

Conformational Dynamics of DNA and Protein-DNA Complexes at Single-Stranded-Double-Stranded DNA Junctions

by

Jack Maurer

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Chemistry

Dissertation Committee:

Jeffery Cina, Chair

Andrew H. Marcus, Advisor

Peter von Hippel, Co-Advisor

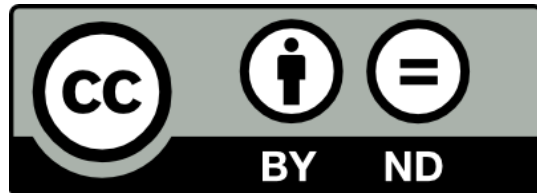
Julia Widom, Core Member

Brian Smith, Institutional Representative

University of Oregon

Fall 2023

© 2023 Jack Maurer



DISSERTATION ABSTRACT:

Jack Maurer

Doctor of Philosophy in Chemistry

Title: Conformational Dynamics of DNA and Protein-DNA Complexes at Single-Stranded-Double-Stranded DNA Junctions

Most biological systems, particularly protein-DNA complexes, leverage a dynamic evolution of their structure to perform a myriad of functions within the context of the cell. Decades of detailed biophysical research have established that the intricacies of such systems stem heavily from their dynamic evolution, abandoning the previous notion of a purely static ‘structure-function’ relationship. This dissertation introduces a new polarization-sensitive methodology for studying the dynamic evolution of local conformation in single-molecules of dsDNA containing an i(Cy3)<sub>2</sub> dimer. The methodology developed during this dissertation is applied to DNA under a variety of experimental conditions as well as protein-DNA complexes. A massively parallel computational pipeline was developed in the course of this work to aid the optimization of kinetic network models, which forms the basis for all current analyses of single-molecule data in the Marcus and von Hippel lab. The primary discovery of this work is the persistence of four relevant conformational macrostates in DNA only systems and five relevant conformational macrostates in the protein-DNA systems examined. The thermodynamic and mechanical stability of these systems is analyzed in detail and structural mechanisms are proposed to merge the observed dynamics with hypothesized local conformations during the dynamic evolution of these ubiquitous biological systems.

This dissertation contains previously published co-authored material.

## CURRICULUM VITAE

NAME OF THE AUTHOR: Jack Maurer

### GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene OR.

University of Denver, Denver CO

### DEGREES AWARDED

Doctor of Philosophy, Chemistry, 2023, University of Oregon

Bachelor of Science, Physics, 2016, University of Denver

Bachelor of Science, Chemistry, 2016, University of Denver

### SPECIAL AREAS OF INTEREST

Single-Molecule Biophysics

Optical Spectroscopy and Metrology

Protein-DNA Interactions

### PROFESSIONAL EXPERIENCE

Physical Chemistry Undergraduate Teaching Assistant, Spring 2022,2023

Undergraduate General Chemistry Teaching Assistant, UO, 2016-2017

Technical Sales Associate, Thyssenkrupp, 2016-2017

### GRANTS, AWARDS AND HONORS

Emmanuel Fellowship, University of Oregon, 2023

Raymund Fellowship, University of Oregon, 2017-2023

Mao Fellowship, University of Oregon, 2017-2023

Dean's First Year Merit Award, University of Oregon, 2017



Outstanding Senior in Physics and Astronomy, University of Denver, 2016

Recipient of Partners in Natural Sciences Grant, University of Denver, 2014, 2015

Dean's List, University of Denver, 2012-2016

## PUBLICATIONS

Marcus, A. H., Heussman, D., Maurer, J., Albrecht, C. S., Herbert, P., & von Hippel, P. H. (2023). Studies of Local DNA Backbone Conformation and Conformational Disorder Using Site-Specific Exciton-Coupled Dimer Probe Spectroscopy. *Annual Review of Physical Chemistry*. Vol 74.

Maurer, J., Albrecht, C.S. , Heussman, D., Herbert, P., von Hippel, P.H. & Marcus, A.H. (2023). Polarization-sweep single-molecule fluorescence microscopy of exciton-coupled (Cy3)<sub>2</sub> dimer-labeled DNA fork constructs. *In Press JCPB*.

Maurer, J., Albrecht, C.S. , von Hippel, P.H. & Marcus, A.H. (2023). 'DNA Breathing' free energy landscape determination via kinetic network analysis of polarization-sweep single-molecule fluorescence microscopy data. *In Preparation*.

Maurer, J., Albrecht, C.S. , P., von Hippel, P.H. & Marcus, A.H. (2023). Multi-state kinetic network modeling of time resolved single-molecule data using a generalized master equation approach. *In Preparation*.

Herbert, P., Maurer, J., Heussman, D., von Hippel, P.H. & Marcus, A.H (2023). Investigating local DNA conformational trapping dynamics during DNA polymerase holoenzyme assembly and exonuclease proof-reading activity. *In Preparation*.

Albrecht, C.S. , Israels, B., Maurer, J., P., von Hippel, P.H. & Marcus, A.H. (2024). Sub-millisecond Single-Molecule FRET Studies of Single-Stranded DNA Conformation Fluctuations Mediated by Single-Stranded DNA Binding Proteins. *Work in Progress*.

## ACKNOWLEDGEMENTS

I am both grateful and lucky to have been a member of the Marcus and von Hippel group for the last six years. I would like to thank all current and former lab members for their advice, insights, and general inspiration over the course of my thesis work. It truly has been a rich and rewarding environment in which to learn and grow as a scientist. I would like to thank my primary advisor, Andy Marcus, for his dedicated support of my research efforts, his willingness to let me explore and investigate new scientific directions and for the pertinent advice he's offered to me which has largely shaped my scientific outlook and thought process. I would also like to thank my co-advisor Peter von Hippel for his constant optimism about life, excitement over the possibilities held by science, pointed research advice, and generosity in supporting both my wife and I during our time in Eugene. Additionally, a special thank you to Larry Scatena for all his help and guidance.

I've been fortunate to have fostered many long-lasting relationships during my time at the University of Oregon. Many of these relationships were forged over a mutual love of food, the outdoors, science, and more often than I would have anticipated, overly complicated board games. To those friends who I met during my time here, thank you for your company, support and all the treasured memories I will hold onto for years to come.

Lastly, thank you to my wife and my family. Your willingness to stand by me on this journey has meant everything to me. Your patience and tolerance for my often overly exuberant personality hasn't gone unnoticed. I believe it is an enormous privilege to pursue a PhD, one which so many around the world can never afford, so thank you for getting me here and continuing to support me along the way. I wouldn't have made it this far without you.

This dissertation is dedicated to my wife and family.

## TABLE OF CONTENTS

Chapter 1 : INTRODUCTION.....	19
1. OVERVIEW.....	19
2. BACKGROUND INFORMATION.....	20
3. SINGLE MOLECULE APPROACHES.....	24
4. MACROMOLECULAR STRUCTURE AND DYNAMICS .....	28
Chapter 2 : POLARIZATION SWEEP SINGLE-MOLECULE FLUORESCENCE MICROSCOPY.....	31
1. OVERVIEW.....	31
2. INTRODUCTION .....	31
3. MATERIALS AND METHODS .....	38
Section 3.1. (iCy3) <sub>2</sub> dimer-labeled single-stranded (ss) – double-stranded (ds) DNA fork constructs.....	38
Section 3.2. Sample preparation.....	40
Section 3.3. Polarization-sweep single-molecule fluorescence (PS-SMF) microscopy experimental setup.....	40
Section 3.4 Theoretical description of Jones vector electric field components.....	43
Section 3.5 Derivation of the polarized signal intensity for PS-SMF experiments .....	46
Section 3.6. Characterization of conformational macrostates using probability distribution functions (PDFs).....	52

Section 3.7 Characterization of conformational dynamics using multi-point time-correlation functions (TCFs).....	57
Section 3.8. Consideration of fluorescence correlation spectroscopy (FCS) as a reporter on conformational dynamics of (iCy3) <sub>2</sub> dimer-labeled single-stranded (ss) – double-stranded (ds) DNA fork constructs .....	63
4. RESULTS AND DISCUSSION.....	70
Section 4.1. Analysis of PS-SMF spectroscopic signals. ....	70
Section 4.2. Studies of local DNA breathing of +1, -1 and -2 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA fork constructs. ....	81
Section 4.3. Salt concentration-dependent breathing of +1 and -2 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA fork constructs. ....	86
5. CONCLUSION.....	92
Chapter 3 : SIMULATING A KINETIC NETWORK TO OBTAIN AN EXPERIMENTALLY DERIVED FREE ENERGY LANDSCAPE.....	96
1. OVERVIEW.....	96
2. INTRODUCTION .....	96
3. COMPUTATIONAL METHODS.....	97
Section 3.1 Simulating a kinetic network using Markov chains .....	97
Section 3.2 Estimating the number of macrostates and minimally necessary model complexity .....	101

Section 3.3 Optimization methods for application of a kinetic network model to experimental single-molecule data .....	104
Section 3.4 Details of calculating the chi-squared across three surfaces .....	115
4. CONCLUSION.....	129
Chapter 4 : ANALYSIS OF PS-SMF DATA FROM SS-DSDNA JUNCTION USING A KINETIC NETWORK MODEL .....	131
1. OVERVIEW.....	131
2. INTRODUCTION .....	131
3. MATERIALS AND METHODS .....	134
4. RESULTS AND DISCUSSION.....	135
4.1 Positional dependence of DNA replication fork free energy landscapes .....	135
4.2 Salt dependence of DNA replication fork free energy landscapes .....	142
5. CONCLUSION.....	152
5.1 Proposed structural model of conformational macrostates at replication fork junctions .....	152
Chapter 5 : APPLICATION OF PS-SMF TO PROTEIN-DNA COMPLEXES.....	158
1. OVERVIEW.....	158
2. INTRODUCTION .....	159
3. MATERIALS AND METHODS .....	164
4. RESULTS AND DISCUSSION.....	166

Section 4.1 Duplex region studies of DNA p/t junctions .....	166
Section 4.2 Protein-DNA complexes studies at and near DNA p/t junctions.....	177
5. CONCLUSION.....	187
Chapter 6 : CONCLUDING SUMMARY .....	193
Appendix: INVESTIGATIONS INTO BROADBAND INTERFEROMETRIC MEASUREMENTS PERFORMED ON SINGLE MOLECULES.....	195

## LIST OF FIGURES

<b>Figure 1.1</b> Depiction of the full T4 bacteriophage replisome and its various protein complexes .....	22
<b>Figure 1.2</b> Depiction of various ss-dsDNA junction topologies. ....	23
<b>Figure 1.3</b> Structure of dsDNA. ....	27
<b>Figure 2.1</b> Labeling chemistry, nomenclature and structural parameters of the internal (iCy3) <sub>2</sub> dimer probes.....	33
<b>Figure 2.2</b> The PS-SMF microscopy experimental layout.....	34
<b>Figure 2.3</b> Schematic of the interferometer and integrated phase-tagged photon counting (PTPC) electronics .....	43
<b>Figure 2.4</b> Control measurements of phase histograms.....	52
<b>Figure 2.5</b> Model demonstration of variable integration time on the emergence of multimodal behavior.....	56
<b>Figure 2.6</b> Experimental demonstration of variable integration time on the emergence of multimodal behavior .....	57
<b>Figure 2.7</b> PS-SMF probability distribution functions (PDFs) for the +1 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA fork constructs.....	64
<b>Figure 2.8</b> PS-SMF probability distribution functions (PDFs) for the +1 iCy3 monomer- labeled ss-dsDNA fork construct.....	65
<b>Figure 2.9</b> PS-SMF time correlation functions (TCFs) for an anisotropic Cy3-film .....	68
<b>Figure 2.10</b> PS-SMF time correlation functions (TCFs) for an isotropic rhodamine sample .....	69
<b>Figure 2.11</b> Example PS-SMF signal trajectories .....	72



<b>Figure 2.12</b> PDFs of the flux, visibility and phase for the +1 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA fork construct. ....	74
<b>Figure 2.13</b> Probability distribution functions (PDFs) of the visibility for the +1 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA construct, and the +1 iCy3 monomer-labeled ss-dsDNA construct .....	76
<b>Figure 2.14</b> Joint PDFs of the flux and visibility signals. ....	77
<b>Figure 2.15</b> Two-point time correlation functions (TCFs) of the visibility and ( <i>B</i> , <i>D</i> ) the flux for the +1 iCy3 monomer-labeled ss-dsDNA construct and the +1 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA construct .....	79
<b>Figure 2.16</b> Probability distribution functions, two-point time-correlation functions and three-point TCFs of the PS-SMF visibility for the +1, -1 and -2 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA fork constructs.. ....	83
<b>Figure 2.17</b> Salt concentration-dependent PDFs, two-point TCFs and three-point TCFs of the signal visibility for the +1 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA construct.....	87
<b>Figure 2.18</b> Salt concentration-dependent PDFs, two-point TCFs and three-point TCFs of the signal visibility for the -2 (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA construct.....	88
<b>Figure 2.19</b> Schematic diagram illustrating hypothesized mechanism of salt-induced instability of (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA fork constructs .....	95
<b>Figure 3.1</b> An explicit form of the rate matrix. ....	105
<b>Figure 3.2</b> Comparison of the results from sampling two separate distributions utilizing an alternate sampling method.....	110
<b>Figure 3.3</b> A diagrammatic depiction of the genetic algorithm workflow. ....	113
<b>Figure 3.4</b> Example of a ‘good’ fit to the data from a model outcome. ....	121
<b>Figure 3.5</b> Weight surfaces applied to the C2 and C3 in ‘good’ fit case.....	121

<b>Figure 3.6</b> ‘Chi-Stats’ overview for reporting on statistical measures of ‘true’ error in the case of a ‘good’ fit.....	122
<b>Figure 3.7</b> ‘Chi-Stats’ overview for reporting on statistical measures of ‘weighted’ error in the case of a ‘good’ fit .....	123
<b>Figure 3.8</b> Example of a ‘bad’ fit to the data from a model outcome. ....	124
<b>Figure 3.9</b> Weight surfaces applied to the C2 and C3 in ‘bad’ fit case. ....	124
<b>Figure 3.10</b> ‘Chi-Stats’ overview for reporting on statistical measures of ‘true’ error in the case of a ‘bad’ fit. ....	125
<b>Figure 3.11</b> ‘Chi-Stats’ overview for reporting on statistical measures of ‘weighted’ error in the case of a ‘bad’ fit.....	126
<b>Figure 4.1</b> Optimized fits from our kinetic network analysis for the position dependence of DNA ‘breathing’ under physiological salt conditions. ....	136
<b>Figure 4.2</b> Depiction of the various modes and extents of DNA ‘breathing’ one might expect . .....	137
<b>Figure 4.3</b> Free energy landscapes and associated kinetic networks for each of the positions examined under physiological salt conditions.. ....	140
<b>Figure 4.4</b> Summary diagrams for the thermodynamic and mechanical stability of the +1, -1 and -2 positions under physiological salt conditions. ....	141
<b>Figure 4.5</b> Optimized fits from our kinetic network analysis for the salt dependence of DNA ‘breathing’ at the +1 position. ....	143
<b>Figure 4.6</b> Free energy landscapes and associated kinetic networks for each of the salt conditions examined at the +1 position.. ....	146

<b>Figure 4.7</b> Optimized fits from our kinetic network analysis for the salt dependence of DNA ‘breathing’ at the -2 position.....	148
<b>Figure 4.8</b> Free energy landscapes and associated kinetic networks for each of the salt conditions examined at the -2 position.....	149
<b>Figure 4.9</b> Summary diagrams for the thermodynamic and mechanical stability of the +1 and -2 and positions under a series of salt conditions.....	151
<b>Figure 4.10</b> Proposed structural model based on the results of our kinetic network analysis of replication fork junctions under numerous solvent conditions. ....	154
<b>Figure 5.1</b> Schematic depiction of the clamp-clamp loader reaction cycle .....	161
<b>Figure 5.2</b> Proposed kinetic scheme obtained of the DNA holoenzyme .....	163
<b>Figure 5.3</b> Linear absorbance of duplex labeled p/t junction constructs.....	167
<b>Figure 5.4 Panels A-E:</b> Experimental single-molecule data and optimized fits for the labeled p/t junction constructs. ....	168
<b>Figure 5.5</b> Linear absorbance and CD spectra of +15, +3, and +1 constructs upon the addition of gp45 clamp and activated gp44/62 clamp loader.....	178
<b>Figure 5.6</b> Experimental data and optimized fits for the +1 dimer and clamp-clamp loader complex.....	179
<b>Figure 5.7</b> CD spectra of gp43 binding to +4 p/t construct upon addition of increasing $Mg^{2+}$ concentration.....	183
<b>Figure 5.8</b> Experimental data and optimized fits for the +4 dimer and DNA polymerase complex.....	185
<b>Figure A.1</b> Schematic of the experimental apparatus used for ultrafast linear and nonlinear single molecule measurements.....	201

<b>Figure A.2</b> Two representative linear interferometric experiments on single molecules of dsDNA containing a $i(\text{Cy}3)_2$ dimer. ....	211
<b>Figure A.3</b> A series of linear Fourier transforms are shown as a function of integration time...	213
<b>Figure A.4</b> Two dimensional spectra obtained on a single dsDNA molecule containing a monomer of Cy3.....	215
<b>Figure A.5</b> Real part of the rephasing spectrum obtained on an ensemble of dsDNA molecules containing a monomer of Cy3. ....	216
<b>Figure A.6</b> Rephasing 2D spectra obtained from single Cy3 monomers incorporated into dsDNA.....	217
<b>Figure A.7</b> Rephasing 2D spectra obtained from single Cy3 dimers incorporated into dsDNA.....	217
<b>Figure A.8</b> Scaling of Nonrephasing and Rephasing signals for variable photon number.....	220
<b>Figure A.9</b> Intensity scaling of the linear and nonlinear signal magnitudes from a Cy3 monomer in dsDNA.....	221
<b>Figure A.10</b> Intensity scaling of the linear and nonlinear signal magnitudes from a Cy3 dimer in dsDNA.....	222
<b>Figure A.11</b> Experimental apparatus for broadband polarization-sweep measurements. ....	223
<b>Figure A.12</b> Simulated linear signal components rotating versus non-rotating polarization .....	224

<b>Figure A.13</b> Simulated signal contrast between two oppositely handed conformers in the case of rotating versus non-rotating polarization.....	225
<b>Figure A.14</b> A histogram of phase-tags obtained on an isotropic sample of rhodamine using a broadband polarization sweep experimental setup.....	227
<b>Figure A.15</b> Non-rotating polarization time-correlation function analysis of the signal visibility for rhodamine and a single molecule containing a Cy3 dimer.....	228
<b>Figure A.16</b> Non-rotating polarization time-correlation function analysis of the signal flux for a single molecule containing a Cy3 dimer.....	229
<b>Figure A.17</b> Two-point TCFs for control data sets with cross-polarized polarization.....	230
<b>Figure A.18</b> Two-point TCFs for a +1 and +15 dimer with cross-polarized polarization.....	231

## LIST OF TABLES

<b>Table 1</b> Base sequences and nomenclature for the ss-dsDNA fork constructs used in DNA only replication fork studies.....	39
<b>Table 2</b> Multi-exponential fit parameters for the flux and visibility 2-point TCFs of the iCy3 monomer- and (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA constructs.....	80
<b>Table 3</b> Multi-exponential fit parameters of the visibility 2-point TCFs for the (iCy3) <sub>2</sub> dimer-labeled ss-dsDNA fork constructs.....	84
<b>Table 4</b> Number of models, per model category, for N=4,5,6 in the course of model generation. ....	107
<b>Table 5</b> Number of models, per model category, for N=4,5,6, when filters are applied during model generation.....	108
<b>Table 6</b> Optimized kinetic network model parameters for the free energy minima and activation barriers for all positions and conditions studied at replication fork junctions.....	154
<b>Table 7</b> Optimized kinetic network model parameters for the kinetic rate constants for all positions and conditions studied at replication fork junctions. ....	157
<b>Table 8</b> Base sequences of p/t-DNA constructs used in protein-DNA studies.....	165
<b>Table 9</b> Optimized kinetic network model parameters for the free energy minima and activation barriers for all protein-DNA conditions studied .....	189
<b>Table 10</b> Optimized kinetic network model parameters for the kinetic rate constants for all for all protein-DNA conditions studied. ....	192

# CHAPTER 1 : INTRODUCTION

## 1. OVERVIEW

The primary focus of this thesis is the study of thermal fluctuations in dsDNA, coined ‘DNA breathing’, both with and without the assembly of protein complexes. The first Chapter of this thesis serves as an introduction to the detailed discussion of research results described in later Chapters. The primary focus of this introductory Chapter is to bolster the readers understanding of the scientific history which led to the studies described here, the key scientific issues addressed by the development of single-molecule techniques during this thesis, and more broadly the field of macromolecular structure and dynamics, which is comprised of many biologically complementary but technologically orthogonal approaches. The contents of later Chapters are devoted to the various investigations undertaken in the labs of Dr. Andrew H. Marcus and Dr. Peter H. von Hippel. Chapter 2 has been submitted to JCPB and covers the development and implementation of polarization-sweep single-molecule fluorescence spectroscopy, done in collaboration with Claire Albrecht, Patrick Herbert, Dylan Heussman, Anabel Chang, Peter von Hippel and Andrew H. Marcus . Chapter 3 is the topic of forthcoming publication dedicated to the development of a computational framework capable of simulating kinetic network models in a massively parallel fashion. This work was done in collaboration with Claire Albrecht, Peter von Hippel and Andrew H. Marcus. Chapter 4 details the analysis of polarization-sweep single-molecule fluorescence spectroscopy experiments using the kinetic network analysis described in Chapter 3. This work was done in collaboration with Claire Albrecht, Peter von Hippel and Andrew H. Marcus. Chapter 5 covers the analysis of protein dependent single-molecule studies carried out by Patrick Herbert. This work was a collaboration between Patrick Herbert, Peter von Hippel and Andrew H. Marcus. The Appendix reports preliminary investigations into the feasibility and utility of ultrafast single-

molecule measurements. This work was done in collaboration with Amr Tamimi, Anabel Chang, Lulu Enkhbaatar, Peter von Hippel and Andrew H. Marcus.

## 2. BACKGROUND INFORMATION

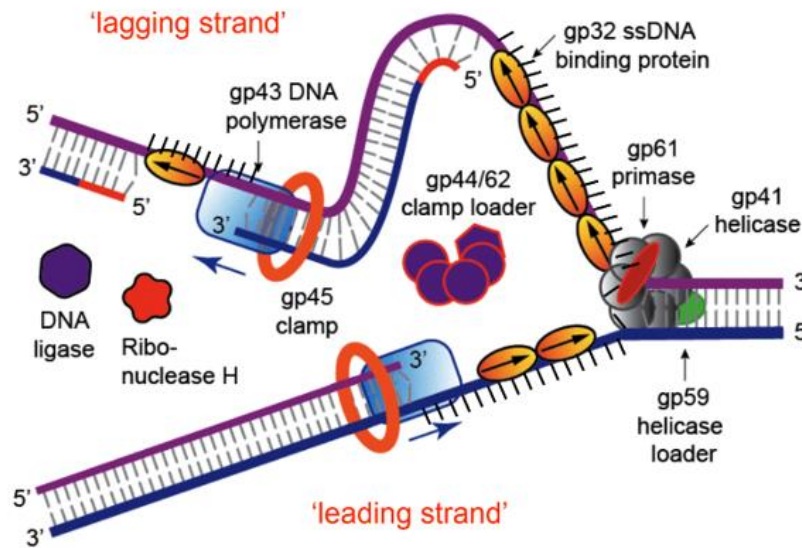
The thermodynamic stability of many biological macromolecules is held in a delicate balance between multiple conformational states under physiological conditions. This inherent lack of highly stable conformations is frequently thought of as a fundamental feature of nucleic acids, proteins, and fully assembled protein complexes, where dynamical rearrangements of their three-dimensional structure are essential for overall function. The importance of thermal fluctuations in the structure of DNA is of great importance for proteins that assemble and function onto a DNA scaffold [1]. The ability to scientifically investigate the role of DNA ‘breathing’ fluctuations began nearly 60 years ago with foundational studies of hydrogen-tritium exchange in nucleic acids obtained from calf thymus samples by von Hippel and coworkers [2], [3]. This work established an experimental basis for the hypothesized presence of local ‘breathing’ events within duplex DNA well below its typical melting temperature of  $\sim 65^{\circ}\text{C}$ . These local ‘breathing’ events were characterized over a range of pH changes and salt concentrations. The trends in pH and salt concentration demonstrated clear relationships between the rates of hydrogen-tritium exchange and the local solvent environment, establishing a basis for future investigations into the mechanisms DNA ‘breathing’. The results offered an elegant picture of the inherent propensity of dsDNA to transiently form and break hydrogen-bonds between complementary base-pairs throughout the double stranded region, which were thought to be important for providing regulatory proteins access to the interior of dsDNA. Future work by Alberts and Nossal demonstrated that despite the ability of the T4 bacteriophage single-stranded DNA binding (ssb)



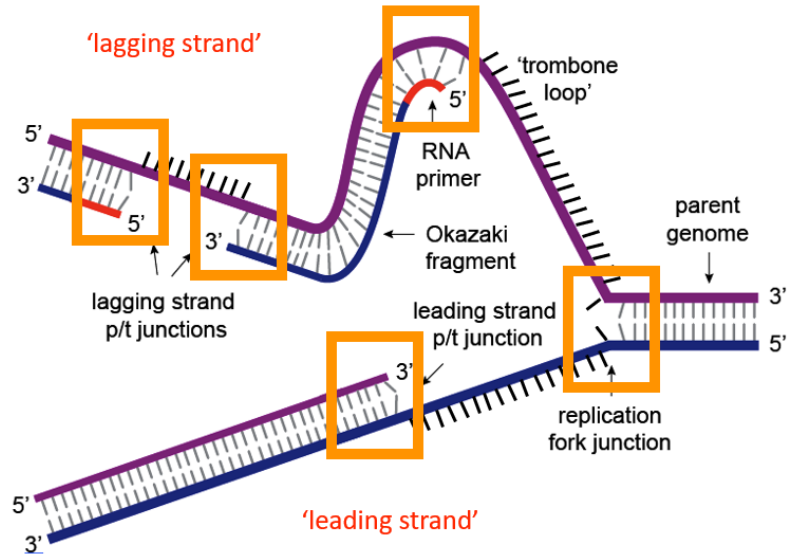
protein (called gene product-32, or gp32) to bind preferentially to ssDNA, gp32 binding to dsDNA was not occurring, suggesting a kinetic block of the fluctuations needed in dsDNA to produce the thermally activated conformational states necessary for the binding of gp32 protein [4], [5]. Followup studies, again by von Hippel and coworkers, focused on the use of formaldehyde as a chemical adduct to prove the ability of base-pairs to not only disrupt the complementary Watson-Crick (W-C) hydrogen bonds, but also to ‘flip-out’ bases into the surrounding solvent [6], [7]. These findings collectively led to many new mechanistic insights regarding the importance of conformational fluctuations in a variety of fundamental biological processes involving dsDNA [1].

Many of these studies focused on the use of protein systems purified and reconstituted from the T4 bacteriophage replication complex (or replisome), a model system for the study of DNA replication [1]. The T4 replisome is an ideal model system for studying DNA replication due to the conservation of its individual protein components in higher organisms, thereby providing a template with potential implications for human health and disease [8]. A cartoon depiction of the fully assembled T4 bacteriophage replisome is shown in Figure 1.1. A few key features can be gleaned from this depiction. First, there are many proteins involved in the form and function of the replisome. Namely, helicase gp41 that unwinds double-stranded DNA, gp59 helicase loading protein which stimulates to binding of gp41, the primase gp61 that synthesizes RNA primers on the lagging strand and the single-stranded DNA binding protein gp32, which coats ssDNA to prevent degradation. The replisome proteins include two gp43 DNA polymerases, which synthesize new DNA complimentary at the origins of the leading and lagging strands, and then gp45 sliding clamp, which contributes to polymerase processivity and fidelity, as well as the gp44/62 clamp-loader. Lagging strand synthesis is completed by RNase H, an exonuclease that removes RNA primers, and gp30 DNA ligase, which ligates the Okazaki fragments together into

a unified strand [9]. The second feature to note in this depiction is the presence of unique DNA topologies at the various junctions formed in the T4 replisome. This is highlighted by Figure 1.2. Notably, the assembly of the protein complexes involved in the T4 replisome is known to be independent of nucleic acid sequence, suggesting that local structure may be the determining factor in the cooperative assembly mechanisms at play [1].



**Figure 1.1** Depiction of the full T4 bacteriophage replisome and its various protein complexes



**Figure 1.2** Depiction of various ss-dsDNA junction topologies, highlighting that local structural differences are thought to play a role in the cooperative assembly and function of each distinct complex in the process of replication.

The presence of distinctly different DNA topologies, each with its own purpose within the replisome, and the notable absence of base sequence specificity in complex assembly begs the question: how do thermal fluctuations of local structure at ss-dsDNA junctions facilitate the binding and assembly of replication protein complexes? To address this question several biophysical considerations need to be made. The pertinent considerations at the level of protein-DNA systems are the length and time scale of relevant conformational fluctuations. These length and time scales ultimately constrain the design and implementation of an experimental protocol which can achieve sensitivity to the structural and dynamical information required. There are many approaches, spanning orders of magnitude in length and time scale, which have been successfully applied to the study of conformational changes in biological macromolecules. The next section focuses on introducing the most employed technique for obtaining primarily dynamical but also structural information from macromolecular systems: single molecule experiments.

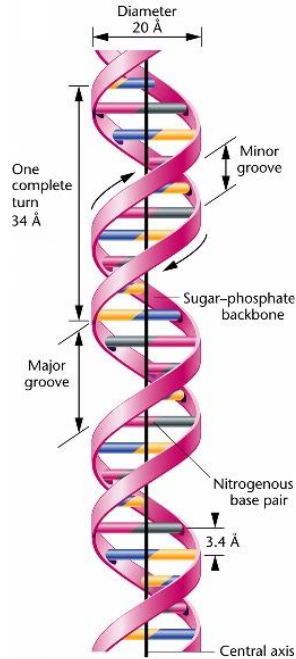
### 3. SINGLE MOLECULE APPROACHES

Measurements of single molecules are an effective route to decoding the heterogeneous nature of countless biological systems. In the absence of synchronization, nearly every dynamical system viewed from the ensemble perspective exists as a weighted average across the numerous observable states occupied by the individual subunits in time. It's recognized that many biological systems rely on time dependent changes in their structural state to either achieve a multitude of functions or more tightly regulate a particular function, in the context of the cell [10]. The most natural approach to studying the time evolution of any system, made up of many subunits, is to isolate and study one single subunit. This intuitive idea was brought into full form with the onset of single molecule imaging in the early 1990's with the first ever optical detection of single molecules [11]. Not long after the initial detection and proof of principle experiments, useful applications of such technological innovations emerged. Most ubiquitous in the field of biophysical science is single molecule Förster Resonant Energy Transfer (smFRET), which was pioneered in the mid 1990's [12]. When the right experimental conditions are met, FRET provides a useful probe of distance between two points in a protein, DNA, RNA or other biological macromolecules. The changes in the FRET efficiency as a function of time then provide a dynamical trajectory of the molecular rearrangements occurring on the length scale of typically ~1-10 nm, and a timescale of  $1\mu\text{s}$ -10 sec, depending on the specific FRET donor-acceptor chromophore pair employed [13]–[15]. Another commonly employed biophysical single molecule strategy is colocalization spectroscopy, which relies on multiplexed imaging of spectrally distinct chromophores to measure the presence and interaction between multiple subunits [16], [17]. Such multiplexed approaches provide a wealth of simultaneous information on the local constituents of hundreds to thousands of individual sub-systems, but often have a time resolution capped at 1ms

due to the use of wide-field imaging cameras. Interestingly, the use of highly multiplexed imaging of single molecules has emerged in recent years as an effective route to achieving sub-diffraction limited optical images of intricate structures within a multitude of systems at the cellular and tissue scale [18]–[20]. And lastly, the use of force-spectroscopy to measure the energetics of protein folding domains, nucleic acid base-pair energies and numerous other biophysical properties of polymers pertinent to molecular packing and overall rigidity [21]–[23]. In the case of force spectroscopy, time resolution is typically unimportant, while sensitivity to small changes in chain length as a function of applied force is often strictly desired.

Taken together, these single molecule techniques can address a wide variety of questions within biophysical science. However, certain questions remain difficult to address with the currently available repertoire of single molecule measurement approaches. Notably, the initially stated research question of this thesis, ‘how do thermal fluctuations of local structure at ss-dsDNA junctions facilitate the binding and assembly of replication protein complexes?’, remains an elusive target for such techniques. This is primarily due to the necessity for a highly local and site-specific measure of conformational changes occurring at ss-dsDNA junctions. To probe the fluctuations of ss-dsDNA junctions, both inside and outside the junction, one could consider monitoring these fluctuations from either the perspective of the base-pairs or the perspective of the sugar-phosphate backbone, as the two primary constituents of DNA structure. In the work described here, the focus is on the sugar-phosphate-backbone, as previously successful experiments using iCy3-dimer probes have demonstrated positional and temperature dependence of the iCy3-dimer structure as a reporter of the local conformations adopted by the nucleobases and sugar-phosphate backbones immediately adjacent to the probes [24]–[27]. To successfully probe these same conformational changes at the single molecule limit, what is needed is an

experiment that can measure small changes in the relative distance and orientation of the two Cy3 probes rigidly inserted into the sugar-phosphate backbones with temporal resolution that meets the timescale of relevant fluctuations. Ideally, this same experiment would delineate changes in signal between single linkages of the sugar-phosphate backbones, as a single linkage defines the transition from single-stranded to double-stranded across the ss-dsDNA junction. While issues of distance and orientation dependence at the limit of single molecules have been routinely handled by FRET in a variety of systems[14], [28]–[31], the length scales involved in the internal rearrangement of the sugar-phosphate backbones are simply too small ( $< 1$  nm) to be within the range of sensitivity of most FRET experiments ( $\sim 5$  nm). Figure 1.3 illustrates the length scales of both the internal duplex of dsDNA and the distance between nearest neighbor base-pairs, emphasizing the need for an observable which can achieve sensitivity in the range of  $\sim 1$  nm internal to the duplex. Also, it is important to consider the ability to maintain signal and sensitivity upon binding of proteins and formation of protein complexes to key sites along the sugar phosphate backbones, as the aim of many, if not all, oligonucleotide single molecule experiments is to better elucidate the biological mechanism governing protein-DNA interactions.



**Figure 1.3** Structure of dsDNA, showing the separation of adjacent bases, diameter of the double helix and definitions for the major and minor grooves of the double helix [32].

The culmination of considering these constraints has led to the development of a polarization sensitive single molecule experiment, called ‘polarization-sweep single molecule fluorescence microscopy’, which can obtain a measure of the internal conformation of iCy3-dimers rigidly inserted into the sugar-phosphate backbones of dsDNA, with demonstrated sensitivity to changes of site-specific labeling on the scale of a single linkage of sugar-phosphate backbone, sensitivity to salt-concentration in the surrounding solvent, and ultimately sensitivity to the binding and functional activity of proteins that form protein-DNA complexes.

## 4. MACROMOLECULAR STRUCTURE AND DYNAMICS

This section is dedicated to a brief overview of the various biophysical methods available to the field of molecular and structural biology which have played and continue to play an important role in the determination of the structure and dynamics of macromolecules. The interested reader is encouraged to further investigate the cited references of this section, as the field of biophysics has grown rapidly over the past few decades with numerous complementary but technological orthogonal techniques coming into existence that can offer a wealth of information about the systems under study. In general, there are a handful of techniques that are well suited for structure determination and a handful of techniques that are well suited for capturing dynamics. It is uncommon for a single approach to offer detailed structural information and fast enough time resolution to be considered a measurement of dynamics. Traditionally, questions of structure have been handled independently of investigations into dynamics. However, the next section briefly highlights progress made into the combination of these two critical components in the study of biological macromolecules.

The field of structural biology is longstanding and thoroughly developed in the fields of X-ray crystallography[33]–[35], NMR[36]–[38], small angle X-ray scattering[39], cryo-electron microscopy [15], [40], [41] and more recently computer simulations [42]. This suite of techniques offers tradeoffs between spatial resolution, addressable system scale and occasionally time resolution, with very few achieving all three to a high degree. Recent advances in cryo-EM have opened the door to stopped-flow experiments of large biomolecules, with time resolution down to a limit of 5ms [43]. Also notable is the use of single-particle picking techniques in cryo-EM [15], [44] as well as single-particle X-ray diffraction methods [35], both aimed at unpacking the inherent heterogeneity of biological macromolecules, which maintaining a high degree of structural



information. Recent developments in structural NMR have allowed the identification and structural assignment of minority protein conformations in solution which play important roles in regulating their function and mediating interactions with therapeutics [36], [45]. Another notable achievement is the use of all optical approaches, such as two-dimensional fluorescence spectroscopy (2DFS), for the determination of local structure with a high degree of specificity, in solution, that also affords quantification of the inherent heterogeneity in the systems under study [24], [25], [46]–[48]. The list of techniques outlined here is by no means exhaustive, mass-spectrometry, EPR and DEER spectroscopy have all been utilized for protein structure determination as well. Ultimately, all these approaches offer a glimpse into the structures which macromolecules freely adopt under physiological conditions, with some approaches coming much closer to reporting on physiological conditions than others.

In the opposing, but intimately related, realm of measuring macromolecular dynamics, the major techniques are time-resolved optical experiments[49], [50], single molecule techniques [8], [13], [14], [51]–[56], computer simulations [57], [58] and also to some extent the previously mentioned developments in NMR, cryo-EM and mass-spectrometry. In the context of this dissertation introduction, the consideration of ‘dynamics’ has been limited to real-time observation and tracking of conformational changes occurring in single molecules and identification of non-majority conformers in ensemble experiments. Importantly, many ultrafast optical techniques exist for probing the fast changes in either the structure or chemical environment of small molecules but are often held to an upper limit of timescales on the order of nanoseconds.

Taken together, these techniques offer a broad range of time resolution depending on the question of interest. At the time of this discussion, single molecule dynamics on the order of milliseconds are routinely accessed, while dynamics on the order of microseconds and even

nanoseconds have been accessed in some cases [14], [59]. Computer simulations have made remarkable progress in producing time resolved conformational changes from both all-atom and coarse-grained approaches, which have greatly aided the pursuits of experimentalists by offering insights beyond the current capabilities of experiment [57], [60]–[62]. However, in nearly all cases the longest trajectories available to such simulated approaches are on the order of microseconds, making direct comparison with slow dynamics observed in experiment difficult. For the purely experimental approaches taken in the field of single molecule spectroscopy, currently only smFRET offers a direct measure of ‘structure’ via the distance dependence of its transfer efficiency. All other structural information is often inferred from the dynamical information measured, as is the case with kinetic binding assays and colocalization studies. In rare instances, the structural resolving power of cryo-EM has been combined with single molecule studies to successfully bridge the gap between the intrinsic dynamics and structural diversity of complex biophysical systems [15], [17], [63]–[65]. This lack of structural sensitivity in predominantly dynamical experiments highlights the need for better real-time estimates of conformation in the field of biophysical science. Given the enormous advantages of studying conformational changes at the limit of single molecules, particular effort should be paid to innovative approaches which either go beyond or compliment the capabilities of FRET. The remaining Chapters of this thesis are dedicated to a detailed discussion of the design, development, optimization, and implemented use cases of the aforementioned ‘single-molecule- polarization sweep microscopy’ method, which achieves microsecond resolved estimates of conformation in the sugar-phosphate backbone of ssDNA molecules, both alone and in complexation with proteins. The concluding Chapter of this thesis details investigations into broadband single molecule experiments, which focuses on technical feasibility and inherent advantages versus disadvantages of such experiments.

## CHAPTER 2 : POLARIZATION SWEEP SINGLE-MOLECULE FLUORESCENCE MICROSCOPY

### 1. OVERVIEW

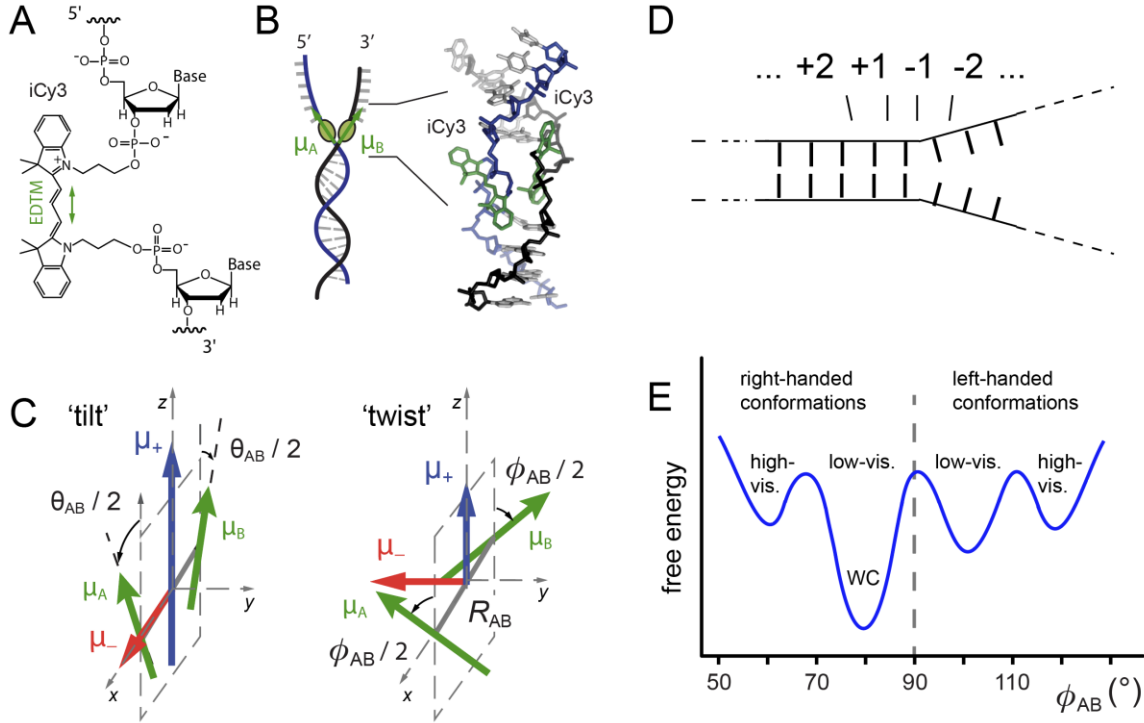
This Chapter contains worked from the article “*Studies of DNA ‘breathing’ by polarization-sweep single-molecule fluorescence microscopy of exciton-coupled (iCy3)<sub>2</sub> dimer-labeled DNA fork constructs*” authored by Jack Maurer, Claire Albrecht, Patrick Herbert, Dylan Heussman, Andrew H. Marcus and Peter von Hippel. This work was funded by the National Institutes of Health (NIGMS Grant GM-215981 to P.H.v.H. and A.H.M.). Andrew H. Marcus was the principal investigator for this work. The contents of the article have been expanded upon here to serve as a comprehensive overview and introduction to the methods which will be used throughout the remaining Chapters. This Chapter will more formally introduce the motivation for the ‘polarization-sweep’ experiment, the experimental setup itself, experimental signal derivation and the statistical analysis approaches employed for all single-molecule data discussed throughout later Chapters.

### 2. INTRODUCTION

The assembly of the protein-DNA complexes that drive the various processes of genome expression involves coordinated interactions between multiple macromolecular species. For example, during the initial stages of DNA replication, proteins recognize and bind to base-sequence-specific sites at the replicative origin, forming a multi-subunit protein-DNA complex that, in turn, recruits additional protein factors to establish the ‘trombone’ shaped framework of

the functional DNA replication-elongation complex. This DNA framework is subject to thermally induced fluctuations that permit local regions of the initially double-stranded (ds) DNA to transiently expose single-stranded (ss) DNA templates to the aqueous surroundings, and thus provide access to the protein components of the DNA replication complex. Although the notion of ‘DNA breathing’ is consistent with free energy landscape models of protein-DNA interactions[66], little is known about the distributions of the Boltzmann-weighted conformational macrostates at and near replication fork junctions, or the associated activation barriers that control the assembly and operation of replisomal protein-DNA complexes.

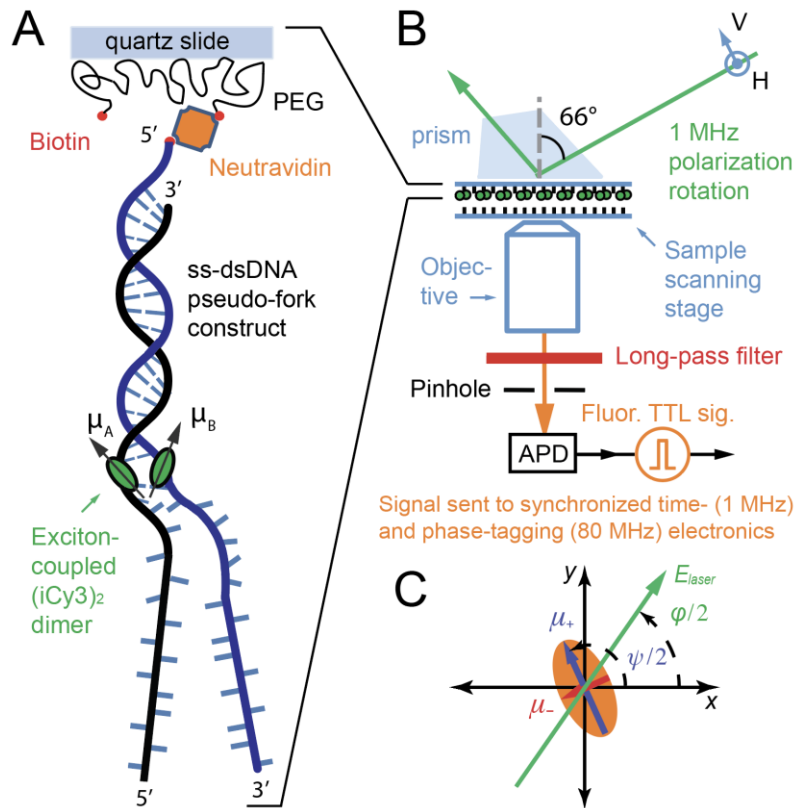
In this work, we present a novel, polarization-selective single-molecule fluorescence method to study the local conformations and conformational dynamics of model DNA fork constructs, which are labeled with pairs of cyanine dyes [(iCy3)<sub>2</sub>] that are rigidly inserted into the sugar-phosphate backbones at various positions relative to the ss-dsDNA fork junction (see Fig. 2.1A and 2.1B)[67]. The closely spaced iCy3 monomers (labeled *A* and *B*) are electrostatically coupled, so that the (iCy3)<sub>2</sub> dimer supports delocalized excitons [symmetric (+) and anti-symmetric (–)] with orthogonally polarized electric dipole transition moments (EDTMs),  $\boldsymbol{\mu}_{\pm} = \frac{1}{\sqrt{2}}(\boldsymbol{\mu}_A \pm \boldsymbol{\mu}_B)$ . The magnitudes of the EDTMs depend sensitively on the local conformation of the (iCy3)<sub>2</sub> dimer probe, which is characterized by the ‘tilt’ angle,  $\theta_{AB}$ , and the ‘twist’ angle,  $\phi_{AB}$  (see Fig. 2.1C) [24], [25], [68].



**Figure 2.1** Labeling chemistry and nomenclature of the internal (iCy3)<sub>2</sub> dimer probes positioned within the sugar-phosphate backbones of model ss-dsDNA constructs. (A) The Lewis structure of the iCy3 chromophore is shown with its 3'- and 5'-linkages to the sugar-phosphate backbone of a local segment of ssDNA. The double-headed green arrow indicates the orientation of the electric dipole transition moment (EDTM). (B) An (iCy3)<sub>2</sub> dimer-labeled DNA construct contains the dimer probe near the single-stranded (ss) – double-stranded (ds) DNA junction. The conformation of the (iCy3)<sub>2</sub> dimer probe reflects the local secondary structure of the sugar-phosphate backbones at the probe insertion site position. The sugar-phosphate backbones of the conjugate DNA strands are shown in black and blue, the bases in gray, and the iCy3 chromophores in green. (C) The structural parameters that define the local conformation of the (iCy3)<sub>2</sub> dimer probe are the inter-chromophore separation vector  $R_{AB}$ , the tilt angle  $\theta_{AB}$ , and the twist angle  $\phi_{AB}$ . The electrostatic coupling between the iCy3 chromophores gives rise to the anti-symmetric (–) and symmetric (+) excitons, which are indicated by the red and blue arrows, respectively, and whose magnitudes and transition energies depend on the structural parameters. (D) The insertion site position of the (iCy3)<sub>2</sub> dimer probe is indicated relative to the pseudo-fork junction using positive integers in the direction towards the double-stranded region, and negative integers in the direction towards the single-stranded region. (E) A hypothetical free energy surface (FES) describing local fluctuations of the (iCy3)<sub>2</sub> dimer-labeled sugar-phosphate backbones near the ss-dsDNA fork junction as a function of the twist angle parameter,  $\phi_{AB}$  [24].

In Fig. 2.2, we show a schematic layout of our experimental approach, which we call polarization-sweep single-molecule fluorescence (PS-SMF) microscopy. A single (iCy3)<sub>2</sub> dimer-labeled DNA construct is resonantly excited using a continuous-wave (cw) laser whose plane polarization is rotated at the frequency  $\Omega/2\pi = 1$  MHz. The symmetric (+) and anti-symmetric (–) excitons of the (iCy3)<sub>2</sub> dimer probe, which are oriented perpendicular to one another, are

alternately excited as the laser polarization direction is rotated across the corresponding EDTMs, and the ensuing modulated fluorescence is detected using the phase-tagged photon-counting (PTPC) method [69]. The resulting PS-SMF signal is directly related to the local conformation of the  $(i\text{Cy}3)_2$  dimer-labeled ss-dsDNA construct, whose fluctuations can be monitored on microseconds and longer time scales.



**Figure 2.2** The PS-SMF microscopy experimental layout. (A) An  $(i\text{Cy}3)_2$  dimer-labeled ss-dsDNA fork construct, which can form a protein-DNA complex, is attached to the microscope slide surface using biotin / neutravidin linkages. (B) A total internal reflection fluorescence (TIRF) microscope is used to illuminate the sample with a continuous-wave (cw) 532 nm laser. The plane-polarized electric field vector of the laser is continuously rotated at the frequency  $\Omega/2\pi = 1$  MHz by sweeping the phase of an interferometer (Fig. 2.3). Individual signal photons (fluorescence) from a single molecule are detected using an avalanche photodiode (APD). Each detection event is registered within a time window  $\Delta t = 1 \mu\text{s}$ , and its phase is assigned to one of 80 incremented values (with bin width  $\Delta\phi = 360^\circ/80 = 4.5^\circ$ ) using the method of phase-tagged photon counting (PTPC) [69]. (C) A single  $(i\text{Cy}3)_2$  dimer-labeled ss-dsDNA construct has symmetric and anti-symmetric EDTMs (labeled  $\mu_\pm$ , shown as blue and red vectors, respectively), which define the major and minor axes of a ‘polarization ellipse’ in the transverse cross-sectional area of the incident laser beam. The orthogonally polarized symmetric and anti-symmetric

excitons are alternately excited as the laser electric field vector (shown in green) rotates in the clockwise direction (with phase  $\varphi/2 = \Omega t/2$ ).

PS-SMF experiments provide both structural and kinetic information about the microscopic configurations of the sugar-phosphate backbones and bases immediately surrounding the (iCy3)<sub>2</sub> dimer probe. PS-SMF trajectories contain information about the equilibrium distribution of conformational macrostates and their associated rates of interconversion, which are directly related to parameters that characterize the site-specific local free energy landscape (FEL) [26]. From the photon count data stream, we determine the following three ensemble-average functions of the PS-SMF signal: (i) the two-point time-correlation function (TCF),  $\bar{C}^{(2)}(\tau)$  that describes the average correlation between two successive measurements separated by the interval  $\tau$ , which is sampled on tens-of-microsecond time scales; (ii) the three-point TCF,  $\bar{C}^{(3)}(\tau_1, \tau_2)$  that describes the average correlation between three successive measurements separated by the intervals  $\tau_1$  and  $\tau_2$  and is sampled on millisecond time scales; and (iii) the probability distribution function (PDF) of the PS-SMF signal visibility  $P(v)$ , which is averaged over the 10 ms time scale required to achieve a signal-to-noise (S/N) ratio of  $\sim 10$ .

Although beyond the scope of this Chapter, the above functions can be analyzed using a kinetic network model to characterize the equilibrium and kinetic properties of multistep conformational transition pathways at equilibrium. While the two-point TCF,  $\bar{C}^{(2)}$ , contains information about the characteristic relaxation times of the system, the three-point TCF,  $\bar{C}^{(3)}$ , contains additional information about the exchange times between pathway intermediates [14], [56], [70], [71]. The use of a kinetic network model will be discussed in Chapters 3 and 4, while the role of time correlation functions and probability distribution functions will be covered in section 3.6 and 3.7 of this Chapter.

Recently, Heussman *et al.* used ensemble absorbance, circular dichroism (CD) and two-dimensional fluorescence spectroscopy (2DFS) to study the *average* local conformations and conformational disorder of iCy3 monomer- and (iCy3)<sub>2</sub> dimer-probes at various positions within ss-dsDNA fork constructs [25]. From these studies it was concluded that the average conformation of the (iCy3)<sub>2</sub> dimer-probe changes systematically as the labelling site is varied across the ss-dsDNA junction from the +2 to the -2 positions (see Fig. 2.1D for site-labelling nomenclature). For positive integer positions the mean local conformation was shown to be right-handed (with mean twist angle,  $\bar{\phi}_{AB} < 90^\circ$ ), and for negative integer positions the local conformation was left-handed ( $\bar{\phi}_{AB} > 90^\circ$ ). Moreover, these experiments indicated that the distributions of conformational states at these key positions are relatively narrow, suggesting that regions of the junction extending towards and into the single-stranded segments exhibit moderate levels of structural order.

A possible explanation for the results of Heussman *et al.* [25] is that the (iCy3)<sub>2</sub> dimer probe, which is covalently linked to the sugar-phosphate backbones and bases immediately surrounding the probe, can adopt only a small number of local conformations whose relative stabilities depend on position relative to the ss-dsDNA fork junction. In Fig. 2.1E, this concept is illustrated using a hypothetical free energy landscape, which depicts four possible local conformations of the (iCy3)<sub>2</sub> dimer probe at a positive integer position, with the right-handed Watson-Crick (WC) B-form conformation being most favored. In this picture the free energy differences between the B-form ground state and the other (non-canonical) local conformations are sufficiently high to ensure that the Boltzmann-weighted distribution of available macrostates is dominated by the B-form structure. At the same time, the moderate free energies of activation allow for infrequent transitions between non-canonical local structures, which are present in trace



amounts. As the probe-labelling position is varied across the ss-dsDNA junction towards the single-stranded region, one might expect the free energy surface to exhibit local minima with approximately the same coordinate values, but with relative stabilities and barrier heights shifted to reflect the presence of non-canonical structures observed in ensemble measurements [24]–[26].

The PS-SMF method presented here is an extension of one previously developed by Phelps *et al.* [13], which simultaneously monitors the linear dichroism (LD) and Förster resonance energy transfer (FRET) signals from an internally labeled iCy3 / iCy5 donor-acceptor pair. The primary advantage of the PS-SMF method is that it can sensitively isolate signals due to the relative internal motions of an exciton-coupled (iCy3)<sub>2</sub> dimer probe from those arising from collective rotation and/or excited state population fluctuations (i.e., ‘blinking’) [13]. As such, PS-SMF experiments provide a direct means to observe the dynamic interconversions between local conformational macrostates of (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs. The PS-SMF method is thus well suited for studies of the kinetic pathways and equilibrium distributions of the conformational fluctuations of the (iCy3)<sub>2</sub> dimer-labeled sugar-phosphate backbones and bases, which are central to understanding — at the molecular level — the different forms of local DNA ‘breathing’ at specified positions relative to ss-dsDNA junctions.

In the following Chapters, we show that PS-SMF experiments can be used to study position- and salt-dependent ‘breathing’ fluctuations of (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs. Among the significant findings of this work is that the local conformations of the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs undergo transient fluctuations between four conformational macrostates, as suggested by the free energy landscape depicted schematically in Fig. 2.1E. Furthermore, at probe labeling positions at and near the ss-dsDNA junction, extending towards the single-stranded region, the local B-form conformation becomes unstable relative to conformations

that exist in the duplex at only trace occupancy levels. Furthermore, altering the salt concentration from physiological (6 mM MgCl<sub>2</sub>, 100 mM NaCl), either to increasing or decreasing levels, also destabilizes the B-form conformation, reminiscent of the concept of cold denaturation [72]. While certain aspects of salt concentration-dependent DNA stability and protein-DNA interactions are well understood [73]–[75], little information is currently available about the microscopic details of how salt concentration affects conformational fluctuations near ss-dsDNA fork junctions. The results of our experiments suggest a possible structural framework for understanding the roles of transient DNA conformational fluctuations at and near ss-dsDNA fork junctions in driving the processes of protein-DNA complex assembly and function.

### 3. MATERIALS AND METHODS

#### **Section 3.1. (iCy3)<sub>2</sub> dimer-labeled single-stranded (ss) – double-stranded (ds) DNA fork constructs.**

The model ss-dsDNA fork constructs that we used in this work have either an iCy3 monomer or an (iCy3)<sub>2</sub> dimer incorporated into the DNA framework, as shown in Figs. 2.1A and 2.1B. The specific nucleic acid base sequences are listed in Table 1. We purchased nanomolar quantities of single-stranded oligonucleotides from Integrated DNA Technologies (IDT, Coralville, IA). The initially dehydrated samples were rehydrated at 100 nM concentrations in aqueous buffer (100 mM NaCl, 6 mM MgCl<sub>2</sub> and 10 mM Tris, pH 8.0). Complementary oligonucleotides were combined in a 1:1.5 molar concentration ratio between the biotinylated and non-biotinylated strands, respectively, before they were heated to 94°C for 4 min and allowed to cool slowly to room temperature (~23°C). The annealed iCy3 monomer and (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA

constructs contained both single-stranded and double-stranded regions with the probe labeling positions indicated by the nomenclature described in Fig. 2.1. The iCy3 monomer-labeled ss-dsDNA constructs contained a thymine base (T) in the complementary strand at the position directly opposite to the probe chromophore. Solution samples were stored at 4°C between experiments conducted over consecutive days, or frozen at -4°C between experiments conducted over more extended periods.

**Table 1** Base sequences and nomenclature for the iCy3 monomer and (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs used in these studies. The horizontal lines indicate regions of complementary base pairing.

DNA construct	Nucleotide base sequence
+1 iCy3 monomer	3'- <u>CAG CGG TCG GAG CGT CG(iCy3)</u> GTT TTT TTT TTT TTT-5' 5'-biotin/GTC GCC AGC CTC GCA GC T CTT-3'
+1 (iCy3) <sub>2</sub> dimer	3'- <u>CAG CGG TCG GAG CGT CG(iCy3)</u> GTT TTT TTT TTT TTT-5' 5'-biotin/GTC GCC AGC CTC GCA GC(iCy3) CTT-3'
-1 (iCy3) <sub>2</sub> dimer	3'- <u>CAG CGG TCG GAG CGT CGG (iCy3)</u> TT TTT TTT TTT TTT-5' 5'-biotin/GTC GCC AGC CTC GCA GCC (iCy3)TT-3'
-2 (iCy3) <sub>2</sub> dimer	3'- <u>CAG CGG TCG GAG CGT CGG T(iCy3)</u> T TTT TTT TTT TTT-5' 5'-biotin/GTC GCC AGC CTC GCA GCC T(iCy3)T-3'

### **Section 3.2. Sample preparation.**

Solutions of annealed ss-dsDNA samples were diluted 1000-fold (~100 pM concentration) before they were introduced into a custom-built microscope sample chamber, as described in prior work [69]. Sample chambers were constructed from microscope slides, which were chemically modified using the procedure of Chandradoss *et al.*[76]. The inner surfaces of the sample chambers were coated with a layer of poly(ethylene glycol) (PEG), which was sparsely labeled with biotin. Neutravidin was used as a linker to bind the biotin-labeled PEG to the biotin molecules covalently attached to the 5' ends of the dsDNA regions of the ss-dsDNA constructs, as shown in Table 1. To reduce the rate of photobleaching in our measurements, an oxygen scavenging solution consisting of glucose oxidase, catalase and glucose was introduced into the sample cell before the PS-SMF data were recorded. To suppress the 'photo-blinking' effects of excited-state intersystem crossing, the triplet state quencher Trolox was used in the solution buffer[77].

### **Section 3.3. Polarization-sweep single-molecule fluorescence (PS-SMF) microscopy experimental setup.**

In Fig. 2.3, we show a schematic of the instrumental setup for PS-SMF experiments on model (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs. The setup is an extension of an approach previously developed to simultaneously monitor single-molecule Förster resonance energy transfer (FRET) and fluorescence-detected linear dichroism (FLD) signals from iCy3/iCy5 (hetero-) dimer-labeled DNA constructs [13]. As we discuss in further detail below, the signals from PS-SMF measurements provide direct information about the distributions of conformational macrostates and conformational dynamics of the (iCy3)<sub>2</sub> dimer probe, which is site-specifically positioned within the ss-dsDNA fork junction.

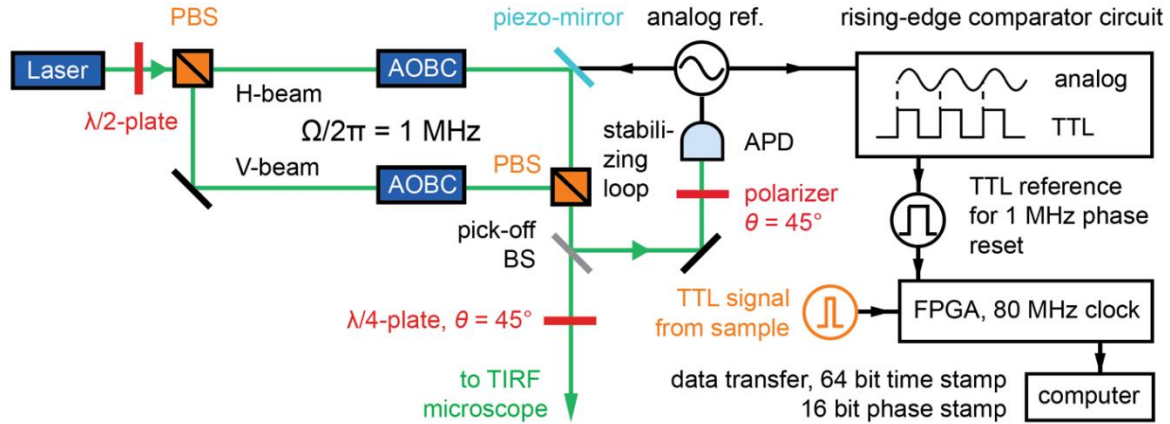
The (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs were chemically attached to the surface of a glass microscope slide using biotin/neutralavidin linkages, as described in section 3.2 (see Fig. 2.2A). The sample chamber was placed on a computer-controlled translation stage with the probe-labeled DNA constructs illuminated in a total internal reflection fluorescence (TIRF) geometry by a plane-polarized continuous-wave laser (with center wavelength  $\lambda_L = 532$  nm, see Fig. 2.2B). The laser beam was focused to a 50  $\mu\text{m}$  diameter spot at the sample, and the incident power was adjusted to  $\sim 15$  mW. Fluorescence from the sample was collected using a high numerical aperture objective lens (N.A. = 1.4) and the image from a single molecule was isolated using a pinhole, a spectral filter (532 nm long-pass) and detected using an avalanche photodiode (APD). The mean total fluorescence photon detection rate (or mean signal flux,  $\bar{f}$ ) was typically  $\sim 8,000 - 10,000$  counts per second (cps). The signal ‘background rate’ was determined to be  $\sim 500$  cps by scanning the stage position away from the center of a molecule. Thus, the signal-to-background ratio of our experiments was  $\sim 20$ . The polarization-dependent signal from each single molecule sample was recorded for a total duration of 30 – 60 s, as described further below.

A Mach-Zehnder interferometer (MZI), which was equipped with acousto-optic phase modulators in each arm, was used to prepare the plane-polarization state of the laser. In Fig. 2.3, we show a schematic of the laser interferometer and the integrated detection electronics used for our PS-SMF experiments. A half-wave plate is used to rotate the plane polarization vector of a continuous wave (cw) 532 nm laser to  $45^\circ$  from vertical. The laser is directed through a polarizing beam-splitter (PBS) to produce balanced vertical and horizontal plane-polarized beams, which traverse separately the two paths of a Mach-Zehnder Interferometer (MZI) before they are recombined at the PBS exit port. Two acousto-optic Bragg cells (AOBCs) are each placed within the paths of the MZI. Each AOBC is driven continuously at a fixed radio frequency so that a

relative phase ‘sweep’ is imparted to the MZI output beam:  $\varphi = \Omega t$  so that the plane polarization (Jones) vector of the electric field was rotated at the constant frequency  $\Omega/2\pi = 1$  MHz. The output beam passes through a quarter-wave plate, which is rotated to  $45^\circ$  from vertical, before it is directed to the TIRF microscope to excite the single-molecule sample, as shown in Fig 2.2. A pick-off window is used to direct a minor component of the beam through a polarizer and detected using an avalanche photodiode (APD) to create a 1 MHz analog reference signal. The reference is utilized in a negative feedback loop to actively minimize the path difference of the MZI using a piezo-controlled mirror, as implemented in a previous experiment[13].

The digital electronics illustrated in Fig. 2.3 are used to implement the phase-tagged photon counting (PTPC) method described by [69]. In this approach, we use a field-programmable gate array (FPGA), which is a manually reconfigurable integrated circuit that contains hardware-enabled signal-processing algorithms. We used the FPGA to discretize the phase of a given modulation cycle into a set of  $m$  ‘phase bins,’ which are numbered and incrementally advanced using an 80 MHz digital counter. The 1 MHz analog reference waveform is first converted into a ‘logical square wave’ (i.e., a periodic TTL waveform) using a custom-built rising-edge comparator circuit. The logical square wave is then used to trigger an 80 MHz phase counter (16-bit width), and to reset the counter at the 1 MHz phase-sweep frequency. The counter reset automatically synchronizes the phase-sweep cycle in the presence of external room vibrations that introduce noise to the MZI reference phase. The average number of phase bins during a modulation cycle is  $m = 80 \text{ MHz} / 1 \text{ MHz} = 80$  bins, and the phase bin interval is  $\Delta\varphi = 360^\circ / 80 = 4.5^\circ \text{ bin}^{-1}$ . Thus, during each phase-sweep cycle the counter advances the phase bin value  $\varphi_j = j\Delta\varphi$  (with  $j = 0, 1, \dots, 79$ ) at the 1 MHz clock speed. We also used a second 1 MHz counter (with 64bit width) to assign a time bin value

$t_k = k\Delta t$  (with  $\Delta t = 1 \mu\text{s}$  and  $k = 0, 1, \dots$ ) to each detection event. Thus, each individual detection event is assigned a 16-bit phase bin value and a 64-bit time bin value, and these data are streamed to computer memory in real time.



**Figure 2.3** Schematic of the interferometer and integrated phase-tagged photon counting (PTPC) electronics [69] used to perform PS-SMF experiments. PBS: polarizing beam-splitter; AOBC: acousto-optic Bragg cell; APD: avalanche photodiode; FPGA: field-programmable gate array.

### Section 3.4 Theoretical description of Jones vector electric field components

In this section, we derive Eq. (7) which appears at the end of this section. We decompose the polarization state of the electric field in the Jones vector basis of horizontal ( $H$ ) and vertical ( $V$ ) plane polarization components. Before entering the MZI (see Fig. 2.3), the plane polarization of the laser is rotated to  $45^\circ$  from vertical using a half-wave plate so that the initial electric field may be written in the  $\begin{bmatrix} H \\ V \end{bmatrix}$  basis:

$$\mathbf{E}_{initial}(\omega) = A(\omega)e^{i\omega L t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (1)$$

In Eq. (1),  $A(\omega)$  is a narrow spectral envelope and  $\omega_L$  is the laser center frequency ( $= 2\pi c/\lambda_L$ , where  $\lambda_L = 532$  nm). At the entrance port of the MZI, the PBS separates the  $H$  and  $V$  polarization components into separate paths. Within each MZI path, an AOBC imparts a time-varying phase shift to its respective beam according to  $\varphi_H = \Omega_H t$  and  $\varphi_V = \Omega_V t$ . Thus, the electric field components within the  $H$  and  $V$  paths can be written:

$$\mathbf{E}_H(\omega) = A(\omega)e^{i\omega_L t} \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{i\varphi_H} \quad (2a)$$

$$\mathbf{E}_V(\omega) = A(\omega)e^{i\omega_L t} \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{i\varphi_V} \quad (2b)$$

At the exit port of the MZI, the  $H$  and  $V$  beams are recombined at a second PBS to produce the total field:

$$\mathbf{E}_{total}(\omega) = \mathbf{E}_H(\omega) + \mathbf{E}_V(\omega) = \frac{A(\omega)}{\sqrt{2}} e^{i\omega_L t} \begin{bmatrix} e^{i\varphi_H} \\ e^{i\varphi_V} \end{bmatrix} \quad (3)$$

Equation (3) can be simplified by defining the relative phase shift  $\varphi = \varphi_V - \varphi_H = \Omega_{VH} t$  between the  $H$  and  $V$  paths (with  $\Omega_{VH} = \Omega = 1$  MHz), and by multiplying by the overall phase shift  $e^{-i(\varphi_V + \varphi_H)/2}$ :

$$\mathbf{E}_{total}(\omega) = \frac{A(\omega)}{\sqrt{2}} e^{i\omega_L t} \begin{bmatrix} e^{-i(\varphi/2)} \\ e^{+i(\varphi/2)} \end{bmatrix} \quad (4)$$



The beam is next directed through a quarter-wave plate that is rotated 45° from vertical. The ‘Jones matrix’  $\mathbf{M}$  for the quarter-wave plate is given by:

$$\mathbf{M} = \frac{e^{i(\pi/4)}}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \quad (5)$$

Thus, after passing through the quarter-wave plate and applying the linear transformation of  $\mathbf{M}$  to the recombined phase modulated field components, the final electric field can be written compactly as:

$$\begin{aligned} \mathbf{E}(\omega, \varphi) &= \mathbf{E}_{total} \mathbf{M} = \frac{A(\omega)}{2} e^{i\omega L t} e^{i(\pi/4)} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \begin{bmatrix} e^{-i(\varphi/2)} \\ e^{+i(\varphi/2)} \end{bmatrix} \\ &= A(\omega) e^{i\omega L t} \begin{bmatrix} +\cos(\varphi/2) \\ -\sin(\varphi/2) \end{bmatrix} \end{aligned} \quad (6)$$

The second equality of Eq. (6) is the electric field incident on the sample, which is equivalent to Eq. (7) below:

$$\mathbf{E}(\omega, \varphi) = A(\omega) e^{i\omega L t} [\cos(\varphi/2)\hat{i} - \sin(\varphi/2)\hat{j}] \quad (7)$$

### Section 3.5 Derivation of the polarized signal intensity for PS-SMF experiments

In Eq. (7),  $A(\omega)$  is the (narrow) spectral envelope,  $\omega_L (=2\pi c/\lambda_L)$  is the laser center frequency and  $\hat{\mathbf{i}}$  and  $\hat{\mathbf{j}}$  are unit vectors that point along the  $x$  and  $y$  axes of the molecular frame, respectively, as shown in Fig. 2.1C. The total electric dipole transition moment (EDTM) of the coupled  $AB$  dimer is  $\boldsymbol{\mu}_{tot} = \boldsymbol{\mu}_+ + \boldsymbol{\mu}_-$ , where  $\boldsymbol{\mu}_\pm = \frac{1}{\sqrt{2}}(\boldsymbol{\mu}_A \pm \boldsymbol{\mu}_B)$  are the EDTMs of the symmetric (+) and anti-symmetric (−) excitons. We note that the magnitudes of the  $\mu_\pm$  EDTMs depend sensitively on the local conformation of the  $AB$  dimer (see Fig. 2.1C and discussions in [24], [25], [27]). The electric field vector,  $\mathbf{E}(\omega, \varphi)$ , is given by Eq. (7) and the symmetric (+) and anti-symmetric (−) electric dipole transition moments (EDTMs) of the exciton coupled dimer,  $\boldsymbol{\mu}_\pm(\omega)$ , are given by

$$\boldsymbol{\mu}_+(\omega) = \sum_{\alpha} \mu_+^{(\alpha)}(\omega) \begin{bmatrix} +\cos(\psi/2) \\ +\sin(\psi/2) \end{bmatrix} \quad (8a)$$

$$\boldsymbol{\mu}_-(\omega) = \sum_{\alpha'} \mu_-^{(\alpha')}(\omega) \begin{bmatrix} -\sin(\psi/2) \\ +\cos(\psi/2) \end{bmatrix} \quad (8b)$$

In Eqs. (8a) and (8b), the angle  $\psi/2$  specifies the orientation of the dimer in the  $x$ - $y$  plane, which we assume to be constant or slowly varying on the time scales of the dimer conformational fluctuations of interest. The terms  $\mu_\pm^{(\alpha)}(\omega)$  are the spectrally-dependent magnitudes of the EDTMs, which depend on the dimer conformation, as discusses previously, and the index  $\alpha$  enumerates the singly-excited dimer states in order of increasing energy [24], [25], [27]. Furthermore, the

directions of  $\boldsymbol{\mu}_{\pm}$  have orthogonal relative orientation, and thus define a ‘polarization ellipse’ in the cross-sectional area in which the laser beam projects onto the molecular frame:

$$\boldsymbol{\mu}_{+}(\omega) = \sum_{\alpha} \boldsymbol{\mu}_{+}^{(\alpha)}(\omega) [\cos(\psi/2)\hat{\mathbf{i}} + \sin(\psi/2)\hat{\mathbf{j}}], \quad (9a)$$

$$\boldsymbol{\mu}_{-}(\omega) = \sum_{\alpha} \boldsymbol{\mu}_{-}^{(\alpha)}(\omega) [-\sin(\psi/2)\hat{\mathbf{i}} + \cos(\psi/2)\hat{\mathbf{j}}], \quad (9b)$$

The  $\boldsymbol{\mu}_{\pm}$  symmetric and anti-symmetric EDTMs ( $\boldsymbol{\mu}_{+}$  and  $\boldsymbol{\mu}_{-}$ , respectively) are depicted graphically in Fig. 2.1C. At any instant, the ‘polarized’ single-molecule fluorescence intensity is given by the spectral overlap and square modulus of the vector projections between the laser electric field and the total EDTM.

$$I_P(\varphi) = \sum_{\alpha} \int d\omega \left| \mathbf{E}(\omega, \varphi) \cdot \boldsymbol{\mu}_{tot}^{(\alpha)}(\omega) \right|^2 \quad (10)$$

Equation (10) can be written as the sum of symmetric (+) and anti-symmetric (–) exciton contributions by substituting  $\boldsymbol{\mu}_{tot} = \boldsymbol{\mu}_{+} + \boldsymbol{\mu}_{-}$ :

$$I_P(\varphi) = \sum_{\alpha} \int d\omega \left| \mathbf{E}(\omega, \varphi) \cdot \boldsymbol{\mu}_{+}^{(\alpha)}(\omega) \right|^2 + \sum_{\alpha'} \int d\omega \left| \mathbf{E}(\omega, \varphi) \cdot \boldsymbol{\mu}_{-}^{(\alpha')}(\omega) \right|^2 \quad (11)$$

We identify the first term on the right-hand side of Eq. (11) as the symmetric exciton contribution,  $I_+(\varphi)$ , and the second term as the anti-symmetric exciton contribution,  $I_-(\varphi)$ . The separation given by Eq. (11) is a consequence of  $\boldsymbol{\mu}_+$  and  $\boldsymbol{\mu}_-$  having orthogonal orientations, which additionally leads to the two signal contributions having a relative phase of  $180^\circ$ . This relative phase is a consequence of the way the rotating electric field is prepared with the use of a quarter wave plate, as described by Eq. 7.

We first calculate the signal contribution from the symmetric excitons. Substitution of Eq. (6) and Eq. (8a) into the first term on the right-hand side of Eq. (11) leads to:

$$I_+(\varphi) = \sum_{\alpha} \int d\omega |E(\omega, \varphi) \cdot \boldsymbol{\mu}_+^{(\alpha)}(\omega)|^2 \quad (12a)$$

$$= |\mu_+|^2 [\cos^2(\psi/2)\cos^2(\varphi/2) + \sin^2(\psi/2)\sin^2(\varphi/2) \quad (12b)$$

$$-2\cos(\psi/2)\sin(\psi/2)\cos(\varphi/2)\sin(\varphi/2)]$$

where we have defined the spectral overlap function  $|\mu_+|^2 = |A(\omega)|^2 \sum_{\alpha} \int d\omega |\boldsymbol{\mu}_+^{(\alpha)}(\omega)|^2$ .

Equation (12b) can be simplified using double angle formulas to yield

$$I_+(\varphi) = \frac{|\mu_+|^2}{2} [1 + \cos(\varphi + \psi)] \quad (13)$$

Following a similar procedure for the anti-symmetric exciton contribution to the polarized signal [by substitution of Eq. (6) and Eq. (8b) into the second term on the right-hand side of Eq. (11)] leads to:

$$I_-(\varphi) = \frac{|\mu_-|^2}{2} [1 - \cos(\varphi + \psi)] \quad (14)$$

where  $|\mu_-|^2 = |A(\omega)|^2 \sum_{\alpha} \int d\omega |\mu_{-}^{(\alpha)}(\omega)|^2$ . Combining Eqs. (13) and (14) gives the expression for the polarized single-molecule fluorescence intensity:

$$\begin{aligned} I_P(\varphi) = I_+(\varphi) + I_-(\varphi) &= \frac{|\mu_+|^2 + |\mu_-|^2}{2} \left\{ 1 + \left[ \frac{|\mu_+|^2 - |\mu_-|^2}{|\mu_+|^2 + |\mu_-|^2} \right] \cos(\varphi + \psi) \right\} \\ &= f \{ 1 + v \cos(\varphi + \psi) \} \end{aligned} \quad (15)$$

In Eq. (15),  $f = \frac{1}{2}[|\mu_+|^2 + |\mu_-|^2]$  is the mean signal rate and  $v = [|\mu_+|^2 - |\mu_-|^2]/[|\mu_+|^2 + |\mu_-|^2]$  is the signal visibility.

From Eq. (15), we see that the signal is a sum of two terms. The first term is the ‘instantaneous’ signal flux  $f$ , which is assumed to be constant during a particular measurement period, and the second ( $\varphi$ -dependent) term varies rapidly at the phase modulation frequency ( $\varphi = \Omega t$ ,  $\Omega = 1$  MHz). The modulated signal component is proportional to the visibility  $v$ , which is directly related to the local conformation of the exciton-coupled (iCy3)<sub>2</sub> dimer probe. The signal phase  $\psi$  represents the orientation of the dimer in the lab frame. Thus, to monitor the internal

conformational fluctuations of the (iCy3)<sub>2</sub> dimer probe, it is necessary to isolate the modulated signal component  $\nu$  from the signal flux  $f$  subject to a well-defined phase  $\psi$ .

Under conditions of high signal intensity, a linear detector, such as a photomultiplier tube (PMT), might be used in place of the APD illustrated in Fig. 2.3. The continuously varying signal described by Eq. (15) could then be measured using a conventional lock-in amplifier [69]. However, under the low-signal-flux conditions of the current work, Eq. (15) describes the probability that the APD detects a single photon at time  $t$ . Thus, at any instant, we consider each detection event to be a discrete sample of the signal phase probability distribution function, sampled by the Dirac delta function  $\delta(\varphi - \varphi_i)$ , where the  $\varphi_i$  are random variables. In our single-molecule experiments, the mean signal flux (integrated over many seconds) was typically  $\bar{f} \cong 8 - 10$  kHz, so that on average only one detection event occurred per 100 – 125 modulation cycles. Therefore, to sample the low flux signal as efficiently as possible, we implemented the phase-tagged photon counting (PTPC) method [69]. In this approach, the electronic waveform used to drive the relative phase of the interferometer was synchronized to an 80 MHz digital counter (see Fig. 2.3). We thus collected a stream of sparsely sampled single-photon detection events in which we assigned to the  $n$ th detection event a ‘time stamp,’  $t_n$ , with resolution  $\Delta t = 1 \mu\text{s}$ , and a ‘phase stamp,’  $\varphi_n$ , with resolution  $\Delta\varphi = 360^\circ / 80 = 4.5^\circ$ . In post-data-acquisition, we calculated the discrete signal intensity from the set of  $N$  detected photons that were measured during an (adjustable) sampling window  $T_w$  centered at a particular time  $t$ .

$$I_P^d(\varphi, t) = \frac{1}{T_w} \int_{t-(T_w/2)}^{t+(T_w/2)} \sum_{n=1}^N \delta(\varphi - \varphi_n) \delta(t' - t_n) dt' \quad (16)$$

We isolated the modulated component of Eq. (16) by calculating the first term of the complex-valued Fourier series expansion with respect to  $\varphi$  [69].

$$Z_P(t) = \frac{1}{2\pi} \int_0^{2\pi} I_P^d(\varphi, t) e^{-i\varphi} d\varphi \quad (17)$$

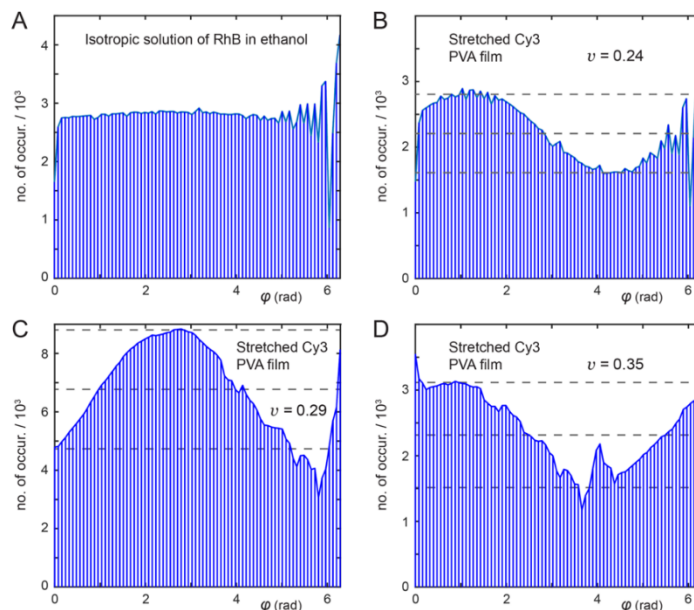
Substitution of Eq. (16) into Eq. (17) shows that the modulated signal component is the sum of  $N$  single photon phase factors:

$$Z_P(t) = \frac{1}{T_w} \sum_{n=1}^{N(t)} e^{-i\varphi_n} = f(t) v(t) e^{-i\psi} \quad (18)$$

The second equality of Eq. (18) relates the visibility to the modulated signal component according to  $v(t) = [Z_P(t)]/f(t)$ , where  $f(t) = N(t)/T_w$  is the instantaneous signal flux and  $N(t) = \int_{t-(T_w/2)}^{t+(T_w/2)} \sum_{n=1}^N \delta(t' - t_n) dt'$  is the number of photons measured at time  $t$  during the sampling window  $T_w$ . We note that the instantaneous flux can be obtained by averaging the signal [Eq. (11)] over the phase variable:  $f = \frac{1}{2\pi} \int_0^{2\pi} I_P(\varphi) d\varphi = N/T_w$ .

To demonstrate that the polarization state of the source laser is well described by Eq. (7), and that the phase ( $\varphi$ )-dependent signal behaves according to Eq. (11), we carried out control measurements on *i*) an isotropic solution of Rhodamine B in ethanol, and *ii*) on an anisotropically-stretched poly vinyl alcohol (PVA) film containing Cy3 (see Fig. 2.4). As expected, the

fluorescence signal of the anisotropic sample exhibited a pronounced  $\varphi$ -dependent modulation, while no significant  $\varphi$ -dependent modulation was observed from the isotropic sample.



**Figure 2.4** Control measurements of phase histograms from (A) an isotropic solution of Rhodamine B in ethanol, and (B – D) an anisotropically-stretched film of Cy3 supported in poly vinyl alcohol (PVA). For each of the panels, B – D, the uniaxial stretch direction and spot location of the PVA film was varied. Histograms were constructed from an average of 10 separate time series, each of  $\sim 45$  s duration. For visualization, the histograms were constructed by re-binning the native 16,000 bin histograms into 79 bins, so that the resolution shown is coarse-grained to  $\sim 0.08$  rad bin $^{-1}$ . The horizontal dashed lines indicate the maximum (top), mean (middle) and minimum (bottom) number of counts per bin. The signal was detected using the phase-tagged photon counting (PTPC) method [69].

### Section 3.6. Characterization of conformational macrostates using probability distribution functions (PDFs)

We monitored the local conformational fluctuations of the (iCy3) $_2$  dimer-labeled ss-dsDNA constructs as reflected by measurements of the PS-SMF signal visibility,  $v(t)$ . We analyzed these data by constructing ensemble averaged probability distribution functions (PDFs) and multi-point



time-correlation functions (TCFs). In the current Chapter, we provide an overview of the statistical analysis approach which can yield these data surfaces. However, by eventually applying a kinetic network model analysis like the approach of Israels *et al.* [14], we may obtain from these data equilibrium distributions of macrostates and kinetic rate constants necessary to parameterize the free energy landscapes of these systems. A quantitative analysis of the free energy landscapes is the focus of Chapter 4. This section serves to introduce the core components needed for a kinetic network analysis.

In order to perform a kinetic network analysis and uncover the free energy landscape which describes the observed thermodynamics, the relative population of conformational macrostates must be captured [14]. Here, we define macrostates (also referred to simply as ‘states’) as groups of configurations, or microstates, that produce similar macroscopic measurements of the observable of interest (i.e. the signal visibility). The microscopic configurations within a macrostate occupy the same free energy minima and are well separated from other macrostates free energy minima by transition energy barriers. We acknowledge that our macrostate assignments obscure interconversion between microstates of the system, which are likely occurring on the nanosecond timescale and faster, but we believe this analysis can capture essential mechanistic details despite the inherent simplifications. The experimentally derived histogram, which serves as the basis for the PDF in our analysis, is a sum over the instantaneous value of the system observable at a particular choice of integration time for all single molecule trajectories in a data set. We define the individual peaks in our experimental histogram using a sum of gaussian distributions, given by Eq. (19).

$$A_j(v) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{(v - v_j)^2}{2\sigma_j^2}\right) \quad (19)$$

The choice of a gaussian to describe the probability of observing a particular value of the visibility within each conformational macrostate is of course the lowest order treatment for such a PDF. Alternative distributions could be used instead of a pure gaussian treatment, such that the effects of anharmonicity in the free energy landscape could be described. But such a treatment would require a greater level of detail than is presently available via the retrievable information in our single-molecule experiments. We set the center of each gaussian defined by Eq. (19) based on a two-step process. First a standard gaussian decomposition is used to establish reasonable boundaries for the allowed values of each peak's center. Then second, the established boundaries are used in a nonlinear optimization to fit the experimentally derived data surfaces to a model, of which the set of visibilities defining each microstate is an output, as discussed in Chapters 3 and 4. We define the probability distribution of conformational macrostates as a sum over Boltzmann weighted peaks of the experimental observable after optimization of the gaussian peak centers and integrated area.

$$p_j(v) = \exp\left(-\frac{G_j(v)}{k_B T}\right) / Z \quad (20)$$

The histogram reports on the probability distribution within each underlying macrostate. The Boltzmann relation, transforms the probabilities of each state to their corresponding free energy distributions, where  $G_j(v)$  is the free energy as a function of the experimental observable,  $v$ ,  $k_B$  is Boltzmann's constant,  $T$  is the temperature and  $Z = \sum_{j=1}^M \exp\left(-\frac{G_j(v)}{k_B T}\right)$  is the partition

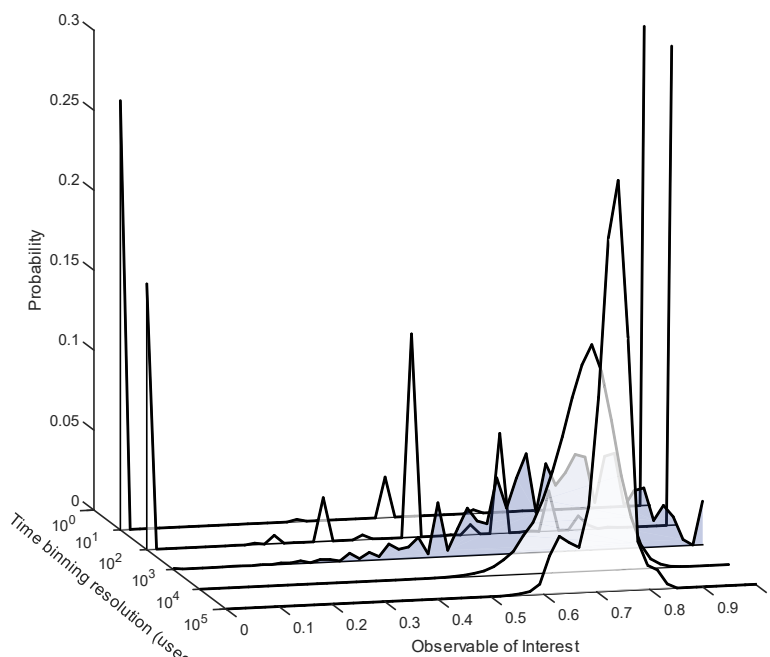
function. Rearranging this expression to solve for the free energy gives the free energy minima as a function of our experimental observable (Eq. 21).

$$\frac{G_j(v)}{k_B T} = -\ln[p_j(v)] = -\ln\left(\frac{p_j^{eq}}{\sigma_j \sqrt{2\pi}}\right) + \frac{1}{2\sigma_j^2}(v - \langle v \rangle_j)^2 \quad (21)$$

Eq. (21) will serve as the central relationship in later Chapters to translate the outcomes of our analysis pipeline into thermodynamic parameters of interest. For the purpose of a pure signal discussion, we define the histogram of our observable,  $A$ , which could be the visibility, FRET or some other quantity, as  $P(A) = \sum_{j=1}^N p_j(A)$  where  $p_j(A) = \alpha_j \exp\left(-\frac{(A - \langle A_j \rangle)^2}{2\sigma_j^2}\right)$  where  $\alpha_j = p_j^{eq}/\sigma_j \sqrt{2\pi}$ ,  $\langle A_j \rangle$  is the mean observable value, and  $\sigma_j$  is the standard deviation of the  $j^{th}$  distribution, as defined by Eq. 19. The area of each of these distributions corresponds to the  $p_j^{eq}$  such that  $p_j^{eq} = \int_{-\infty}^{\infty} p_j(A) dA = \alpha_j \sigma_j \sqrt{2\pi}$  and that the sum over the  $p_j^{eq}$  distributions is unity, i.e.,  $\sum_{j=1}^N p_j^{eq} = 1$ . We require that  $\mathbf{p}^{eq}$  values resulting from the histogram fits must be consistent with the  $\mathbf{p}^{eq}$  values we obtain from the TCF simulations, as discussed below.

The final discussion point concerning experimental histograms is the choice of an integration time. The integration time sets the bin size used to sum up individual photons phase factors,  $T_w$  in Eq. (18), to obtain an estimate of the polarized signal intensity. Under typical conditions of sampling from a fixed, singularly valued, distribution describing some system observable, the standard error associated with the sampled distribution goes as  $\frac{1}{\sqrt{n}}$ , where  $n$  is the number of independent sampling events performed on the underlying distribution. This standard situation suggests that sampling for as long as possible is optimal, to limit the standard error of the sampling to some minimum value. However, in the case of a dynamic, multimodal distribution, as

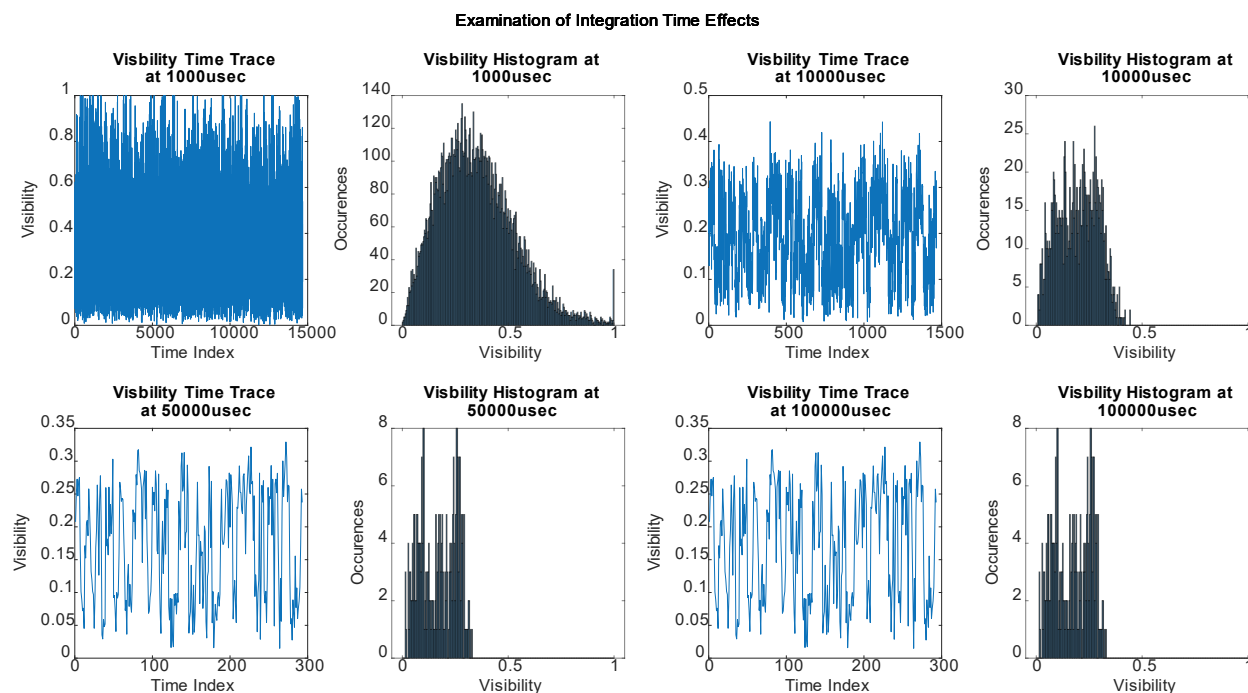
is the case for the single-molecule observables under consideration here, an optimal choice of integration time comes down to a compromise between the lower standard error associated with longer integration times and the loss of important features in the histogram to the effects of time averaging multiple distinct values together, a consequence of the central limit theorem. These effects, and others pertinent to the TCFs discussed later, have been nicely describe by Berg and coworkers [78] An illustrative example of various choices in integration time is shown below in Fig. 2.5.



**Figure 2.5** Model demonstration of variable integration time on the emergence of multimodal behavior and suppression of noise in the experimental histogram of a single molecule observable. Each curve displayed is the same underlying experimental data, integrated with using various bin sizes.

The earliest choice of 10usec in this example yields highly discrete peaks at the limits of the domain due to single photon statistics, giving a poor estimate of the macrostates in the system. At the level of 10ms, a clear picture of the underlying PDF governing the macrostates emerges. In general, the choice of integration time depends on the data set and system under

consideration. The optimal choice will always be the fastest integration time which affords a robust estimate of the signal (SNR  $\sim 10$ ). But given the variability in signal intensity, or photon flux, the choice must be made on a case-by-case basis. A panel of experimentally derived histograms, for a single PS-SMF example trace, at various integration times are shown in Fig. 2.6 as the corresponding example of non-idealized binning time effects.



**Figure 2.6** Experimental demonstration of variable integration time on the emergence of multimodal behavior and suppression of noise in the experimental histogram of a single molecule observable. Each panel is the same experimental data, integrated with using various bin sizes. Both the trajectory of the visibility and the resulting PDF are shown.

### Section 3.7 Characterization of conformational dynamics using multi-point time-correlation functions (TCFs)

Time correlation functions are an invaluable tool for isolating meaningful fluctuations from stochastic noise in a time series. Here, we use them to extract the rates of interconversion between macrostates and characterize the kinetic information contained in our PS-SMF experiments using

two-point and three-point TCFs [13], [14], [70]. We define the time-dependent fluctuation of the visibility  $\delta v(t) = v(t) - \langle v \rangle$ , where  $\langle v \rangle$  is the mean value determined from a time series of measurements performed on a single molecule for all later discussions of PS-SMF data. For the discussion which presently follows, a general experimental observable  $A$  will be used, which could be the visibility, FRET or any other equilibrium observable obtained from single molecule studies. The simplest TCF is the two-point time correlation function (Eq. 22), which is the average product of two successive measurements separated by the interval  $\tau$ :

$$\bar{C}^{(2)}(\tau) = \langle \delta A(\tau) \delta A(0) \rangle \quad (22)$$

In Eq. (22), the angle brackets indicate a running average over all possible initial measurement times. The function  $\bar{C}^{(2)}$  contains information about the number of quasi-stable macrostates of the system, the mean value of each macrostate, and the characteristic time scales of interconversion between macrostates [13], [56], [70]. If the signal rate is continuous, we can use the Weiner-Khinchin theorem to calculate the two-point TCF from the experimental data (Eq. 23),

$$\bar{C}_E^{(2)}(\tau) = \frac{1}{N_{binned}^2} \mathcal{F}^{-1}\{|\mathcal{F}[\delta A(\tau)]|^2\} \quad (23)$$

where  $N_{binned}$  is the number of counts per integration period and  $\mathcal{F}$  and  $\mathcal{F}^{-1}$  are the fast Fourier transform and inverse fast Fourier transform, respectively. Otherwise, if the sampling rate is fast with respect to the frequency of signal events, producing a sparse time series, we use the average product of the discrete observable values (Eq. 24),

$$\bar{C}_E^{(2)}(\tau) = \frac{1}{N_{2P}(\tau)} \sum_{i=0}^{N_{tot}-1} [\delta A(t_i) \delta A(t_i + \tau)] \quad (24)$$

where  $N_{2P}$  is the number of two-point pair-wise products where both bins contain valid data,  $N_{tot}$  is the total number of possible pairs spaced apart by the delay  $\tau$  and  $t_i$  are the starting indexes. We note that in the limit that the signal becomes continuous the results of Eq. 23 are identical to Eq. 24. The analysis presented here is recommended for systems that contain dynamics with well separated timescales in the two-point TCF, such as the data presented in section 4. As discussed in detail in Chapter 3, section 3.1, the number of well separated decay components in the two-point TCF can be used to determine the minimum number of macrostates required to model the data with a kinetic network model.

While the two-point TCF reports on direct state-to-state transitions, we also use higher-order TCFs to study the effects of multi-step kinetic pathways that influence the observed dynamics in our experiments. In previous studies [14], [56], [70], our group has previously applied four-point TCFs to calculate the average product of the signal fluctuation across four points separated by three-time intervals,  $\tau_1 = t_2 - t_1$ ,  $\tau_2 = t_3 - t_2$  and  $\tau_3 = t_4 - t_3$  (Eq. 25).

$$\bar{C}^{(4)}(\tau) = \langle \delta A(\tau_1 + \tau_2 + \tau_3) \delta A(\tau_1 + \tau_2) \delta A \delta(\tau_1) A(0) \rangle \quad (25)$$

As before, the Weiner-Khinchin theorem can be used to calculate higher order TCFs for experiments with high enough acquisition rates to be considered continuous. However, for the studies discussed here, data sets are sparse at the chosen binning times, so we must use the discrete product form with  $N_{4P}$  as the number of four-point products where all bins contain valid data (Eq 26).

$$\bar{C}_E^{(4)}(\tau_1, \tau_2, \tau_3) = \frac{1}{N_{4P}(\tau_1, \tau_2, \tau_3)} \sum_{i=0}^{N_{tot}-1} [\delta A(t_i) \delta A(t_i + \tau_1) \delta A(t_i + \tau_1 + \tau_2) \delta A(t_i + \tau_1 + \tau_2 + \tau_3)] \quad (26)$$

Our applications of four-point time correlation functions have typically allowed the  $\tau_2$  interval to be zero, reducing Eq 25 to:

$$\bar{C}^{(4)}(\tau_1, \tau_2 = 0, \tau_3) = \langle \delta A(\tau_1 + \tau_3) \delta A(\tau_1)^2 \delta A(0) \rangle \quad (27)$$

Recently, Berg and coworkers showed that when performing these TCF calculations, the effect of noise must be considered [78]. Here, we define our signal fluctuation as  $\delta \bar{A}(t) = A(t) + \epsilon(t) + b - \langle A(t) \rangle - \langle \epsilon(t) \rangle - \langle b \rangle \approx \delta A(t) + \epsilon(t)$  where  $A(t)$  is our observable of interest,  $\epsilon(t)$  is time dependent noise (e.g., gaussian distributed dark counts, errors in phase and time tagging electronics, etc.)  $b$  is some constant background and  $\langle \dots \rangle$  indicate the time average. We note that the time dependent noise overall averages to zero,  $\langle \epsilon \rangle = 0$ , it is uncorrelated at different times,  $\langle \epsilon(t) \epsilon(t') \rangle = 0$  where  $t \neq t'$ , and it is uncorrelated with the signal  $\langle A(t) \epsilon(t') \rangle = 0$  for all  $t, t'$ . However, the time dependent noise *is* correlated with itself at the same instant in time, i.e.,  $\langle \epsilon(t) \epsilon(t) \rangle = \sigma_\epsilon^2 \neq 0$ . Plugging  $\delta \bar{A}(t)$  into Eq 27 and simplifying produces Eq. 28:

$$\bar{C}^{(4)}(\tau_1, \tau_2 = 0, \tau_3) = \langle \delta A(\tau_1 + \tau_3) \delta A(\tau_1)^2 \delta A(0) \rangle + \sigma_\epsilon^2 \langle \delta A(\tau_1 + \tau_3) \delta A(0) \rangle + \sigma_\epsilon^2 \delta(\tau_1) \sigma_\epsilon^2 \delta(\tau_3) \quad (28)$$

where  $\sigma_\epsilon^2 = \langle \epsilon(t) \rangle^2$ , and  $\delta(t)$  is the Kroneker Delta function. The first term in Eq. 28 is the signal correlation we are interested in. The third term is a noise term that is only nonzero when  $\tau_1$  and  $\tau_3$  are zero, at the first point of the TCF. But the second term is a contribution to the higher-ordered correlation which is proportional to the variance of the time dependent noise in the system. Thus, when using a four-point time correlation function, with one delay set to zero, there is a non-negligible component of correlated time dependent noise.



We resolved this problem by switching our analysis to calculating three-point TCFs (Eq. 29), instead of four-point TCFs with one delay at zero (Eq. 27). The three-point TCF is the time-averaged product of three consecutive measurements separated by the intervals  $\tau_1$  and  $\tau_2$ .

$$\begin{aligned}\bar{C}^{(3)}(\tau_1, \tau_2) &= \langle \delta A(\tau_1 + \tau_2) \delta A(\tau_1) \delta A(0) \rangle \\ &= \frac{1}{N_{3P}(\tau)} \sum_{i=0}^{N_{\text{tot}}-1} [\delta A(t_i) \delta A(t_i + \tau_1) \delta A(t_i + \tau_1 + \tau_2)]\end{aligned}\quad (29)$$

This function is sensitive to the roles of intermediates, like four-point TCF, whose presence can facilitate or hinder successive transitions between macrostates. However, unlike even-moment TCFs, the three-point TCF,  $\bar{C}^{(3)}$ , contains no underlying noise-related background term that could potentially obscure the experimentally derived surface.  $\bar{C}^{(3)}(\tau_1, \tau_2)$  does not result in a correlation of the time dependent noise when  $\delta \bar{A}(t)$  is plugged into Eq. 29. Instead, we are left with the three-point correlation of interest and the noise term, which vanishes after the  $\tau_1 = \tau_3 = 0$  point, thus avoiding the issues with the  $\bar{C}^{(4)}(\tau_1, \tau_2 = 0, \tau_3)$  [71]. This is given below by Eq. 30. Going forward we proceed with the three-point TCF to study higher order correlations in single-molecule time series.

$$\bar{C}^{(3)}(\tau_1, \tau_2) = \langle \delta A(\tau_1 + \tau_2) \delta A(\tau_1) \delta A(0) \rangle + \sigma_E^2 \delta(\tau_1) \sigma_E^2 \delta(\tau_2) \quad (30)$$

A notable feature of the three-point correlation function is its inherently higher degree of noise compared with the two-point correlation function. This is due to the reduced number of available products of the signal observable at three unique points in time, compared to two unique points in time. Due to the limiting SNR of the three-point TCF, decisions about minimal integration times in experimental data sets should be based on the availability of an analyzable surface, for kinetic network modeling, in the case of the three-point TCF.

In the analysis, which is presented in Chapter 2, section 4, we constructed our time-dependent measurements of the PS-SMF visibility the PDF,  $P(v)$ , using the sampling window  $T_w = 10$  ms. We chose this value for  $T_w$  because it is roughly equal to the time required for the signal-to-noise ratio (S/N) to acquire a value of  $\sim 10$ . In addition, we calculated two-point and three-point TCFs,  $\bar{C}^{(2)}$  and  $\bar{C}^{(3)}$ , respectively, using the sampling window  $T_w = 250 \mu\text{s}$ . Typically,  $\sim 150 - 250$  individual single-molecule data sets were combined to construct each of the experimentally-derived statistical functions.


While TCFs are a powerful tool for extracting dynamics from experimental time series, there are some conditions that need to be satisfied. First, the main source of information should be obtained from deviations in the observable of interest from the mean, i.e., fluctuations,  $\delta A$ . The analysis could be applied to pure observable values, but we expect the sensitivity to decrease as the dynamic range of the observable correlation will be greatly reduced in the cases where the dynamic fluctuation is small compared to the average. Additionally, these fluctuations need to come about in an experiment where the system and surroundings are in equilibrium. Stopped-flow kinetic experiments and force-pulling single molecule measurements would not typically meet these criteria, except during periods of constant solvent environment or net zero force applied, respectively [15], [23]

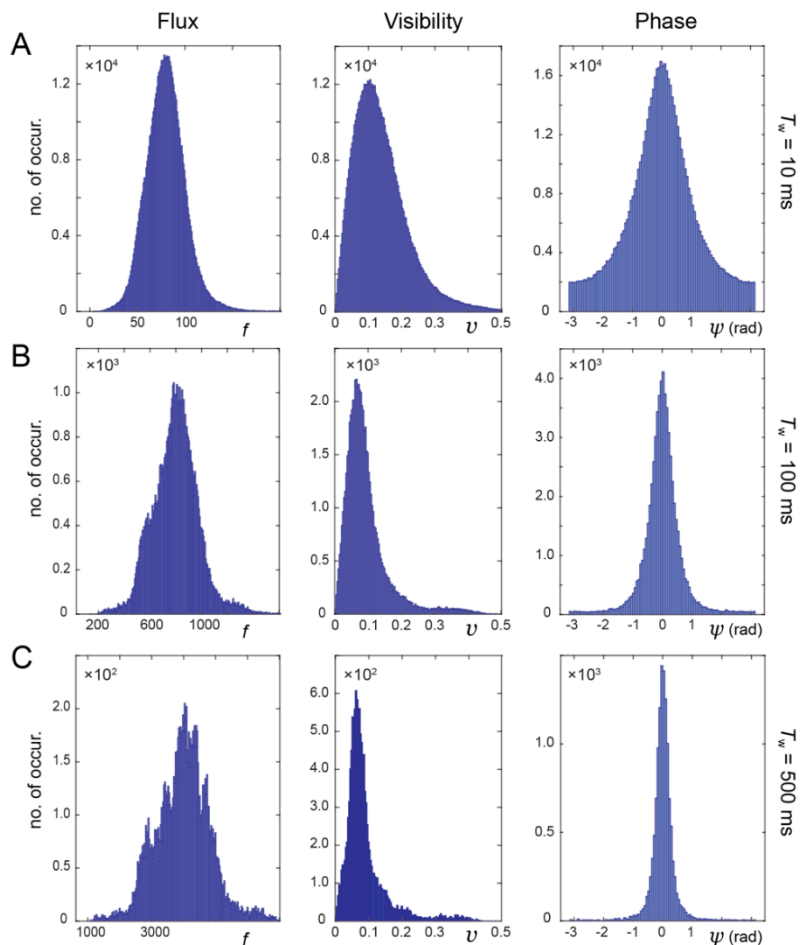
Another consideration is the balance between the fluorescence lifetime and quantum yield of the molecular probe and the measurement speed of the instrument. The instrument needs to have a sampling rate fast enough to capture the dynamics of interest, but a point of diminishing returns is reached when the sampling rate well surpasses the flux from the fluorescent probe, such that the desired integration periods are mostly empty or poorly resolved. Brighter dye molecules provide more frequent sampling of the state of the system and, ultimately, better signal-to-noise of fast

dynamics can be achieved when in combination with fast measurement speed. The importance of the measurement speed and fluorescent flux needs to be balanced based on experimental constraints and the timescales of the relevant dynamics. Ultimately, time correlation functions can be an excellent way to study the dynamics of single-molecule data from an experiment designed for an informative fluctuating observable, a sufficiently bright fluorescent probe and correspondingly fast measurement capabilities.

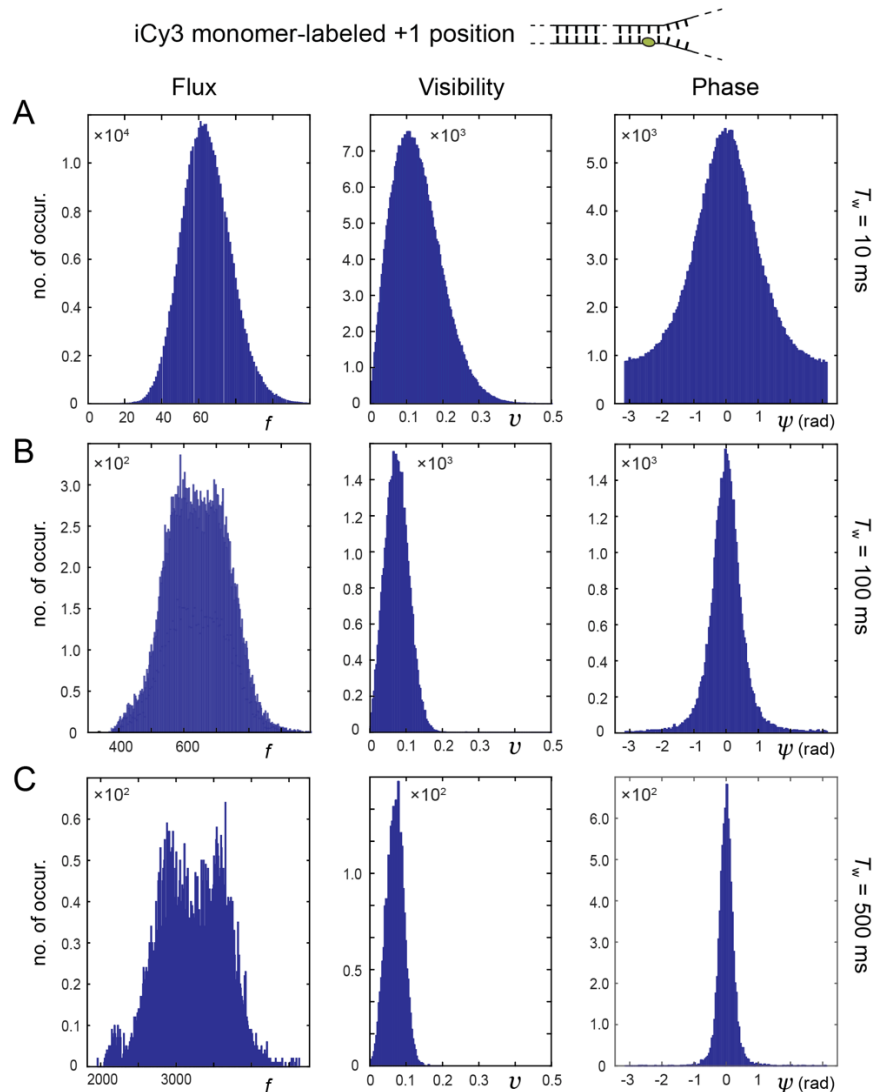
### **Section 3.8. Consideration of fluorescence correlation spectroscopy (FCS) as a reporter on conformational dynamics of (iCy3)<sub>2</sub> dimer-labeled single-stranded (ss) – double-stranded (ds) DNA fork constructs**

A commonly employed approach to investigating macromolecules and their dynamics is to study correlations in the pure fluorescence intensity from a single molecule, known as fluorescence correlation spectroscopy. While successful at determining the concentration, diffusive properties, hydrodynamics radius and in some limited cases molecular interactions, the ability to parse generic photo-physics from true molecularly induced changes to the intensity remain difficult [79]. This section will discuss key features of the visibility signal,  $v(t)$ , versus the pure signal rate, or intensity,  $f = \frac{1}{2\pi} \int_0^{2\pi} I_P(\varphi) d\varphi = N/T_w$ , and their respective abilities to act as a reporter on the internal conformational changes of ss-dsDNA junctions. It is natural to think that fluctuations in the signal flux might also be a reporter on the internal conformations of dsDNA given that the spectral overlap terms determining the overall fluorescence intensity of a single Cy3 dimer are also a function of the dimer's geometry.

(iCy3)<sub>2</sub> dimer-labeled +1 position 



**Figure 2.7** PS-SMF probability distribution functions (PDFs) for the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork construct. Histograms were constructed from the raw photon data streams for the signal flux (left column), signal visibility (middle column) and signal phase (right column). The mean signal flux is  $\bar{f} = 7,500 \text{ s}^{-1}$ . Comparisons are shown varying the integration window (**A**)  $T_w = 10$  ms, (**B**)  $T_w = 100$  ms and (**C**)  $T_w = 500$  ms. The corresponding  $S/N = \sqrt{N} = \sqrt{\bar{f}T_w} = 8.7, 27$  and  $61$ , respectively.



**Figure 2.8** PS-SMF probability distribution functions (PDFs) for the +1 iCy3 monomer-labeled ss-dsDNA fork construct. Panels are the same as described in Fig. 2.7.

As discussed in section 3.6 and 3.7, our approach to analyzing and interpreting dynamics in macromolecular assemblies relies on the ability to construct a PDF and calculate TCFs. The need for a well-behaved PDF shouldn't be understated, given the often-complicated deconvolution process which can influence the centers and number of macrostate assignments within a dataset. Figures 2.7 and 2.8 demonstrate the effects of variable integration time for both a Cy3 dimer and Cy3 monomer, which is an important factor in the

SNR and dynamical resolution of the PDF, for three separate signal quantities: the flux, the visibility, and the phase. For both the Cy3 dimer and Cy3 monomer cases, the flux and visibility appear reasonably well behaved at the lower integration time limit of 10ms. However, as the integration time is increased, the visibility nicely converges to a distinct set of values with well defined averages, while the flux begins to take on multi-modal characteristics which neither converge to a well-defined average nor appreciably decrease in their standard deviation. The width of the phase distribution in all cases tracks with the standard deviation of the visibility as it is an internal measure of the visibility's certainty. From a pure PDF perspective, if one were to use the signal flux or intensity to establish a discrete set of observable values which correspond to distinct conformational macrostates within the system, then both the number of states assigned as well as their central averages would be a function of integration time. This poses an immense challenge in the configuration of any analysis protocol given the strong variability of model outcome with variable integration time. The visibility signal demonstrably does not suffer from this same issue and is therefore far better suited for analysis approaches which attempt to leverage the discretization of conformational macrostates within biophysical systems. This is owed in part to the visibility being photon number normalized, such that it is impervious to fluctuations in intensity arising from purely photophysical phenomena as well as the variability in average intensity which occurs within and across data sets due to instrumental alignment, laser stability and optical clarity of the sample.

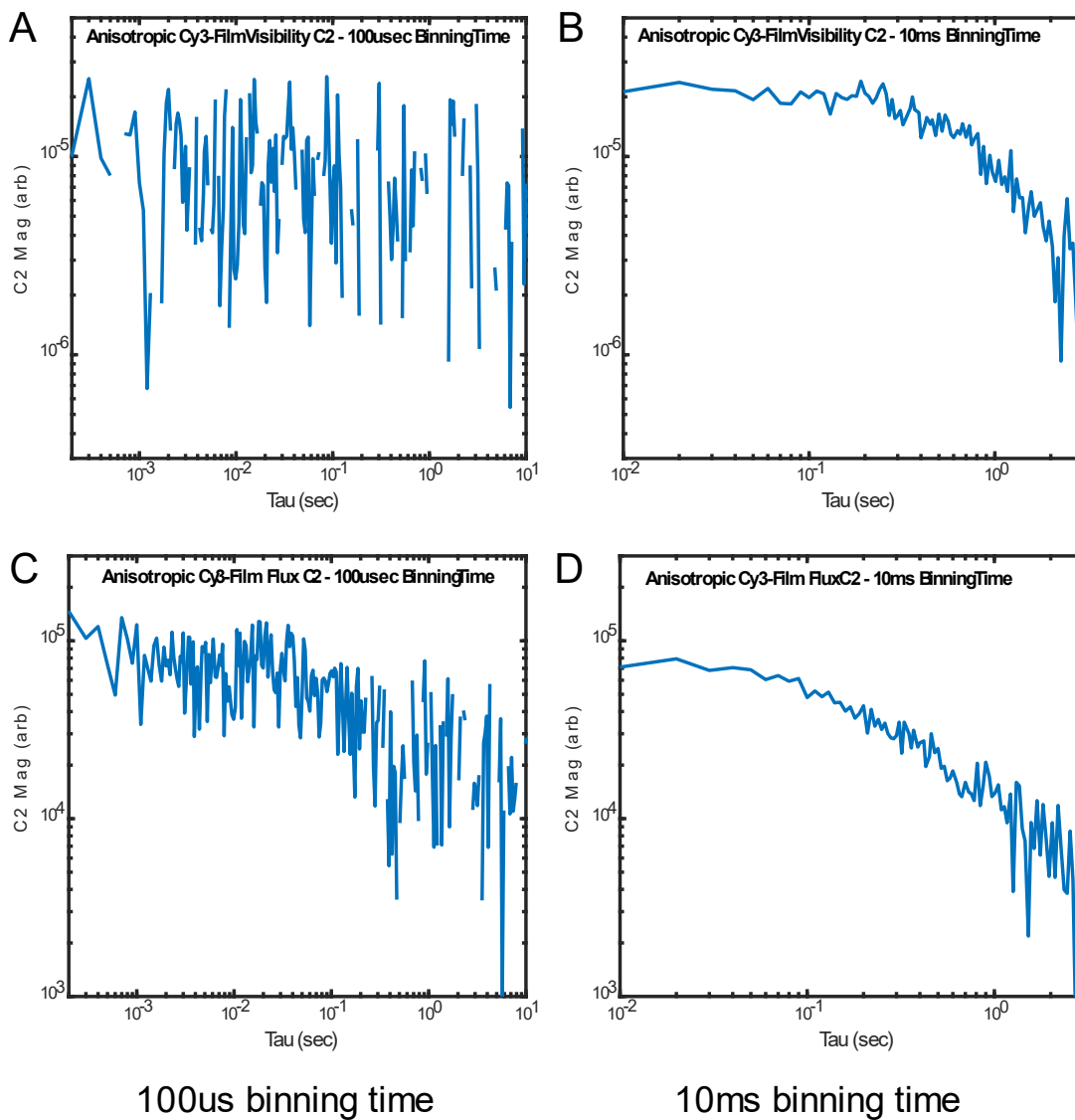
An additionally important consideration in the use of any single molecule observable is, whether systematic sources of random or correlated noise persist on the timescales used to investigate single molecules. In order to address this possibility, control experiments can be

performed for a variety of samples that capture the potential dependence of any signal observable on the lab environment or instrumentation. The use of TCFs to determine the timescales and relative magnitude of spurious correlations in the experimental apparatus is a useful tool for comparing differing signal quantities. A set of TCFs examining the correlations in both the visibility signal and signal flux are shown in figures 2.9 and 2.10.

Figure panels 2.9, A-D and 2.9, A-D provide the experimental basis for the inclusion of a control time scale in our single molecule polarization sweep measurements. Two control samples were measured, a stretched film of Cy3 embedded in a soft polymer matrix and a solution of rhodamine in methanol. The stretched film is an anisotropic sample, displaying a significant collective orientation of individual Cy3 transition dipole moments within the film, as illustrated in Fig. 2.4. The completely isotropic rhodamine sample provides a check on the electronics and phase tagging method, as any non-zero visibility or correlations within the visibility ought to be absent in this case. The photon streams from both samples were separately recorded and time correlation functions were calculated from those data streams.

What can be seen immediately from examination of plots A and B of Fig. 2.9 and 2.10 is that only the Cy3 film displays any correlation in the signal visibility, with an exponential decay timescale on the order of 2-4 seconds. Additionally, this correlation arises only at the limit of long binning time, suggesting that the spurious correlations in the signal visibility are only on slow timescales and not affecting the faster dynamics observed in dsDNA studies. As stated in the main text, we attribute this slow correlation timescale to instrumental drift due to room vibrations. In the case of rhodamine, no visibility correlations are present at either the short or long binning times examined, suggesting that any pure contribution from phase-tagging electronics to the visibility or its correlations are not present.

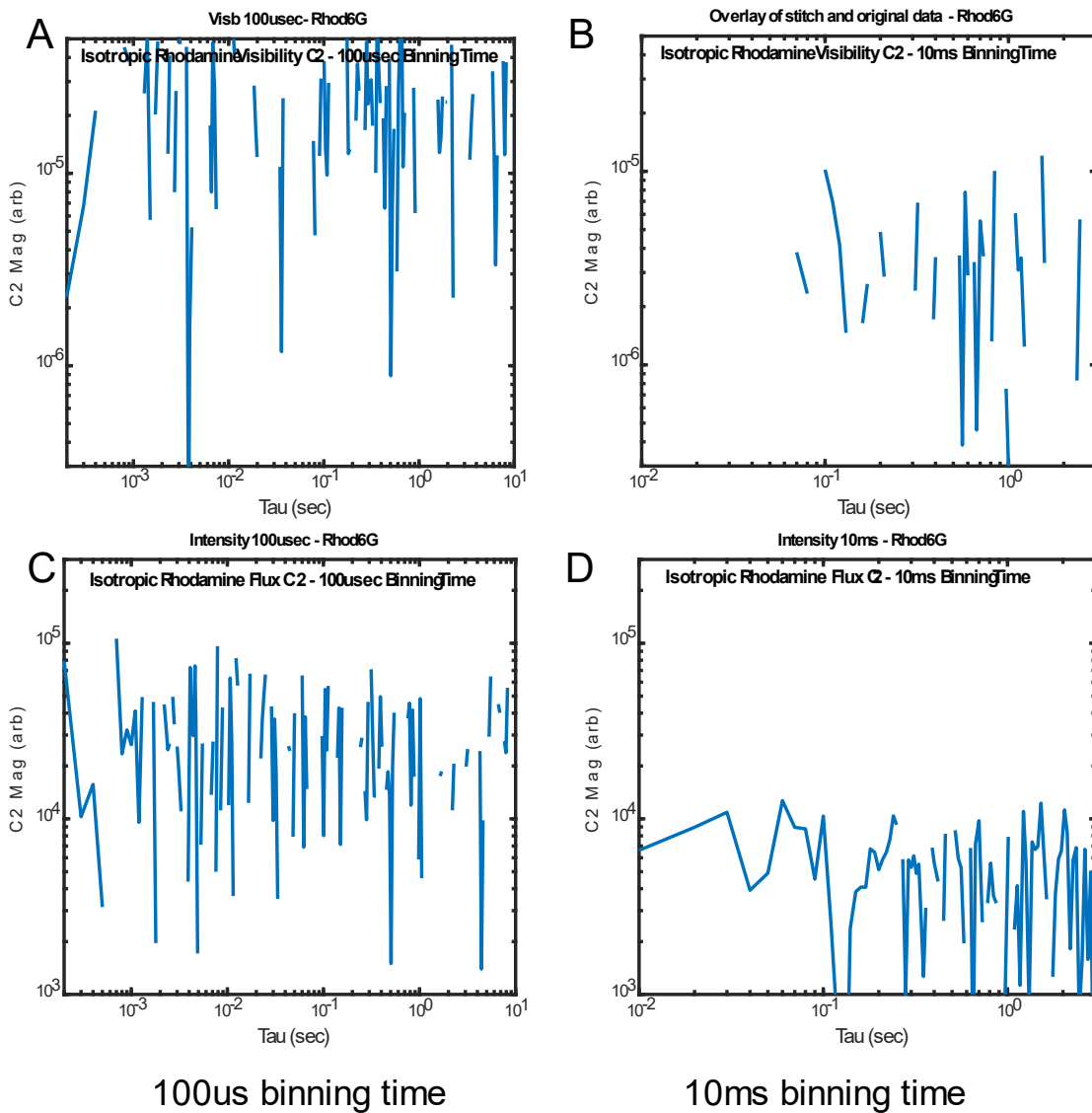
# Cy3 Films



**Figure 2.9** PS-SMF time correlation functions (TCFs) for an anisotropic Cy3-film, both the visibility and flux TCF are calculated and shown in panels A-D. Integration times of 100usec and 10ms are examined for both the visibility and flux case.



# Rhodamine



**Figure 2.10** PS-SMF time correlation functions (TCFs) for an isotropic rhodamine sample, both the visibility and flux TCF are calculated and shown in panels A-D. Integration times of 100usec and 10ms are examined for both the visibility and flux case.

Two additional points should be made regarding these control studies. The first is that the correlations in the flux TCF for either the Cy3 film or rhodamine contain multiple timescales that do not arise in the visibility signal and cannot be easily deconvolved from the relevant timescales

observed in single molecule data. This presents a strong argument against attempting the kinetic network analysis outlined in this work using only the time dependent flux of the molecule. Secondly, the overall magnitude of the correlations seen in the visibility signal from the Cy3 film are on the order of  $10^{-5}$  whereas the correlations seen from single molecule samples are typically on the order of  $10^{-4} - 10^{-3}$ , demonstrating that this contribution from instrument drift is at most a full order of magnitude less than the correlations arising from conformational fluctuations of single molecules.

Taken all together, it is clear that both the PDFs and TCFs of the signal flux are plagued by a host of issues when using the PS-SMF methodology, despite the potential relationship between signal flux and internal geometry of dsDNA reported on by a single Cy3 dimer. Given the longstanding approach of using FCS to study molecular interactions, we believe that the visibility signal in our PS-SMF is a more faithful reporter of the conformational fluctuations in dsDNA and can be used without significant complications, unlike the signal flux, in the kinetic network analysis discussed in Chapter 3.

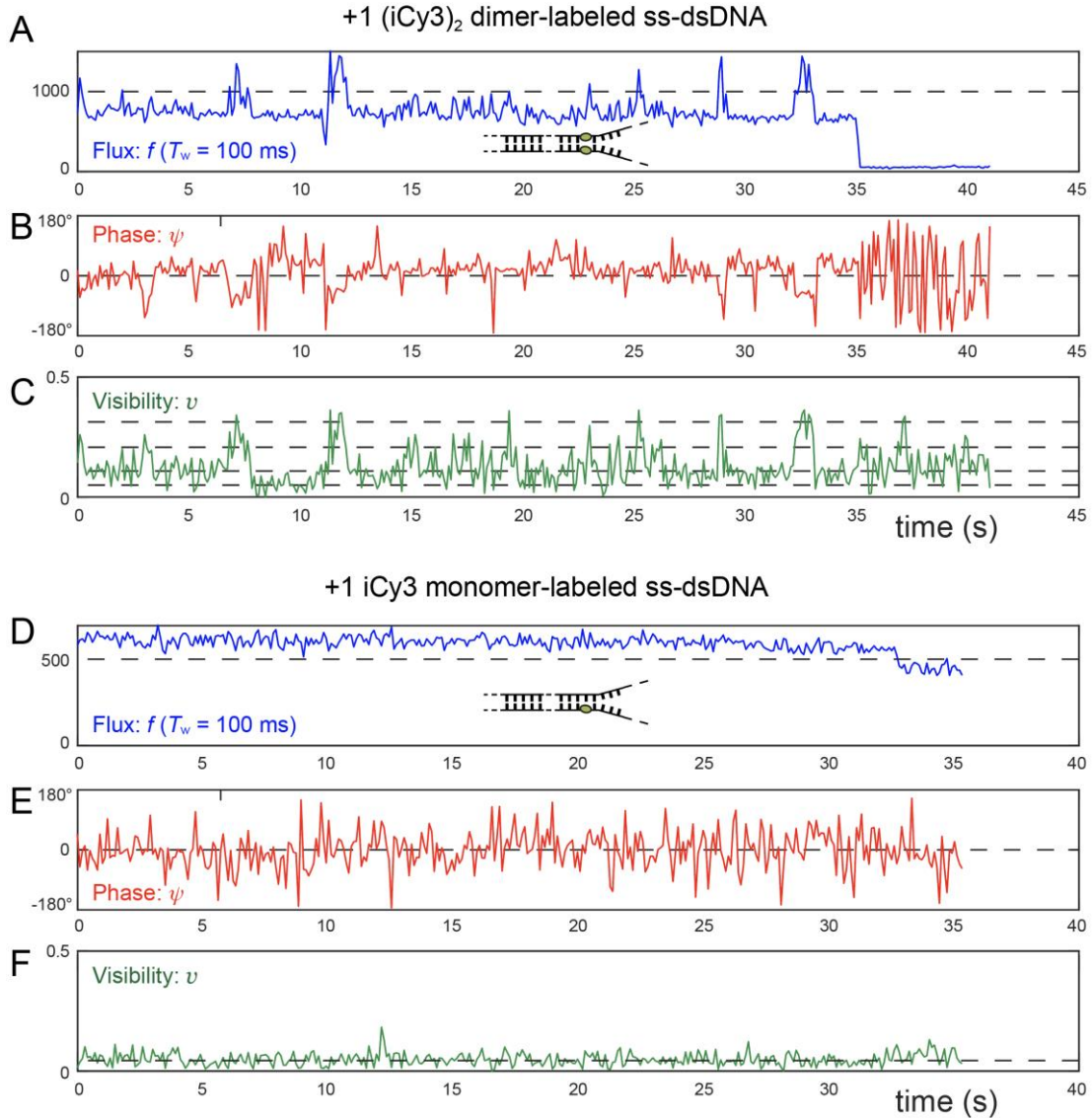
## 4. RESULTS AND DISCUSSION

### **Section 4.1. Analysis of PS-SMF spectroscopic signals.**

For each of the iCy3 monomer and (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs that we studied (see Table 1), we recorded PS-SMF data streams for a typical duration of  $\sim 30 - 50$  s. In Fig. 2.11, we present examples of PS-SMF signal trajectories for the +1 (iCy3)<sub>2</sub> dimer- (see Figs. 2.11, *A - C*) and the +1 iCy3 monomer-labeled ss-dsDNA fork construct (Figs. 2.11, *D - F*) using the integration window  $T_w = 100$  ms. In these examples, the mean flux is  $\bar{f} \approx 7,500$  s<sup>-1</sup> for the

(iCy3)<sub>2</sub> dimer and  $\bar{f} \approx 6,000 \text{ s}^{-1}$  for the iCy3 monomer. We note that the apparent ‘noise’ in the signal trajectories is due to the measurement uncertainty associated with the finite number of photons detected at these low flux levels during the integration window [69]. Thus, the mean signal-to-noise ratio is  $S/N = \sqrt{\bar{f}T_w} \approx \sqrt{750} \approx 27.4$  for the trajectory of the (iCy3)<sub>2</sub> dimer-labeled construct, and  $S/N \approx 24.5$  for the trajectory of the iCy3 monomer-labeled construct.

The three signal components are the instantaneous flux,  $f(t) = N(t)/T_w$  (Figs. 2.11 *A* and *D*), the phase,  $\psi(t)$  (Figs. 2.11 *B* and *E*) and the visibility,  $v(t) = [Z_P(t)]/f(t)$  (Figs. 2.11 *C* and *F*), which are described by Eqs. (16) – (18), respectively. The flux,  $f(t)$ , is equivalent to the fluorescence intensity, a quantity that is independent of the laser polarization since it is averaged over the polarization phase variable, as discussed in the previous section. We note that fluctuations of the flux reflect changes in the (iCy3)<sub>2</sub> dimer probe’s local environment, which influence excited state deactivation pathways. The phase,  $\psi(t)$ , is shown ‘unwrapped’ with its mean value set (arbitrarily) to  $\bar{\psi} = 0$ . The phase represents the orientation of the major axis of the (iCy3)<sub>2</sub> dimer ‘polarization ellipse’ in the laboratory frame (see Fig. 2.2 *C*), which undergoes intermittent jumps about its mean value ( $\bar{\psi} = 0$ ) over the duration of the trajectory. We note that most of these phase jumps appear to traverse a full 360° revolution, which we attribute to mechanical room vibrations.

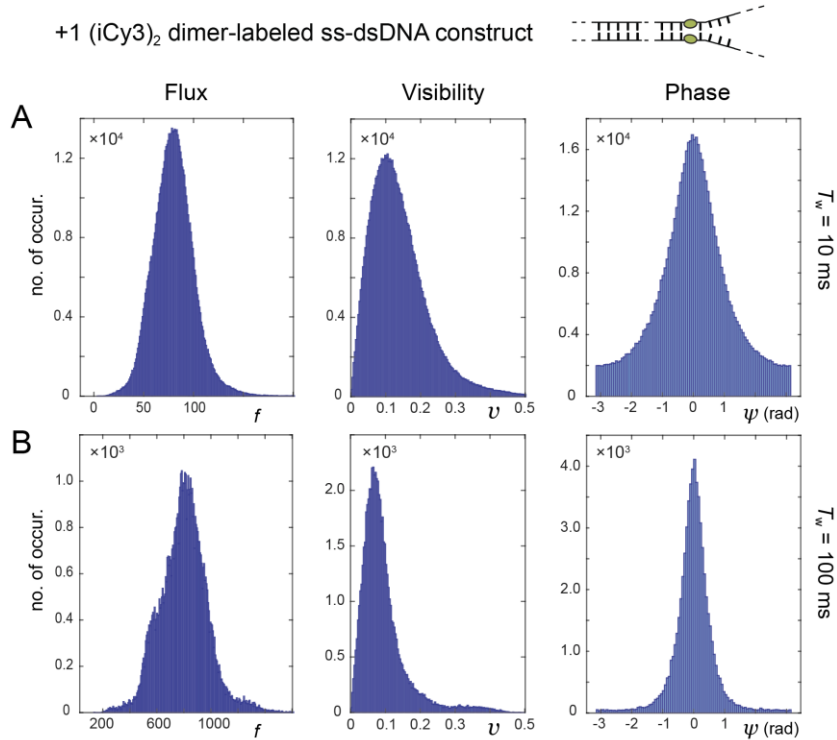


**Figure 2.11** Example PS-SMF signal trajectories using the integration window  $T_w = 100$  ms for the (A – C) +1 (iCy3)<sub>2</sub> dimer- and (D – F) +1 iCy3 monomer-labeled ss-dsDNA fork constructs. Experiments were performed using 100 mM NaCl and 6 mM MgCl<sub>2</sub>. For the +1 (iCy3)<sub>2</sub> dimer-labeled construct the photon data stream was recorded with a mean signal flux  $\bar{f} = \sim 7,500$  s<sup>-1</sup>, corresponding to  $S/N \approx 27.4$ . For the +1 iCy3 monomer-labeled construct the mean signal flux was  $\bar{f} = \sim 6,000$  s<sup>-1</sup>, corresponding to  $S/N \approx 24.5$ . (A, D) The instantaneous signal flux,  $f(t) = N(t)/T_w$ , is shown in blue. (B, E) The instantaneous signal phase,  $\psi$ , is shown in red, ‘unwrapped’ with its mean value set equal to zero,  $\bar{\psi} = 0$ . (C, F) The signal visibility,  $v(t)$ , is plotted in green. Horizontal dashed lines are shown as a guide to the eye to indicate the presence of multiple discrete conformational states for the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct (panel C), while only a single discrete state is observed for the iCy3 monomer-labeled ss-dsDNA construct (panel F).

From Fig. 2.11C, we see that the visibility trajectory for the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct undergoes discontinuous transitions between a small number of discrete values

within the range  $0 < v < 0.4$  (dashed lines are guides to the eye). In this case, the visibility is a direct measure of the internal conformation changes of the (iCy3)<sub>2</sub> dimer probe, as discussed in Sect. 3.4 [see Eq. (11)]. In contrast, the visibility trajectory for the +1 iCy3 monomer-labeled ss-dsDNA construct shown in Fig. 2.11*F* exhibits relatively uniform fluctuations about a single discrete value,  $v \sim 0.08$ . Here, the visibility measures only the mean projection of the iCy3 monomer's 'polarization ellipse' within the transverse plane of the polarized laser beam (Fig. 2.2*C*), which remains relatively constant over the duration of the trajectory. For example, were the EDTM of the iCy3 monomer to be oriented orthogonally to the laser polarization, the visibility would be zero. In principle, the visibility can assume values within the continuous range  $-1 < v < +1$ , with sign inversion corresponding to a 180° phase shift. Nevertheless, the PS-SMF experiment cannot distinguish between negative and positive values of the visibility, so that in practice we measure the absolute value  $|v|$ .

From the PS-SMF data streams, we constructed probability distribution functions (PDFs) for different values of the integration time window,  $T_w$ . In Fig. 2.12, we compare PDFs of the flux,  $f$  (left column), visibility,  $v$  (middle), and phase,  $\psi$  (right), for the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct. Additional PDFs for the +1 (iCy3)<sub>2</sub> dimer- and the +1 iCy3 monomer-labeled ss-dsDNA constructs using  $T_w = 500$  ms are shown in Figs. 2.7 and 2.8, respectively.



**Figure 2.12** PDFs of the flux (left column), visibility (middle) and phase (right) for the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork construct. PDFs were constructed from the raw photon data streams (see, e.g., Fig. 2.10). The mean signal flux is  $\bar{f} = 8,000 \text{ s}^{-1}$ . Comparisons are shown varying the integration window (*A*)  $T_w = 10 \text{ ms}$  and (*B*)  $T_w = 100 \text{ ms}$ . The corresponding  $S/N = \sqrt{N} = \sqrt{\bar{f}T_w} = 8.9$  and 28, respectively.

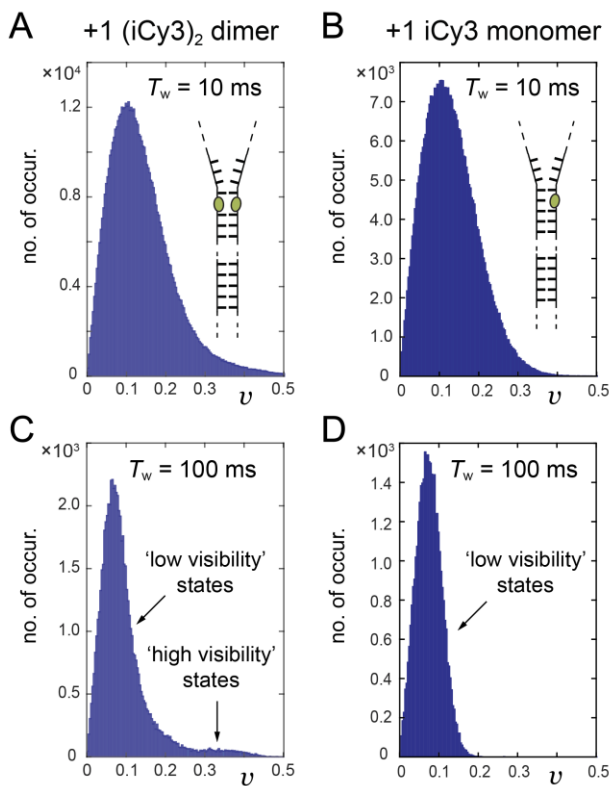
For all the samples that we studied, the visibility PDFs are asymmetrically distributed and bounded within the range  $0 < v < 0.5$ , with mean value  $\bar{v} = \sim 0.1$ . The widths of the visibility PDFs narrow with increasing  $T_w$ , revealing the presence of distinct features. The effect of increasing  $T_w$  is to average over random noise associated with the sparsely detected (photon counting) signal. However, choosing too large of an integration period can potentially lead to additional, undesirable narrowing of the PDFs by averaging over the state-to-state interconversion events that we wish to monitor and study [78]. It is therefore important to choose a value of  $T_w$  that is large enough to ensure that  $S/N \gtrsim 10$ , but small enough to retain the necessary time resolution to monitor the relevant kinetics. Since the mean flux in our experiments is typically  $\bar{f} = 8,000 \text{ s}^{-1}$ , a suitable

integration period is  $T_w = 10$  ms, which corresponds to  $S/N = \sqrt{\bar{f}T_w} \cong 8.9$ . However, we find  $T_w = 100$  ms to be useful for display purposes.

Like the visibility, the phase PDFs narrow with increasing  $T_w$ , indicating a singular value of the phase that defines the visibility over the integration time window (Fig. 2.12, right column). However, the flux PDFs do not converge to well-behaved distributions, but rather exhibit increasingly complex structure with increasing  $T_w$  (Fig. 2.12, left column). While the instantaneous flux depends on multiple environmental factors, which influence chromophore absorbance and fluorescence efficiency (e.g., probe orientation and excited electronic state dynamics), the normalized visibility depends primarily on the relative strengths and orientations of the probe EDTMs.

In Fig. 2.13, we compare the visibility PDF of the +1 (iCy3)<sub>2</sub> dimer-labeled construct to the +1 iCy3 monomer-labeled construct. For the +1 monomer, the PDF exhibits a single ‘low visibility’ feature in the range  $0 \lesssim v \lesssim 0.1$  (see Figs. 2.13 *B* and *D*), which results from the projection of the monomer EDTM onto the plane of the rotating laser polarization. This contrasts with the +1 dimer-labeled construct, which exhibits – in addition to a dominant low visibility feature – a sparsely populated ‘high visibility’ feature in the range  $0.1 \lesssim v \lesssim 0.4$  (see Figs. 2.13 *A* and *C*). The PDF for the dimer appears to resolve in multiple underlying features when the integration window is increased to  $T_w = 500$  ms (see Figs. 2.7 and 2.8). The broadly distributed features that we observe for the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs indicate the presence of both stable and thermally activated local conformations of the dimer probes. Transitions between low and high visibility macrostates are evident in the trajectories for the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct, as shown in Fig. 2.11 *C*, but not for the iCy3 monomer-labeled ss-dsDNA construct (Fig. 2.11 *F*). The above observations suggest that the (iCy3)<sub>2</sub> dimer probes

interconvert between a small number of local conformations, which likely depend on the relative stabilities and dynamics of the bases and sugar-phosphate backbones immediately adjacent to the (iCy3)<sub>2</sub> dimer probes. Furthermore, these results agree with recent ensemble studies of (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs, which concluded that only a small number of local conformational macrostates of the (iCy3)<sub>2</sub> dimer are populated under room temperature and physiological salt conditions due to steric restrictions of the surrounding nucleic acid bases and sugar-phosphate backbones [24].

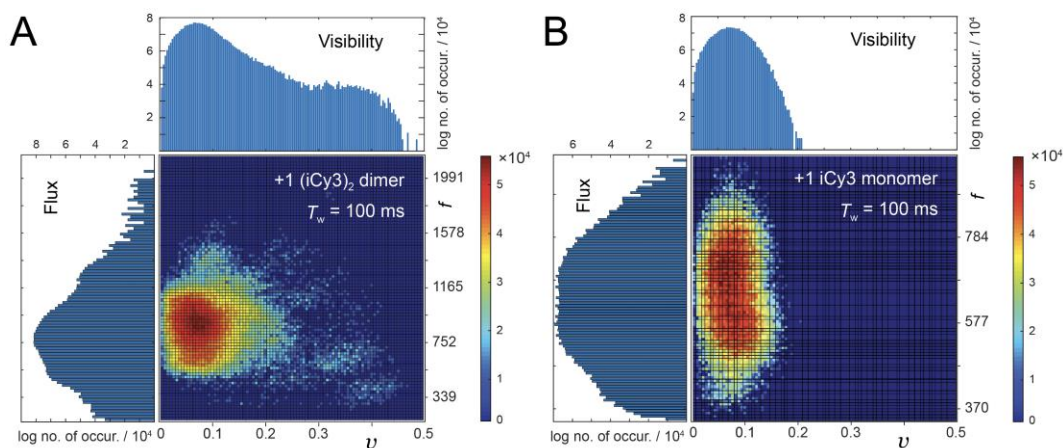


**Figure 2.13** Probability distribution functions (PDFs) of the visibility for (A, C) the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct, and (B, E) the +1 iCy3 monomer-labeled ss-dsDNA construct. (A, B)  $T_w = 10$  ms. (C, D)  $T_w = 100$  ms.

To determine whether a unique correspondence exists between the instantaneous flux and visibility, we examined the bivariate PDFs for the +1 (iCy3)<sub>2</sub> dimer- and the +1 iCy3 monomer-



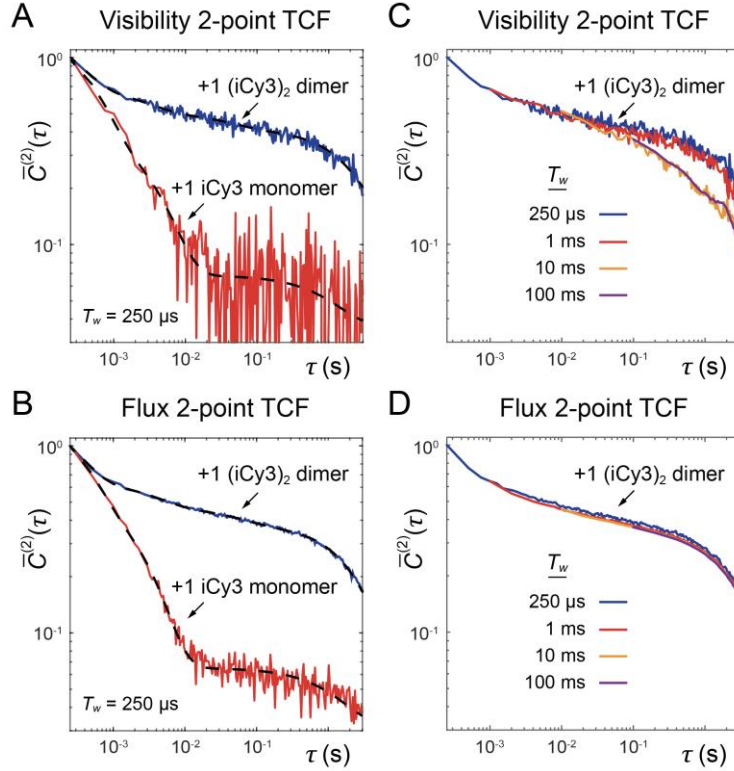
labeled ss-dsDNA constructs. In Fig. 2.14, we plot the bivariate distributions for these constructs as two-dimensional contour diagrams. The bivariate PDF for the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct (Fig. 2.14 A) shows that low visibility conformations (with  $0 \lesssim v \lesssim 0.1$ ) correspond to a broad range of flux values, while high visibility states (with  $0.1 \lesssim v \lesssim 0.4$ ) correspond to relatively low flux values. Similarly, the low visibility states of the iCy3 monomer-labeled ss-dsDNA construct (Fig. 2.14 B) correspond to a broad range of flux values. Thus, there does not appear to be a unique correspondence between flux and visibility values, which suggests that the visibility signal alone, and not the flux, contains information that can be interpreted in terms of the local conformational fluctuations of the iCy3 monomer and (iCy3)<sub>2</sub> dimer probes. This conclusion is also supported by the earlier discussion of section 3.8 which detailed the comparison of visibility and flux on the basis of the PDFs and TCFs of differing control samples.



**Figure 2.14** Joint PDFs of the flux and visibility signals (**A**) the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct and (**B**) the +1 iCy3 monomer-labeled ss-dsDNA construct using  $T_w = 100$  ms.

We next examined how the PS-SMF signals report on the dynamics of the site-specifically-labeled positions within ss-dsDNA fork constructs. In Fig. 2.15, we plot separately the two-point TCFs of the visibility and the flux for the +1 iCy3 monomer- and the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs. We present these data using  $T_w = 250 \mu\text{s}$  alongside model fits to multi-exponential functions,  $\sum_i \alpha_i \exp(-\tau/t_i)$ , where the number of decay components for the dimer is 4 and the number of decay components for the monomer is 3. The values that we obtained for the fitting parameters are given in Table 2. Although the time dependencies of the visibility and flux TCFs are qualitatively similar, we found that the decay components of the flux TCFs are generally faster than those of the visibility for all the ss-dsDNA constructs that we studied. Thus, although the dynamic processes that give rise to fluctuations of the flux and visibility signals are related to one another, they are clearly not identical.

From Figs. 2.15 *A* and 2.15 *B*, we see that the decay of the +1 iCy3 monomer-labeled ss-dsDNA construct is significantly faster than the +1 (iCy3)<sub>2</sub> dimer-labeled construct. The decay of the monomer-labeled construct is dominated by two relatively fast components ( $t_1 \lesssim 0.5 \text{ ms}$ ,  $t_2 \lesssim 4 \text{ ms}$ ), and exhibits a slowly-decaying, low amplitude, ‘baseline’ with  $t_3 \lesssim 1.2 \text{ s}$  and amplitude  $A_3 \sim 0.03$ . In comparison, the (iCy3)<sub>2</sub> dimer-labeled construct exhibits much slower relaxation behavior with four well-separated decay components:  $t_1 \sim 0.28 \text{ ms}$ ,  $t_2 \sim 2.9 \text{ ms}$ ,  $t_3 \sim 44 \text{ ms}$  and  $t_4 \sim 2.5 \text{ s}$ . We note that the slowest time scale component (on the order of seconds) is present in all our data sets. In our control measurements and previous studies [14] we attributed this slow, low-amplitude process to room vibrations.



**Figure 2.15** Two-point time correlation functions (TCFs) of (A, C) the visibility,  $\langle \delta v(\tau) \delta v(\mathbf{0}) \rangle$ , and (B, D) the flux,  $\langle \delta f(\tau) \delta f(\mathbf{0}) \rangle$ , for the +1 iCy3 monomer-labeled ss-dsDNA construct and the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct. (A) Visibility and (B) flux TCFs, respectively, using the integration time window  $T_w = 250 \mu\text{s}$ . Solid dashed curves are multi-exponential fits to the data with fitting parameters given in Table 2. (C) Visibility and (D) flux TCFs, respectively, of the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct as a function of  $T_w$ . Note that all TCFs and fits are shown normalized to the point  $\tau = 250 \mu\text{s}$ .

Our finding that the two-point TCF for the +1 iCy3 monomer-labeled ss-dsDNA construct exhibits a relatively fast decay is consistent with ensemble spectroscopic measurements [24]–[26], from which we concluded that the local environment immediately surrounding the iCy3 monomer probe is structurally disordered. We speculated that the disorder is due to misalignment of complementary Watson-Crick base pairing, which is induced by the presence of the iCy3 monomer in one strand and a thymidine spacer at the opposite position in the complementary strand. In contrast, the relatively slow decay of the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct suggests that the local environment immediately surrounding the (iCy3)<sub>2</sub> dimer probes is relatively well

ordered at room temperature and under physiological buffer salt conditions, which is also in agreement with results of our prior studies [24].

**Table 2** Multi-exponential fit parameters for the flux and visibility 2-point TCFs of the iCy3 monomer- and (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs. Note that the TCFs and fits are normalized to the point  $\tau = 250 \mu\text{s}$ .

DNA construct	$A_1$	$t_1$ ( $\mu\text{s}$ )	$A_2$	$t_2$ (ms)	$A_3$	$t_3$ (ms)	$A_4$	$t_4$ (s)
visibility two-point TCFs								
+1 (iCy3) <sub>2</sub> dimer	0.8	275	0.18	2.9	0.09	43.8	0.31	2.5
+1 iCy3 monomer	0.98	493	0.32	4.2	0.03	1,200	–	–
flux two-point TCFs								
+1 (iCy3) <sub>2</sub> dimer	0.92	265	0.17	3.9	0.09	56.7	0.32	2.7
+1 iCy3 monomer	0.99	373	0.45	3.0	0.03	1,300	–	–

We next considered the effects of varying the integration window,  $T_w$ , on the TCFs of the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct (see Figs. 2.15 C and 2.15 D). On the shortest time scale ( $0.25 \text{ ms} < \tau < 10 \text{ ms}$ ), the flux and visibility TCFs exhibit similar decays for all values of  $T_w$ . However, for larger values of  $T_w$  ( $= 10 \text{ ms}, 100 \text{ ms}$ ), the decay of the visibility TCFs at longer times ( $10 \text{ ms} < \tau < 100 \text{ ms}$ ) is faster than those using small values of  $T_w$  ( $= 250 \mu\text{s}, 1 \text{ ms}$ , see Fig. 2.15 C). In contrast, the long-time decay behavior of the flux TCFs do not vary with  $T_w$  (Fig. 2.15

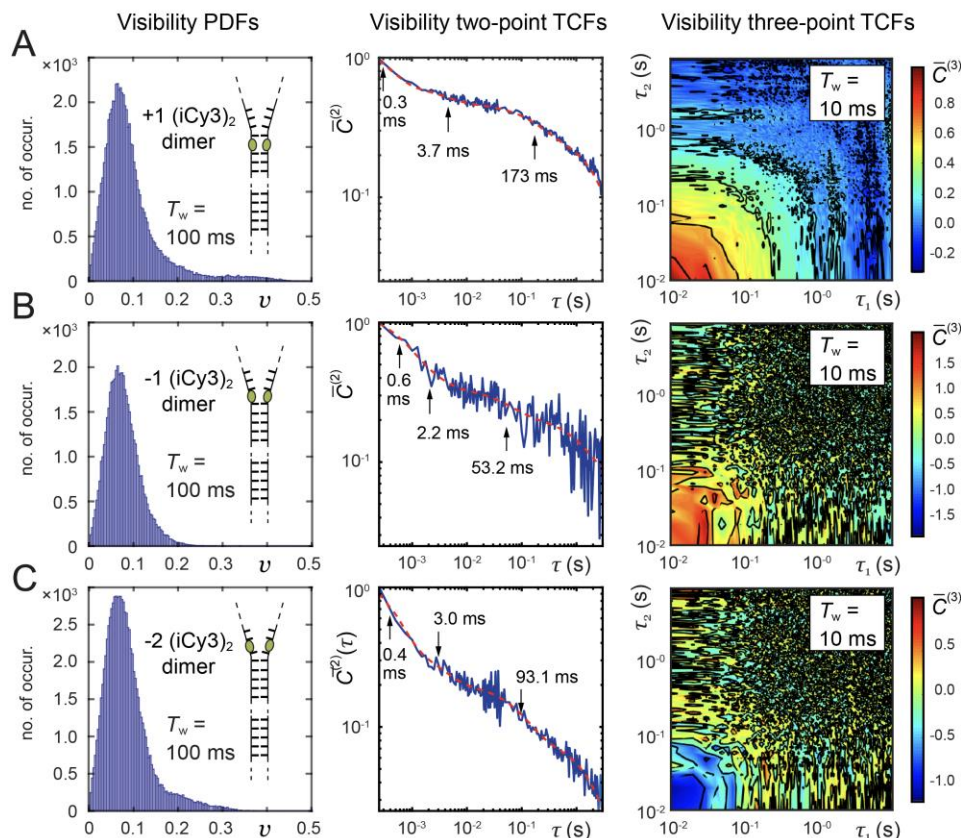
*D*). Thus, the conformational dynamics of the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs on time scales less than 10 ms appear to be reflected by fluctuations of both the flux and visibility signals, while the dynamics on time scales greater than 10 ms are most accurately reflected by fluctuations of the visibility alone. In the calculations that follow, we constructed two-point TCFs of the visibility by stitching together the decays using  $T_w = 250 \mu\text{s}$  over the range  $250 \mu\text{s} - 25 \text{ ms}$ , and using  $T_w = 10 \text{ ms}$  over the range  $25 \text{ ms} - 2.5 \text{ s}$ .

#### **Section 4.2. Studies of local DNA breathing of +1, -1 and -2 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs.**

Having established protocols for constructing experimentally-derived functions of the visibility (i.e., the PDFs, and the two-point and three-point TCFs), we next examined the sensitivity of these functions to varying (iCy3)<sub>2</sub> dimer probe position relative to the ss-dsDNA fork junction. In Figs. 2.16 *A - C*, we present results for the +1, -1 and -2 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs, which were obtained at room temperature (23°C) and physiological buffer salt conditions ([NaCl] = 100 mM, [MgCl<sub>2</sub>] = 6 mM). The left column shows the visibility PDFs,  $P(v)$ , the middle column shows the two-point TCFs,  $\bar{C}^{(2)}(\tau)$ , which are overlaid with multi-exponential fits (parameters listed in Table 3), and the right column shows contour plots of the three-point TCFs,  $\bar{C}^{(3)}(\tau_1, \tau_2)$ . Like the PDF of the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct (discussed above and shown in Fig. 2.16 *A*), the PDFs of the -1 and -2 dimer-labeled constructs exhibit a major ‘low visibility’ feature in the region  $0 \lesssim v \lesssim 0.1$ , and minor ‘high visibility’ features in the region  $0.1 \lesssim v \lesssim 0.4$ . However, the relative weights of the ‘high visibility’ features depend on (iCy3)<sub>2</sub> dimer probe labeling position relative to the ss-dsDNA fork junction with ‘high visibility’ features notably less prominent for the -1 construct in

comparison to the +1 and -2 constructs. Furthermore, the PDF of the -2 construct is significantly different than that of the +1 construct, as expected given the differential stabilities of stacked bases within dsDNA versus ssDNA regions spanning the ss-dsDNA fork junction.

In the second and third columns of Fig. 2.16, respectively, we present the two-point and three-point TCFs, which characterize the conformational dynamics of the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs as a function of probe labeling position. The two-point TCFs are shown overlaid with multi-exponential fits (dashed red curves) whose parameters are listed in Table 3. For each of the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs, the four decay components are well-separated in time:  $t_1 \sim 0.3 - 0.6$  ms,  $t_2 \sim 2 - 4$  ms,  $t_3 \sim 50 - 200$  ms and  $t_4 \sim 1 - 2$  s. As mentioned previously, the seconds-long decay  $t_4$  is due to mechanical room vibrations. We thus find that there are three relevant, well-separated decay components present in all the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs that we studied under physiological salt conditions and for most elevated and reduced salt conditions (see Table 3).



**Figure 2.16** Probability distribution functions (PDFs, left column), two-point time-correlation functions (TCFs, middle column) and three-point TCFs (left column) of the PS-SMF visibility for the (A) +1, (B) -1 and (C) -2 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs. The integration windows used for the PDFs and three-point TCFs are indicated in the insets. The two-point TCFs were stitched together over the range 250  $\mu$ s – 25 ms using  $T_w = 250 \mu$ s, and from 25 ms – 2.5 s using  $T_w = 10$  ms. The two-point TCFs are shown overlaid with multi-exponential fits (dashed red curves) whose parameters are listed in Table 3. The three-point TCFs are plotted as two-dimensional contour diagrams. Experiments were performed at room temperature (23  $^{\circ}$ C) and physiological buffer salt conditions ([NaCl] = 100 mM, [MgCl<sub>2</sub>] = 6 mM).

From the two-point TCFs of Fig. 2.16, we see that the + 1 construct exhibits the slowest overall decay (Fig. 2.16 A). As the position of the (iCy3)<sub>2</sub> dimer is varied across the ss-dsDNA junction from +1 to -1 (see Fig. 2.15 B), and again from -1 to -2 (Fig. 2.16 C), the relaxation dynamics become faster. These observations are consistent with the notion that the conformational motions of the (iCy3)<sub>2</sub> dimer probe depend on the stabilities and dynamics of the DNA bases and sugar-phosphate backbones immediately adjacent to the probe. The Watson-Crick (WC) base pairs within the duplex side of the ss-dsDNA junction are largely stacked, while stacking interactions

between bases within the ssDNA side of the junction are much weaker due to the lack of complementary base pairing. We note that the three-point TCFs,  $\bar{C}^{(3)}$ , exhibit at short times ( $10 \text{ ms} < \tau_1, \tau_2 < 100 \text{ ms}$ ) positive amplitude for the +1 and -1 constructs, and negative amplitude for the -2 construct. These observations suggest that the dominant kinetic pathways that govern conformational transitions for the +1 and -1 constructs, which undergo exchange dynamics relatively slowly, are distinct from the pathways that govern the transitions of the -2 construct, which undergoes relatively fast exchange dynamics.

**Table 3** Multi-exponential fit parameters of the visibility 2-point TCFs for the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs studied in this work. Note that the TCFs and fits are normalized to the point  $\tau = 250 \mu\text{s}$ .

construct	$\alpha_1$	$t_1$ ( $\mu\text{s}$ )	$\alpha_2$	$t_2$ (ms)	$\alpha_3$	$t_3$ (ms)	$\alpha_4$	$t_4$ (s)	$\alpha_5$	$t_5$ (s)
+1 (iCy3) <sub>2</sub> dimer, [NaCl] = 300 mM, [MgCl <sub>2</sub> ] = 6 mM	0.90	250	0.24	2.2	0.18	18	0.14	0.15	0.20	2.0
+1 (iCy3) <sub>2</sub> dimer, [NaCl] = 100 mM, [MgCl <sub>2</sub> ] = 6 mM	0.8	300	0.18	3.7	0.20	170	0.30	3.1	NA	NA
+1 (iCy3) <sub>2</sub> dimer, [NaCl] = 20 mM, [MgCl <sub>2</sub> ] = 6 mM	1.0	270	0.21	2.8	0.13	90	0.23	1.5	NA	NA
+1 (iCy3) <sub>2</sub> dimer, [NaCl] = 20 mM, [MgCl <sub>2</sub> ] = 0 mM	1.0	290	0.26	3.7	0.19	82	0.12	1.1	NA	NA
-1 (iCy3) <sub>2</sub> dimer, [NaCl] = 100 mM, [MgCl <sub>2</sub> ] = 6 mM	0.62	600	0.27	2.2	0.13	53	0.14	1.4	NA	NA
-2 (iCy3) <sub>2</sub> dimer, [NaCl] = 300 mM, [MgCl <sub>2</sub> ] = 6 mM	1.0	360	0.13	5.1	0.12	200	0.17	2.5	NA	NA



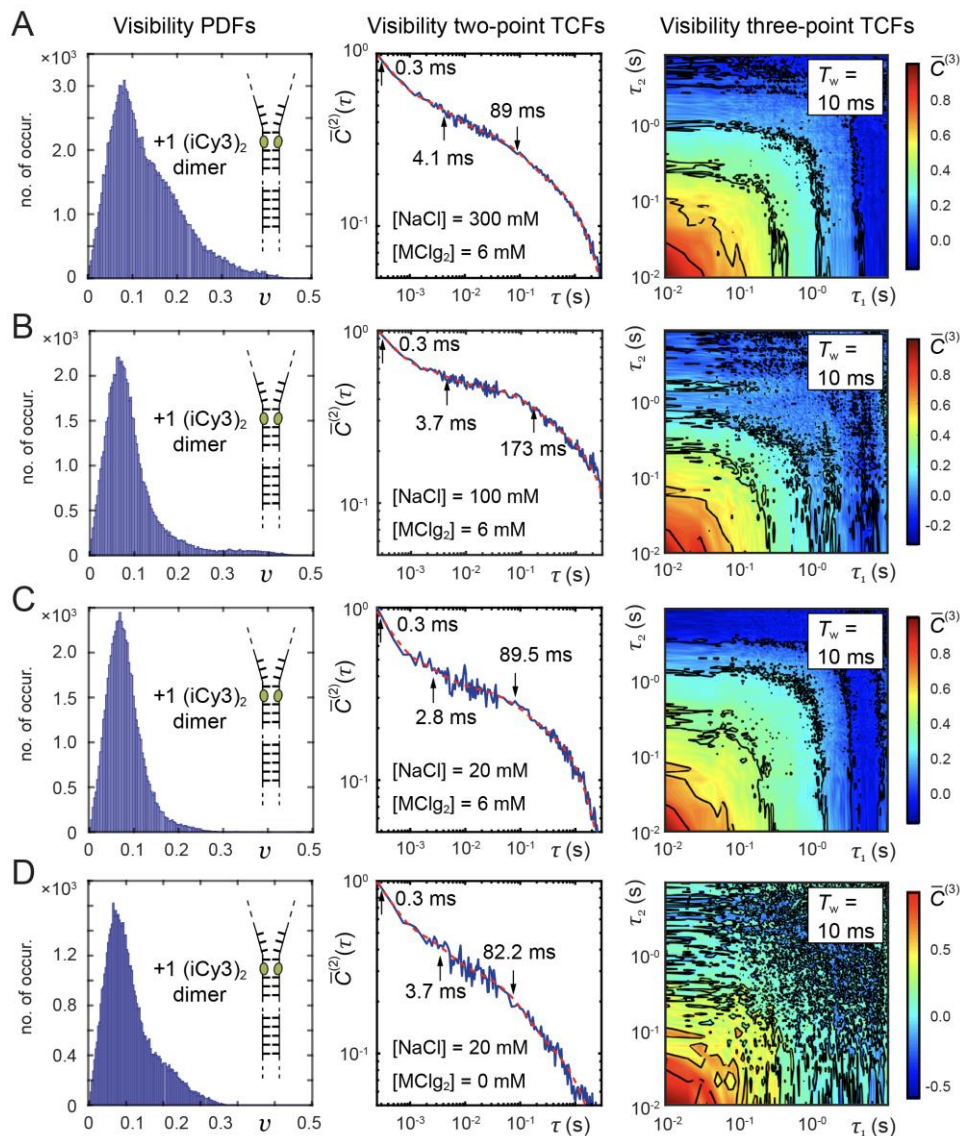
-2 (iCy3) <sub>2</sub> dimer, [NaCl] = 100 mM, [MgCl <sub>2</sub> ] = 6 mM	1.0	390	0.19	3.0	0.12	93	0.08	2.0	NA	NA
-2 (iCy3) <sub>2</sub> dimer, [NaCl] = 20 mM, [MgCl <sub>2</sub> ] = 6 mM	1.0	310	0.18	5.9	0.13	100	0.21	1.6	NA	NA
-2 (iCy3) <sub>2</sub> dimer, [NaCl] = 20 mM, [MgCl <sub>2</sub> ] = 0 mM	1.0	360	0.20	2.9	0.10	22.0	0.07	1.2	0.12	4.6
construct	$\alpha_1$	$t_1$ ( $\mu$ s)	$\alpha_2$	$t_2$ (ms)	$\alpha_3$	$t_3$ (ms)	$\alpha_4$	$t_4$ (s)	$\alpha_5$	$t_5$ (s)
+1 (iCy3) <sub>2</sub> dimer, [NaCl] = 300 mM, [MgCl <sub>2</sub> ] = 6 mM	0.90	250	0.24	2.2	0.18	18	0.14	0.15	0.20	2.0
+1 (iCy3) <sub>2</sub> dimer, [NaCl] = 100 mM, [MgCl <sub>2</sub> ] = 6 mM	0.8	300	0.18	3.7	0.20	170	0.30	3.1	NA	NA
+1 (iCy3) <sub>2</sub> dimer, [NaCl] = 20 mM, [MgCl <sub>2</sub> ] = 6 mM	1.0	270	0.21	2.8	0.13	90	0.23	1.5	NA	NA
+1 (iCy3) <sub>2</sub> dimer, [NaCl] = 20 mM, [MgCl <sub>2</sub> ] = 0 mM	1.0	290	0.26	3.7	0.19	82	0.12	1.1	NA	NA
-1 (iCy3) <sub>2</sub> dimer, [NaCl] = 100	0.62	600	0.27	2.2	0.13	53	0.14	1.4	NA	NA

mM, [MgCl <sub>2</sub> ] = 6 mM										
-2 (iCy3) <sub>2</sub> dimer, [NaCl] = 300 mM, [MgCl <sub>2</sub> ] = 6 mM	1.0	360	0.13	5.1	0.12	200	0.17	2.5	NA	NA
-2 (iCy3) <sub>2</sub> dimer, [NaCl] = 100 mM, [MgCl <sub>2</sub> ] = 6 mM	1.0	390	0.19	3.0	0.12	93	0.08	2.0	NA	NA
-2 (iCy3) <sub>2</sub> dimer, [NaCl] = 20 mM, [MgCl <sub>2</sub> ] = 6 mM	1.0	310	0.18	5.9	0.13	100	0.21	1.6	NA	NA
-2 (iCy3) <sub>2</sub> dimer, [NaCl] = 20 mM, [MgCl <sub>2</sub> ] = 0 mM	1.0	360	0.20	2.9	0.10	22.0	0.07	1.2	0.12	4.6

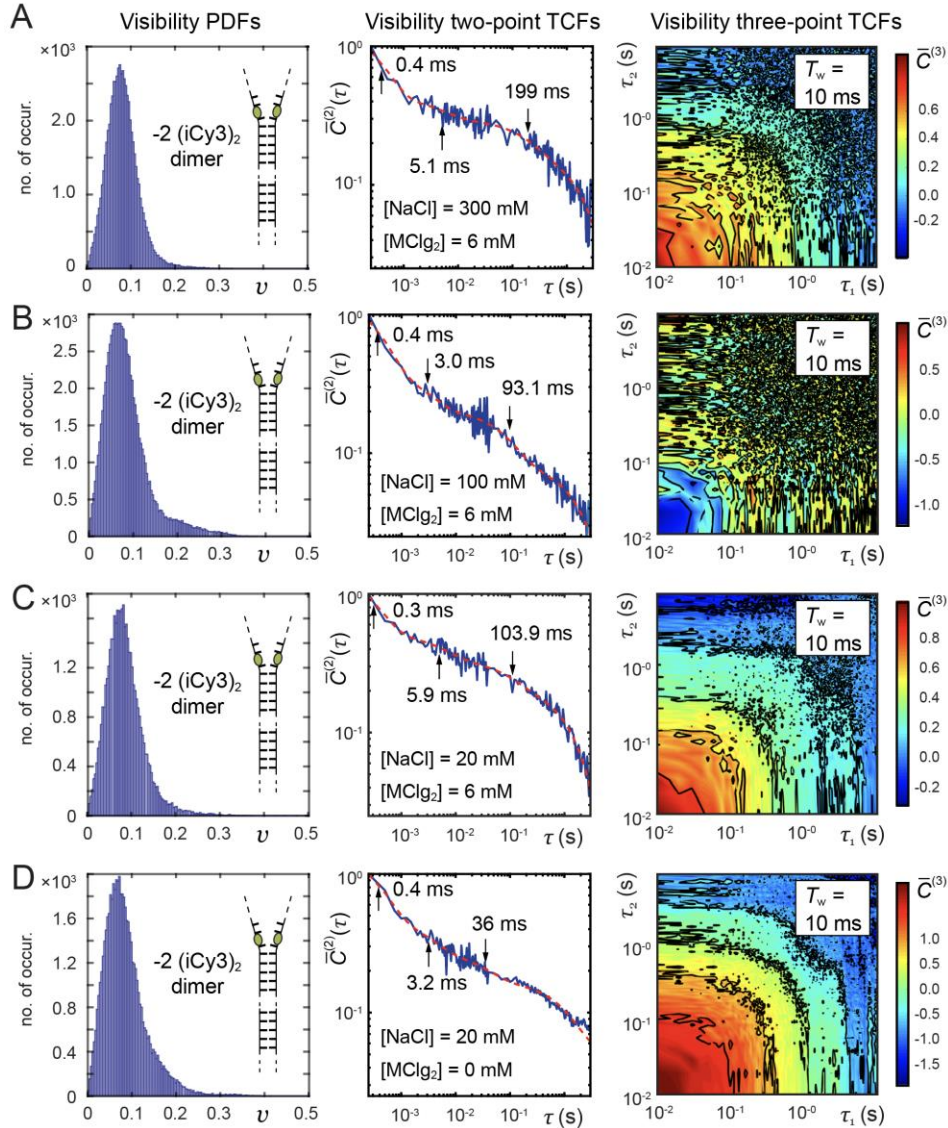
### Section 4.3. Salt concentration-dependent breathing of +1 and -2 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs.

We next consider the effects of varying buffer salt concentration on the equilibrium properties and dynamics of the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs. We examined the effects of varying salt concentration on the slowest (+1 position) and the fastest (-2 position) of the ss-dsDNA constructs (see Fig. 2.17 and Fig. 2.18, respectively). In the case of the +1 construct the dimer probe is positioned within the dsDNA region immediately adjacent to the ss-dsDNA fork

junction, while the -2 construct has the dimer probe positioned within the ssDNA region immediately adjacent to the junction.



**Figure 2.17** Salt concentration-dependent PDFs (left column), two-point TCFs (middle column) and three-point TCFs (right column) of the signal visibility for the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct. Experiments were performed at room temperature (23°C) and using (A) [NaCl] = 300 mM, [MgCl<sub>2</sub>] = 6 mM, (B) [NaCl] = 100 mM, [MgCl<sub>2</sub>] = 6 mM, (C) [NaCl] = 20 mM, [MgCl<sub>2</sub>] = 6 mM, and (D) [NaCl] = 20 mM, [MgCl<sub>2</sub>] = 0 mM.



**Figure 2.18** Salt concentration-dependent PDFs (left column), two-point TCFs (middle column) and three-point TCFs (right column) of the signal visibility for the -2 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct. Experiments were performed at room temperature (23°C) and using (A) [NaCl] = 100 mM, [MgCl<sub>2</sub>] = 6 mM, (B) [NaCl] = 20 mM, [MgCl<sub>2</sub>] = 6 mM, (C) [NaCl] = 20 mM, [MgCl<sub>2</sub>] = 0 mM, and (D) [NaCl] = 300 mM, [MgCl<sub>2</sub>] = 6 mM.

At physiological salt conditions ([NaCl] = 100 mM, [MgCl<sub>2</sub>] = 6 mM) the DNA duplex is thought to adopt primarily the Watson-Crick right-handed conformation, although thermally excited conformational macrostates may be populated transiently. The marginal stability of the DNA duplex results from a near balance between opposing thermodynamic forces (i.e., enthalpy-entropy compensation)[72]. For example, the forces that favor base-stacking are due primarily to

the positive solvent entropy change upon expulsion of the water molecules that are otherwise confined between flat base surfaces (i.e., the hydrophobic effect). However, the forces that oppose base stacking are enthalpic. These are due primarily to the strain of the sugar-phosphate backbones and electrostatic repulsion between negatively charged phosphate groups, which are 70% screened by the diffuse ‘ion cloud’ of monovalent sodium ions within the condensation layer immediately surrounding the negatively charged DNA-water interface. Decreasing salt concentration (relative to physiological conditions) destabilizes the DNA duplex by reducing the electrostatic screening between negatively charged phosphate groups [72], [73], [75]. On the other hand, increasing salt concentration above physiological conditions can also destabilize the DNA duplex. The origins of duplex destabilization at moderately elevated salt concentrations ( $> 100$  mM NaCl) are not fully understood, although this may be partly due to a reduction of the water entropy associated with the formation of structured solvation shells around increasing concentrations of ions and counterions in bulk solution. At salt concentrations much higher than physiological ( $\sim 1$  M), the duplex may be either stabilized or destabilized through Hoffmeister effects [80], [81]. Few details are currently available about how salt concentration affects the equilibrium distribution of conformational macrostates at and near ss-dsDNA fork junctions, or the transition barriers that mediate their interconversion.

In Fig. 2.17, we present the results of our salt concentration-dependent studies for the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct. These results are organized, from top row to bottom, in order of decreasing salt concentration. We note that for all the salt concentrations that we studied, the PDFs exhibit a pattern of low and high visibility features, and most two-point and three-point TCFs exhibit three well-separated decay components, like those discussed in previous sections. An exception is the TCF corresponding to the highest salt condition shown in Fig. 2.17

*A*, which exhibits four well-separated decay components (see Table 3). As described in previous work [13], [14], the presence of three well-separated decay components suggests that the simplest kinetic network model that can be used to simulate these systems involve four macrostates in thermal equilibrium.

At the highest salt concentrations that we studied ( $[\text{NaCl}] = 300 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ , Fig. 2.17 *A*), the PDF of the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct exhibits increased population of ‘high visibility’ macrostates within the range  $0.15 \lesssim v \lesssim 0.45$  in comparison to physiological salt conditions (see Fig. 2.17 *B*). Moreover, the TCFs decay significantly faster, suggesting that the transition barriers for state-to-state interconversion are reduced. When the salt concentration is decreased relative to physiological conditions to  $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$  (Fig. 2.17 *C*), the PDF exhibits a slight narrowing of the distribution of ‘low visibility’ macrostates within the range  $0 \lesssim v \lesssim 0.15$  and a reduced population of ‘high visibility’ macrostates within the range  $0.25 \lesssim v \lesssim 0.45$ . Furthermore, the TCFs exhibit faster dynamics at these low salt concentrations, suggesting that the most heavily weighted macrostates exhibit shorter lifetimes in comparison to physiological conditions. Finally, elimination of divalent magnesium at the lowest sodium concentration ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 0 \text{ mM}$ , Fig. 2.16 *D*) leads to the reappearance in the PDF of ‘high visibility’ macrostates in the range  $0.1 \lesssim v \lesssim 0.3$ , albeit with a Boltzmann-weighted distribution quite different than observed at physiological conditions. The dynamics exhibited by the TCFs are fast, suggesting that the transition barriers of the most heavily weighted macrostates are reduced relative to physiological salt conditions. In summary, the effects of raising and lowering salt concentration relative to physiological conditions for the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct are to shift the equilibrium balance of ‘high’ and ‘low visibility’

conformational macrostates to favor ‘high visibility’ states and to lower the free energy of activation for state-to-state interconversion.

We next examined the effects of varying salt concentration on the -2 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct (see Fig. 2.18). The TCF of the -2 DNA construct at physiological salt concentrations exhibits three well-separated decay components, which are significantly faster than for the +1 construct. Moreover, the PDF of the -2 construct at physiological salt concentrations exhibits significant population of ‘high visibility’ states within the range  $0.15 \lesssim v \lesssim 0.35$ , indicating that the bases and sugar-phosphate backbones immediately adjacent to the dimer probes at the -2 position adopt a different distribution of conformational macrostates than that of the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct. The relatively fast dynamics of the -2 construct indicates that the transition barriers of the most heavily weighted macrostates are relatively low compared to the +1 construct.

Interestingly, the effects of increasing and decreasing salt concentration on the -2 dimer-labeled DNA construct relative to physiological conditions are opposite to those we observed for the +1 construct. At elevated monovalent salt concentration ( $[\text{NaCl}] = 300 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ , Fig. 2.17 *A*), the PDF exhibits a narrowing of the ‘low visibility’ states within the range  $0 \lesssim v \lesssim 0.15$  and reduced population of ‘high visibility’ states (Fig. 2.17 *A*). Moreover, the relaxation dynamics are significantly slowed, suggesting that the activation barriers for state-to-state interconversion are elevated. A similar effect occurs for reduced monovalent salt concentration ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ , Fig. 2.17 *C*). Elimination of divalent magnesium at the reduced sodium concentration ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 0 \text{ mM}$ , Fig. 2.18 *D*) leads to a slight broadening of the PDF and faster relaxation dynamics, like those observed at physiological salt conditions. In addition, at this low salt concentration the TCFs exhibit four well-separated decay components

(see Table 3). We note that the primary negative amplitude feature of the three-point TCF, seen at physiological salt concentrations, becomes positive at elevated and reduced salt concentrations under conditions in which the dynamics have slowed. These observations indicate that the effects of varying salt concentration relative to physiological conditions for the -2 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct are to shift the distribution of conformational macrostates to favor ‘low visibility’ conformational macrostates and to raise the free energy of activation for state-to-state interconversion.

## 5. CONCLUSION

In this work, we have introduced a novel experimental method, called polarization-sweep single-molecule fluorescence (PS-SMF) microscopy, to study site-specific DNA ‘breathing’ fluctuations at and near ss-dsDNA fork junctions. PS-SMF uses exciton-coupled (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs to directly monitor the fluctuations of the dimer probes, which depend sensitively on the local conformations of the DNA bases and sugar-phosphate backbones immediately adjacent to the dimer probes at specific site positions.

Our results (summarized in Fig. 2.16 – Fig. 2.18) indicate that the bases and sugar-phosphate backbones sensed by the (iCy3)<sub>2</sub> dimer probes can adopt four quasi-stable local conformations whose relative stabilities and transition state barriers depend on probe labeling position relative to the ss-dsDNA fork junction. Under physiological buffer salt conditions ([NaCl] = 100 mM, [MgCl<sub>2</sub>] = 6 mM), the +1 (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA construct exhibits a relatively broad distribution of local base and backbone conformations within the duplex region of the ss-dsDNA junction (Fig. 2.16). The conformational macrostates at this position undergo the slowest dynamics of all the systems that we investigated, indicating that the most thermodynamically stable states

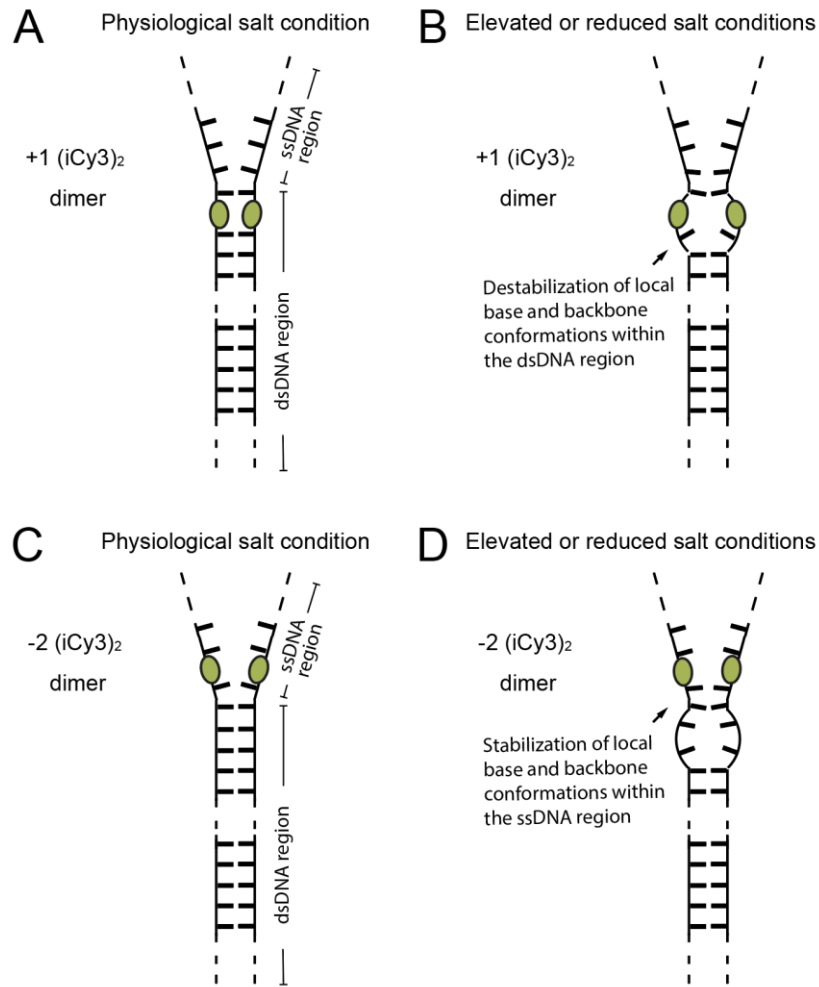


within the duplex side of the fork junction are also mechanically stable with relatively long population lifetimes. In comparison, the -2 dimer-labeled ss-dsDNA construct exhibits significantly faster dynamics and a distinctly different distribution of conformational macrostates, indicating that the thermodynamically favored conformations of bases and sugar-phosphate backbones sensed by the dimer probes within the ssDNA region of the fork junction are mechanically unstable and distinct from the duplex.

The energetics of the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs can be systematically varied by increasing or decreasing salt concentrations relative to physiological conditions. The position-dependent distribution of conformational macrostates is affected by salt concentration in complementary ways. For the +1 construct, increasing or decreasing salt concentration relative to physiological conditions shifts the equilibrium distribution to favor macrostates that are mechanically unstable (with relatively low transition barriers, see Fig. 2.17). For the -2 construct, varying salt concentration shifts the equilibrium distribution of macrostates to favor those that are mechanically stable (with relatively high transition barriers, see Fig. 2.18).

In Fig. 2.19, we illustrate a hypothetical mechanism to account for these observations. In the case of the +1 dimer-labeled ss-dsDNA construct, the local base and backbone conformations adjacent to the dimer probes are within the DNA duplex region of the fork junction (Fig. 2.19 *A*). At physiological salt conditions, the majority of conformational macrostates sensed by the dimer probe are dominated by WC base stacking, and these conformations are mechanically stable. Increasing or decreasing salt concentration leads to disruption of the local WC conformations within the duplex region and the reduction of mechanical stability (Fig. 2.19 *B*). In the case of the -2 dimer-labeled ss-dsDNA construct, the local base and backbone conformations adjacent to the dimer probes are within the ssDNA region (Fig. 2.19 *C*). At physiological salt conditions, the

majority of conformational macrostates sensed by the dimer in the ssDNA region are dominated by unstacked base conformations, which are distinct from the duplex region and mechanically unstable. Further evidence for the presence of unstacked and dynamically labile base conformations within the ssDNA region of oligo(dT)<sub>15</sub> tails of ss-dsDNA fork constructs was determined in microsecond-resolved single-molecule FRET experiments [14] and corroborated by small-angle x-ray scattering experiments [82]. At elevated or reduced salt concentration, the distribution of conformational macrostates sensed by the dimer probes within the ssDNA region of the fork junction is shifted to mechanically stable base-stacked conformations (Fig. 2.19 D). The above hypothetical mechanism is consistent with the findings of previous ensemble studies of (iCy3)<sub>2</sub> dimer labeled ss-dsDNA fork constructs. PS-SMF data, such as those presented in the current work, can be further analyzed using a kinetic network model [13], [56] to provide quantitative information about the relative stabilities of the various macrostates and their free energies of activation. The structural and kinetic properties of the ss-dsDNA fork junction revealed by these experiments can provide mechanistic insights about how replication and repair proteins that assemble at these sites carry out their biological functions. The analysis of these data with a kinetic network model is detailed in Chapters 3 and 4. The PS-SMF method monitors the signal visibility,  $v$ , from exciton-coupled (iCy3)<sub>2</sub> dimer probes that are rigidly inserted within the sugar-phosphate backbones of the ss-dsDNA fork construct. Unlike single-molecule optical experiments that solely monitor fluorescence intensity, the visibility observable is a reduced quantity that is directly related to the alignment of the monomeric subunits of the dimer probe. Therefore, PS-SMF is a useful means to measure local structure at the single-molecule level on the scale of a few Angstroms, which cannot be otherwise studied using other single-molecule optical approaches.



**Figure 2.19** Schematic diagram illustrating hypothesized mechanism of salt-induced instability of (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA fork constructs labeled at (**A**, **B**) the +1 position and (**C**, **D**) the -2 position. Left column panels (**A**, **C**) represent the physiological salt condition and right column panels (**B**, **D**) represent elevated or reduced salt conditions.

The PS-SMF experiments presented here focus on local DNA conformations at and near model ss-dsDNA fork junctions, which are relevant to the non-base-sequence specific assembly of replication and repair proteins at these sites. However, the PS-SMF approach could be applied advantageously to study the local conformations and conformational fluctuations relevant to base-sequence specific recognition of proteins that function to regulate genes in addition to the thermodynamic and kinetic effects of epigenetic modification of bases.

## CHAPTER 3 : SIMULATING A KINETIC NETWORK TO OBTAIN AN EXPERIMENTALLY DERIVED FREE ENERGY LANDSCAPE

### 1. OVERVIEW

This Chapter contains worked from the forthcoming article “*Multi-state kinetic network modeling of time resolved single-molecule data using a generalized master equation approach.*” authored by Jack Maurer, Claire Albrecht, Andrew H. Marcus and Peter von Hippel. This work was funded by the National Institutes of Health (NIGMS Grant GM-215981 to P.H.v.H. and A.H.M.). Andrew H. Marcus was the principal investigator for this work. The contents of the article have been expanded upon here to serve as a comprehensive overview and introduction to the development of computational methods which were used for the analysis of all single-molecule data presented in this thesis. Additionally, the computational methods discussed here will serve as the analytical basis for gp32 FRET studies and future PS-SMF studies.

### 2. INTRODUCTION

The methodology described in Chapter 2, sections 3.6 and 3.7, served to introduce the statistical framework used to analyze experimental data, generate the relevant data surfaces, and then form an interpretation of the observed dynamics using kinetic networks and a generalized master equation. This Chapter expands on those sections and introduces the concepts of a memoryless Markov chain and the rate matrix, with its associated eigenvectors, as the primary tools for simulating a kinetic network of interconverting macrostates. This methodology can retrieve an experimentally measured free energy landscape with only a handful of statistical data surfaces at its disposal. The use of a kinetic network to model this free energy surface makes the

underlying assumption that our experimental observable reports only on thermodynamic macrostates, with the many microstates within those energetic minima being obscured by either the time resolution of the instrument or the sensitivity of the measurement itself to small fluctuations in the signal observable. The final portion of this Chapter focuses on the implementation of our kinetic network simulations in a highly parallel optimization pipeline. This computational approach alleviates the inherent bottleneck on optimization due to the sheer number of possible networks. Our approach also allows the exploration of the parameter space to be carried out in a two-fold fashion, utilizing an initially coarse search of the parameter space with a custom built genetic algorithm followed by refinement of the initially optimized values in a more conventional steepest descent method. The nuances, strengths and weaknesses of this approach are discussed.

### 3. COMPUTATIONAL METHODS

#### Section 3.1 Simulating a kinetic network using Markov chains

The kinetic network model proposed here is a Markov chain of jumps between macrostates. The Markov nature of the chain implies memory-less system where there is no memory of previous states, only the current state and its probability of jumping to another [70]. We employ this model (Eqns. 31 and 32) to simulate our experimental TCFs (Eqns. 22 and 29).

$$\bar{C}^{(2)}(\tau) = \sum_{i,j=1}^M \delta A_j p_{ij}(\tau) \delta A_i p_i^{eq} \quad (31)$$

$$\bar{C}^{(3)}(\tau_1, \tau_2) = \sum_{i,j,k=1}^M \delta A_k p_{jk}(\tau_2) \delta A_j p_{ij}(\tau_1) \delta A_i p_i^{eq} \quad (32)$$

where  $\delta A_i$  is the fluctuation in the  $i^{th}$  observable, as defined in the previous section,  $p_i^{eq}$  is the probability of state- $i$  in the long-time limit, i.e., the equilibrium probability, and  $p_{ij}(\tau)$  is the conditional probability of the system going from state- $i$  to state- $j$  in time  $\tau$ . Each  $A_i$  is an optimization parameter, which will be further constrained by the experimental histogram.

We simulate the histogram of our observable,  $A$ , as  $P(A) = \sum_{j=1}^N p_j(A)$  where  $p_j(A) = \alpha_j \exp\left(-\frac{(A-\langle A_j \rangle)^2}{2\sigma_j^2}\right)$  where  $\alpha_j = p_j^{eq}/\sigma_j\sqrt{2\pi}$ ,  $\langle A_j \rangle$  is the mean observable value, and  $\sigma_j$  is the standard deviation of the  $j^{th}$  distribution. The area of each of these distributions corresponds to the  $p_j^{eq}$  such that  $p_j^{eq} = \int_{-\infty}^{\infty} p_j(A) dA = \alpha_j \sigma_j \sqrt{2\pi}$  and that the sum over the  $p_j^{eq}$  distributions is unity, i.e.,  $\sum_{j=1}^N p_j^{eq} = 1$ . We require that  $\mathbf{p}^{eq}$  values resulting from the histogram fits must be consistent with the  $\mathbf{p}^{eq}$  values we obtain from the TCF simulations, as discussed below. This allows us to constrain our model to three experimental data surfaces: the histogram, the two-point TCF and three-point TCF.

The conditional probabilities and equilibrium probabilities are calculated by solving the kinetic master equation,

$$\dot{\mathbf{p}}(t) = \mathbf{K} \mathbf{p}(t) \quad (33)$$

This describes the flow of probability,  $\dot{\mathbf{p}}(t)$ , at time,  $t$ , given the current probability,  $\mathbf{p}(t)$ , and the rates of interconversion that exist in the system, which are contained within the rate matrix  $\mathbf{K}$ . As written in Eqn. 33,  $\dot{\mathbf{p}}(t)$ ,  $\mathbf{p}(t)$  and  $\mathbf{K}$  are  $M \times M$  matrices, where  $M$  is the number of states in the system. The  $\dot{\mathbf{p}}(t)$  matrix contains all  $p_{ij}$  elements describing the conditional probability of going from state  $i$  to state  $j$  where  $i$  and  $j$  can be any of the  $M$  states in the system. The  $\mathbf{p}(t)$  matrix contains the current probability of the system, where the  $i^{th}$  column is the array of  $M$  probabilities

at time  $t$  given that the system started in state- $i$ . The rate matrix,  $\mathbf{K}$ , contains the microscopic rate constants,  $k_{ij}$ , which are the inverse of the time for the system to go from state  $i$  to state  $j$ , i.e.,  $k_{ij} = t_{ij}^{-1}$ , and the layout of the matrix – or where the nonzero elements lie – sets the connectivity (Eqn. 34).

$$\mathbf{K} = \begin{bmatrix} -\sum_{i \neq j=1}^M k_{1j} & k_{21} & \cdots & k_{M1} \\ k_{12} & -\sum_{i \neq j=1}^M k_{2j} & \cdots & k_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1M} & k_{2M} & \cdots & -\sum_{i \neq j=1}^M k_{Mj} \end{bmatrix} \quad (34)$$

The on-diagonal elements contain the negative sum of rates out of the  $i^{th}$  state, and the off-diagonal elements contain the rates from the  $i^{th}$  to  $j^{th}$  state where  $i$  is the column and  $j$  is the row. A properly constructed rate matrix will have each column sum to zero. We also ensure that any loops within the network satisfy detailed balance, preserving the flow of probability of the fluctuations within the system at thermal equilibrium [56], [70], [83].

The solution to the master equation (Eqn. 33) gives a matrix of all possible conditional probabilities and the long-time limit of the conditional probabilities provides the equilibrium probabilities (i.e.  $p_{ij}(\tau) \rightarrow p_i^{eq}$  as  $\tau \rightarrow \infty$ ), which correspond to each of the sub-distributions in the experimental histogram. So, to simulate our TCFs we solve the master equation for all possible initial conditions, i.e., the system is allowed to begin in any of the available states. The technique we use to solve the master equation is called a similarity transform [79]. We start by diagonalizing  $\mathbf{K}$  to obtain the eigenvalues,  $\lambda_i$ , and eigenvectors  $\mathbf{v}_i$ . Using the eigenvalues, we construct a matrix exponential,

$$[\mathbf{e}^{\lambda t}] = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_M t} \end{bmatrix}$$

The structure of the rate matrix,  $\mathbf{K}$ , requires that one eigenvalue always be zero, we let this be  $\lambda_1$ , and all other eigenvalues,  $\lambda_i$ , are negative. Then we construct a matrix,  $\mathbf{U}$ , where each column is an eigenvector of  $\mathbf{K}$ , giving  $\mathbf{U} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M]$ . Using these components, we formulate our similarity transform,

$$\mathbf{p}(t) = \mathbf{U}[e^{\lambda t}]\mathbf{U}^{-1}\mathbb{I} \quad (35)$$

where  $\mathbf{U}^{-1}$  is the inverse of  $\mathbf{U}$  and  $\mathbb{I}$  is the  $M \times M$  identity matrix, accounting for all possible initial conditions [70] The resulting  $\mathbf{p}(t)$  is an  $M \times M$  matrix of conditional probabilities, where the elements  $p_{ij}(t)$  give the probability of the system transferring from state  $i$  to state  $j$  in the time  $t$ . As we let  $t \rightarrow \infty$ , each column of  $\mathbf{p}(t)$  gives the  $M \times 1$  vector of equilibrium probabilities,  $\mathbf{p}^{eq}$ . The  $\mathbf{p}^{eq}$  will be further constrained by the experimental histogram as discussed in more detail below.

To better understand the time dependence of the solution, we can write the solution explicitly for the  $n^{th}$  column vector as,

$$\mathbf{p}_n(t) = c_1^n \mathbf{v}_1 e^{\lambda_1 t} + c_2^n \mathbf{v}_2 e^{\lambda_2 t} + \dots + c_M^n \mathbf{v}_M e^{\lambda_M t}$$

As mentioned above, the first eigenvalue of  $\mathbf{K}$  is always zero, so this expression simplifies to

$$\mathbf{p}_n(t) = \mathbf{p}^{eq} + c_2^n \mathbf{v}_2 e^{\lambda_2 t} + \dots + c_M^n \mathbf{v}_M e^{\lambda_M t} \quad (36)$$

Now that we have expressions for our conditional and equilibrium probabilities, we can plug them back into Eqn. 31 to see the time dependence of the simulated two-point TCF.

$$\bar{C}^{(2)}(\tau) = \mathcal{A}_2 e^{\lambda_2 \tau} + \mathcal{A}_3 e^{\lambda_3 \tau} + \dots + \mathcal{A}_M e^{\lambda_M \tau} \quad (37)$$

Here the  $\mathcal{A}_i$ s are the relative weights of the  $M - 1$  collective relaxation processes. This form of  $\bar{C}^{(2)}$  explicitly shows that the simulation is comprised of  $M - 1$  decay components for a model of



$M$  states. We use this as our criteria for determining the minimum number of macrostates for a given system. An important point in our analysis is that the  $M - 1$  decay components indeed the need for minimally  $M$  states to explain the data. This means that the lower limit on the number of possible conformational macrostates is set by the number of decays seen in the two-point TCF. It also implies that a robust and reproducible method is needed for evaluating the proper number of decays to represent the system. For our evaluation of the number of decays, we turn to the use of general estimates for model complexity in the form of Akaike and Bayesian information criterion.

### **Section 3.2 Estimating the number of macrostates and minimally necessary model complexity**

We construct our two-point time correlation function from experimental data, as described previously, and fit it with a sum of exponential decays, starting with one decay and we obtain the R-squared value and calculate the Akaike and Bayesian information criterion (AIC and BIC, respectively). These calculations are done as follows:

$$AIC = 2k - 2\ln(\mathcal{L}) \quad (38a)$$

$$BIC = k\ln(n) - 2\ln(\mathcal{L}) \quad (38b)$$

Where  $k$  is the number of parameters in the model,  $\ln(\mathcal{L})$  is the log-likelihood of the model and  $n$  is the number of data points. These criteria balance the complexity of the model with the goodness of fit. The interplay between the complexity of the model being used to construct a fit and the potential information content in the underlying data ensures that overfitting is avoided [84]. For the purposes of fitting the two-point TCF with a sum of exponentials, we use the model,

$$h(t) = \sum_{i=1}^H \alpha_i \exp\left(-\frac{t}{\tau_i}\right) + \beta \quad (39)$$

where  $h(t)$  is our exponential fit,  $H$  is the total number of exponentials,  $\alpha_i$  is the amplitude of the  $i^{th}$  decay,  $\tau_i$  is the decay time of the  $i^{th}$  exponential and  $\beta$  is an overall offset. This has  $2H + 1$  parameters for a model with  $H$  exponentials, which gives us our value for  $k$  in the AIC and BIC calculations. The value for  $n$  can be obtained by just counting the number of data points in the two-point TCF, and so what we are left with is calculating the log-likelihood. We maintain a consistent data point spacing and density in the TCFs to avoid issues relates to variable data point number in the AIC/BIC estimations.

The log-likelihood describes the probability of the data given a set of known parameters [84], and it is aimed at trying to determine how well the data and the fit agree. Or in other words, it aims to address how good of an explanation the model is for the data. To calculate this, we are first going to determine the squared residuals of the sum of exponentials fit to the data. We assume that for a good fit, the residuals should be normally distributed about zero. We first optimize a gaussian distribution to the residuals to estimate the value of the standard deviation. Then we can write down our likelihood function as.  $\mathcal{L}_i = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_i^2}{2\sigma^2}}$ , where  $x_i$  are each of the data points from the residual calculation ( $i \in [1, n]$ ) and  $\sigma$  is the value from the optimized fit. We calculate this for all  $n$  data points in the residual and then we take the sum over the log of the likelihood, i.e.,  $\sum_{i=1}^n \ln(\mathcal{L}_i)$ . Now we have all the pieces required to calculate the AIC and BIC (Eqn. 17a, 17b). So, we perform fits of one, two, three, etc. exponentials until we find which case is most favored based on our information criterion calculations. To compare models with the information criterion, take the minimum value of AIC (BIC) across all models sampled and subtract it from all the other AIC (BIC) values. Typically, the rule of thumb is that if the difference of AIC (BIC) is about 10 or

more, the model with the smaller value is the preferred case. An alternative, and often valuable, way of calculating AIC and BIC is to use either built-in functions within the software Origin, or another statistical software interface, that can compare two models against one another for the relative strength of model fit. This provides an additional estimate of the AIC and BIC beyond our own approach, which can differ due to the nature of the log-likelihood function that is employed for model comparison.

Another consideration that must be made is if there are any decays that are present in the data but are not due to the conformational dynamics. For example, there could be drift from the instrument or photophysical phenomena that manifests itself in the two-point TCF, which needs to be carefully characterized and removed. One way to test for these parasitic fluctuations is to design control measurements. For example, in our PS-SMF measurements we use an isotropic sample (e.g. rhodamine dye in solution) and a highly oriented sample (e.g., stretched film of a dye molecule) to identify the control timescales. If there is a decay in the TCFs of the control samples, which are selected such that they are expected to not have interesting conformational dynamics probed by the experiments, then we assign these decay timescales to instrumental effect and we do not incorporate them into our model estimate for the number of states.

The total number of states will therefore be,  $M = n_{exp} + 1 - n_c$ , where  $n_c$  are the number of timescales assigned to the instrument from the control experiments. Then when we construct our simulated TCFs we add these control time scales before comparing with the experimental data. This transforms Eqns. 31 and 32 to

$$\bar{C}^{(2)}(\tau) = \sum_{i,j=1}^M \delta A_j p_{ij}(\tau) \delta A_i p_i^{eq} + \sum_{s=1}^{n_c} \beta_s e^{-\eta_s \tau} \quad (40)$$

$$\bar{C}^{(3)}(\tau_1, \tau_2) = \sum_{i,j,k=1}^M \delta A_k p_{jk}(\tau_2) \delta A_j p_{ij}(\tau_1) \delta A_i p_i^{eq} + \sum_{s=1}^{n_c} \Gamma_s e^{-\eta_s \tau_1} e^{-\eta_s \tau_2} \quad (41)$$

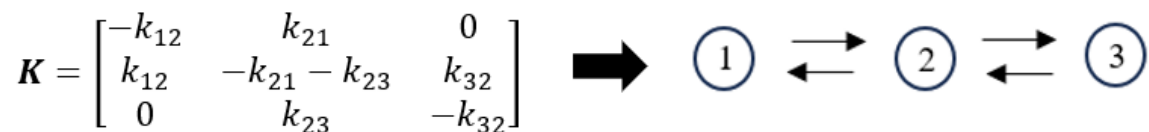
where  $n_c$  is the number of controls,  $\eta_s$  are the time scales of the control decays,  $\beta_s$  and  $\Gamma_s$  are the amplitudes of the control decays in the two-point and three-point TCF, respectively. Typically, we allow some variation in the control amplitudes and decay values when optimizing our models to account for variations due to the amplitudes coming from the decays in our kinetic network models.

With the minimally necessary number of microstates determined on a dataset-by-dataset basis, we continue on to optimize those same datasets within our kinetic network analysis. It is occasionally true that the determined number of macrostates via AIC and BIC comparisons fails to yield optimized fits using a kinetic network approach. In these rare cases, optimization attempts are halted and then restarted using one additional state beyond the originally prescribed number. Next, we will dive into the details of the various computational tools we employed to apply this framework to analyze single-molecule data.

### **Section 3.3 Optimization methods for application of a kinetic network model to experimental single-molecule data**

We have discussed the mathematics needed to simulate time-correlations functions, probability distribution functions and ultimately leverage that information for reconstruction of a free energy landscape describing the system. As previously discussed, obtaining the free energy landscape requires determination of the equilibrium probabilities and kinetic rates which interconnect all macrostates of the system. To determine these quantities using our approach, we employ a large-scale optimization of potential kinetic networks, allowing each network to be

simulated separately. In this section we discuss the fundamental components needed to define a single network and our approach to generalizing the simulation of all possible network connectivity's. Optimization of the kinetic network that best represents the experimental data necessitates that all possible network topologies be generated and tested for their ability to reproduce the three aforementioned data surfaces. The exception to this bottom-up exhaustive approach is when physical insights or logically motivated mechanisms necessitate or exclude certain transitions within the network. The fundamental setup for any network stems from Eqn. 34, which defines the rate matrix  $\mathbf{K}$ . Fig. 3.1 illustrates the relation between a single form of the rate matrix and a particular network connectivity. Where the non-zero elements of  $\mathbf{K}$  define each connection in the network.



**Figure 3.1** An explicit form of the rate matrix is depicted, with the only zero elements being those which would close the connection between state 1 and state 3 in this linear chain model. A schematized network is shown as the illustrated outcome of such a rate matrix to demonstrate the relationship between a particular form of  $\mathbf{K}$  and the resulting network of macrostates.

We considered two methods to produce all possible network topologies: a reductive based ‘top-down’ method and an additive based ‘bottom-up’ method. The reductive approach would start with a fully connected network, and stepwise eliminate connectivity between any two-states yielding a new network. The process of sequential reductions in connectivity could then be repeated until all possible networks were generated. Due to the complicated nature of properly handling microscopic reversibility in our kinetic network models [83], as detailed below, we have instead chosen to employ the additive ‘bottom-up’ method.

### Section 3.3.1 Generating kinetic network models

We begin by generating the sequence space for all permutations of  $N$  distinguishable objects with a generic permutation generating function within MATLAB. For the kinetic networks of interest, there is two-fold redundancy in the uniqueness of all permutations, since the same sequence run forward or backward is equivalent in a kinetic network approach, but distinct in the case of pure permutations. We remove these redundancies to obtain the complete set of linear connectivities, which we build upon for more complex networks.

With the set of linear models, we generate higher-ordered networks with connections between various points in the chain. The result is to obtain unique networks containing all possible loop pathways, given a base linear model. An important consideration in this process is the need to satisfy detailed-balance conditions, also known as microscopic reversibility [83]. This network dependent constraint on the microscopic rates is the primary driver behind the decision to generate higher-ordered networks from the basic linear chains, rather than start with a fully connected network and eliminate connections to generate all possible networks. Each time a connection is added to an existing network, which yields a closed loop, a single microscopic rate must be redefined in terms of a product between the forward and backward microscopic rates contained in the newly established loop.

We transform each higher order network into a Boolean mask using the criterion  $\mathbf{K} > 0$ . This maps each connectivity into a binary matrix form of zero and non-zero entries. With each newly obtained higher-ordered network, we compare the Boolean mask to the currently accumulated set of possible Boolean masks, such that each unique network appears only once in the final set. Using the same set of Boolean masks, we can later populate each  $k_{ij}$  value in the

proper  $i^{th}$  column,  $j^{th}$  row. Finally, the on-diagonal elements are calculated as the negative sums of the off-diagonal elements in each column (Eqn. 34). This generates the proper starting point for each unique kinetic network model by defining the rate matrix,  $K$ , from our generated connectivity's. The only remaining components to define a model are the observable values defining macrostates,  $A$ , and the standard deviations of those same macrostates,  $\sigma$ , for which guesses can easily be generated.

To better understand the fundamental need for a generalized approach to modeling networks, we include Table 4 below, which shows the connectivity categories and number of models for a steadily increasing number of macrostates. We note that for our present use cases, we have not employed 'branching' network models, which are models with higher connectivity than linear chain models but no closed loops. The model generating algorithm has the capability to produce such models, but their inclusion in our investigations has thus far been unwarranted.

**Table 4** Number of models, per model category, for N=4,5,6 in the course of model generation. Model categories are defined by their connectivity, with linear models having no closed loops and all higher level model categories being defined by the number of loops they contain.

<b>Number of Macrostates</b>	<b>Linear Models</b>	<b>Single Loop Models</b>	<b>Multi-loop Models</b>	<b>Total models</b>
4	12	15	7	34
5	60	188	363	611
6	360	2148	12303	14811

In the cases of M=5 the number of models becomes burdensome, and in the case of M=6, the number of total models possibles becomes intractable even with highly parallel computing capabilities. To address this issue of rapidly increasing model numbers with increasing M, we have a filtering method to remove models from the total set of simulated possibilities. The imposition of filters can be either including or excluding a particular state-to-state transition within the set of

possible networks. Table 5 shows the newly calculated number of models when two filters are included in the generation of the model space.

**Table 5** Number of models, per model category, for N=4,5,6, when filters are applied during model generation. Filters applied in most cases are based on physically intuitive connections, or lack thereof, in the set of kinetic networks.

<b>Number of Macrostates</b>	<b>Linear Models After Filtering</b>	<b>Single Loop Models After Filtering</b>	<b>Multi-loop Models After Filtering</b>	<b>Total models</b>
4	4	4	6	14
5	18	55	89	162
6	85	601	3267	3953

Clearly, the inclusion of a few physically intuitive filters, be it transitions that are strictly forbidden or highly likely to occur, can vastly reduce the model space to a level that is tractable with HPC resources. Once all models are generated, appropriately filtered, and rate matrices initialized, we can simulate the aforementioned data surfaces and compare experimentally derived results using a custom nonlinear least-squares fitting routine.

### **Section 3.3.2 Broad exploration with home-built genetic algorithm**

The first phase of the nonlinear least squares optimization uses a home built genetic algorithm to broadly search the parameter space. The inherent advantage to using a genetic algorithm at the onset of parameter refinement is to better ensure the global chi-squared minima is explored during the optimization, rather than exploring a subset of local chi-squared minima. We employ steepest descent refinement methods after an initial round of optimization in a genetic algorithm to ensure we have reached the bottom of the minima. We have found it to be true in all cases that close fits are required from the genetic algorithm for the secondary refinement steps to



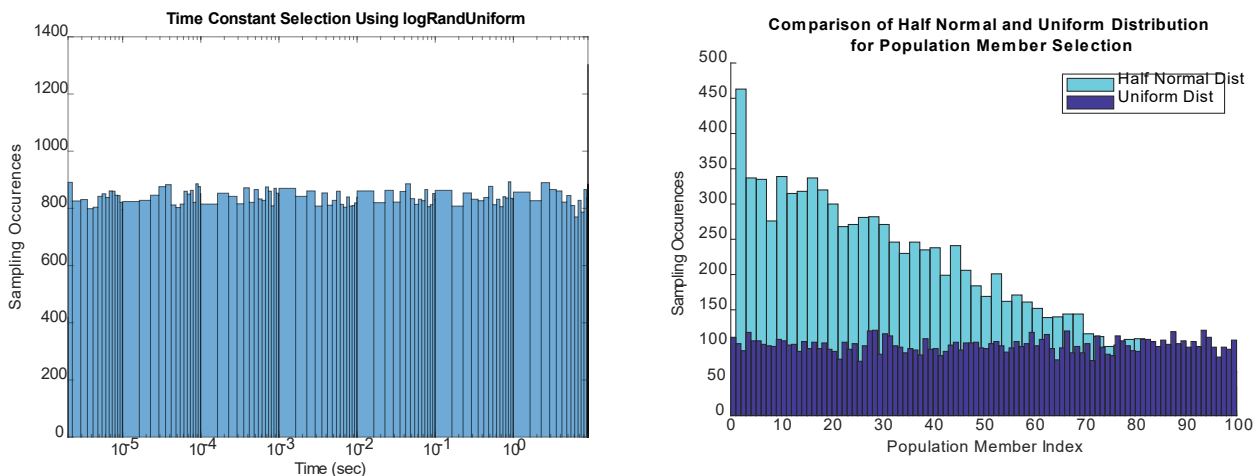
be effective, whereas poor fits or random starting points simply cannot be effectively refined using the steepest descent approach.

The genetic algorithm developed in our lab works on the same basic principles as any genetic algorithm. A set of initial guesses for the parameters are made, they are sorted by their chi-squared, then a subset of those cases are retained based on quality of fit to the data (Fig. 3.3). Those not retained are replaced by random mixtures of the retained guess parameters, then all guesses are subject to random mutations of variable magnitude. This process is repeated until the overall best outcome fails to improve after a chosen number of parameter mixing and mutation events. We have introduced a few variations and home-built methods which have led to a uniqueness compared with other more conventional approaches and have been instrumental in obtaining sufficient fits to our single-molecule data [14]. In the remainder of this section, we will discuss the newly introduced unique features.

### **New sampling methods**

The first step in the genetic algorithm is the initialization of many possible parameters within a single model. In this case, this constitutes unique guesses of each parameter that comprises a single kinetic network model. In our approach, we let the size of the initial guess pool be roughly ten times larger than the subsequent “generations” of guesses (Fig. 3.3). This is due in part to the overall increase of the parameter space size which is necessitated by moving from  $M = 3$ , as was done in [14], to  $M > 3$  states. This increase in parameter space dimension leads to a significant barrier in the combinatorics of best fit possibilities. An additional consideration on the combinatorics is the span of each parameters allowed values, which can require a highly open boundary on the inverse rate constants ( $k_{ij}^{-1}$ ) to ensure each parameter takes on its optimal value.

By opening the boundary on the inverse rate constants to span multiple decades in time, the uniformity of parameter space sampling becomes a concern. The use of standard random number generation, with a multiplier on the uniform (0,1) interval to meet the maximum allowed value on each parameter, fails to give equal probabilities for each power of ten interval of the allowed space (Fig 3.2). This is particularly problematic in the modeling efforts described here, since the inverse rate constants are given the interval of  $(10^{-6}, 10^2)$  seconds to explore. To combat the systematic oversampling of long times compared to short times in the inverse rate constants, a new sampling method was developed which ensures an equal probability and number of sample occurrences within each decade of the allowed parameter space. This method simply determines the number of decades being spanned, devises a probability distribution function that can be sampled from, and then returns a uniform number of sampling occurrences within each decade of the allowed boundary.



**Figure 3.2** Two panels compare the results of sampling two separate distributions. The first panel spanning many orders of magnitude is improperly uniform when using conventional RNG. To overcome this issue, we custom built the LogRandomUniform sampling method. The results of sampling with our method are shown in the first panel. The second panel depicts the results of selecting population members from either a half-normal or uniform distribution within the mixing and mutation phase of of the genetic algorithm.

### **Sort and select good fits**

Once all parameters are initialized and a population set of simulated outcomes is produced, the next step is evaluation of individual guess quality (Fig. 3.3). The selection process for retention versus elimination of modeled outcomes at each step of the overall algorithm typically comes down to a single choice of a cutoff percentage. In this case, the set of chi-squared sorted guesses which below the cutoff percentage line are removed, while those above the line are retained and their parameters used for mixing in the subsequent step. In a typical genetic algorithm, a uniform distribution is used to select a model outcome from the chi-squared sorted list for mixing. But to better leverage the large number of model outcomes being calculated, a half-gaussian distribution function was used for selection of model outcomes to be mixed. This leads to more rapid convergence of each iteration of the genetic algorithm, at the cost of slightly narrowing the exploration of the parameter space due to the preference for a subset of potential parameters to be mixed. In principle this approach also aims to avoid mixing poor guess outcomes into later generations, which fall very near to the cutoff percentage line.

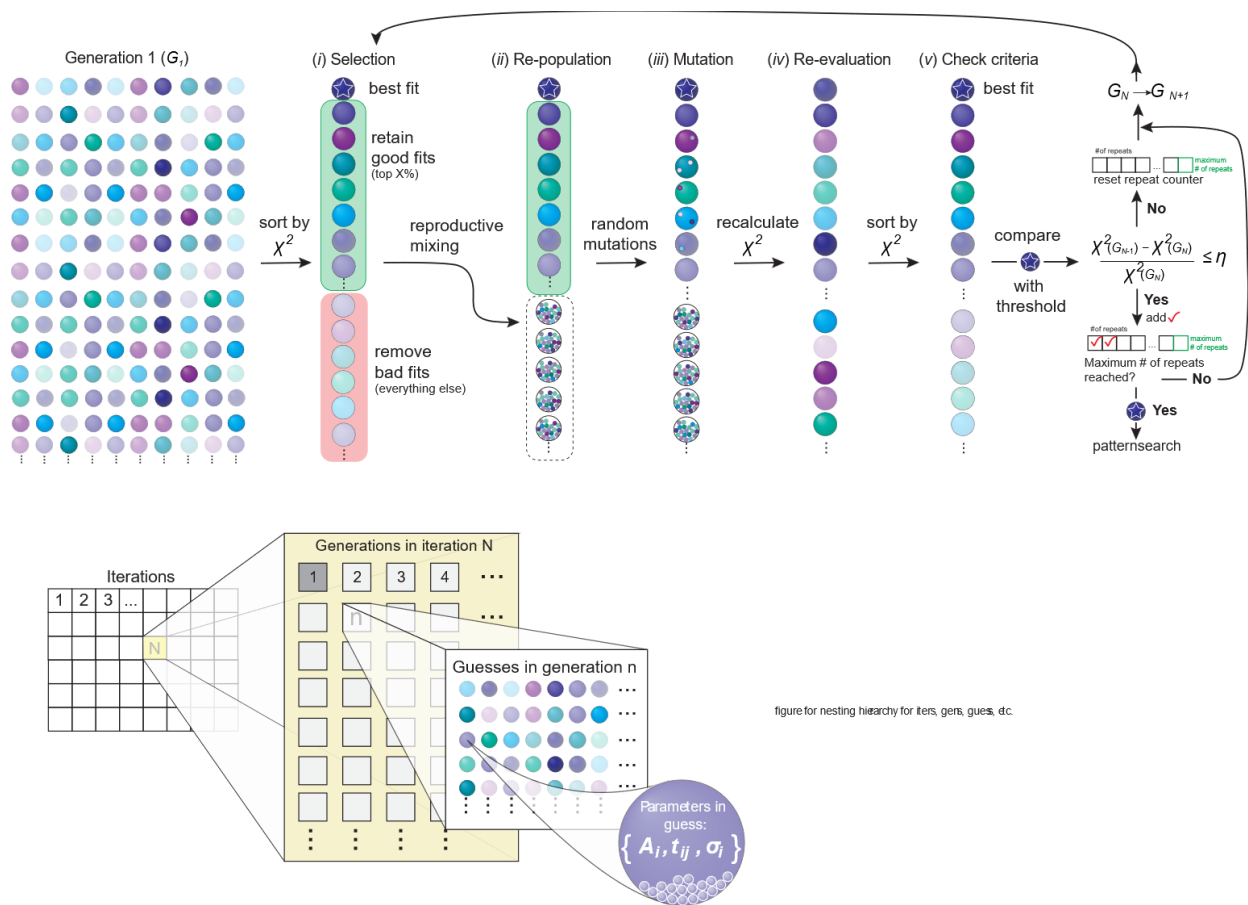
### **Mix and mutate**

Once a set of models has been chosen for retention, a small fraction is kept without alteration to the parameters, typically the top 2-4 individual outcomes. The fraction of the guess population thrown out at each iteration is regenerated with random mixtures of the parameters from all retained members of the previous iteration. This step generates entirely new guesses independent of the previous iteration's fit quality, which helps prevent trapping within a local chi-squared minima of the large combinatoric space. Once the set of models are regenerated, all parameters are probabilistically subjected to random mutations of their parameter values from the previous iteration (Fig. 3.3). Again, because the span of each parameter can be quite significant,

our logarithmically sampled random number generator is used to ensure a minimal mutation on the order of the parameter space lower bound can be achieved as well as a maximal mutation on the order of 10% the upper bound.

Lastly, all best fit outcomes after typically 35-40 repeated generations of mixing and mutations steps within a single iteration of the genetic algorithm that fail to improve the chi-squared by more than 1% are saved and sorted based on their chi-squared. Variable control over the number of repeated generations within an iteration exists within the routine. By increasing the number of repeats or decreasing the percentage-wise chi-squared improvement cutoff, iterations can be refined longer for successively better outcomes. This does come at the expense of total iterations, which limits the extent of parameter space explored by the genetic algorithm within a fixed amount of time. This allows for the best outcome across tens of thousands of guesses to be examined, while the statistics across best fit outcomes are used to motivate adjustment of the allowed parameter space.

The algorithm is run for a fixed number of iterations, typically chosen to be ~500-1000 depending on the number of states  $M$ , and degree of span within the allowed parameter space. Individual iterations are stopped when the best fit chi-squared across the total population meets the criteria discussed in the previous paragraph. These criteria currently serve as a place holder for a stricter convergence condition. However, despite the lack of a strict convergence condition, the optimized results of the genetic algorithm routinely approach a level which can be safely refined using a more traditional steepest decent method, as will be discussed below.



**Figure 3.3** A diagrammatic depiction of the genetic algorithm workflow. Starting from initialization of random guesses to form a population. Followed by simulation and model fit quality assessment. Subsequent mixing, mutation and re-evaluation steps are carried out until one possible stopping criteria is triggered.

### Section 3.3.3 Deterministic optimization with Pattern Search

The second phase of the nonlinear least squares optimization for determination of the optimal kinetic network employs a commercially available optimization routine from MATLAB known as ‘Pattern Search’. ‘Pattern Search’ utilizes a mesh grid refinement routine across all parameters allowed to float in the model under consideration. If the model under consideration has  $N$  floating parameters being optimized, then at every step of the ‘Pattern Search’ optimization an  $N$ -dimensional grid is generated about the current parameter values in both the positive and

negative directions, yielding  $2N$  total possibilities for the next best fit result. The initial starting point for the routine is the set of values found during the genetic algorithm. Each parameter receives a positive or negative shift in its value, with the magnitude being proportional to the current size of the mesh grid in all  $N$  dimensions. The mesh grid size starts off at its maximally allowed value, which is based on the allowed boundaries for each parameter. As the optimization evolves, the mesh grid size contracts to smaller values if no lower chi-squared is observed across the  $2N$  possibilities of the current refinement step. This mesh grid size contraction propagates until the result being refined reaches the level of the precision allowed by the lower bounds on all parameters, thus resulting in a plateaued minimization of the chi-squared between target data and model. Given that the mesh grid is only allowed to contract in this routine, the ability of the routine to reach a globally optimized result is highly dependent on the quality of input fit parameters determined in the genetic algorithm. Whether these input parameters correspond to the global or local minima in the chi-squared determines the success of the ‘Pattern Search’ refinement. For this reason, the working protocol is to allow the genetic algorithm roughly 500-1000 iterations of optimization prior to submitting any result to ‘Pattern Search’ for refinement. 1000 iterations are likely excessive for three state models but may not be sufficient for five or six states models given the higher dimensional parameter space. However, for four states we have observed that 500-1000 iterations of the genetic algorithm have achieved robust refinement results in most cases.

Multiple polling methods for the generation of the  $2N$  possible outcomes at each step of ‘Pattern Search’ exist. A polling method in this context is how the generation and evaluation of the  $N$  dimensional mesh grid at every step is performed. In this work, we use the ‘GPS Poll Method’, which causes the mesh grid size, and therefore degree of perturbation to each of the  $N$  parameters, to maintain its value in both the positive and negative directions at each step. Additionally, inverse

proportionality across all dimensions is maintained with the current step number, leading to the largest shifts in any parameter being made early in the refinement, while small shifts occur later. This ensures a funneling of the chi-squared to its local minima is achieved. Other polling methods were used early on in this work, namely the ‘MADS’ polling method, which is a random magnitude, random direction method that aims to avoid the deterministic trajectory of the ‘GPS’ method for a given input. However, the performance of the ‘MADS’ method was inferior to the ‘GPS’ method in nearly all cases and was not explored further for that reason.

Variable control over either the number of optimization iterations, simulation evaluations or total run time performed in a single run of ‘Pattern Search’ is used as a stopping criterion. We use the total run time of ‘Pattern Search’ as our stopping criterion given the amenability to HPC job submission run times. While convergence within the allowed number of model calculations is not guaranteed, a typical value of 50,000 model calculations, with each step of ‘Pattern Search’ making  $2N$  such calculations, is seen to be sufficient for four states. Given that the number of model calculations at each step grows with the size of the model under consideration, a value of 50,000 maximum model calculations may not be sufficient for five states and beyond. For this reason we regularly employ a 24 hour run time across all data sets for single runs of ‘Pattern Search’.

### **Section 3.4 Details of calculating the chi-squared across three surfaces**

A critical component of the optimization routines described above is how the calculation and evaluation of the best-fit chi-squared is performed. For each surface the chi-squared is calculated in a standard way,

$$\chi^2 = \frac{\sum_i (\psi_i - \bar{\psi}_i)^2}{\psi_i} \quad (42)$$

where  $\psi_i$  is the data and  $\bar{\psi}_i$  is the simulated fit. As will be discussed further below, dividing by the data is sometimes ignored, and otherwise is carefully tailored to avoid dividing by zero. The chi-squared values for each surface are then summed together into a total chi-squared value which is the parameter minimized by the optimization routines, i.e.,  $\chi_{tot} = \chi_{hist} + \chi_{C2} + \chi_{C3}$ .

Since the minimization of the chi-squared guides the optimization at every level, it is paramount to achieving an accurate representation of the data via a model fit. There are two central issues at play in simultaneously fitting the multiple statistical functions described here. The first issue is the disparate magnitudes of the data comprising these statistical functions, because the total chi-squared score across distinct models is compared based on the sum over all three surfaces individual chi-squared scores. The magnitude of the curve describing the probability distribution function is on the order of  $10^{-1}$  while the three-point time correlation function is order  $10^{-5}$ . When all statistical functions experience a similar percent error residual on the model outcome versus data, the chi-squared only reflects the residual of the highest magnitude statistical surface, while the remaining functions can have a lesser contribution by orders of magnitude. Untreated, this leads the optimization to fit only the highest magnitude statistical function without regard for the remaining functions. Therefore, a constant multiplier must be applied to the individual chi-squared of each statistical function so that their contribution to the total chi-squared is of similar magnitude given a similar degree of residual error ( $\alpha_{hist}, \alpha_{C2}, \alpha_{C3}$ ).

$$\chi_{tot} = \alpha_{hist} \chi_{hist} + \alpha_{C2} \chi_{C2} + \alpha_{C3} \chi_{C3} \quad (43)$$

The second issue that arises in the fitting of these statistical functions is the density of points used to represent the time correlation functions across the individual surfaces. As was



discussed in [14] a logarithmic sampling of  $\tau$  in both the two and three-point correlation functions is used to achieve enough data points for a robust representation, without calculating every possible  $\tau$  product available at a particular data binning resolution. This avoids an overly dense representations of the two and three-point TCF at late  $\tau$ , which greatly increases calculation time from raw data and simulation time during modeling. The simple result of this approach is that the first decade of  $\tau$  contains very few points compared to the final few decades (Fig. 3.2). It is worth noting that this problem would not only persist but be amplified if  $\tau$  were to be exhaustively sampled at the data binning resolution. The calculation of chi-squared performed for each statistical function is an average over all squared residuals between data and model. Given the much greater number of points in the average from long values of  $\tau$  compared to short values, a simple average is dominated by the long  $\tau$  values. To compensate for this density issue, we apply an artificial weighing function to the square of the residuals prior to taking the average. These arbitrary weight functions, ( $\eta_\beta$  where  $\beta$ =histogram, C2 or C3) or mask, are applied to each element of the  $\chi_{tot}$  calculation (see eq) and is calculated assuming a steadily increasing noise profile of the data. This originates from both the decaying dynamical correlation in the data as well as the loss of pair-wise products which contribute at late  $\tau$  compared to early  $\tau$ , leading to a statistically less accurate measure of the correlation functions at late  $\tau$ .

$$\chi_{tot} = \alpha_{hist}\chi_{hist}\eta_{hist} + \alpha_{C2}\chi_{C2}\eta_{C2} + \alpha_{C3}\chi_{C3}\eta_{C3} \quad (44)$$

To best capture the essential features of each unique data set, a home-built chi-squared weight function is created to address the subtle differences between inputs being optimized. The basic elements of the chi-squared weight function are inputs of a constant noise floor  $\gamma_0$ , a variable noise profile  $\gamma(\tau)$  across the various decades of  $\tau$ , as well as a choice in the discretization  $N_{segments}$  of the chi-squared surface to address data point density. The underlying assumption of

the noise mask  $\varepsilon(\tau)$  is that the noise profile of the correlation function surfaces strictly increases and that the idealized weight function applied to the chi-squared residuals would result in equal contributions from each decade, or segment of a decade, across all  $\tau$ . This is in opposition to a blind application of a squared residual error function, which is dominated by the latest decades of the correlation functions. The complete noise mask is simply defined by the sum:

$$\varepsilon(\tau) = \sum_{i=1}^{N_{segments}} \gamma_i(\tau) + \gamma_0 \quad (45)$$

The weight surface for the correlation functions is then defined by applying the noise mask to the underlying data and establishing equal contributions from every segment to the pure chi-squared. This is done by multiplying each segment's chi-squared contribution in the presence of the noise mask by an appropriate scalar  $S$  to meet the magnitude of the maximally contributing segment.

$$\eta_{C2/C3} = \sum_{j=1}^{N_{segments}} \frac{(\psi_j - \varepsilon_j \psi_j)^2}{\psi_i} S_j \quad (46)$$

The result of this approach is to obtain unique weight functions for every data set, which only rely on steadily increasing noise and differences in data point density. For the case of the histogram, the weight function is simplified as the inverse of the data surface itself:

$$\eta_{Hist} = \frac{1}{PDF} \quad (47)$$

As is often the case, one set of parameters does not work for every data set. Therefore, the noise floor, noise profile and discretization of the calculated functions are left as inputs to be adjusted for on a data set-by-data set basis. Additionally, cutoffs in the correlation times are established for all correlation functions to avoid dividing the error residuals by zero, as commonly occurs in the

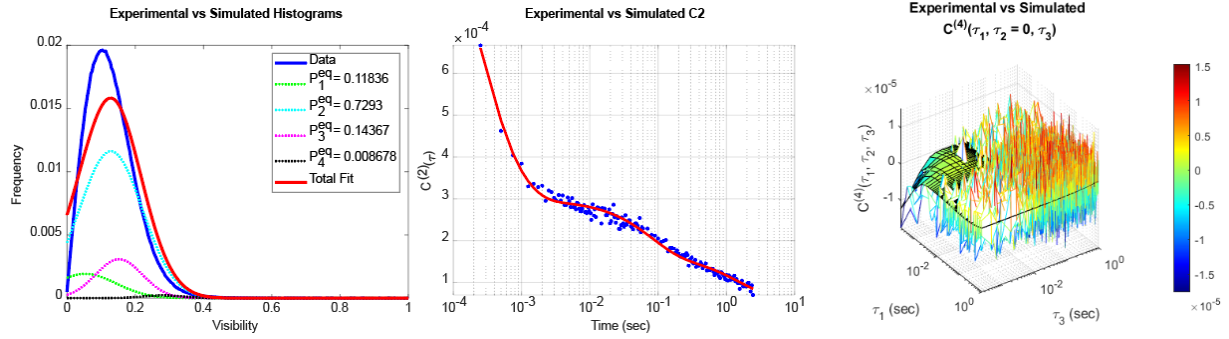
data at the longest values of  $\tau$ . A typical set of values is regularly used at the onset of the optimization process, with adjustments being made once the optimization reaches a certain threshold in the balance of statistical contributions from all three data surfaces. It should be noted that the interplay between these inputs does allow for the ability to tailor the weight functions given a particularly problematic or unique data set in a manner which was not possible for the simpler and more conventional  $\tau^{-n}$  approach. The resulting weight functions and artificial noise profile are then applied to the data to determine an average chi-squared contribution across all three surfaces. This determines the starting point for the scalar multipliers to be used in the optimization. Further adjustment of scalar multipliers by the user is typically necessary ( $\alpha_i$ 's in Eq. 44). However, this approach further reduces computational run time, as the number of unique optimization trials to achieve proper scalar multipliers is dramatically reduced.

The final statistical tool we employ to guide the optimization of our kinetic networks via updates to the weight function parameters is a graphical interface which visualizes the error from multiple perspectives. The perspectives are the perceived error from the use of the weight functions, the true error from both the present error and raw residual, as well as the form of the weight functions themselves. The combination of these statistical data surfaces offers a complete view of where the optimization is missing its target outcome, how that target outcome can be better configured and lastly whether the interpretation of the unweighted error is improving or worsening via alterations to the weight function parameters. Taken in totality, the outputs of this user interface are a convenient and concise way of motivating updates to the optimization across all data sets. The details of this approach are discussed below in section 3.4.1.

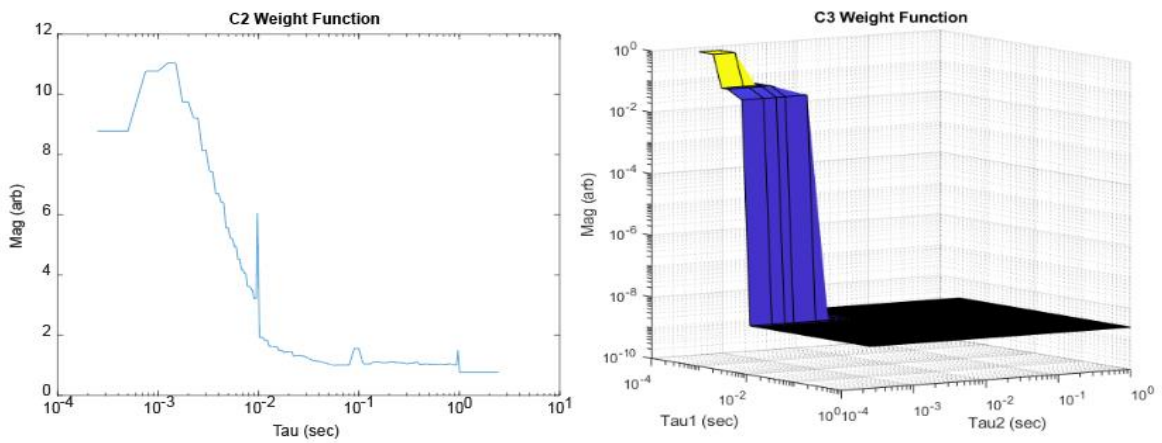
### Section 3.4.1 Quality of fit assessment via Global Chi-squared Statistics

In order to assess the quality of fits and deal with the pitfalls of simultaneously fitting three data surfaces to a single model, a modeled chi-squared weight function is generated for every data set using a noise profile based on the experimental data. We refer to this approach as a ‘weights calculator’, where the tabulation of the chi-squared outcome for each model within a data set is optimized based on a constructed interpretation of the residuals rather than a pure interpretation. However, the pure residuals and percent error between model and data are still tabulated and tracked at the initial steps of refinement to help guide the proper selection of ‘weights calculator’ parameters. To better elucidate this process the outcome of calculating the chi-squared using both a ‘weights calculator’ and raw error interpretation is performed for both a high quality and lower quality fit to the same data. The visualization of this approach for both cases is shown in Figs. 3.4-3.11. The actual weight functions used for these examples are shown in Figs. 3.5 and 3.9. Notably, these surfaces differ from any smooth analytical function which might otherwise be employed for the purpose of weighting the chi-squared.

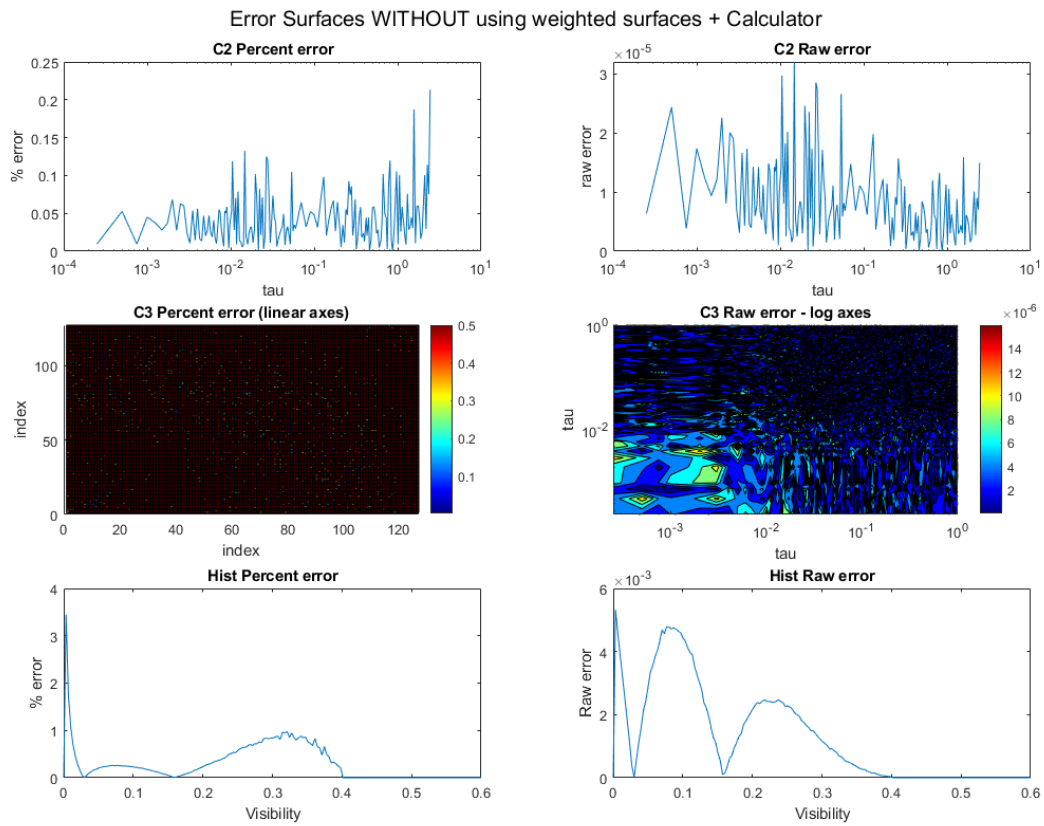
In Fig. 3.4, a high-quality fit to experimental single molecule data can be seen. The resulting ‘weights calculator’ interpretation of this data (Fig. 3.7) and the raw error interpretation (Fig. 3.6) of the data are in relative agreement, with both displaying low values of their respective error functions in all regions where the fit to data is good and the data itself contains little noise. Visual inspection of the actual fits to the data shows a convincingly close representation of the underlying dynamics and conformational states which yielded the experimental results.



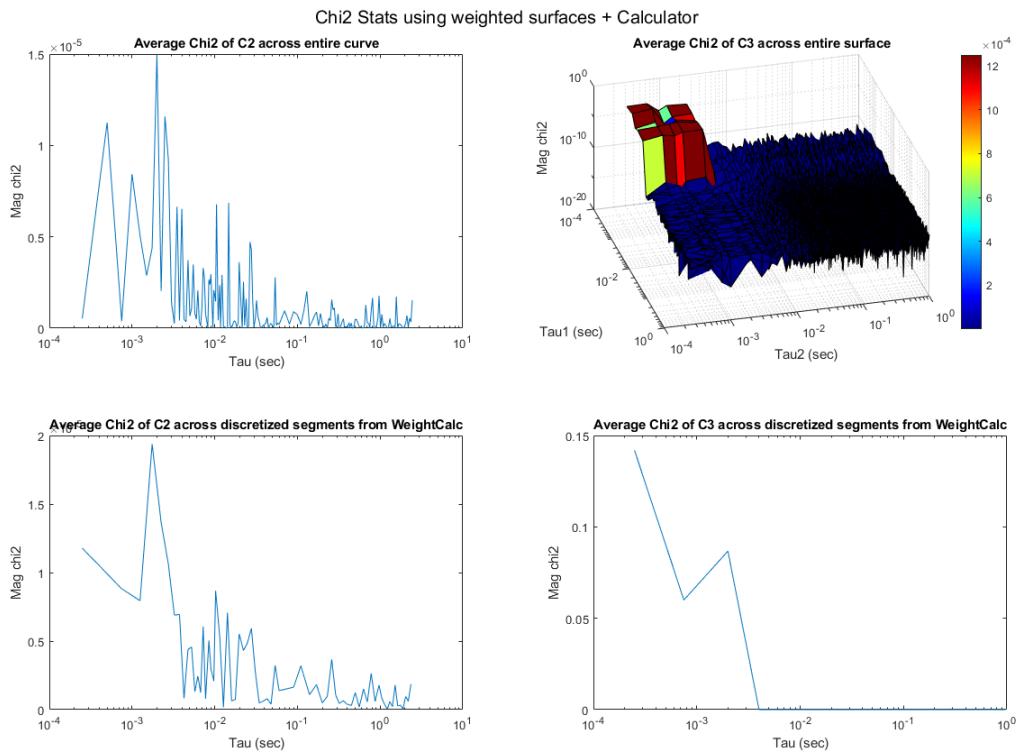
**Figure 3.4** Example of a ‘good’ fit to the data from a model outcome. The PDF, C2 and C3 are shown against their respective model outcomes. The residuals are, on average, near their minimum.



**Figure 3.5** Weight surfaces applied to the C2 and C3 respectively. These surfaces are applied to the residuals of the data versus the model prior to calculating the chi-squared (see Eq. 44).



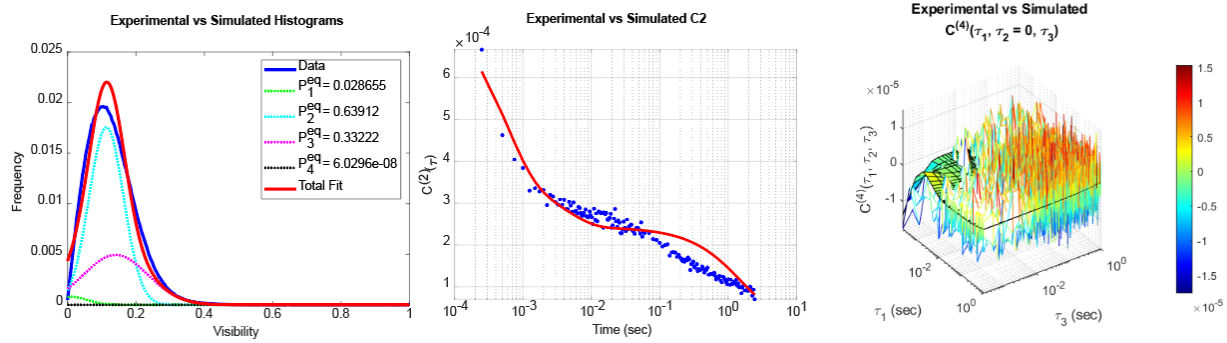
**Figure 3.6** ‘Chi-Stats’ overview for the panels that report on statistical measures of ‘true’ error (no weight function applied). The case shown here is for the high-quality fit of Fig 3.4.



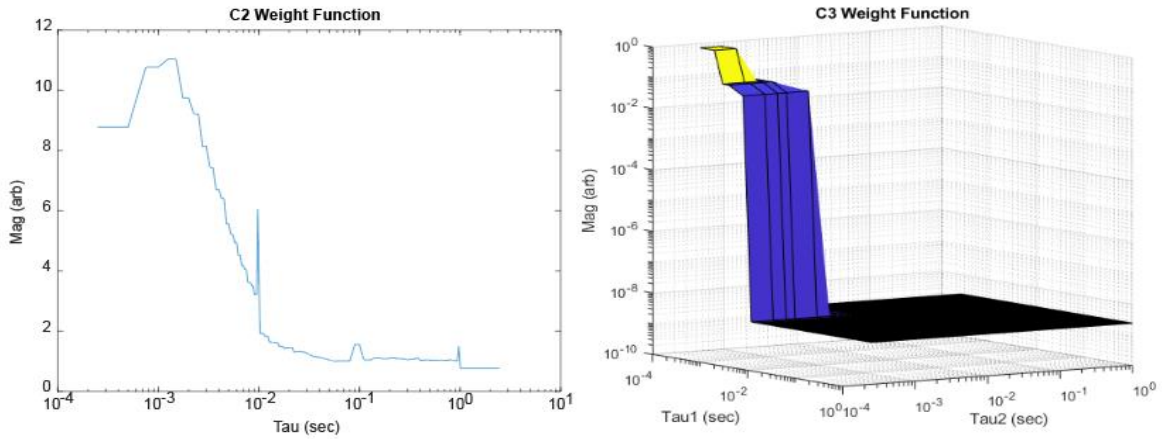
**Figure 3.7** ‘Chi-Stats’ overview for the panels and statistical measures of the ‘weighted’ error via the application of weight functions to the data residuals. The case shown here is for the high-quality fit of Fig. 3.4.

In Figs. 3.8-3.11, a lower quality fit to the same experimental data is shown, using the same ‘weights calculator’ parameters to inspect the chi-squared and error. The raw error in this case is much greater than before and this is captured by the ‘weights calculators’ as a greater magnitude in the interpretation of chi-squared from both the two and three-point correlation functions, in the regions where the model deviates most significantly from the data. Notably, some regions of the ‘weights calculators’ surfaces are greater in magnitude than the raw error would suggest is necessary. This is due to the issue of data point density and dilution of the statistical average chi-squared that was mentioned previously. Ultimately, these panels of interpreted versus actual

residual error are used to guide the attenuation of parameters used to construct the ‘weights calculator’ surfaces.

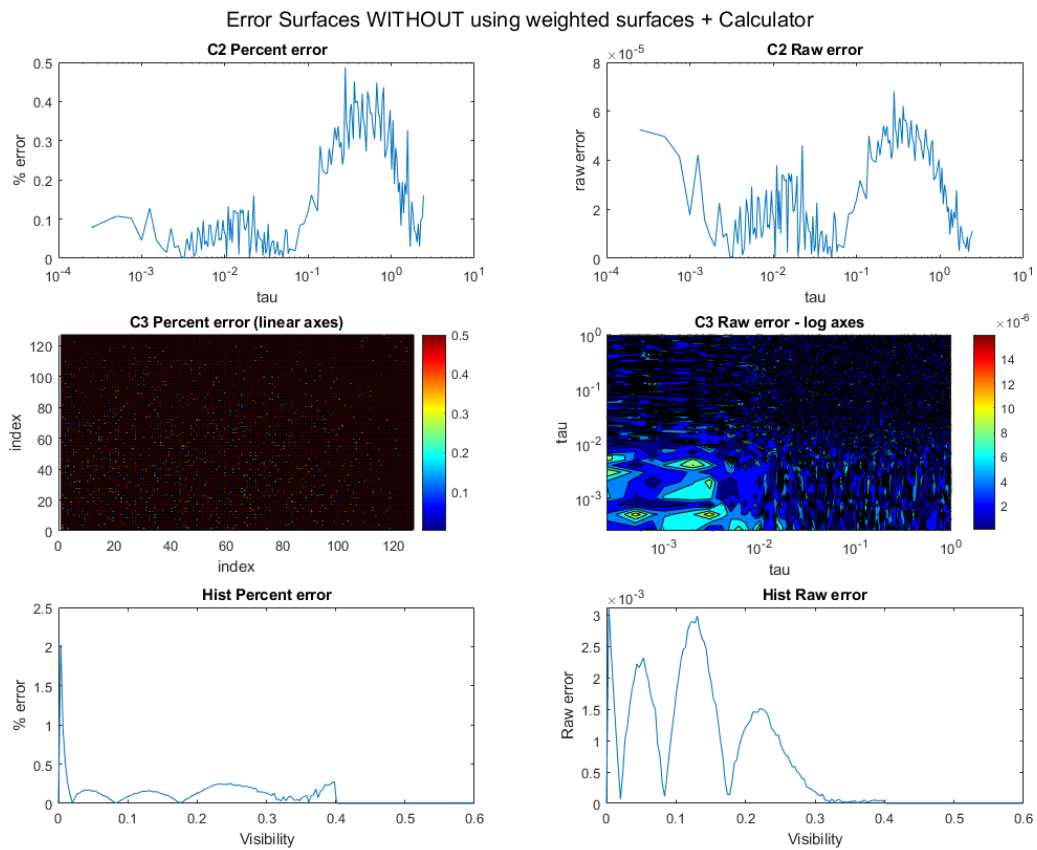


**Figure 3.8** Example of a ‘bad’ fit to the data from a model outcome. The PDF, C2 and C3 are shown against their respective model outcomes. The residuals are, on average, not near their minimum.

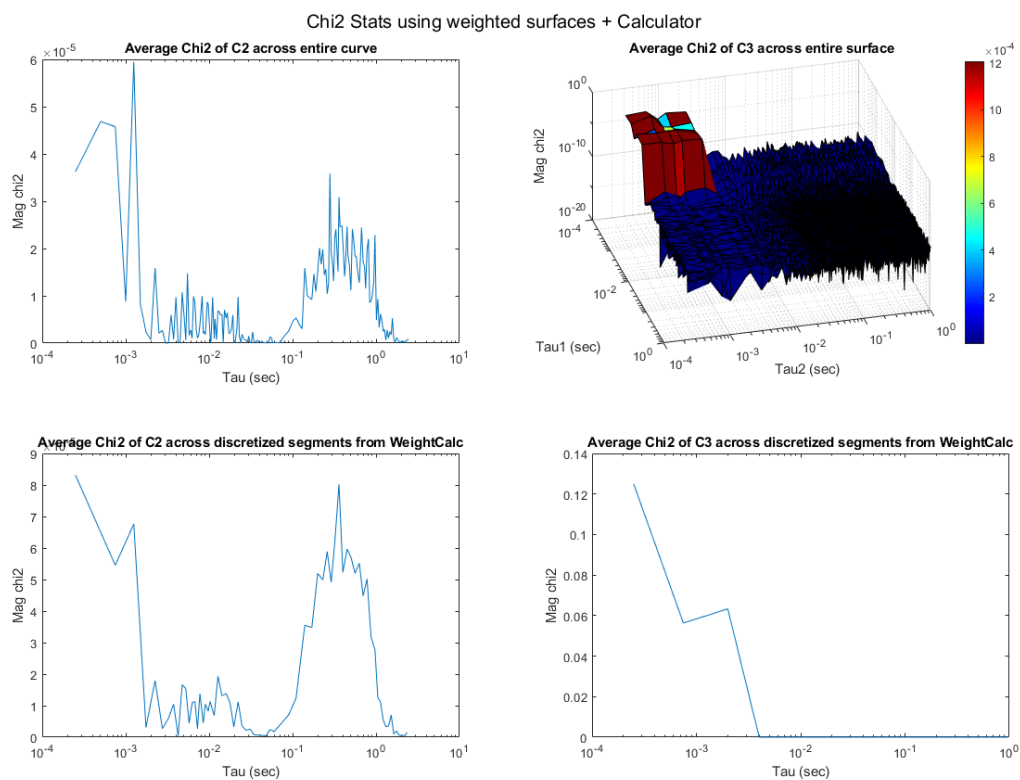


**Figure 3.9** Weight surfaces applied to the C2 and C3 respectively. These surfaces are applied to the residuals of the data versus the model prior to calculating the chi-squared (see Eq. 44). The surfaces shown here are notably not the same as those of Fig. 3.5, leading to the discrepancy in observed fit quality between the two examples.





**Figure 3.10** ‘Chi-Stats’ overview for the panels that report on statistical measures of ‘true’ error (no weight function applied). The case shown here is for the low-quality fit of Fig 3.8.



**Figure 3.11** ‘Chi-Stats’ overview for the panels and statistical measures of the ‘weighted’ error via the application of weight functions to the data residuals. The case shown here is for the low-quality fit of Fig. 3.8.

### Section 3.4.2 Optimizing models in parallel on a high-performance computing cluster

Traditionally, we have used physically motivated intuition to guess the proper network, or a subset of probable networks then proceed with a limited number of optimizations on a handful of local PCs [14]. While this approach certainly reduces the need for large scaling computing, it does leave out many possibilities which may appear unintuitive but represent the data better than physically motivated expectations. With the ability to arbitrarily generate all network models for  $M$  states within a class of network topologies, as discussed in section 3.3.1, the sheer number of possible models requiring optimization quickly becomes unwieldy. Here, we used a local high performance computing cluster (HPC) to scale the optimization of the best fit kinetic networks

across a dozen or more data sets simultaneously. The limitation on the number of models being run simultaneously is simply a matter of hardware availability and utilization, as is the case with any shared computing resource. With the use of the HPC, the entire allowable model space can be evaluated at a sufficient level of sampling within a few days for the genetic algorithm. The underlying MATLAB optimization routines at both the genetic algorithm and pattern search phases were wrapped into bash routines to automate the submission of jobs as well as automate the deployment of all possible data sets onto the HPC simultaneously, achieving optimization of all models in parallel with one another.

The constraints on typical laboratory computing resources are of course primarily driven by the sheer size of the optimization, which typically limits optimization of a single model to a single CPU core. Random access memory requirements also limit the total number of models capable of being optimized on a single PC, irrespective of the clock rate and number of independent cores available on the CPU. No GPU based computing was entertained during this work, although in principle could aid the deployment of more models simultaneously onto local computing resources or greater parallelization within a single model, possibly reducing the need for the HPC.

For more than five states the number of models grows so substantially that even typical HPC resources cannot handle simultaneous optimization of all possible models, unless unduly long periods of time are tolerable to achieve results. In these cases, it is beneficial to introduce a filter onto the entire model space, which is motivated by simple physical considerations of the system under study, as was discussed in section IIIA.

### Section 3.4.3 Construction of ‘weights calculator’ surfaces

The approach taken in this work to handle the simultaneous fitting of the three data surfaces is to construct a separate three surfaces which individually weight the chi-squared contribution from each segment of data across the entirety of each surface. The basic premise of the ‘weights calculator’ is to model the noise profile of the data and then calculate the anticipated chi-squared contribution from each data segment. In order to do this, a few parameters must be defined to construct the model noise profile. These are called the ‘noise floor’, ‘noise ramp’ and ‘number of data segments.’ Each parameter will be discussed and defined below.

A data segment is defined by a short section of consecutive data points from either the two- or three-point correlation function over some discrete length of independent variable  $\tau$ . Segments can be as coarse as entire decades in  $\tau$  or as fine as individual data points. Coarsely segmenting the data alleviates effects from outliers in producing unusual contributions to the final weight surface. However, averaging together noise contributions using coarse segments results in some loss of sensitivity to the nuances of each underlying data surface and its unique contours.

We have composed the noise profile with two main elements: a ‘noise floor’ and a ‘noise ramp.’ The ‘noise floor’ is defined as the baseline noise contribution to the entire underlying data surface. This is a simple percentage of the data surface included in the noise mask applied to the data. The ‘noise ramp’ is a linearly increasing noise contribution as a function of  $\tau$ , which allows for the latest points in the correlation functions to have a greater percentage wise error than earlier points. This gain in noise with increasing  $\tau$  is expected based on the signal to noise in these statistical functions growing with the total number of valid pair or triple products. As well as the loss of correlation at late  $\tau$ .

Altering the ‘noise floor’ or ‘noise ramp’ has a dramatic effect on the resulting weight functions. The basic principle behind alterations to these parameters is to better weight segments of the data surfaces that are underfit relative to other data segments which are overfit. The interpretation of the relative under or overweighting of certain data segments is made through examination of the underlying fits to the data as well as inspection of the chi-squared results through the lens of both the ‘weights calculator’ and raw error representations.

The approach taken to produce high quality optimized fits to single molecule data is to first run the genetic algorithm, use the statistical tools described previously to guide a balanced optimization across all three data surfaces and then allow the optimization to evolve for some time. Once reasonably high-quality fits are achieved in the genetic algorithm, the results serve as starting points for refinement in ‘Pattern Search’. Once again, outcomes of ‘Pattern Search’ are subjected to the same statistical examination, with balance changes or alterations to the weight surfaces being motivated by the combination of the statistics and visual inspection of the optimized fits.

#### 4. CONCLUSION

We have developed an analytical and computational framework for extraction of kinetic network parameters from single molecule measurements, independent of system observable. This information can be leveraged to construct the free energy landscape governing system dynamics and offers a mechanistic view of the conformational changes occurring within macromolecular assemblies. The approach is generalizable to a wide variety of experimental systems, after considering both the relevant control experiments to identify sources of noise as well as the degree to which equilibrium has been established during periods of data acquisition. The ability to analyze data in a parallel fashion opens the possibility of leveraging large data set statistics and best fit

parameters to tailor the optimization of a large combinatoric space toward achieving a sensible and robust result. The ability to implement parallel optimizations hinges on the availability of computational resources. But for systems where physically intuitive constraints can be placed on the interconnectivity of the underlying kinetic network, the demand for computational resources can be significantly reduced. In cases where the system is well understood from a conformational macrostate perspective, but no current methodology affords the assignment of microscopic kinetic rate constants, our framework presents the opportunity to uncover a richer picture of the underlying dynamics in such systems. We plan to apply this analysis framework to problems pertaining to the assembly and function of protein-DNA complexes, where the unique assignment of free energy landscape parameters can elucidate the biochemical mechanisms driving the core function of these macromolecular machines.

## CHAPTER 4 : ANALYSIS OF PS-SMF DATA FROM SS-DSDNA JUNCTION USING A KINETIC NETWORK MODEL

### 1. OVERVIEW

This Chapter contains worked from the article “*DNA ‘Breathing’ free energy landscape determination via kinetic network analysis of polarization-sweep single-molecule fluorescence microscopy data*” authored by Jack Maurer, Claire Albrecht, Andrew H. Marcus and Peter von Hippel. This work was funded by the National Institutes of Health (NIGMS Grant GM-215981 to P.H.v.H. and A.H.M.). Andrew H. Marcus was the principal investigator for this work. The contents of the article have been expanded upon here to serve as a comprehensive overview and introduction to the methods which will be used throughout the remaining Chapters. This Chapter will more formally analyze the results of ‘polarization-sweep’ experiments. A series of structural inferences are made based on both ensemble and computational work to assign conformational macrostates to a local geometry in the resulting kinetic networks.

### 2. INTRODUCTION

Obtaining an experimental measure of free energy has been of much interest to biophysical studies of heterogeneous systems, namely protein-DNA and protein-protein assemblies, where the dynamical evolution of macromolecular structure determines biological function. The role of DNA ‘breathing’ in the assembly and function of protein complexes at DNA replication forks within the T4 bacteriophage replisome has long been appreciated based on ensemble biophysical studies, which quantified the presence and extent of such thermally populated conformational states in DNA [2], [3]. Ensemble studies using fluorescent nucleobase analogues were able to show that

structural differences in base stacking and base-pairing persist between adjacent sites at ss-dsDNA junctions, with unique spectral signatures inside and outside of the junction [85]. More recent ensemble studies employing (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs demonstrated that the local structure and degree of disorder near ss-dsDNA junctions was a function of the specific sites investigated[24]. Taken together, there is substantial evidence for the occurrence of thermally populated DNA ‘breathing’ states at replication fork junctions, whose conformations and relative stabilities are a sensitive function of labeling position. Our recent work studying DNA ‘breathing’ using polarization-sweep single-molecule fluorescence (PS-SMF) microscopy demonstrates that these thermally activated DNA conformers can be measured directly and are transiently populated in the vicinity of a DNA replication fork. Our results indicate that the bases and sugar-phosphate backbones sensed by the (iCy3)<sub>2</sub> dimer probes can adopt four quasi-stable local conformations, whose relative stabilities and transition state barriers depend on probe labeling position relative to the ss-dsDNA fork junction. Additionally, the energetics of the (iCy3)<sub>2</sub> dimer-labeled ss-dsDNA constructs were shown to be systematically varied by increasing or decreasing salt concentration relative to physiological conditions. The position-dependent distribution of conformational macrostates was shown to be affected by salt concentration in complementary ways. In this current work, we present the results of a kinetic network model analysis of the same PS-SMF data using a generalized master equation approach in combination with a massively-parallel optimization scheme deployed on high performance computing cluster resources, which is the topic of a forthcoming article. We quantify the thermodynamic minima of all conformational states and the activation barrier heights connecting states within each unique kinetic network, under all experimental conditions examined. The combination of our experimental and analytical approach can be used to sensitively study the mechanism of protein-DNA complex recognition, assembly,



and function, with single molecule sensitivity and microsecond time resolution, allowing for full quantification of the free energy landscape governing fundamental biological processes in replication, repair and beyond.

The results of our kinetic network model analysis reveal that the thermodynamically most stable macrostate is typically B-form DNA, with the secondarily most stable macrostate being attributed to a left-handed conformation that is supported by ensemble studies using circular dichroism measurements [24]. Recent computational studies and previous ensemble experimental studies suggest that local Hoogsteen base-pairing conformations can persist as the thermodynamically most stable macrostate at sites where nucleic acids have undergone methylation [86], [87]. These Hoogsteen base-pairing interactions form through a flipping out of a base-pair from the double-stranded region, followed by rotation of that base-pair about its glycosidic bond, to eventually form a Hoogsteen-base pair, where the pucker of the sugar-phosphate backbone can undergo a conformational transition depending on the site and identity of the methylated base-pair[86]. However, even in the absence of methylation, the ability to form Hoogsteen base-pairs at between complementary nucleic acids still occurred in the free energy surfaces of the same simulations. We hypothesize that such a Hoogsteen base-pairing interaction is responsible for the left-handed state that we observe in our experiments. This phenomenon is diagrammatically captured in Fig. 4.10. The effects of altering salt concentration cause the thermodynamic stability of each conformational macrostate to be reorganized and in some cases results in a re-ordering of the most stable conformational macrostate. The thermodynamic transition barriers obtained from our analysis are modulated by the concentration of salt in a manner which supports the inferred structural assignments given to the four conformational macrostates, based on varying degrees of hydrogen bonding and base-stacking interactions in the

local vicinity of (iCy3)<sub>2</sub> dimer probes that label the ss-dsDNA fork junction. We use the culmination of these results obtained from our model simulations to postulate structural details of each macrostate and offer a proposed mechanism for the interconversion of each conformer within the optimized kinetic network of all experimental data sets.

### 3. MATERIALS AND METHODS

All experimental data analyzed in this paper was obtained using the methodologies described in [120]. For the analysis that is presented here, the theoretical framework employed has been previously used to successfully obtain free energy landscape parameters from single molecule FRET data [14]. A detailed description of the theoretical framework and computational optimization pipeline is the subject Chapter 3. The simulation and optimization of kinetic network models was carried out in MATLAB using the combination of a home-built genetic algorithm and customized mesh-grid refinement based on MATLAB's 'PatternSearch' toolbox. Our MATLAB routine is intended to simulate a single network connectivity, from the set of all possible network connectivity's, for any given data set. In order to scale our optimization to meet the demands of simultaneously accounting for many different network connectivity's, across many distinct data sets, we developed a Linux Bash routine that is capable of deploying thousands of model optimizations simultaneously on a high-performance computing cluster. By removing the bottleneck on the optimization of many model networks, each with large parameter spaces defining them, we were able to obtain the globally optimal kinetic network that best explains the observed macrostate probability distribution and conformational dynamics at DNA replication forks.

To achieve robust optimized fits across all data surfaces in our routine, as shown in Fig. 4.1, we devised a nonlinear chi-squared weighting scheme which accounts for the disparity in

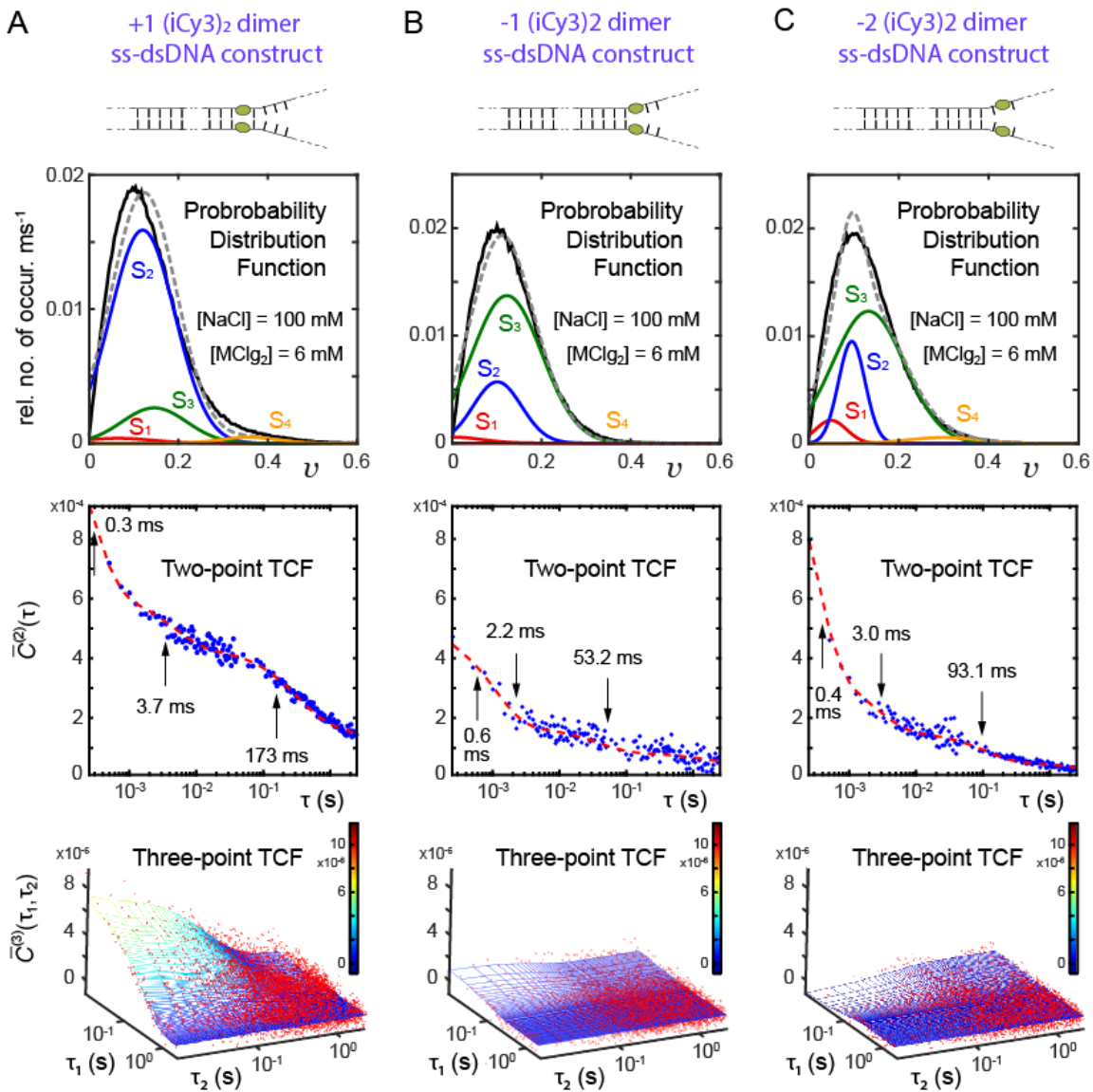
magnitude, noise and data point density across the three data surfaces. The details of this scheme are found in Chapter 3.

## 4. RESULTS AND DISCUSSION

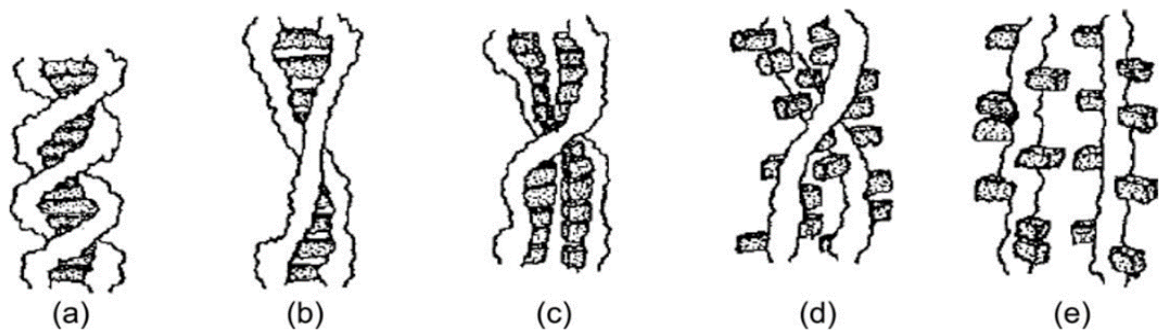
### 4.1 Positional dependence of DNA replication fork free energy landscapes

The optimized kinetic network outcomes under physiological salt conditions ( $[\text{NaCl}] = 100$  mM,  $[\text{MgCl}_2] = 6$  mM) at the +1, -1 and -2 labeling positions are shown in Fig. 4.1 and superimposed onto the experimental data in each panel. The fit outcomes are shown in dashed grey, dashed red and dotted red for the probability distribution function, two-point TCF and three-point TCF respectively. We emphasize that the quality of fits shown is high and significantly better than the dozens, sometimes hundreds, of lower quality model outcomes, which are not shown for the sake of brevity.

Fig. 4.1 illustrates the positional dependence of the four conformational macrostates thermodynamic stability, obtained from optimizing a kinetic network to our experimental data. Before discussing the relative stabilities and making inferred structural assignments to the macrostates depicted in Fig. 4.1, we briefly review the experimental and theoretical basis for such assignments. Previous studies of DNA ‘Breathing’, ensemble studies of average structure, as well as conventional knowledge about the thermodynamics of nucleic acid chemistry offer a handful of structural motifs which can be used to infer structural assignments of our conformational macrostates [2], [3], [24], [72]. Fig. 4.2 shows a logical progression of DNA ‘Breathing’ modes which can occur over a range of temperatures, or salt concentrations.



**Figure 4.1** Optimized fits from our kinetic network analysis for the position dependence of DNA ‘breathing’ under physiological salt conditions. Panels A, B, and C are for the +1, -1 and -2 positions respectively.



**Figure 4.2** Depiction of the various modes and extents of DNA ‘breathing’ one might expect (taken from [1]). Panel (a) shows fully stable B-form DNA. Panel (b) depicts a breathing mode where the bases lift off one another. Panel (c) depicts a breathing mode where the bases break their complementary hydrogen bonding interaction. Panel (d) is a pre-melting transition where a mixture of base-unstacking and base-unpairing is observed. Panel (e) is a fully melted case, where the two strands have completely unannealed.

In panel A of Fig. 4.2 the canonical B-form DNA is depicted, characterized by fully stacked and hydrogen-bonded base-pairs. In panel B the bases have lifted off one another, disrupting the stacking interactions. In panel C, the bases have broken their complementary hydrogen bonding such that the duplex is opened up to the surrounding environment. In panel D some combination of base unstacking and loss of hydrogen bonding is depicted, leading to a highly disordered and unstable conformation with many bases flipping outside the duplex. Finally, panel E depicts a nearly fully melted structure which has lost all resemblance to dsDNA. The energetics of each possibility in this series has been addressed in detail, while also accounting for the role of water in the entropy of the system, through a series of calorimetric studies [72]. Treating these various degrees of freedom as energetically distinct structural fluctuations which perturb the arrangement of the sugar-phosphate backbone, we can begin to make assignments to our kinetic network results obtained on DNA replication fork junctions.

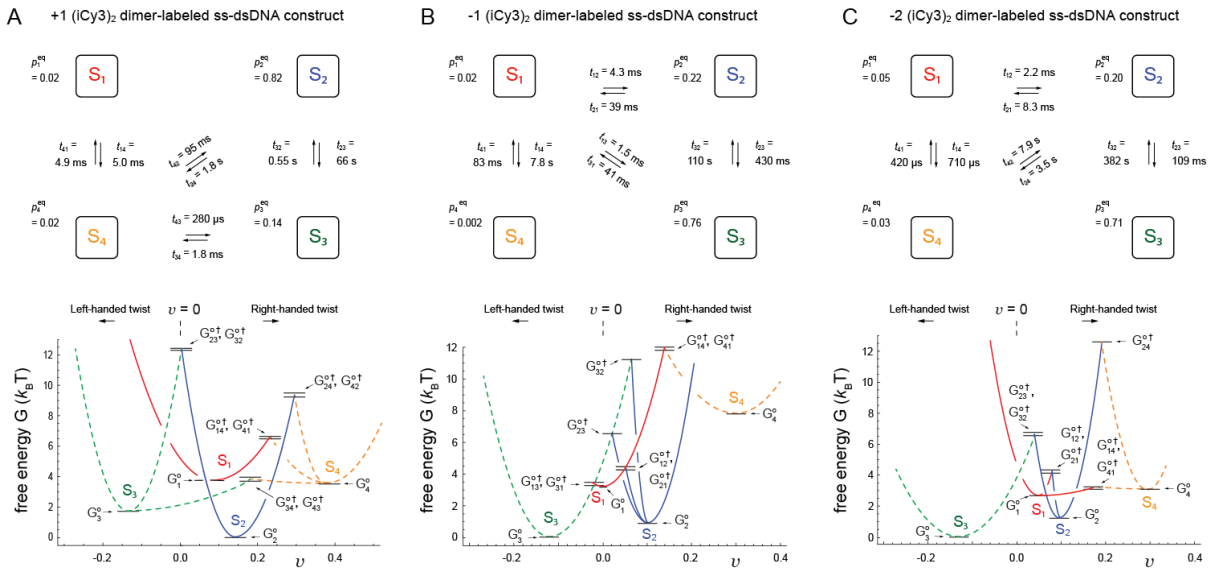
Starting with the +1 position under physiological conditions, we attribute the state  $S_2$  to B-form DNA, which is most stable at the +1 position and is the expected predominant conformer in

duplex DNA. The state  $S_3$  is secondarily most stable, which we attribute to a left-handed conformation, consistent with ensemble studies as well as recent computational studies on the role of Hoogsteen base-pairing [24], [86]. The remaining two states  $S_1$  and  $S_4$  are relatively unstable compared to  $S_2$  and  $S_3$  at the +1 position but remain essential to explaining the observed conformational dynamics in the context of a generalized master equation. The low visibility state  $S_1$  is attributed to a propped-open state at the junction, where the loss of complementary hydrogen bonding leads to a locally open conformation that is consistent with the loss of excitonic coupling between the two Cy3 monomers, yielding a low visibility signal. Finally,  $S_4$  is attributed to the bases surrounding the junction lifting off one another, thereby disturbing base-stacking interactions, leading to a more parallel alignment of the sugar-phosphate backbone, that is consistent with a stronger polarized response in our Cy3 dimer probes, and a higher visibility value, in our model. It is interesting to note that the equilibrium probability of  $S_1$  exceeds that of  $S_4$  at all positions under physiological conditions, consistent with the notion that base-stacking interactions rather than base-pairing interactions are the predominant enthalpic contribution to duplex stabilization in dsDNA [72], [88].

At the -1 and -2 positions the thermodynamically most stable state shifts to  $S_3$  from  $S_2$ , seen in panels B and C of Fig. 4.1. As previously mentioned, we assign the state  $S_3$  to a left-handed conformation. This redistribution in the equilibrium probabilities of the two most stable, and oppositely handed, states is consistent with ensemble studies of (iCy3)<sub>2</sub> dimer labeled fork constructs, where the handedness of the (iCy3)<sub>2</sub> dimer probe was shown to invert from right- to left-handed, moving across the junction from the +1 to -1 position [24]. Notably,  $S_2$  becomes the secondarily most stable state at the -1 and -2 positions, causing the relative stability of  $S_2$  and  $S_3$  in the duplex to become the mirror image of the relative stabilities of  $S_2$  and  $S_3$  outside the junction

in the single-stranded region. The reduced stability of  $S_1$  and  $S_4$  at the -1 position, relative to both the +1 and -2 positions, suggests that the free energy landscape at the -1 position is unique and dominated by mostly two states. The uniqueness of the -1 position in our analysis is in agreement with previous ensemble measurements made on DNA systems containing base analogues, which showed that the degree of base-stacking was greatly reduced at the -1 position relative to all neighboring positions [85]. As was noted in the initial work which these analyses are based on, the dynamical trends in the two-point and three-point TCFs are expected given the differential stabilities of stacked bases within dsDNA versus ssDNA regions spanning the ss-dsDNA fork junction [120].

The free energy surfaces under physiological conditions for the +1, -1, and -2 positions, calculated using Eqn. 21, are shown in Fig 4.3. The diagrammatic depiction of networks composed of states  $S_1$  through  $S_4$  illustrate the microscopic rate constants determined from our kinetic network analysis, which are on the order of microseconds to seconds. Similarly, it can be appreciated that many of the transition barriers between conformational macrostates are within the range of  $\sim 2-6 k_bT$ , with only a few of the rarer transitions have higher transition barriers. The optimization results for all studies are summarized in Tables 6, 7 and 8.

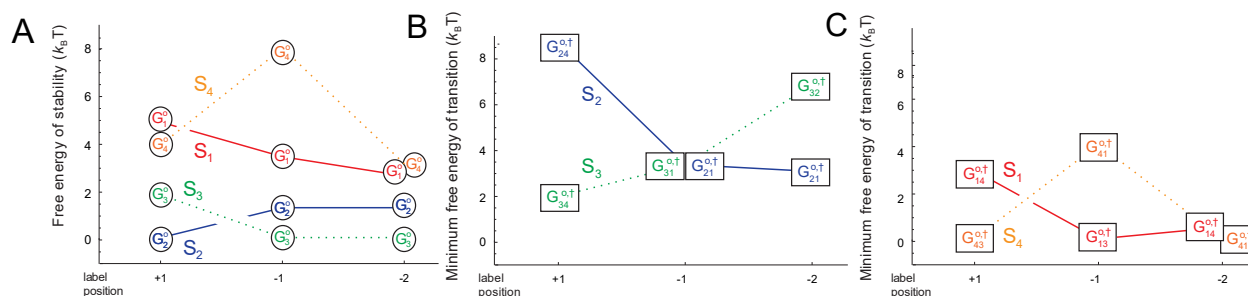


**Figure 4.3** Free energy landscapes and associated kinetic networks for each of the positions examined under physiological salt conditions. Microscopic rate constants are shown and labeled by the state-to-state transitions they correspond to.

A summary of the thermodynamic and mechanical stability at each position studied under physiological conditions is presented in Fig. 4.4. Panel A of Fig. 4.4 summarizes the free energy minima associated with each of the four conformational macrostates as a function of position. Panels B and C of Fig. 4.4 show the minimum transition barrier height out of each macrostate, at each position. Panels B and C separate the minimum barrier heights for states  $S_2$ ,  $S_3$  and  $S_1$ ,  $S_4$  respectively. By plotting the minimum transition barrier height as a function of labeling position, for each macrostate, the mechanical stability of each underlying conformer can be better understood. The mechanical stability of  $S_2$  is significantly higher than any of the remaining macrostates at the +1 position, suggesting that transitions out of  $S_2$  are unlikely to occur compared to all other transition pathways (Fig. 4.4 Panel B). The mechanical stability of  $S_3$  is relatively low at the +1 position, but steadily increases across the junction to reach a maximum at the -2 position. Simultaneously, the mechanical stability of  $S_2$  decreases moving outward across the junction,



reaching a minimum at the -2 position. Notably, the mechanical stability of  $S_2$  and  $S_3$  are in close competition at the -1 position, indicating more rapid interconversion of the two conformers is occurring, further supporting previous studies which suggested a uniqueness of the conformational landscape at the -1 position compared to neighboring positions [85]. The mechanical stability of  $S_1$  and  $S_4$  are significantly lower on average than either  $S_2$  or  $S_3$ . The mechanical stability of  $S_1$  drops off moving outward across the junction, while  $S_4$  becomes more mechanically stable moving from the +1 to the -1 position before returning to a level of mechanical stability at the -2 position which reflects the +1 position (Fig. 4.4 Panel C).



**Figure 4.4** Summary diagrams for the thermodynamic and mechanical stability of the +1, -1 and -2 positions under physiological salt conditions. Panel A illustrates the free energy minima associated with of the four conformational macrostates as a function of probe position. Panels B and C depict the minimum transition barrier governing the mechanical stability of each macrostate, as the probe position is varied.

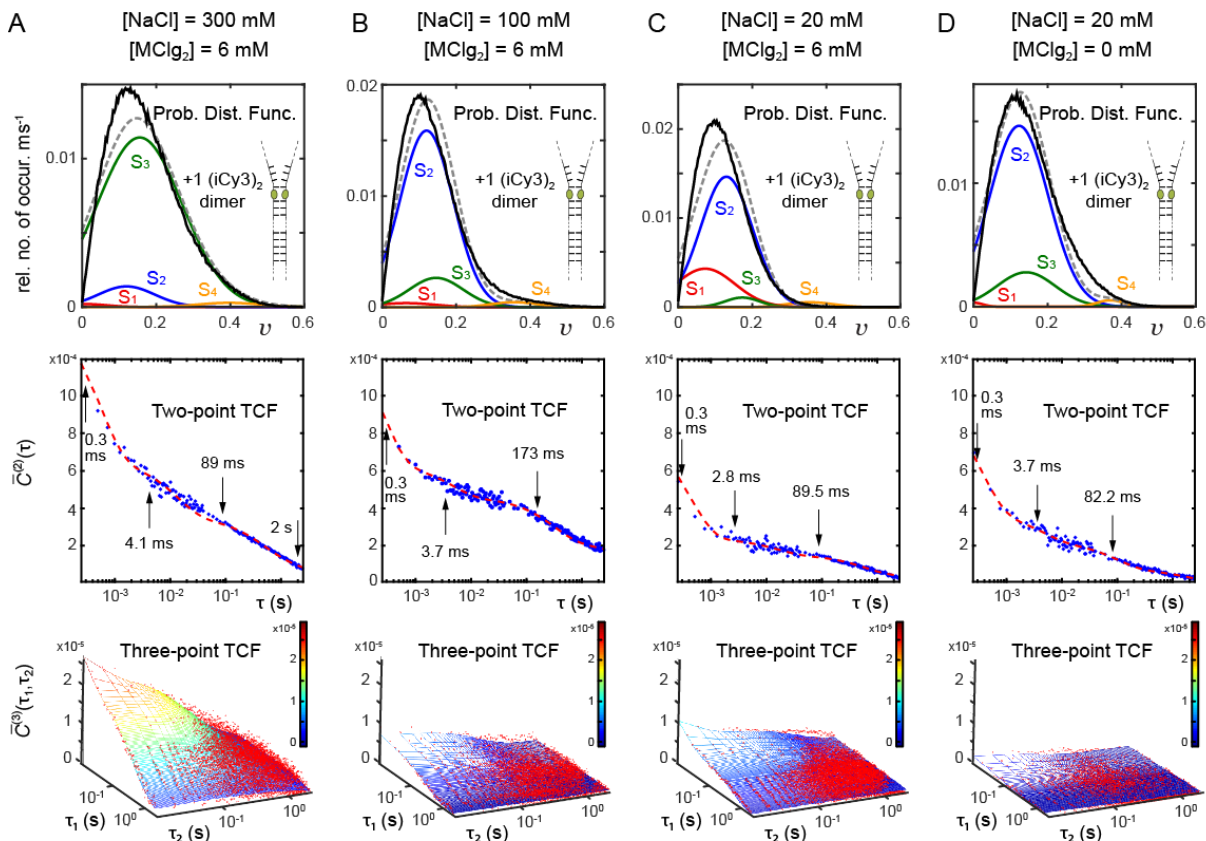
The continuous decrease in mechanical stability under physiological conditions of  $S_2$ , moving outward across the junction, is consistent with the assignment of  $S_2$  to B-form DNA, where the degree of complementary base pairing decreases moving from the +1 to -2 positions in our model replication fork junctions. The steady increase in the mechanical stability of  $S_3$  can be understood from the loss of base-stacking interactions in the single stranded region that occurs moving outward across the junction. This loss of an enthalpic penalty from the disruption of base-

stacking interactions tends to stabilize the formation of an unstacked, left-handed conformer at the -1 and -2 positions [72], [82], in which a Hoogsteen base-pair has formed, relative to the predominantly B-form conformer seen in the duplex [86]. The decreasing mechanical stability of  $S_1$  across the junction is consistent with the loss of enthalpically favored hydrogen bonded base-pairs and base-stacking interactions at both nearest neighbor positions that would tend to stabilize a locally propped-open conformation. The trend in mechanical stability for  $S_4$  is perhaps most difficult to interpret. At the +1 and -2 positions, the mechanical stability of  $S_4$  is quite low, consistent with the disruption of base-stacking interactions in the duplex being the largest enthalpic penalty to overcome during a conformational fluctuation. The sharp rise in the mechanical stability of  $S_4$  at the -1 position must be due to some unique interplay between the enthalpic penalty of disrupting base-stacking interactions in the duplex and a higher degree of conformational and solvent entropy upon formation of the thermodynamically unstable  $S_4$  macrostate.

## 4.2 Salt dependence of DNA replication fork free energy landscapes

The optimized outcomes for the salt dependent studies at the +1 and -2 positions are shown in Fig. 4.5 and Fig. 4.6, Panels A-D, with the same color coding as described for the positional dependence of Fig. 4.1. We carried out four unique sets of salt dependent experiments, at the both the +1 and -2 position, with variable NaCl and MgCl<sub>2</sub> concentrations, namely, [NaCl] = 100 mM, [MgCl<sub>2</sub>] = 6 mM; [NaCl] = 20 mM, [MgCl<sub>2</sub>] = 6 mM; [NaCl] = 20 mM, [MgCl<sub>2</sub>] = 0 mM; [NaCl] = 300 mM, [MgCl<sub>2</sub>] = 6 mM. This spanning set of conditions is intended to address the role of elevated and depleted monovalent ions as well as bivalent ions, as they influence the propensity for conformational fluctuations to occur at replication fork junctions. Notably, few details are currently available about how salt concentration affects the equilibrium distribution

of conformational macrostates at and near ss-dsDNA fork junctions, or the transition barriers that mediate their interconversion.



**Figure 4.5** Optimized fits from our kinetic network analysis for the salt dependence of DNA ‘breathing’ at the +1 position. Panels A-D are labeled by the salt condition they represent.

Fig. 4.5, Panels A-D, show the analysis results for salt dependent effects on the conformational macrostates observed at the +1 position, across all four sets of conditions. The analysis results for physiological salt concentrations ( $[\text{NaCl}] = 100 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ) at the +1 position were already discussed in section 3.1. Reduction of monovalent salt concentration ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ), results in the slight destabilization of the B-form macrostate  $S_2$  and an increase in the thermodynamic stability of the locally propped-open macrostate  $S_1$ , relative to physiological salt conditions. Previous ensembles studies of DNA melting temperature

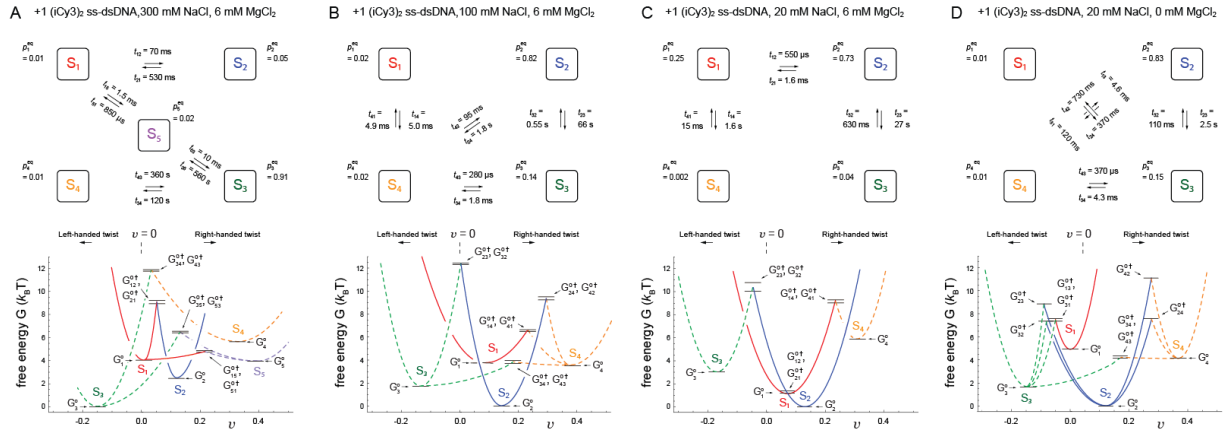
curves as a function of salt concentration suggest a general destabilization of DNA at lower salt concentrations, consistent with a higher propensity for non B-form conformers as salt concentration is decreased in our measurements [80], [89], [90]. The thermodynamic stability of  $S_3$  is decreased by the loss of monovalent salt at the +1 position, as is the thermodynamic stability of  $S_4$ . The loss of thermodynamic stability of  $S_3$  and  $S_4$  at decreased monovalent salt is true at both the +1 and -2 positions (Fig. 4.5 Panel C, Fig. 4.6 Panel C), suggesting a role played by monovalent salt in the stabilization of the left-handed, bases unstacked, Hoogsteen base-paired  $S_3$  conformer as well as the right-handed, bases unstacked,  $S_4$  conformer. Given our proposition that both  $S_3$  and  $S_4$  are mediated by alteration of the base-stacking interactions, one would expect the degree of enthalpic penalty to unstacking bases to be mutually affected by the loss of monovalent salt for either handedness of the (iCy3)<sub>2</sub> dimer probe, both in the duplex and single-stranded regions.

The subsequent loss of magnesium under already low monovalent salt conditions ([NaCl] = 20 mM, [MgCl<sub>2</sub>] = 0 mM), leads to the thermodynamic stability of  $S_3$  and  $S_4$  being restored to levels which are similar to physiological salt conditions, while  $S_1$  experiences its global free energy maximum across all conditions and constructs examined. It is thought that magnesium plays a special role in the stabilization of key structures in DNA, with a significantly higher degree of sensitivity in the melting temperature dependence of DNA due to alterations in magnesium concentration than sodium or potassium concentration [80]. In contrast to the effects of lowering monovalent salt, where the stability of conformers mediated by unstacked bases was decreased, the effect of lowering bivalent salt appears to raise the stability of  $S_3$  and  $S_4$ , suggesting that the role of monovalent and bivalent salt in mediating base-stacking interactions have an opposite effect on free energy at the +1 position.

Raising the concentration of monovalent salt at the +1 position relative to physiological conditions ( $[\text{NaCl}] = 300 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ), causes the thermodynamic minimum of the free energy landscape to shift towards  $S_3$ , while  $S_4$ ,  $S_2$  and  $S_1$  all become thermodynamically less stable. Interestingly, a fifth macrostate,  $S_5$ , emerges under elevated salt concentrations. This additional macrostate is minimally necessary to explain the results of our PS-SMF experiments using a kinetic network analysis. While the apparent conformational heterogeneity at the +1 position is increased by raising the concentration of monovalent salt, the thermodynamic stability of  $S_3$  is lower and further removed from the thermodynamic stability of neighboring macrostates than any of the remaining salt concentrations examined at the +1 position. The greater stability of  $S_3$  compared to all other macrostates present the +1 position, under elevated monovalent salt, is consistent with increased stability of duplex DNA during both melting curve and force pulling experiments of DNA under elevated monovalent salt [80], [88].

The appearance of an additional conformational macrostate at the +1 position under elevated monovalent salt could simply be the result of  $S_5$  dwell times becoming long enough to be experimentally accessible by the range of integration times available to our PS-SMF experiment [120]. This is consistent with the slowest overall dynamics observed at the +1 position being under elevated monovalent salt (Fig. 4.5 Panels A-D), as well as the emergence of a fifth macrostate at the -2 position under decreased salt and the loss of magnesium, where similarly the slowest dynamics are observed across all conditions explored at the -2 position (Fig. 4.6 Panels A-D). The complete free energy landscapes for the salt dependence of the +1 position are shown in Fig. 4.6, Panels A-D, along with the timescales of state-to-state transitions in the optimized kinetic

networks. A convenient summary of the thermodynamic and mechanical stability of conformational macrostates at the +1 position is given in Fig. 4.9, Panels A-C.

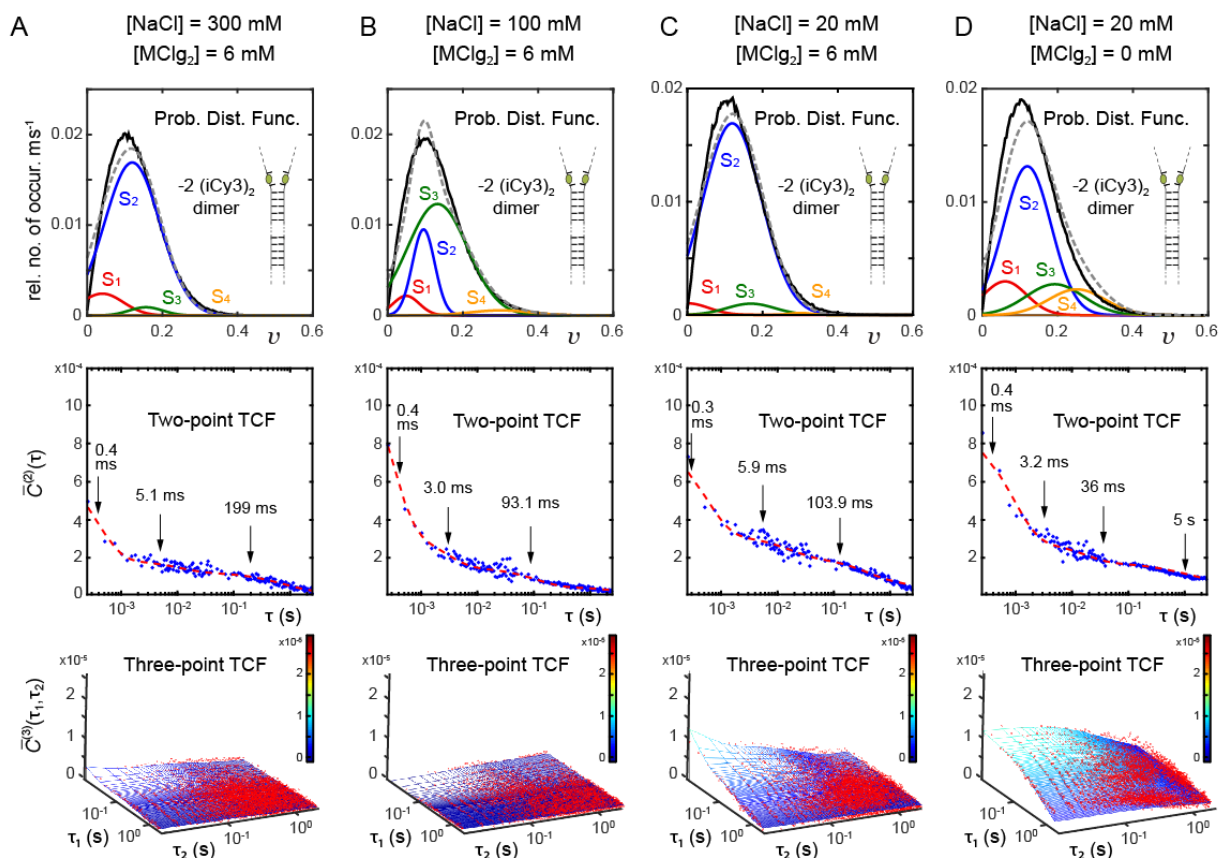


**Figure 4.6** Free energy landscapes and associated kinetic networks for each of the salt conditions examined at the +1 position. Microscopic rate constants are shown and labeled by the state-to-state transitions they correspond to.

The salt induced effects on the mechanical stability of the +1 position are summarized in Fig. 4.9, panels B, C. Raising the concentration of monovalent salt at the +1 position ( $[\text{NaCl}] = 300 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ) tends to increase the mechanical stability of  $S_3$  and  $S_4$ , the conformers mediated by base-stacking interactions. Whereas the mechanical stability of  $S_1$  and  $S_2$  is decreased, which are conformers mediated by hydrogen bonding of complementary base-pairs within our model. Lowering the concentration of monovalent salt away from physiological conditions at the +1 position ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ) has the same effect on the mechanical stability of each macrostate compared with raising the concentration. The mechanical stability of  $S_3$  and  $S_4$  is increased and the mechanical stability of  $S_1$  and  $S_2$  is decreased. This sort of interplay is reminiscent of the concept of cold denaturation, where the optimal conditions for the thermodynamic and mechanical stability of biologically relevant conformations occur precisely at physiological conditions, and any alteration to those conditions tends to destabilize the system

[72]. The loss of bivalent salt, at already low monovalent salt concentration ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 0 \text{ mM}$ ), causes the mechanical stability of each conformer at the +1 position to revert toward the picture seen under physiological conditions. This trend in mechanical stability echoes the similar restoration of relative thermodynamic stabilities at the +1 position upon loss of bivalent salt, at already depleted monovalent salt concentrations. It is notable that the mechanical stability of  $S_3$  and  $S_4$  as well as  $S_1$  and  $S_2$  are modulated by the effects of altering salt concentration in equal and opposite ways at the +1 position, supporting the notion that the enthalpic and entropic contributions to the intramolecular interactions mediating these pairs of conformational macrostates are interrelated.

Fig. 4.7, Panels A-D, show the analysis results for salt dependent effects on the conformational macrostates observed at the -2 position, across all four sets of conditions. Physiological salt conditions ( $[\text{NaCl}] = 100 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ) at the -2 position were previously described in section 3.1. Decreasing the concentration of monovalent salt at the -2 position ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ) causes the thermodynamically most stable macrostate to shift away from  $S_3$  toward  $S_2$ . Much like the effects observed at the +1 position, the loss of monovalent salt tends to destabilize the states  $S_3$  and  $S_4$  which are governed by base-stacking interactions in our model. The stability of the propped-open  $S_1$  macrostate is decreased by the loss of monovalent salt at the -2 position. This destabilization occurs in an opposite manner to the +1 position, suggesting that the salt induced instability of base-pairing interactions has an oppositional effect on the duplex and single-stranded regions, much like the hypothesized mechanism outlined in our initial work [120].



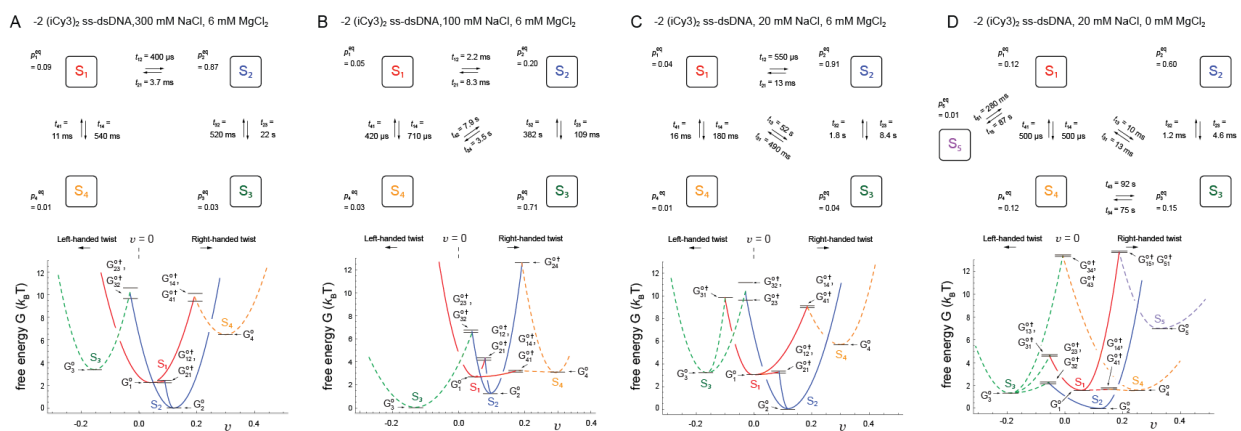
**Figure 4.7** Optimized fits from our kinetic network analysis for the salt dependence of DNA ‘breathing’ at the -2 position. Panels A-D are labeled by the salt condition they represent.

The loss of bivalent salt at lowered monovalent salt concentrations ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 0 \text{ mM}$ ) causes the thermodynamic stability of  $S_1, S_3$  and  $S_4$  to all increase while  $S_2$  remains the thermodynamic minima. This change brings the relative thermodynamic stability of each conformational macrostate into closer agreement with what was observed under physiological  $\tau$  conditions, much like the results of losing bivalent salt at the +1 position, albeit with slightly less strict correspondence between the two salt concentration regimes at the -2 position. Interestingly, a fifth macrostate is required to explain these experimental results at the -2 position, much like the results of elevated monovalent salt at the +1 position.



As previously mentioned, these two salt concentration regimes, at each respective position, correspond to the slowest overall dynamics observed across all conditions investigated.

Finally, raising monovalent salt concentration at the -2 position ( $[\text{NaCl}] = 300 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ) also causes the thermodynamic minima to shift away from  $S_3$  toward  $S_2$ , further suggesting the role played by salt parallels the effects seen in cold denaturation. Under these same conditions,  $S_1$  is slightly stabilized while  $S_4$  is destabilized. At the -2 position, the effect of increasing monovalent salt tends to stabilize conformers mediated by hydrogen bonding,  $S_1$  and  $S_2$ , while destabilizing conformers mediated by base-stacking interactions,  $S_3$  and  $S_4$ . The free energy landscapes and timescales connecting macrostates within each kinetic network are shown in Fig. 8 for the -2 position salt series results.



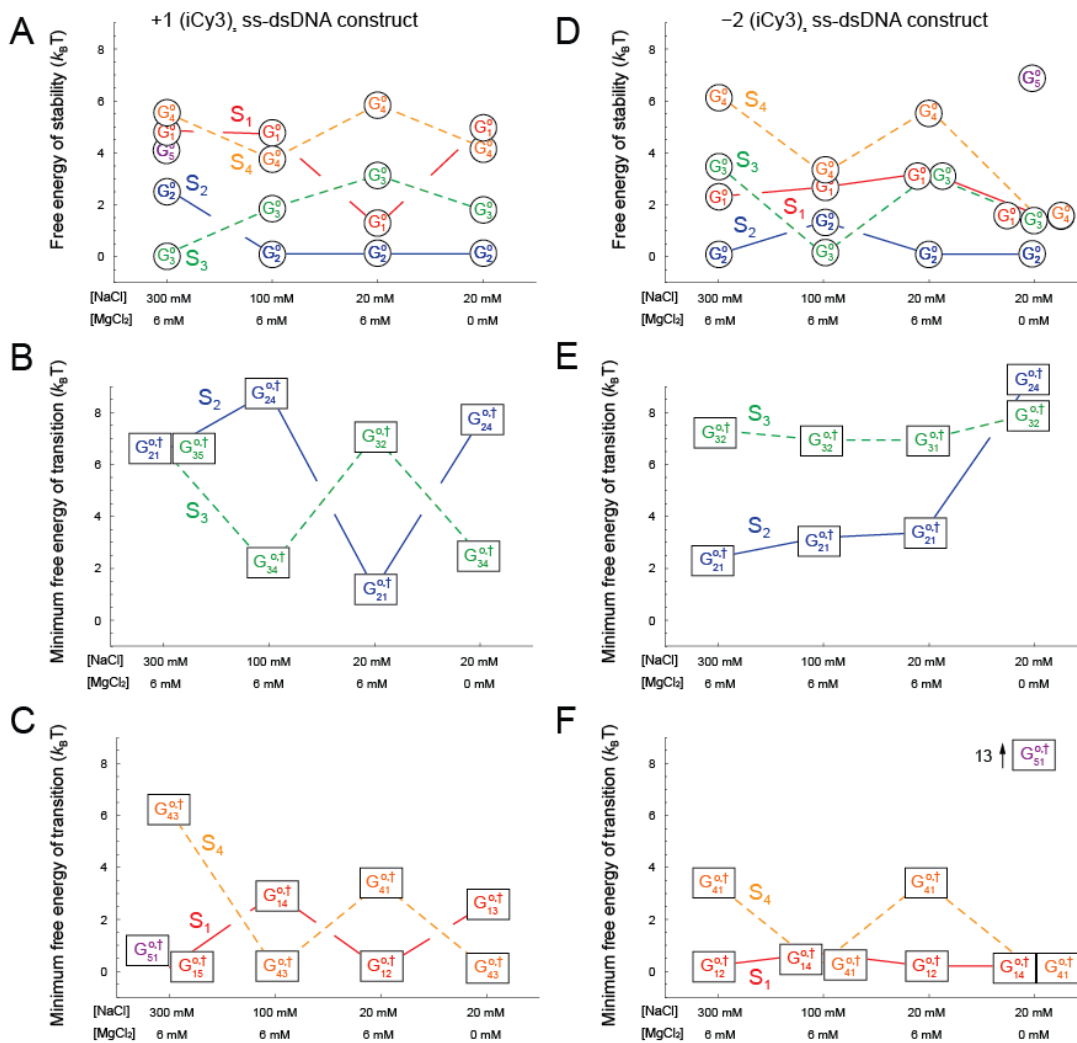
**Figure 4.8** Free energy landscapes and associated kinetic networks for each of the salt conditions examined at the -2 position. Microscopic rate constants are shown and labeled by the state-to-state transitions they correspond to.

The salt induced effects on the mechanical stability of the -2 position are summarized in Fig. 4.9, panels E, F. On average, the mechanical stability of the -2 position is modulated to a lesser extent than the +1 position. One might expect the mechanical stability of conformational macrostates to be less sensitive to the alteration of salt in the single-stranded region (-2 position),

as the conformational landscape is already quite disordered and more metastable than duplex DNA (+1 position). Raising the concentration of monovalent salt at the -2 position ( $[\text{NaCl}] = 300 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ) tends to increase the mechanical stability of  $S_3$  and  $S_4$ , the conformers mediated by base-stacking interactions. A similar trend to that observed at the +1 position. The mechanical stability of  $S_1$  and  $S_2$  are decreased by increasing the concentration of monovalent salt, conformers mediated by hydrogen-bonding interactions, again reflecting the trend observed at the +1 position under elevated monovalent salt.

Decreasing the concentration of monovalent salt at the -2 position ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 6 \text{ mM}$ ) tends to increase the mechanical stability of  $S_4$  and decrease the mechanical stability of  $S_1$ , while leaving the mechanical stability of  $S_2$  and  $S_3$  relatively unchanged. The decrease in mechanical stability of  $S_1$ , as well as the increased mechanical stability of  $S_4$ , agree with the salt dependent results at the +1 position. This suggests that the mechanisms governing changes to the enthalpy and entropy of conformational changes between these macrostates is similar between the duplex and single-stranded region upon lowering monovalent salt, whereas the same does not appear to be true for the mechanical stability of  $S_2$  and  $S_3$  at the -2 position.

The loss of bivalent salt ( $[\text{NaCl}] = 20 \text{ mM}$ ,  $[\text{MgCl}_2] = 0 \text{ mM}$ ) causes the mechanical stability of  $S_2$  and  $S_3$  to sharply increase, while the mechanical stability of  $S_1$  and  $S_4$  taken on their global minima at the -2 position. Again, two of the four conformational macrostates, namely  $S_2$  and  $S_4$ , see changes in their mechanical stability which reflects the trends observed at the +1 position upon loss of magnesium. While the remaining two macrostates,  $S_1$  and  $S_3$ , appear to be differentially affected by the loss of bivalent salt compared with the results obtained in the duplex.



**Figure 4.9** Summary diagrams for the thermodynamic and mechanical stability of the +1 and -2 and positions under a series of salt conditions. Panel A and D illustrate the free energy minima associated with of the four conformational macrostates as a function of probe position (+1,-2). Panels B, C, E and F depict the minimum transition barrier governing the mechanical stability of each macrostate, as the salt concentration is varied, at the +1 (B and C) and -2 (E and F) positions respectively.

The general trends in the thermodynamic and mechanical stability of conformational macrostates at the +1 and -2 positions under a variety of salt conditions suggests that similar interactions mediate  $S_1$  and  $S_2$  as well as  $S_3$  and  $S_4$ . Notably, in most optimized kinetic networks, these pairs of states ( $S_1$  and  $S_2$ ,  $S_3$  and  $S_4$ ) are directly connected as nearest

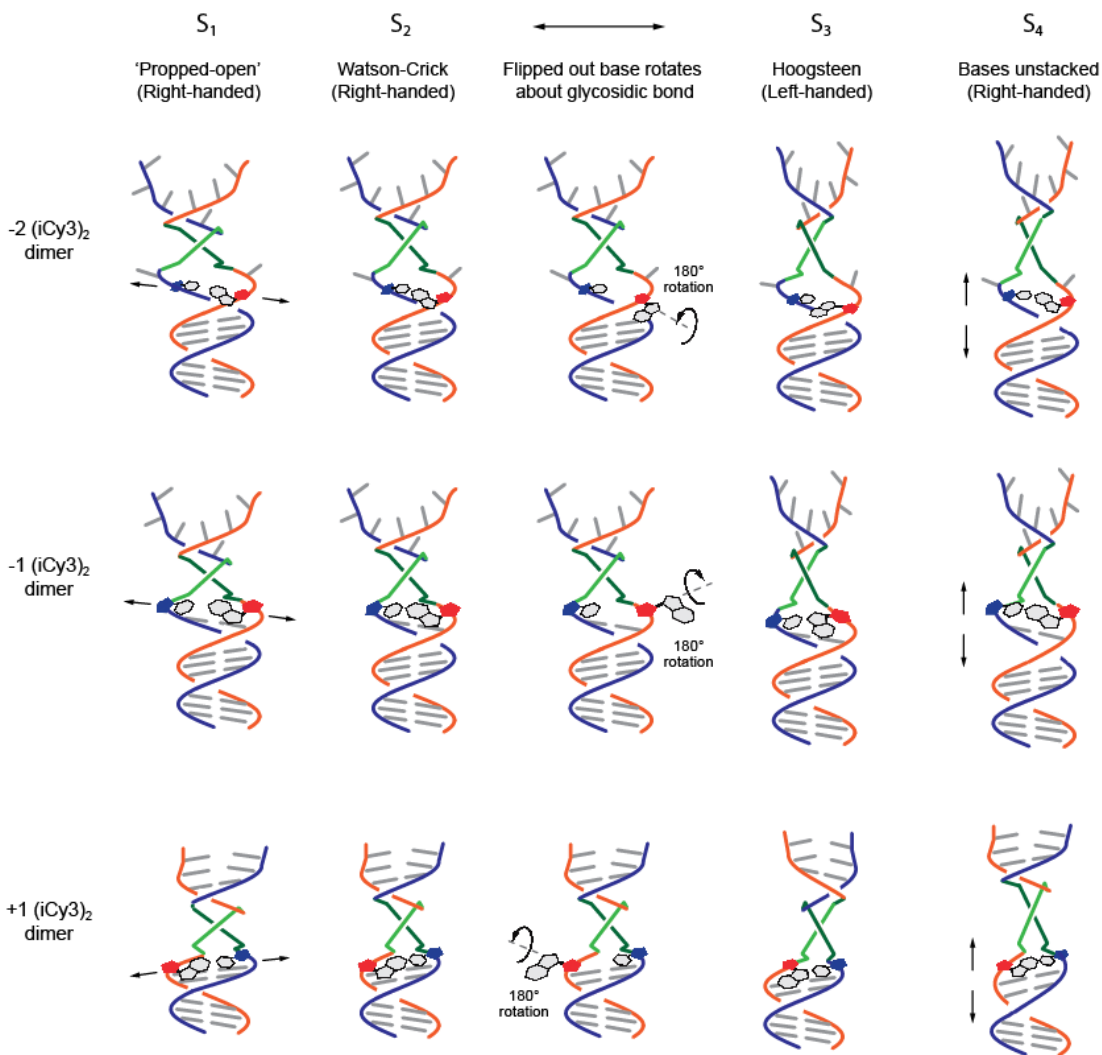
neighbors in the layout of the network. Suggesting regularly observed transitions between these conformers must occur. As previously stated, we propose that  $S_1$  is a right-handed, locally propped open conformation and  $S_2$  is B-form DNA. We also propose that  $S_3$  is a left-handed, bases unstacked, Hoogsteen base-paired conformation and  $S_4$  is a right-handed, bases-unstacked conformation. Using the insights gained from the results of our kinetic network analysis, we can propose a structural model to accompany the free energy landscape governing the conformational fluctuations we observe at replication fork junctions in ss-dsDNA.

## 5. CONCLUSION

### **5.1 Proposed structural model of conformational macrostates at replication fork junctions**

With the results of our kinetic network model analysis, we may now begin to fill in a proposed structural model for the conformational macrostates occurring in the vicinity of a ss-dsDNA replication fork junction. Our analysis suggests that four thermodynamically stable macrostates persist under physiological conditions inside and outside of the junction. The assignments given on a structural basis are depicted in Fig 4.10, with each position examined in this study shown separately to emphasize the structural differences from the duplex side out into the single stranded region. Furthermore, our analysis suggests that two of these macrostates dominate the free energy landscape, namely  $S_2$ , B-form DNA, and  $S_3$ , a left-handed Hoogsteen base-paired conformer [86]. These two macrostates are the most thermodynamically stable within the duplex and single stranded regions, under all conditions examined. The remaining two macrostates  $S_1$  and  $S_4$  are

present in all data sets with various degrees of occupancy, owing to the trends observed in the thermodynamic and mechanical stability of all macrostates under various salt concentrations. We attribute  $S_1$  to a locally propped-open state where the complementary hydrogen bonds have been broken to allow for a locally form a bubble, much like the mechanism of DNA ‘breathing’ first proposed [2]. The final macrostate,  $S_4$ , is attributed to a right-handed bases unstacked conformation which is often rare in DNA only cases but can be important in the formation of protein-DNA complexes. The structural model depicted in Fig. 4.10 summarizes these results and emphasizes the sensitivity we have to the site-specific labeling position of the DNA using  $i(\text{Cy}3)_2$  dimer probes. Building off these DNA only studies, we can now begin to examine the effects of protein binding on the conformational landscape of DNA. We can also imagine examining sequence and polarity dependent dynamics in the vicinity of junctions or other such locally important regions of DNA, where the enthalpy and entropy effects of various mixtures of AT versus GC base pairs, base-stacks and locations within the major and minor grooves of DNA can all influence the energetics.



**Figure 4.10** Proposed structural model based on the results of our kinetic network analysis of replication fork junctions under numerous solvent conditions. Rows are labeled by position of the (iCy3)<sub>2</sub> dimer probe while columns label the conformation of the sugar-phosphate backbone, as sensed by the probes in each of the sites examined. The central column depicts the mechanism of action for formation of a Hoogsteen base-pair at the replication fork junction.

**Table 6** Optimized kinetic network model parameters for the free energy minima and activation barriers (left column) and optimized PDF parameters (right column) for all positions and conditions studied. Blank entries denote connections or states which did not exist in the optimized network models.

Construct	$A_1$	$\langle v \rangle_1$	$\sigma_1$	$p_1^{eq}$	$A_2$	$\langle v \rangle_2$	$\sigma_2$	$p_2^{eq}$	$A_3$	$\langle v \rangle_3$	$\sigma_3$	$p_3^{eq}$	$A_4$	$\langle v \rangle_4$	$\sigma_4$	$p_4^{eq}$	$A_5$	$\langle v \rangle_5$	$\sigma_5$	$p_5^{eq}$
+1 Dimer 100mM NaCl	0.146	0.057	0.059	0.021	6.111	0.112	0.053	0.819	1.019	0.139	0.053	0.137	0.176	0.35	0.048	0.021	-	-	-	-
+1 Dimer 20mM NaCl	1.787	0.064	0.055	0.249	6.068	0.114	0.048	0.744	0.316	0.166	0.032	0.025	0.02	0.295	0.053	0.002	-	-	-	-
+1 Dimer 20mM NaCl 0mM MgCl2	0.121	0.0001	0.019	0.005	5.307	0.121	0.062	0.825	1.012	0.141	0.06	0.154	0.192	0.368	0.028	0.013	-	-	-	-
+1 Dimer 300mM NaCl	0.083	0.003	0.037	0.007	0.587	0.114	0.049	0.073	4.837	0.15	0.075	0.91	0.021	0.321	0.072	0.003	0.119	0.391	0.054	0.016
-1 Dimer 100mM NaCl	0.225	0.0001	0.045	0.025	2.286	0.097	0.037	0.217	5.5	0.118	0.054	0.756	0.002	0.296	0.049	0.0003	-	-	-	-
-2 Dimer 100mM NaCl	0.783	0.045	0.027	0.053	3.455	0.091	0.023	0.203	4.475	0.128	0.063	0.71	0.202	0.296	0.063	0.032	-	-	-	-
-2 Dimer 20mM NaCl	0.404	0.0001	0.038	0.039	6.71	0.117	0.054	0.914	0.392	0.166	0.044	0.043	0.067	0.298	0.02	0.003	-	-	-	-
-2 Dimer 20mM NaCl 0mM MgCl2	1.432	0.055	0.034	0.123	6.317	0.114	0.037	0.601	1.313	0.187	0.045	0.151	1.094	0.245	0.045	0.123	0.005	0.315	0.046	0.0006
-2 Dimer 300mM NaCl	0.974	0.038	0.039	0.095	6.808	0.118	0.051	0.881	0.369	0.154	0.031	0.028	0.019	0.298	0.039	0.001	-	-	-	-

Construct	$G_1^{0E}$	$G_2^{0E}$	$G_3^{0E}$	$G_4^{0E}$	$G_5^{0E}$	$G_{1,2}^{0E}$	$G_{2,1}^{0E}$	$G_{1,3}^{0E}$	$G_{3,1}^{0E}$	$G_{1,4}^{0E}$	$G_{4,1}^{0E}$	$G_{1,5}^{0E}$	$G_{5,1}^{0E}$	$G_{2,3}^{0E}$	$G_{3,2}^{0E}$	$G_{2,4}^{0E}$	$G_{4,2}^{0E}$	$G_{2,5}^{0E}$	$G_{5,2}^{0E}$	$G_{3,4}^{0E}$	$G_{4,3}^{0E}$	$G_{3,5}^{0E}$	$G_{5,3}^{0E}$	$G_{4,5}^{0E}$	$G_{5,4}^{0E}$	
+1 Dimer 100mM NaCl	3.625	0	1.781	3.648	"	"	"	"	"	2.862	2.84	"	"	12.357	7.565	8.739	5.816	"	"	1.869	0	"	"	"	"	"
+1 Dimer 20mM NaCl	1.093	0	3.03	5.763	"	0	1.093	"	"	8.001	3.331	"	"	10.802	7.036	"	"	"	"	"	"	"	"	"	"	
+1 Dimer 20mM NaCl 0mM MgCl2	4.937	0	1.675	4.096	"	"	2.524	5.787	"	"	"	"	"	8.817	5.679	7.584	6.897	"	"	2.45	0	"	"	"	"	
+1 Dimer 300mM NaCl	4.755	2.522	0	5.437	4.028	4.36	6.444	"	"	"	"	0	0.727	"	"	"	"	"	"	11.846	6.06	6.486	2.459	"	"	
-1 Dimer 100mM NaCl	3.375	1.247	0	7.916	"	1.079	3.292	0	3.347	8.584	4.043	"	"	5.698	11.258	"	"	"	"	"	"	"	"	"	"	
-2 Dimer 100mM NaCl	2.579	1.251	0	3.098	"	1.645	2.973	"	"	0.518	0	"	"	5.549	6.801	11.317	12.126	"	"	"	"	"	"	"	"	
-2 Dimer 20mM NaCl	3.149	0	3.037	5.591	"	0	3.15	11.458	6.793	5.781	3.339	"	"	9.633	8.117	"	"	"	"	"	"	"	"	"	"	
-2 Dimer 20mM NaCl 0mM MgCl2	1.56	0	1.381	1.56	6.895	"	"	9.894	10.064	0	0	18.954	13.254	9.137	7.756	"	"	"	"	18.709	18.888	"	"	"	"	
-2 Dimer 300mM NaCl	2.223	0	3.421	6.128	"	0	2.223	"	"	7.195	3.29	"	"	10.894	7.155	"	"	"	"	"	"	"	"	"	"	

+1 Dimer 100mM NaCl	Inf	Inf	Inf	Inf	4.969	4.86	Inf	Inf	Inf	66105.8	548.238	1772.93	95.961	Inf	Inf	Inf	Inf	1.841	0.283	Inf	Inf	Inf	Inf	Inf	
+1 Dimer 20mM NaCl	0.551	1.645	Inf	Inf	1644.82	15.427	Inf	Inf	Inf	27098.6	627.095	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
+1 Dimer 20mM NaCl OmM MgCl2	Inf	Inf	4.639	121.15	Inf	Inf	Inf	Inf	Inf	2507.56	108.761	730.973	367.723	Inf	Inf	Inf	Inf	4.308	0.371	Inf	Inf	Inf	Inf	Inf	Inf
+1 Dimer 300mM NaCl	66.327	533.203	Inf	Inf	Inf	Inf	0.847	1.754	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	118212.6	363.249	556.226	9.913	Inf	Inf	Inf	Inf
-1 Dimer 100mM NaCl	4.266	38.997	1.449	41.228	7757.04	82.706	Inf	Inf	432.642	112459	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
-2 Dimer 100mM NaCl	2.205	8.321	Inf	Inf	0.714	0.425	Inf	Inf	109.372	382.505	34997	78604	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
-2 Dimer 20mM NaCl	0.553	12.931	52475	493.97	179.687	15.628	Inf	Inf	8453.7	1857.8	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
-2 Dimer 20mM NaCl OmM MgCl2	Inf	Inf	10.413	12.722	0.0005	0.0005	87195.31	285.266	4.649	1.168	Inf	Inf	Inf	Inf	Inf	Inf	Inf	75055.88	91691.59	Inf	Inf	Inf	Inf	Inf	Inf
-2 Dimer 300mM NaCl	0.404	3.737	Inf	Inf	539.457	10.864	Inf	Inf	21797	518.401	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf



**Table 7** Optimized kinetic network model parameters for the kinetic rate constants for all positions and conditions studied. Entries with 'Inf' denote connections which did not exist in the optimized network models.

## CHAPTER 5 : APPLICATION OF PS-SMF TO PROTEIN-DNA COMPLEXES

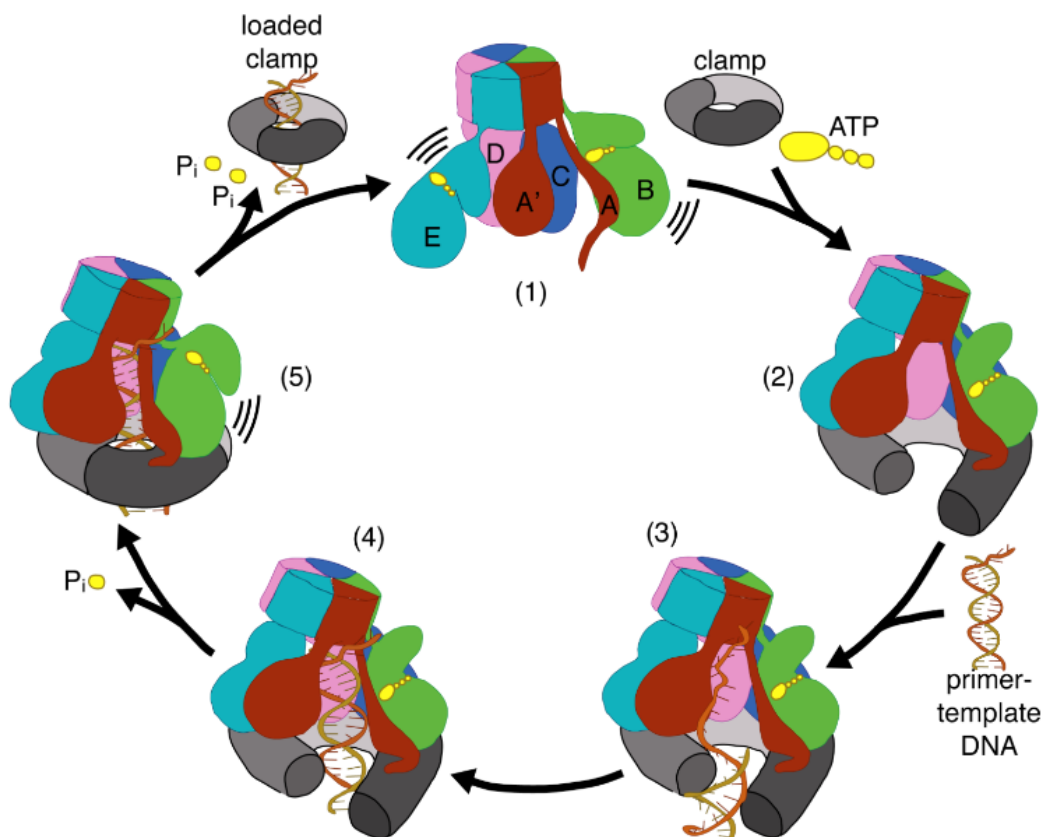
### 1. OVERVIEW

During this thesis work, numerous collaborations were undertaken between the primary author, Jack Maurer, and other members of the Marcus and von Hippel Lab. The biochemically focused collaborations were primarily carried out with Dr. Patrick Herbert and pertinent to the study of key protein-DNA complexes formed during the assembly and function of replisome machinery at ss-dsDNA junctions. The first section of this Chapter aims to introduce the systems studied, detail their biochemical and biophysical importance and then motivate the studies internal to our laboratory, where unique dynamical insights were gained. Some of the information presented here is the topic of a forthcoming research article which at this time is still in preparation for journal submission titled, “*Investigating local DNA conformational trapping dynamics during DNA polymerase holoenzyme assembly and exonuclease proof-reading activity*”. Primarily two protein-DNA complexes were studied by Dr. Patrick Herbert using the PS-SMF method and numerous ensemble approaches. These complexes were the clamp-clamp loader (gp45 clamp, gp44/62 clamp loader) and DNA polymerase holoenzyme (gp43 polymerase).

## 2. INTRODUCTION

The T4 bacteriophage clamp-clamp loader system is a well-studied, biologically relevant, macromolecular machine which can load the sliding clamp (gp45) onto a DNA p/t junction for eventually complexation with DNA polymerase (gp43). Progressive chromosomal replication relies on the presence of the clamp to achieve recognition and nucleotide addition in the DNA polymerase holoenzyme with rates of addition that are orders of magnitudes faster in the presence of the clamp than in the absence of the clamp [91]. To achieve this enormous reduction in the time required to make a copy of DNA, the clamp must first be successfully loaded and positioned at a p/t junction, then complex with DNA polymerase to initiate replication. The clamp itself cannot load onto a DNA strand because it is circular and closed in the absence of the clamp loader. Therefore some conformational rearrangement must first facilitate binding of the clamp-clamp loader complex to DNA, where the clamp is at least partially open to facilitate loading, and then another conformation which sees the clamp close down onto the DNA strand and the clamp-loader depart in preparation for DNA polymerase activity and function. There has been significant effort dedicated to the structure and biochemistry of the clamp-clamp loader system of T4 bacteriophage, as well as the corresponding clamp-clamp loader system of eukaryotes and archaea [91]–[94]. Structural studies have been able to elucidate key conformations in the formation of the clamp-clamp loader complex as well as the fully accommodated clamp onto a DNA template strand. But many of these structural studies have been based on crystallography data and lack the dynamical information needed to report on the kinetics of these processes, while also suffering from the potential bias which crystallization conditions can induce in preferentially producing a particular structure from a slew of possibilities [33]. Similar efforts have been made toward elucidation of the structure and biochemical mechanisms of the DNA polymerase holoenzyme [93]–[97].

The highest quality structure of the T4 bacteriophage clamp-clamp loader system bound to a DNA template was reported in 2011 by Kelch et.al [91]. The insights gained in their study are depicted schematically in Fig. 5.1. The detailed atomistic map produced from their work led to a hypothesized mechanism for the assembly of the clamp, by the clamp loader, onto a DNA template. Step 1 in Fig. 5.1, shows that in the absence of ATP, the clamp loader AAA+ modules cannot organize into a spiral shape. Then in step 2, ATP binding causes the AAA+ modules to form a spiral that can bind and open the clamp. Next, in step 3 the primer-template DNA is threaded through the gaps between the clamp subunits I & III and the clamp loader A and A' domains. In step 4, DNA binding in the interior chamber of the clamp loader, activates ATP hydrolysis. They hypothesize this most likely occurs through flipping of the switch residue and release of the Walker B glutamate seen in their crystal structures. The 5<sup>th</sup> and final step is ATP hydrolysis at the B subunit which breaks the interface at the AAA+ modules of the B and C subunits and allows closure of the clamp around primer-template DNA. Further ATP hydrolyses at the C and D subunits dissolve the symmetric spiral of AAA+ modules, thus ejecting the clamp loader because the recognition of DNA and the clamp is broken. The clamp is now loaded onto primer-template DNA and the clamp loader is free to recycle for another round of clamp loading. This postulated mechanism of assembly and function is elegant and physically conceivable. But the ability to link static crystal structures with highly dynamical processes occurring in a solution phase context is difficult at best. To better understand the mechanisms that underly these critical protein-DNA complexes, Dr. Patrick Herbert performed experiments on these reconstituted systems in vitro, using a variety of experimental methods.

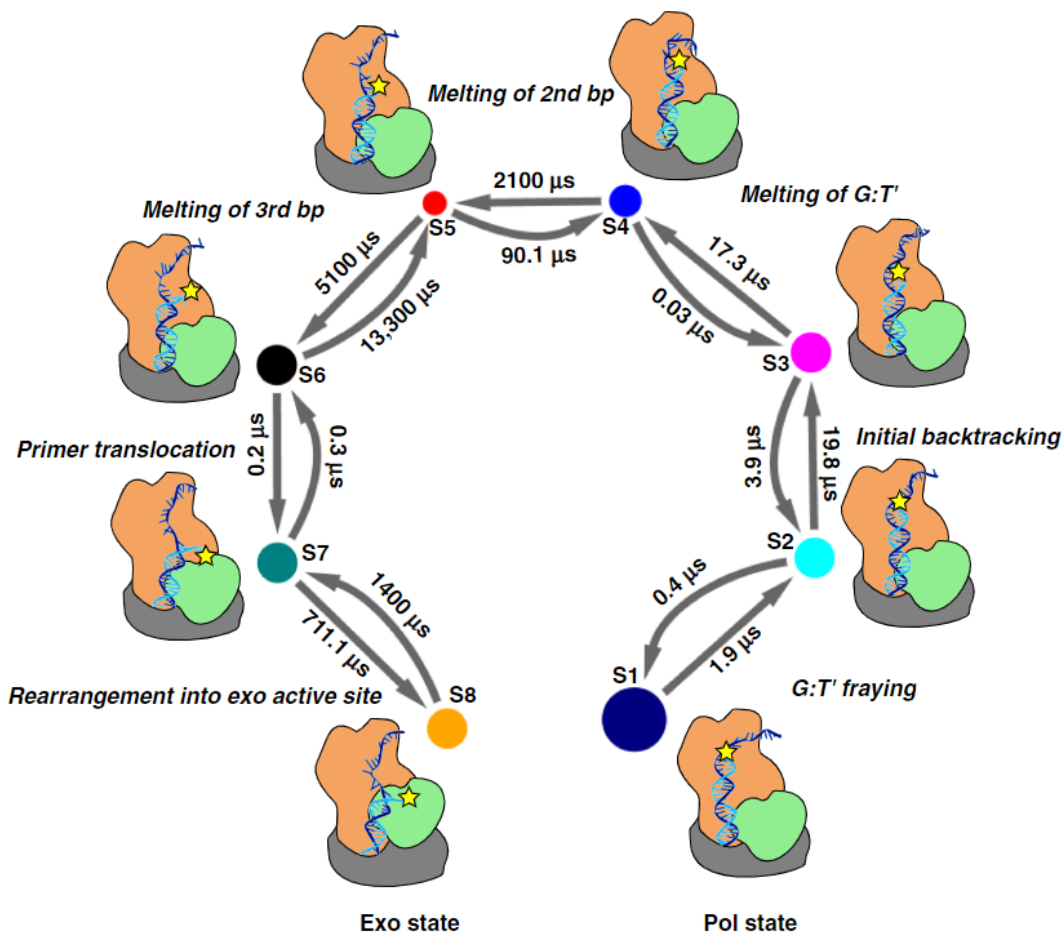


**Figure 5.1** Schematic depiction of the clamp-clamp loader reaction cycle taken from Kelch et. al [91].

But most critically, the recent development of PS-SMF opened the possibility of studying the rapid interconversion of multiple conformational macrostates, local to a DNA p/t junction, with high temporal resolution and exquisite-specificity for probe labeling site. The aim of these studies was to uncover the relevant conformations of DNA p/t junctions which are leveraged by the clamp-clamp loader to achieve binding and function. Ultimately, determining the free energy landscape in the presence of the clamp-clamp loader at a p/t junction yields unique insights that can aide in the discovery of new mechanistic insights and offer support toward previous structural assignments made on static complexes.

The logical next system to study, given the results obtained on a clamp-clamp loader complex, is DNA polymerase bound to a p/t junction. DNA polymerase can polymerize a copy of DNA from a template strand with enormous accuracy and relatively high-speed. This occurs at both the leading and lagging strands of an unwound DNA template, where the leading strand proceeds in a more straightforward fashion while the lagging strand experiences a series of start-stop cycles that leads to the formation of Okazaki fragments. Typically, base pairs are improperly incorporated only 1 time per  $10^5$  to  $10^8$  nucleotide additions leading to an ultra high fidelity in the process of DNA replication. A critical feature of this protein-DNA complex is the known switching behavior between two functional modes when bound to DNA. Where each mode of operation is structural distinct [98], [99]. These are the exonuclease mode, where the polymerase rearranges the DNA to perform an important proofreading and editing step to eliminate possible misincorporation of an incorrect nucleotide, and the polymerase mode, where nucleotides are added to the growing DNA copy. This system is a classic example of a macromolecular machine which leverages conformational rearrangements as the basis for the multitude of functions it performs. Additionally, these conformational rearrangements occur on a very local scale within the DNA template, namely at the level of single nucleotides. Owing to the highly local dynamics of DNA polymerase and the logical extension of the clamp-clamp loader work, Dr. Patrick Herbert also obtained PS-SMF data on DNA p/t junctions bound with DNA polymerase, where it was shown that the degree of conformational changes could be turned by the concentration of magnesium in the local solvent environment. This result agrees with previous studies that examined the degree of reversible exo-to-pol switching dynamics using circular dichroism and base analogue probes [99]. Crystal structures of DNA polymerase have been obtained, but obtaining structures bound to a DNA template in either the exonuclease or polymerase mode

remain elusive. Recent computational studies have offered a hypothetical view of exo-to-pol switching in DNA polymerase, which suggests a kinetic pathway that requires transition times on the order of tens-of-milliseconds to switch between the two configurations (Fig. 5.2) [98].



**Figure 5.2** Proposed kinetic scheme obtained by computational studies of Dodd, et. al.[100] The end-to-end transition is between the exo state and pol state of the bound DNA holoenzyme, which contains the clamp as well as DNA polymerase. Notably, the results suggest a net transition time of ~10ms between the two conformational macrostates.

Both the clamp-clamp loader and DNA polymerase complexes offer an exciting opportunity to obtain detailed kinetic and thermodynamic information on the local conformational changes driving key functions within the replisome machinery. This Chapter focuses on the PS-SMF results obtained on these complexes. These results were analyzed with a large-scale optimization to

determine the kinetic network which best explained the observed thermodynamic behavior. The kinetic network results are largely presented here, drawing occasionally on results obtained by Dr. Patrick Herbert using more traditional ensemble techniques that aid in the explanation and interpretation of single-molecule studies. Single-molecule studies of DNA only samples labeled well within the duplex side of the p/t junctions are also discussed here, as they are foundational to building up an understanding of the conformational landscape that is leveraged by the fully assembled protein-DNA complexes.

### 3. MATERIALS AND METHODS

Table 8 shows the sequences and nomenclature of the internally labeled (iCy3)<sub>2</sub> dimer-labeled p/t-DNA constructs. Oligonucleotide samples were purchased from Integrated DNA Technologies (IDT, Coralville, IA) and used as received. For our absorbance and circular dichroism (CD) measurements, solutions were prepared with sample concentrations of 1 mM and a standard aqueous buffer of 10 mM Tris, 100 mM NaCl, and 6 mM MgCl<sub>2</sub>, unless otherwise stated. We combined complementary oligonucleotide strands to form (Cy3)<sub>2</sub> dimer-labeled p/t-DNA constructs, which contain both ds and ss DNA regions. For ‘duplex’ constructs, the (iCy3)<sub>2</sub> dimer probes were positioned deep within the double-stranded region (at the +15 position) of the p/t-DNA construct. For ‘primer-template’ (p/t) DNA constructs, the (iCy3)<sub>2</sub> dimer probes were positioned at the +4, +3, +2 and +1 positions relative to the ss–dsDNA junction. Prior to the experiments, the sample solutions were annealed by heating to 95°C for 3 minutes before they were allowed to slowly cool to room temperature. The expression and purification methods of gp45, gp44/62, and gp32 proteins have been previously reported [101]. T4 Bacteriophage DNA



Polymerase gp43 was purchased from Molecular Cloning Laboratories (MCLAB, San Francisco, CA).

**Table 8** Base sequences of p/t-DNA constructs used in these studies.

**+1 Construct**

3' CTC CCT CGT GTC GTC CGT TCT TGG CTG G(CY3)T TTG GTT GGT TAG GTC TTG TTT TGC CGG TCA 5'

5' GAG GGA GCA CAG CAG GCA AGA ACC GAC C(CY3)A 3'

**+2 Construct**

3' CTC CCT CGT GTC GTC CGT TCT TGG CTG (CY3)GT TTG GTT GGT TAG GTC TTG TTT TGC CGG TCA 5'

5' GAG GGA GCA CAG CAG GCA AGA ACC GAC (CY3)CA 3'

**+3 Construct**

3' CTC CCT CGT GTC GTC CGT TCT TGG CT(CY3) GGT TTG GTT GGT TAG GTC TTG TTT TGC CGG TCA 5'

5' GAG GGA GCA CAG CAG GCA AGA ACC GA(CY3) CCA 3'

**+4 Construct**

3' CTC CCT CGT GTC GTC CGT TCT TGG C(CY3)T GGT TTG GTT GGT TAG GTC TTG TTT TGC CGG TCA 5'

5' GAG GGA GCA CAG CAG GCA AGA ACC G(CY3)A CCA 3'

**+15 Construct**

3' CTC CCT CGT GTC GT(CY3) CCG TTC TTG GCT GGT TTG GTT GGT TAG GTC TTG TTT TGC CGG TCA 5'

5' GAG GGA GCA CAG CA(CY3) GGC AAG AAC CGA CCA 3'

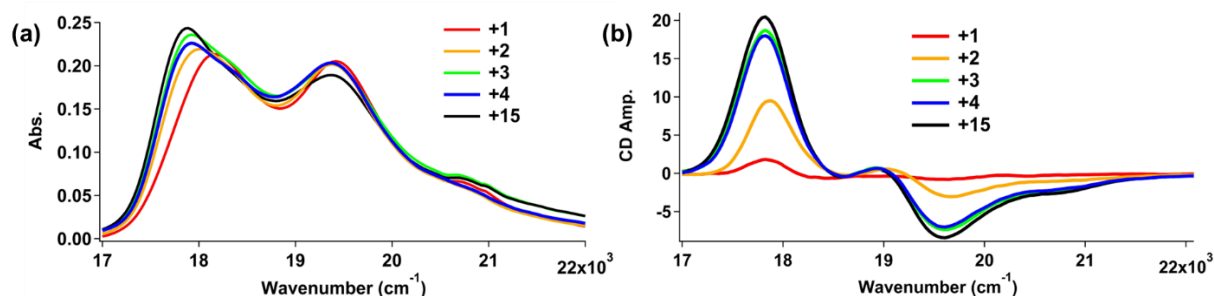
All sample chamber preparation, incubation times and experimental PS-SMF protocols are the same as outlined in Chapter 3 for DNA only studies of replication fork junctions at the +1, -1 and -2 positions. All ensemble data referenced in this Chapter, primarily linear absorbance, and circular dichroism (CD) data, have experimental protocols which have been well documented previously [24], [25], [68].

## 4. RESULTS AND DISCUSSION

### Section 4.1 Duplex region studies of DNA p/t junctions

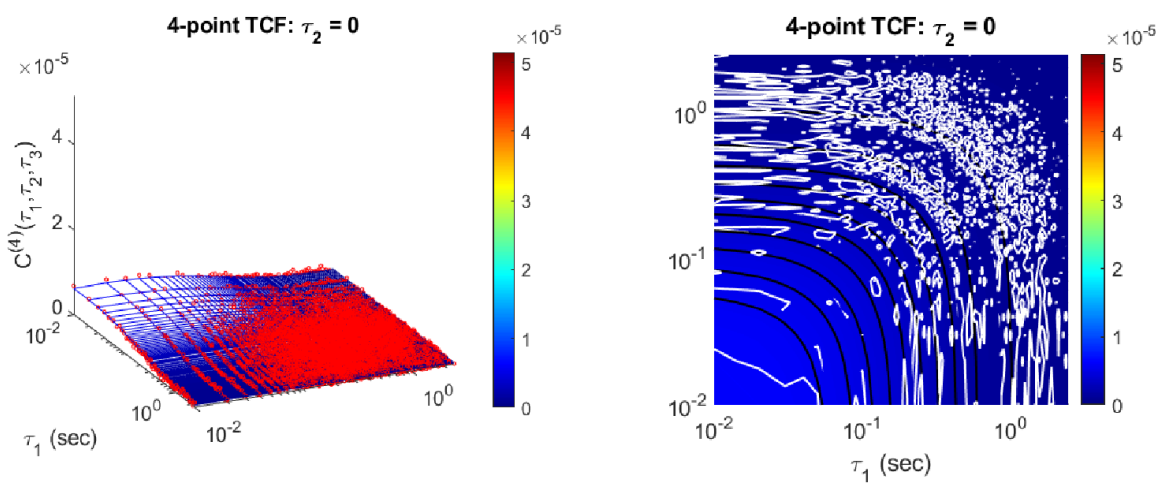
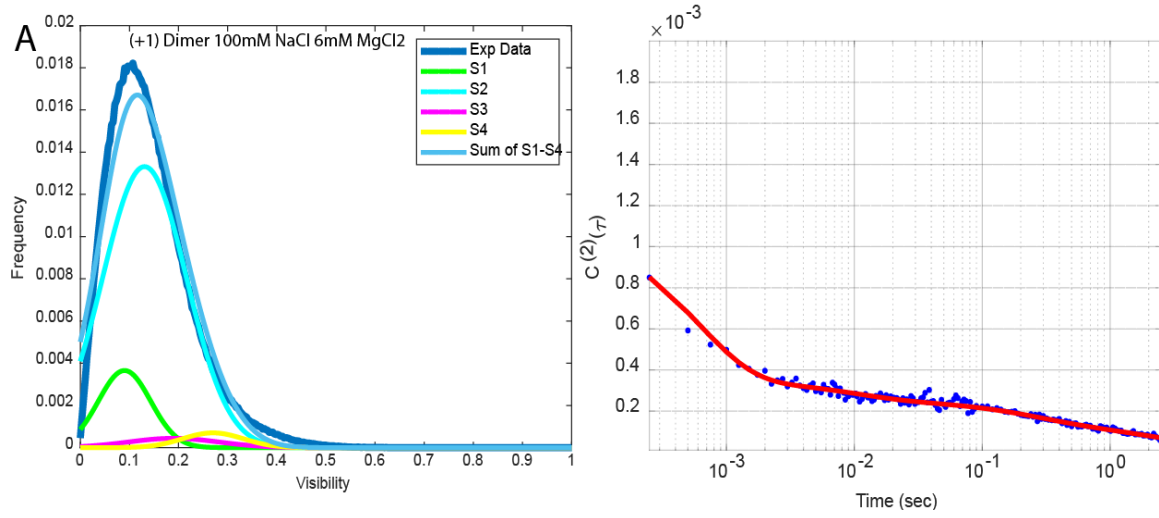
The results of DNA only studies are presented first as the basis for later assignments and inferences in the presence of bound protein assemblies. DNA constructs labeled with (iCy3)<sub>2</sub> dimer probes at the +1, +2, +3, +4 and +15 position were studied using PS-SMF to obtain thermodynamic, kinetic parameters and construct free energy landscapes to report on the conformational mechanisms within the duplex region of p/t junctions. The corresponding ensemble studies on these constructs were also performed and are shown below in Fig. 5.3, panels A and B.

It can be immediately appreciated from these ensemble studies that the +1 and +2 positions are unique compared to the rest of the duplex positions examined. The steady drop in the CD signal moving from +3 to +1 suggests that the conformational freedom of the sugar-phosphate backbone must be steadily increasing as the junction is approached.

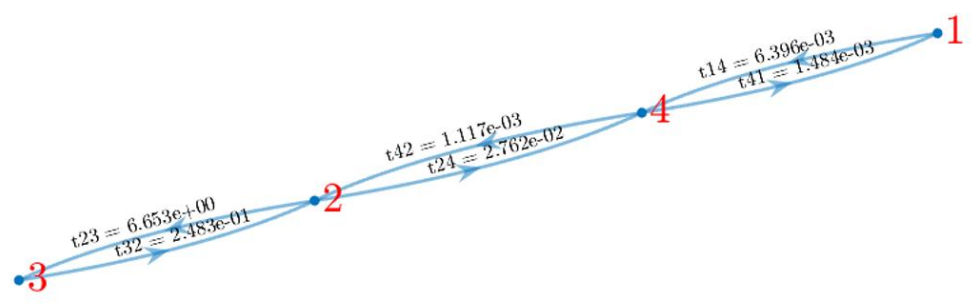


**Figure 5.3** Linear absorbance (a) and CD (b) spectra of +1 (red), +2, (orange), +3 (green), +4 (blue) and +15 (black) constructs.

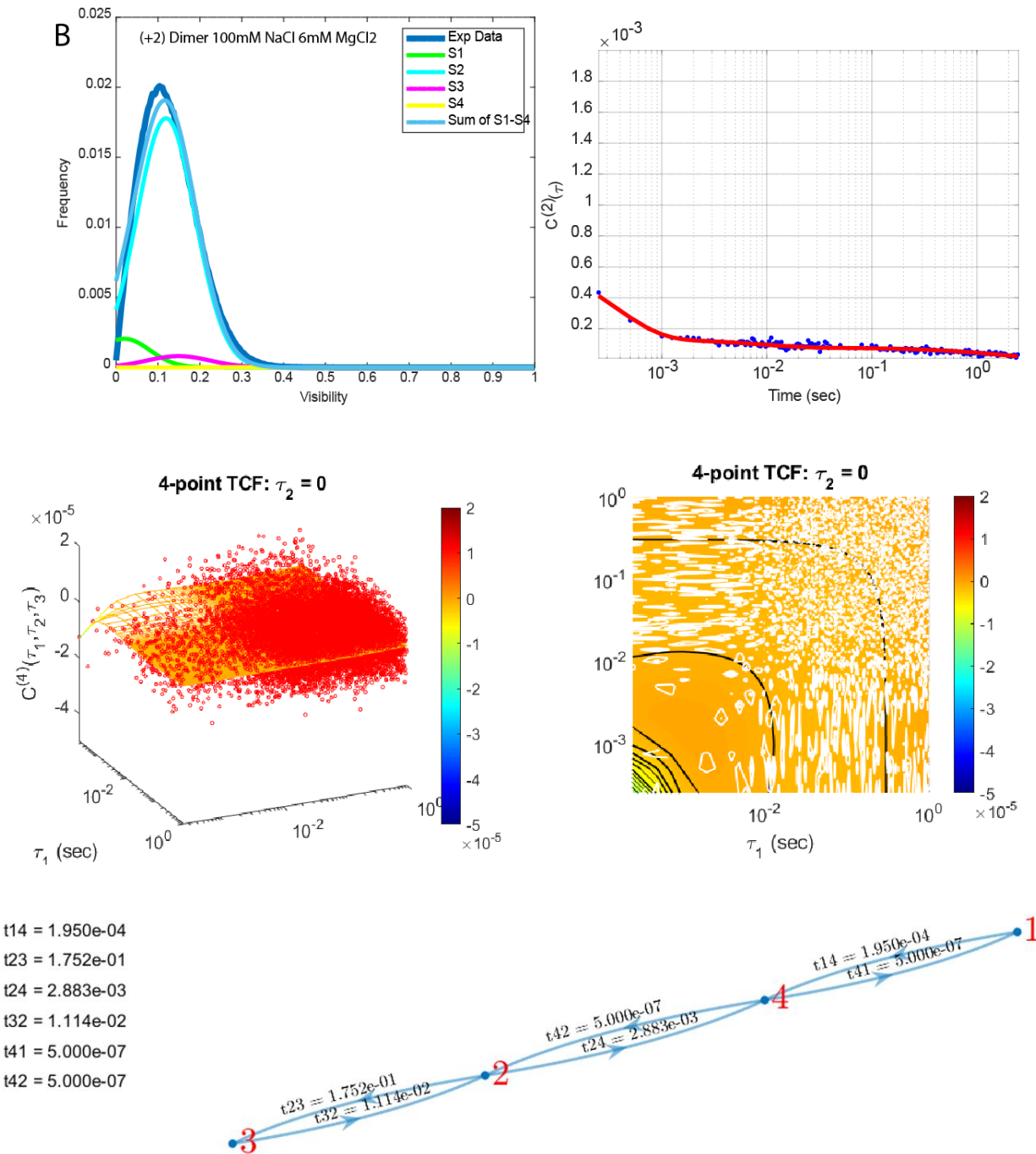
Additionally, the clear shift in the linear absorbance signal from a strongly coupled right-handed signature at the +15 and +4 positions to a less coupled signature at the +1 and +2 positions suggests an increase in dynamics and decrease in relative stability as the junction is approached from the duplex side of these constructs. This inference in the absorbance spectra stems from similarities of the +2 and +1 position to the -1 position in previous ensemble studies, which is known to be uniquely disordered [1], [24]. The results for DNA only PS-SMF experiments, with optimized fits overlaid, for all positions studied, are given in Figure 5.4, panels A-E. Network topologies and kinetic rate constants are shown below each set of optimized fits.



$t_{14} = 6.396e-03$   
 $t_{23} = 6.653e+00$   
 $t_{24} = 2.762e-02$   
 $t_{32} = 2.483e-01$   
 $t_{41} = 1.484e-03$   
 $t_{42} = 1.117e-03$



**Figure 5.4 Panel A:** Experimental data and optimized fits for the +1 dimer (sequence given in Table 8). PDF, C2, C3, and kinetic network shown.



**Figure 5.4 Panel B:** Experimental data and optimized fits for the +2 dimer (sequence given in Table 8). PDF, C2, C3, and kinetic network shown.

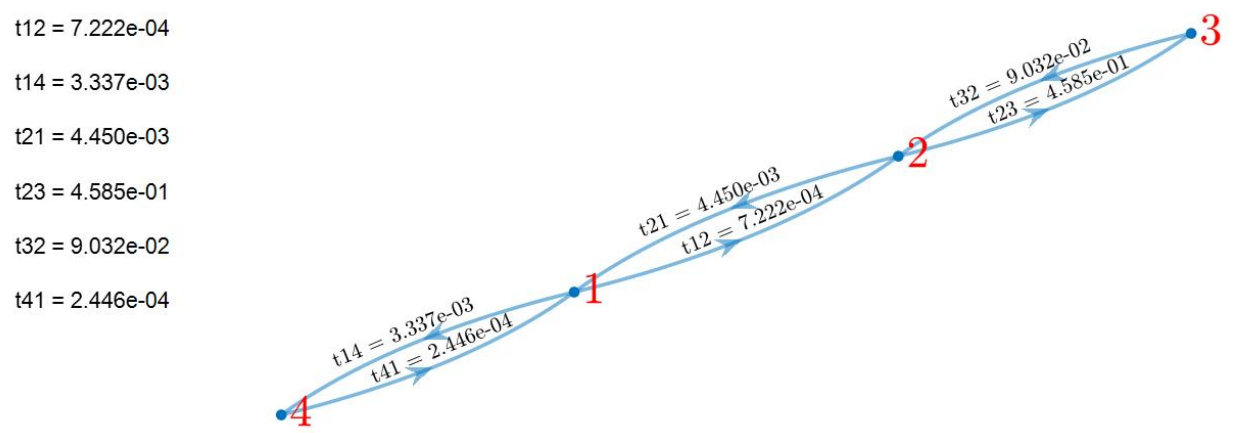
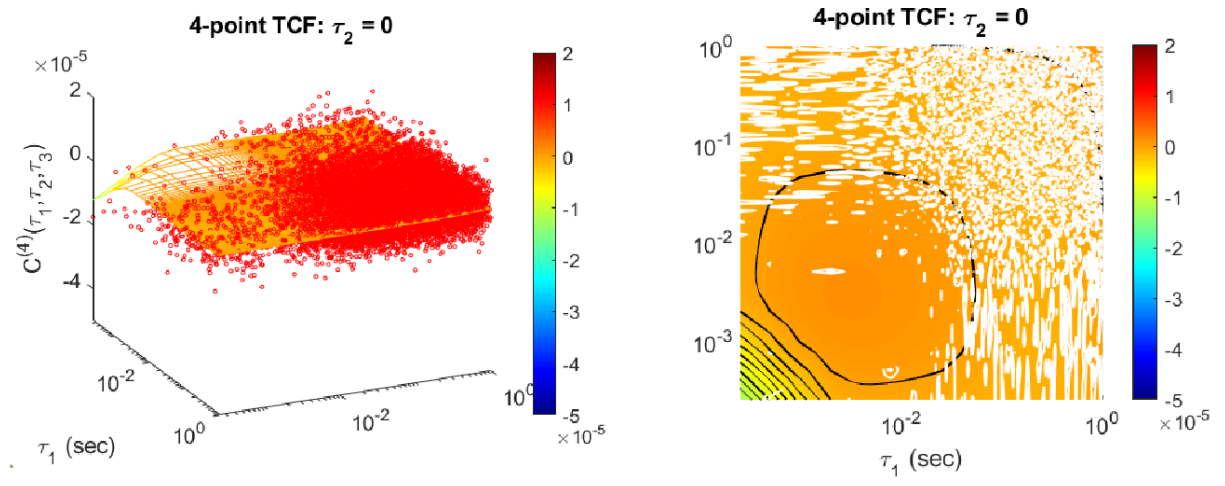
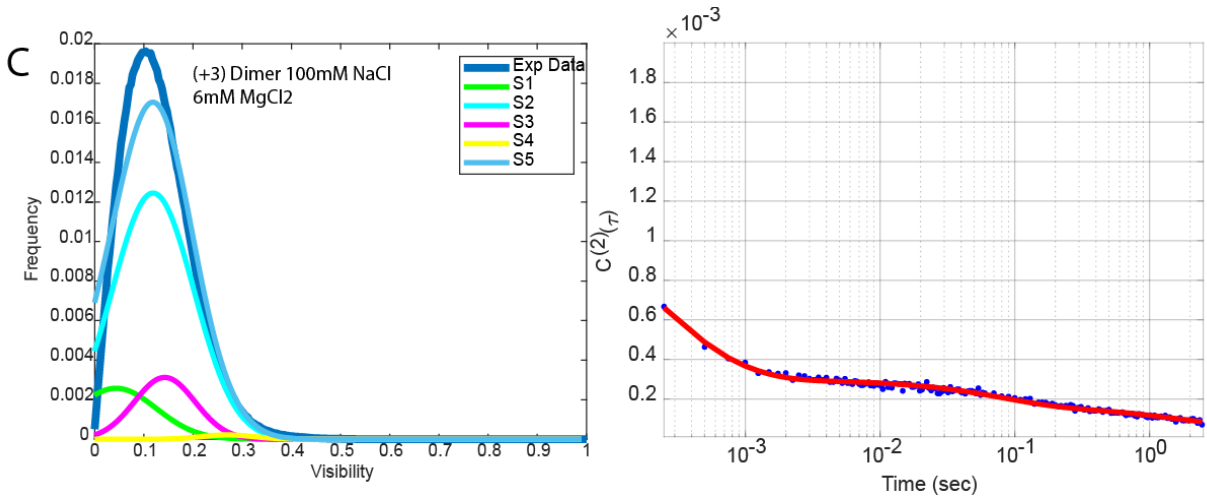
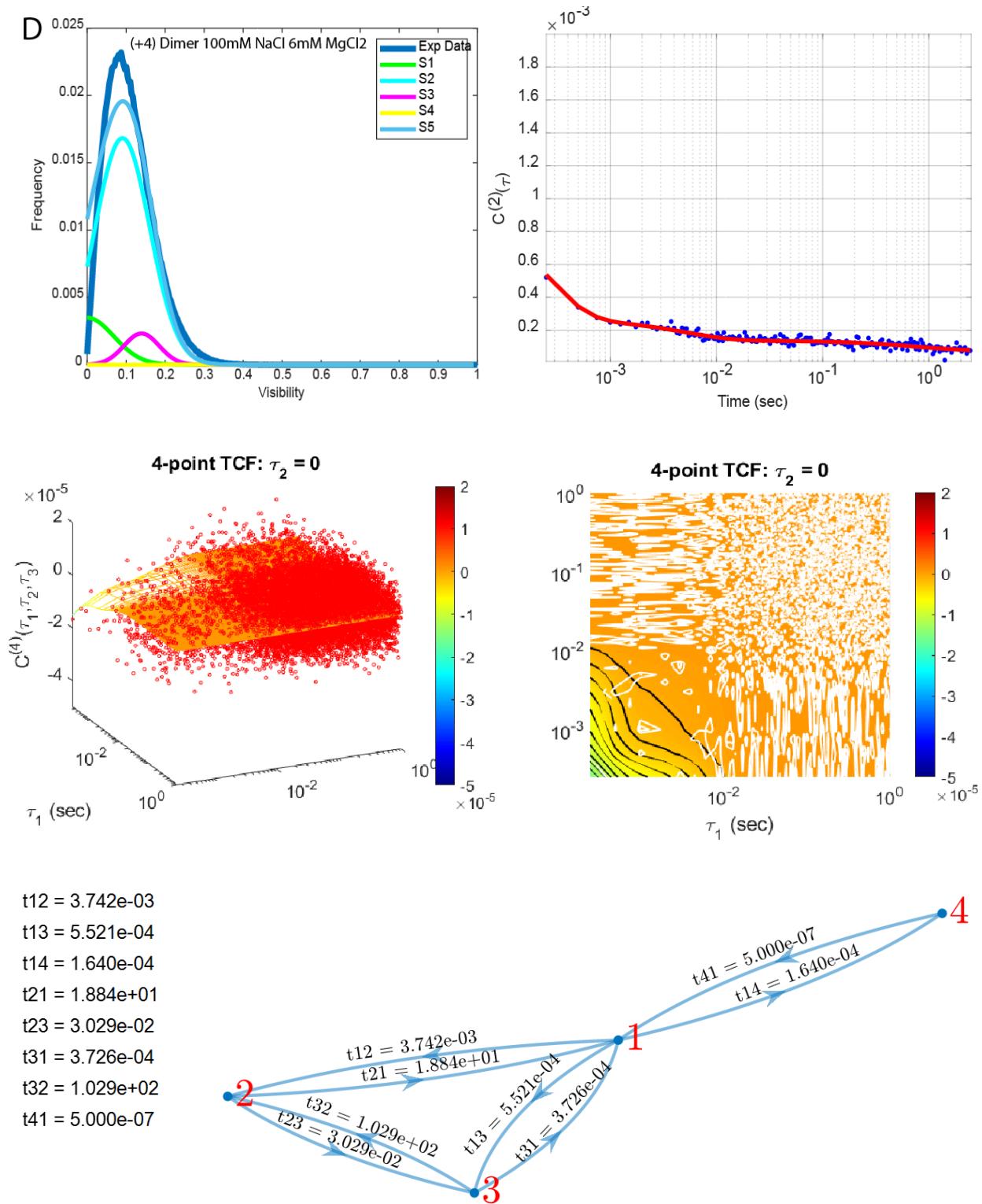
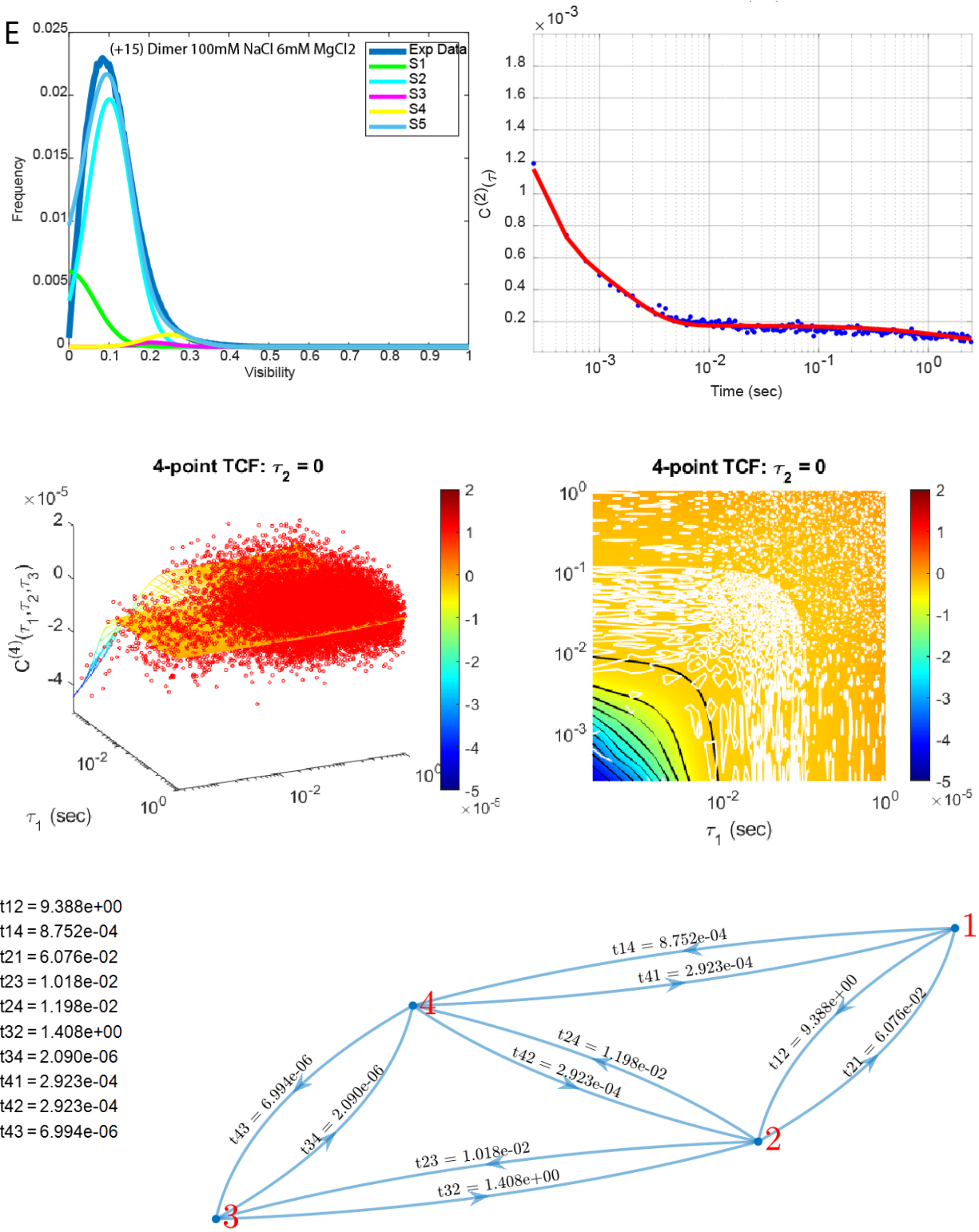


Figure 5.4 Panel C: Experimental data and optimized fits for the +3 dimer (sequence given in Table 8). PDF, C2, C3, and kinetic network shown.



**Figure 5.4 Panel D:** Experimental data and optimized fits for the +4 dimer (sequence given in Table 8). PDF, C2, C3, and kinetic network shown.



**Figure 5.4 Panel E:** Experimental data and optimized fits for the +15 dimer (sequence given in Table 8). PDF, C2, C3, and kinetic network shown.



From the results of these optimized fits, we can construct free energy landscapes and begin to make inferences about the potential structure based on considerations of thermodynamic and mechanical stability. The structural assignments inferred in Chapter 4 will be drawn upon here to make similar conclusions about the conformational macrostates which underlie the visibility assignments in our PS-SMF measurements.

Starting at the +1 position, we see a relatively large sub population of  $S_1$ ,  $S_3$  and  $S_4$  relative to  $S_2$ . We assign  $S_2$  to B-form DNA, as was done in Chapter 4 and is consistent with the CD results shown for this construct. Comparison of the +1 position in this Chapter to the +1 position examined in Chapter 4 reveals a key difference between them, which is the propensity for  $S_1$  versus  $S_3$ . The +1 position in Chapter 4 has a terminal GC base-pair directly at the junction, versus a terminal AT base pair in the dimer of Chapter 5. The extra hydrogen bond of a GC base-pair is indicative of a greater propensity to maintain complementary base-pairing. As such, the lower population of  $S_1$ , the propped-open state due to loss of complementary hydrogen bonding, is more probable with a terminal AT base-pair than a GC base-pair. The relative stability of  $S_3$  is lower for the dimer with a terminal AT base-pair than for the GC dimer. This is likely due to the base-stacking interactions further inside the duplex which act to stabilize most right-handed conformations in the presently considered case versus the case considered in Chapter 4, given the assignment of  $S_3$  to a left-handed conformer. This hypothesis agrees with the increased probability of  $S_4$  in the presently considered +1 dimer, since this state was previously attributed to a right-handed, bases-unstacked conformation. The dynamics at the +1 position are the slowest seen in all DNA only studies of the p/t junctions considered in this Chapter. The +1 position in Chapter 4 also displayed the overall slowest dynamics compared to positions within the vicinity of a replication fork junction (-1 and -

2). This general trend seen at the +1 suggests a unique free energy landscape governs conformational fluctuations right at the site of ss-dsDNA junctions.

Examination of the +2 position results shows that most of the probability still resides in  $S_2$ , the right-handed B-form DNA conformation, consistent with ensemble results. Interestingly the relative probability of  $S_1$ ,  $S_3$  and  $S_4$  are all reduced compared to the probabilities seen in the duplex at the +3, +4 and +15 positions. This result is consistent with the observed dynamics, since the +2 displays the most rapid loss of correlation in both the two- and three-point TCF compared to the surrounding positions. The +2 position is also the only dimer labeled position flanked by GC base-pairs at both nearest neighbor sites. The inability to resolve very rapid interconversion events occurring at the +2 is likely the reason for this apparent greater stability of  $S_2$  at the +2 position compared to the +3, +4 and +15, all of which have stronger right-handed CD signals. Notably, the optimized network topology for the +1 and +2 positions were identical in the presently considered data, despite a blind search of the complete network topology space, suggesting some conservation in the allowed transitions of these conformational fluctuations near p/t junctions. Also notable is the conserved trend in the persistent connection between states  $S_2$  and  $S_3$  as well as  $S_1$  and  $S_4$  in previous and current kinetic networks results at both replication fork and p/t junctions.

The results of our kinetic network analysis at the +3 and +4 position are generally consistent with the emerging picture of conformational dynamics in the duplex region of p/t junctions. State  $S_2$  is thermodynamically most stable, while  $S_1$  and  $S_3$  are relatively close in terms of their respective probability, while  $S_4$  remains least stable. The local sequence context of the +3 and +4 positions is quite similar (see Table 8), suggesting that the observed dynamics ought to be similar as well as the propensity for particular conformations to occur. This appears to be the case based on our kinetic network modeling, which suggests a highly similar picture for the +3 and +4

position in terms of the thermodynamic stability dictating the free energy landscapes at these local sites. This is also in agreement with CD and linear absorbance data on these constructs which suggest the tightest agreement between the +3 and +4 position relative to all other pairwise comparisons (Fig. 5.3).

Lastly, the analysis results at the +15 position show a high probability for  $S_2$  relative to all other states, completing the retention of a B-form DNA conformer as the predominant structure at all duplex positions, as the ensemble data would suggest is the case. The probability of  $S_1$  is greatest at the +15 position, other than  $S_2$ , suggesting that the locally propped-open conformation is available deep in the duplex to a greater extent than either of the oppositely handed unstacked base conformations,  $S_3$  and  $S_4$ . The dynamics are also very rapid at the +15 position, evidenced by the fast component in the C2 and significant negative decay at 250usec in the C3. This aligns well with previous hypotheses that the rapid loss and gain of complementary hydrogen-bonding, consistent with a locally propped-open state, is occurring on the microseconds timescale in duplex DNA [1]–[3].

Examining the mechanical stability of these thermally populated conformational macrostates can offer valuable insights into the important mechanistic pathways which yield the observed kinetics. To do this, we will make use of the same approach developed in Chapter 4, where the mechanical stability of each macrostate is discussed in terms of the lowest transition barrier connecting it to any other macrostate in the kinetic network. The mechanical stability of each macrostate can be inferred based on the fastest kinetic rate driving transitions away from the macrostate under consideration. All rates and connections are shown in the kinetic network diagrams of Fig. 5.4 panels A-E, and Table 10 at the end of this Chapter.

Starting at the +1 position, the mechanical stability of the propped-open macrostate  $S_1$  is relatively low and determined exclusively by its connection to  $S_4$ . The mechanical stability of B-form DNA  $S_2$ , the thermodynamically most stable macrostate throughout the entire duplex, is determined by transitions to  $S_4$ , which are an order of magnitude faster than transitions toward  $S_3$  in the presently consider +1 positional data. The presence of a terminal AT base-pair could offer some insight into the relative propensity for  $S_2$  to prefer fluctuations toward a bases-unstacked right-handed structure of  $S_4$  which is then able to be propped-open into  $S_1$ . Whereas previous studies at the +1 position, where the terminal base-pair was GC, tended to prefer the left-handed  $S_3$  macrostate over the right-handed  $S_4$  or  $S_1$  macrostates. The mechanical stability of  $S_3$  in the presently considered data is slightly higher than that of  $S_4$ , but determined exclusively by its transition back toward  $S_2$ .

The relative mechanical stabilities of all four conformational macrostates at the +2 position are quite similar to the +1 position, as is the network topology. The key difference in the two kinetic networks is the reduced mechanical stability of  $S_4$  at the +1 position, which makes very rapid transitions to either  $S_2$  or  $S_1$ , causing the relative abundance of  $S_4$  in the PDF to fall off moving from +1 to +2 in the duplex. The mechanical stability of  $S_2$  is also slightly lower at the +2 position but given the decreased mechanical stability of all non-majority conformers, the relative abundance of  $S_2$  isn't altered much.

Moving deeper into the duplex to the +3, +4 and +15 positions, the network topology rearranges slightly compared to the +1 and +2 positions, yielding a new core set of connections between all four macrostates in the system. This shift going from +1/+2 toward deeper duplex sites is seemingly supported by the ensemble CD and absorbance data showing a systematic difference between the +1/+2 and all sites deeper into the duplex (Fig. 5.3). The mechanical stability of  $S_2$  is

carefully balanced by its transitions to both  $S_1$  and  $S_3$  at the +3 and +4 positions.  $S_1$  has low mechanical stability in general, with transitions toward either  $S_4$  or back to  $S_2$  being preferable at both the +3 and +4 positions.  $S_4$  is mechanically unstable at both the +3 and +4 positions and is dominated by transitions to  $S_1$ .

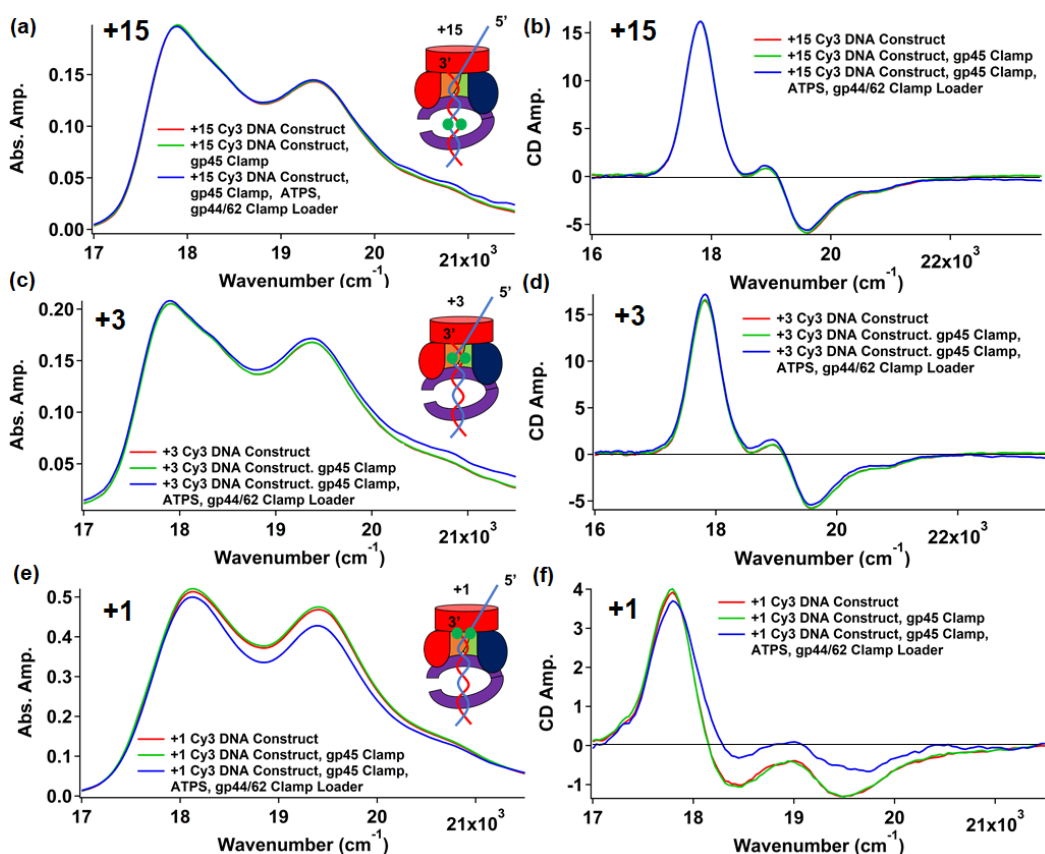
At the +15 position, the mechanical stability of  $S_2$  is carefully balanced by transitions toward all 3 remaining macrostates, with subtle differences distinguishing the predominant pathway for a fluctuation to occur. The mechanical stability of  $S_1$  is dominated by transitions toward  $S_4$ , similar to  $S_3$  which also prefers transitions toward  $S_4$  over an immediate return to  $S_2$ . Fluctuations out of  $S_4$  occur very rapidly with  $S_3$ , while transitions toward  $S_2$  and  $S_1$  are relatively balanced. Notably, the mechanical stability of  $S_2$  is significantly greater than all 3 remaining macrostates at the +15 position, as would be expected for B-form DNA deep inside the duplex. The analysis of these DNA only data allows us to now consider the effects of protein binding events to p/t junctions, as the free energy landscape is rearranged by the functional action of such macromolecular machines.

## **Section 4.2 Protein-DNA complexes studies at and near DNA p/t junctions**

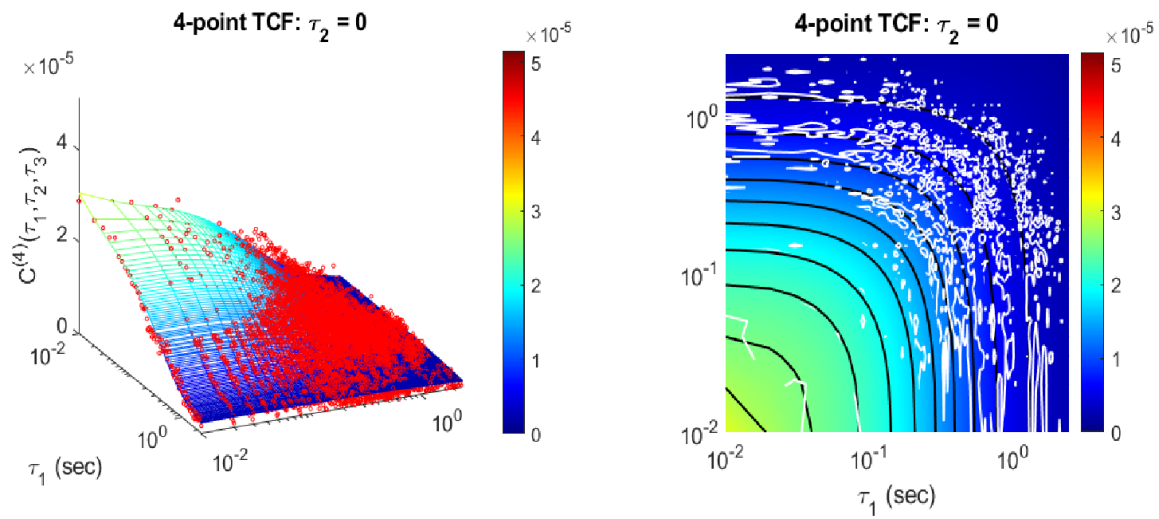
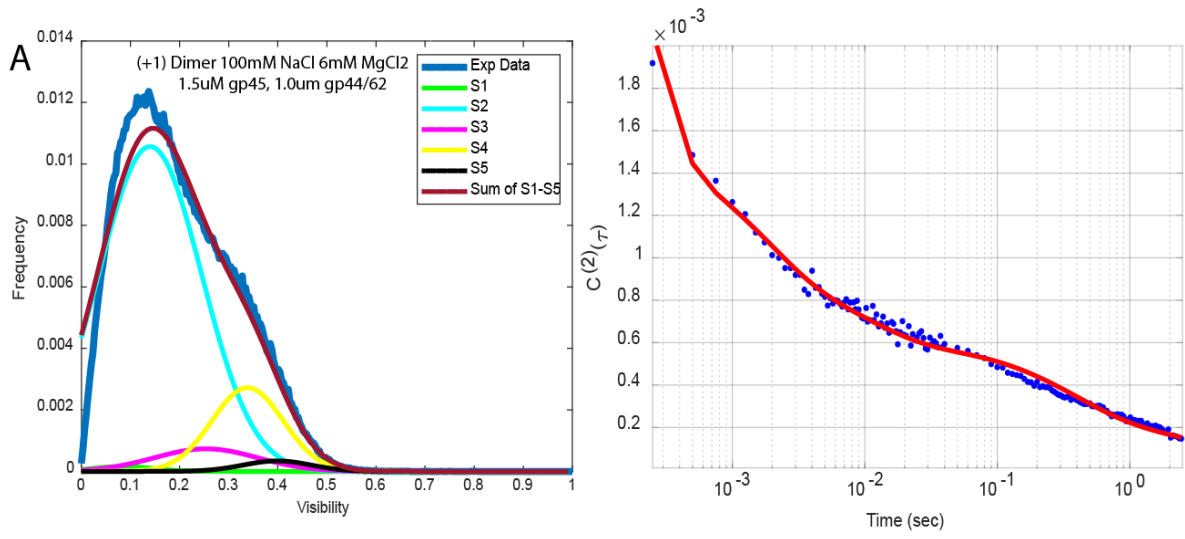
### **Section 4.2.1 Clamp-Clamp Loader Studies**

The PS-SMF experimental results for DNA complexed with the clamp-clamp loader, with optimized fits overlaid, are given in Fig. 5.6. Linear absorption and circular dichroism spectra of the clamp-clamp loader protein-DNA complex is shown in Fig. 5.5. Examination of the ensemble data for the clamp-clamp loader shows that the +1 position is uniquely affected by binding of the clamp, whereas the +3 and +15 position are weakly perturbed. Notably, the handedness of the

ensemble remains right-handed at the +1 position in the presence of the clamp loader, suggesting that B-form DNA and other such right-handed conformers remain the dominant species. For these reasons, only the +1 position was investigated in the presence of the clamp-clamp loader complex, given its unique response to binding. ATP $\gamma$ s was added to the experimental samples for the clamp-clamp loader, as this is known to activate the clamp-loader and enable binding of the clamp, but not allow the clamp-loader to fully dissociate at normal rates due to the non-hydrolyzable nature of ATP $\gamma$ s. This ultimately enables dynamical studies of the stalled complex at a p/t junction. The clear shift in the CD spectra upon titration of ATP $\gamma$ s suggests that the fully assembled complex is uniquely forming at the +1 position only when the clamp-loader is activated.



**Figure 5.5** Linear absorbance and CD spectra of +15 (a, b), +3 (c, d), and +1 (e, f) constructs (red) upon the addition of gp45 clamp (green) and activated gp44/62 clamp loader (blue).



t15 = 2.484e-03  
t23 = 1.000e+01  
t24 = 7.452e-04  
t25 = 3.024e-01  
t32 = 1.030e+01  
t34 = 3.808e-01  
t42 = 1.436e-04  
t43 = 1.919e+00  
t51 = 6.371e-03  
t52 = 6.926e-03

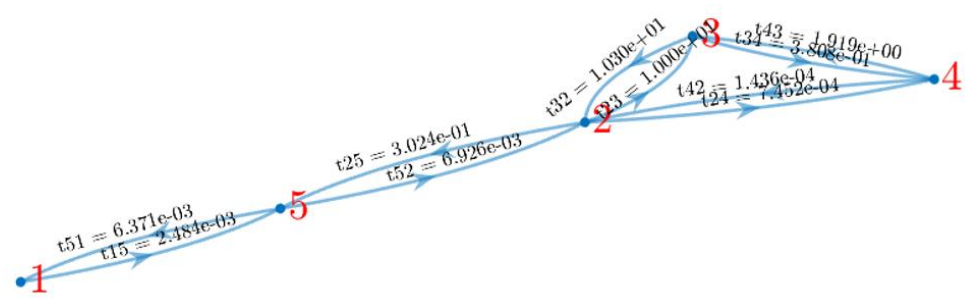


Figure 5.6 Experimental data and optimized fits for the +1 dimer and clamp-clamp loader complex. PDF, C2, C3, and kinetic network shown.

The results of our kinetic network analysis for the +1 position in complex with the clamp-clamp loader demanded a minimum of five conformational macrostates to explain the observed dynamics. Much like high monovalent salt conditions at the +1 position, discussed in Chapter 4, demanded an additional conformational macrostate to explain the observed dynamics. The thermodynamically most stable state upon clamp-clamp loader binding is still  $S_2$ . It is known that the open clamp adopts a spiral geometry to accommodate the native helical structure of B-form DNA, suggesting that initial loading of the clamp to a p/t junction should favor the native B-form structure, which our results also suggest [91], [92], [102], [103]. The overall thermodynamic stability of  $S_1$  is lowered at the +1 position by the presence of the clamp-clamp loader. It has been hypothesized, drawing on crystal structures, that ionic interactions within the central channel of the clamp help to stabilize the interaction between charged regions of DNA and the charged interior of the clamp pore [91], [92]. In this case, the ability of the DNA to be propped-open at the junction is likely reduced, because of the steric constraint placed on the newly encapsulated DNA junction in the presence of the clamp-clamp loader. The thermodynamic stability of  $S_3$  remains close to the +1 position DNA only results, suggesting that the clamp-clamp loader does little to leverage or deplete the interactions leading to formation of a left-handed conformer during its assembly. Most notably, the thermodynamic stability of  $S_4$  is significantly increased in the presence of the clamp-clamp loader. Given this increased stability of  $S_4$ , it is likely that this conformational macrostate plays an important role in the formation of the fully bound protein-DNA complex. We have previously assigned  $S_4$  to a right-handed, bases-unstacked conformation, consistent with observed trends in salt-dependent DNA only studies and indicative of a more parallel geometry of the sugar-phosphate backbone, as would be required for a high visibility observation (Eq. 15). Interestingly, previous crystallographic studies have shown that p/t DNA junctions must be thread through the

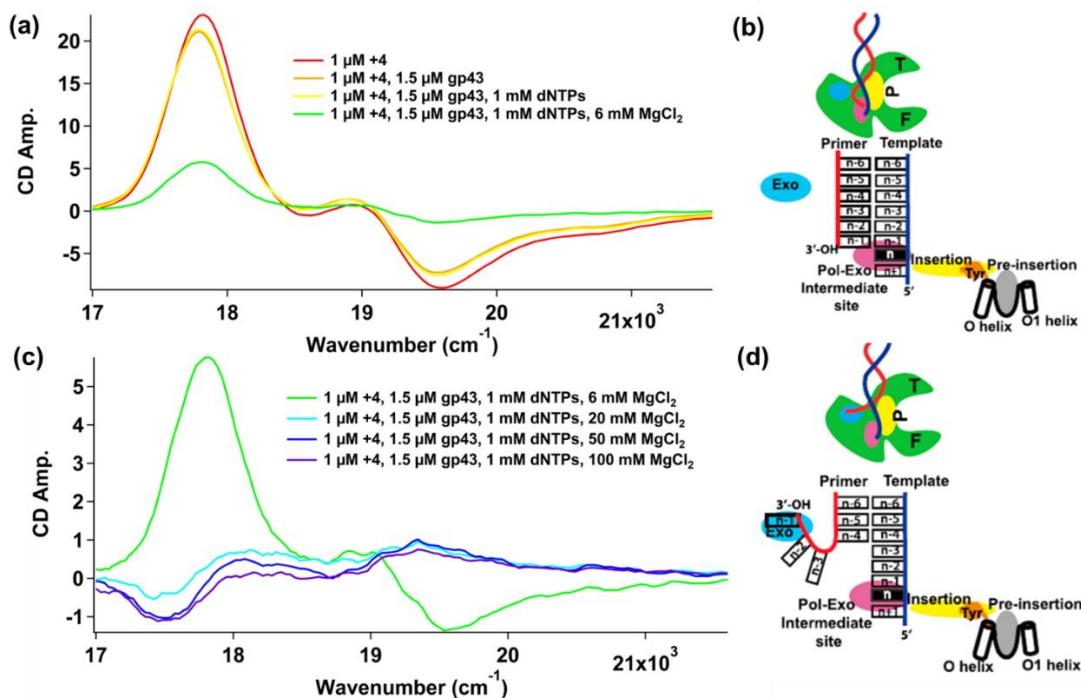


clamp subunits I & III and the clamp loader A and A' domains [91] (see Fig. 5.1). This gap in the two-proteins forming the complex is relatively small compared to the diameter of B-form DNA. We hypothesize that the enriched signal from  $S_4$  in our PS-SMF experiments, consistent with a right-handed bases unstacked conformation, could be the conformation selected for in the accommodation of DNA into the clamp-clamp loader complex. Such a conformation would have its sugar-phosphate backbone extended along the central axis of the DNA to a greater extent than B-form, possibly leading to a reduction in the overall diameter of the double helix at the +1 position, such that loading could occur through the clamp subunits I & III and the clamp loader A and A' domains. The fifth macrostate  $S_5$ , has an appreciable thermodynamic stability on the order of  $S_3$ . This higher visibility state could also be involved in the loading process of DNA into the protein complex, as it further corroborates the necessary elongation of the DNA double helical axis and accompanied narrowing of DNA's diameter to fit into the protein complex. Alternatively, the emergence of a fifth macrostate could be due to the slower dynamics in the presence of the clamp-clamp loader, which much like the elevated monovalent salt conditions at the +1 position of Chapter 4, lead to longer dwell times in the  $S_5$  conformational macrostate, such that it falls within the achievable temporal resolution of our PS-SMF experiments. It is clear from the redistribution of the free energy landscape at the +1 position in the presence of the clamp-clamp loader that  $S_4$  plays an important role in the assembly and function of the overall complex and that  $S_2$ , native B-form DNA, remains the most thermodynamically stable, owing to the spiral geometry the clamp is known to adopt upon binding and activation of the clamp loader.

Examination of the mechanical stability of the clamp-clamp loader complex offers unique mechanistic insights. The mechanical stability of B-form DNA,  $S_2$ , is dominated by transitions to  $S_4$ , the same macrostate which is enriched and seemingly leveraged by complex formation. The

mechanical stability of  $S_4$  is similarly dominated by its transition to  $S_2$ , demonstrating that the free energy landscape is dominated by transitions between these two conformational macrostates, evidenced by the relative height of all remaining transition barriers given in Table 9. This rapid interconversion between  $S_2$  and  $S_4$  supports the notion that  $S_4$  could be the p/t junction conformation that the clamp-clamp loader selects for as it tries to dynamically load the DNA into the central pore. Since the complex is stalled and binding/unbinding transiently, the lower mechanical stability of both  $S_2$  and  $S_4$  could be reflecting this selection mechanism during binding/unbinding events. The mechanical stability of  $S_3$  is relatively high compared to other macrostates. It could be that this left-handed state forms rarely due to the influence of the clamp-clamp loader, but intermolecular contacts tend to stabilize it upon formation. Alternatively, the enthalpy-entropy compensation of forming  $S_3$  could be uniquely stabilizing, despite high transition barriers into  $S_3$ . The mechanical stability of  $S_5$  is closely balanced by transitions to either  $S_1$  or  $S_2$ . Notably, since  $S_5$  is a high visibility state like  $S_4$ , its conformation could be similar to  $S_4$ , but not proper for binding of the clamp-clamp loader. Interestingly, a similar connection between high visibility states in DNA polymerase data exists, suggesting a productive versus unproductive route to protein assembly. The mechanical stability of  $S_1$  is low and determined exclusively by its transition back to  $S_5$ .

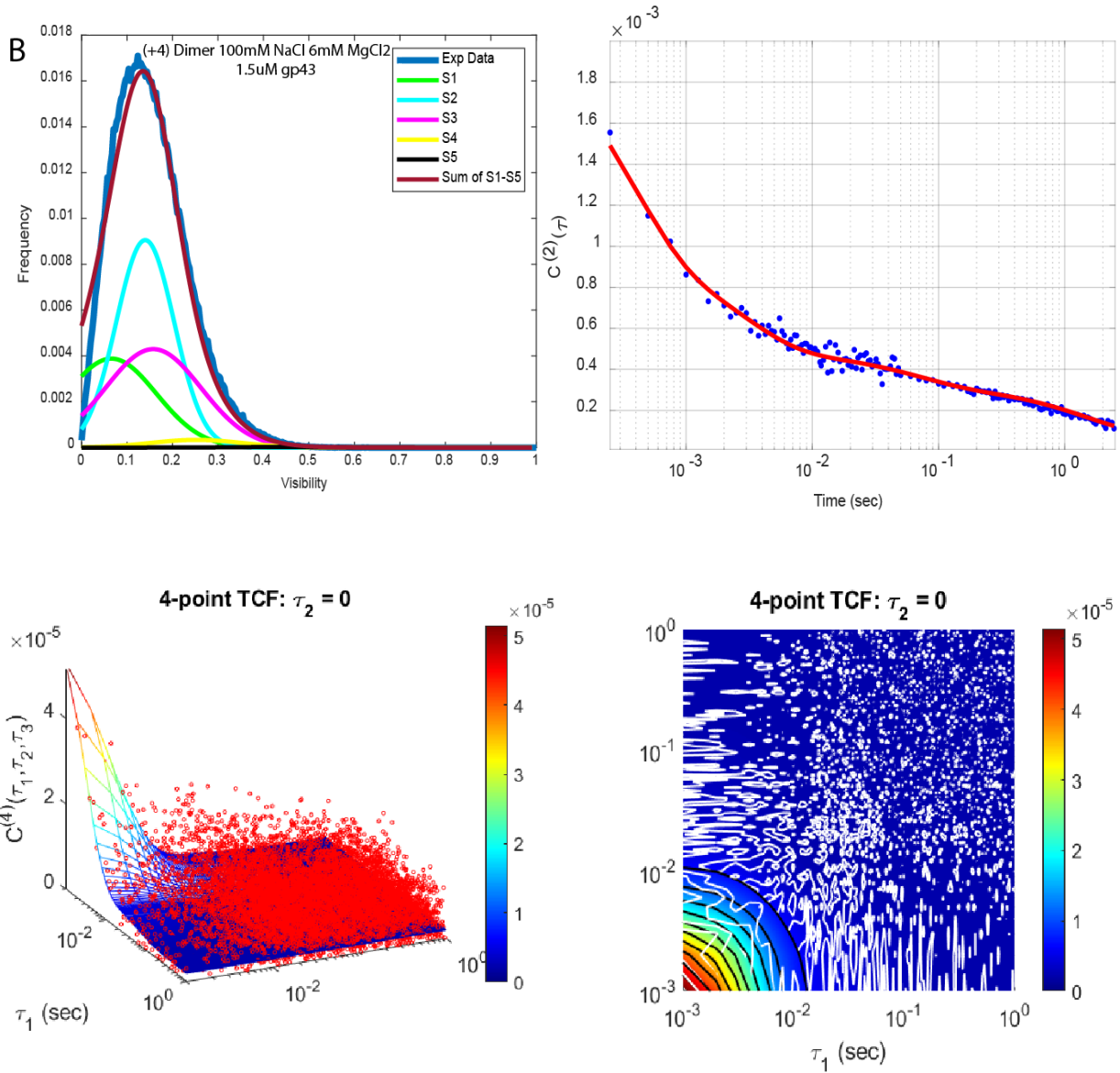
## Section 4.2.2 DNA polymerase studies

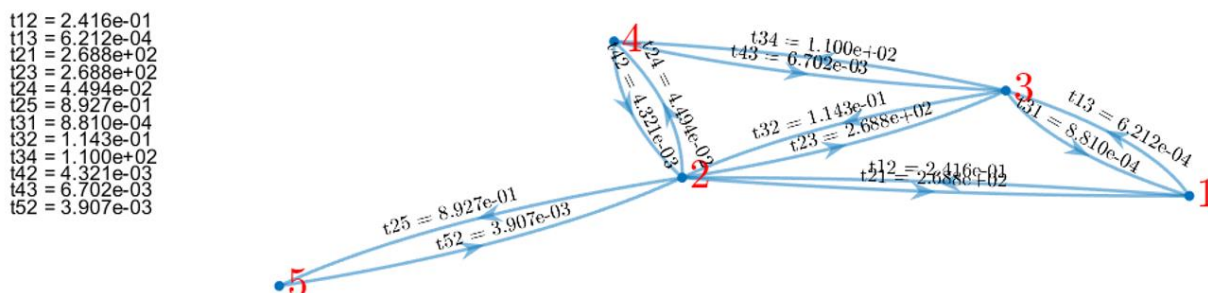


**Figure 5.7** (a) CD spectra of gp43 binding to +4 p/t construct with at 0  $\text{Mg}^{2+}$  concentration. Upon addition of physiological  $\text{Mg}^{2+}$  (6 mM), gp43 becomes active (green trace). (b) Further increasing  $\text{Mg}^{2+}$  concentration results in predominately left-handed conformation at +4 position.

Ensemble circular dichroism data for the +4 position upon addition of DNA polymerase (gp43) and dNTPs, with increasing levels of titrated magnesium, are shown in Fig. 5.7. The +4 position is strongly right-handed in the DNA case. The titration of DNA polymerase and dNTPs initially causes a very slight shift in the CD spectra, suggesting some weak association. However, the addition of bivalent salt in the form of magnesium causes the CD spectra to sharply respond and eventually invert to a left-handed conformation at highly elevated levels of magnesium. This sort of interconversion in DNA polymerase between two oppositely handed conformers is reminiscent of trends observed in the CD spectra of coupled base analogues over a range of variable calcium concentrations, which tuned the degree of exonuclease versus polymerase activity in the bound complex at a p/t junction [99]. Taking the observed balance between the two active

modes of DNA polymerase at a p/t junction as a starting point for our PS-SMF studies, we can identify which conformational macrostates in our model correspond to the exonuclease and polymerase configurations.





**Figure 5.8** Experimental data and optimized fits for the +4 dimer and DNA polymerase complex. PDF, C2, C3, and kinetic network shown.

The thermodynamic stability of  $S_1, S_2$  and  $S_3$  at the +4 position are all significantly altered by the binding of DNA polymerase, seen in Fig. 5.8. Macrostates  $S_4$  and  $S_5$  are considerably less stable but necessary to explain the observed dynamics. Additionally,  $S_4$  plays an important intermediate role in the optimized kinetic network, supporting its necessity in the observed dynamics. Given the assignment of  $S_3$  to a left-handed bases unstacked state, this is likely the exonuclease mode within the complex. This is supported by a few key pieces of information. First, ensemble CD results show that a left-handed conformation becomes favored as bivalent salt is titrated into solution to drive the DNA polymerase bound p/t junction into the exonuclease pocket (Fig. 5.7). Additionally, crystallographic studies as well as computational studies suggest that the arrangement of DNA in the exonuclease mode is perturbed away from B-form, with the sugar-phosphate backbone being rearranged locally [91], [98], [104]. This local rearrangement is more consistent with the high visibility left-handed conformer  $S_3$  than the low visibility propped-open conformer of  $S_1$ , since the crystal geometry of the DNA sugar-phosphate backbone in the exonuclease pocket is well off from 90 degrees (low signal visibility). The computational studies performed on the DNA holoenzyme suggested a timescale of interconversion between exonuclease and polymerase modes on the order of milliseconds [98]. The states  $S_1$  and  $S_3$  are directly

connected in the optimized kinetic network, both of which directly transition back to B-form DNA  $S_2$ . The time scale of interconversion between  $S_1$  and  $S_3$  is on the order of 1ms and is the most rapid interconverting pair of macrostates within the network. This, as well as the previously mentioned structural considerations, suggests that if  $S_3$  is the exonuclease mode then  $S_1$  is the polymerase mode, leaving  $S_2$  as B-form DNA. In this case,  $S_4$  plays an important role as the intermediate which tends to drive transitions from  $S_2$  into the rapidly interconverting  $S_3$  and  $S_1$ . Given our assignment of  $S_4$  as a right-handed, bases unstacked conformer, it could be that the loss of base-stacking interactions allows the transition to exonuclease activity to proceed without a significant enthalpic penalty from disruption of favorable stacking interactions. Upon forming the exonuclease pocket, the complex rapidly interconverts between its proof-reading and polymerizing functions, until returning to native B-form DNA. Also of note is that  $S_2$  and  $S_5$  rapidly interconvert between one another, but  $S_5$  is not connected with any other states in the network. It could be that  $S_5$  is structurally related to the other high visibility macrostate  $S_4$ , which  $S_2$  also rapidly exchanges with, but  $S_5$  isn't a productive structure for the formation of the exonuclease mode. In this case B-form DNA is rapidly sampling between two closely related conformers,  $S_4$  and  $S_5$ , but only leverages  $S_4$  to induce dynamic switching between exonuclease and polymerase modes.

The mechanical stability of a DNA p/t junction when bound to DNA polymerase is highly suggestive of a quasi-one-way loop mechanism. The thermodynamically most stable macrostate,  $S_2$ , is mechanically less stable transitioning toward  $S_5$  and  $S_4$  than it is transitioning toward  $S_1$  and  $S_3$ , the proposed polymerase and exonuclease modes respectively. As mentioned previously, it could be that  $S_5$  is structurally related to  $S_4$  but differs enough that it proves unproductive in inducing rapid switching between exonuclease and polymerase activity within the complex. When  $S_2$  makes a transition into  $S_4$ , the mechanical stability of  $S_4$  barely favors a transition directly back

to  $S_2$  over a transition to  $S_3$ . However, when  $S_4$  does transition to  $S_3$ , the mechanical stability of  $S_3$  is dominated by rapid fluctuations toward  $S_1$ , with the mechanical stability of  $S_1$  also being dominated by transitions back toward  $S_3$ . Transition barriers back to  $S_4$  from  $S_3$  are significant and tend to cause  $S_3$  and  $S_1$  to rapidly interconvert until transitioning back to  $S_2$  where the cycle begins anew. Notably,  $S_2$  has high transition barriers to directly forming either  $S_3$  or  $S_1$ , leading to the formation of a quasi-one-way loop mechanism. In this loop, the intermediate macrostate  $S_4$  is required to induce either of the two active modes of the polymerase starting from B-form DNA  $S_2$ . Additionally, both the direct formation of the active modes ( $S_3$  and  $S_1$ ) from  $S_2$  and backward transitions from active modes to the intermediate  $S_4$  are rare. This leads to conformational fluctuations within the DNA polymerase complex occurring mostly in a single direction within the kinetic network once the productive intermediate  $S_4$  successfully forms the exonuclease macrostate.

## 5. CONCLUSION

The interpretation of these results suggests that the DNA p/t junction at the +1 position rapidly fluctuates between multiple thermodynamically stable macrostates, which are selected for by the clamp-clamp loader complex for binding. The macrostate  $S_4$  appears to be the predominant conformer leveraged by the clamp-clamp loader during binding. Our proposals for the structures of  $S_2$  and  $S_4$  are consistent with previous crystallographic studies which have pointed out the small channel which DNA must pass through to load into the central pore of the clamp-clamp loader complex, as well as the spiral geometry of the clamp itself upon activation by the clamp loader to fit the helical backbone of B-form DNA [91], [92], [102]. The free energy landscape of the +1

position is significantly rearranged by the action of the clamp-clamp loader, suggesting that observed structural changes in ensemble CD and absorbance measurements performed on this same system are being recapitulated at the single-molecule level.

The proposed mechanism of DNA polymerase at a DNA p/t junction appears to be consistent with both previous ensemble biochemical studies and recent computational studies detailing the role of exonuclease versus polymerase modes as they dynamically interconvert [91], [98], [99]. Interestingly, the mechanism born out of our kinetic network analyses suggests a highly one-directional flow of probability around the optimized network of macrostates. The timescales of interconversion and relative abundance of these macrostates matches the intuitive picture put forth by previous studies and are again consistent with both CD and absorbance results that suggest a close competition between active polymerase modes. The free energy landscape of the polymerase data is significantly different than that of the +4 position DNA only data, suggesting we retain sensitivity to protein binding in our single molecule assays, even without the clamp present to tether the polymerase to the junction.

The experimental and computational methods described in Chapters 3-5, detailing the PS-SMF approach and its potential utility, can be readily extended to other protein-DNA systems. The most exciting and meaningful targets for such studies will surely be protein-DNA complexes which are highly dynamic on a local scale, precluding the possibility of more traditional FRET measurements and demanding time resolution not typically available to most camera-based imaging systems. The extension of these results might come in the form of additional independent studies beyond the scope of the clamp-clamp loader and DNA polymerase, but interesting insights might be gained by altering both the solvent environment during single-molecule experiments or considering alternate sequences and junction topologies on the ability of these same protein complexes to form



and remain stable. The often-accepted notion that replication machinery is agnostic to sequence could be tested further to examine possible issues of fidelity or rate limiting steps in replisome assembly due to alterations of both the junction itself and the solvent surround.

**Table 9** Optimized kinetic network model parameters for the free energy minima and activation barriers (left column) and optimized PDF parameters (right column) for all positions and conditions studied. Blank entries denote connections or states which did not exist in the optimized network models.

Construct	$A_1$	$\langle v \rangle_1$	$\sigma_1$	$p_1^{eq}$	$A_2$	$\langle v \rangle_2$	$\sigma_2$	$p_2^{eq}$	$A_3$	$\langle v \rangle_3$	$\sigma_3$	$p_3^{eq}$	$A_4$	$\langle v \rangle_4$	$\sigma_4$	$p_4^{eq}$	$A_5$	$\langle v \rangle_5$	$\sigma_5$	$p_5^{eq}$
+1 Dimer 100nm NaCl, 6mM MgCl2	1.48	0.089	0.037	0.139	5.293	0.13	0.06	0.798	0.183	0.189	0.065	0.029	0.296	0.27	0.043	0.032	-	-	-	-
+2 Dimer 100nm NaCl, 6mM MgCl2	0.566	0.019	0.042	0.06	7.233	0.119	0.049	0.895	0.331	0.149	0.053	0.044	0.001	0.324	0.057	0.0002	-	-	-	-
+3 Dimer 100nm NaCl, 6mM MgCl2	0.817	0.043	0.057	0.118	4.96	0.118	0.058	0.729	1.313	0.142	0.043	0.143	0.092	0.27	0.037	0.008	-	-	-	-
+4 Dimer 100nm NaCl, 6mM MgCl2	0.832	0.0001	0.048	0.101	6.6	0.09	0.049	0.819	0.981	0.139	0.031	0.078	0.002	0.249	0.045	0.0003	-	-	-	-
+15 Dimer 100nm NaCl, 6mM MgCl2	1.292	0.0001	0.045	0.148	8.06	0.101	0.039	0.791	0.146	0.205	0.037	0.013	0.413	0.25	0.044	0.045	-	-	-	-
+1 Dimer 100nm NaCl, 6mM MgCl2 1.5uM gp45, 1.0uM gp44/62	0.064	0.084	0.042	0.006	4.139	0.139	0.074	0.77	0.308	0.253	0.074	0.057	1.152	0.338	0.051	0.148	0.144	0.398	0.048	0.017
+4 Dimer 100nm NaCl, 6mM MgCl2 1.5uM gp43	1.284	0.065	0.069	0.225	3.795	0.139	0.044	0.427	1.718	0.157	0.074	0.32	0.141	0.252	0.07	0.025	0.009	0.437	0.079	0.001

Construct	$G_1^0$	$G_2^0$	$G_3^0$	$G_4^0$	$G_5^0$	$G_{1,2}^{0E}$	$G_{2,1}^{0E}$	$G_{1,3}^{0E}$	$G_{3,1}^{0E}$	$G_{1,4}^{0E}$	$G_{4,1}^{0E}$	$G_{1,5}^{0E}$	$G_{5,1}^{0E}$	$G_{2,3}^{0E}$	$G_{3,2}^{0E}$	$G_{2,4}^{0E}$	$G_{4,2}^{0E}$	$G_{2,5}^{0E}$	$G_{5,2}^{0E}$	$G_{3,4}^{0E}$	$G_{4,3}^{0E}$	$G_{3,5}^{0E}$	$G_{5,3}^{0E}$	$G_{4,4}^{0E}$	$G_{4,5}^{0E}$	$G_{5,4}^{0E}$	
+1 Dimer 100nm NaCl, 6mM MgCl2	1.747	0	3.287	3.208	∞	∞	∞	∞	∞	1.745	0.284	∞	∞	8.692	5.404	3.208	0	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
+2 Dimer 100nm NaCl, 6mM MgCl2	2.693	0	3.009	8.659	∞	∞	∞	∞	∞	5.966	0	∞	∞	12.873	9.863	8.659	0	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
+3 Dimer 100nm NaCl, 6mM MgCl2	1.818	0	1.624	4.431	∞	∞	∞	∞	∞	2.612	0	∞	∞	7.535	5.911	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
+4 Dimer 100nm NaCl, 6mM MgCl2	2.089	0	2.345	7.882	∞	∞	∞	∞	∞	8.92	17.444	7.006	6.613	5.793	0	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
+15 Dimer 100nm NaCl, 6mM MgCl2	1.672	0	4.045	2.848	∞	∞	∞	∞	∞	15.317	10.277	∞	∞	6.037	4.940	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
+1 Dimer 100nm NaCl, 6mM MgCl2 1.5uM gp45, 1.0uM gp44/62	4.718	0	2.995	1.646	3.776	∞	∞	∞	∞	∞	∞	∞	∞	2.85	3.792	11.151	11.181	1.646	0	7.652	3.876	7.883	9.5	∞	∞	∞	∞
+4 Dimer 100nm NaCl, 6mM MgCl2 1.5uM gp43	∞	∞	∞	∞	∞	5.963	12.977	0	0.349	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞

Construct	$k_{12}^{-1}(ms)$	$k_{21}^{-1}(ms)$	$k_{13}^{-1}(ms)$	$k_{31}^{-1}(ms)$	$k_{14}^{-1}(ms)$	$k_{41}^{-1}(ms)$	$k_{15}^{-1}(ms)$	$k_{51}^{-1}(ms)$	$k_{23}^{-1}(ms)$	$k_{32}^{-1}(ms)$	$k_{24}^{-1}(ms)$	$k_{42}^{-1}(ms)$	$k_{25}^{-1}(ms)$	$k_{52}^{-1}(ms)$	$k_{34}^{-1}(ms)$	$k_{43}^{-1}(ms)$	$k_{35}^{-1}(ms)$	$k_{53}^{-1}(ms)$	$k_{45}^{-1}(ms)$	$k_{54}^{-1}(ms)$
+1 Dimer 100mM NaCl, 6mM MgCl2	Inf	Inf	Inf	Inf	6.395	1.484	Inf	Inf	6553.488	248.292	27.621	1.116	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
+2 Dimer 100mM NaCl, 6mM MgCl2	Inf	Inf	Inf	Inf	0.194	0.001	Inf	Inf	194.886	9.611	2.883	0.001	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
+3 Dimer 100mM NaCl, 6mM MgCl2	0.722	4.449	Inf	Inf	3.336	0.244	Inf	Inf	458.494	90.32	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
+4 Dimer 100mM NaCl, 6mM MgCl2	3.741	18839.93	0.552	0.372	0.164	0.001	Inf	Inf	30.292	102940	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
+15 Dimer 100mM NaCl, 6mM MgCl2	9387.791	60.757	Inf	Inf	0.875	0.292	Inf	Inf	10.183	1407.743	11.978	0.292	Inf	Inf	0.002	0.006	Inf	Inf	Inf	Inf
+1 Dimer 100mM NaCl, 6mM MgCl2 1.5uM gp45, 1.0uM gp44/62	Inf	Inf	Inf	Inf	Inf	Inf	2.484	6.371	10004.41	10304.19	0.745	0.143	302.385	6.926	380.788	1918.868	Inf	Inf	Inf	Inf
+4 Dimer 100mM NaCl, 6mM MgCl2 1.5uM gp45	241.577	268810.1	0.621	0.881	Inf	Inf	Inf	Inf	268810.1	114.325	44.935	4.32	892.662	3.906	109999.8	6.701	Inf	Inf	Inf	Inf

**Table 10** Optimized kinetic network model parameters for the kinetic rate constants for all positions and conditions studied. Entries with 'Inf' denote connections which did not exist in the optimized network models.

## CHAPTER 6 : CONCLUDING SUMMARY

The technical developments of PS-SMF detailed in Chapters 2 and 3 enable the possibility for dynamical estimates of the local conformational free energy landscape of ss-dsDNA junctions as well as protein-DNA complexes to be made. Owing to the high degree of local specificity and microsecond time resolution, the persistence of four conformational macrostates emerged as the minimally necessary picture to explain the observed data. The sensitivity of the measurement to both position and salt concentration was demonstrated in Chapter 2 and 4. The analysis protocol developed for these single molecule experiments is robust in terms of the spanning set of possible kinetic networks optimized during analysis. Another strength of the approach is the ability to address data sets with largely disparate timescales in the two- and three-point TCFs as well as largely different PDFs. This inherent lack of bias for any particular experimental data set facilitates large scale optimization without much alteration to optimization parameters, thereby removing most of the tedious effort required to meet the needs of each individual data set.

The application of PS-SMF to protein-DNA complexes in Chapter 5 proves to be a useful analytical tool for investigating the biochemical mechanisms at play in such complexes. The availability of conformations at ss-dsDNA junctions appears to play an important role in the binding and function of these macromolecular machines. The role of intermediates in the formation of the productive modes of each complex appear important, as do the thermodynamic barriers which drive directed action through particular paths in the optimized networks. Potential structural and steric arguments exist which might elucidate the role each conformational macrostate plays in the process of protein complex assembly and function.

The appendix details various forays into the use of broadband excitation onto single molecules for potentially useful information that is uniquely available to broadband and higher

order spectroscopies. Despite the technical successes of such efforts, no new insights emerged which weren't already available from ensemble experiments. Some room does exist to improve and redesign these experimental approaches to better harness the inherent strengths of single molecule spectroscopy, namely dynamics and resolution of heterogeneity.

Lastly, interesting future directions drawing on combinations and extensions of these works can be imagined. The role of therapeutics in the disruption or enhancement of the conformational space of a protein, DNA or protein-DNA complex could be examined and understood through the lens of PS-SMF experiments. Efforts are currently underway in a variety of industrial and academic settings to leverage SM-FRET as a tool for drug discovery and drug mechanism in a variety of biophysical systems. Owing in part to the longstanding success and significant development of FRET protocols, analyses and biophysical implications, these efforts are making headway. A similar direction could be envisioned for PS-SMF experiments. Additionally, multiplexed experiments employing both a FRET pair and Cy3 homodimer for simultaneous FRET and PS-SMF measurements could be used to examine issues of allosteric control, colocalization effects from multi-subunit complexes and potentially be coordinated with therapeutics candidates to address a wide variety of biophysical questions pertinent to human health and disease.

# APPENDIX : INVESTIGATIONS INTO BROADBAND INTERFEROMETRIC MEASUREMENTS PERFORMED ON SINGLE MOLECULES

## 1. OVERVIEW

The research results presented in this final appendix are a compilation of various unpublished ultrafast single-molecule approaches developed within the Marcus and von Hippel labs by Jack Maurer and Amr Tamimi, with help of Anabel Chang and Lulu Enkhbataar. Broadly speaking, few labs have achieved ultrafast spectroscopy results on single molecules. This is largely due to the need for both novel detection schemes of single molecule signals as well as the often-fragile nature of individual molecules, which in a fluorescence-based experiment can present significant limitations on available acquisition time. Two main areas of focus were taken on by the contributing authors in this appendix. The first area of focus is the development and implementation of linear and nonlinear interferometric measurements on single molecules, in both one and two dimensions. The second area of focus is the design and implementation of novel time-correlated experiments on single molecules using broadband laser sources. These time-correlated experiments closely echo the approach taken in Chapter 3 to probe the inherently cross-polarized transitions of an electronically coupled  $i(\text{Cy}3)_2$  dimer. During these investigations, a rather intuitive but robust single molecule sample setup was devised by Jack Maurer to enable the long-term imaging of fluorescent fluorophores on the time scale of hours, which was essential to performing these experiments.

An important point should be made regarding these studies and the possibilities offered up by their initial success. While the technological achievement that these results represent is a first of

its kind, the biophysical insights gained from the data sets are either redundant when comparing to ensemble results or lacking in sensitivity compared to simpler experiments using CW laser sources. The rather high level of effort and difficulty required to make these experiments successful begs the question, how valuable must these experimental insights be to outweigh their cost in terms of time and resources? In the case of the protein-DNA or DNA only systems investigated in the Marcus and von Hippel group, the answer is surely that the value of the results doesn't outweigh the costs of performing the experiments. In the closing remarks of this appendix, alternate systems and use cases will be briefly discussed where the approaches described herein might find enhanced utility.

## 2. INTRODUCTION

In many chemical systems a degree of heterogeneity persists both statically and temporally, which tends to obscure the underlying structures and dynamics at the individual molecule level. This is particularly true of biological systems, where a high degree of conformational disorder exists across multiple scales and numerous subsystems within the cell [1]. The spectroscopic study of single molecules permits real time measurement of system observables, enabling study of their time evolution. Gaining access to system subpopulations and the instantaneous value of these spectroscopic observables can negate the effects of ensemble averaging, as demonstrated by a variety of well-studied approaches [11], [14], [105], [106]. Recently, advancements have been reported in the detection and analysis of one-dimensional linear and nonlinear spectroscopic signals obtained from single molecules under broadband pulsed excitation [52], [54], [107], [108]. Meanwhile, multidimensional approaches have exclusively been applied to traditional molecular



ensembles or spatially localized ensembles on the order of tens of nanometers [24], [25], [47], [109], [110]. Multidimensional spectroscopic techniques can uniquely provide information such as the presence or absence of coupling between optical transitions in photophysical aggregates, monitor the evolution of excited state population dynamics, and provide insight into the relative contributions of homogenous and inhomogeneous line broadening effects in heterogeneous systems.

In the presently discussed work, we successfully performed linear, nonlinear, and multidimensional experiments on single molecules of dsDNA containing monomers and dimers of Cy3 fluorophores rigidly inserted into the sugar-phosphate backbone of dsDNA. The signals reported here are acquired using a collinear pulse train and phase modulation technique developed by Marcus and coworkers [48], [111]. Acquisition of the linear response along a single interferometric time delay, followed by Fourier transform, yields the linear absorption spectra while tracking of the nonlinear response over two independent time delays yields the two-dimensional interferogram of a single molecule and the accompanying two-dimensional spectrum via Fourier transform. The linear as well as nonlinear signals in our multidimensional data sets are obtained through single photon time and phase tagging of the Stokes shifted fluorescence signal of individual molecules [69]. This phase-tagging method is based on earlier techniques which similarly employed single photon time and phase tagging to probe the optical response of single molecules [13], [57].

In higher ordered time-resolved spectroscopic approaches, the set of possible observables and the information they contain is much broader than conventional continuous-wave or temporally static spectroscopies [112]–[115]. However, these higher ordered approaches rely on lengthy experimental acquisition times, often across multiple time delays, which results in a

measurement time scale that has not been accessible to single molecule techniques due to the rapid and irreversible photobleaching that ubiquitously occurs in fluorescence-based approaches using organic dyes [77]. The work presented here overcomes the traditionally inaccessible regime of higher ordered measurements on single molecules by devising a scheme which permits the fluorophores under study to sustain emission for many hours on end. Recent work to better understand the mechanism of dark-state formation as well as improve the overall yield and stability of organic dyes could aid in the refinement of the approach outlined here [116], [117]. Notably, the lengthy integration times required for successfully obtaining the full two-dimensional histogram tends to cause the results to closely resemble ensemble studies when measuring the traditional rephasing and non-rephasing signals for estimates of the conformational heterogeneity [24]. Alternative approaches to obtaining these data sets, which may enable better access to dynamical effects at the single-molecule limit, are discussed in the conclusions section.

The work contained here is largely an investigative project to test the possibilities of each potential measurement for its utility within the broader field of biophysical science. As benchmarks of utility, we examine the effects of time averaging and total experimental acquisition time on the demodulation of linear and nonlinear responses of single molecules by time and phase tagging of single photons. We also present preliminary attempts to perform time-correlated analyses on a variety of linear signals obtained from single molecules, as both a comparison with the established PS-SMF approach and as a unique path toward isolating dynamical fluctuations in the presence of broadband excitation. We compare single molecule results to those obtained in bulk for the same dsDNA systems in the case of fully sampled interferometric measurements. The thresholds for acceptable SNR are established by examination of the resulting one- and two-dimensional spectra over a wide range of integration times and total photon numbers. The uncertainty in our

measurements is established through well-known boot-strapping methods for statistical characterization of error when sampling from an underlying probability distribution [69]. Lastly, we introduce a novel method for fluorophore lifetime enhancement using a closed-loop nitrogen purged imaging system.

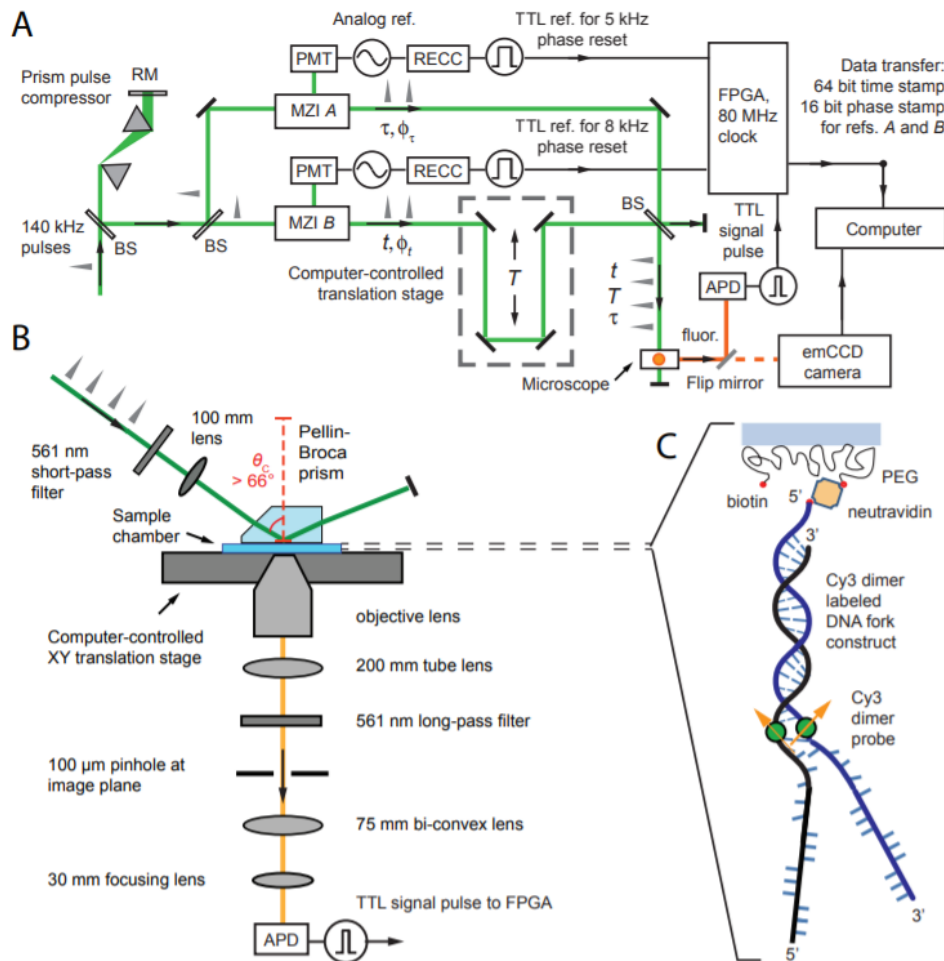
### 3. MATERIALS AND METHODS

#### Section A.3.1: Experimental Setup

Linear and nonlinear spectroscopic measurements are carried out by sending the output of a high-repetition-rate non-collinear optical parametric amplifier (NOPA) into either a single, or pair of, Mach-Zender interferometers (MZI) that each contain two acousto-optic modulators (AOMs) that sweep the relative phase of the two MZI arms according to  $\phi_{ij}(t) = \nu_{ij}mT$  with  $\nu_{ij} = \nu_i - \nu_j$  being the difference frequency of the two AOMs within the arms of each MZI and  $mT$  describing the  $m$ -th pulse in the pulse train with an interpulse separation period of  $T$  [48], [111]. The recombined pulse train from each MZI is sent to a 50/50 beam splitter, and then relayed to a Nikon TE2000 microscope where a total internal reflection (TIRF) geometry is used to excite a sample of immobilized single molecules. Fluorescence is spatially isolated by use of a 100 $\mu$ m pinhole and collected by a high NA immersion oil objective (100x) which directs the transmitted photon stream to either an em-CCD camera for spatial imaging or an APD for time and phase tagging. A schematic of the experimental apparatus is shown in Fig. A.1.

Single molecule samples are prepared by passivating the surface of a quartz microscope slide with a combination of mPEG (M.W.=500) and Biotin mPEG. A glass coverslip is passivated using mPEG (M.W.=500). The chemical details of this procedure are described elsewhere [14]. A sample

chamber is assembled from a passivated quartz slide and an appropriately cut section of passivated glass coverslip using a diamond tipped score. The quartz slide is drilled using a diamond tipped drill bit to achieve two holes with approximately ~2cm separation along the long axis of the slide. Two 100uL pipet tips are inserted into the quartz slide and the tips are shaven off from the biotinylated surface using a razor blade, actively avoiding the imaging area. Four 1uL droplets of TRIS buffer (100mM NaCl, 10mM TRIS, 6mM MgCl<sub>2</sub>) are cast between the two drilled inlets on the quartz slide, the prepared section of glass coverslip is carefully placed to fall onto the drop cast region. The edges of the glass cover slip are sealed on the slide using a two-part rapid-cure 3M epoxy. Two short (~4-6cm) segments of Tygon tubing (1/8" internal diameter) are inserted into the remaining ends of the pipet tips and sealed in place using epoxy. The open ends of the Tygon tubing are fit with 1/8"-to-male luer lock adapters. Once all epoxy has dried fully, a 300uL portion of TRIS buffer is injected into the sample chamber using a standard 1mL syringe to check for leaks or clogs. Fully assembled sample chambers may be stored in a fridge for up to 10 days. Assuming no issues exist, 100uL of a Neutravidin solution (0.1mg/mL) is injected into the sample chamber and allowed to react for ~1min prior to a complete flush of the sample chamber using 300uL of TRIS buffer. Next, 100uL of 1pM biotinylated dsDNA solution is injected into the sample chamber and allowed to react for ~1min prior to a complete flush with 300uL of TRIS buffer. The completed slide sample chamber is placed in a protective dark environment to reduce photodegradation of the bound Neutravidin.



**Figure A.1** Schematic of the experimental apparatus used for ultrafast linear and nonlinear single molecule measurements.

A solution of 30mg of Trolox, 120 $\mu\text{L}$  of 1M NaOH, 100mg Glucose and 10mL of TRIS buffer is prepared and mixed for a minimum of 2 hours to ensure complete dissolution. Separate solutions of 89mg of glucose oxidase in 1mL of buffer and 57mg Catalase in 1mL of buffer are both prepped and divided into 60 $\mu\text{L}$  aliquots, with a single aliquot of each needed per sample. The remaining aliquots can be kept at  $-20^\circ\text{C}$  for long term storage. A INSI tech series 720p peristaltic pump with a C60 Flex hairpin loop is attached to two long pieces ( $\sim 1.5'$  each) of 1/8" Tygon tubing at each port of the hairpin loop. One of the Tygon tubes is equipped with a 1/8"-to-male luer lock adapter with a 1.5" two way stop-cock attached. The other Tygon tube is attached to a 1/8"-to-male syringe adapter. A third piece of Tygon tubing is cut (1-1.5') and fitted with a 1/8"-to-male luer lock adapter

and a 1.5" two way stop-cock at one end, with the opposite end being attached to a 1/8"-to-male syringe adapter. One 5mL reservoir with airtight septum, two 20-gauge needles and two 18-gauge needles are acquired for final assembly. Tygon tubing lines are cleaned with ethanol and dried prior to use. Once peristaltic pump assembly is finished, a mixture is prepared with 4.9mL Trolox solution, 50uL Catalase and 50uL oxidase (TOSS). TOSS is mixed thoroughly and centrifuged, separating out the supernatant into manageable portions.

The pump system is assembled by securing the empty reservoir with septa in place using a ring stand or other device to be near the experimental apparatus. The septa is pierced just once using one 18-gauge needle, inserted completely into the reservoir. The septa is again pierced just once using one 20-gauge needle, inserted completely into the reservoir. Using a syringe with a separate needle attached, aliquots of the TOSS solution are moved into the reservoir using the 18-gauge needle until nearly full (~3.5-4mL). A second 18-gauge needle is inserted and both pump lines are attached to the two 18-gauge needles. TOSS solution is flowed through the dead volume of the pump until full. N<sub>2</sub> gas, flowed in via the 20-gauge needle, can be used to generate back pressure in the Tygon pump lines if peristalsis fails to extract TOSS solution from the reservoir. Tygon tubing leads are attached to sample slide, stop-cocks closed, and the reservoir is brought up to full volume. The final 20-gauge needle attached to a pure N<sub>2</sub> tank is inserted and a gentle stream of N<sub>2</sub> gas inside the reservoir is initiated; with the opposing 20-gauge needle and Tygon tubing line inserted into a vial of water to avoid pressure buildup in the reservoir. The pump system is gently run with slide attached, stop cocks open and nitrogen flowing for ~20-30min to ensure no clogs or leaks form. If intact after waiting period, the pump is shut down. N<sub>2</sub> gas is left running and the system is allowed to purge under N<sub>2</sub> overnight. After a full purge period, the sample slide is imaged

and gently pumped every few hours for ~5min to replenish the working volume of TOSS inside the sample chamber.

### Section A.3.2 Linear Signal Derivation

In the high flux regime at a single MZI delay of  $t_{ij}$ , the linear signal obtained by integration of the fluorescent photon stream is described by Eq. (53), which comes from considering the effects of imparting a phase modulated pulse train onto the transition dipole moment of a chromophore [24], [68], [69].

Starting with the form of the electric field for any one pulse that has a complex valued phase,  $iv_jmT$ , imparted to it by the AOMs:

$$\vec{E}_j(\omega, t) = \vec{a}_j(\omega)e^{i\omega t_j - iv_jmT} \quad (48)$$

Utilizing the results of Appendix A.10 in [48], we can write the resulting product of the j-th field matter interaction with transition dipole moment of the molecule,  $\vec{\mu}_{ng}$ , as a ket, where the possible transition frequencies,  $\omega_{ng}$ , are summed over:

$$|\psi_j\rangle = i \sum_n \vec{\mu}_{ng} \vec{a}_j(\omega_{ng}) e^{i\omega_{ng}t_j - iv_jmT} |n\rangle \quad (49)$$

Next, we can consider the overlap between any two such field-matter interactions to produce a population capable of fluorescence by radiative decay. These interactions can come from either two distinctly modulated pulses, from separate arms of the MZI, or from the same pulse. We will assume a non-zero directional overlap between vector elements in the sum.

$$\langle \psi_i | \psi_j \rangle = \sum_n |\mu_{ng}|^2 \alpha_i(\omega_{ng})^* \alpha_j(\omega_{ng}) e^{i\omega_{ng}t_j - iv_jmT} e^{-i\omega_{ng}t_i + iv_imT} \langle n | n \rangle \quad (50)$$

Assuming the field of the pulses to be spectrally identical, and their times of arrival independent from one another:

$$= \sum_n |\mu_{ng}|^2 \alpha(\omega_{ng})^2 e^{i\omega_{ng}(t_j-t_i)-i(v_j-v_i) mT} \quad (51)$$

Utilizing the simplifications states above, we arrive at:

$$\langle \psi_i | \psi_j \rangle = \sum_n |\mu_{ng}|^2 \alpha(\omega_{ng})^2 e^{i\omega_{ng}(t_{ij})-i\phi_{ij}} \quad (52)$$

When taking  $i \neq j$ , we obtain a phase modulated term, where  $\phi_{ij} \neq 0$ . In the case of  $i = j$  we obtain a background term which does not contribute to the modulated portion of the time varying fluorescence signal. Combing terms from both cases yields an overall expression for the linear signal given by Eq. (53):

$$\begin{aligned} A(\phi_{ij}, t_{ij}) &= A_{\text{bkgd}} + A_{\text{lin}}(\phi_{ij}, t_{ij}) \\ &= 2 \sum_n |\alpha(\omega_{ng})|^2 |\mu_{ng}|^2 + 2 \text{Re} \sum_n |\alpha(\omega_{ng})|^2 |\mu_{ng}|^2 \exp\{i[\phi_{ij}(t) - \omega_{ng} t_{ij}]\} \end{aligned} \quad (53)$$

The sum is carried over all molecular excited state levels (labeled  $n$ ),  $\mu_{ng}$  is still the transition dipole matrix element that couples the ground and  $n$ th excited states,  $\omega_{ng}$  is the complex-valued optical transition frequency and  $|\alpha(\omega)|^2$  is the intensity of the laser pulse spectrum [111]. The two terms contributing to Eq. (53) are  $A_{\text{bkgd}}$  and  $A_{\text{lin}}(\phi_{ij}, t_{ij})$ .

$A_{\text{bkgd}}$  represents the constant background that arises from single pulse interactions, stray light and detector noise; all of which are independent of the phase sweep in the MZI. The second term,  $A_{\text{lin}}(\phi, \tau)$ , depends on the phase of the MZI and represents the modulated contribution to the time varying fluorescence that can be detected and demodulated to obtain the signal quadratures.



In the regime of single molecules, where the flux of fluorescent photons is low and on the order of  $10^3$  to  $10^5$  per second, an alternate approach for demodulating the signal amplitude and phase is required [13], [57], [69]. In this regime, we regard each TTL detection event registered at the APD as a Dirac delta function  $\delta(\phi - \phi_k)$  that instantaneously samples the signal phase  $\phi_{ij}$  from a probability distribution specified by the high-flux photon count rate of Eq.(53). Over the course of a fixed integration period  $T_{PT}$ , the phase dependent photon rate arising from  $N$  detection events can be treated as the discrete sum in Eq.(54):

$$A^{PT}(\phi, t_{ij}) = \frac{1}{T_{PT}} \sum_{k=1}^N \delta(\phi - \phi_k) \quad (54)$$

In order to isolate the linear signal from the background via Fourier transform, the photon rate must be phase-tagged with respect to the phase coordinate. This is mathematically equivalent to summing over the  $N$  photon phase factors:

$$Z_{\text{lin}}^{PT}(t_{ij}) = \frac{1}{2\pi} \int_0^{2\pi} A^{PT}(\phi, t_{ij}) e^{-i\phi} d\phi \quad (55a)$$

$$= \frac{1}{\pi T_{PT}} \int_0^{2\pi} \left[ \sum_{k=1}^N \delta(\phi - \phi_k) \right] e^{-i\phi} d\phi = \frac{1}{\pi T_{PT}} \sum_{k=1}^N e^{-i\phi_k} \quad (55b)$$

$$= f^{PT} |v^{PT}(\tau)| e^{i[\gamma^{PT}(\tau) - \omega_R \tau]} \quad (55c)$$

Where  $v^{PT}(\tau)$  is the visibility of the signal,  $\gamma^{PT}(\tau)$  is the phase of the signal and  $\omega_R$  is the frequency of the optical reference used for down sampling[48], [69], [111]. The quadrature's

typically obtained from demodulation of analog signals correspond to the real and imaginary parts of the complex valued linear signal obtained by discrete sum over the  $N$  photon phase factors.

$$Z_{\text{lin}}^{PT}(\tau) = X_{\text{lin}}^{PT}(\tau) - iY_{\text{lin}}^{PT}(\tau) \equiv \frac{1}{\pi T_{PT}} [\sum_{k=1}^N \cos(\phi_k) - i \sum_{k=1}^N \sin(\phi_k)] \quad (56)$$

Where the usual definition for amplitude and phase of a complex valued vector applies. The complex valued molecular susceptibility can then be obtained by Fourier transformation of the complex valued linear signal with respect to the MZI delay, which yields the molecular overlap spectrum [111].

### Section A.3.3 Nonlinear Signal Derivation

In the high flux regime at two unique MZI delays of  $t_{ij}$  and  $t_{lk}$  the nonlinear signal obtained by integration of the fluorescent photon stream is described by Eq. (60 and 61), which comes from the same considerations as were discussed for Eq. 53. The form of each field is the same as before, but owing to the combinatorics of four pulse sequences versus two pulse sequences, more possible terms exist which can contribute to the time varying fluorescence signal modulated at combinations of the underlying linear phase signatures, yielding nonlinear terms. For the sake of brevity a single term out of all possible terms will be derived and considered. The approach taken is generally applicable to all remaining terms.

Utilizing the results of Appendix A.12 in [48], we can write the resulting product of the  $i$ -th,  $j$ -th and  $k$ -th field matter interaction with transition dipole moment of the molecule,  $\vec{\mu}_{ng}$ , as a ket, where the possible transition frequencies are individually summed over. We consider only the rephasing and nonrephasing pathways in this appendix. This limits the scope of signal terms to

those arising from the overlap of first and third order wave packets [48]. Whereas the terms arising from overlap of two mutually second order wave packets are exclusive to double quantum coherence terms, which are not considered in this work [48], [110]. Notably, the middle interaction in all third order wave packet terms of this approach are taken to act oppositely the other two interactions, yielding a difference in sign from the phase imparted by the middle pulse:

$$|\psi_{ijk}\rangle = |\psi_i\rangle|\psi_j^*\rangle|\psi_k\rangle = i \sum_a \vec{\mu}_{ag} \vec{a}_i(\omega_{ag}) e^{i\omega_{ag}t_i - iv_i mT} |a\rangle \sum_b \vec{\mu}_{bg} \vec{a}_j(\omega_{bg}) e^{-i\omega_{bg}t_j + iv_j mT} \langle b| \sum_c \vec{\mu}_{cg} \vec{a}_k(\omega_{cg}) e^{i\omega_{cg}t_k - iv_k mT} |c\rangle \quad (57)$$

Given the constraint in our experiments that relevant signal terms must result in an excited state population, two of these transition dipole elements need be the same (i.e.  $b = c$ ), with the remaining transition dipole element being equal to the first order wave packet that overlaps this third order wave packet, in the course of a four-pulse interaction sequence [48].

$$= i \sum_a \vec{\mu}_{ag} \vec{a}_i(\omega_{ag}) e^{i\omega_{ag}t_i - iv_i mT} |a\rangle \sum_b \vec{\mu}_{bg} \vec{a}_j(\omega_{bg}) e^{-i\omega_{bg}t_j + iv_j mT} \vec{\mu}_{bg} \vec{a}_k(\omega_{bg}) e^{i\omega_{bg}t_k - iv_k mT} \langle b|b\rangle \quad (58)$$

If we take the pulses to be spectrally identical:

$$= i \sum_{a,b} \vec{\mu}_{ag} \vec{a}(\omega_{ag}) e^{i\omega_{ag}t_i - iv_i mT} |\vec{\mu}_{bg}|^2 \vec{a}(\omega_{bg})^2 e^{i\omega_{bg}t_j - iv_j mT} e^{i\omega_{bg}t_k - iv_k mT} |a\rangle \quad (59)$$

We can now overlap this third order wave packet with a first order wave packet (Eq. 49) to obtain the general form of the rephasing and nonrephasing signal terms:

$$\begin{aligned} & \langle \psi_l | \psi_{ijk} \rangle \\ &= \sum_{a,b} |\vec{\mu}_{ag}|^2 \vec{a}(\omega_{ag})^2 e^{i\omega_{ag}t_i - iv_i mT} e^{-i\omega_{ag}t_l + iv_l mT} |\vec{\mu}_{bg}|^2 \vec{a}(\omega_{bg})^2 e^{-i\omega_{bg}t_j + iv_j mT} e^{i\omega_{bg}t_k - iv_k mT} \\ &= \sum_{a,b} |\vec{\mu}_{ag}|^2 \vec{a}(\omega_{ag})^2 |\vec{\mu}_{bg}|^2 \vec{a}(\omega_{bg})^2 e^{i\omega_{ag}(t_i - t_l) - i(v_i - v_l)mT} e^{-i\omega_{bg}(t_j - t_k) + i(v_j - v_k)mT} \end{aligned} \quad (60)$$

The identification of purely rephasing versus nonrephasing terms comes from the explicit consideration of which pulse produces the first order wave packet. In general, these are four such relevant terms (plus their complex conjugates):

$$\begin{aligned}
S_{RP} + S_{NRP} &= \\
&\langle \psi_1 | \psi_{234} \rangle + \langle \psi_2 | \psi_{134} \rangle + \langle \psi_3 | \psi_{124} \rangle + \langle \psi_4 | \psi_{123} \rangle + \langle \psi_{123} | \psi_4 \rangle + \langle \psi_{124} | \psi_3 \rangle + \langle \psi_{134} | \psi_2 \rangle + \langle \psi_{234} | \psi_1 \rangle \\
&= 2 \operatorname{Re} [ \langle \psi_1 | \psi_{234} \rangle + \langle \psi_2 | \psi_{134} \rangle + \langle \psi_3 | \psi_{124} \rangle + \langle \psi_4 | \psi_{123} \rangle ]
\end{aligned} \tag{61}$$

Isolation of the terms which contain differences of the underlying linear signal phases ( $\phi_{21} - \phi_{43}$ ) leads to identification of the rephasing signal, while isolation of terms containing the sum of the underlying linear signal phases ( $\phi_{21} + \phi_{43}$ ) leads to identification of the nonrephasing signal [48]. In a typical ensemble experiment, both linear signal phases are tracked independently, then multiplied and low-pass filtered to achieve mixing and produce the necessary reference signals for demodulation of nonlinear signals using a lock-in amplifier. But in the case of low-flux single-molecule experiments, every photon is tagged for its time of arrival and phase with respect to both linear references signals. The construction of the nonlinear signal in the low-flux regime is very similar to the approach taken for the linear signal case (Eqns. 54-56), but instead of mixing analog signals as is done in the case of ensemble experiments, simple multiplication of the photon phase tags arising from the two independent linear references yields the nonlinear signal. This results in a total phase factor for the rephasing and nonrephasing signals which appears at the difference and sum of the linear phase tags, respectively.

We can define two linear photon phase tag terms, one for each MZI:

$$A_{21}^{PT}(\phi_{21}, t_{21}) = \frac{1}{T_{PT}} \sum_{k=1}^N \delta(\phi_{21} - \phi_{21}^k) \tag{62}$$

$$A_{43}^{PT}(\phi_{43}, t_{43}) = \frac{1}{T_{PT}} \sum_{z=1}^N \delta(\phi_{43} - \phi_{43}^z) = \frac{1}{\pi T_{PT}} \sum_{z=1}^N e^{-i\phi_{43}^z} \quad (63)$$

Here,  $\phi_{21}$  is the general coordinate that tracks the phase relationship between the two arms of the first MZI, while  $\phi_{21}^k$  is the instantaneous phase assigned to the k-th photon detection event. The same follows for  $\phi_{43}$  which tracks the second MZI. Each of these terms is a discrete sum over the instantaneous phase of each photon arriving at the APD, with respect to either of the two linear reference signals. Multiplication of these linear phase-tagged sums, prior to averaging over the phase coordinate, yields the nonrephasing photon tags:

$$A_{NRP}^{PT}(\phi_{21}, \phi_{43}, t_{21}, t_{43}) = \frac{1}{T_{PT}} \sum_{k=1}^N \delta(\phi_{21} - \phi_{21}^k) \delta(\phi_{43} - \phi_{43}^k) \quad (64)$$

While exchange of the sign on the second reference signal yields the rephasing photon tags:

$$A_{RP}^{PT}(\phi_{21}, \phi_{43}, t_{21}, t_{43}) = \frac{1}{T_{PT}} \sum_{k=1}^N \delta(\phi_{21} - \phi_{21}^k) \delta(\phi_{43} + \phi_{43}^k) \quad (65)$$

Averaging over each phase coordinate respectively, we obtain:

$$\begin{aligned} Z_{NRP}^{PT}(\phi_{21}, \phi_{43}, t_{21}, t_{43}) &= \\ \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2\pi} A_{NRP}^{PT}(\phi_{21}, \phi_{43}, t_{21}, t_{43}) e^{-i(\phi_{21} + \phi_{43})} d\phi_{21} d\phi_{43} &= \frac{1}{\pi T_{PT}} \sum_{k=1}^N e^{-i(\phi_{21}^k + \phi_{43}^k)} \end{aligned} \quad (66)$$

$$\begin{aligned} Z_{RP}^{PT}(\phi_{21}, \phi_{43}, t_{21}, t_{43}) &= \\ \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2\pi} A_{RP}^{PT}(\phi_{21}, \phi_{43}, t_{21}, t_{43}) e^{-i(\phi_{21} - \phi_{43})} d\phi_{21} d\phi_{43} &= \frac{1}{\pi T_{PT}} \sum_{k=1}^N e^{-i(\phi_{21}^k - \phi_{43}^k)} \end{aligned} \quad (67)$$

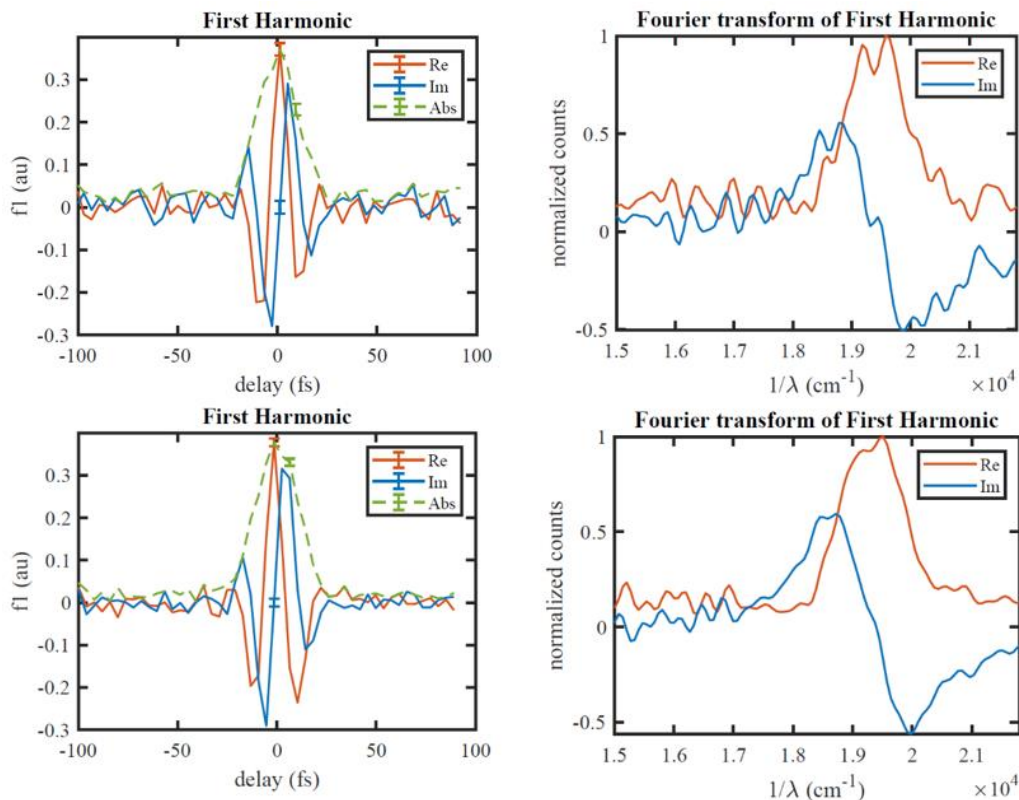
Both these terms can be separated into their real and imaginary parts to obtain the quadratures for both the rephasing and nonrephasing signals, the same as was done for the linear signals in Eq. 56.

With demodulation of both the linear and nonlinear phase tagging signals, we can examine the results of single-molecule experiments leveraging the discrete sums of Eqns. 55b, 66 and 67.

## 4. RESULTS AND DISCUSSION

### Section A.4.1 Linear Ultrafast Single Molecule Measurements

Representative linear interferometric measurements performed on single molecules are shown in Fig. A.2. These measurements were obtained with 1-second integration time and a sampling of 51 time point along the  $t_{21}$  axis. The resulting  $\sim$ 1-minute scans display a similarly well resolved spectral distribution centered at the laser bandwidth. Yet, the spectral features appear to differ between these two scans. The inherent differences between scans are likely due to noise effects at the level of 1-minute acquisition times, but in principle could arise from unique sets of conformational fluctuations during data acquisition. Further refinement of the SNR in these experiments, leading to a reduction of the integration time, as well as compressed sensing in the time domain could be used to resolve dynamics at the single molecule level in the form of time resolved linear absorption spectra.



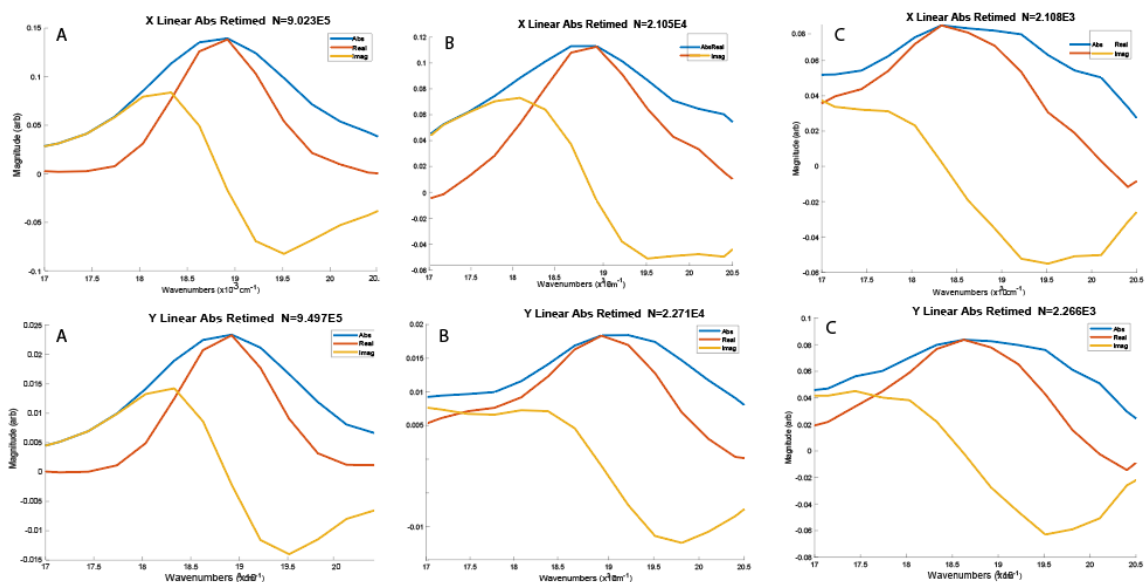
**Figure A.2** Two representative linear interferometric experiments on single molecules of dsDNA containing a  $i(\text{Cy}3)_2$  dimer. Interferogram data is shown in the first column and the resulting Fourier transform is shown in the second column.

Recent insights gleaned from single molecule studies of protein-DNA interactions reveal that the mechanical stability of certain macromolecular complexes obey transition time scales on the order of hundreds of milliseconds to tens of seconds. These sorts of timescales could be easily amenable to rapid linear interferometric experiments, to obtain a time resolved sampling of the linear absorbance spectrum. However, the information gained from such efforts would need to exceed the dynamical information already available from more conventional single molecule techniques. In the case of rapidly sampled linear absorption spectra, one could model the dipole-dipole interactions occurring within the  $i(\text{Cy}3)_2$  dimer, using the tools developed by the Marcus lab [24], [46], [47], to obtain structural parameters on a scan-by-scan basis. The limitations here are primarily due to the SNR of such rapidly obtained spectra, but the inherently greater capacity

to address structure as well as dynamics could prove valuable. Whereas in more conventional single molecule measurements the structural details are often inferred or limited to a small set of geometric parameters. One important consideration in improving these measurements is the optical setup employed. In the experiments presented here, the output of the NOPA is set at a repetition rate of 144kHz. Given that the pulse duration in our instrument is on the order of  $\sim 30$ fs, the number of fluorescent photons detected is certainly dominated by the fluorescent lifetime of the  $i(\text{Cy}3)_2$  dimer, which is orders of magnitude longer in time. Higher repetition rates at the pulse generation source could provide a better photon budget during experiments. Notably groups with demonstrated success in related measurements have often employed MHz repetition rate systems [107], [108].

As a demonstration of the limitations on linear measurements with increasing or decreasing the number of detected photons, a series of linear absorption spectra are shown in Fig. A.3, where the integration time at each step in the  $t_{21}$  delay space was varied 42seconds (Column A) to 1sec (Column B) and finally 100ms (Column C). The data were obtained during a full two-dimensional experiment, such that both MZI 1 (top row) and MZI 2 (bottom row) could be examined simultaneously for their relative loss in SNR as the allowable integration time was varied.





**Figure A.3** A series of linear Fourier transforms are shown as a function of integration time, ranging from 42sec (left column) to 100ms (right column). The top and bottom rows are the linear transform obtained by demodulation of each interferometers linear signal contribution, which are prepared simultaneously in the course of a two-dimensional experimental protocol.

The lower limit of 100ms shows a relatively poorly resolved spectra, while the upper limit at the full 42 second integration time is well resolved. Importantly, these scans were performed with only 9 steps in either the  $t_{21}$  or  $t_{43}$  domain, to minimally satisfy the Nyquist sampling frequency such that total acquisition time of the full interferogram could be kept at a minimum. From these initial attempts, it appears that experiments utilizing integration times on the order of 100ms are possible, but will be limited by their SNR.

### Section A.4.2 Nonlinear Ultrafast Single Molecule Measurements

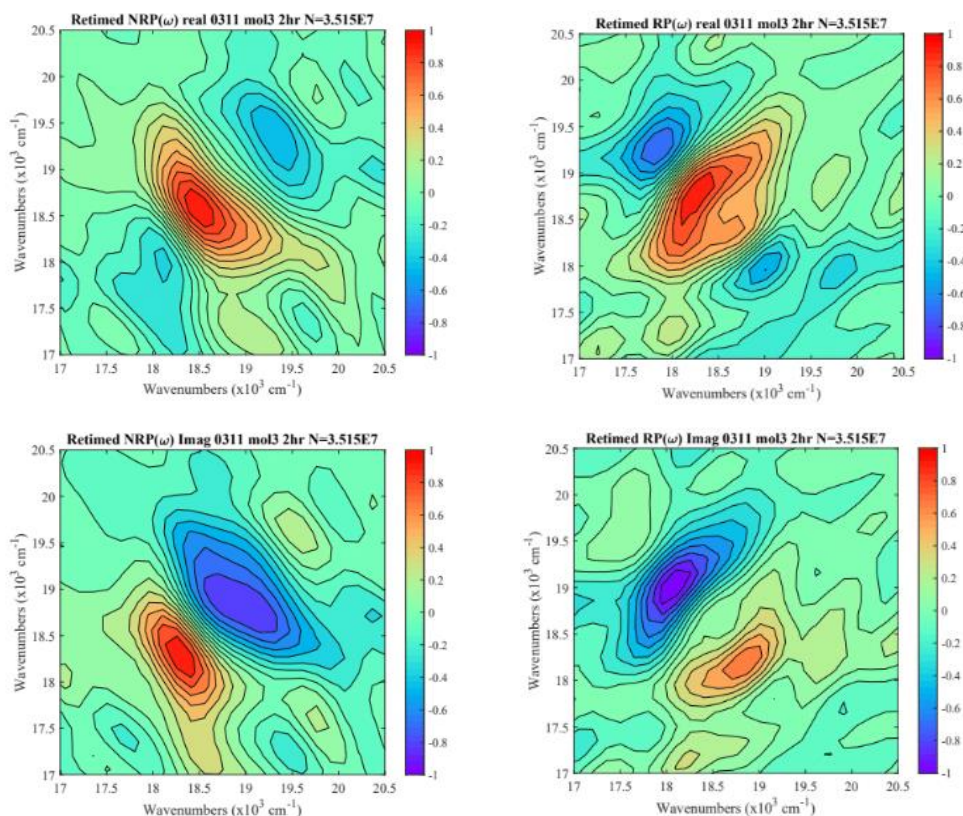
Nonlinear experiments were also performed on single molecules, enabled both by novel time and phase-tagging of single photons as well as the closed-loop sample chamber discussed in the previous section [69]. Initially, these experiments were investigated for their feasibility given

the technical hurdle they presented. Later, attempts were made to average together multiple scans from the same molecule to obtain high SNR reconstructions of the rephasing and nonrephasing spectrum. The minimal interferometric sampling and integration time at each point in the delay space was used for individual two-dimensional scans, owing to the long-term drift which can occur in the instrument over the course of one experiment that tends to obscure the true time-zero of each interpulse delay.

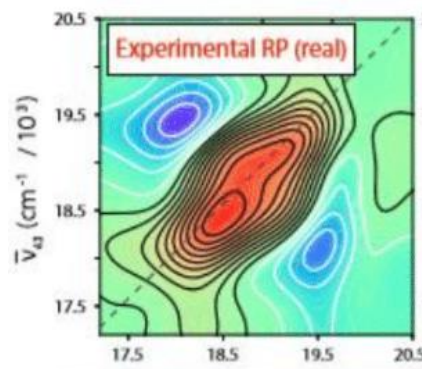
An example two-dimensional spectrum of a single molecule is shown in Fig. A.4, all individual scans were obtained by taking ten steps of 2.5fs along the  $t_{21}$  dimension and 9 steps of 2.5fs along the  $t_{43}$  dimension. At each point in the delay space, the signal is integrated for 42 seconds. This results in a total of 90 steps and a scan time of 63 minutes. A single 63-minute experiment can yield results without further averaging, but best results are obtained from averaging over multiple 63-minute interferograms. A symmetric grid in  $t_{21}$  and  $t_{43}$  is constructed from the asymmetric sampling in post processing using a retiming and rephasing procedure. This result is then passed to a two-dimensional Fourier transform, yielding the two-dimensional spectrum. Multiple retimed and rephased data sets from the same molecule were averaged to obtain the results displayed below in Fig. A.4.

For comparison to Fig. A.4, a Cy3 monomer spectra obtained in ensemble studies is shown in Fig. A.5, courtesy of Dr. Dylan Heussman [25]. The results of single molecule experiments and ensemble experiments closely resemble one another. While their equivalence provides an excellent check on the validity of the single molecule approach, it also suggests that any ability to study conformational subpopulations or dynamics, without the effects of ensemble averaging, is severely limited. As has been noted in numerous other studies regarding ultrafast spectroscopic measurements on single molecules [54], [107], [112], the typically low flux of fluorescent photons

from a single molecule demands a lengthy integration time to achieve reasonable SNR in the demodulation of linear or nonlinear signals. Of course, these signals must also be sufficiently sampled in the time domain to avoid aliasing from frequencies beyond Nyquist, which places a constraint on the minimum separation and number of points sampled in the interpulse delay space. This innate limitation on data acquisition time juxtaposed with the traditionally short-lived fluorescent emission from single molecules demands an alternative experimental approach if sub-ensemble averaged information is desired. Compressed sensing approaches offer reduced acquisition times without loss of spectral information [112], [118], [119], but have not yet been applied at the level of single molecules.



**Figure A.4** Two dimensional spectra obtained on a single dsDNA molecule containing a monomer of Cy3 rigidly inserted into the sugar-phosphate backbone. The real (upper row) and imaginary parts (lower row) of the nonrephasing and rephasing spectra are shown on the left and right respectively.

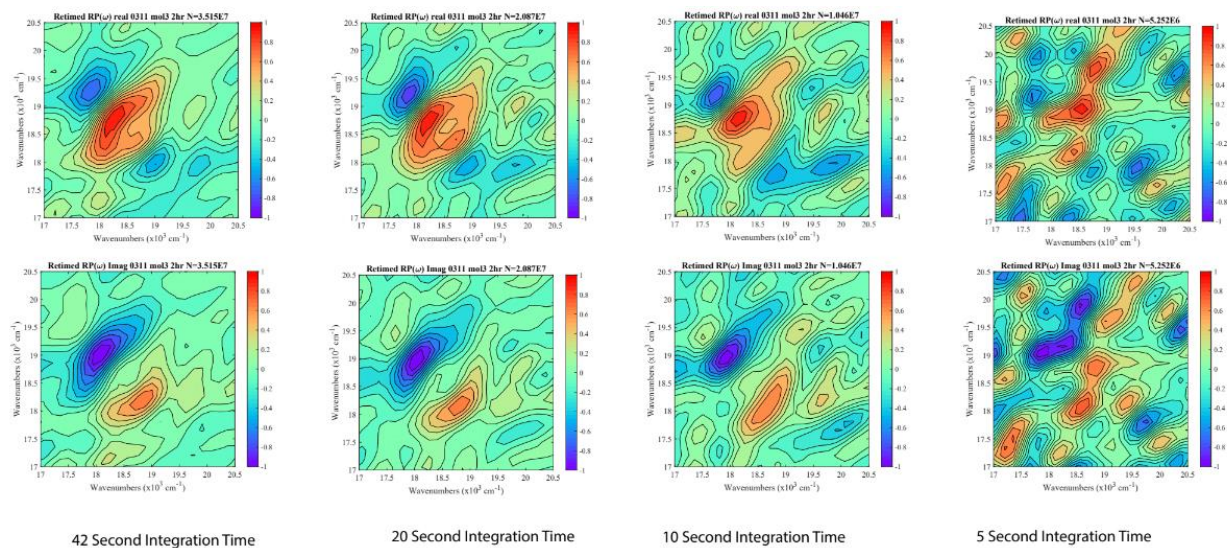


**Figure A.5** Real part of the rephasing spectrum obtained on an ensemble of dsDNA molecules containing a monomer of Cy3. This ensemble result can be directly compared with the upper right panel of Fig. A.4.

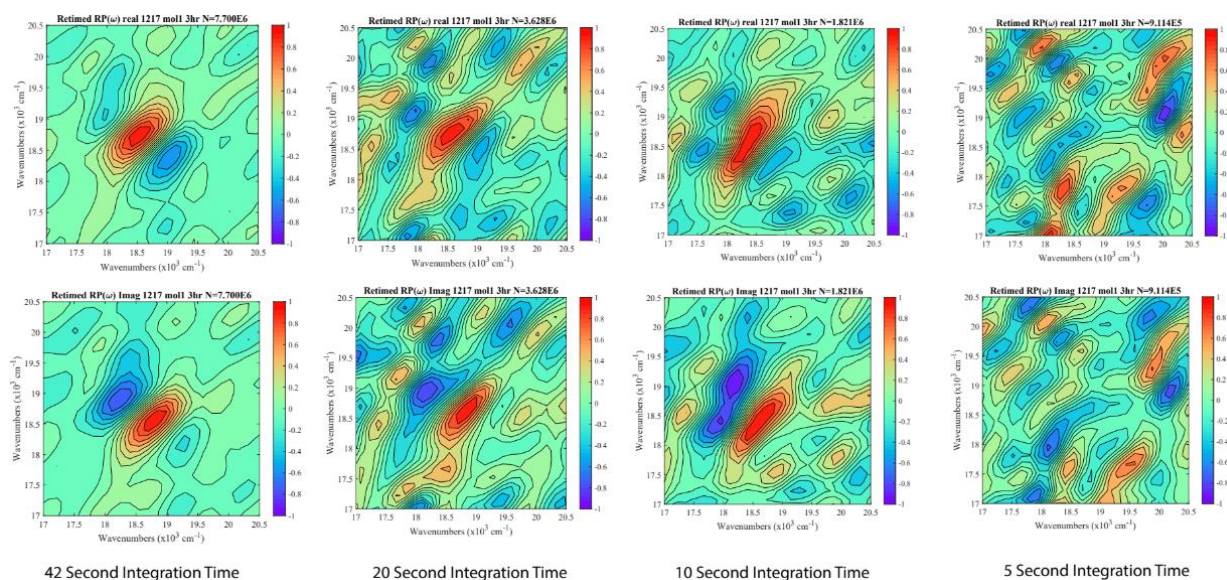
#### Section A.4.2.1 Averaging effects on signal-to-noise and error analysis of nonlinear single molecule data sets

Figures (A.6) and (A.7) show the real (top row) and imaginary (bottom row) rephasing spectra obtained from single dsDNA molecule containing either a Cy3 monomer (Fig. A.6) or dimer (Fig. A.7) incorporated at the +1 position. These spectra are subjected to a similar test of SNR as the linear spectra of Figure (A.3), where the integration time is variably adjusted from a maximum of 42 seconds down to a minimum of 5 seconds at each point in the time domain.





**Figure A.6** Rephasing 2D spectra obtained from single Cy3 monomers incorporated into dsDNA. The integration time at each time step is variably adjusted to demonstrate the limits on SNR.



**Figure A.7** Rephasing 2D spectra obtained from single Cy3 dimers incorporated into dsDNA. The integration time at each time step is variably adjusted to demonstrate the limits on SNR.

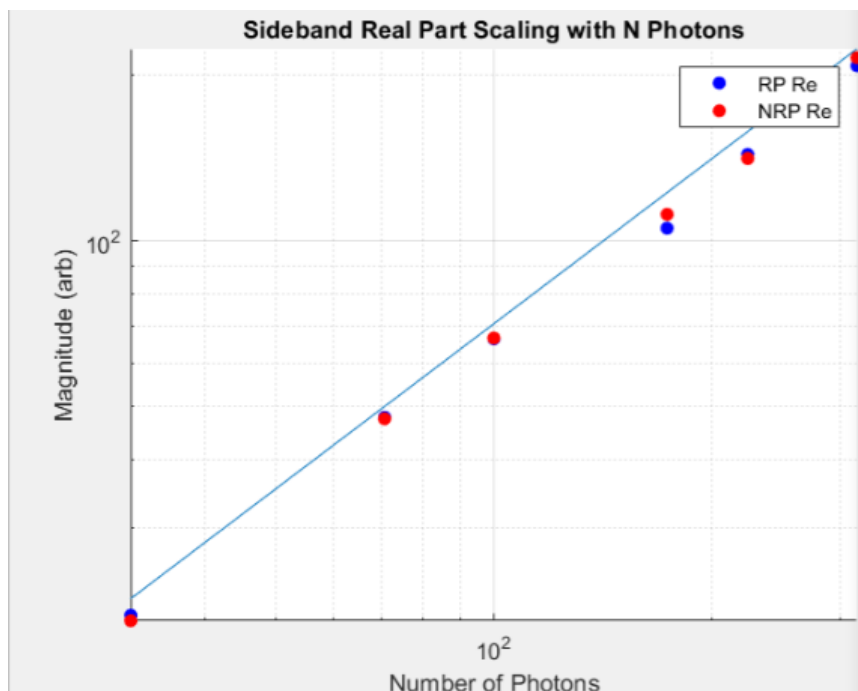
As can be seen from the resulting rephasing spectra for both the monomer and dimer constructs, the SNR of the rephasing spectra drops dramatically as the integration time is lowered. This drop in SNR occurs well before an integration time of 5 seconds is reached. This is somewhat expected

given the higher ordered contributions to nonlinear signals, which greatly reduces the magnitude of nonlinear components demodulated from the time varying fluorescent photon stream. These results also reveal that experimental acquisition times are doubtful to get below the timescale of tens-of-minutes, utilizing the presently discussed protocols, which reduces utility at the single molecule limit.

With the use of photon time and phase tagging on single molecules, errors present in the discrete detection can greatly influence experimental results given the relatively low number of photons detected. Therefore, the uncertainty in demodulated signal quadrature's resulting from low photon numbers requires quantification. This quantification is based on scaling of signal magnitude with the number of photons detected in the window  $T_{PT}$  of Eqns. 66 and 67. Assuming  $N$  photon detection events corresponding to  $N$  measurements of the quantity  $x$  ( $x_1, x_2, x_3 \dots x_N$ ) are statistically independent, then the distribution of measurement outcomes ( $x_1, x_2, x_3 \dots x_N$ ) obey gaussian statistics. In this case, the SNR of these measurements can be assigned as the ratio of the distribution average to the standard deviation of the distribution,  $SNR = \bar{x}/\sigma_x$ , which scales with  $\sqrt{N}$  [69]. To demonstrate that the nonlinear signals demodulated from single molecules are both valid and not subject to additional sources of error beyond statistical sampling uncertainty, we plot the demodulated magnitude of the real valued rephasing and nonrephasing signals versus the number of photons used to calculate the average. As can be seen in Fig. A.8, the magnitude of the real valued rephasing and nonrephasing signals scale as  $\sqrt{N}$  over the range of photon numbers examined. The standard deviation of the signal at each photon number  $N$  is weakly determined by means of typical error estimation. This occurs when the underlying probability distribution is not sufficiently sampled from. In this work, we have estimated the standard deviation of the underlying distribution by employing a bootstrapping approach which randomly samples, with replacement,

from the same underlying data set of  $N$  photons to form a more reliable estimate of the standard deviation at each photon number examined [69].

In order to assess the validity of the signals obtained via discrete time and phase tagging, an examination of the power dependence for both linear and nonlinear signals is necessary. This can effectively rule out detector saturation effects, photon pile-up and the possibility of dark counts forming the major contribution to the demodulated amplitudes. We expect that in our regime of single molecules, the number of photons  $N$  emitted from the sample is dominated by the square of the molecular dipole projected onto the electric field,  $N \propto \sum_n |\alpha(\omega_{ng})|^2 |\mu_{ng}|^2$ . Based on Eq. (53), we expect a linear dependence between the field intensity and magnitude of the linear signal. Therefore, if both the number of photons  $N$  and the linear signal magnitude both vary linearly with the field intensity, their ratio should be constant and independent of the incident field intensity. A non-constant ratio would imply signal artifacts are present to a significant degree. In the case of the nonlinear signals, the magnitude is proportional to the square of the field intensity, as shown in Eq. (60). In this instance, we expect the ratio of the demodulated nonlinear signal magnitude and incident field intensity to obey a linear relationship. Therefore, as the incident field intensity is attenuated, the magnitude of the nonlinear signal should scale linearly. A final test of both the linear and nonlinear signals validity is to attenuate

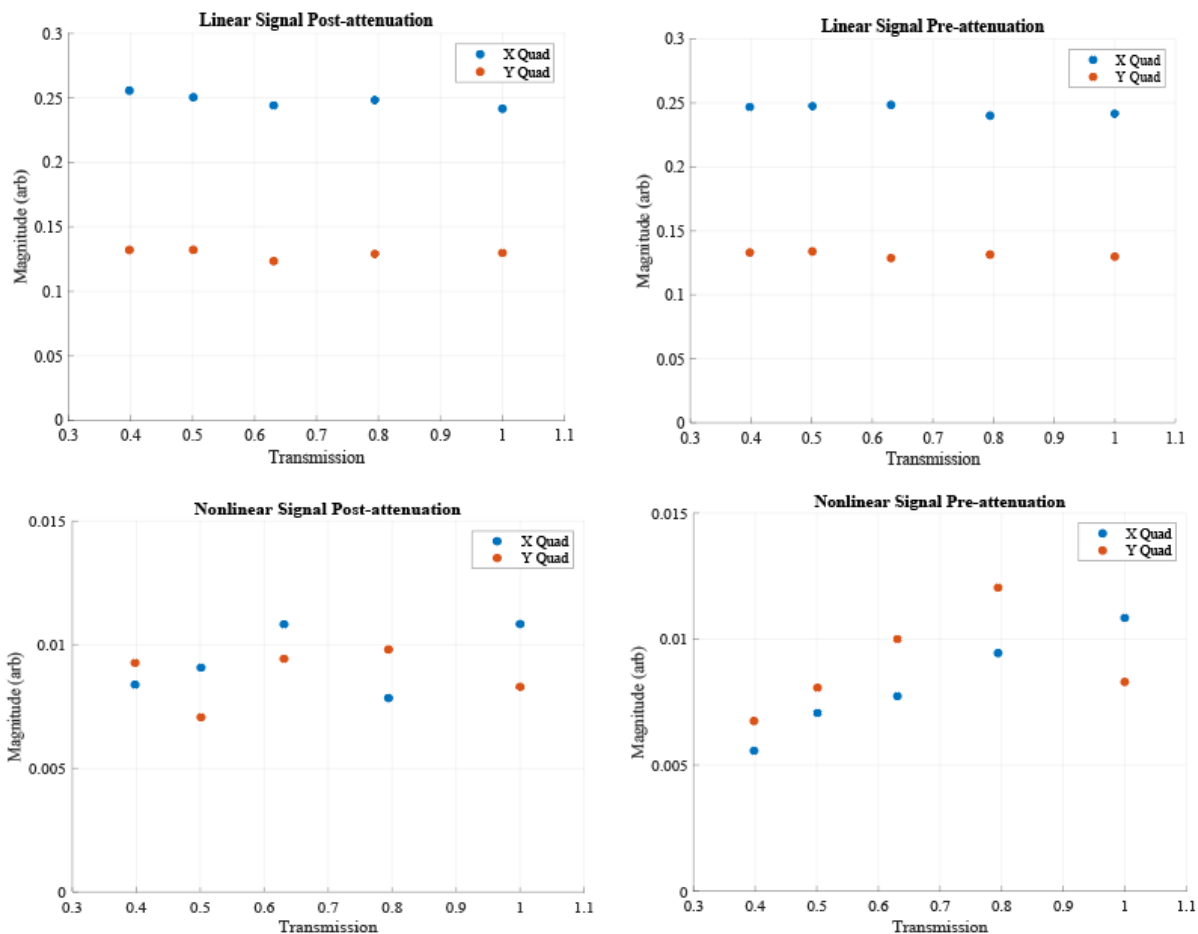


**Figure A.8** Scaling of Nonrephasing and Rephasing signals for variable photon number  $N$  used to average the signal from single molecules.

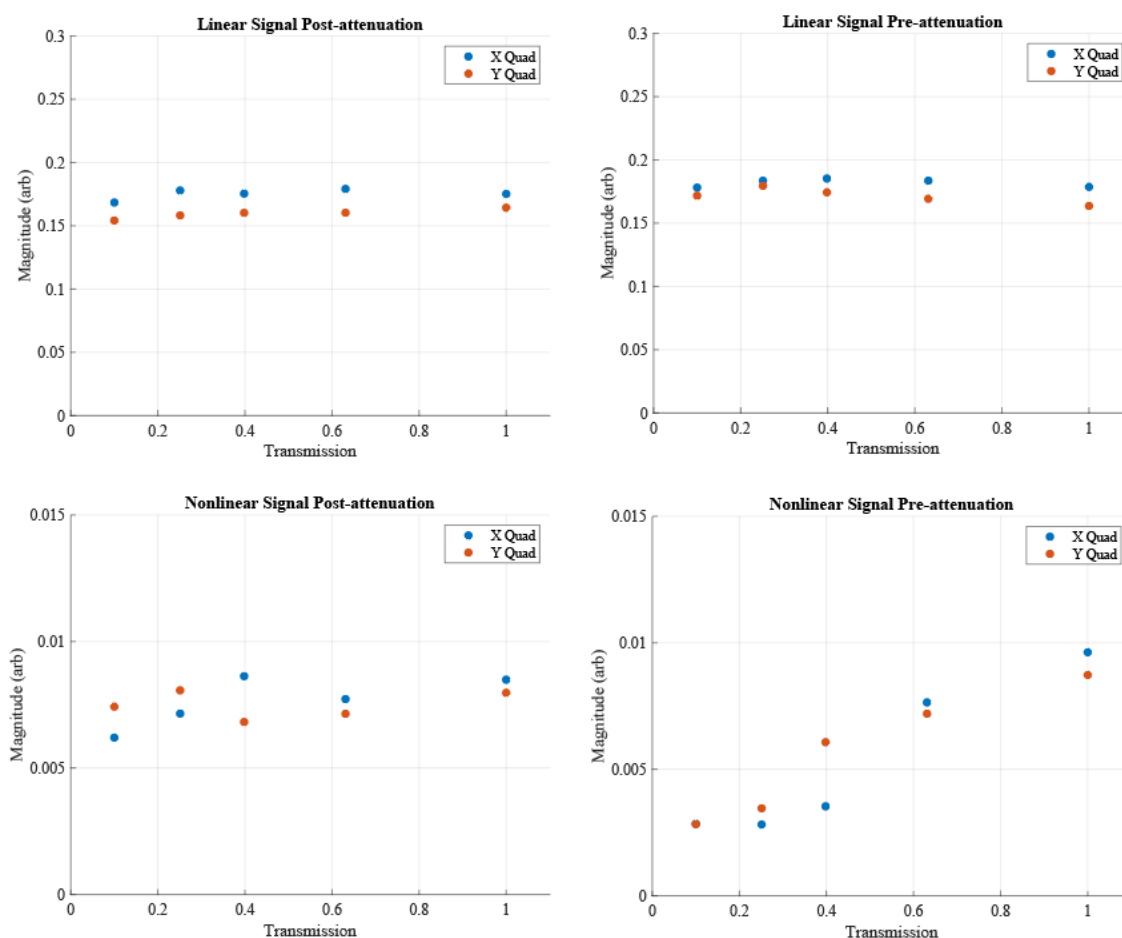
the emitted photon stream of a single molecule, holding excitation constant. This attenuation of the fluorescent photons can test whether the photon flux arriving at the detector is introducing saturation artifacts. Saturation and photon pile up can result in an artificial overestimate of the signal magnitudes. If no spurious effects are present, then the magnitude of both the linear and nonlinear signals should be constant when attenuating the emitted fluorescent photons. The signals are integrated over the same fixed number of photons at each stage of attenuation to guarantee a similarly reliable estimate of the quadrature magnitudes. Figure A.9 shows the linear and nonlinear signal magnitude scaling of a Cy3 monomer in dsDNA for both pre and post attenuation, corresponding to attenuation of the incident field intensity and emitted photon flux, respectively. Figure A.10 shows the same scaling for a Cy3 dimer in dsDNA. Both cases reveal the predicted trends in signal scaling, affirming that the measured signals are predominantly due to the time



varying fluorescence emitted from the excited state populations we prepare and not systematic artifacts of the experimental setup.



**Figure A.9** Scaling of the linear and nonlinear signal magnitudes from a Cy3 monomer in dsDNA. The pre and post attenuation reflects a reduction in the incident laser power and emitted photon flux, respectively.

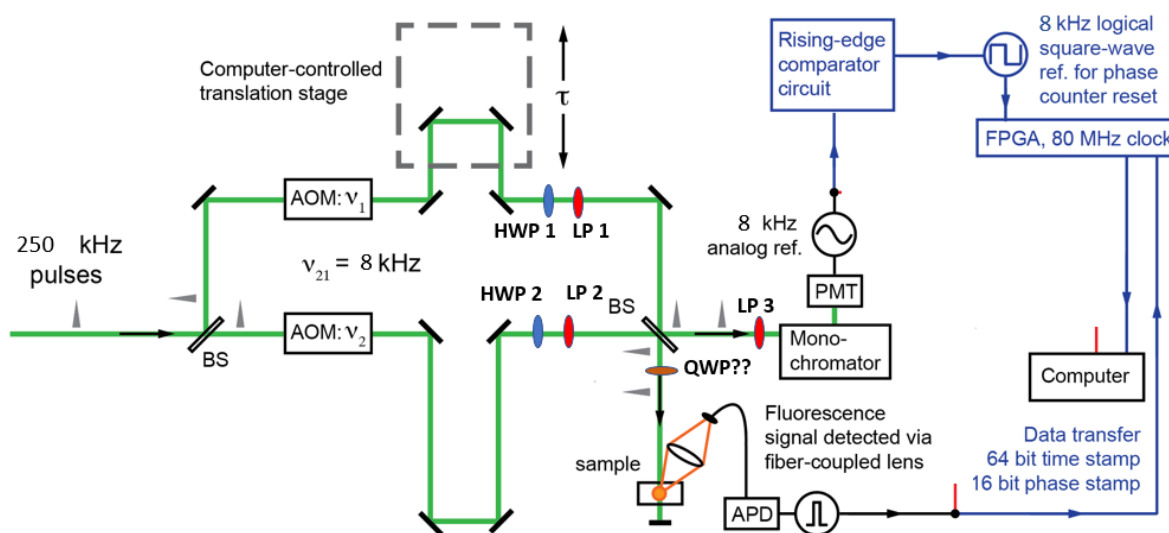


**Figure A.10** Scaling of the linear and nonlinear signal magnitudes from a Cy3 dimer in dsDNA. The pre and post attenuation reflects a reduction in the incident laser power and emitted photon flux, respectively.

### Section A.4.3 Time-correlated measurements of linear signals from single molecules under broadband excitation

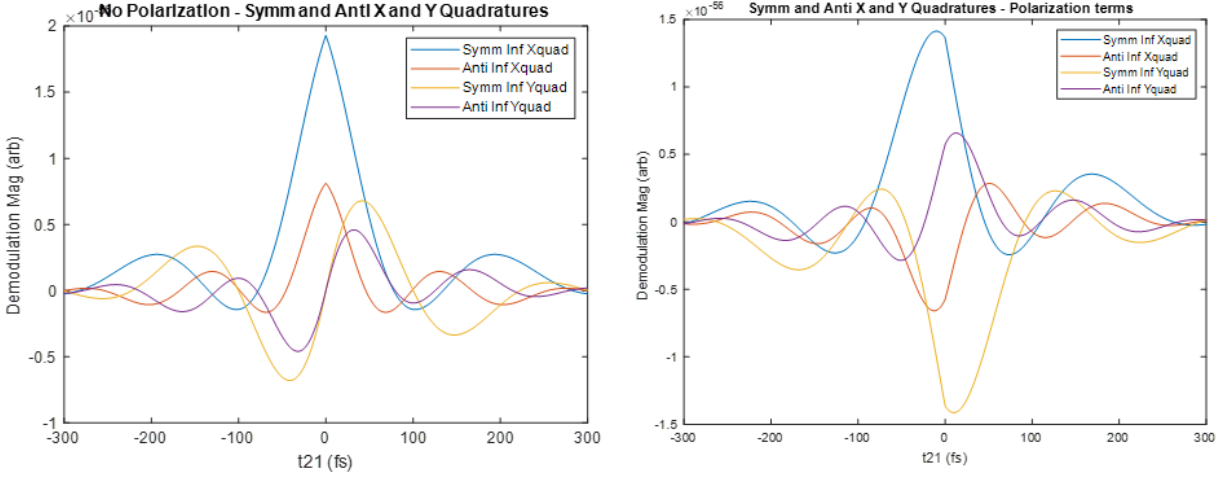
To test the possible limits on obtaining dynamical estimates of conformation in dsDNA systems, time-correlated measurements were taken on single molecules under broad-band excitation. The motivation for these investigations is the possible increase in sensitivity to conformational fluctuations afforded by a broader probe of the distribution describing the absorption spectra of  $i(\text{Cy}3)_2$  dimers. Additionally, simulated studies of polarization sweep

experiments utilizing broadband excitation suggested possible interdependency between the measured signal visibility and interpulse delay,  $t_{21}$ , within one MZI. The combination of enhanced spectral sampling with a broad source and potential encoding of conformation dynamics as function of interpulse delay led to the setup of such an experiment. A diagram of the experimental setup is shown in Fig. A.11. The visibility, or complex amplitude, of the demodulated linear signal components (Eq. 56) was analyzed for fluctuations about its mean value, the same analytical approach employed in Chapter 3 for PS-SMF experiments.



**Figure A.11** Experimental apparatus for broadband polarization-sweep measurements. Half-wave plates (HP), linear polarizers (LP) and a quarter waveplate (QWP) are labeled according to their position within the interferometer.

The simulated results of a non-rotating (left panel) versus rotating (right panel) broadband polarization experiment are shown as a function of  $t_{21}$  in Fig. A.12. The simulation assumes a constant geometry and handedness for the  $i(\text{Cy}3)_2$  dimer.



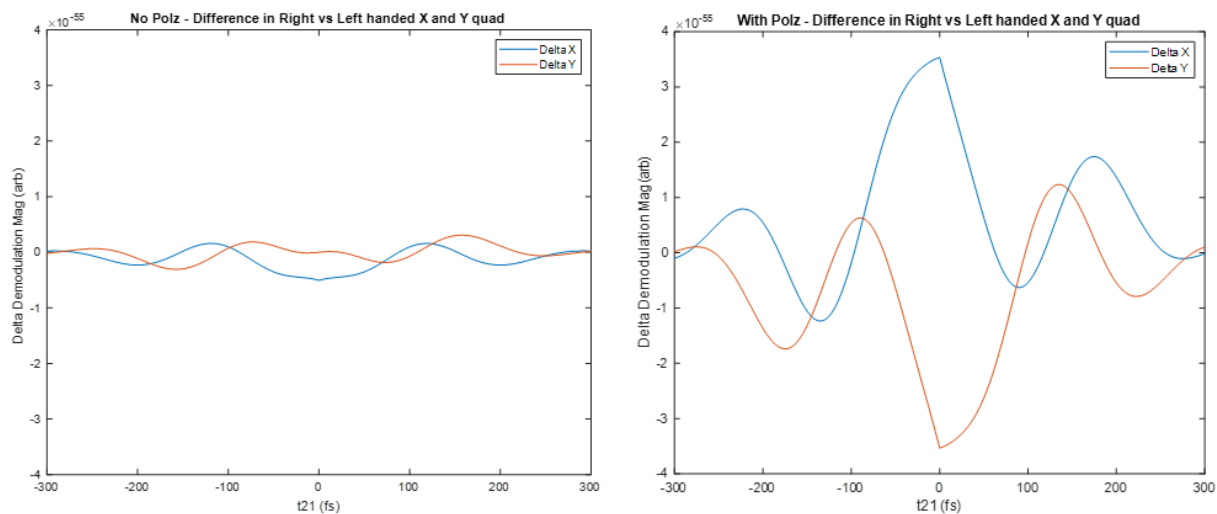
**Figure A.12** Linear signal components arising from either symmetric or anti-symmetric transitions, in either a rotating (right panel) or non-rotating (left panel) polarization scheme.

Eqns. 68 and 69 show the dependence of the signal arising from the symmetric versus antisymmetric manifolds of the coupled  $i(\text{Cy}3)_2$  dimer, in the non-rotating and rotating cases respectively. Eqn. 68 is equivalent to Eqn. 53 of the appendix, section 3, except that the orientation of the single molecule,  $\varphi_0$ , causes the signal magnitude to drop as overlap with the linear polarization of the laser is lost. The derivation of this signal is identical to the approach shown for Eqn. 53 of the appendix, section 3, but first a dipole-field inner product is considered to account for the singular orientation of all dipoles probed in the experiment.  $A^+$  is the signal from the symmetric manifold while  $A^-$  is the signal from the anti-symmetric manifold.

$$A^\pm(\varphi, \tau_{21}) = \sum_n |\mu_{ng}^\pm \cdot E(\omega)|^2 [1 + \cos(\omega\tau_{21} - \Omega_m T)] \cos^2(\varphi_0) \quad (68)$$

$$A_{Rot}^\pm(\varphi, \tau_{21}) = \sum_n |\mu_{ng}^\pm \cdot E(\omega)|^2 [1 \pm 2 \sin(\omega\tau_{21} - \Omega_m T + 2\varphi_0)] \quad (69)$$

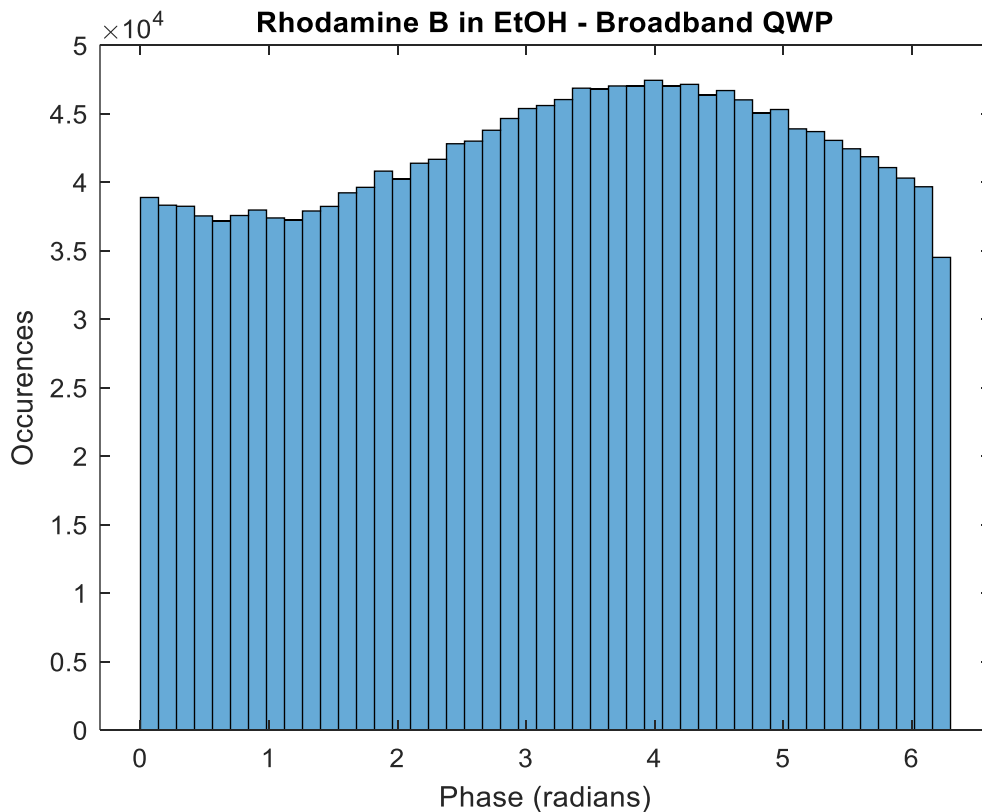
Rotation of the lasers polarization results in the dipole orientation giving rise to an overall signal phase, rather than an overall signal diminishment in the case of a non-rotating polarization. In both cases, one can ask whether the measured signal from a single molecule would fluctuate if the molecule changed its conformation. To test this, two oppositely handed geometries of a coupled  $i(\text{Cy}3)_2$  dimer were assumed in the simulation and the difference between the signal quadrature's arising from one conformer versus the other were calculated. This was done for both the rotating and non-rotating case. It can be seen in Fig. A.13 that the rotating case appears to have much greater signal contrast between the oppositely handed conformers, suggesting sensitivity in the signal visibility to the simulated conformational change occurring during the course of the experiment.



**Figure A.13** The signal contrast between two oppositely handed conformers in the case of rotating (right panel) versus non-rotating (left panel) polarization under broadband excitation. Direct differences in demodulated quadrature data are plotted.

However, experimental implementation of this rotating polarization scheme resulted in the measurement of a persistent signal visibility when exciting a sample of isotropic rhodamine (Fig

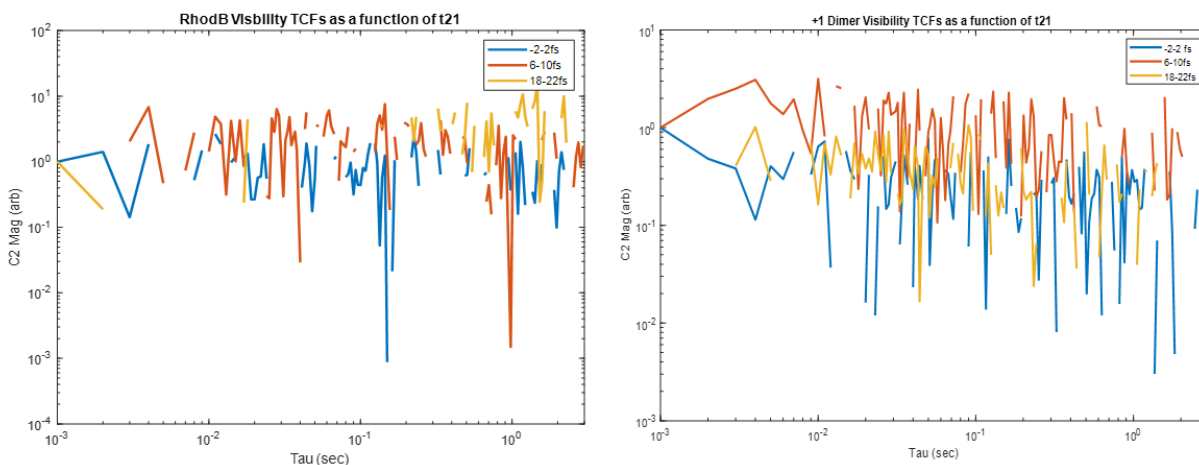
A.14). This is unexpected within our model and not consistent with the isotropic nature of rhodamine in solution. Notably, this persistent visibility only occurred when employing a broadband QWP, which maintains a quarter wave of retardance across the entire laser bandwidth. No tuning of the relative polarization in the two arms MZI could eliminate this artifact, i.e. there was no interpulse interference contributing to the signal modulation as one might expect in the non-rotating scheme. This was confirmed by removal of the QWP. Since the aim with a rotating polarization scheme is to examine fluctuations of the signal visibility, this persistent contribution from the role of the broadband QWP prevented reliable experimental results from being obtained. It should be noted that multiple broadband QWPs, with varying degrees of accuracy and precision, were tested to try and eliminate this effect. No such effect is observed in the PS-SMF experiments of Chapter 3 where a narrowband QWP is employed for use at 532nm exclusively.



**Figure A.14** A histogram of phase-tags obtained on an isotropic sample of rhodamine at  $t_{21} = 0$  using the experimental setup shown in Fig. A.11. A ~15% modulation of the signal can be observed.

The inability to reliably perform rotating polarization experiments under broadband excitation led to the test of both non-rotating time-correlated experiments (conventional parallel polarization setup) and a time-correlated experiment reminiscent of linear dichroism studies, where the two arms of the MZI are held cross polarized but no QWP is employed, much like Phelps et.al [13].

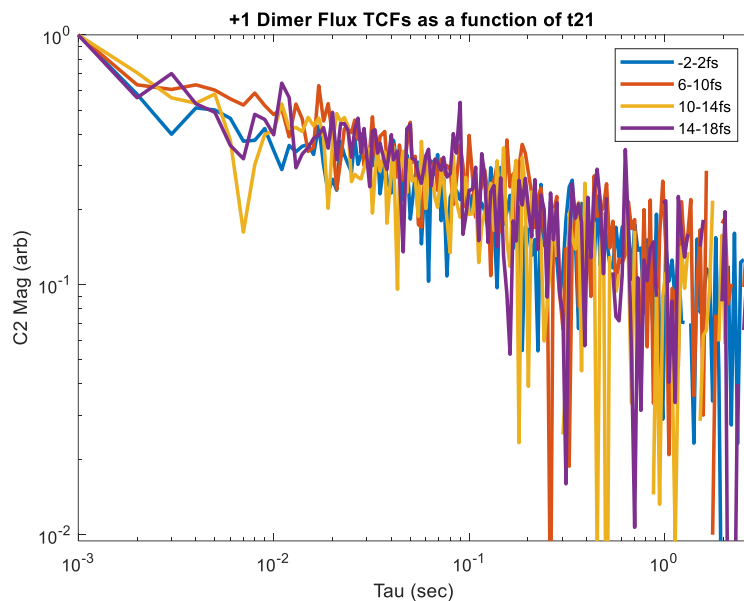
The non-rotating experiment attempts to leverage the unique set of spectral interference patterns, which evolve at the modulation frequency, as a function of the interpulse delay,  $t_{21}$ . The hypothesis in this case is, if there is an optimal interpulse delay which produces the best (or worst) spectral overlap with the fluctuating absorption profile of a single molecule, then the time-correlated statistics of the signal visibility should yield unique outcomes as a function of  $t_{21}$ . In essence, this experiment attempts to encode conformational dynamics in the signal fluctuations observed as a function of  $t_{21}$ . The results of time-correlation function analyses in the non-rotating case are shown for both the rhodamine control and a +1 dimer in Fig. A.15.



**Figure A.15** Non-rotating polarization time-correlation function analysis of a control (rhodamine) and single molecule containing a Cy3 dimer. Data are binned within a narrow range of  $t_{21}$  to boost the statistics available at low signal levels.

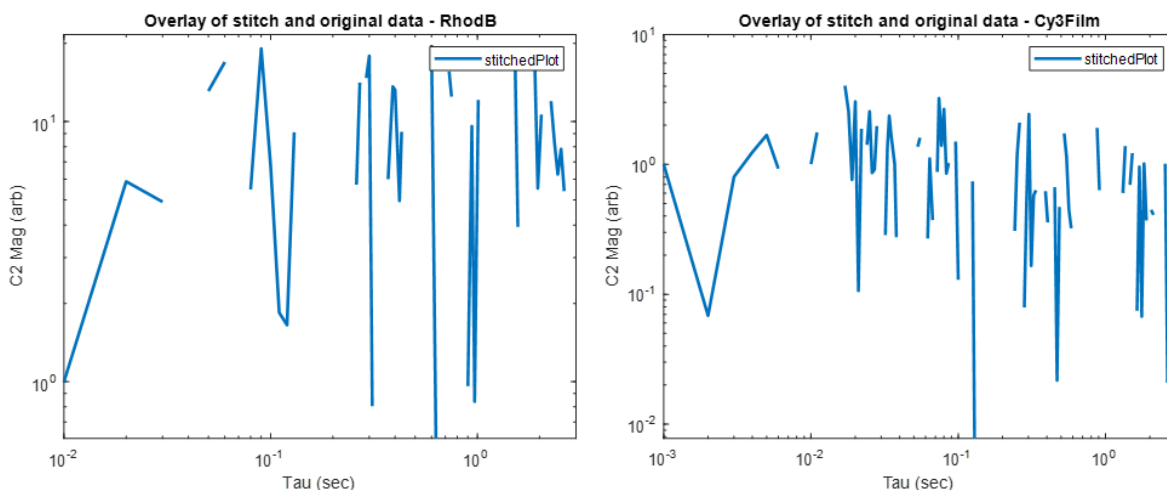
The TCFs above are complete noise in the case of the rhodamine control and are very weakly resolved in the case of the +1 dimer. Notably these data are integrated to a binning window of 1ms, which typically results in high SNR TCFs in the more conventional PS-SMF experiment, at similar levels of fluorescent photon flux. These initial experimental attempts suggested that a non-rotating time-correlated approach would yield noisy results at best, without much sensitivity to the interpulse delay. Due to the noise profile present, we choose to also examine the fluctuations in the pure photon flux, which in previous single molecule studies has shown to be sensitive to experimental conditions and therefore might show distinction between TCFs as a function of  $t_{21}$ . This was done to motivate refinement or improvement of such experimental results, in the case that a clear  $t_{21}$  dependence was observed. The pure flux TCFs as a function of  $t_{21}$  are shown in Fig. A.16. The SNR of these TCFs is certainly better than those performed on the visibility signals in Fig. A.15. But there is no clear separation between the TCFs in the pure flux case, further suggesting that any inherent dependence of the measured fluctuations as a function of  $t_{21}$  is either lacking or too weak to resolve. However, this test is partially incomplete given the limited number of molecules included in the statistics. But nonetheless, no obvious distinction in the TCFs as a function of  $t_{21}$  emerges to motivate further investigation of such phenomena.





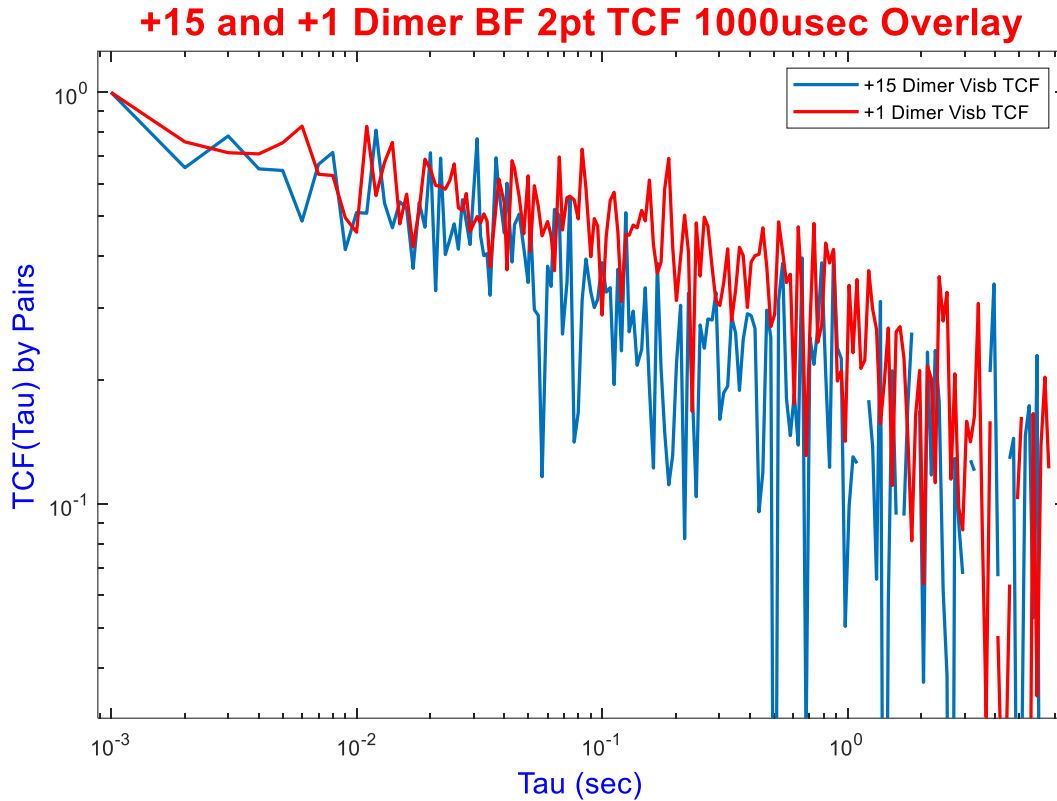
**Figure A.16** Non-rotating polarization time-correlation function analysis of photon flux for a single molecule containing a Cy3 dimer. Data are binned within a narrow range of  $t_{21}$  to boost the statistics available at low signal levels.

The final experimental protocol tested during time-correlated measurements employing a broadband source was two cross-polarized pulses in the two arms of an MZI and no QWP, examined over the range of various  $t_{21}$ . This experiment is much like linear dichroism studies touched on previously, but notably lacks any background signal visibility during control experiments which complicated the interpretation of experiments employing broadband QWP and holds potential for uniquely encoded dynamics as a function of interpulse delay (unlike previous linear dichroism experiments). Control experiments were performed on both rhodamine and a stretched Cy3 film. Both cases show no correlation in the TCFs of Fig. A.17.



**Figure A.17** Two-point TCFs for both control data sets, in the cross-polarized case, binned at 1ms. Rhodamine (left) and Cy3 film (right).

The same experiment was then performed on single molecules containing a Cy3 dimer at both the +1 and +15 position. The TCFs for both cases when  $t_{21} = 0$  are shown in Fig. A.18. This case is the first and only experimental protocol to display time correlated fluctuations in the signal visibility, when employing a broadband source. While binning at 1ms results in distinguishable TCFs above the noise floor, these results are significantly lower in their SNR than similar experiments performed for these same constructs with the PS-SMF instrument. Additionally, there is weak separation between the +1 and +15 TCFs, which in previous studies were shown to be distinctly different in their conformational landscape.



**Figure A.18** Two-point TCFs of the signal visibility for a +1 and +15 dimer in the cross-polarized experimental protocol. Binned at 1ms with  $t_{21} = 0$ .

Whether an exhaustive sampling of the  $t_{21}$  domain would yield unique insights for this cross-polarized protocol remains undetermined, as those experiments were not attempted due to the inability to perform accurate retiming of the interferometric signals prior to time-correlated analyses. Ultimately it was deemed unlikely that time-correlated experiments of signal visibility under broadband excitation would yield meaningful results above and beyond those already available from CW PS-SMF experiments.

## 5. CONCLUSION

The results of these studies demonstrate the feasibility of detecting linear and nonlinear signals from single molecules under broadband excitation, in either one- or two-dimensional protocols. Notably, the detection of nonlinear signals from single molecules is not new [107]. However, multi-dimensional measurements on a single molecule do appear to be the first of their kind. Additionally, these studies place limits on SNR as a function of integration time, which informs the potential for dynamically resolved spectra to be obtained. Time correlated analysis of linear signals failed to produce robust statistic or meaningfully different outcomes under a variety of experimental protocols, although improvements to the instrument or uniquely creative implementation of alternate approaches may still prove useful.

The possible applications of this work are likely outside the scope of single molecule biophysics, where fluctuations are rapid and unlikely to be interferometric sampled on relevant time scales. The possibility to address single crystals or semiconductor nanostructures within an ensemble of statically heterogeneous members could prove useful, although advanced approaches already exist for characterization of most solid-state materials. Systems exhibiting slow dynamics could be examined here using fully sampled interferometric techniques, or sparsely sampled compressed sensing approaches, should the relevant timescales be sub-minutes. It appears unlikely that time-correlated analyses of these sort will exceed the utility offered by simpler, more conventional CW experiments. Systems with more significant line shape fluctuations, or generally broader excited state manifolds might be well probed by such approaches. But again, dynamics would need to be sufficiently slow to obtain well resolved photon statistics for each configuration of the system.

- [1] P. H. Von Hippel, N. P. Johnson, and A. H. Marcus, ‘Fifty years of DNA “breathing”: Reflections on old and new approaches’, *Biopolymers*, vol. 99, no. 12, pp. 923–954, 2013, doi: 10.1002/bip.22347.
- [2] M. P. Printz and P. H. von Hippel, ‘On the Kinetics of Hydrogen Exchange in Deoxyribonucleic Acid, pH and Salt Effects’, *Biochemistry*, vol. Vol 7, no. No. 9, pp. 3194–3206, 1964.
- [3] B. McConnell and P. H. Von Hippel, ‘Hydrogen Exchange as a Probe of the Dynamic Structure of DNA II. Effects of Base Composition and Destabilizing Salts’, 1970. doi: J. Mol. Bio.
- [4] N. G. Nossal and B. M. Peterlin, ‘DNA replication by bacteriophage T4 proteins. The T4, 43, 32, 44-62 and 45 proteins are required for strand displacement synthesis at nicks in duplex DNA’, *Journal of Biological Chemistry*, vol. 254, no. 13, pp. 6032–6037, 1979, doi: 10.1016/s0021-9258(18)50515-9.
- [5] C. F. Morris, N. K. Sinha, and B. M. Alberts, ‘Reconstruction of bacteriophage T4 DNA replication apparatus from purified components: Rolling circle replication following de novo chain initiation on a single-stranded circular DNA template (replication forrK/riDonucleoside-tripnosphate-dependent priming/DNA strand’, 1975.
- [6] J. D. McGhee and P. H. von Hippel, ‘Formaldehyde as a Probe of DNA Structure. IV. Mechanism of the Initial Reaction of Formaldehyde with DNA.’, UTC, 1977. [Online]. Available: <https://pubs.acs.org/sharingguidelines>
- [7] J. D. McGhee and P. H. von Hippel, ‘Formaldehyde as a Probe of DNA Structure. II. Reaction with Endocyclic Imino Groups of DNA Basest’, UTC, 1975. [Online]. Available: <https://pubs.acs.org/sharingguidelines>
- [8] W. Lee, J. P. Gillies, D. Jose, B. A. Israels, P. H. Von Hippel, and A. H. Marcus, ‘Single-molecule FRET studies of the cooperative and non-cooperative binding kinetics of the bacteriophage T4 single-stranded DNA binding protein ( gp32 ) to ssDNA lattices at replication fork junctions’, vol. 44, no. 22, pp. 10691–10710, 2016, doi: 10.1093/nar/gkw863.
- [9] T. C. Mueser, J. M. Hinerman, J. M. Devos, R. A. Boyer, and K. J. Williams, ‘Structural analysis of bacteriophage T4 DNA replication: A review in the Virology Journal series on bacteriophage T4 and its relatives’, *Virology Journal*, vol. 7. 2010. doi: 10.1186/1743-422X-7-359.
- [10] T. L. Carus, D. R. Natura, S. Greek, R. Empire, T. S. Chymist, and R. Boyle, ‘Introduction to Single Molecule Imaging and Mechanics : Seeing and Touching Molecules One at a Time’.
- [11] W. E. Moerner and L. Kador, ‘Optical detection and spectroscopy of single molecules in a solid’, *Phys Rev Lett*, vol. 62, no. 21, pp. 2535–2538, 1989, doi: 10.1103/PhysRevLett.62.2535.

- [12] T. Ha, T. H. Enderle, D. F. Ogletreet, D. S. Chemla, P. R. Selvin, and S. Weiss, ‘Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor’, 1996.
- [13] C. Phelps, W. Lee, D. Jose, P. H. von Hippel, and A. H. Marcus, ‘Single-molecule FRET and linear dichroism studies of DNA breathing and helicase binding at replication fork junctions’, *Proceedings of the National Academy of Sciences*, vol. 110, no. 43, pp. 17320–17325, 2013, doi: 10.1073/pnas.1314862110.
- [14] B. Israels, C. S. Albrecht, A. Dang, M. Barney, P. H. Von Hippel, and A. H. Marcus, ‘Sub-millisecond conformational transitions of short single-stranded DNA lattices by photon correlation single-molecule FRET’.
- [15] M. F. Juetten *et al.*, ‘Didemnin B and ternatin-4 differentially inhibit conformational changes in eEF1A required for aminoacyl-tRNA accommodation into mammalian ribosomes’, *Elife*, vol. 11, 2022, doi: 10.7554/eLife.81608.
- [16] S. R. Hansen, D. S. White, M. Scalf, I. R. Corrêa, L. M. Smith, and A. A. Hoskins, ‘Multi-step recognition of potential 5’ splice sites by the *Saccharomyces cerevisiae* U1 snRNP’, *Elife*, vol. 11, Aug. 2022, doi: 10.7554/eLife.70534.
- [17] C. P. Lapointe, R. Grosely, A. G. Johnson, J. Wang, I. S. Fernández, and J. D. Puglisi, ‘Dynamic competition between SARS-CoV-2 NSP1 and mRNA on the human ribosome inhibits translation initiation’, doi: 10.1073/pnas.2017715118/-/DCSupplemental.
- [18] M. Y. Lee *et al.*, ‘Fluorescent labeling of abundant reactive entities (FLARE) for cleared-tissue and super-resolution microscopy’, *Nature Protocols*, vol. 17, no. 3. Nature Research, pp. 819–846, Mar. 01, 2022. doi: 10.1038/s41596-021-00667-2.
- [19] A. M. Koester, K. Tao, M. Szczepaniak, M. J. Rames, and X. Nan, ‘Nanoscope Spatial Association between Ras and Phosphatidylserine on the Cell Membrane Studied with Multicolor Super Resolution Microscopy’, *Biomolecules*, vol. 12, no. 8, Aug. 2022, doi: 10.3390/biom12081033.
- [20] L. A. Barner *et al.*, ‘Multiresolution nondestructive 3D pathology of whole lymph nodes for breast cancer staging’, *J Biomed Opt*, vol. 27, no. 03, Mar. 2022, doi: 10.1117/1.jbo.27.3.036501.
- [21] K. C. Neuman and A. Nagy, ‘Single-molecule force spectroscopy: Optical tweezers, magnetic tweezers and atomic force microscopy’, *Nature Methods*, vol. 5, no. 6. pp. 491–505, Jun. 2008. doi: 10.1038/nmeth.1218.
- [22] D. R. Jacobson and T. T. Perkins, ‘Free-energy changes of bacteriorhodopsin point mutants measured by single-molecule force spectroscopy’, doi: 10.1073/pnas.2020083118/-/DCSupplemental.
- [23] D. T. Edwards, M.-A. Leblanc, and T. T. Perkins, ‘Modulation of a protein-folding landscape revealed by AFM-based force spectroscopy notwithstanding instrumental limitations’, doi: 10.1073/pnas.2015728118/-/DCSupplemental.
- [24] D. Heussman, J. Kittell, P. H. Von Hippel, and A. H. Marcus, ‘Temperature-dependent local conformations and conformational distributions of cyanine dimer labeled single-stranded-

- double-stranded DNA junctions by 2D fluorescence spectroscopy’, *Journal of Chemical Physics*, vol. 156, no. 4, 2022, doi: 10.1063/5.0076261.
- [25] D. Heussman, J. Kittell, L. Kringle, A. Tamimi, P. H. Von Hippel, and A. H. Marcus, ‘Measuring local conformations and conformational disorder of ( Cy3 ) 2 dimer labeled DNA fork junctions using absorbance , circular dichroism and two- dimensional fluorescence spectroscopy I . Introduction’, pp. 1–38.
- [26] A. H. Marcus, D. Heussman, J. Maurer, C. S. Albrecht, P. Herbert, and P. H. Von Hippel, ‘Studies of Local DNA Backbone Conformation and Conformational Disorder Using Site-Specific Exciton-Coupled Dimer Probe Spectroscopy’, *Annu Rev Phys Chem*, 2023, doi: 10.1146/annurev-physchem-090419.
- [27] L. Kringle *et al.*, ‘Temperature-dependent conformations of exciton-coupled Cy3 dimers in double-stranded DNA’, *Journal of Chemical Physics*, vol. 148, no. 8, 2018, doi: 10.1063/1.5020084.
- [28] W. Lee, P. H. Von Hippel, and A. H. Marcus, ‘Internally labeled Cy3 / Cy5 DNA constructs show greatly enhanced photo-stability in single-molecule FRET experiments’, vol. 42, no. 9, pp. 5967–5977, 2014, doi: 10.1093/nar/gku199.
- [29] N. F. Dupuis, E. D. Holmstrom, and D. J. Nesbitt, ‘Single-molecule kinetics reveal cation-promoted DNA duplex formation through ordering of single-stranded helices’, *Biophys J*, vol. 105, no. 3, pp. 756–766, Aug. 2013, doi: 10.1016/j.bpj.2013.05.061.
- [30] E. D. Holmstrom, N. F. Dupuis, and D. J. Nesbitt, ‘Kinetic and Thermodynamic Origins of Osmolyte-In fl uenced Nucleic Acid Folding’, 2015, doi: 10.1021/jp512491n.
- [31] N. F. Dupuis, E. D. Holmstrom, and D. J. Nesbitt, ‘Single-molecule kinetics reveal cation-promoted DNA duplex formation through ordering of single-stranded helices’, *Biophys J*, vol. 105, no. 3, pp. 756–766, Aug. 2013, doi: 10.1016/j.bpj.2013.05.061.
- [32] Bio 3400, ‘DNA Structure’, Bio3400. (n.d.). DNA Structure. How much DNA do you have? <http://bio3400.nicerweb.com/Locked/media/ch10/DNA.html> .
- [33] S. Gellerl, ‘X-RAY STRUCTURAL CRYSTALLOGRAPHY’. [Online]. Available: [www.annualreviews.org](http://www.annualreviews.org)
- [34] H. Ki, K. Young Oang, J. Kim, and H. Ihee, ‘Ultrafast X-Ray Crystallography and Liquidography’, *The Annual Review of Physical Chemistry is online at*, vol. 68, pp. 473–97, 2017, doi: 10.1146/annurev-physchem.
- [35] B. Von Ardenne, M. Mechelke, and H. Grubmüller, ‘Structure determination from single molecule X-ray scattering with three photons per image’, *Nat Commun*, vol. 9, no. 1, pp. 1–9, 2018, doi: 10.1038/s41467-018-04830-4.
- [36] T. Xie, T. Saleh, P. Rossi, and C. G. Kalodimos, ‘Conformational states dynamically populated by a kinase determine its function’, *Science (1979)*, vol. 370, no. 6513, Oct. 2020, doi: 10.1126/science.abc2754.

- [37] A. V. Reshetnyak *et al.*, ‘Mechanism for the activation of the anaplastic lymphoma kinase receptor’, *Nature*, vol. 600, no. 7887, pp. 153–157, Dec. 2021, doi: 10.1038/s41586-021-04140-8.
- [38] K. H. Gardner and L. E. Kay, ‘THE USE OF 2 H, 13 C, 15 N MULTIDIMENSIONAL NMR TO STUDY THE STRUCTURE AND DYNAMICS OF PROTEINS’, 1998. [Online]. Available: [www.annualreviews.org](http://www.annualreviews.org)
- [39] M. E. Wall, S. C. Gallagher, and J. Trehwella, ‘LARGE-SCALE SHAPE CHANGES IN PROTEINS AND MACROMOLECULAR COMPLEXES’, 2000. [Online]. Available: [www.annualreviews.org](http://www.annualreviews.org)
- [40] M. E. Mäeots and R. I. Enchev, ‘Structural dynamics: review of time-resolved cryo-EM’, *Acta Crystallographica Section D: Structural Biology*, vol. 78. International Union of Crystallography, pp. 927–935, Aug. 01, 2022. doi: 10.1107/S2059798322006155.
- [41] R. M. Glaeser, ‘How Good Can Single-Particle Cryo-EM Become? What Remains Before It Approaches Its Physical Limits?’, 2019, doi: 10.1146/annurev-biophys-070317.
- [42] J. Jumper *et al.*, ‘Highly accurate protein structure prediction with AlphaFold’, *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.
- [43] S. Torino, M. Dhurandhar, A. Stroobants, R. Claessens, and R. G. Efremov, ‘Time-resolved cryo-EM using a combination of droplet microfluidics with on-demand jetting’, *Nat Methods*, Aug. 2023, doi: 10.1038/s41592-023-01967-z.
- [44] M. Holm *et al.*, ‘mRNA decoding in human is kinetically and structurally distinct from bacteria’, *Nature*, 2023, doi: 10.1038/s41586-023-05908-w.
- [45] T. Xie, T. Saleh, P. Rossi, D. Miller, and C. G. Kalodimos, ‘Imatinib can act as an allosteric activator of Abl kinase’, 2021.
- [46] J. R. Widom, ‘Local Conformations and Excited State Dynamics of Porphyrins and Nucleic Acids By 2-Dimensional Fluorescence Spectroscopy’, *Doctoral Dissertation*, no. December, 2013.
- [47] J. R. Widom, N. P. Johnson, P. H. Von Hippel, and A. H. Marcus, ‘Solution conformation of 2-aminopurine dinucleotide determined by ultraviolet two-dimensional fluorescence spectroscopy’, *New J Phys*, vol. 15, pp. 1–16, 2013, doi: 10.1088/1367-2630/15/2/025028.
- [48] P. F. Tekavec, G. A. Lott, and A. H. Marcus, ‘Fluorescence-detected two-dimensional electronic coherence spectroscopy by acousto-optic phase modulation’, *Journal of Chemical Physics*, vol. 127, no. 21, pp. 1–21, 2007, doi: 10.1063/1.2800560.
- [49] K. S. Wilson and C. Y. Wong, ‘Calibrating a spatially encoded time delay for transient absorption spectroscopy’, pp. 3–7.
- [50] X. Fu, H. Kaur, M. L. Rodgers, E. J. Montemayor, S. E. Butcher, and A. A. Hoskins, ‘Identification of transient intermediates during spliceosome activation by single molecule fluorescence microscopy’, *Proc Natl Acad Sci U S A*, vol. 119, no. 48, Nov. 2022, doi: 10.1073/pnas.2206815119.



- [51] N. F. Dupuis, E. D. Holmstrom, and D. J. Nesbitt, ‘Single-molecule kinetics reveal cation-promoted DNA duplex formation through ordering of single-stranded helices’, *Biophys J*, vol. 105, no. 3, pp. 756–766, Aug. 2013, doi: 10.1016/j.bpj.2013.05.061.
- [52] L. Piatkowski, E. Gellings, and N. F. Van Hulst, ‘Broadband single-molecule excitation spectroscopy’, *Nat Commun*, vol. 7, pp. 1–9, 2016, doi: 10.1038/ncomms10411.
- [53] G. U. Nienhaus, ‘Single-molecule fluorescence studies of protein folding.’, *Methods Mol Biol*, vol. 490, pp. 311–337, 2009, doi: 10.1007/978-1-59745-367-7\_13.
- [54] D. Brinks *et al.*, ‘Ultrafast dynamics of single molecules’, *Chem Soc Rev*, vol. 43, no. 8, pp. 2476–2491, 2014, doi: 10.1039/c3cs60269a.
- [55] A. S. Backer, A. S. Biebricher, G. A. King, G. J. L. Wuite, I. Heller, and E. J. G. Peterman, ‘Single-molecule polarization microscopy of DNA intercalators sheds light on the structure of S-DNA’, no. March, 2019.
- [56] C. Phelps, B. Israels, D. Jose, M. C. Marsh, P. H. von Hippel, and A. H. Marcus, ‘Using microsecond single-molecule FRET to determine the assembly pathways of T4 ssDNA binding protein onto model DNA replication forks’, *Proceedings of the National Academy of Sciences*, vol. 114, no. 18, pp. E3612–E3621, 2017, doi: 10.1073/pnas.1619819114.
- [57] M. C. Fink, K. V. Adair, M. G. Guenza, and A. H. Marcus, ‘Translational diffusion of fluorescent proteins by molecular fourier imaging correlation spectroscopy’, *Biophys J*, vol. 91, no. 9, pp. 3482–3498, 2006, doi: 10.1529/biophysj.106.085712.
- [58] M. K. Knowles, M. G. Guenza, R. A. Capaldi, and A. H. Marcus, ‘Cytoskeletal-assisted dynamics of the mitochondrial reticulum in living cells’, *Proceedings of the National Academy of Sciences*, vol. 99, no. 23, pp. 14772–14777, 2002, doi: 10.1073/pnas.232346999.
- [59] V. Glembockyte, L. Grabenhorst, K. Trofymchuk, and P. Tinnefeld, ‘DNA Origami Nanoantennas for Fluorescence Enhancement’, *Acc Chem Res*, vol. 54, no. 17, pp. 3338–3348, 2021, doi: 10.1021/acs.accounts.1c00307.
- [60] E. R. Beyerle and M. G. Guenza, ‘Identifying the leading dynamics of ubiquitin: A comparison between the tICA and the LE4PD slow fluctuations in amino acids’ position’, *J Chem Phys*, vol. 155, no. 24, p. 244108, Dec. 2021, doi: 10.1063/5.0059688.
- [61] M. G. Saunders and G. A. Voth, ‘Coarse-graining methods for computational biology’, *Annu Rev Biophys*, vol. 42, no. 1, pp. 73–93, May 2013, doi: 10.1146/annurev-biophys-083012-130348.
- [62] J. Wang *et al.*, ‘Twisting and swiveling domain motions in Cas9 to recognize target DNA duplexes, make double-strand breaks, and release cleaved duplexes’, *Frontiers in Molecular Biosciences*, vol. 9. Frontiers Media S.A., Jan. 09, 2023. doi: 10.3389/fmolb.2022.1072733.
- [63] G. A. Fitzgerald, D. S. Terry, A. L. Warren, M. Quick, J. A. Javitch, and S. C. Blanchard, ‘Quantifying secondary transport at single-molecule resolution’, *Nature*, vol. 575, no. 7783, pp. 528–534, Nov. 2019, doi: 10.1038/s41586-019-1747-5.

- [64] W. B. Asher *et al.*, ‘Single-molecule FRET imaging of GPCR dimers in living cells’, *Nat Methods*, vol. 18, no. 4, pp. 397–405, Apr. 2021, doi: 10.1038/s41592-021-01081-y.
- [65] C. P. Lapointe *et al.*, ‘eIF5B and eIF1A reorient initiator tRNA to allow ribosomal subunit joining’, *Nature*, vol. 607, no. 7917, pp. 185–190, Jul. 2022, doi: 10.1038/s41586-022-04858-z.
- [66] P. H. Von Hippel and E. Delagoutte, ‘Review A General Model for Nucleic Acid Helicases and Their “Coupling” within Macromolecular Machines Most well-studied proteins that interact with ssNA’, 2001.
- [67] E. M. S. Stennett, N. Ma, A. Van Der Vaart, and M. Levitus, ‘Photophysical and dynamical properties of doubly linked Cy3-DNA constructs’, *Journal of Physical Chemistry B*, vol. 118, no. 1, pp. 152–163, 2014, doi: 10.1021/jp410976p.
- [68] L. Kringle, N. P. D. Sawaya, J. Widom, C. Adams, M. G. Raymer, and A. H. Marcus, ‘Temperature-dependent conformations of exciton-coupled Cy3 dimers in double-stranded DNA’, no. Cd.
- [69] A. Tamimi, T. Landes, J. Lavoie, B. J. Smith, M. G. Raymer, and A. H. Marcus, ‘Fluorescence-detected Fourier transform electronic spectroscopy by photon counting phase-tagging’, pp. 1–22.
- [70] C. Phelps, B. Israels, M. C. Marsh, P. H. Von Hippel, and A. H. Marcus, ‘Using Multiorder Time-Correlation Functions (TCFs) to Elucidate Biomolecular Reaction Pathways from Microsecond Single-Molecule Fluorescence Experiments’, *Journal of Physical Chemistry B*, vol. 120, no. 51, pp. 13003–13016, 2016, doi: 10.1021/acs.jpcc.6b08449.
- [71] M. Dhar, J. A. Dickinson, and M. A. Berg, ‘Efficient, nonparametric removal of noise and recovery of probability distributions from time series using nonlinear-correlation functions: Additive noise’.
- [72] P. Vaitiekunas, C. Crane-Robinson, and P. L. Privalov, ‘The energetic basis of the DNA double helix: A combined microcalorimetric approach’, *Nucleic Acids Res*, vol. 43, no. 17, pp. 8577–8589, Sep. 2015, doi: 10.1093/nar/gkv812.
- [73] M. T. Record, P. L. Dehaseth, and T. M. Lohman, ‘Interpretation of Monovalent and Divalent Cation Effects on the lac Repressor-Operator Interaction<sup>^</sup>’. [Online]. Available: <https://pubs.acs.org/sharingguidelines>
- [74] ‘Record-QtrlyrevofBiophysics-1978’.
- [75] T. M. Lohman, P. L. Dehaseth, and R. M. Thomas, ‘ANALYSIS OF ION CONCENTRATION EFFECTS ON THE KINETICS OF PROTEIN-NUCLEIC ACID INTERACTIONS. APPLICATION TO LAC: REPRESSOR-OPERATOR INTERACTIONS’, North-Holland Publishing Company, 1978.
- [76] S. D. Chandradoss, A. C. Haagsma, Y. K. Lee, J. H. Hwang, J. M. Nam, and C. Joo, ‘Surface passivation for single-molecule protein studies’, *Journal of Visualized Experiments*, no. 86, Apr. 2014, doi: 10.3791/50549.

- [77] T. Cordes, J. Vogelsang, and P. Tinnefeld, ‘On the Mechanism of Trolox as Antiblinking and Antibleaching Reagent’, no. iv, pp. 5018–5019, 2009.
- [78] M. Dhar, J. A. Dickinson, and M. A. Berg, ‘Efficient, nonparametric removal of noise and recovery of probability distributions from time series using nonlinear-correlation functions: Additive noise’, *J Chem Phys*, vol. 159, no. 5, Aug. 2023, doi: 10.1063/5.0158199.
- [79] L. Yu *et al.*, ‘A Comprehensive Review of Fluorescence Correlation Spectroscopy’, *Frontiers in Physics*, vol. 9. Frontiers Media S.A., Apr. 12, 2021. doi: 10.3389/fphy.2021.644450.
- [80] R. Owczarzy, B. G. Moreira, Y. You, M. A. Behlke, and J. A. Walder, ‘Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations’, *Biochemistry*, vol. 47, no. 19, pp. 5336–5353, May 2008, doi: 10.1021/bi702363u.
- [81] Z. J. Tan and S. J. Chen, ‘Nucleic acid helix stability: Effects of salt concentration, cation valence and size, and chain length’, *Biophys J*, vol. 90, no. 4, pp. 1175–1190, 2006, doi: 10.1529/biophysj.105.070904.
- [82] C. H. Mak, ‘Unraveling base stacking driving forces in DNA’, *Journal of Physical Chemistry B*, vol. 120, no. 26, pp. 6010–6020, Jul. 2016, doi: 10.1021/acs.jpccb.6b01934.
- [83] D. Colquhoun, K. A. Dowsland, M. Beato, and A. J. R. Plested, ‘How to impose microscopic reversibility in complex reaction mechanisms’, *Biophys J*, vol. 86, no. 6, pp. 3510–3518, 2004, doi: 10.1529/biophysj.103.038679.
- [84] K. P. Burnham and D. R. Anderson, ‘Multimodel inference: Understanding AIC and BIC in model selection’, *Sociological Methods and Research*, vol. 33, no. 2. pp. 261–304, Nov. 2004. doi: 10.1177/0049124104268644.
- [85] D. Jose *et al.*, ‘Spectroscopic studies of position-specific DNA “breathing” fluctuations at replication forks and primer-template junctions’, 2009.
- [86] C. Yang, E. Kim, M. Lim, and Y. Pak, ‘Computational Probing of Watson-Crick/Hoogsteen Breathing in a DNA Duplex Containing N1-Methylated Adenine’, *J Chem Theory Comput*, vol. 15, no. 1, pp. 751–761, Jan. 2019, doi: 10.1021/acs.jctc.8b00936.
- [87] J. Douglas and P. H. Von Hippel, ‘Effects of Methylation on the Stability of Nucleic Acid Conformations STUDIES AT THE POLYMER LEVEL\*’, 1978. [Online]. Available: <http://www.jbc.org/>
- [88] J. M. Hugueta, C. V. Bizarro, N. Forns, S. B. Smith, C. Bustamante, and F. Ritort, ‘Single-molecule derivation of salt dependent base-pair free energies in DNA’, *PNAS*, vol. 107, 2010, doi: 10.1073/pnas.1001454107/-/DCSupplemental.
- [89] R. Owczarzy *et al.*, ‘Effects of Sodium Ions on DNA Duplex Oligomers: Improved Predictions of Melting Temperatures’, *Biochemistry*, vol. 43, no. 12, pp. 3537–3554, Mar. 2004, doi: 10.1021/bi034621r.
- [90] J. M. Hugueta, C. V. Bizarro, N. Forns, S. B. Smith, C. Bustamante, and F. Ritort, ‘Single-molecule derivation of salt dependent base-pair free energies in DNA’, *Proc Natl Acad Sci U S A*, vol. 107, no. 35, pp. 15431–15436, 2010, doi: 10.1073/pnas.1001454107.

- [91] B. A. Kelch, D. L. Makino, M. O'Donnell, and J. Kuriyan, 'How a DNA polymerase clamp loader opens a sliding clamp', *Science (1979)*, vol. 334, no. 6063, pp. 1675–1680, Dec. 2011, doi: 10.1126/science.1211884.
- [92] C. Gaubitz *et al.*, 'Structure of the human clamp loader bound to the sliding clamp: a further twist on AAA+ mechanism', doi: 10.1101/2020.02.18.953257.
- [93] E. Delagoutte and P. H. Von Hippel, 'Molecular mechanisms of the functional coupling of the helicase (gp41) and polymerase (gp43) of bacteriophage T4 within the DNA replication fork', *Biochemistry*, vol. 40, no. 14, pp. 4459–4477, Apr. 2001, doi: 10.1021/bi001306l.
- [94] E. Delagoutte and P. H. Von Hippel, 'Function and assembly of the bacteriophage T4 DNA replication complex: Interactions of the T4 polymerase with various model DNA constructs', *Journal of Biological Chemistry*, vol. 278, no. 28, pp. 25435–25447, Jul. 2003, doi: 10.1074/jbc.M303370200.
- [95] S. W. Nelson, R. Kumar, and S. J. Benkovic, 'RNA primer handoff in bacteriophage T4 DNA replication: The role of single-stranded DNA-binding protein and polymerase accessory proteins', *Journal of Biological Chemistry*, vol. 283, no. 33, pp. 22838–22846, Aug. 2008, doi: 10.1074/jbc.M802762200.
- [96] A. Yuzhakov, Z. Kelman, and M. O'donnell, 'Trading Places on DNA-A Three-Point Switch Underlies Primer Handoff from Primase to the Replicative DNA Polymerase onto DNA (Figure 1A). This primase displacement task', 1999.
- [97] P. Pietroni, M. C. Young, G. J. Latham, and P. H. Von Hippel, 'Structural Analyses of gp45 Sliding Clamp Interactions during Assembly of the Bacteriophage T4 DNA Polymerase Holoenzyme I. CONFORMATIONAL CHANGES WITHIN THE gp44/62-gp45-ATP COMPLEX DURING CLAMP LOADING\*', 1997. [Online]. Available: <http://www.jbc.org/>
- [98] T. Dodd, M. Botto, F. Paul, R. Fernandez-Leiro, M. H. Lamers, and I. Ivanov, 'Polymerization and editing modes of a high-fidelity DNA polymerase are linked by a well-defined path', *Nat Commun*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-19165-2.
- [99] K. Datta, N. P. Johnson, and P. H. Von Hippel, 'DNA conformational changes at the primer-template junction regulate the fidelity of replication by DNA polymerase', doi: 10.1073/pnas.1012277107/-/DCSupplemental.
- [100] T. Dodd, M. Botto, F. Paul, R. Fernandez-Leiro, M. H. Lamers, and I. Ivanov, 'Polymerization and editing modes of a high-fidelity DNA polymerase are linked by a well-defined path', *Nat Commun*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-19165-2.
- [101] K. Pant *et al.*, 'The role of the C-domain of bacteriophage T4 gene 32 protein in ssDNA binding and dsDNA helix-destabilization: Kinetic, single-molecule, and cross-linking studies', *PLoS One*, vol. 13, no. 4, Apr. 2018, doi: 10.1371/journal.pone.0194357.
- [102] G. J. Latham, D. J. Bacheller, P. Pietroni, and P. H. Von Hippel, 'Structural Analyses of gp45 Sliding Clamp Interactions during Assembly of the Bacteriophage T4 DNA Polymerase Holoenzyme II. THE gp44/62 CLAMP LOADER INTERACTS WITH A SINGLE

- DEFINED FACE OF THE SLIDING CLAMP RING\*', 1997. [Online]. Available: <http://www.jbc.org/>
- [103] G. J. Latham, D. J. Bacheller, P. Pietroni, and P. H. Von Hippel, 'Structural Analyses of gp45 Sliding Clamp Interactions during Assembly of the Bacteriophage T4 DNA Polymerase Holoenzyme III. THE gp43 DNA POLYMERASE BINDS TO THE SAME FACE OF THE SLIDING CLAMP AS THE CLAMP LOADER\*', 1997. [Online]. Available: <http://www.jbc.org/>
- [104] Y. Shamoo and T. A. Steitz, 'Building a Replisome from Interacting Pieces: Sliding Clamp Complexed to a Peptide from DNA Polymerase and a Polymerase Editing Complex RB69 DNA polymerase shows structural similarities to the pol I family polymerases in the polymerase catalytic domain', 1999.
- [105] W. E. Moerner and D. P. Fromm, 'Methods of single-molecule fluorescence spectroscopy and microscopy', *Review of Scientific Instruments*, vol. 74, no. 8, pp. 3597–3619, 2003, doi: 10.1063/1.1589587.
- [106] W. Lee, D. Jose, C. Phelps, A. H. Marcus, and P. H. Von Hippel, 'A Single-Molecule View of the Assembly Pathway, Subunit Stoichiometry, and Unwinding Activity of the Bacteriophage T4 Primosome (helicase – primase) Complex', 2013, doi: 10.1021/bi400231s.
- [107] M. Liebel, C. Toninelli, and N. F. Van Hulst, 'Nonlinear spectroscopy of a single molecule', *Nat Photonics*, vol. 12, no. 1, pp. 1–6, 2018, doi: 10.1038/s41566-017-0056-5.
- [108] R. Moya, A. C. Norris, T. Kondo, and G. S. Schlau-Cohen, 'Observation of robust energy transfer in the photosynthetic protein allophycocyanin using single-molecule pump–probe spectroscopy', *Nat Chem*, vol. 14, no. 2, pp. 153–159, 2022, doi: 10.1038/s41557-021-00841-9.
- [109] V. Tiwari, Y. A. Matutes, A. T. Gardiner, T. L. C. Jansen, R. J. Cogdell, and J. P. Ogilvie, 'Spatially-resolved fluorescence-detected two-dimensional electronic spectroscopy probes varying excitonic structure in photosynthetic bacteria', no. 2018, doi: 10.1038/s41467-018-06619-x.
- [110] A. Perdomo-Ortiz, J. R. Widom, G. A. Lott, A. Aspuru-Guzik, and A. H. Marcus, 'Conformation and electronic population transfer in membrane-supported self-assembled porphyrin dimers by 2D fluorescence spectroscopy', *Journal of Physical Chemistry B*, vol. 116, no. 35, pp. 10757–10770, 2012, doi: 10.1021/jp305916x.
- [111] P. F. Tekavec, T. R. Dyke, and A. H. Marcus, 'Wave packet interferometry and quantum state reconstruction by acousto-optic phase modulation', *Journal of Chemical Physics*, vol. 125, no. 19, 2006, doi: 10.1063/1.2386159.
- [112] J. C. Phys, S. Roeding, N. Klimovich, and T. Brixner, 'Optimizing sparse sampling for 2D electronic spectroscopy', vol. 084201, no. December 2016, 2017, doi: 10.1063/1.4976309.
- [113] T. Brixner, G. Krampert, P. Niklaus, and G. Gerber, 'Generation and characterization of polarization-shaped femtosecond laser pulses', *Appl Phys B*, vol. 74, no. SUPPL., pp. 133–144, 2002, doi: 10.1007/s00340-002-0911-y.

- [114] A. J. Kiessling and J. A. Cina, ‘Monitoring the evolution of intersite and interexciton coherence in electronic excitation transfer via wave-packet interferometry’, *Journal of Chemical Physics*, vol. 152, no. 24, 2020, doi: 10.1063/5.0008766.
- [115] S. M. Hart *et al.*, ‘Activating charge-transfer state formation in strongly-coupled dimers using DNA scaffolds’, *Chem Sci*, pp. 13020–13031, 2022, doi: 10.1039/d2sc02759c.
- [116] S. M. Hart, J. L. Banal, M. Bathe, and G. S. Schlau-Cohen, ‘Identification of Nonradiative Decay Pathways in Cy3’, *J Phys Chem Lett*, pp. 5000–5007, 2020, doi: 10.1021/acs.jpcclett.0c01201.
- [117] A. K. Pati *et al.*, ‘Tuning the Baird aromatic triplet-state energy of cyclooctatetraene to maximize the self-healing mechanism in organic fluorophores’, *Proc Natl Acad Sci U S A*, vol. 117, no. 39, pp. 24305–24315, 2020, doi: 10.1073/pnas.2006517117.
- [118] A. G. Numerik and / Optimierung, ‘Introduction to Compressed Sensing’.
- [119] J. N. Sanders *et al.*, ‘Compressed sensing for multidimensional spectroscopy experiments’, *Journal of Physical Chemistry Letters*, vol. 3, no. 18, pp. 2697–2702, 2012, doi: 10.1021/jz300988p.
- [120] Maurer, J., Albrecht, C.S. , Heussman, D., Herbert, P., von Hippel, P.H. & Marcus, A.H. (2023). Polarization-sweep single-molecule fluorescence microscopy of exciton-coupled (Cy3)<sub>2</sub> dimer-labeled DNA fork constructs. *Submitted to JCPB*.