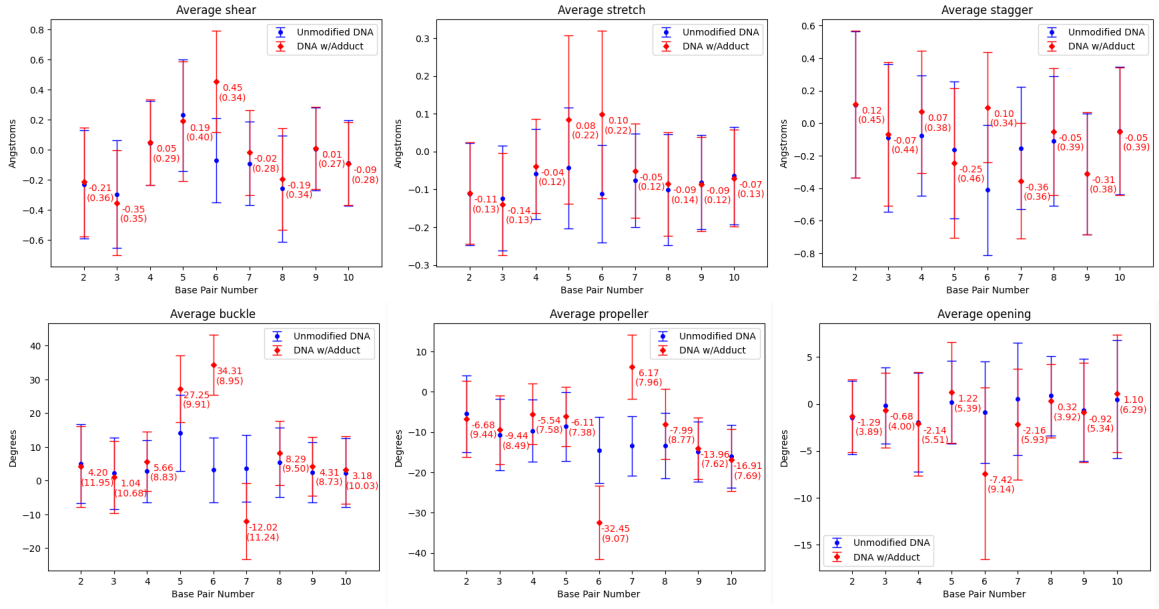


(6.181) DB[a,c]A-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

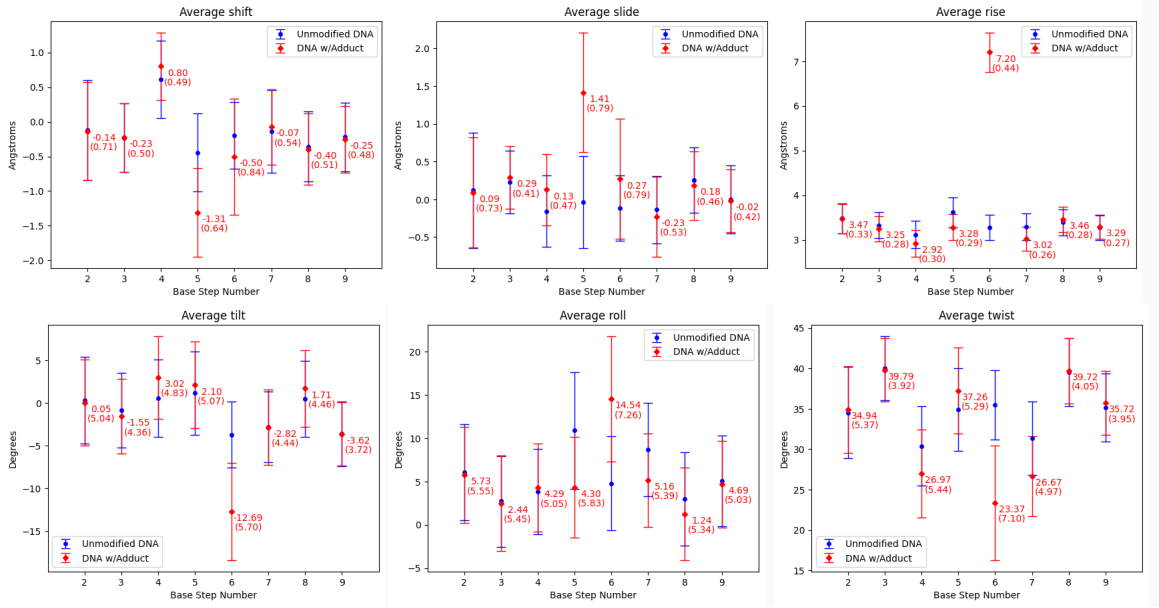
6.2.1.13 DB[a,h]A-DNA



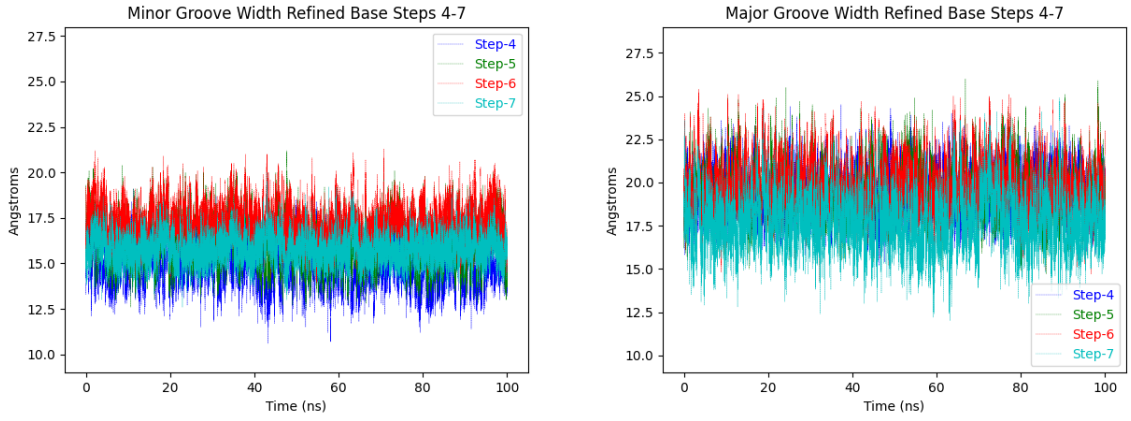
(6.182) DB[a,h]A-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



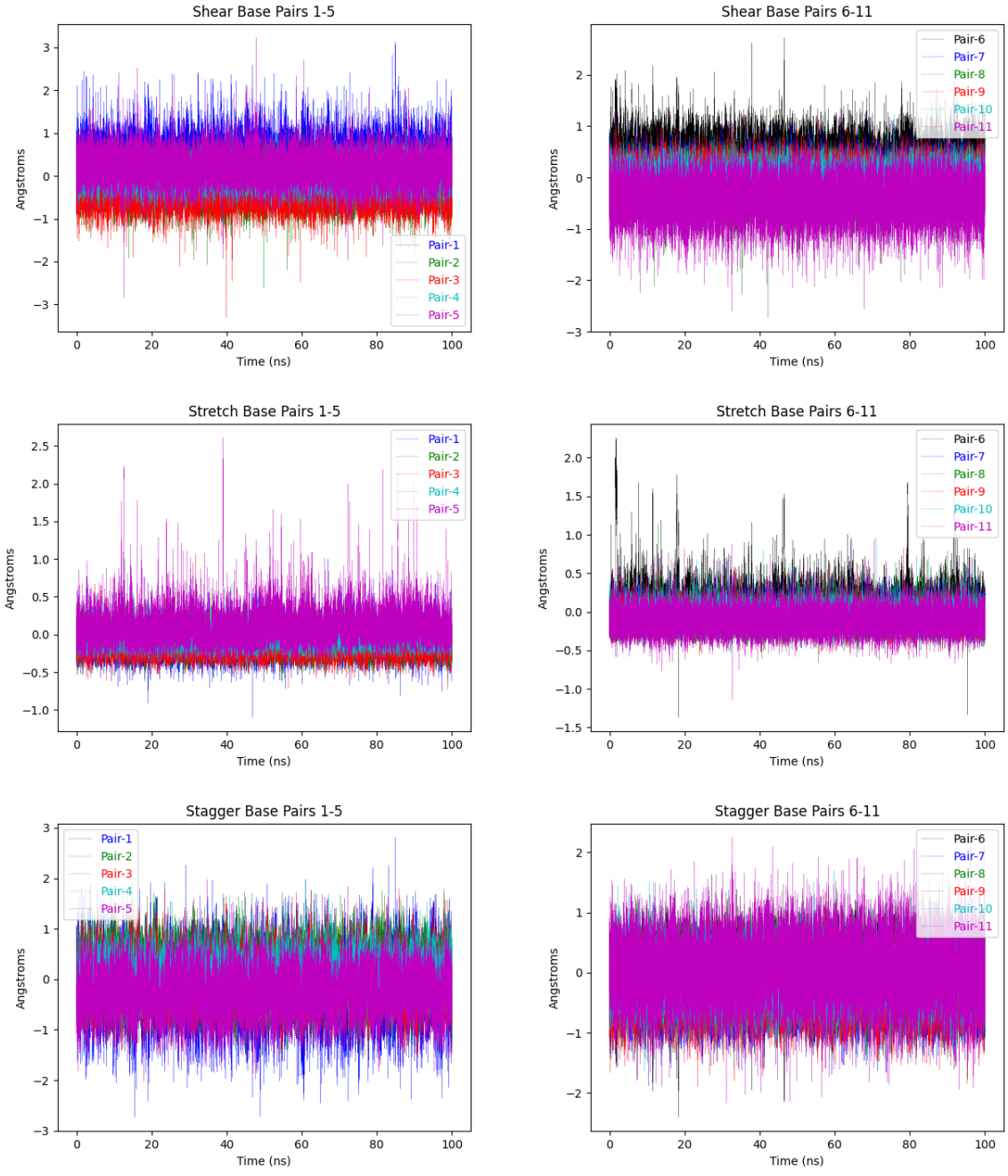
(6.183) DB[a,h]A-DNA: Average values of base pair rigid-body parameter, standard deviation in parenthesis.



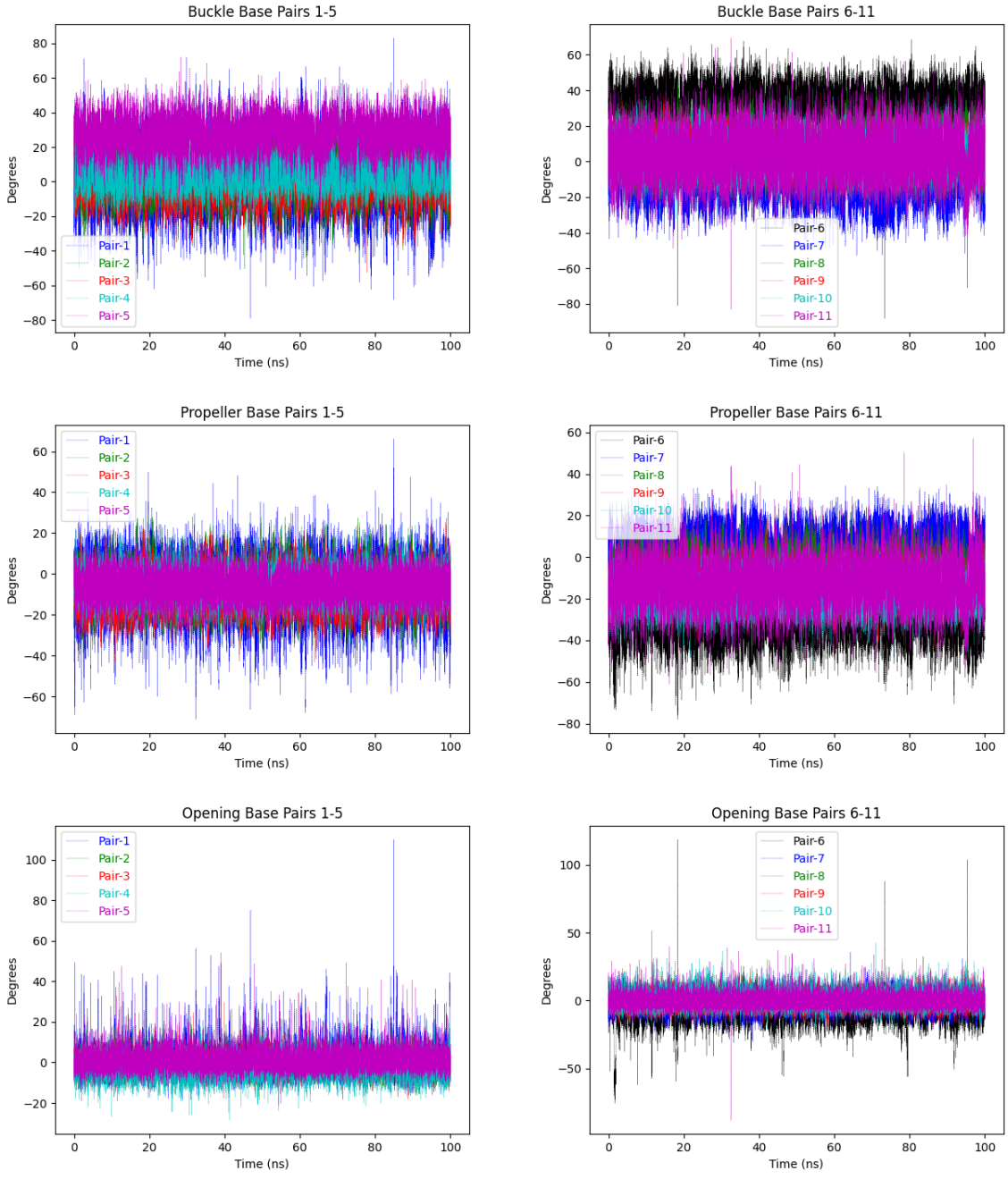
(6.184) DB[a,h]A-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



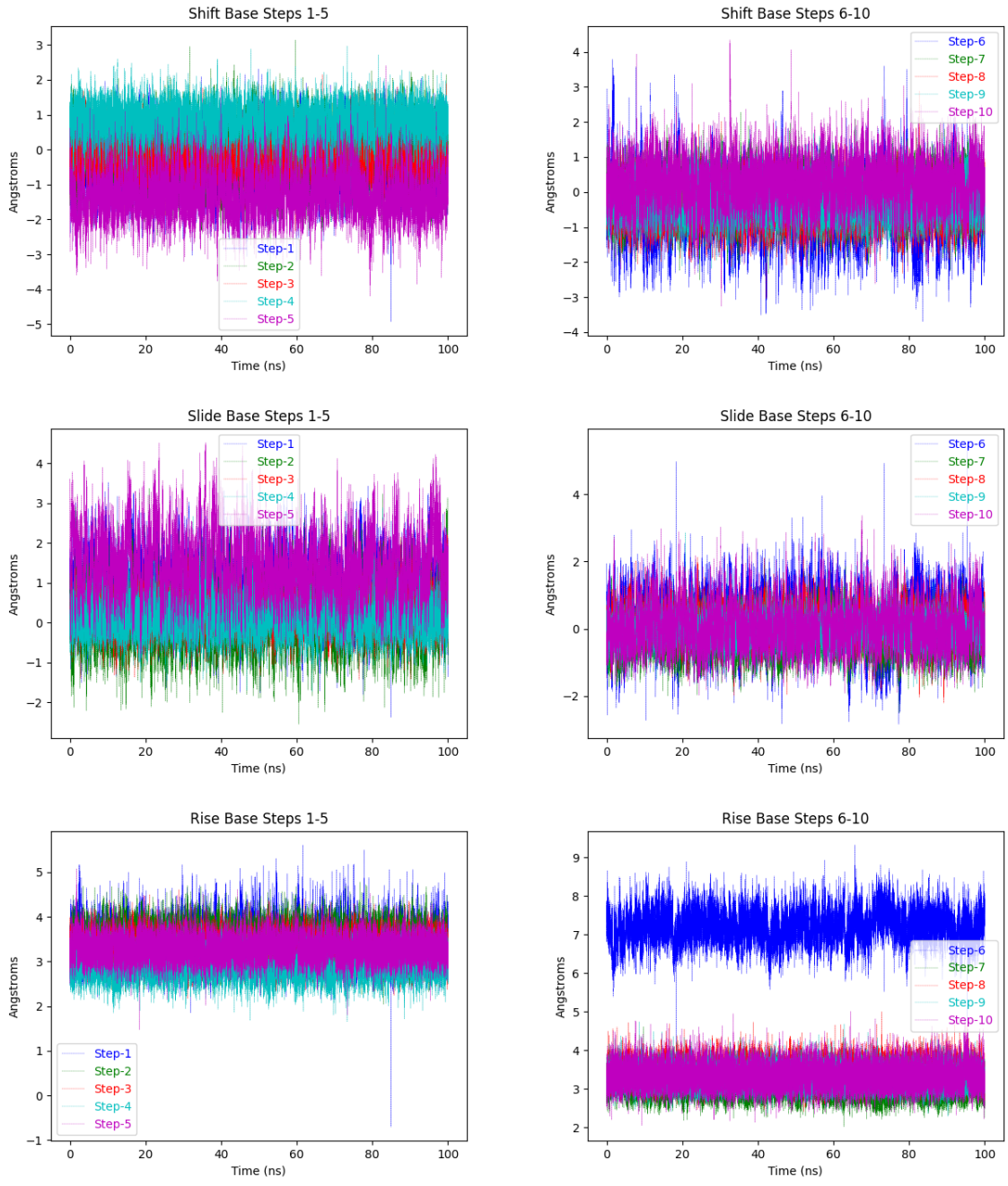
(6.185) DB[a,h]A-DNA: Refined major and minor groove trajectories



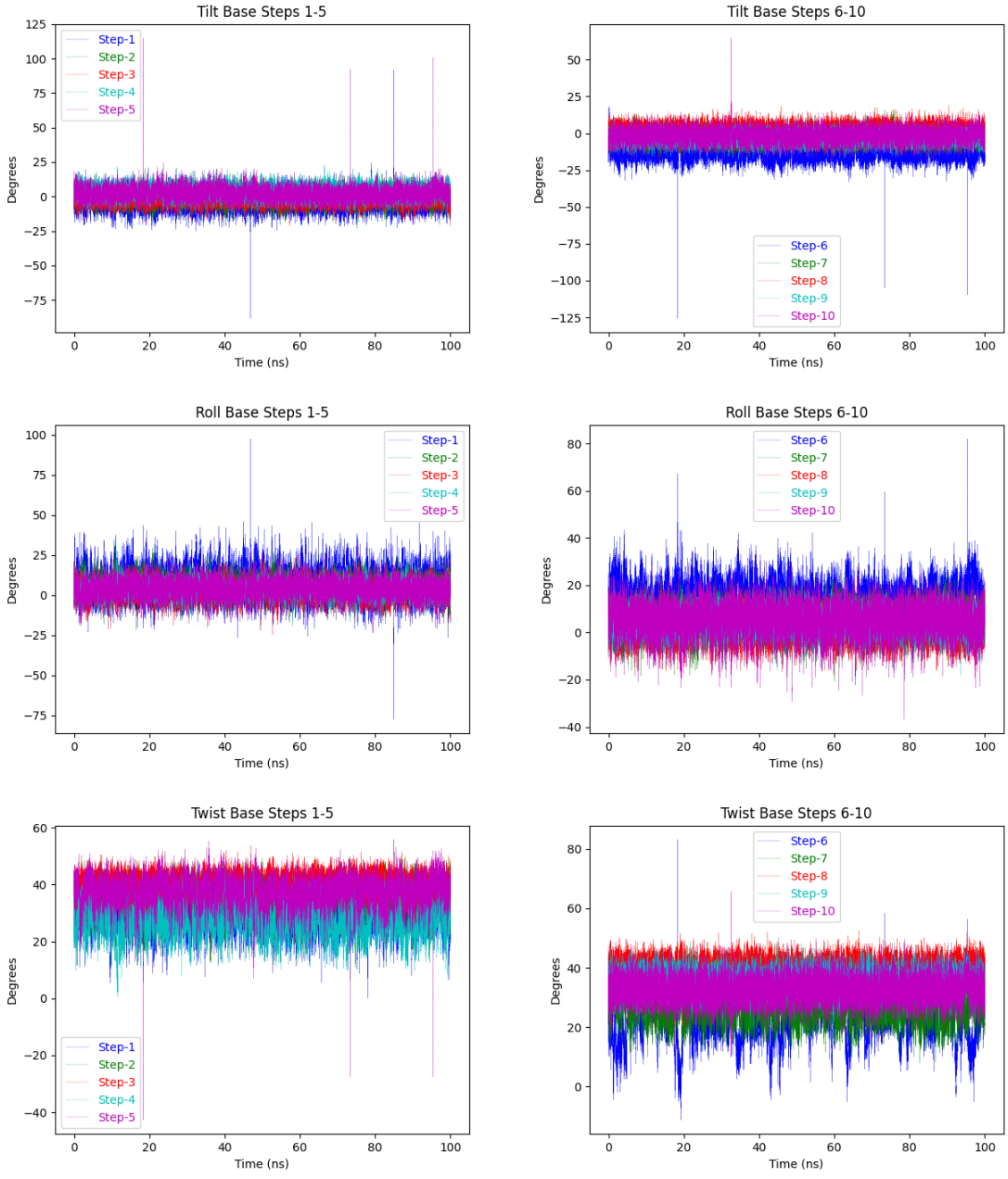
(6.186) DB[a,h]A-DNA: Base pair trajectories



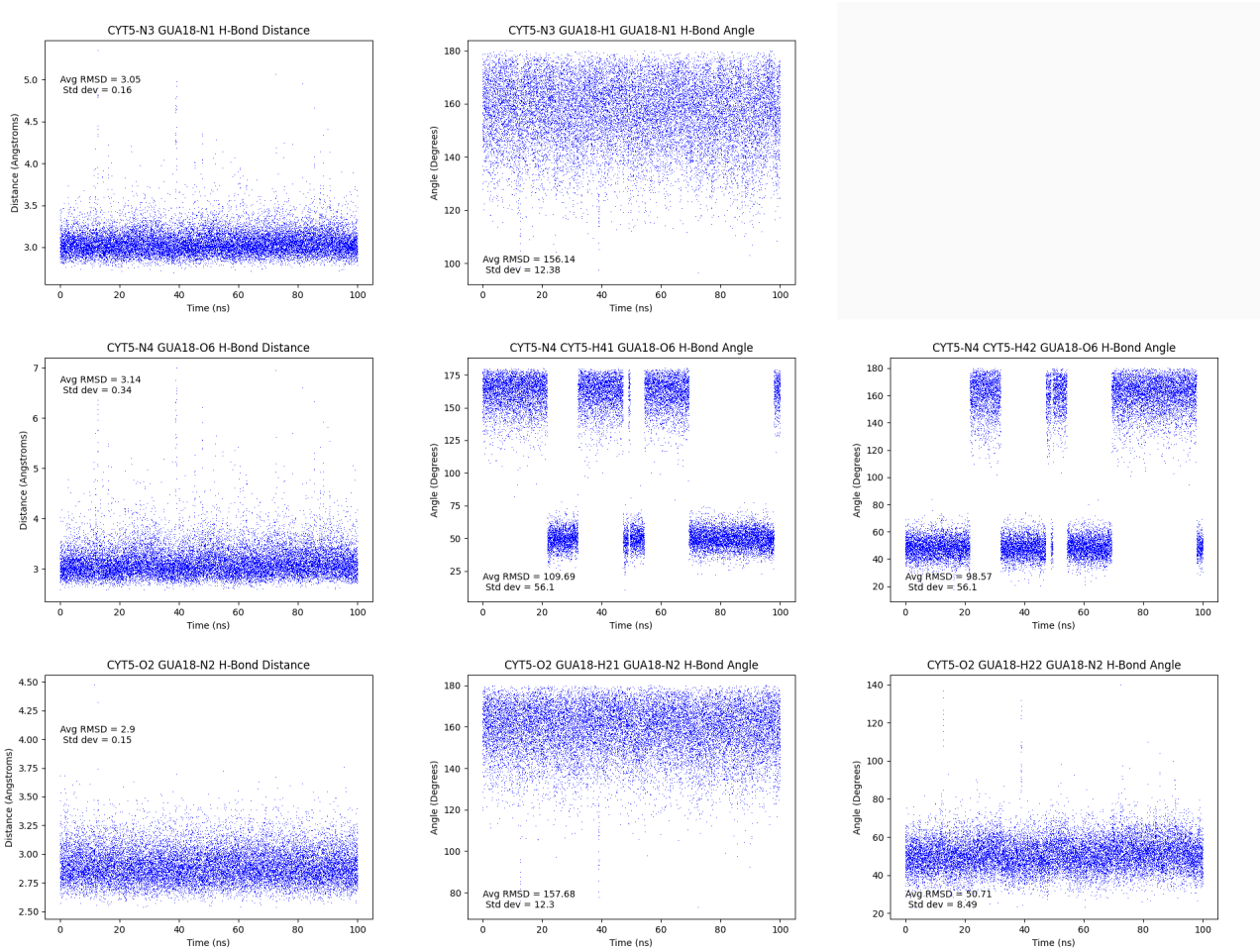
(6.187) DB[a,h]A-DNA: Base pair trajectories



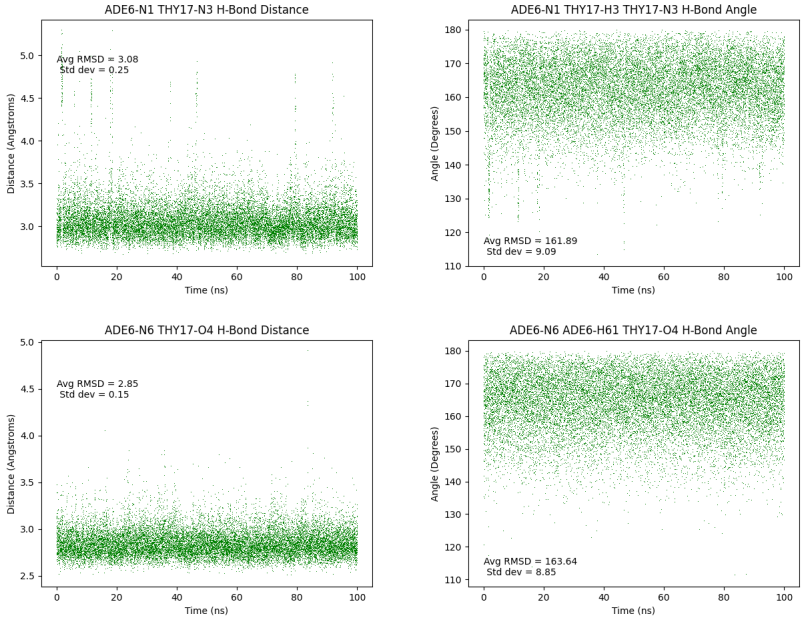
(6.188) DB[a,h]A-DNA: Base step trajectories



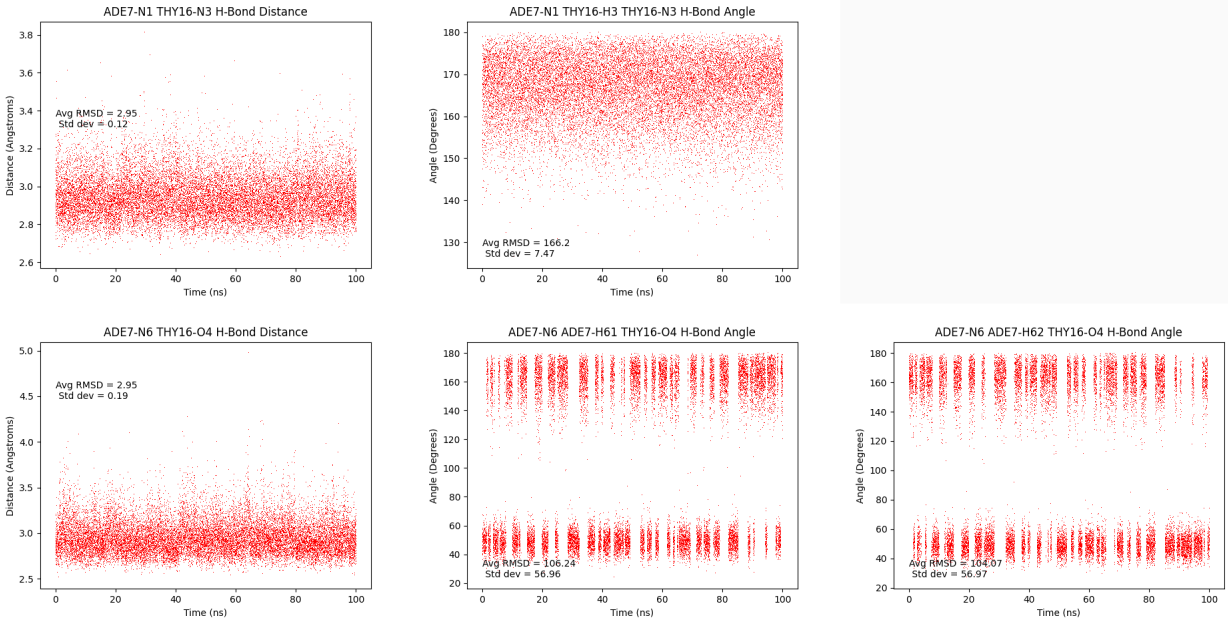
(6.189) DB[a,h]A-DNA: Base step trajectories



(6.190) DB[a,h]A-DNA: dC_5:dG_{18} hydrogen bond trajectories

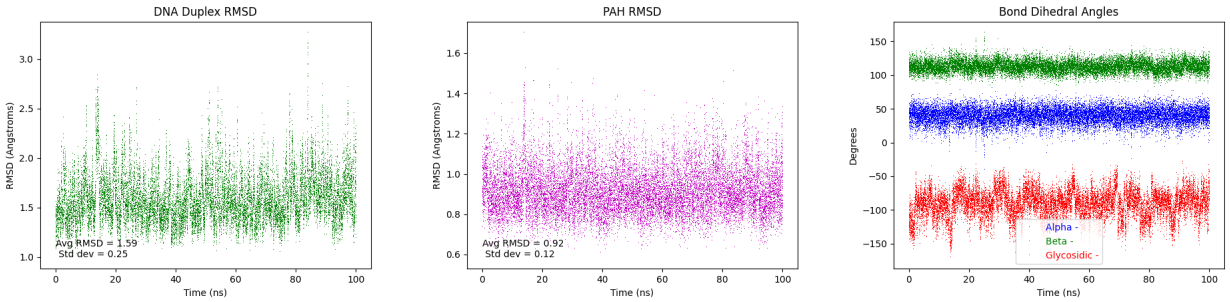


(6.191) DB[a,h]A-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

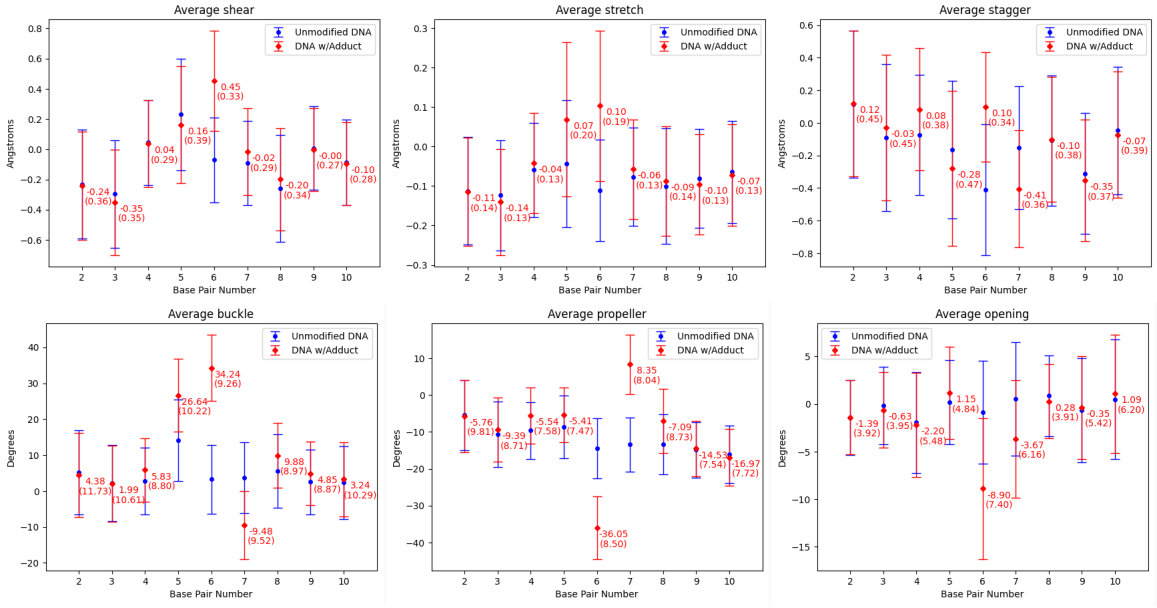


(6.192) DB[a,h]A-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

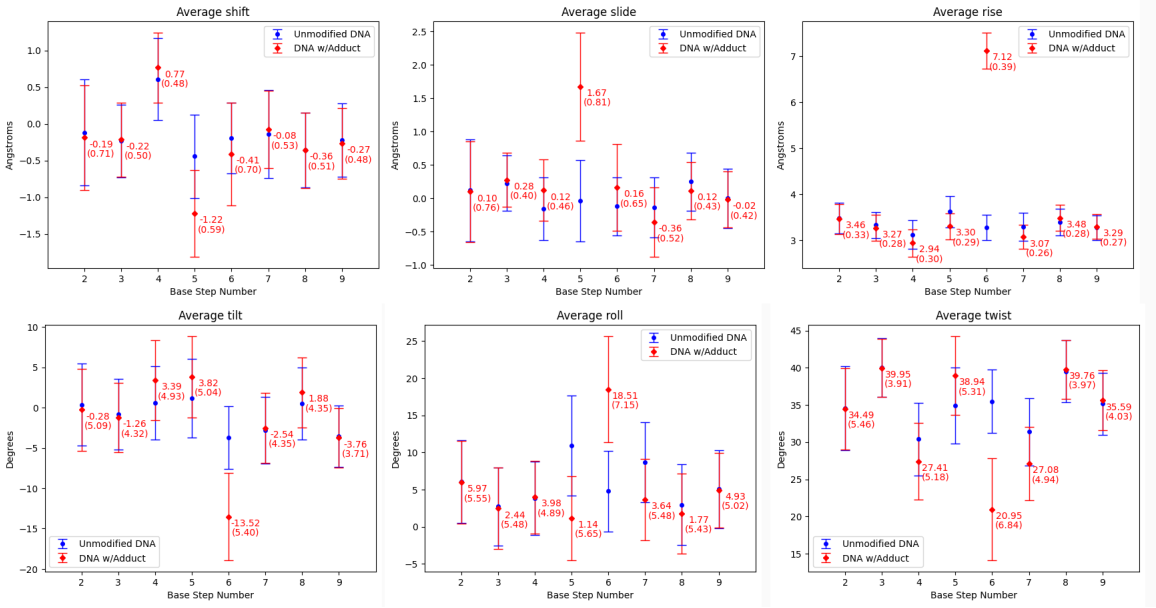
6.2.1.14 DB[a,j]A-DNA



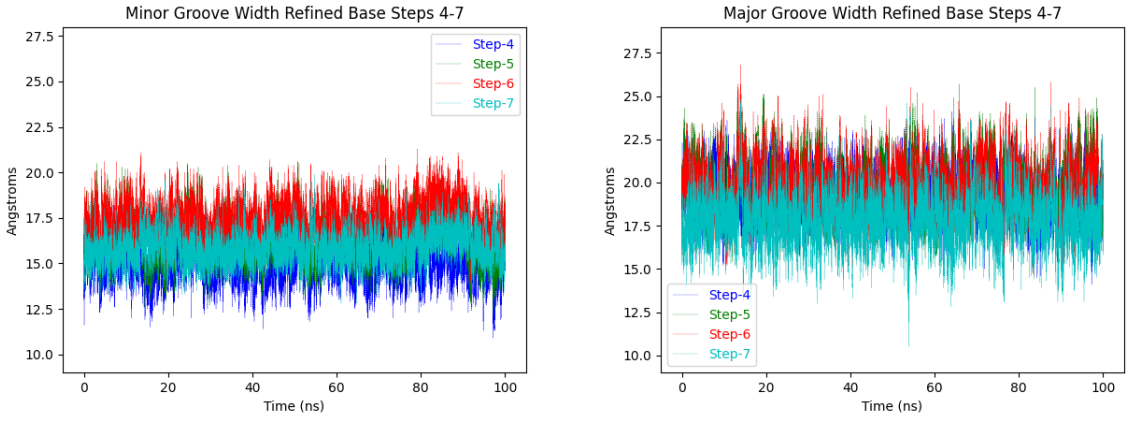
(6.193) DB[a,j]A-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



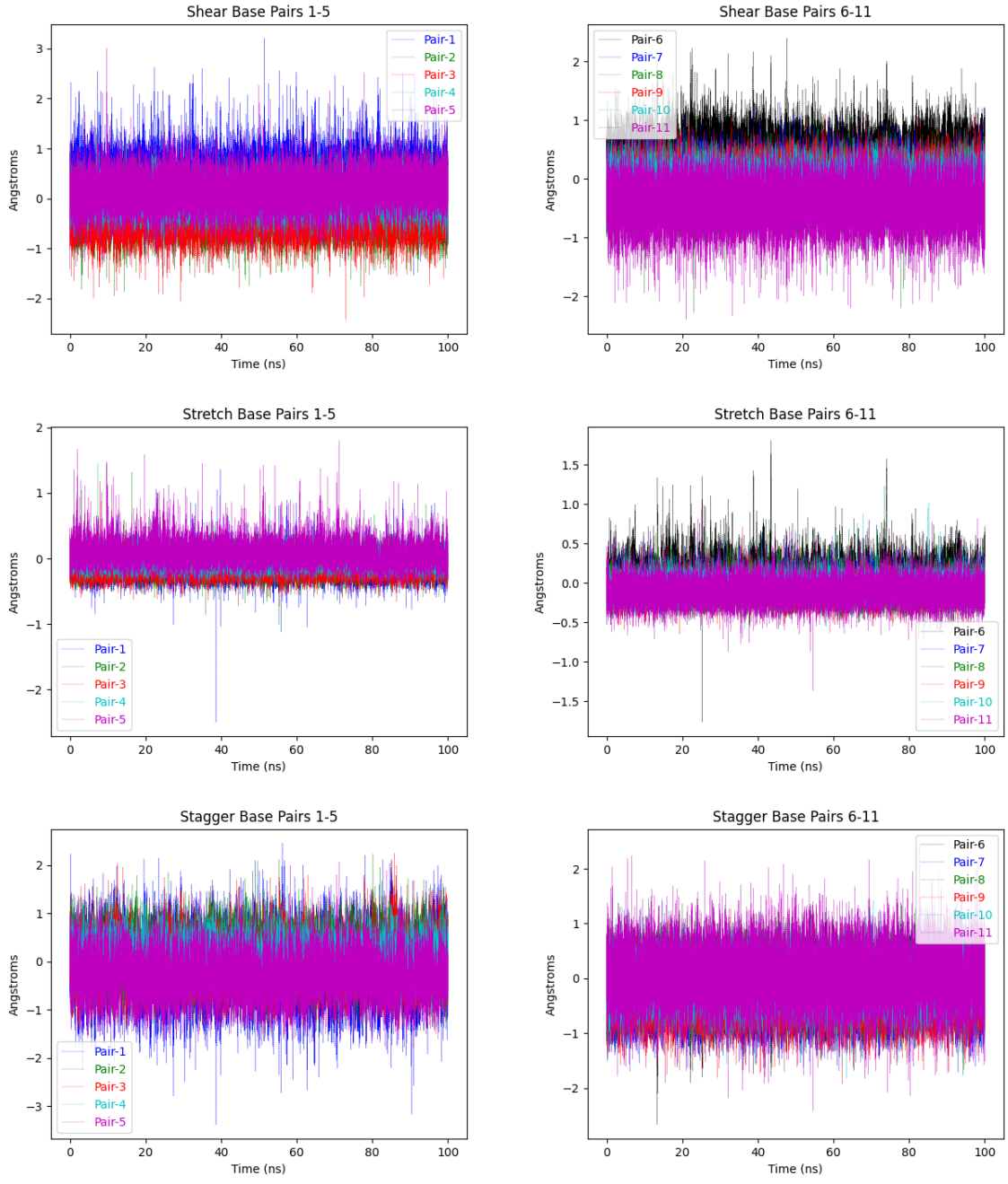
(6.194) DB[a,j]A-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



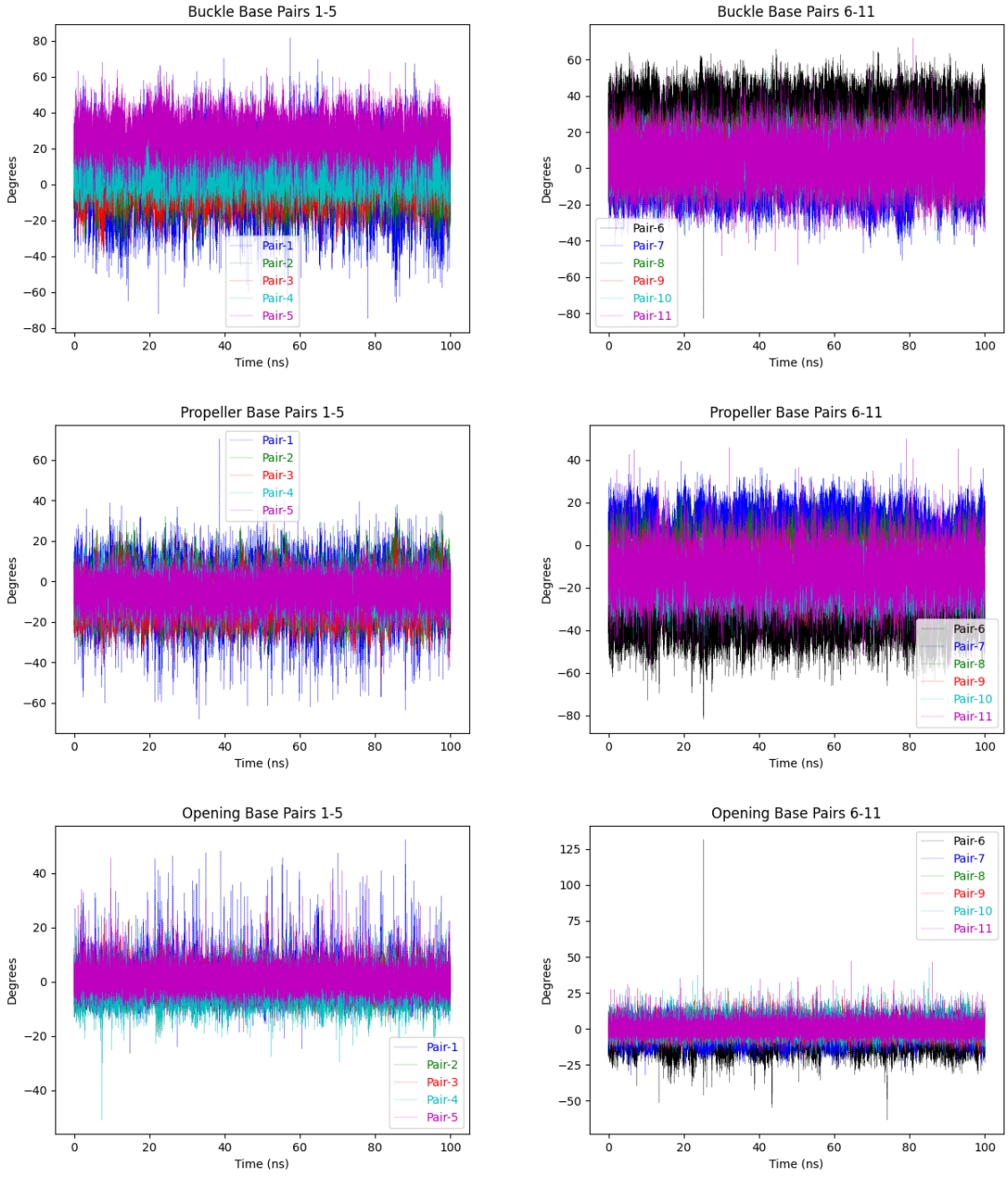
(6.195) DB[a,j]A-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



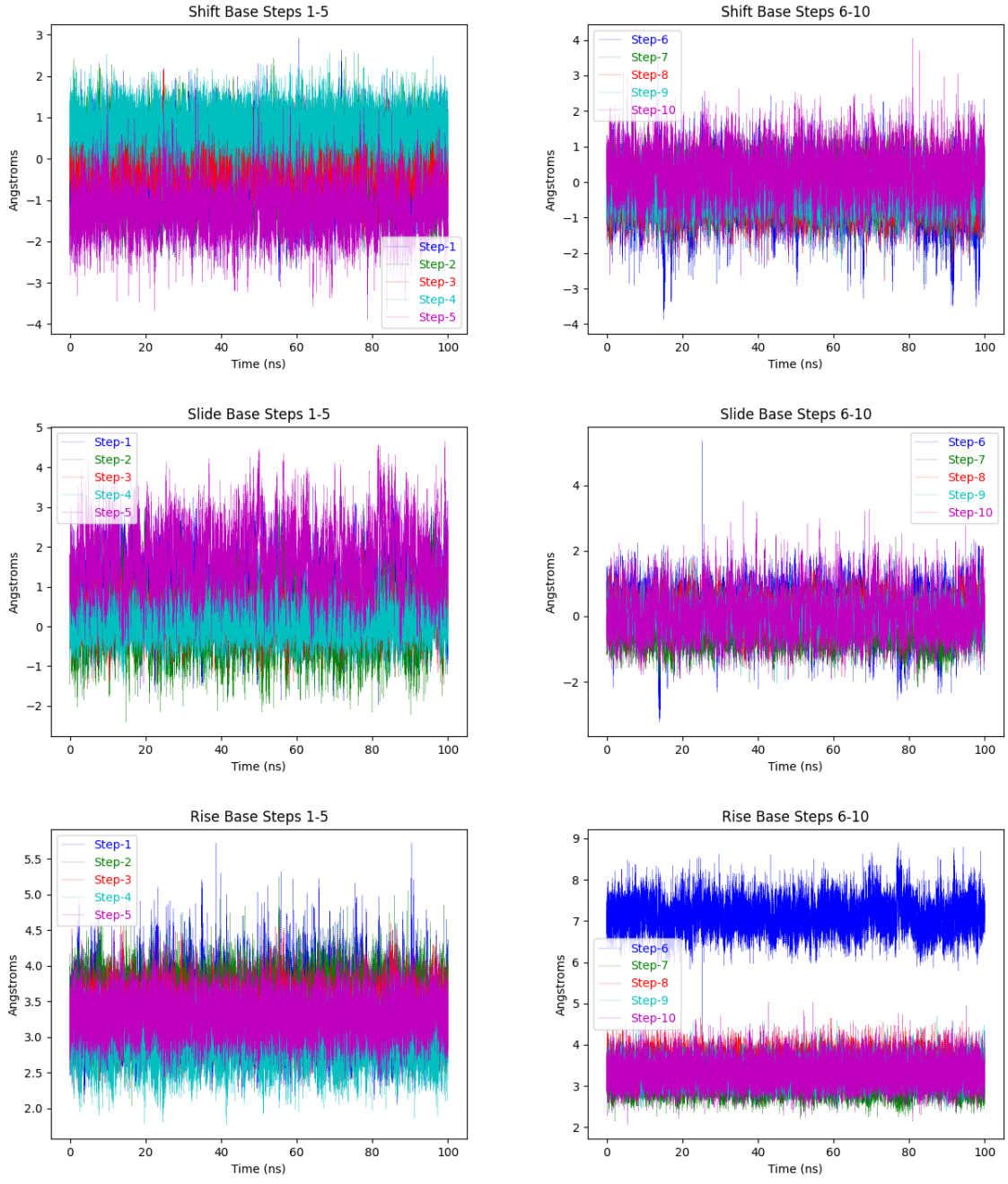
(6.196) DB[a,j]A-DNA: Refined major and minor groove trajectories



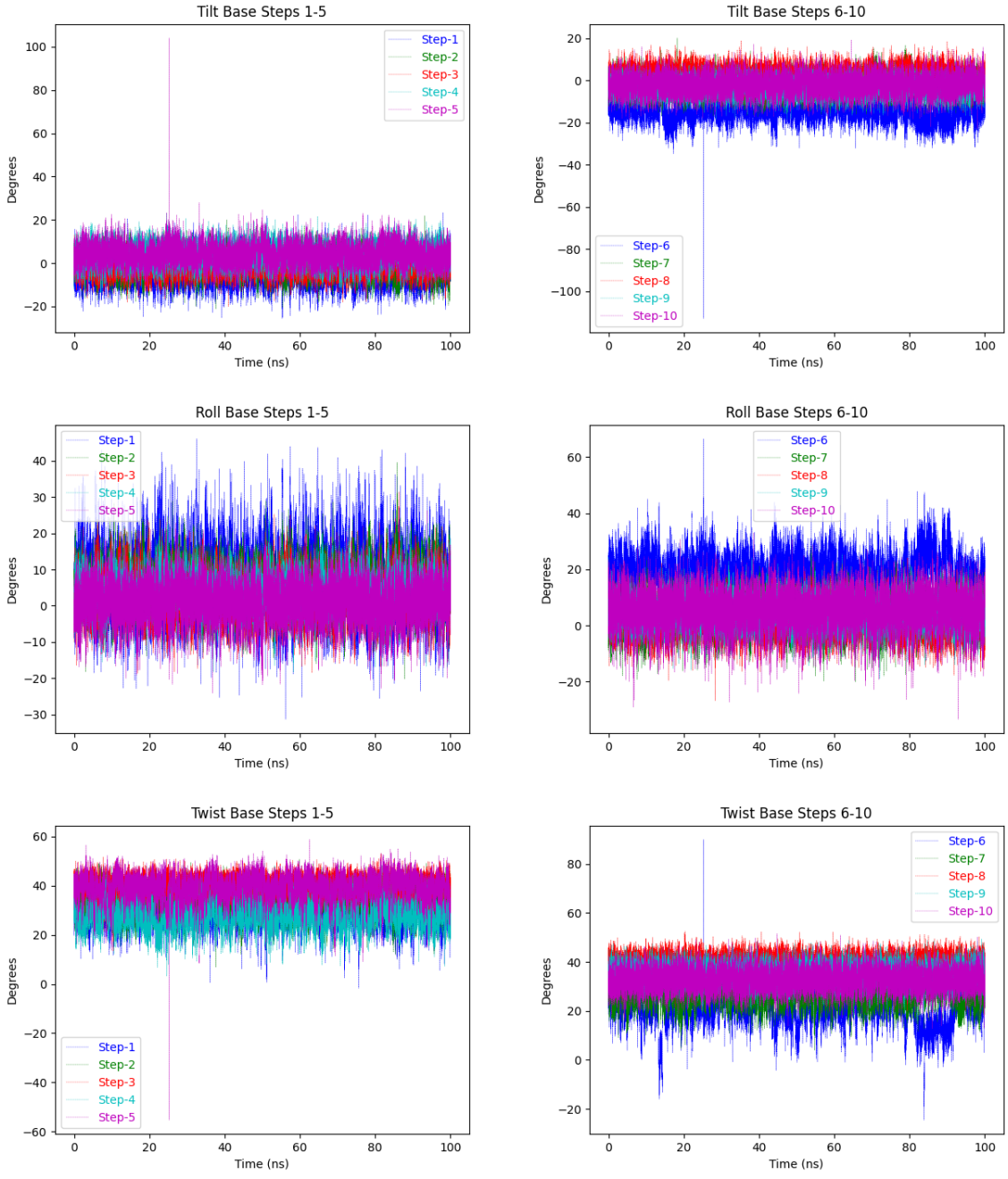
(6.197) DB[a,j]A-DNA: Base pair trajectories



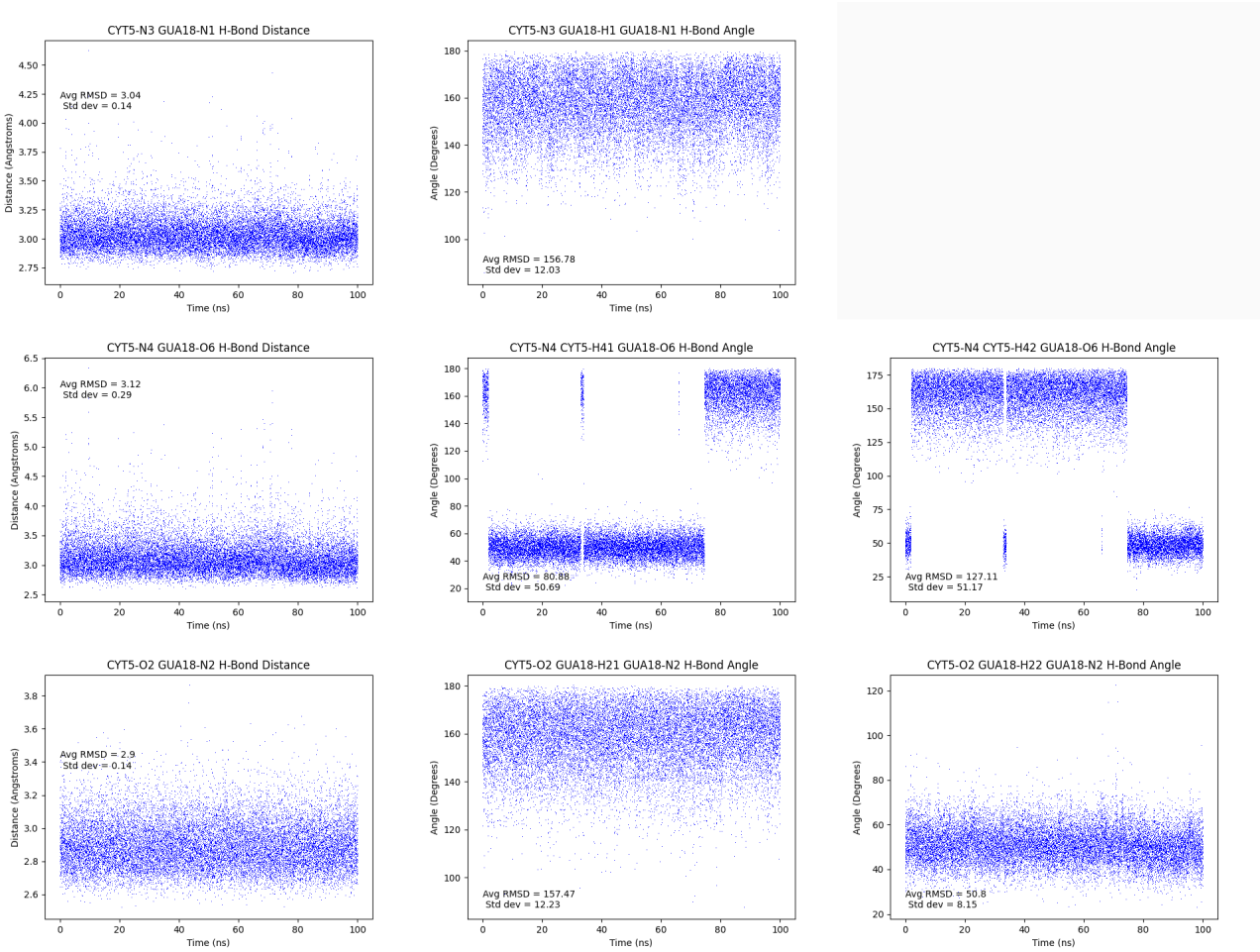
(6.198) DB[a,j]A-DNA: Base pair trajectories



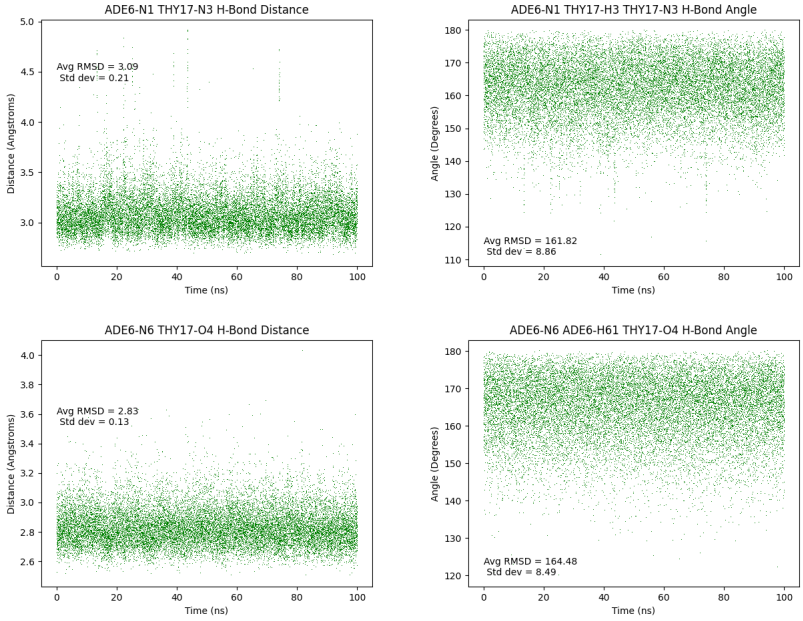
(6.199) DB[a,j]A-DNA: Base step trajectories



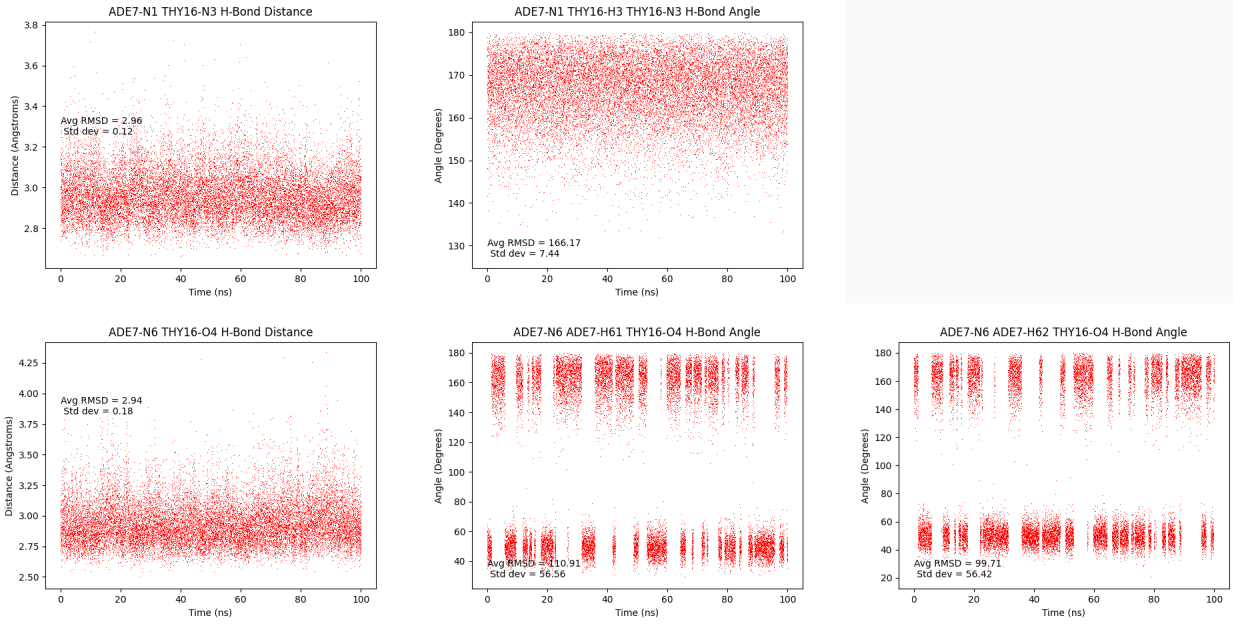
(6.200) DB[a,j]A-DNA: Base step trajectories



(6.201) DB[a,j]A-DNA: dC_5:dG_{18} hydrogen bond trajectories

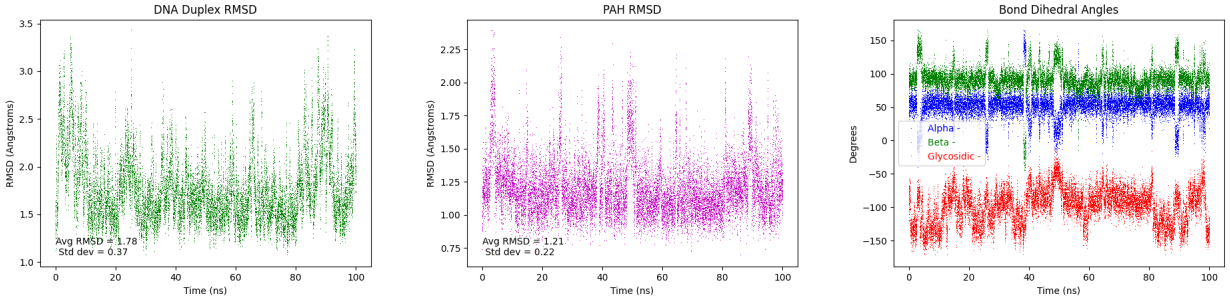


(6.202) DB[a,j]A-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

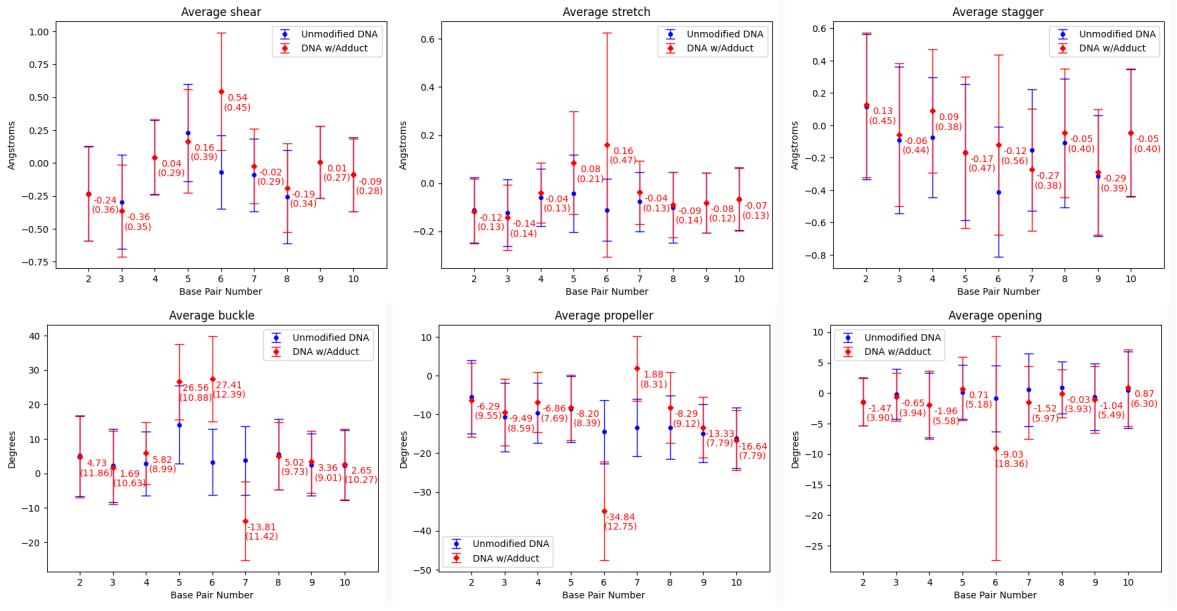


(6.203) DB[a,j]A-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

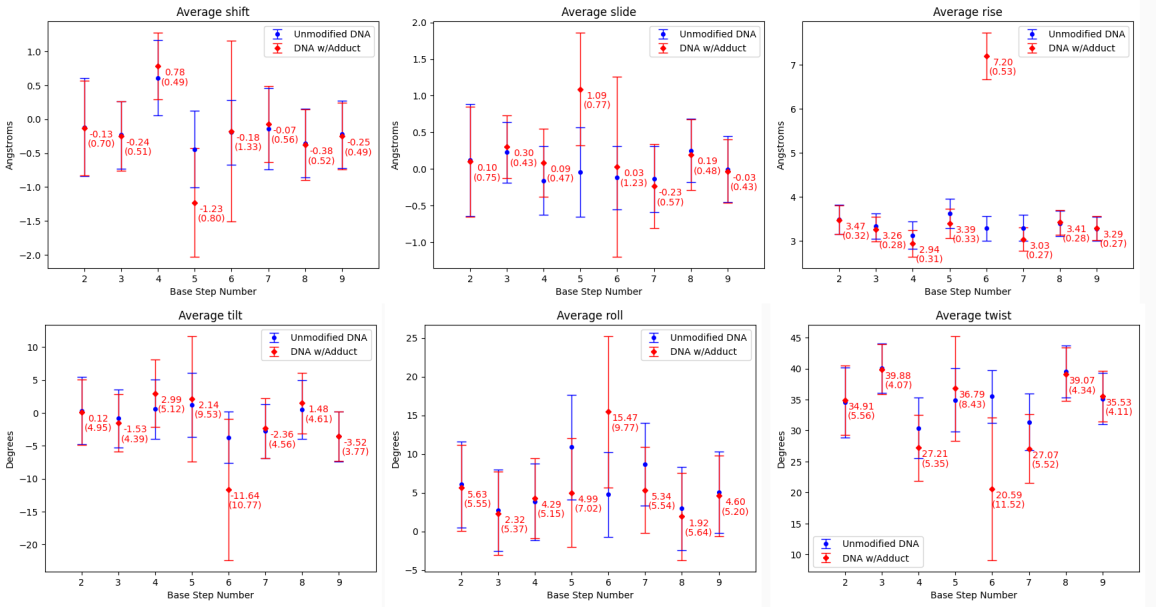
6.2.1.15 B[b]C-DNA



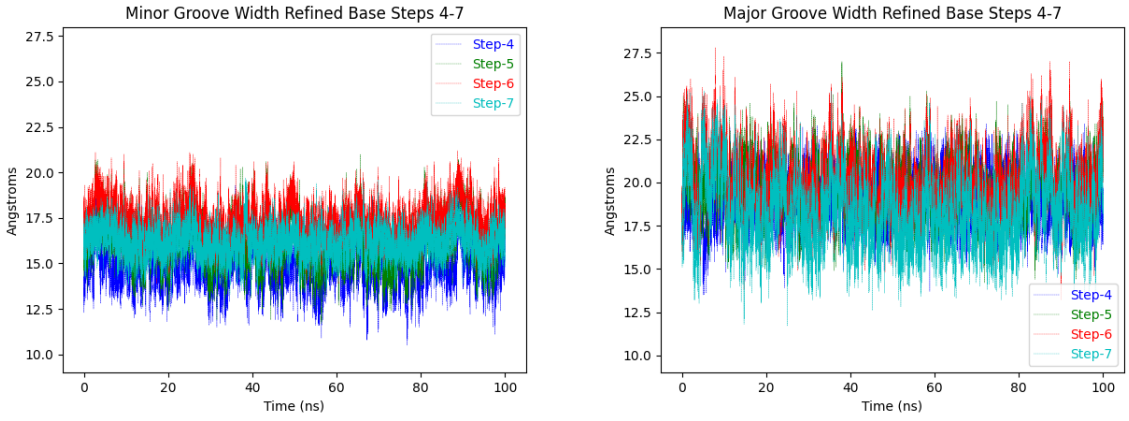
(6.204) B[b]C-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



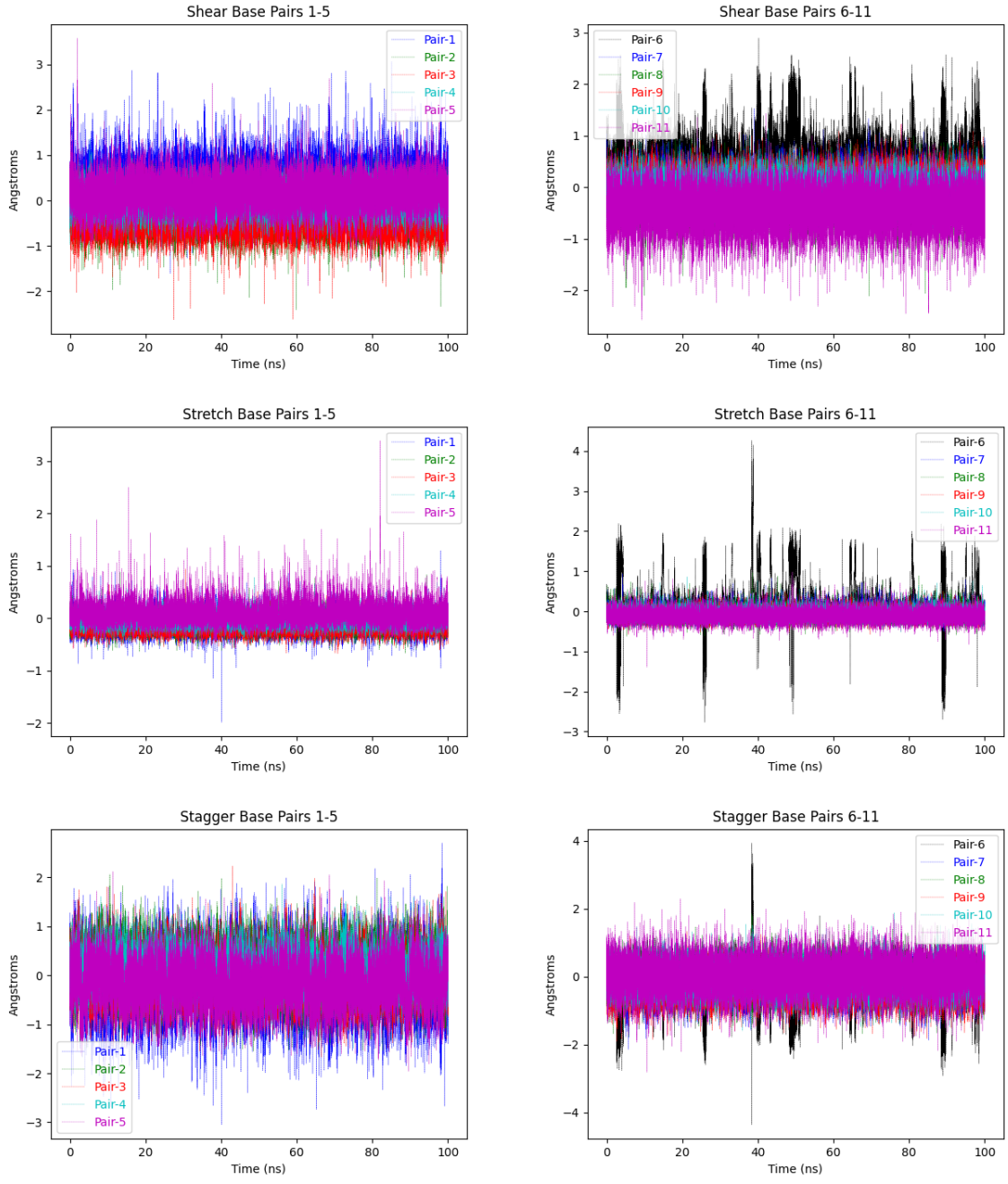
(6.205) B[b]C-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



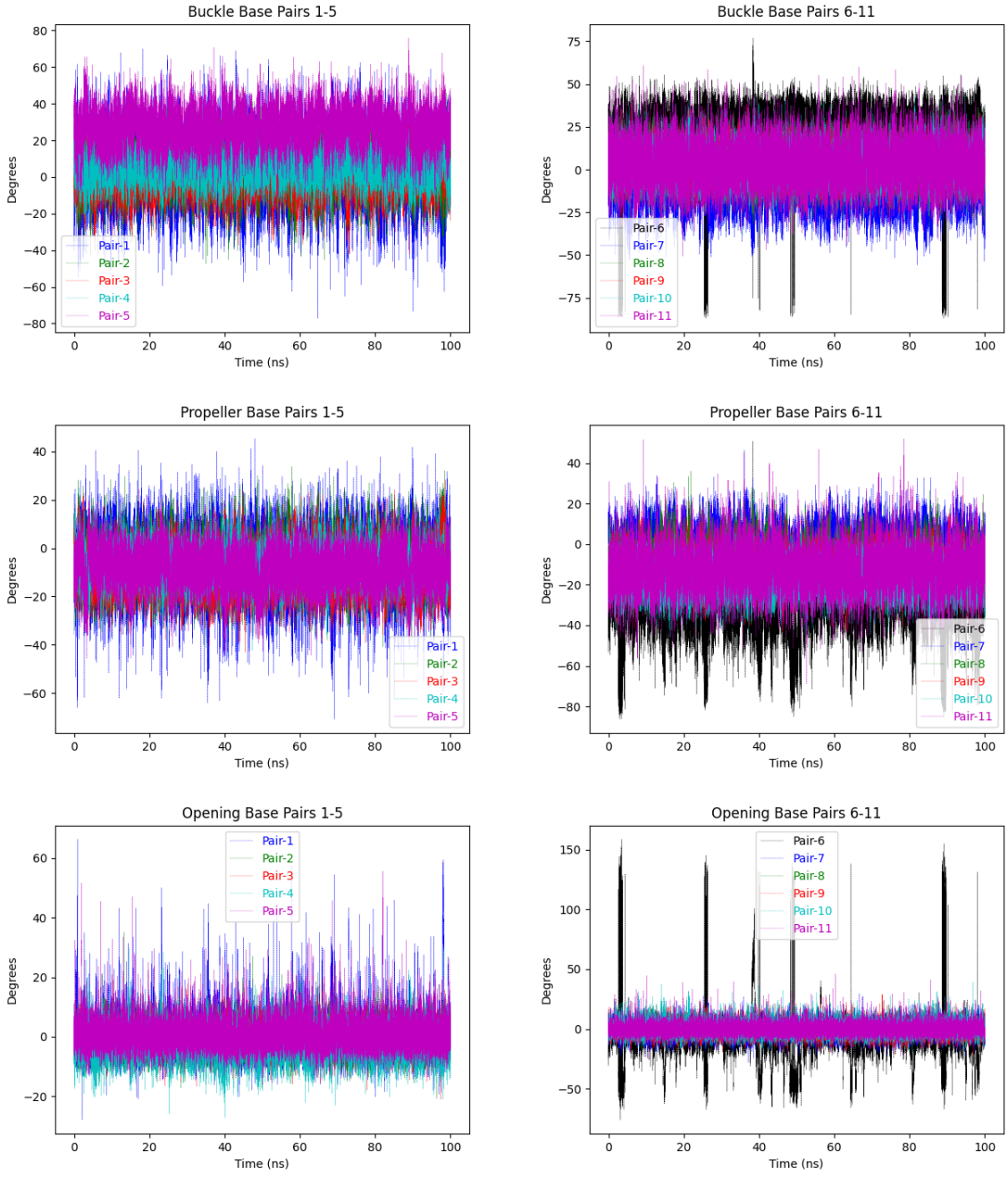
(6.206) B[b]C-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



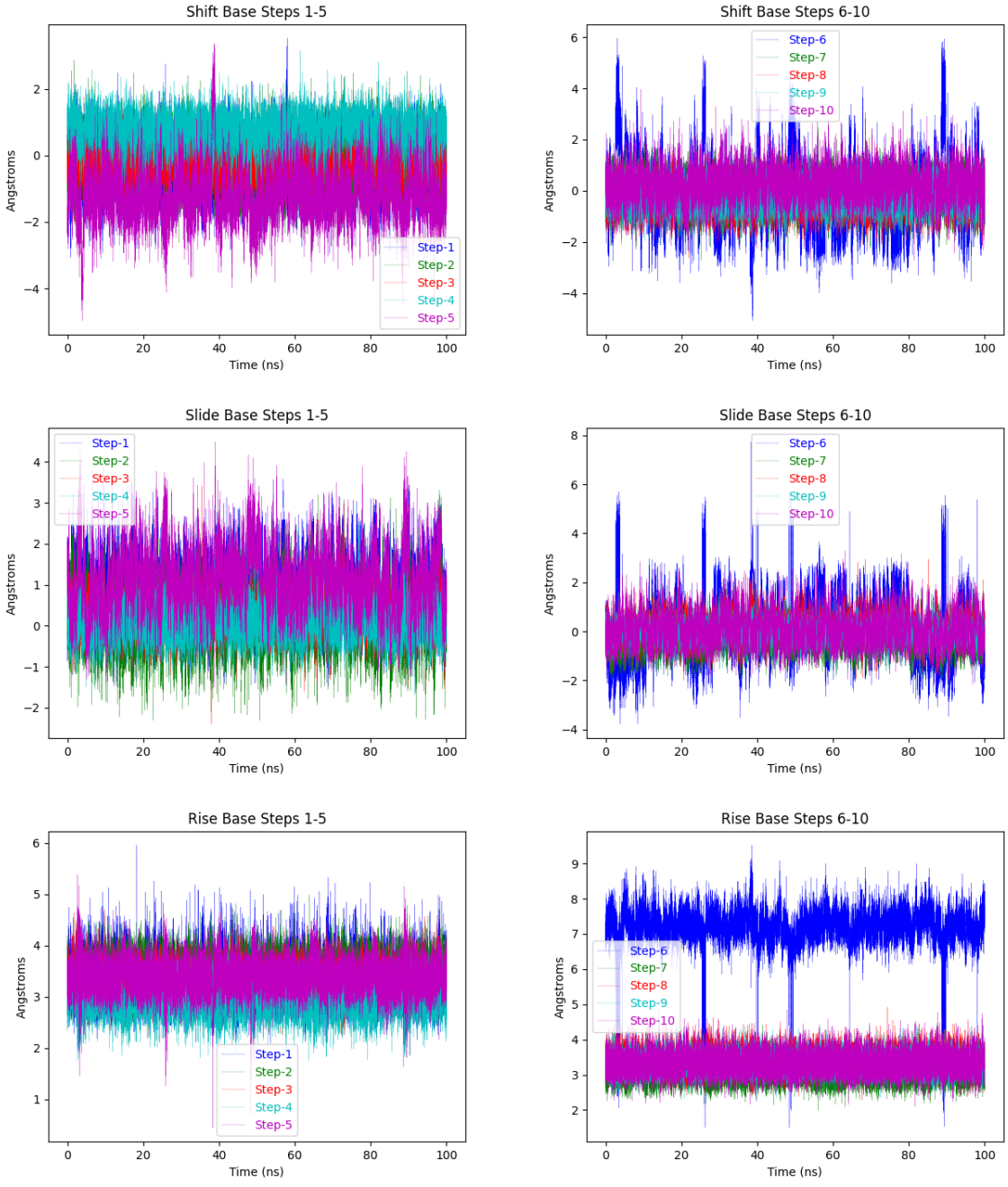
(6.207) B[b]C-DNA: Refined major and minor groove trajectories



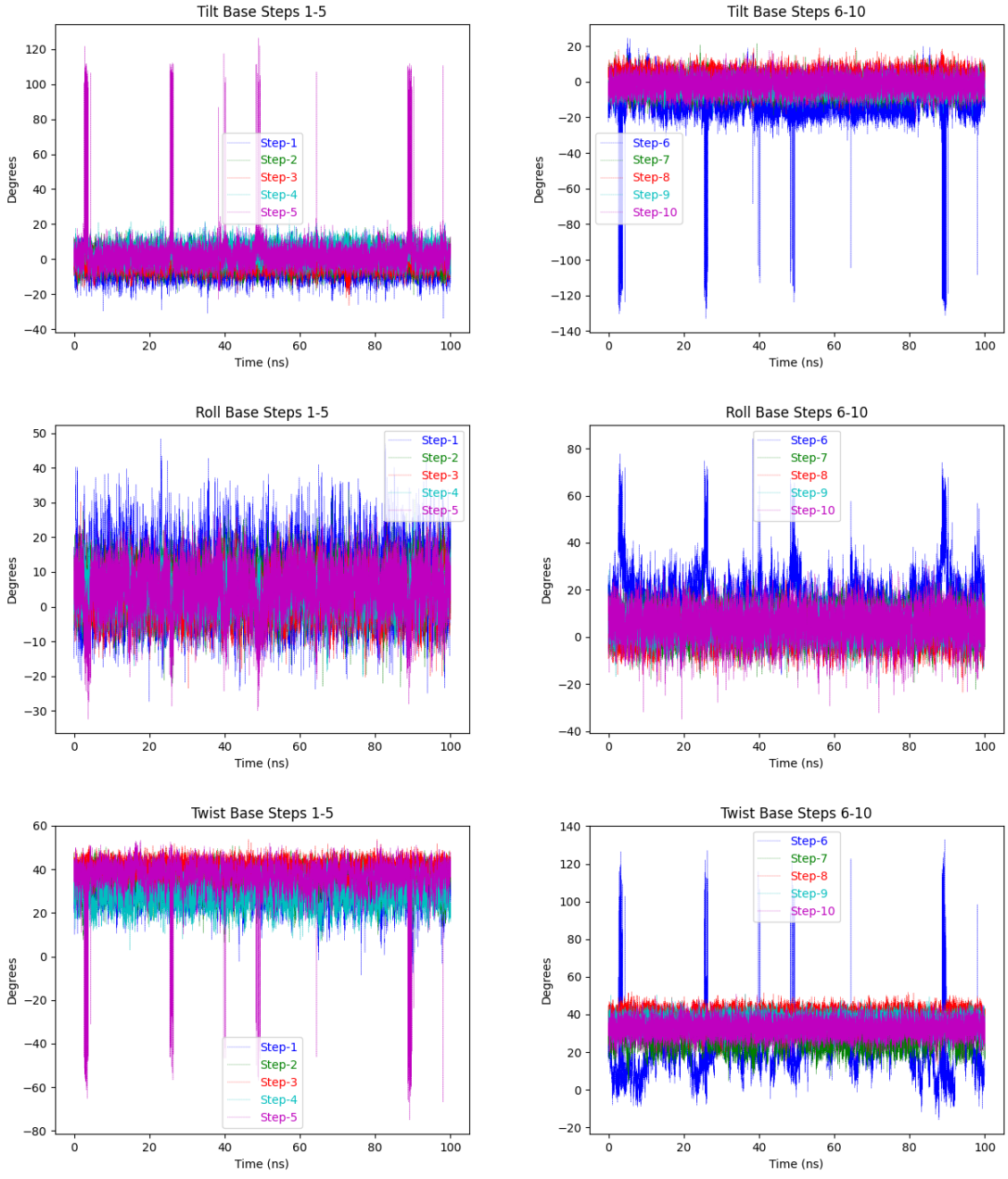
(6.208) B[b]C-DNA: Base pair trajectories



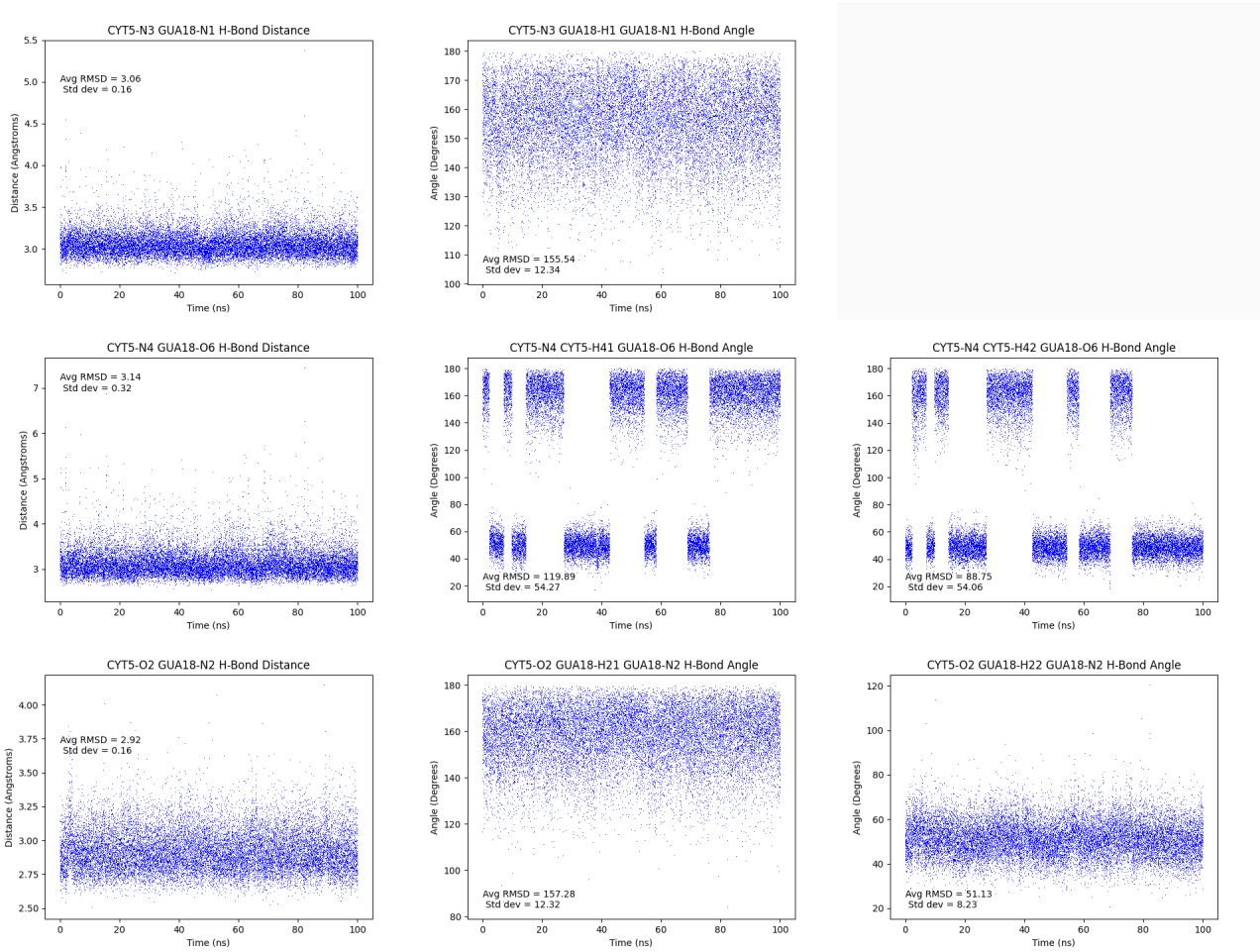
(6.209) B[b]C-DNA: Base pair trajectories



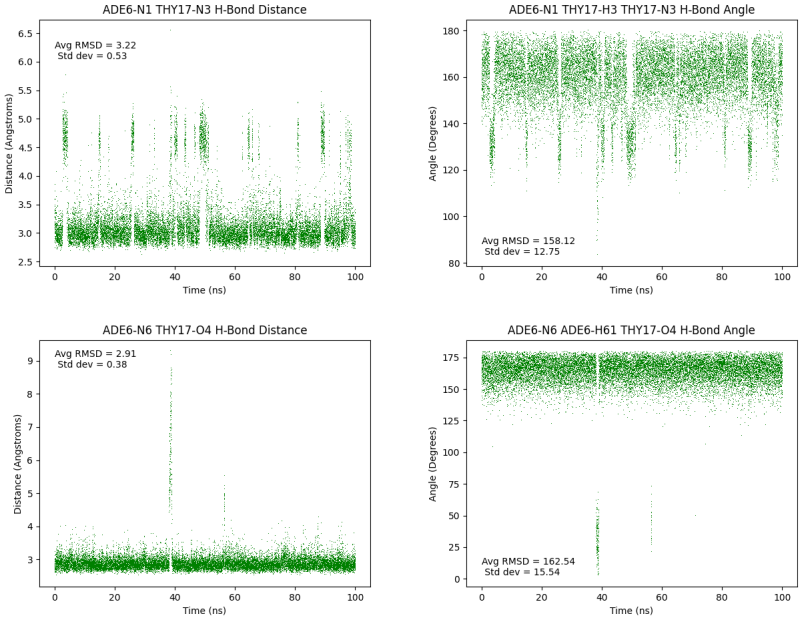
(6.210) B[b]C-DNA: Base step trajectories



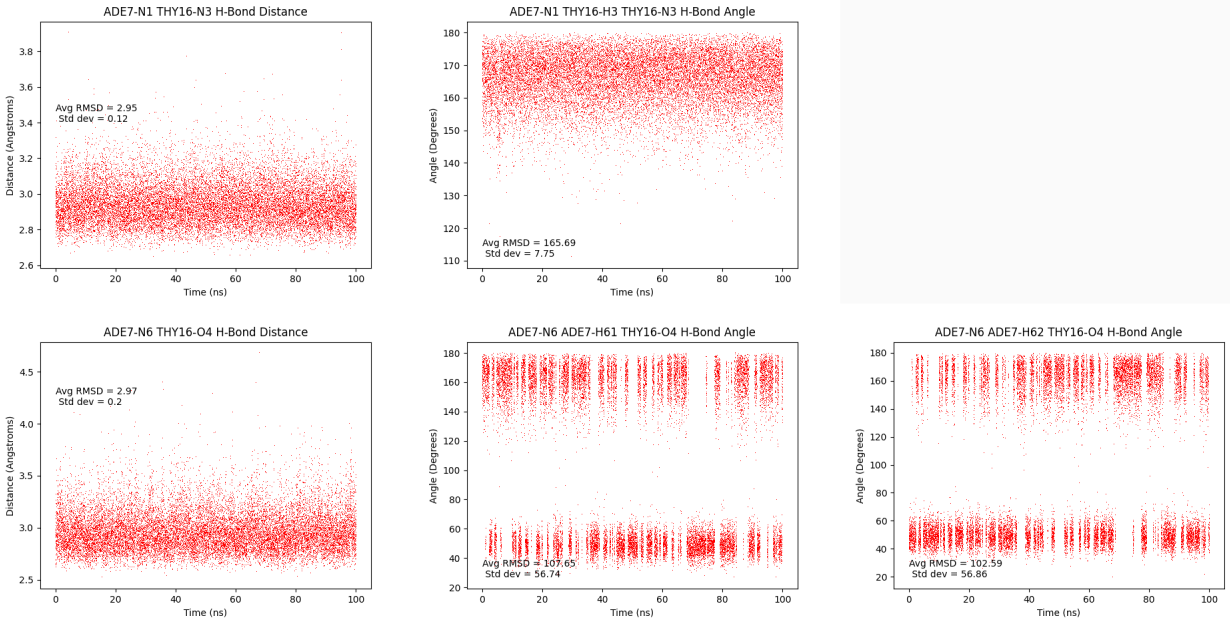
(6.211) B[b]C-DNA: Base step trajectories



(6.212) B[b]C-DNA: dC_5:dG_{18} hydrogen bond trajectories

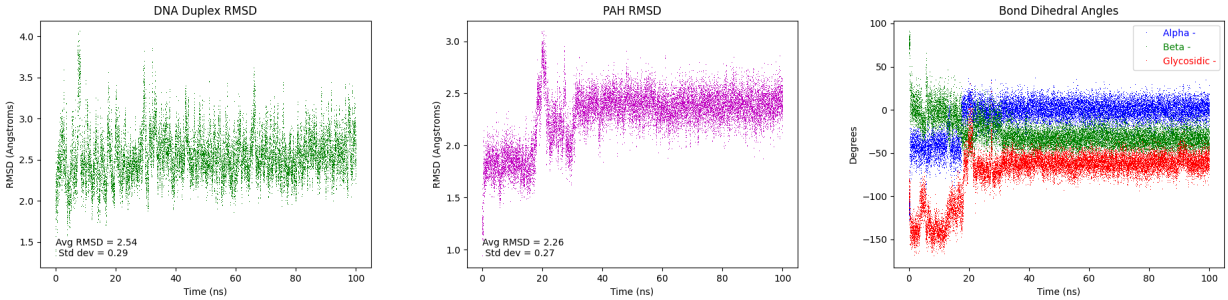


(6.213) B[b]C-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

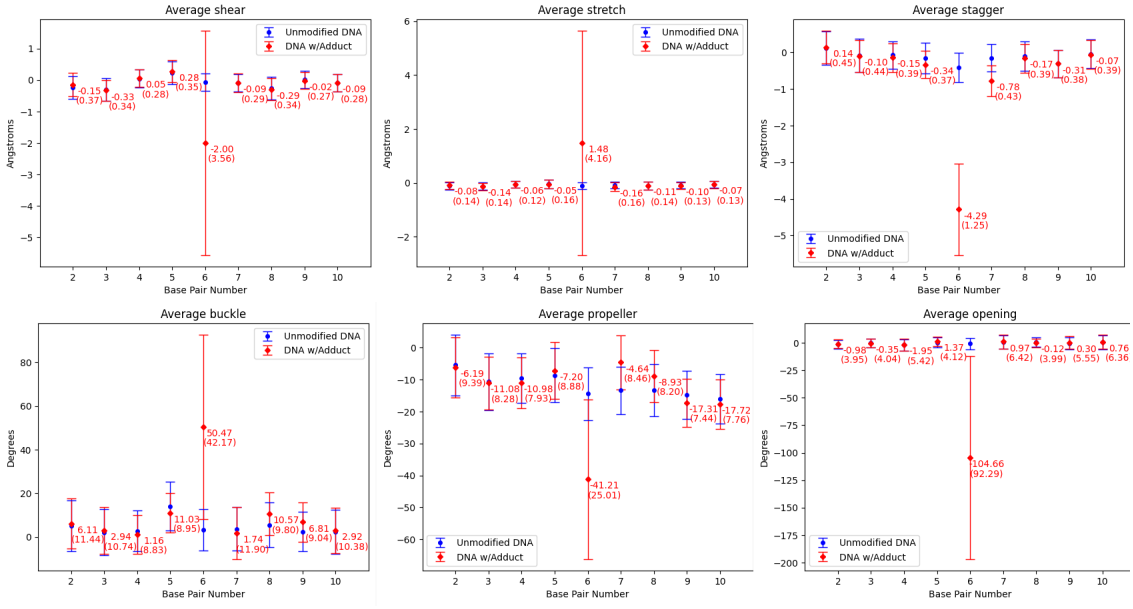


(6.214) B[b]C-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

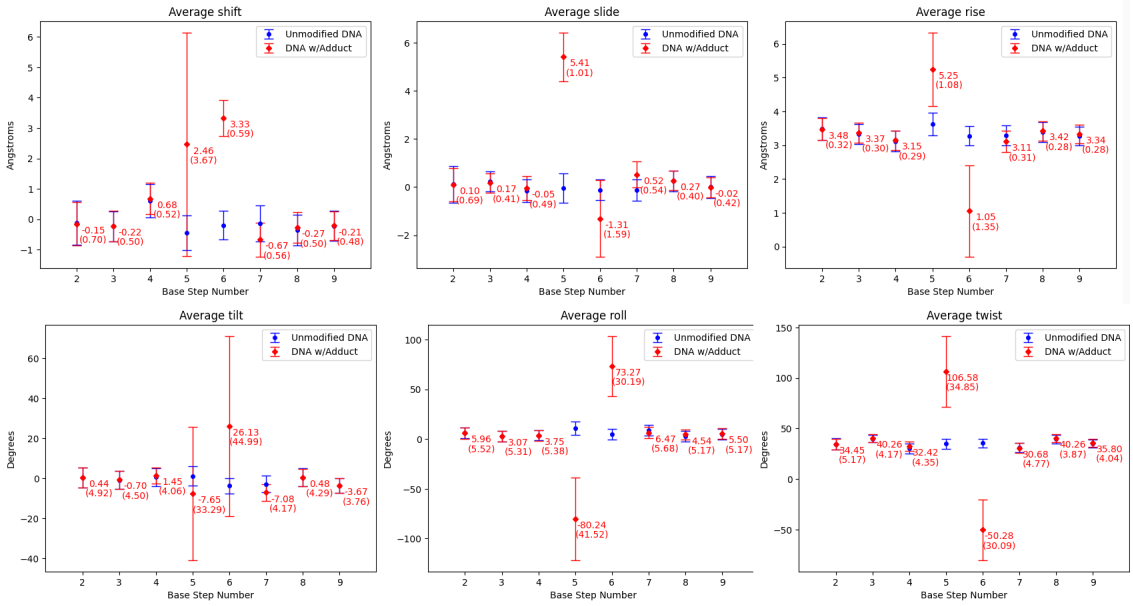
6.2.1.16 DB[e,l]P-DNA



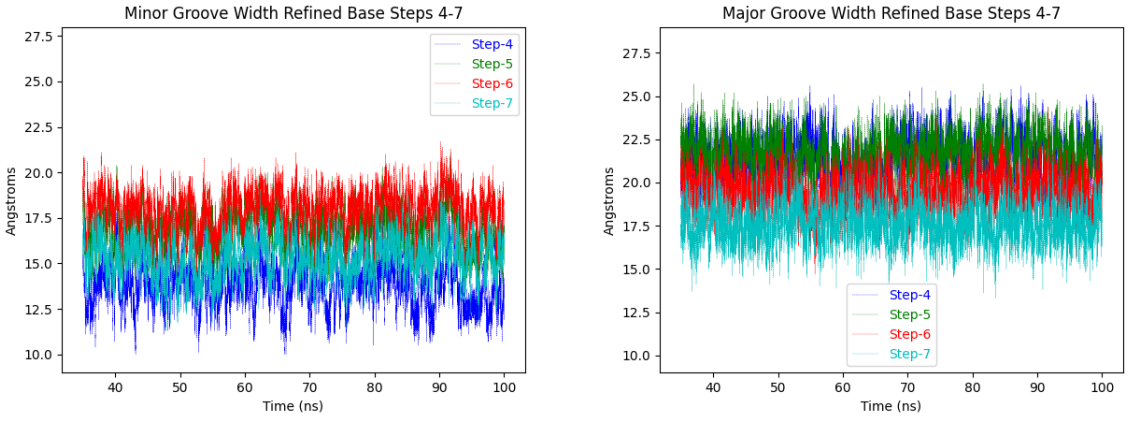
(6.215) DB[e,l]P-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



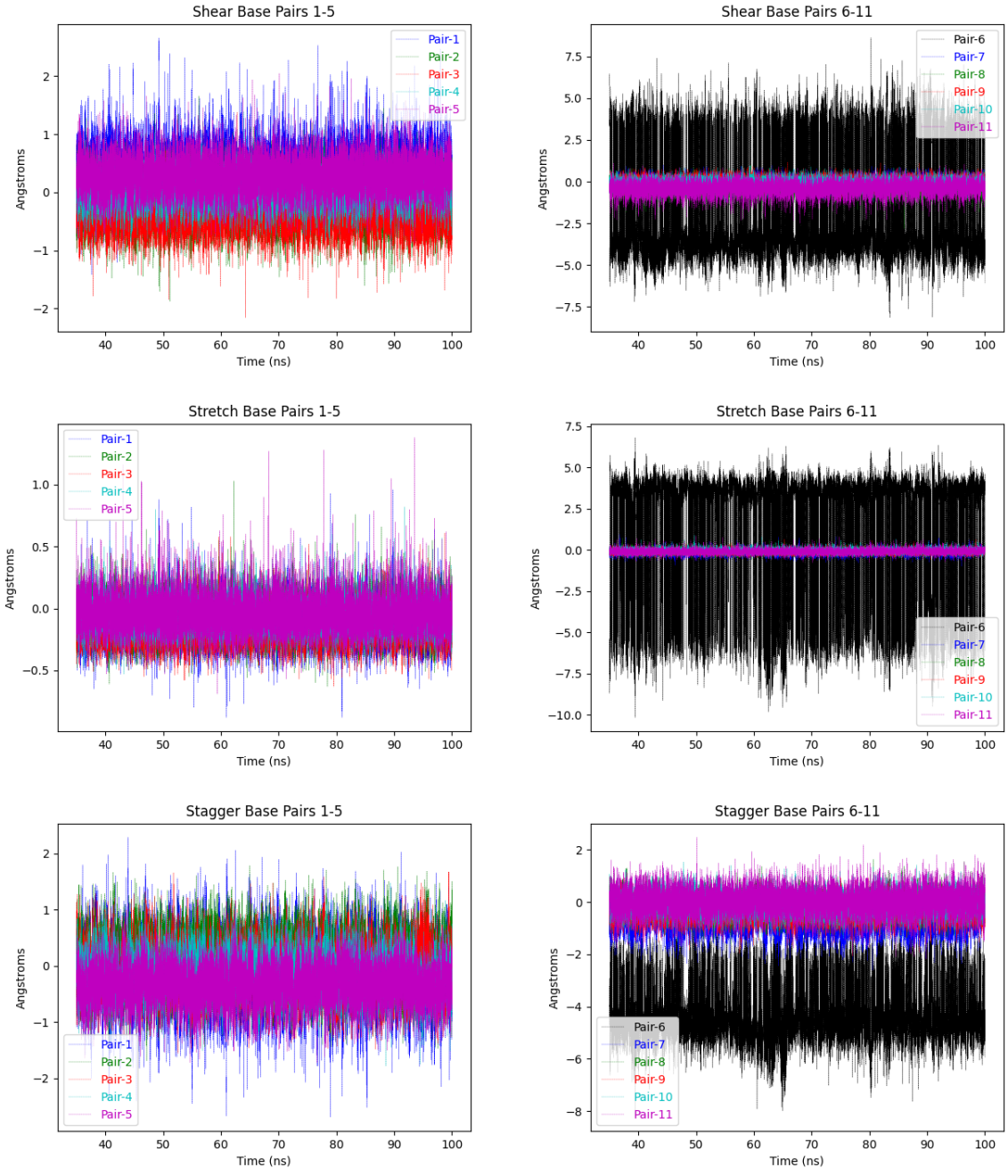
(6.216) DB[e,l]P-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



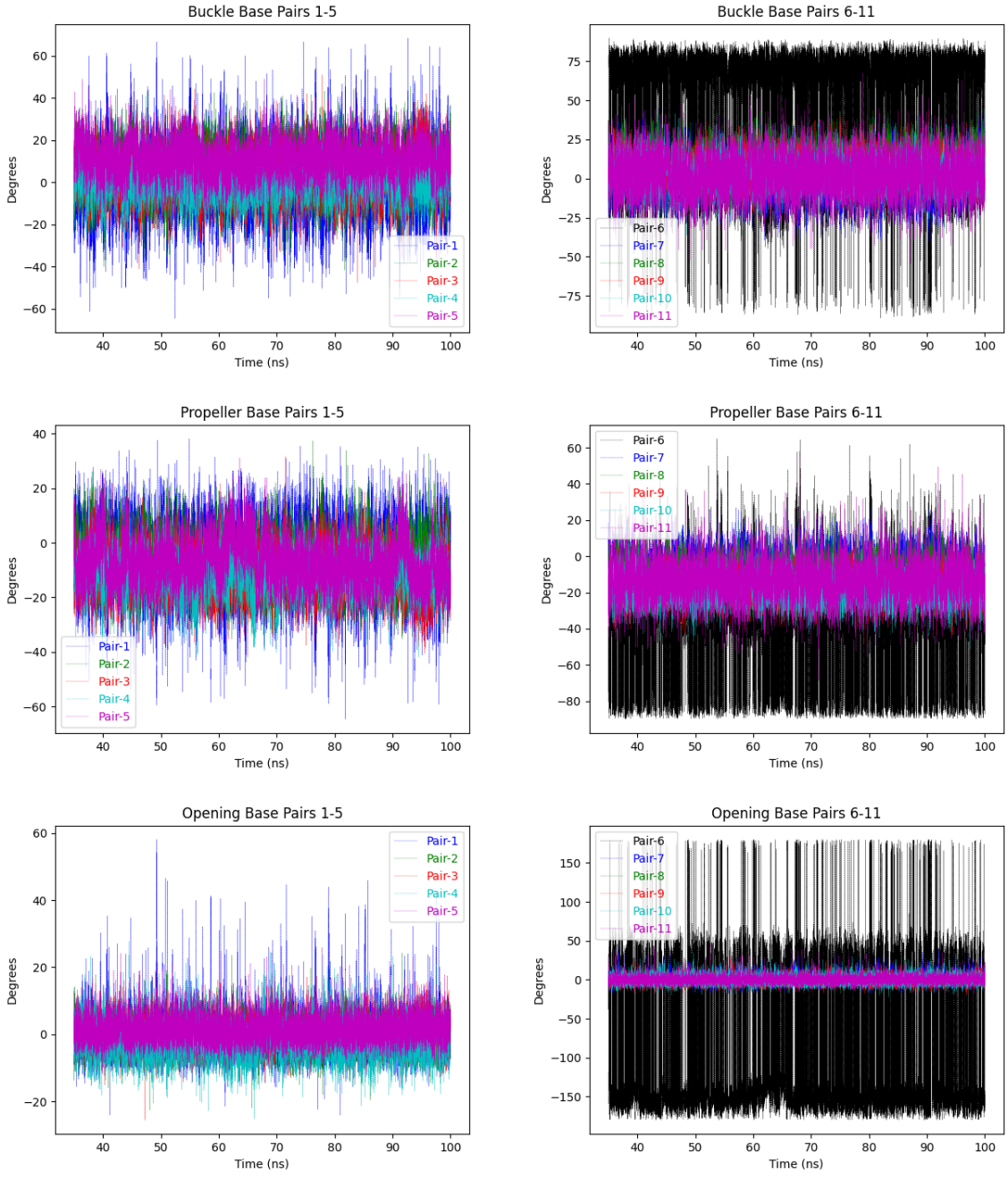
(6.217) DB[e,l]P-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



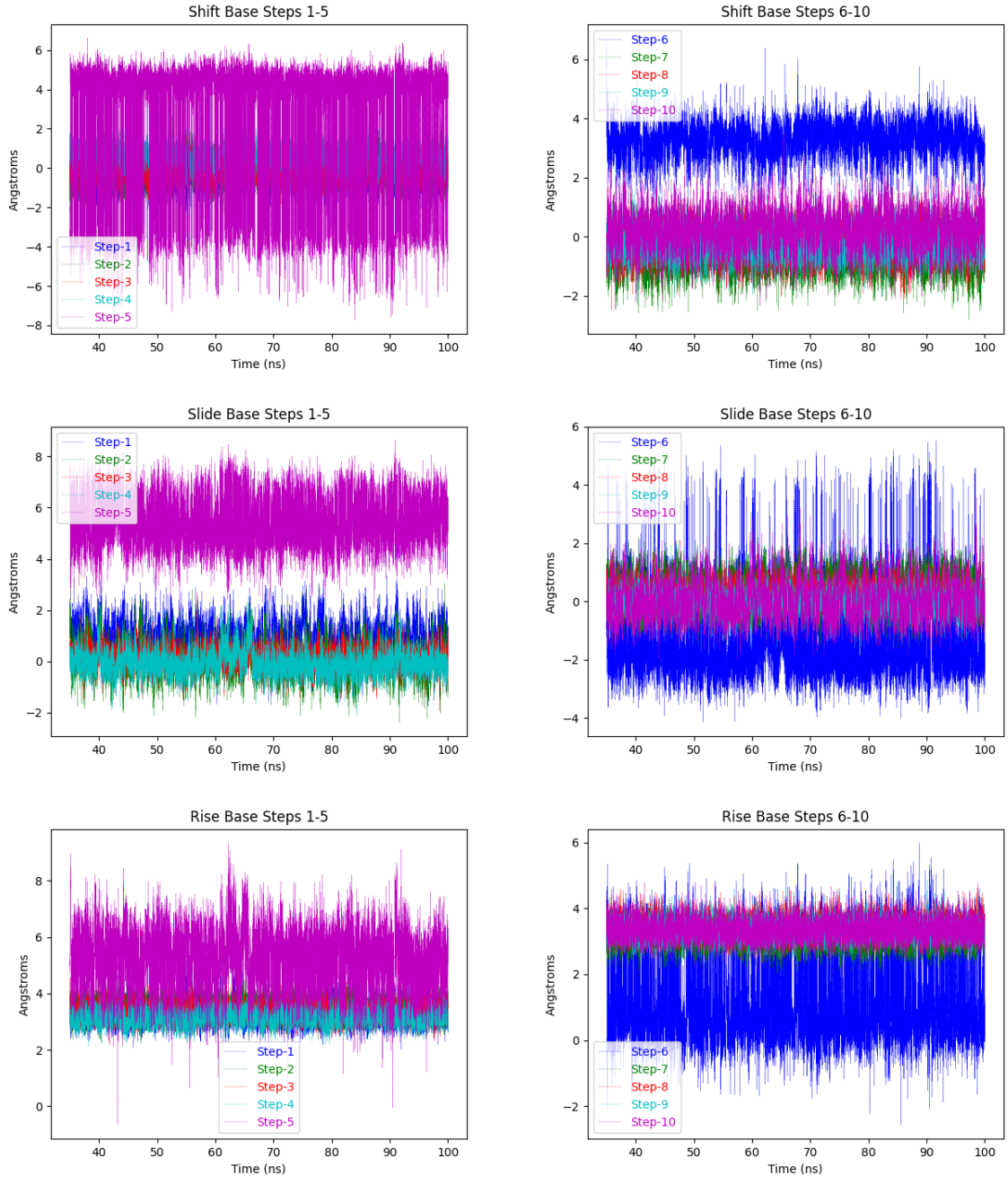
(6.218) DB[e,l]P-DNA: Refined major and minor groove trajectories



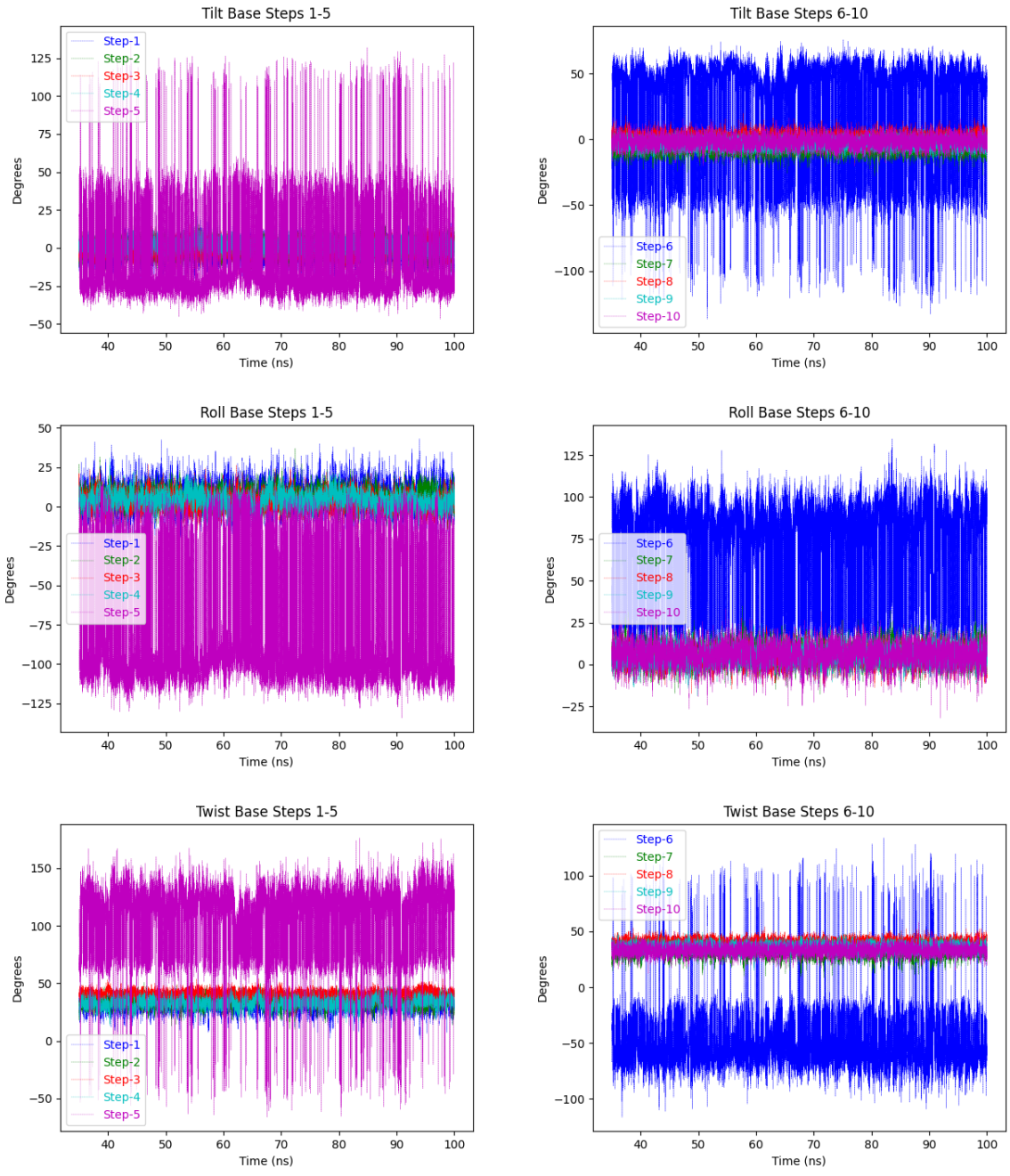
(6.219) DB[e,l]P-DNA: Base pair trajectories



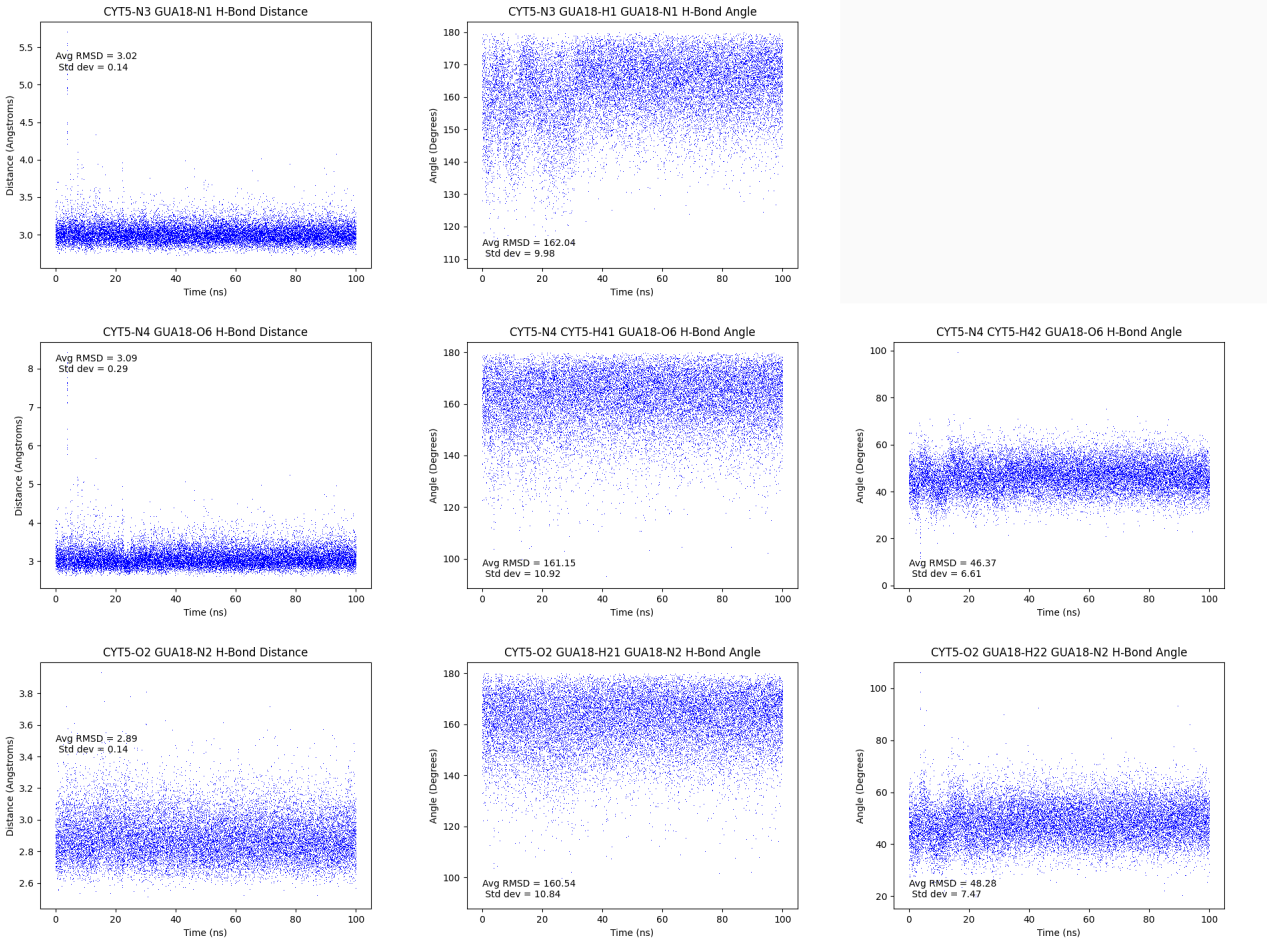
(6.220) DB[e,l]P-DNA: Base pair trajectories



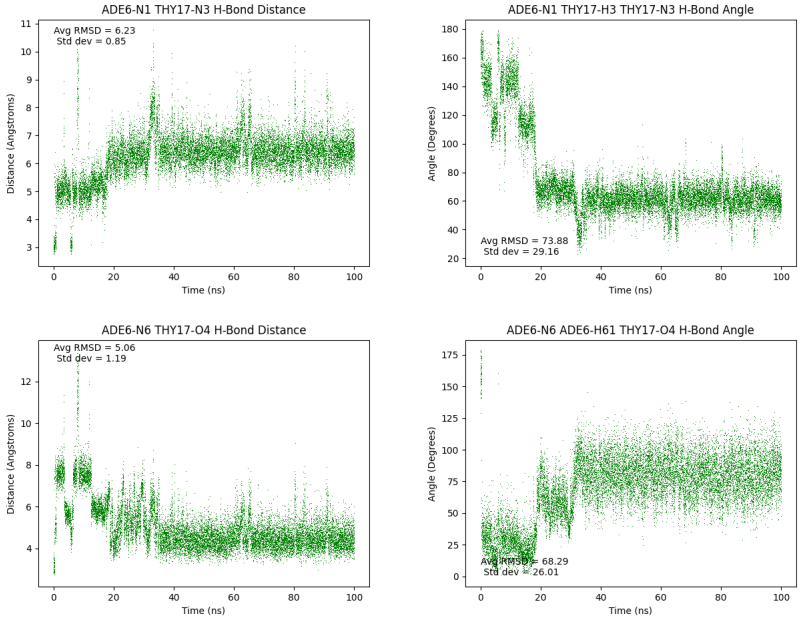
(6.221) DB[e,l]P-DNA: Base step trajectories



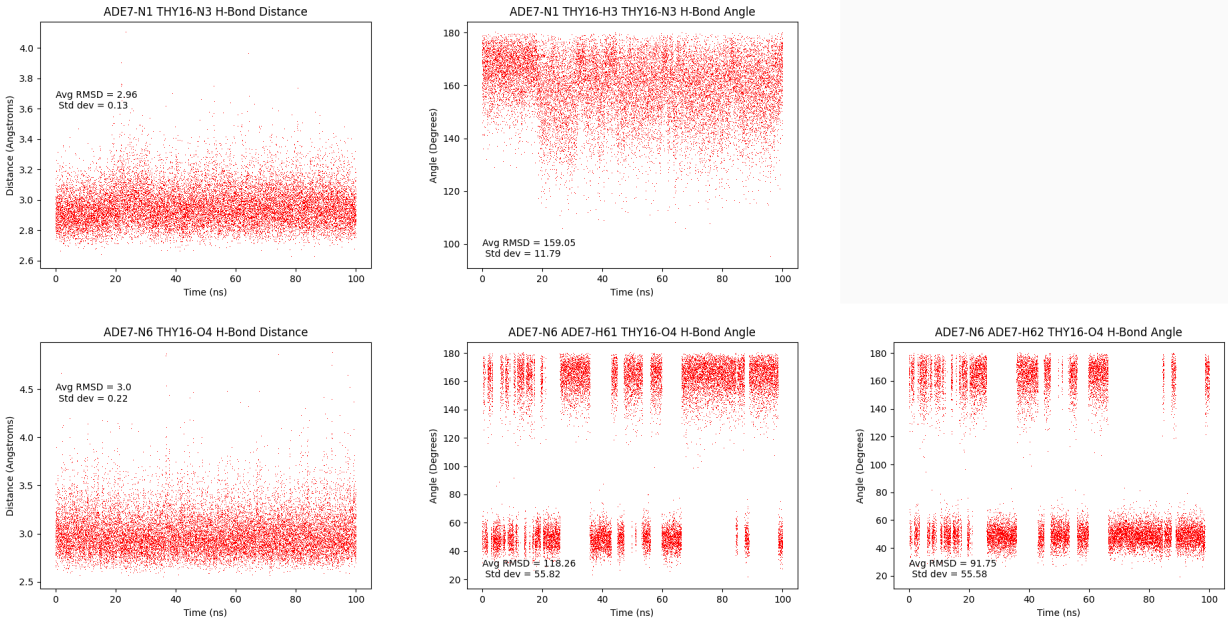
(6.222) DB[e,l]P-DNA: Base step trajectories



(6.223) DB[e,l]P-DNA: dC_5:dG_{18} hydrogen bond trajectories

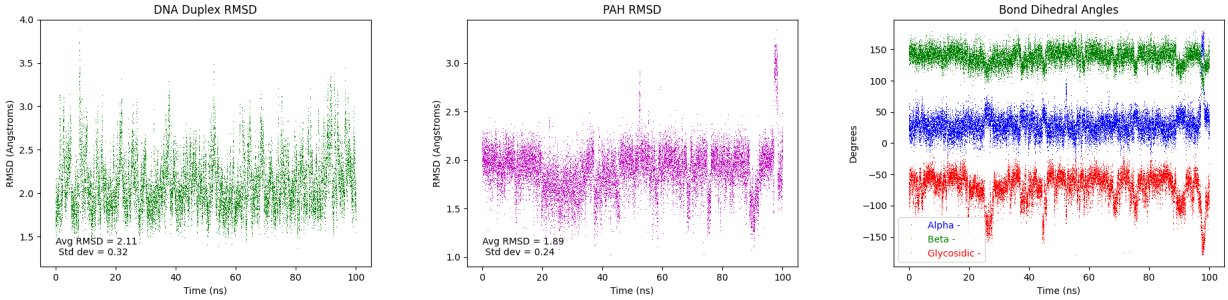


(6.224) DB[e,l]P-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories

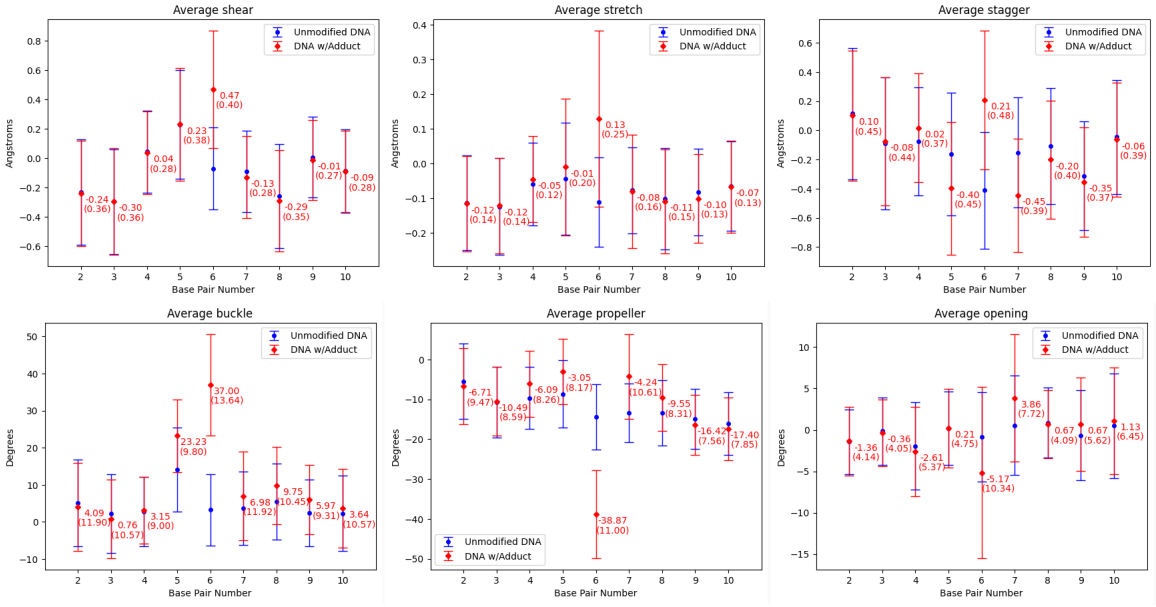


(6.225) DB[e,l]P-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

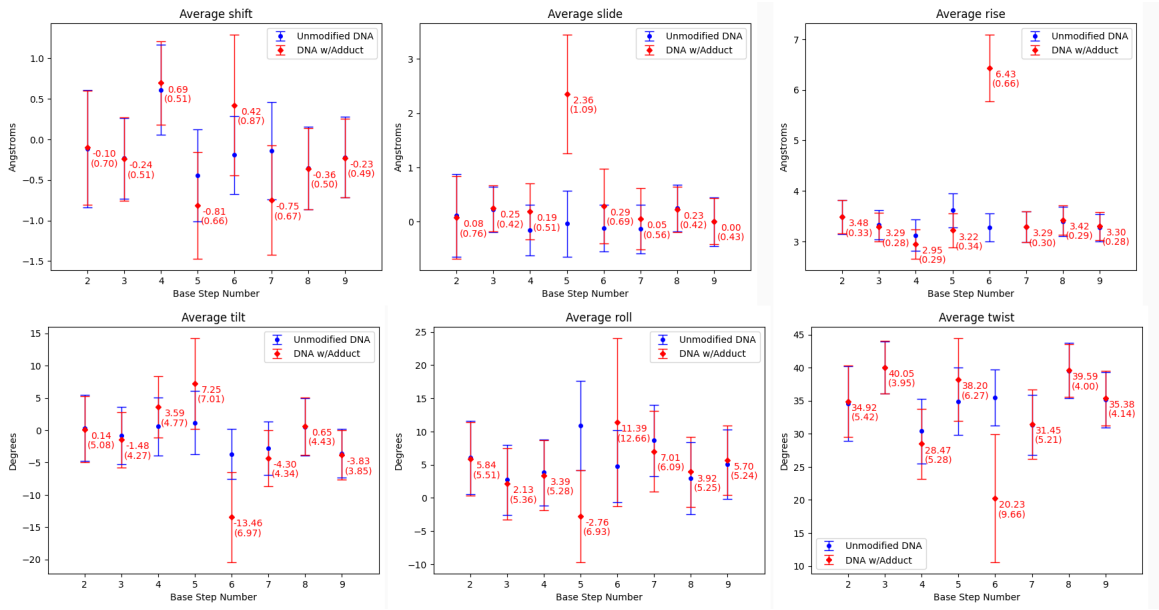
6.2.1.17 B[e]P-DNA



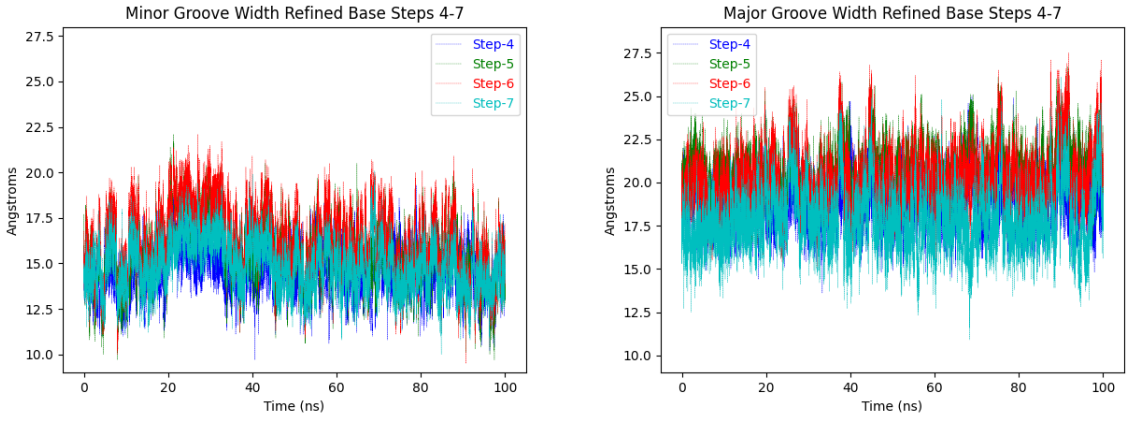
(6.226) B[e]P-DNA: duplex RMSD; PAH RMSD; α, β, χ trajectories



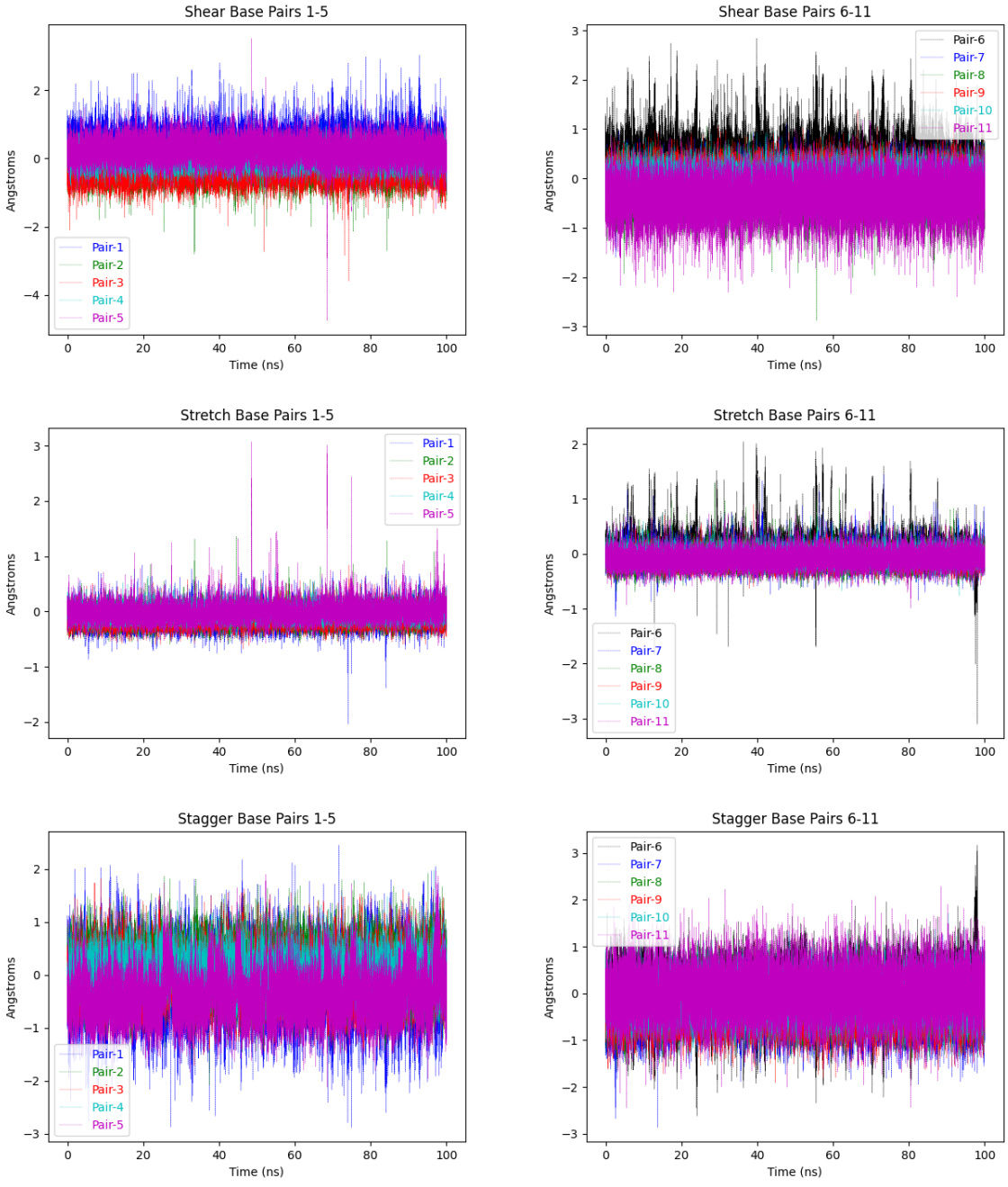
(6.227) B[e]P-DNA: Average values of base pair rigid-body parameters, standard deviation in parenthesis.



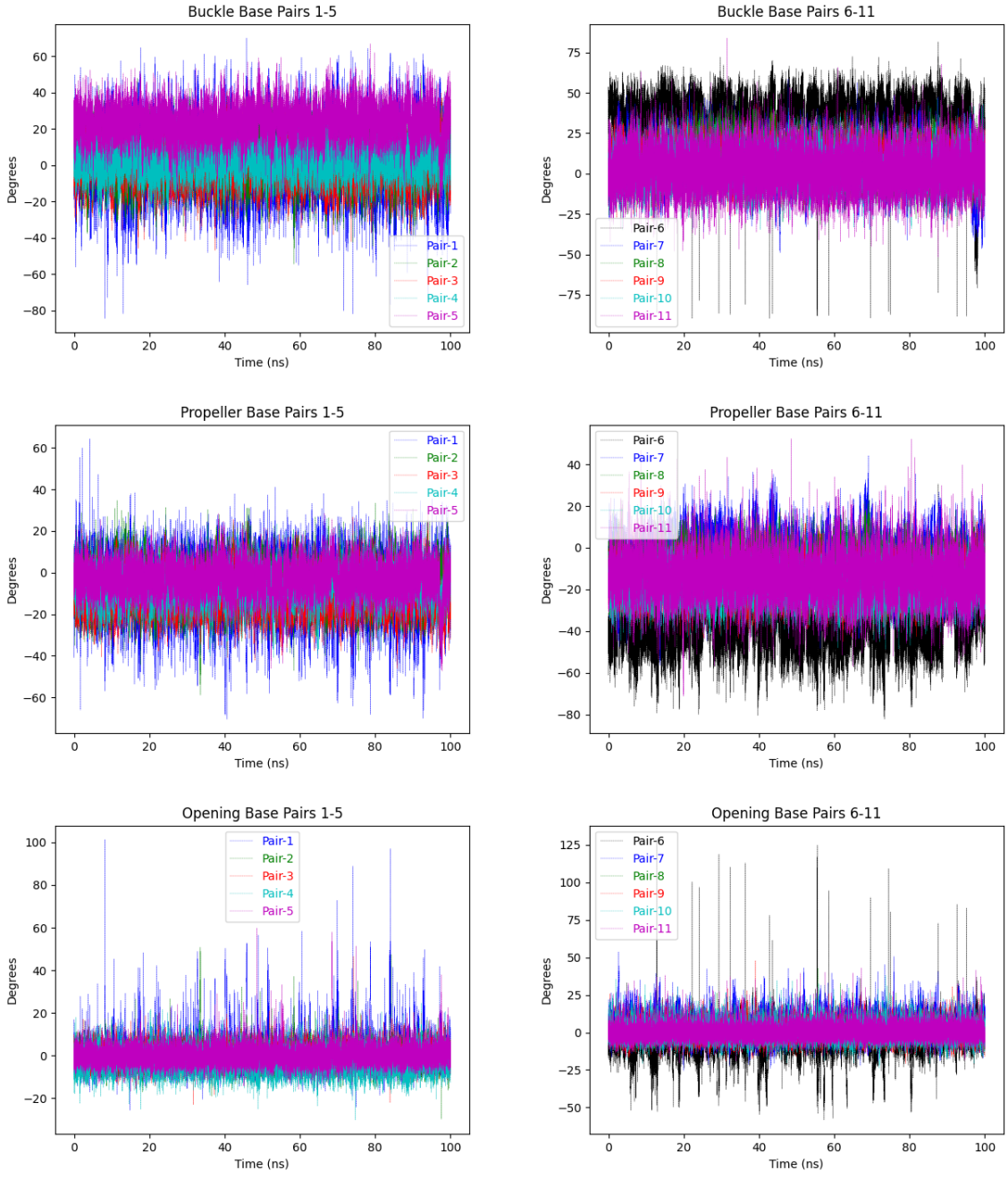
(6.228) B[e]P-DNA: Average values of base step rigid-body parameters, standard deviation in parenthesis.



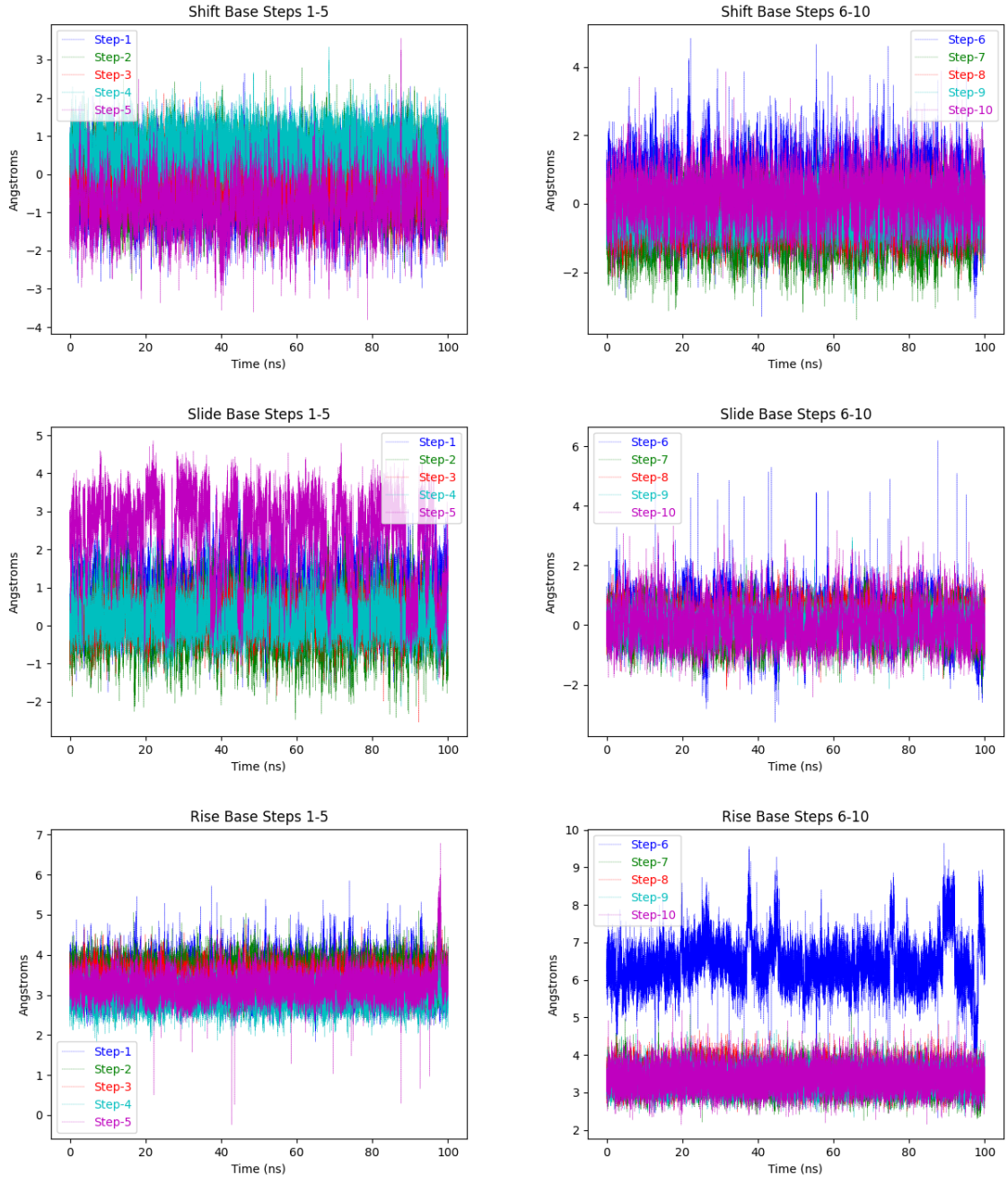
(6.229) B[e]P-DNA: Refined major and minor groove trajectories



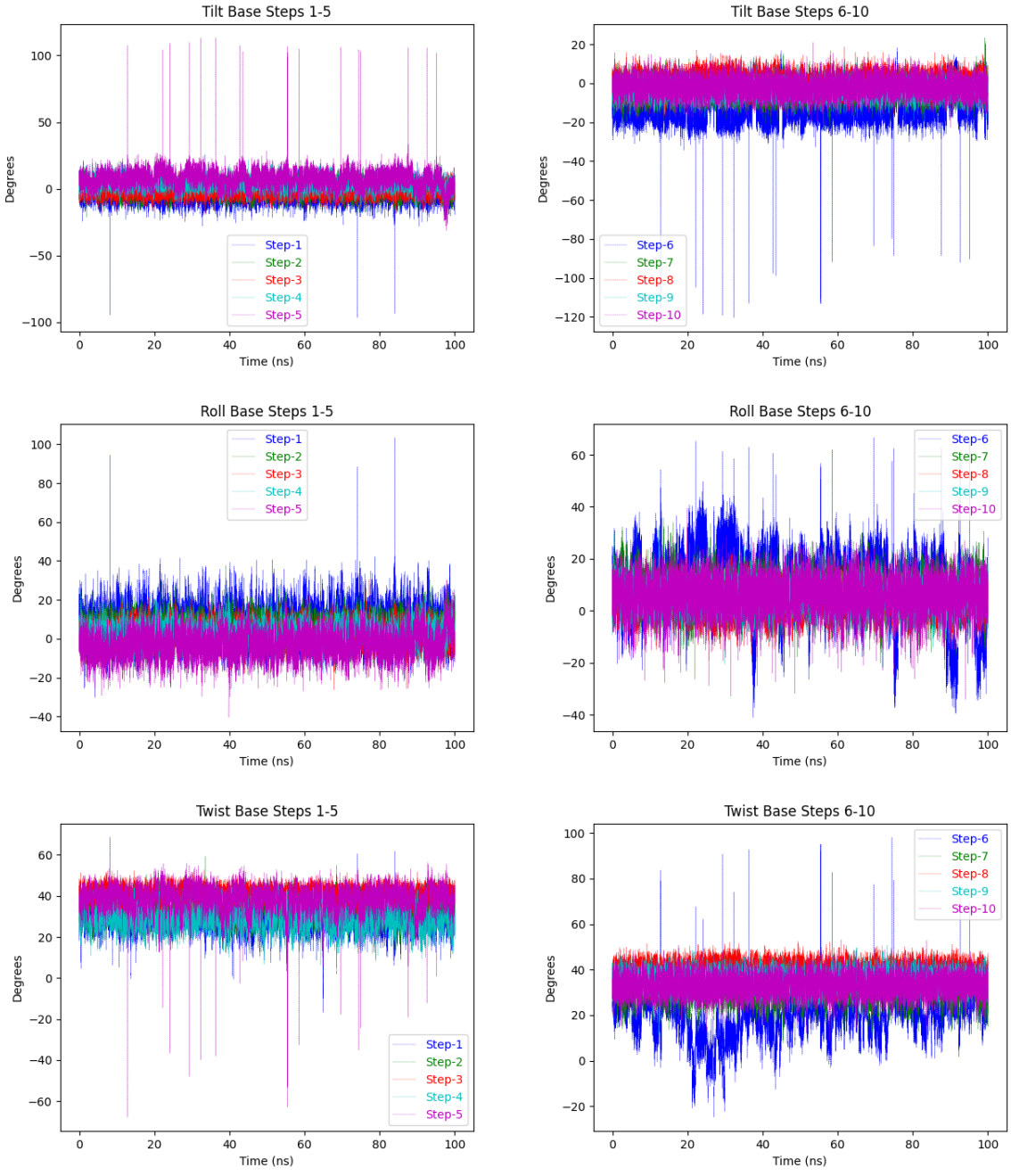
(6.230) B[e]P-DNA: Base pair trajectories



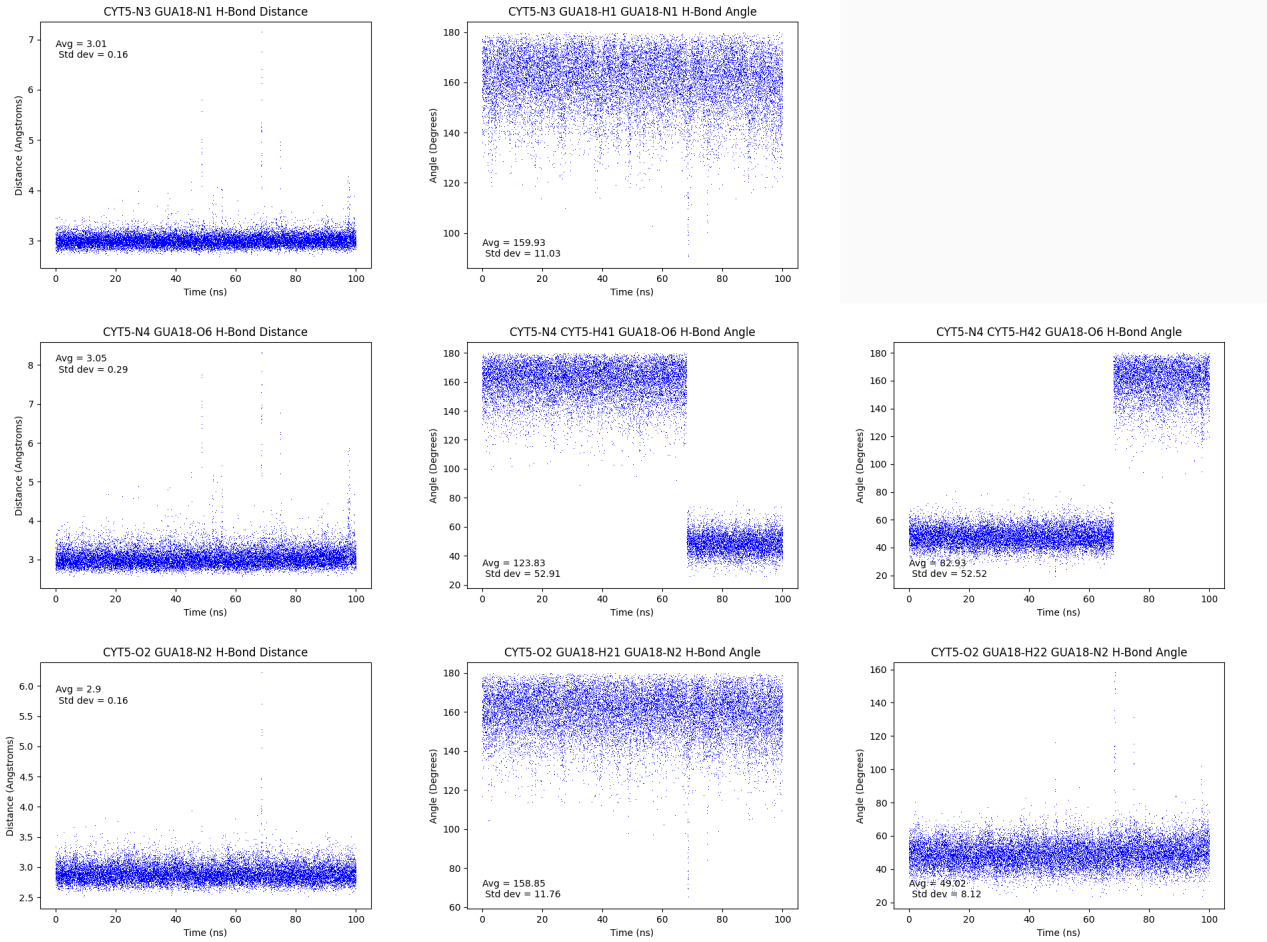
(6.231) B[e]P-DNA: Base pair trajectories



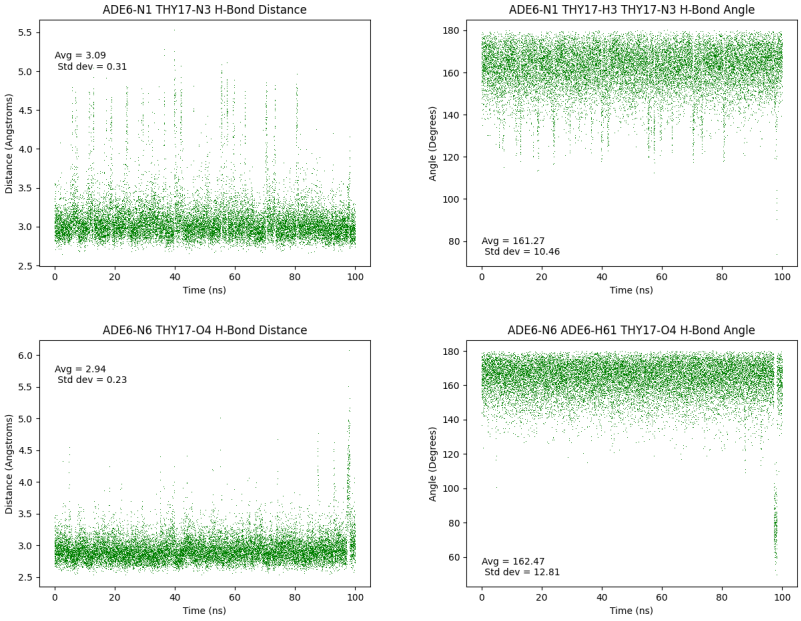
(6.232) B[e]P-DNA: Base step trajectories



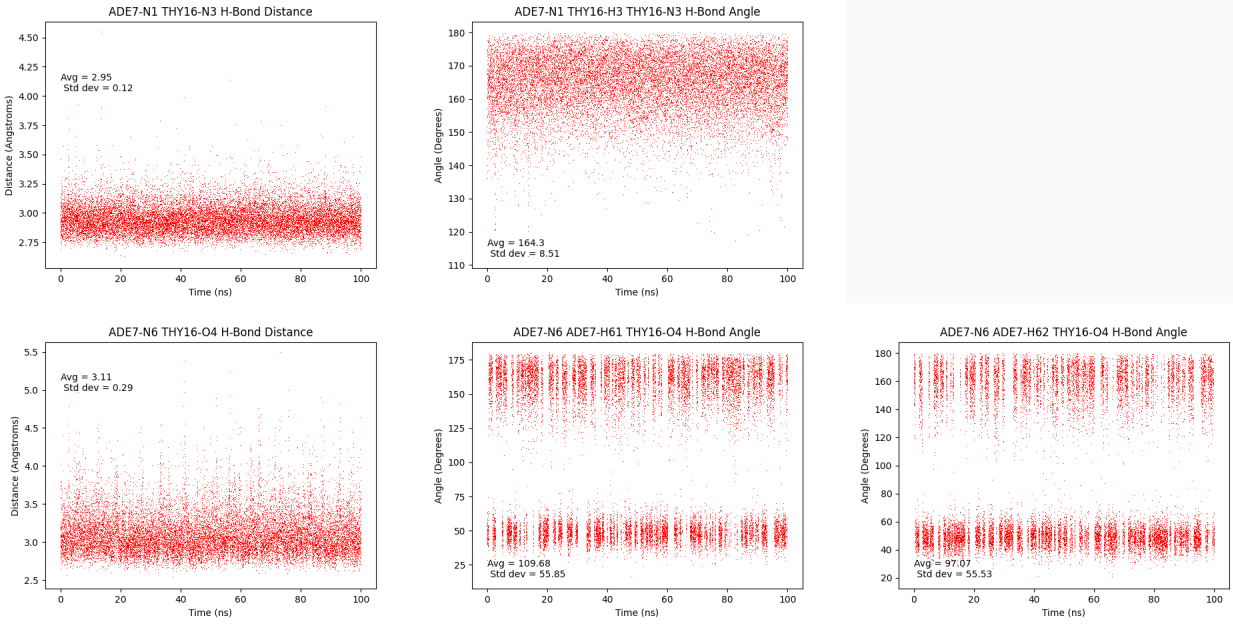
(6.233) B[e]P-DNA: Base step trajectories



(6.234) B[e]P-DNA: dC_5:dG_{18} hydrogen bond trajectories



(6.235) B[e]P-DNA: dA*$_6$:dT$_{17}$ hydrogen bond trajectories



(6.236) B[e]P-DNA: dA$_7$:dT$_{16}$ hydrogen bond trajectories

REFERENCES

[1] H. J. Yeh, J. M. Sayer, X. Liu, A. S. Altieri, R. A. Byrd, M. K. Lakshman, H. Yagi, E. J. Schurter, D. G. Gorenstein, and D. M. Jerina, "NMR Solution Structure of a Nonanucleotide Duplex with a dG Mismatch Opposite a 10S Adduct Derived from Trans Addition of a Deoxyadenosine N6-Amino Group to (+)-(7R,8S,9S,10R)-7,8-Dihydroxy-9,10-epoxy-7,8,9,10-tetrahydrobenzo [a] pyrene: An Unusual syn Glycosidic Torsion Angle at the Modified dA," *Biochemistry*, vol. 34, no. 41, pp. 13570–13581, 1995.

[2] "Iarc monographs on the identification of carcinogenic hazards to humans." https://monographs.iarc.fr/list-of-classifications. Accessed: 02-13-2022.

[3] "Iarc monographs on the evaluation of carcinogenic risks to humans. volume 92: Some non-heterocyclc polycyclic aromatic hydrocarbons and some related exposures," tech. rep., 2010.

[4] J. G. VanRooij, J. H. De Roos, M. M. Bodelier-Bade, and F. J. Jongeneelen, "Absorption of polycyclic aromatic hydrocarbons through human skin: Differences between anatomical sites and individuals," *J. Toxicol. Environ. Health*, vol. 38, no. 4, pp. 355–368, 1993.

[5] J. G. VanRooij, M. M. Bodelier-Bade, and F. J. Jongeneelen, "Estimation of the dermal and respiratory uptake of PAH among 12 coke oven workers," *Hum. Exp. Toxicol.*, vol. 12, no. 4, p. 352, 1993.

[6] E. Dybing, P. E. Schwarze, P. Nafstad, K. Victorin, and T. M. Penning, "Polycyclic aromatic hydrocarbons in ambient air and cancer ," in *Air Pollution and Cancer. IARC Scientific Publication No. 161* (K. Straif, A. Cohen, and J. Samet, eds.), pp. 75–94, International Agency for Research on Cancer, 2013.

[7] K. W. Fent, J. Eisenberg, D. Evans, D. Sammons, S. Robertson, C. Striley, J. Snawder, C. Mueller, V. Kochenderfer, J. Pleil, and M. Stiegel, "Niosh hhe - evaluation of dermal exposure to polycyclic aromatic hydrocarbons in fire fighters: Report no. 2010-0156-3196," tech. rep., 2013.

[8] D. J. Lee, T. Koru-Sengul, M. N. Hernandez, A. J. Caban-Martinez, L. A. McClure, J. A. Mackinnon, and E. N. Kobetz, "Cancer risk among career male and female Florida firefighters: Evidence from the Florida Firefighter Cancer Registry (1981-2014)," *Am. J. Ind. Med.*, vol. 63, no. 4, pp. 285–299, 2020.

[9] R. J. Tsai, S. E. Luckhaupt, P. Schumacher, R. D. Cress, D. M. Deapen, and G. M. Calvert, "Risk of cancer among firefighters in California, 1988-2007," *Am. J. Ind. Med.*, vol. 58, no. 7, pp. 715–729, 2015.

[10] R. D. Daniels, T. L. Kubale, J. H. Yiin, M. M. Dahm, T. R. Hales, D. Baris, S. H. Zahm, J. J. Beaumont, K. M. Waters, and L. E. Pinkerton, "Mortality and cancer incidence in a pooled

cohort of US fire fighters from San Francisco, Chicago and Philadelphia (1950-2009)," *Occup. Environ. Med.*, vol. 71, no. 6, pp. 388–397, 2014.

[11] K. W. Fent, J. Eisenberg, J. Snawder, D. Sammons, J. D. Pleil, M. A. Stiegel, C. Mueller, G. P. Horn, and J. Dalton, "Systemic exposure to pahs and benzene in firefighters suppressing controlled structure fires," *Ann. Occup. Hyg.*, vol. 58, no. 7, pp. 830–845, 2014.

[12] N. E. Geacintov and S. Broyde, "Repair-Resistant DNA Lesions," *Chem. Res. Toxicol.*, vol. 30, no. 8, pp. 1517–1548, 2017.

[13] A. Luch, "On the impact of the molecule structure in chemical carcinogenesis," in *Molecular, Clinical and Environmental Toxicology. Volume 1: Molecular Toxicology* (A. Luch, ed.), pp. 151–178, Birkhäuser Verlag, 2009.

[14] H. Zheng, Y. Cai, S. Ding, Y. Tang, K. Kropachev, Y. Zhou, L. Wang, S. Wang, N. E. Geacintov, Y. Zhang, and S. Broyde, "Base flipping free energy profiles for damaged and undamaged DNA," *Chem. Res. Toxicol.*, vol. 23, no. 12, pp. 1868–1870, 2010.

[15] J. H. Min and N. P. Pavletich, "Recognition of DNA damage by the Rad4 nucleotide excision repair protein," *Nature*, vol. 449, no. 7162, pp. 570–575, 2007.

[16] Y. Cai, H. Zheng, S. Ding, K. Kropachev, A. G. Schwaid, Y. Tang, H. Mu, S. Wang, N. E. Geacintov, Y. Zhang, and S. Broyde, "Free energy profiles of base flipping in intercalative polycyclic aromatic hydrocarbon-damaged DNA duplexes: Energetic and structural relationships to nucleotide excision repair susceptibility," *Chem. Res. Toxicol.*, vol. 26, no. 7, pp. 1115–1125, 2013.

[17] S. Broyde, L. Wang, Y. Cai, L. Jia, R. Shapiro, D. J. Patel, and N. E. Geacintov, "Covalent Polycyclic Aromatic Hydrocarbon–DNA Adducts: Carcinogenicity, Structure, and Function," in *Chemical Carcinogenesis* (T. M. Penning, ed.), ch. 9, pp. 181–207, Springer, 2011.

[18] C. Posch, M. Sanlorenzo, I. Vujic, and et. al, "Phosphoproteomic Analyses of NRAS(G12) and NRAS(Q61) Mutant Melanocytes Reveal Increased CK2a Kinase Levels in NRAS(Q61) Mutant Cells," *Journal of Investigative Dermatology*, vol. 136, pp. 2041–2048, 2016.

[19] H. Ling, J. M. Sayer, B. S. Plosky, H. Yagi, F. Boudsocq, R. Woodgate, D. M. Jerina, and W. Yang, "Crystal structure of a blenzo[a]pyrene diol epoxide adduct in a ternary complex with a DNA polymerase," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 8, pp. 2265–2269, 2004.

[20] B. Hwa Yun, J. Guo, M. Bellamri, and R. J. Turesky, "Dna adducts: Formation, biological effects, and new biospecimens for mass spectrometric measurements in humans," *Mass spectrometry reviews*, vol. 39, no. 1-2, pp. 55–82, 2020.

[21] S. Wei, R. L. Chang, C.-Q. Wong, N. Bhachech, X. X. Cui, E. Hennig, H. Yagi, J. M. Sayer, D. M. Jerina, and B. D. Preston, "Dose-dependent differences in the profile of mutations induced by an ultimate carcinogen from benzo [a] pyrene," *Proceedings of the National Academy of Sciences*, vol. 88, no. 24, pp. 11227–11230, 1991.

[22] S.-J. Wei, R. Chang, E. Hennig, X. Cui, K. Merkler, C.-Q. Wong, H. Yagi, D. Jerina, and A. Conney, "Mutagenic selectivity at the hprt locus in v-79 cells: comparison of mutations caused by bay-region benzo [a] pyrene 7, 8-diol-9, 10-epoxide enantiomers with high and low carcinogenic activity," *Carcinogenesis*, vol. 15, no. 8, pp. 1729–1735, 1994.

[23] D. Chakravarti, J. C. Pelling, E. L. Cavalieri, and E. G. Rogan, "Relating aromatic hydrocarbon-induced dna adducts and ch-ras mutations in mouse skin papillomas: the role of apurinic sites," *Proceedings of the National Academy of Sciences*, vol. 92, no. 22, pp. 10422–10426, 1995.

[24] D. Chakravarti, D. Venugopal, P. C. Mailander, J. L. Meza, S. Higginbotham, E. L. Cavalieri, and E. G. Rogan, "The role of polycyclic aromatic hydrocarbon–dna adducts in inducing mutations in mouse skin," *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, vol. 649, no. 1-2, pp. 161–178, 2008.

[25] B. Mahadevan, W.-M. Dashwood, A. Luch, A. Pecaj, J. Doehmer, A. Seidel, C. Pereira, and W. M. Baird, "Mutations induced by (-)-anti-11r, 12s-dihydrodiol 13s, 14r-epoxide of dibenzo [a, l] pyrene in the coding region of the hypoxanthine phosphoribosyltransferase (hprt) gene in chinese hamster v79 cells," *Environmental and molecular mutagenesis*, vol. 41, no. 2, pp. 131–139, 2003.

[26] J.-H. Yoon, A. Besaratinia, Z. Feng, M.-s. Tang, S. Amin, A. Luch, and G. P. Pfeifer, "Dna damage, repair, and mutation induction by (+)-syn and (-)-anti-dibenzo [a, l] pyrene-11, 12-diol-13, 14-epoxides in mouse cells," *Cancer research*, vol. 64, no. 20, pp. 7321–7328, 2004.

[27] H. Mu, N. E. Geacintov, Y. Zhang, and S. Broyde, "Recognition of Damaged DNA for Nucleotide Excision Repair: A Correlated Motion Mechanism with a Mismatched cis-syn Thymine Dimer Lesion," *Biochemistry*, vol. 54, no. 34, pp. 5263–5267, 2015.

[28] Y.-C. Lee, Y. Cai, H. Mu, S. Broyde, S. Amin, X. Chen, J.-H. Min, and N. E. Geacintov, "The relationships between xpc binding to conformationally diverse dna adducts and their excision by the human ner system: is there a correlation?," *DNA repair*, vol. 19, pp. 55–63, 2014.

[29] J. T. Andersson and C. Achten, "Time to Say Goodbye to the 16 EPA PAHs? Toward an Up-to-Date Use of PACs for Environmental Purposes," *Polycycl. Aromat. Compd.*, vol. 35, no. 2-4, pp. 330–354, 2015.

[30] "Epa priority pollutant list." https://www.epa.gov/sites/production/files/2015-09/documents/priority-pollutant-list-epa.pdf. Accessed: 02-13-2022.

[31] C. C. Valentine, R. R. Young, M. R. Fielden, R. Kulkarni, L. N. Williams, T. Li, S. Minocherhomji, and J. J. Salk, "Direct quantification of in vivo mutagenesis and carcinogenesis using duplex sequencing," *Proceedings of the National Academy of Sciences*, vol. 117, no. 52, pp. 33414–33425, 2020.

[32] Z. Cournia, B. Allen, and W. Sherman, "Relative binding free energy calculations in drug discovery: recent advances and practical considerations," *Journal of chemical information and modeling*, vol. 57, no. 12, pp. 2911–2937, 2017.

[33] E. J. Schurter, J. M. Sayer, T. Oh-hara, H. J. Yeh, H. Yagi, B. A. Luxon, D. M. Jerina, and D. G. Gorenstein, "Nuclear Magnetic Resonance Solution Structure of an Undecanucleotide Duplex with a Complementary Thymidine Base opposite a 10R Adduct Derived from Trans Addition of a Deoxyadenosine N6-Amino Group to (-)-(7R,8S,9R,10S)-7,8-Dihydroxy-9,10-epoxy-7,8,9,10-tetrahydrobenzo[a]pyrene," *Biochemistry*, vol. 34, no. 28, pp. 9009–9020, 1995.

[34] M. Cosman, R. Fiala, B. E. Hingerty, A. Laryea, H. Lee, R. G. Harvey, S. Amin, N. E. Geacintov, S. Broyde, and D. Patel, "Solution Conformation of the (+)-trans-anti-[BPh]dA Adduct opposite dT in a DNA Duplex: Intercalation of the Covalently Attached Benzo[c]phenanthrene to the 5'-Side of the Adduct Site without Disruption of the Modified Base Pair," *Biochemistry*, vol. 32, no. 46, pp. 12488–12497, 1993.

[35] M. Cosman, A. Laryea, R. Fiala, B. E. Hingerty, S. Amin, N. E. Geacintov, S. Broyde, and D. J. Patel, "Solution Conformation of the (-)-trans-anti-Benzo[c]phenanthrene-dA ([BPh]dA) Adduct opposite dT in a DNA Duplex: Intercalation of the Covalently Attached Benzo[c]phenanthrenyl Ring to the 3'-Side of the Adduct Site and Comparison with the (+)-trans-anti-[BPh]dA opposite dT Stereoisomer," *Biochemistry*, vol. 34, no. 4, pp. 1295–1307, 1995.

[36] Y. Cai, S. Ding, N. E. Geacintov, and S. Broyde, "Intercalative conformations of the 14R(+)- and 14S(-)- trans-anti-DB[a,l]P- N6-dA adducts: Molecular modeling and MD simulations," *Chem. Res. Toxicol.*, vol. 24, no. 4, pp. 522–531, 2011.

[37] Y. Cai, N. E. Geacintov, and S. Broyde, "Nucleotide excision repair efficiencies of bulky carcinogen-DNA adducts are governed by a balance between stabilizing and destabilizing interactions," *Biochemistry*, vol. 51, no. 7, pp. 1486–1499, 2012.

[38] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general Amber force field," *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, 2004.

[39] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations," *J. Mol. Graph. Model.*, vol. 25, no. 2, pp. 247–260, 2006.

[40] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. Mackerell Jr, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *J. Comput. Chem.*, vol. 31, no. 4, pp. 671–690, 2010.

[41] K. Vanommeslaeghe and A. MacKerell Jr, "Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing," *J. Chem. Inf. Model.*, vol. 52, pp. 3144–3154, 2012.

[42] K. Vanommeslaeghe, E. Raman, and A. MacKerell Jr, "Automation of the CHARMM General Force Field (CGenFF) II: Assignment of bonded parameters and partial atomic charges," *J. Chem. Inf. Model.*, vol. 52, pp. 3155–3168, 2012.

[43] V. Zoete, M. A. Cuendet, A. Grosdidier, and O. Michielin, "SwissParam: A Fast Force Field Generation Tool for Small Organic Molecules," *J. Comput. Chem.*, vol. 32, no. 11, pp. 2359–2368, 2011.

[44] C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, and J. C. Gumbart, "Rapid parameterization of small molecules using the force field toolkit," *J. Comput. Chem.*, vol. 34, no. 32, pp. 2757–2770, 2013.

[45] A. Kumar, O. Yoluk, and A. D. MacKerell Jr., "FFParam: Standalone package for CHARMM additive and Drude polarizable force field parametrization of small molecules," *J. Comput. Chem.*, vol. 41, no. 9, pp. 958–970, 2020.

[46] C. Grill and L. Larue, "NRAS, NRAS, Which Mutation is Fairest of Them All?," *Journal of Investigative Dermatology*, vol. 136, pp. 1936–1938, 2016.

[47] K. Vanommeslaeghe, M. Yang, and A. D. Mackerell, "Robustness in the fitting of molecular mechanics parameters," *J. Comput. Chem.*, vol. 36, pp. 1083–1101, may 2015.

[48] O. Guvench and A. D. MacKerell, "Automated conformational energy fitting for force-field development," *J. Mol. Model.*, vol. 14, no. 8, pp. 667–679, 2008.

[49] J. L. Schwartz, J. S. Rice, B. A. Luxon, J. M. Sayer, G. Xie, H. J. Yeh, X. Liu, D. M. Jerina, and D. G. Gorenstein, "Solution structure of the minor conformer of a DNA duplex containing a dG mismatch opposite a benzo[a]pyrene diol epoxide/dA adduct: Glycosidic rotation from syn to anti at the modified deoxyadenosine," *Biochemistry*, vol. 36, no. 37, pp. 11069–11076, 1997.

[50] N. Foloppe and A. D. MacKerell, "All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data," *J. Comput. Chem.*, vol. 21, no. 2, pp. 86–104, 2000.

[51] A. D. MacKerell and N. K. Banavali, "All-Atom Empirical Force Field for Nucleic Acids: II. Application to Molecular Dynamics Simulations of DNA and RNA in Solution," *J. Comput. Chem.*, vol. 21, no. 2, pp. 105–120, 2000.

[52] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, "Gaussian~16 Revision C.01," 2016. Gaussian Inc. Wallingford CT.

[53] H. Mu, N. E. Geacintov, J. H. Min, Y. Zhang, and S. Broyde, "Nucleotide Excision Repair Lesion-Recognition Protein Rad4 Captures a Pre-Flipped Partner Base in a Benzo[a]pyrene-Derived DNA Lesion: How Structure Impacts the Binding Pathway," *Chem. Res. Toxicol.*, vol. 30, no. 6, pp. 1344–1354, 2017.

[54] E. Glendening, J. Badenhoop, A. Reed, J. Carpenter, J. Bohmann, C. Morales, P. Karafiloglou, C. Landis, and F. Weinhold, "Nbo7.0." Theoretical Chemistry Institute, University of Wisconsin, Madison, 2018.

[55] A. D. MacKerell, "The CHARMM Force Field-CECAM Workshop: Advances in Biomolecular Modelling and Simulations using CHARMM." https://mackerell.umaryland.edu/ kenno/cgenff/downloader.php?filename=CHARMM_FF_-Mackerell4.pdf, June 2012. Accessed: 02-13-2022.

[56] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–1802, 2005.

[57] J. C. Phillips, D. J. Hardy, J. D. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid, "Scalable molecular dynamics on CPU and GPU architectures with NAMD," *J. Chem. Phys.*, vol. 153, no. 4, 2020.

[58] K. Vanommeslaeghe, "CGenFF FAQs." https://mackerell.umaryland.edu/ kenno/cgenff/faq.-php#compile. Accessed: 02-13-2022.

[59] C. W. Hopkins and A. E. Roitberg, "Fitting of dihedral terms in classical force fields as an analytic linear least-squares problem," *J. Chem. Inf. Model.*, vol. 54, no. 7, pp. 1978–1986, 2014.

[60] G. Golub and C. Van Loan, *Matrix Computations*. Baltimore: Johns Hopkins University Press, 1996.

[61] J. Demmel, *Applied Numerical Linear Algebra*. Philadelphia: Siam, 1997.

[62] W. Humphrey, A. Dalke, and K. Schulten, "VMD - Visual Molecular Dynamics," *J. Molec. Graphics*, vol. 14, pp. 33–38, 1996.

[63] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

[64] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks, "Constant pressure molecular dynamics simulation: The Langevin piston method.," *J. Comput. Phys.*, vol. 103, no. 11, pp. 4613–4621, 1995.

[65] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993.

[66] H. C. Andersen, "Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations," *J. Comput. Phys.*, vol. 52, no. 1, pp. 24–34, 1983.

[67] S. Miyamoto and P. A. Kollman, "Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models," *J. Comput. Chem.*, vol. 13, no. 8, pp. 952–962, 1992.

[68] V. Minhas, T. Sun, A. Mirzoev, N. Korolev, A. Lyubartsev, and L. Nordenskiold, "Modeling DNA Flexibility: Comparison of Force Fields from Atomistic to Multiscale Levels," *J. Phys. Chem. B*, vol. 124, pp. 38–49, 2020.

[69] S. Dasgupta, T. Yamasaki, and W. A. Goddard III, "The Hessian Biased Singular Value Decomposition Method for Optimization and Analysis of Force Fields," *J. Chem. Phys.*, vol. 104, pp. 2898–2920, 1996.

[70] L. Elden, "Algorithms for Regularization of Ill-Conditioned Least Squares Problems," *BIT*, vol. 17, pp. 134–145, 1977.

[71] P. C. Hansen, "The Truncated SVD as a Method for Regularization," *BIT*, vol. 27, pp. 534–553, 1987.

[72] P. C. Hansen, "Truncated Singular Value Decomposition Solutions to Discrete Ill-Posed Problems with Ill-Determined Numerical Rank," *SIAM J. Sci. Stat. Comput.*, vol. 11, pp. 503–518, 1990.

[73] A. N. Tikhonov, "Solution of Incorrectly Formulated Problems and the Regularization Method.," *Dokl. Akad. Nauk. SSSR*, vol. 151, pp. 501–504, 1963.

[74] D. L. Phillips, "A Technique for the Numerical Solution of Certain Integral Equations of the First Kind.," *J. ACM*, vol. 9, pp. 84–97, 1962.

[75] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice Hall, 1974.

[76] J. M. Varah, "On the Numerical Solution of Ill-conditioned Linear Systems with Applications to Ill-Posed Problems.," *SIAM J. Numer. Anal.*, vol. 10, pp. 257–267, 1973.

[77] J. M. Varah, "A Practical Examination of Some Numerical Methods for Linear Discrete Ill-Posed Problems.," *SIAM Rev.*, vol. 21, pp. 100–111, 1979.

[78] J. M. Varah, "Pitfalls in the Numerical Solution of Linear Ill-Posed Problems.," *SIAM J. Sci. statist. Comput.*, vol. 4, pp. 164–176, 1983.

[79] V. Gapsys, S. Michielssens, J. H. Peters, B. L. d. Groot, and H. Leonov, "Calculation of binding free energies," in *Molecular Modeling of Proteins*, pp. 173–209, Springer, 2015.

[80] C. Chipot and A. Pohorille, "Free energy calculations," *Springer series in chemical physics*, vol. 86, pp. 159–184, 2007.

[81] J. Henin, J. Gumbart, and C. Chipot, "In Silico Alchemy: A Tutorial for Alchemical Free-Energy Perturbation Calculations with NAMD." https://www.ks.uiuc.edu/Training/Tutorials/namd/FEP/tutorial-FEP.pdf, Sept 2017. Accessed: 02-13-2022.

[82] P. Liu, F. Dehez, W. Cai, and C. Chipot, "A toolkit for the analysis of free-energy perturbation calculations," *Journal of chemical theory and computation*, vol. 8, no. 8, pp. 2606–2616, 2012.

[83] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. Van Gunsteren, "Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations," *Chemical physics letters*, vol. 222, no. 6, pp. 529–539, 1994.

[84] Z. Li, H.-Y. Kim, P. J. Tamura, C. M. Harris, T. M. Harris, and M. P. Stone, "Intercalation of the (1S,2R,3S,4R)-N6-[1-(1,2,3,4-Tetrahydro-2,3,4-trihydroxybenz[a]anthracenyl)]-2'-deoxyadenosyl Adduct in an Oligodeoxynucleotide Containing the Human NRAS Codon 61 Sequence," *Biochemistry*, vol. 38, pp. 16045–16057, 1999.

[85] W. Levin, R. L. Chang, A. W. Wood, H. Yagi, D. R. Thakker, D. M. Jerina, and A. H. Conney, "High Stereoselectivity among the Optical Isomers of the Diastereomeric Bay-Region Diol-Epoxides of Benz(a)anthracene in the Expression of Tumorigenic Activity in Murine Tumor Models," *Cancer Res.*, vol. 44, pp. 929–933, 1984.

[86] E. Pettersen, T. Goddar, C. Huang, G. Couch, D. Greenblatt, E. Meng, and T. Ferrin, "UCSF Chimera - A Visualization System for Exploratory Research and Analysis," *J. Comput. Chem.*, vol. 13, pp. 1605–1612, 2004.

[87] D. Paul, H. Mu, H. Zhao, O. Ouerfelli, P. D. Jeffrey, S. Broyde, and J.-H. Min, "Nar breakthrough article structure and mechanism of pyrimidine–pyrimidone," *Nucleic Acids Research*, vol. 47, no. 12, pp. 6015–6028, 2019.

[88] V. Gapsys, M. Khabiri, B. L. de Groot, and P. L. Freddolino, "Comment on "deficiencies in molecular dynamics simulation-based prediction of protein-dna binding free energy landscapes"," *The Journal of Physical Chemistry B*, vol. 124, no. 6, pp. 1115–1123, 2018.

[89] W. Jiang, C. Chipot, and B. Roux, "Computing relative binding affinity of ligands to receptor: An effective hybrid single-dual-topology free-energy perturbation approach in namd," *Journal of chemical information and modeling*, vol. 59, no. 9, pp. 3794–3802, 2019.

[90] S. Liu, L. Wang, and D. L. Mobley, "Is ring breaking feasible in relative binding free energy calculations?," *Journal of chemical information and modeling*, vol. 55, no. 4, pp. 727–735, 2015.

[91] E. J. Denning, U. D. Priyakumar, L. Nilsson, and A. D. Mackerell Jr, "Impact of 2-hydroxyl sampling on the conformational properties of rna: update of the charmm all-atom additive force field for rna," *Journal of computational chemistry*, vol. 32, no. 9, pp. 1929–1943, 2011.

[92] "Implementation of the free energy methods in namd." https://www.ks.uiuc.edu/Research/namd/2.10/ug/node62.html. Accessed: 02-13-2022.

[93] C. H. Bennett, "Efficient estimation of free energy differences from monte carlo data," *Journal of Computational Physics*, vol. 22, no. 2, pp. 245–268, 1976.

[94] X.-J. Lu and W. K. Olson, "3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures," *Nucleic acids research*, vol. 31, no. 17, pp. 5108–5121, 2003.

[95] X.-J. Lu and W. K. Olson, "3dna: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures," *Nature protocols*, vol. 3, no. 7, pp. 1213–1227, 2008.

[96] A. V. Colasanti, X.-J. Lu, and W. K. Olson, "Analyzing and building nucleic acid structures with 3dna," *JoVE (Journal of Visualized Experiments)*, no. 74, p. e4401, 2013.

[97] M. El Hassan and C. Calladine, "Two distinct modes of protein-induced bending in dna," *Journal of molecular biology*, vol. 282, no. 2, pp. 331–343, 1998.

[98] A. Hospital, I. Faustino, R. Collepardo-Guevara, C. Gonzalez, J. L. Gelpí, and M. Orozco, "Naflex: a web server for the study of nucleic acid flexibility," *Nucleic Acids Research*, vol. 41, no. W1, pp. W47–W55, 2013.