Dynamics of Stop-codon Recognition by Release Factor 1

Colin Donald Kinz-Thompson

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Dynamics of Stop-codon Recognition by Release Factor 1

Colin Donald Kinz-Thompson

Translation of an mRNA template into its corresponding protein is necessarily a highly accurate process. In all organisms, this translation is performed by the universally conserved macromolecular machine, the ribosome. However, the mechanisms through which the ribosome is able to regulate translation, and therefore ensure its fidelity, are not well understood. Often these types of mechanisms, which ensure molecular fidelity, utilize multiple, transient states over which cognate and non-cognate substrates are discriminated multiple times. However, such transient and/or rarely populated states are difficult to study by conventional, ensemble experimental techniques. In this thesis, single-molecule fluorescence resonance energy transfer (smFRET), which alleviates many of these limitations, is used in order to interrogate the dynamics of a translation factor, release factor 1 (RF1), and how they are organized to ensure accurate and efficient recognition of stop-codons during the termination stage of translation.

In order to observe the dynamics of the RF1 binding and codon discrimination processes with smFRET, a relatively high concentration of fluorophore-labeled RF1 must be used in order to observe significant binding to sense-codons; however, such high concentrations are not accessible with traditional smFRET total internal fluorescence microscopy. Therefore, in Chapter 2 a novel approach to breaking this concentration barrier is presented, in which robustly-passivated gold-based nanoaperture arrays are developed to limit the excitation volume used in smFRET measurements of RF1. Unfortunately, as in the case of RF1 binding to sense-codon programmed ribosomes, many of the ribosomal dynamics that are in principal observable using smFRET are too fast to observe using current wide-field detectors. Therefore, Chapter 3 investigates the precision and accuracy with which transient conformational dynamics can be quantified using single-molecule techniques such as smFRET. As a case study, these approaches were used to analyze the dynamics of the GS1-GS2 equilibrium of the pretranslocation (PRE) ribosome–a situation where transient intermediate states that can be observed using single-particle cryo-electron microscopy are not seen using smFRET.

In Chapter 4, a novel computational method is developed to address such temporally-limited single-molecule data, and in doing so, it is used to analyze the structural contributions of tRNA to ribosomal transition state energy barriers using temperature-dependent smFRET with temporal super-resolution. The temperature-dependence of reaction rate constants is governed by the underlying thermodynamic landscape of the molecular system. To investigate the energy landscape over which the PRE ribosome operates,

temperature-dependent smFRET experiments were performed on PRE complexes containing different tRNAs. By investigating the relative temperature-dependence of the rate constants involved in the GS1 - GS2 equilibrium as a function of tRNA identity, nascent polypeptide chain presence, and A and P site occupation, relative thermodynamic contributions of the different structural elements were quantified. Unfortunately, this investigation was complicated by fast rate constants which approach the time resolution limitations of smFRET TIRF experiments, especially with the increased temperatures used in these experiments. Additionally, it is complicated by the heterogeneity within the ensemble of ribosomes that is created when some of the enzymatically-prepared ribosomal complexes fail to undergo, or undergo additional rounds of translation. To overcome these complications, a novel computational method to achieve temporal super-resolution. This method uses Bayesian inference for the analysis of sub-temporal resolution data (BIASD). By integrating this approach with a Bayesian variational mixture model, the fast dynamics of heterogenous populations can be accurately and precisely quantified. This then allowed the contributions of the structural differences that the various tRNA make to the underlying PRE complex energy landscape to be determined.

The conformational dynamics that regulate the binding affinity and codon discrimination ability of RF1 are investigated in Chapter 5. During the elongation stage of translation, class I release factors compete with aminoacyl-tRNAs to interrogate the mRNA triplet-nucleotide codon that is located in the ribosomal aminoacyl-tRNA binding (A) site. To avoid deleterious effects, class I RFs must be able to accurately discriminate stop-codons from sense-codons, only triggering the termination stage of translation and catalyzing the release of the nascent polypeptide chain from the peptidyl-tRNA located in the ribosomal peptidyl-tRNA binding (P) site upon recognition of a stop-codon. Despite its importance for ensuring the accuracy of gene expression, the high fidelity mechanism through which class I RFs discriminate sense codons remains elusive. Using smFRET, the kinetics with which a fluorophore-labeled, bacterial RF1 binds to the A site of bacterial ribosomal release complexes carrying a fluorophore-labeled peptidyl-tRNA in the P site and either a stop-codon, or a sense-codon that differs from a stop-codon by a single nucleotide (*i.e.*, a near-stop codon) programmed in the A-site are investigated. The results of these experiments, as well as analogous experiments performed using RF1 mutants or antibiotic inhibitors of RF1 function, reveal that RF1 binding affinity and codon discrimination occurs *via* a multistep process. Taken together with molecular dynamics simulations of wildtype and mutant RF1, these data demonstrate how the conformation dynamics of the switch loop modulate RF1 binding affinity and codon discrimination–enabling the elucidation of some of the molecular details through which class I RFs ensure the integrity of translation elongation and the fidelity of translation termination.

Table of Contents

List of Figures

List of Tables

Acknowledgments

There is a quite long list of people that I must thank, for without their support, belief, encouragement, and guidance throughout years, I would never have reached this point here. Foremost, I should thank my mother, Ms. Judy Kinz, brother, Mr. Patrick Kinz-Thompson, and grandmother, Ms. Joy Kinz, for the support only a family provides. Similarly, I also need to thank the rest of my family, extended and adopted. I'd like to thank Ms. Jackie Kim for being a wonderful partner through these past years.

As for early influences, I need to acknowledge a Ms. Mee and Mr. Dupra for encouraging a budding scientist. Similarly, I need to sincerely thank Prof. Thomas Krugh for the non-compulsory mentoring of an undergraduate chemistry student. Additionally, I would like to thank Prof. Dave McCammant, and Prof. Lewis Rothberg for their scientific encouragement and mentoring. Words cannot express my gratitude to Prof. Esther Conwell, whose guidance and instruction shaped me into a scientist. Finally, I somehow have to express that same gratitude to Prof. Ruben Gonzalez for the unending guidance and encouragement that polished, refined, and launched me into so many new directions.

There are many friends and lab mates that I must thank for making the past few years bearable; in particular, though not limited to, Mr. Nathan Daly and Dr. Mark Hendricks for being so understanding, Dr. Justin Ambramson and Dr. Dileep Pulukkunat for teaching me *everything* in lab, Dr. Kelvin Caban for taking over when Dileep left, Dr. Lindsay Leone and Dr. Glen Hocky for being outside experts, and especially Ms. Bridget Huang for putting up with a lot.

I have also had the benefit of many great collaborations over the years here, and must thank the many collaborators for making them so enjoyable and fruitful. Those who have not yet been mentioned: Nanoapertures – Dr. Alexander Gondarenko, Prof. Matteo Palma, Mr. Daniel Chenet, Prof. James Hone, and Prof Shalom Wind. Dwell-times – Dr. Ajeet Sharma, Prof. Joachim Frank, Prof. Debashish Chowdhury. Rate Constants – Ms. Nevette Bailey. PRE Complexes – Prof. Jingyi Fei. Quenchers – Dr. Andrew Anzalone. Nanodiamonds – Prof. Abe Wolcott. SPAD Arrays – Mr. Nick Bertone, Dr. Simone Tisa, Dr. Andrea Giudice. Btu-CDF – Dr. Jinrang Kim, Mr. Kun Leng, Mr. Lingwei Zhu, Prof. John Hunt.

From my time spent at Columbia University during my graduate work, I have to thank the entire department of chemistry – especially staff, coordinators, professors, historic Havemeyer Hall, other members of the Physical Chemistry Student Seminar Series, and many department events. Also, for guiding and presiding over the apotheosis of my time spent here, I must thank my dissertation committee, alphabetically: Professors Ruben Gonzalez, Eric Greene, John Hunt, Ann Mcdermott, and Wei Min. Finally, I must thank Columbia University's National Institutes of Health Training Program in Molecular Biophysics (T32-GM008281), and the Department of Energy Office of Science Graduate Fellowship Program (DE-AC05-06OR23100) for their support.

Chapter 1

Introduction*

1.1 Overview of Protein Synthesis

Proteins are the ubiquitous workhorses of the cell [1, 2]. They come in diverse classes, such as enzymes, structural components, or molecular motors, and they function in even more diverse roles, such as in translation, transcription, replication, repair, recombination, cell division, membrane biogenesis, secretion, ion transport, signal transduction, energy production and conversion, and metabolism and transport [3]. In living organisms, the information required to create each protein is stored as a gene in the form of DNA. In order to put this DNA-based information to use and make new proteins, nature has converged upon a common, two step approach to the organization of life [4]. First, the DNA, which was created by templating the replication of an exact copy from the DNA of a parental cell, is used to template the transcription of an intermediate form of the gene called messenger RNA (mRNA) [5]. After being transcribed, the mRNA is then used to template the translation of the final protein product of the gene [6]. Interestingly, these two reactions, mRNA transcription and protein translation, which together achieve the chemical realization of a DNA gene product, are both successive linear-polymerization reactions of a relatively small number of different monomers–four ribonucleotides, or 20 common amino acids, respectively [7].

In order to create their respective high-molecular weight polymers, both of these reactions must catalyze the thermodynamically unfavorable monomer-polymer bond formation reactions as each monomer is added during the elongation stages of the polymerization reactions [7]. However, unlike many chemi-

* In part, adapted with permission from Kinz-Thompson, C.D., Sharma, A.K., Frank, J., Gonzalez, Jr., R.L., Chowdhury, D Quantitative Connection Between Ensemble Thermodynamics and Single-Molecule Kinetics: A Case Study Using Cryo-EM and smFRET Investigations of the Ribosome *Journal of Physical Chemistry B*, **2015**, 119(34), 10888-10901. Copyright 2015 American Chemical Society.
Additionally, in part reprinted from FEBS Letters, 588(19), Kinz-Thompson, C.D., Gonzalez, Jr., R.L., smFRET studies of the 'encounter' complexes and subsequent intermediate states that regulate the selectivity of ligand binding, 3526-3538, 2014, with permission from Elsevier.

cal polymerization reactions, there is a stringent requirement that the specific monomer added is dictated by the underlying sequence of the template, since errors at any step can compromise the structure and function of the emergent protein [8–11]. This process of templating or 'biological copying' is central to both transcription and translation, but because they occur on a nanoscopic scale where thermal fluctuations can overwhelm the small energetic differences between the correct, template-defined monomer and the other incorrect monomers, errors cannot be avoided [1, 12–14]. Fortunately, large macromolecular complexes have evolved that ensure the accurate and efficient copying of the templates [1, 9, 15–17].

In particular, in all organisms, mRNAs are translated into proteins by the ribosome [18], a 2.5 to 4.3 MDa [19, 20] molecular machine [21, 22] that is composed of a large and a small ribonucleoprotein subunit (50S and 30S in bacteria, respectively) [23]. With the help of many translation factors, the ribosome assembles on an mRNA, and recruits the aminoacyl-tRNA (aa-tRNA) that matches each successive codon of the mRNA [24] in order to deliver each successive amino acid to the elongating, nascent polypeptide chain that will become a protein [25, 26]. Thus, the tRNAs function as 'adaptor' molecules [6, 7, 27] that can read the codon of the mRNA, because they are also made of ribonucleotides and can therefore exploit Watson-Crick base-pairing to form a codon-anticodon interaction [24], and that can deliver the proper amino acid that corresponds to that particular codon, because aminoacyl-synthetases can specifically aminoacylate their corresponding tRNA with a high degree of accuracy [1, 28, 29]. Therefore, instead of having to discriminate between the correct or incorrect amino acids during this templating process, the ribosome is evolved to select the correct tRNA (reviewed in Ref. 9). However, the energetic differences between anticodons of tRNA that, for example, might differ by a single ribonucleotide substitution of an adenine for a guanine, which are both purines, are still small energetic differences, and therefore mistakes will still occur by selecting the incorrect tRNA to translate the mRNA [15, 16]. Similarly, when protein synthesis is complete, class I release factors recognize specific stop-codons and catalyze termination of the translation, but these proteins must recognize only stop-codons, as stopping the protein synthesis early yields an abortive product with deleterious effects *in vivo* [30, 31]. While these types of error are thermodynamically unavoidable, *in vivo* the error rate for aa-tRNA selection at non-sense codons is 10^{-3} to 10^{-4} [9], and the error rate for translation termination at non-stop codons is less than 10^{-5} [32]. In order to achieve such high fidelity in translation, the ribosome and its associated translation factors play an active role to discriminate between cognate and non-cognate substrates.

1.1.1 Accuracy in Templating Processes

The ability of a biomolecule to selectively bind its cognate ligand when faced with the vast array of structurally similar near- and non-cognate ligands present in the cellular environment is critical for maintaining the fidelity of templating and other biological processes. Specific examples include critical processes spanning all of biology: a transcription factor binding a DNA promoter site, a small nuclear ribonucleoprotein binding a precursor mRNA splice site, a microRNA binding a target mRNA, or a ribosome binding an aa-tRNA substrate. On the thermodynamic level, the maximum amount of discrimination between cognate and non-cognate ligands in a templating process is determined by the free energy difference (thermodynamic-control), as well as by the energy barrier difference (kinetic-control) between the incorporation of cognate and non-cognate ligands [1, 14]. However, on the molecular level, biomolecules must be organized in such a way as to exploit these thermodynamic differences in order to efficiently achieve the highest amount of discrimination [12]; indeed, the maximal amount of discrimination between a cognate and non-cognate ligand occurs when the templating process is performed adiabatically, and under thermodynamic-control, but such a process would have to occur with zero incorporation rate [14, 33]. How can biomolecules efficiently interact with their cognate targets? How can biomolecules attain accuracies higher than dictated by the thermodynamic differences between targets? In this section, I will review theories that allow biomolecules to attain such specificity and accuracy.

Molecular recognition during binding events is often thought to occur through two distinct, but complementary mechanism (reviewed in Ref. [34]). In Koshland's 'induced fit' model, binding of a ligand to a biomolecule drives conformational changes within the complex [35]. With the binding of cognate ligands, these conformational changes correctly position the complex for further function (*e.g.*, catalysis or tight-binding), while, with the binding of non-cognate ligands, these conformational changes incorrectly position the complex for further function [35]. Notably, induced fit suggest a mechanism through which biomolecular function can be allosterically regulated by the conformational changes induced by ligand binding, even if the ligand-binding allosteric site is distal from the function-related active site [36, 37]. However, biomolecules are inherently dynamic and exist in many different, functionally-active conformations [38, 39]; in fact, both conformations involved in an allosteric transition can be accessed in the absence of the allosteric effector ligand, and while they probably have different population distributions[34, 40], in some cases both conformations can be substantially populated such as with enzymes like dihydrofolate reductase [41] or triosephosphate isomerase [42], or even with large ribonucleoprotein complexes such as the ribosome [43]. This suggests a

second mode through which molecular recognition can occur. If an ensemble of different biomolecular conformations dynamically interconvert in solution, then a conformation which is more competent for cognate ligand binding than it is for non-cognate ligand binding can be selected by the binding of cognate ligand [44, 45]. Thus, molecular recognition can also occur through 'conformational selection' in which ligand binding results in conformational ensemble population shifts towards the bound-conformation state. It is worth noting that both the induced fit and conformational selection models are found to be utilized for molecular recognition [34], however both are constrained to the thermodynamically-determined cognate versus non-cognate specificity limit.

High specificity of ligand binding is typically achieved through multistep, ligand-binding reactions in which the first step encompasses the reversible formation of a nearly non-specific, weakly interacting, transient, intermediate 'encounter' complex between the biomolecule and a potential ligand [46]. Subsequent steps lead to the formation of a final biomolecule-ligand complex [46, 47]. Multistep ligand-binding reactions such as these are governed by free-energy landscapes such as the one shown in Figure 1.1, which depicts a minimal, two-step, ligand-binding reaction. Regardless of the number of steps, formation of an encounter complex reduces the search that is required to form the final biomolecule-ligand complex from three spatial dimensions to two [46]. This results in a lowering of the considerably large entropic barrier(s) that would otherwise separate the unbound state of the biomolecule from the bound state and potentially allows the search for the bound state to be guided along a primarily enthalpic energy funnel [46–49]. As a consequence, formation of an encounter complex significantly increases the rate with which biomolecules can conformationally screen potential ligands. In addition, the weakly interacting, transient, and reversible nature of an encounter complex allows it to rapidly dissociate into its component biomolecule and potential ligand. The reversibility and selectivity of ligand-binding reactions can therefore be precisely regulated by coupling the identity of the ligand to the probability of dissociating versus the probability of proceeding along the reaction pathway. Near- and non-cognate ligands will have a higher probability of dissociating while cognate ligands will have a higher probability of proceeding along the reaction pathway and forming the final biomolecule-ligand complex. Here, more complex ligand-binding reactions encompassing three or more steps allow additional opportunities for the ligand to dissociate from the post-transition state ($\ddagger_{EC-B}$ in Figure 1.1) intermediate states [47]. If one of more of these steps can be coupled to the dissipation of some amount of energy, for instance from small changes in configurational entropy [50] or the hydrolysis of a nucleoside triphosphate, then the accuracy of selecting cognate ligands can be increased beyond the limit imposed by the thermodynamic differences between cognate and non-cognate ligands [14, 51]. These types of multistep

mechanisms effectively 'proofread' the molecular interaction, and utilize the thermodynamic differences between cognate and non-cognate ligands multiple times in order to increase the overall accuracy of the ligand binding process. For instance, in the 'kinetic proofreading' mechanism that operates in aa-tRNA selection during translation elongation [52], an initial selection step favors cognate ligand over non-cognate ligand according to the thermodynamically-determined accuracy, but then GTP hydrolysis occurs, which creates a non-equilibrium population difference between already-bound cognate and near-cognate ligand, and thus creates the opportunity to perform an additional selection step from the previously selected ligands [15, 16]. Because these proofreading mechanisms utilize some form of energy dissipation, they are able to attain increased specificity of ligand binding, and also do not significantly decrease the rate of the ligand-binding reaction.

Although they play a decisive role in ensuring that biomolecules can rapidly and selectively bind their cognate ligands, the encounter complexes and, if present, subsequent intermediate states that form during ligand-binding reactions have traditionally been very difficult to experimentally observe and characterize [46]. This is because the stochastic nature of ligand-binding kinetics and the transient nature of the encounter complexes and subsequent intermediate states that are formed give rise to a situation in which such complexes and states comprise exceedingly low-population states that are extremely difficult to detect using traditional, ensemble biophysical techniques that report only on the average behavior of the entire ensemble (see Section 1.5.1). Further exacerbating this situation is the fact that, due to their partially non-specific nature, encounter complexes and, possibly, subsequent early intermediate states, can exhibit significant heterogeneity in their structures and other physical properties. Despite these challenges, the emergence in the early 2000s of nuclear magnetic resonance (NMR) spectroscopic techniques, particularly paramagnetic relaxation enhancement, that enable the detection of low-population states have allowed the partial characterization of several encounter complexes and/or subsequent intermediate states formed during ligand-binding reactions [53–55]. Almost simultaneously, the advent of single-molecule biophysical techniques, particularly single-molecule fluorescence approaches, has provided a powerful complement to NMR spectroscopic methods in detecting such low-population states (see Section 1.5). For my thesis work, I have utilized just such a single-molecule approach to investigate the origins and regulation of ligand-binding at the end of protein synthesis – translation termination.

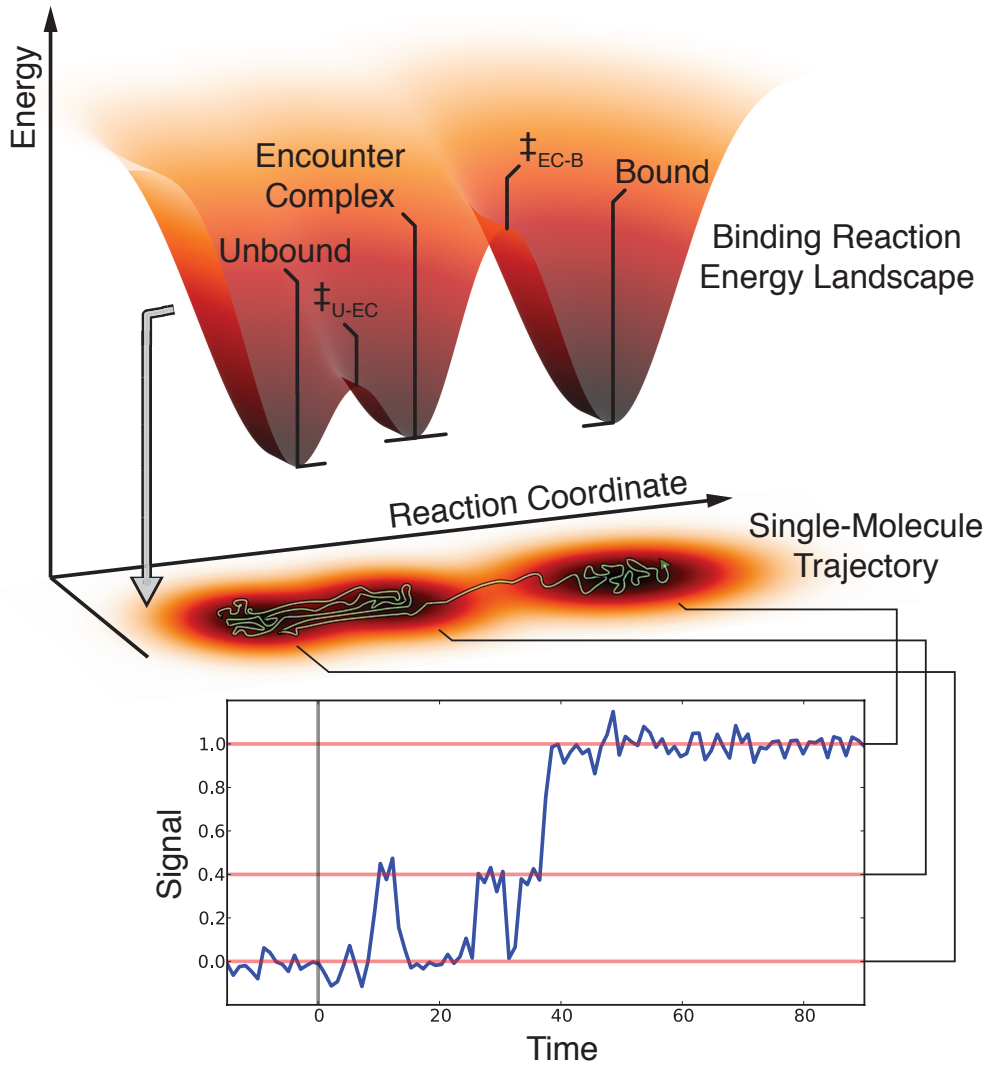

Figure 1.1: Schematic of a ligand-binding reaction. The minimal, three-state, energy landscape describes the energy of a ligand-binding reaction along the reaction coordinate. The free biomolecule and ligand begin in the unbound state, and may collide and cross an initial energy barrier ($\ddagger_{U-EC}$) to form an encounter complex. The encounter complex may then either dissociate to reestablish the free biomolecule and ligand or cross a second energy barrier ($\ddagger_{EC-B}$) to form the final bound state. The energy landscape is projected onto a plane below it, where the hypothetical trajectory of a single ligand-binding reaction is depicted diffusing along the reaction coordinate (gray curved line). Below this trajectory, a hypothetical signal versus time trajectory corresponding to the ligand-binding reaction trajectory is shown, where the signal of the unbound state, encounter complex, and bound state are denoted (red horizontal lines). The reaction is initiated at t = 0, crosses $\ddagger_{U-EC}$ several times to transiently sample the encounter complex, and ultimately crosses $\ddagger_{EC-B}$ to form the bound state.

1.2 Bacterial Ribosomal Architecture

The bacterial ribosome is a 2.3 MDa ribonucleoprotein complex that sediments as a 70S particle composed of a large subunit (50S), and a small subunit (30S) (Fig. 1.2A) [3, 23, 56, 57]. The 50S subunit is composed of 34 ribosomal proteins (rProteins), 19 of which are conserved among all domains of life [58], as well as a highly conserved 23S ribosomal RNA (rRNA) ($\sim$2,900 nt) and a 5S rRNA ($\sim$120 nt) [57]. The 30S subunit is composed of 23 rProteins, 15 of which are conserved among all domains of life [58], and a highly conserved 16S rRNA ($\sim$1,540 nt) [57]. Among different bacteria, X-ray crystallography structures show that 70S bacterial ribosomes are structurally very similar [59, 60]. Interestingly, along with the universally conserved rProteins, over 95% of the rRNA comprises a universally conserved 'common core' that is present and expanded upon in the ribosomes of all lower and higher eukaryotes. [57, 61–63].

Thus, the bacterial 70S ribosomal architecture is organized in a manner that is ubiquitous for all steps of protein synthesis: initiation, elongation, termination, and recycling (Fig. 1.2B). These steps center upon two reaction centers, one called the decoding center (DC), where aa-tRNA anticodons base pair to mRNA codons to determine which amino acid should be added to the nascent polypeptide chain [9, 64], and one called the peptidyl transferase center (PTC), where peptide bond formation occurs and an amino acid is added to the nascent polypeptide chain [65, 66] (Fig. 1.2). Interestingly, highly conserved rRNA in the DC, such as G529, G530, A1492, and A1493, as well as in the PTC, such as U2506, G2583, U2584, and U2585, play active roles in these reactions [9, 64–66]. Additionally, after nucleolytic processing of rRNA operon transcripts during ribosome biogenesis, at least 11 bases of the 16S rRNA, and 25 bases of the 50S are then post-transcriptionally modified – mostly in the regions surrounding the DC and the PTC respectively [56]. The mRNA binds to the 30S subunit, in part in the DC, within a cleft along the bottom of the three tRNA binding sites: the aminoacyl binding (A) site containing the incoming aa-tRNAs that bind the ribosome, the peptidyl-tRNA binding (P) site containing the tRNA holding the nascent polypeptide chain, and the deacylated tRNA exit (E) site containing tRNA that are about to dissociate from the ribosome [67]. Many of the structural dynamics required for translation occur in these sites [68].

Translation begins with initiation (Fig. 1.2), the rate-limiting step of protein synthesis *in vivo* (reviewed in Ref. 69). During initiation, the ribosome must locate an open reading frame within the mRNA, which includes making base-pairing interactions between the Shine-Dalgarno sequence of the mRNA and the anti-Shine-Dalgarno sequence in the 3' end of the 16S rRNA [70], and bind a formylated initiator tRNA to the 30S subunit at an inframe start-codon [11, 71]. This process, and the subsequent binding of the 50S subunit

to form a 70S elongation complex, is facilitated and regulated by initiation factors IF1, IF2, and IF3 [11, 69, 71]. Following initiation, successive rounds of translation elongation occur, in which the nascent polypeptide chain is synthesized.

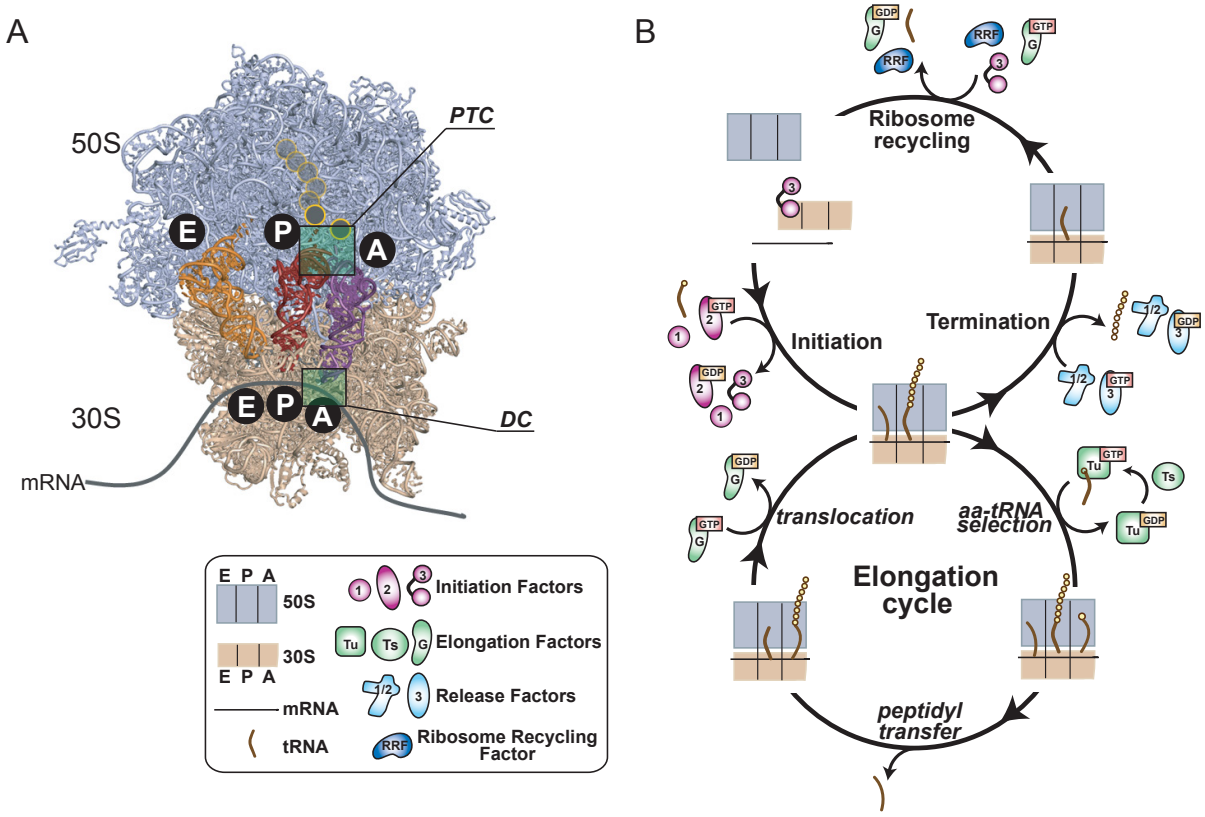

Figure 1.2: Prokaryotic Ribosomal Architecture and a Schematic of Translation. (A) The structure of the prokaryotic ribosome (*Thermus Thermophilus* (*T. Thermophilus*)) as determined by X-ray crystallography (PDB ID: 2J00 and 2J01). The 50S (lavender) and 30S (tan) subunits bind mRNA (gray) and provide the binding sites for the aminoacyl-tRNA (A-site, purple), the peptidyl-tRNA (P-site, red), and decacylated tRNA (E-site, orange). Codon recognition during tRNA selection and translation termination occurs in the decoding center (DC), and peptide bond formation and hydrolysis occurs in the peptidyl-transferase center (PTC). (B) Schematic cartoon of the translation cycle in which initiation is followed by successive rounds of elongation, then termination, and finally ribosome recycling. Figure reproduced from Ref. 72.

1.3 Prokaryotic Translation Elongation

During the elongation stage of translation, the ribosome undergoes consecutive rounds of an elongation cycle (Fig. 1.2B) in which it successively adds amino acids to the nascent polypeptide chain in the order

dictated by the sequence of the mRNA. In the first step of the elongation cycle, the correct mRNA-encoded aa-tRNA is delivered to the A-site of the ribosome in the form of a ternary complex (TC) that is composed of the ribosomal guanosine triphosphatase (GTPase) elongation factor (EF) Tu, guanosine triphosphate (GTP), and aa-tRNA [8, 11, 21, 64, 67]. Upon delivery of the mRNA-encoded aa-tRNA into the A-site, peptide bond formation results in transfer of the nascent polypeptide chain from the peptidyl-tRNA bound at the ribosomal P-site to the aa-tRNA at the A-site, generating a ribosomal pretranslocation (PRE) complex carrying a newly deacylated tRNA at the P site and a newly formed peptidyl-tRNA, extended by one amino acid, at the A-site [65, 73, 74]. Subsequently, the ribosome must translocate along the mRNA, moving the newly deacylated tRNA from the P-site to the ribosomal E-site, and the newly formed peptidyl-tRNA from the A-site to the P-site [21, 64, 75–81]. While translocation can occur spontaneously, albeit slowly, *in vitro* [82], it is accelerated by orders of magnitude *in vivo* through the action of EF-G, another ribosomal GTPase [64, 81].

1.3.1 Aminoacyl-tRNA Selection

Despite the pool of 41-55 different aa-tRNA isoacceptors that is found in cells [83], the ribosome is able to accurately select the aa-tRNA whose anticodon correctly basepairs to the mRNA codon at the A site (*i.e.*, the cognate aa-tRNA), only misincorporating aa-tRNAs with one-base mismatches (*i.e.*, near-cognate aa-tRNAs), or with two-base or greater mismatches (*i.e.*, non-cognate aa-tRNAs) with a frequency of 1 in 10^3-10^4 [84–87]. Notably, this level of fidelity greatly exceeds the maximum misincorporation frequency of 1 in 10^2 that would be expected from just the thermodynamic stability differences between the anticodon-codon interactions formed by a cognate aa-tRNA and those formed by the corresponding near-cognate aa-tRNAs [88–90]. Extensive studies of the mechanisms through which the ribosome achieves the observed high fidelity of aa-tRNA selection have led to the development of a widely accepted, multistep mechanism (Fig. 1.3) [91, 92] in which kinetic proofreading [15–17, 52, 93, 94] and induced fit [35, 94] strategies are employed to increase the otherwise relatively low fidelity of aa-tRNA selection that would be expected [9, 11].

Over the past decade, the mechanism of aa-tRNA selection has been the subject of numerous single-molecule fluorescence resonance energy transfer (smFRET) studies (see Section 1.5) [43, 95–99]. smFRET studies of aa-tRNA selection are typically performed using ribosomal elongation complexes (RECs) that are biotinylated at the 5' end of the mRNA and that have been labeled with a donor fluorophore either within the fMet-tRNA^{fMet} that is bound at the ribosomal P-site [95, 100], or within ribosomal protein L11 [96, 101]. RECs are then tethered to the surface of a microfluidic, observation flowcell via their 5'-biotinylated mRNA such

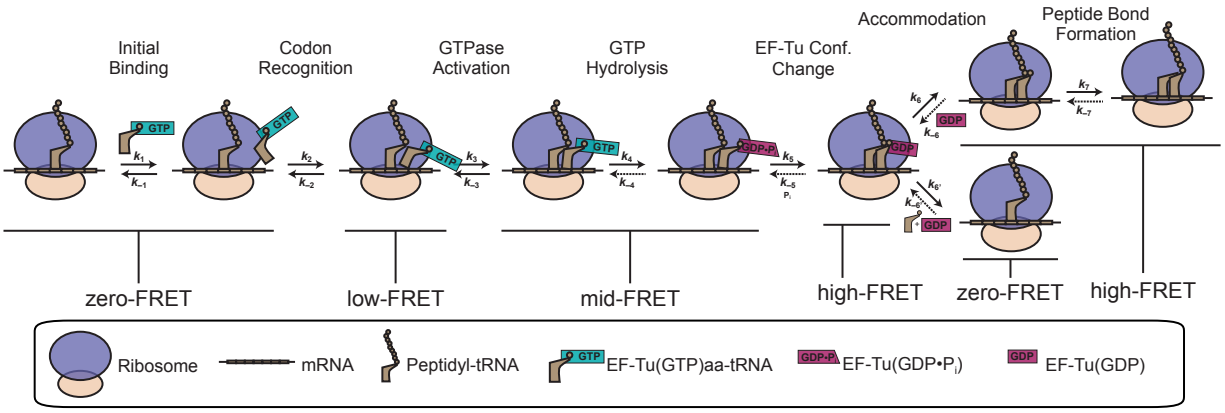

Figure 1.3: Cartoon schematic of the mechanism of aa-tRNA selection during translation elongation. The E_{FRET} values for the states that are observed by smFRET are denoted below the corresponding states. The rate constants shown are likely composite rate constants that describe several events that occur during a step. Dashed arrows represent steps believed to have exceedingly low probabilities of occurring. smFRET experiments have utilized both tRNA-tRNA and tRNA-ribosome donor-acceptor labeling schemes. This mechanistic scheme is based on a similar scheme appearing in Ref. 21.

that they can be imaged with single-molecule resolution using total internal fluorescence (TIRF) microscopy (Sec. 1.5). Stopped-flow delivery of a TC carrying an acceptor-labeled aa-tRNA to an REC carrying a donor-labeled P-site tRNA and a cognate A-site codon then yields acceptor- and donor intensities versus time trajectories that are used to calculate pre-steady-state E_{FRET} versus time trajectories. E_{FRET} versus time trajectories initiate at zero-E_{FRET} and evolve through transiently sampled low- and mid-E_{FRET} states before arriving at a high-E_{FRET} final state that is consistent with structures approximating the final state of the reaction in which the aa-tRNA has been accommodated into the A site (Fig. 1.3) [43, 95–99]. The mid-E_{FRET} state has been assigned to a mixture of at least two intermediate states that had been previously observed in biochemical [94] and structural studies [102, 103], and that correspond to the conformations of the TC-bound REC that immediately precede and immediately follow ribosome-catalyzed GTP hydrolysis by EF-Tu (Fig. 1.3) [95]. The state that precedes GTP hydrolysis can be biochemically 'captured' and stabilized using a non-hydrolyzable GTP analog [104] or a GTP hydrolysis-deficient EF-Tu mutant [105]. Likewise, the state that immediately follows GTP hydrolysis can be captured and stabilized using the EF-Tu-targeting antibiotic kirromycin [106]. Such approaches allow the populations of these ordinarily transient and low-population states to be increased such that they can be easily studied using ensemble biochemical and structural methods [8, 10, 11, 21, 64].

In contrast to the mid-E_{FRET} state, the low-E_{FRET} state has been assigned to a structurally novel intermediate state that has thus far eluded capture and stabilization using mutations, biochemical analogs, or small-molecule inhibitors, thereby precluding its direct detection using ensemble biochemical or structural

studies (Fig. 1.3). Nonetheless, the conformation of the TC-bound REC corresponding to the low-E_{FRET} state is a critical, codon-dependent intermediate state during aa-tRNA selection. Experiments in which the TC is delivered to an REC carrying a non-cognate A-site codon, for example, do not result in any detectable sm-FRET signals, including even the detection of highly transient sampling of the low-E_{FRET} state. Analogous experiments using RECs carrying a near-cognate A-site codon, on the other hand, result in the detection of a highly transient low-E_{FRET} state that corresponds to the formation of a weakly interacting, transient TC-bound REC from which TC has a much higher probability of dissociating from the REC than of progressing along the reaction pathway. In contrast, experiments using RECs carrying a cognate A-site codon result the detection of a slightly longer-lived low-E_{FRET} state that corresponds to the formation of a slightly more stably interacting, less transient TC-bound REC from which the TC has a much higher probability of progressing along the reaction pathway than of dissociating from the REC. Interestingly, the precise E_{FRET} values of the low-E_{FRET} state differ for cognate and near-cognate TC-bound RECs, indicating that the TC in a cognate TC-bound REC is positioned in a manner that differs slightly from how it is positioned in a near-cognate TC-bound REC. This suggests that positioning of the TC within the low-E_{FRET} TC-bound REC state is an important, yet transient, structural response to the recognition of a cognate codon at the A-site of the REC.

Despite the success of smFRET in identifying and characterizing the highly transient low-E_{FRET} TC-bound REC state, it is important to note that there are one or more states preceding this state in the reaction pathway that play important roles during aa-tRNA selection and that remain to be identified and characterized. For example, the very first state resulting from the codon-independent, initial binding of the TC to the REC, which is likely to have properties that are similar or identical to those of an encounter complex, has yet to be observed (Fig. 1.3). In addition, the fact that the low-E_{FRET} state is not detected in smFRET experiments in which a TC is delivered to an REC carrying a non-cognate A-site codon suggests that, either: (i) the low-E_{FRET} state is sampled too rarely to be detected due to current experimental limitations (*e.g.*, the low concentrations of acceptor-labeled TC that must be used to maintain the signal-to-background required to observe single fluorophores, the rate with which the donors on the RECs photobleach, etc.); (ii) the low-E_{FRET} state is sampled too transiently to be detected due to current limitations on the time resolution of the experiments; (iii) non-cognate TCs are recognized and discriminated against within a state that precedes and is physically distinct from the low-E_{FRET} state, but in which the distance between the donors and acceptors in the currently available fluorophore labeling schemes are too far away to generate a detectable E_{FRET} ; or (iv) a combination of these possibilities. Recent single-molecule fluorescence co-localization microscopy experiments that use nanofabricated, microfluidic, observation flow-cells to overcome the concentration bar-

rier (see Sec. 1.5.4), for example, have demonstrated that fluorophore-labeled TCs do indeed transiently colocalize with fluorophore-labeled RECs carrying non-cognate A-site codons. Although these were not smFRET experiments in which the measured E_{FRET} could be compared to what is observed in RECs carrying near-cognate or cognate A-site codons, they represent an important step towards understanding where in the pathway and how RECs discriminate against non-cognate TCs [107]. Thus, weakly interacting, highly transient states that are critical for fully understanding the physical and molecular mechanisms underlying aa-tRNA selection during translation remain to be identified and characterized – representing an ongoing challenge for the field.

1.3.2 Pre-translocation Dynamics

Prior to translocation and in the absence of EF-G (Fig. 1.2), at least three individual structural elements of the PRE complex undergo thermally driven conformational fluctuations: (i) the P- and A-site tRNAs fluctuate between their classical P/P and A/A configurations and their hybrid P/E and A/P configurations (where, relative to the classical P/P and A/A configurations, the hybrid P/E and A/P configurations are characterized by the movement of the acyl acceptor ends of the P- and A-site tRNAs from the P and A sites of the 50S subunit into the E and P sites of the 50S subunit, respectively); (ii) the ribosome fluctuates between its non-rotated and rotated subunit orientations (where, relative to the non-rotated subunit orientation, the rotated subunit orientation is characterized by a counterclockwise rotation of the 30S subunit relative to the 50S subunit when viewed from the solvent-accessible side of the 30S subunit) [108]; and (iii) the L1 stalk of the 50S subunit fluctuates between its open and closed conformations (where, relative to the open L1 stalk conformation, the closed L1 stalk conformation is characterized by movement of the L1 stalk into the intersubunit space such that it can make a direct contact with the hybrid P/E-configured tRNA) (Fig. 1.4) [109]. Because of the stochastic nature of thermally driven processes, the tRNAs, ribosomal subunits, and L1 stalk within an ensemble of PRE complexes will asynchronously fluctuate between these transiently populated states in the absence of EF-G. While this structural heterogeneity impedes ensemble studies of these dynamics, they have been successfully characterized by single-molecule methods [43, 80, 96, 110–118].

Remarkably, smFRET studies performed by Fei and coworkers have observed PRE complexes fluctuating between two discrete states: (i) global state 1 (GS1), characterized by classically configured tRNAs, non-rotated subunits, and an open L1 stalk, and (ii) global state 2 (GS2), characterized by hybrid-configured tRNAs, rotated subunits, and a closed L1 stalk [110, 111]. The observation that the PRE complex fluctuates between just two states in the smFRET studies of Fei and coworkers is consistent with numerous

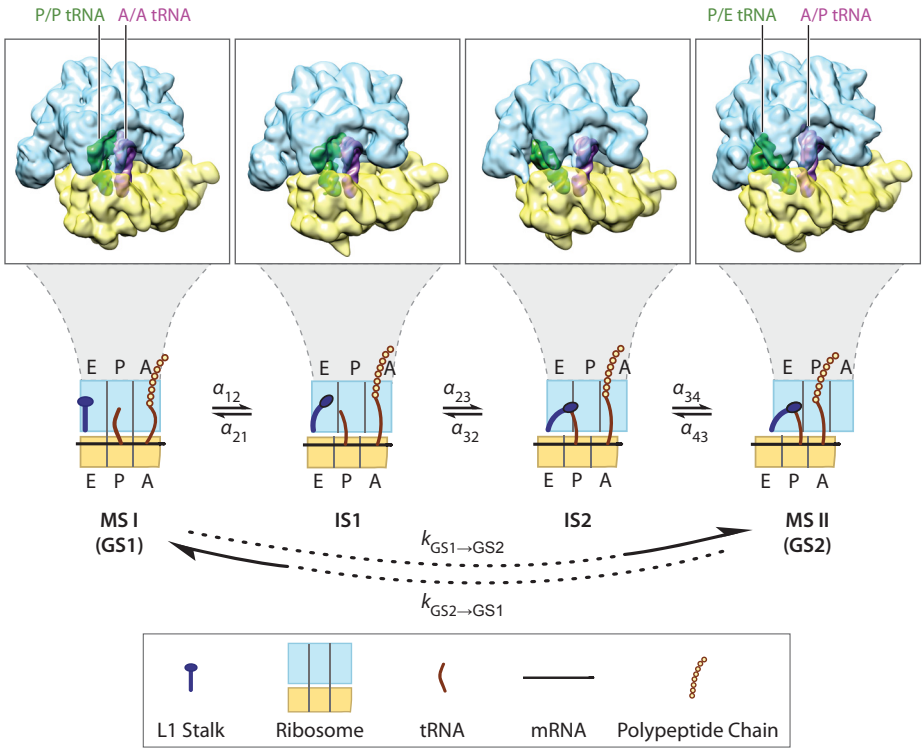

Figure 1.4: Cartoon schematic mechanism of PRE complex fluctuations. After peptide-bond formation, the PRE complex fluctuates between MS I/GS1 and MS II/GS2, passing through IS1 and IS2, until EF-G catalyzed translocation occurs.

subsequent smFRET studies from several other groups in which the tRNAs, ribosomal subunits, or L1 stalk elements of PRE complexes are also observed to fluctuate between just two states corresponding to the classical and hybrid tRNA configurations, the non-rotated and rotated subunit orientations, or the open and closed L1-stalk conformations, respectively [110–114, 118, 119]. Furthermore, the initial observation by Fei and coworkers that the PRE complex fluctuates between GS1 and GS2 is consistent with the more recent observation that fluctuations of the tRNAs between their classical and hybrid configurations, the ribosomal subunits between their non-rotated and rotated orientations, and the L1 stalk between its open and closed conformations are physically coupled, and coordinated by the ribosome in order to maximize and regulate the efficiency of translocation [118]. smFRET studies reveal that the thermodynamics and kinetics of the equilibrium between GS1 and GS2 are sensitive to: (i) the presence, identity, and acylation status of the P-site tRNA [43, 80, 96, 110–116]; (ii) the presence and acylation status of the A-site tRNA [43, 80, 96, 110–116]; (iii) the binding of EF-G [96, 110–114]; (iv) Mg^{2+} concentration [115]; (v) temperature [117]; (vi) the binding of ribosome-targeting antibiotic inhibitors of translocation [115, 118, 120]; and (vii) the perturbation

of intersubunit rotation via disruption of specific ribosomal intersubunit interactions [116, 118]. Collectively, these studies have provided deep insights into the roles that the P- and A-site tRNAs, EF-G, antibiotics, cooperative conformational changes, and allostery play in regulating translocation.

Cryo-electron microscopy (cryo-EM) studies of PRE complexes performed by Agirrezabala and coworkers have directly observed two states that are presumably the structural equivalents of GS1 and GS2, termed macrostate 1 (MS I) and macrostate 2 (MS II), respectively [108, 121–123]. More recent molecular dynamics simulations of cryo-EM-derived structural models of MS I and MS II support this presumption, and found that, in the studies of Fei and coworkers, the E_{FRET} observed in GS1 and GS2 are consistent with the simulations of MS I and MS II, respectively [124]. However, in addition to MS I and MS II, a more recent analysis of the cryo-EM data set of Agirrezabala *et al.* has revealed the presence of two additional states that are presumably intermediate between MS I/GS1 and MS II/GS2 along the reaction coordinate [125]. Neither of these two intermediate states, referred to here as intermediate state 1 (IS1) and intermediate state 2 (IS2), nor any others, were detected in the smFRET studies of Fei and coworkers [110, 111, 118, 119] or several other groups [112–114]. In contrast, smFRET studies of PRE complexes by Munro and coworkers identified and characterized two additional states that presumably lie along the reaction coordinate between MS I/GS1 and MS II/GS2 [80, 116]. However, these studies relied heavily on the use of smFRET data collected using ribosomes in which a substitution mutation disrupts a critical ribosome-tRNA interaction, and consequently causes the P-site tRNA in the resulting intermediate states to adopt conformations that are very different from those observed in either IS1 or IS2 in PRE complexes formed using wild-type ribosomes [125].

Since the smFRET experiments of Fei and coworkers and the cryo-EM experiments of Agirrezabala and coworkers interrogate PRE complexes composed of wild-type ribosomes (*i.e.,* without mutations that disrupt ribosome-tRNA interactions), it is highly likely that IS1 and IS2 are also present in the smFRET data, but that, given the spatial resolution, time resolution, and/or signal-to-background ratio (SBR) of the smFRET experiments, IS1 and IS2 do not produce large enough changes in E_{FRET} to be distinguished from GS1 or GS2. By connecting the results from static, equilibrium-state, ensemble experiments, such as cryo-EM, with the results from dynamic, time-dependent, single-molecule experiments, such as smFRET, through a theoretical framework, these hypotheses can be tested, and the energy landscape where the PRE complex exists can be characterized more precisely (see Sec. 3.4).

1.4 Prokaryotic Translation Termination

By noticing that mutations to certain codons in a gene resulted in truncated protein products in a site-specific manner [126], the codons UAA, UAG, and UGA were recognized as encoding the signal for nascent polypeptide chain termination (stop-codons) [127–129]. While early workers imagined that a special tRNA not carrying an amino acid would recognize these mRNA stop-codons through base pairing and then catalyze chain termination, Capecchi showed that a protein 'release factor' was actually responsible recognizing stop-codons and catalyzing nascent polypeptide chain termination [30]. Further work demonstrated that, in *Escherichia coli* (*E. coli*), nascent polypeptide chain termination is performed by two different release factors, RF1 and RF2, with distinct stop-codon specificities; RF1 is responsible for termination at the UAA and UAG stop-codons, while RF2 is responsible for termination at the UAA and UGA stop-codons[130]. These observations raised the questions of if and how a protein can recognize an mRNA stop-codon in the absence of base-pairing, if and how a release factor can both recognize stop-codons in the DC and catalyze chain termination in the PTC, and how RF1 and RF2 can both recognize the UAA stop-codon, but have differential specificities for the UAG and UGA stop-codons?

In general, translation termination occurs when, in competition with TC carrying aa-tRNA, a class I release factor binds to the A-site of a ribosome containing a peptidyl-tRNA in the P-site, and then directly recognizes a stop-codon in the DC [131–135] (Fig. 1.2). Upon stop-codon recognition, the class I release factor catalyzes hydrolysis of the ester bond between the nascent polypeptide chain and the tRNA in the PTC [66, 136, 137]. Following hydrolysis, a GTPase class II release factor, RF3 in bacteria, catalyzes dissociation of the class I release factor [138–140], priming the ribosomal post-termination complex for ribosome recycling, so that it can begin another round of protein synthesis [141–143] (Fig. 1.2B).

In bacteria, class I release factors undergo premature termination at a non-stop codon about 1 in 10^5 codons [32]. Notably, this *in vivo* error rate of 10^{-5} ensures that on average over 250 proteins are completely translated in *E. coli* before a premature termination event occurs [144], and additionally, this error rate is greater than the misincorporation rate during translation elongation of 10^{-3} to 10^{-4}, since premature termination is much more deleterious than random misincorporation [9]. Moreover, extensive biochemical measurements of the accuracy of termination at all codons differing from a stop-codon by one nucleotide (near-stop codons) using an optimized *in vitro* translation system corroborates that the high level of accuracy for translation termination seen *in vivo* is due to the release factors [145]. For near-stop codons, large $\mathrm{K_M}$ increase account for the majority of the observed accuracy of the class I release factors, with some small

contribution from decreased values of k_{cat} [145]. Regardless, this error rate is somehow achieved in the absence of a traditinonal kinetic proofreading scheme, since RF3, the only GTPase involved in termination, decreases the accuracy of translation termination *in vitro* [145]. Moreover, it is difficult to imagine that the thermodynamic energy differences between some stop and near-stop codons can account for the level of discrimination observed [10, 11, 145, 146]. Additionally, it is known that aminoglycoside antibiotics, which affect the accuracy of translation elongation [9, 147, 148], also affect the accuracy of translation termination [10, 149, 150], despite having different binding sites [150–152]. Furthermore, the A-site codon context (*i.e.*, the bases before (...-2, -1), and after (+4, +5...) the A-site codon) also modulates the efficiency with class I release factors catalyze chain termination [153, 154]. Furthermore, class I release factor mutations to residues that are distal from the DC when the release factor is bound to the ribosome at a stop-codon can change the release factor codon specificity, creating 'omnipotent' release factors that recognize all three stop-codons (UAA, UAG, and UGA) [155, 156]. Finally, chemical probing experiments of RF1 binding to stop-codon and near-stop codon programmed ribosomes show conformational rearrangements within and around the DC and PTC of the stop-codon programmed ribosomes, suggesting that codon recognition involves distinct conformational changes. These observations all suggest that there are additional molecular requirements for accurate translation termination other than just the direct RNA-protein interactions made between the stop-codon and the highly conserved 'tripeptide anticodon' of the class I release factors that directly contacts the stop-codons [131–135, 157–160], and that codon recognition is a multistep process, providing opportunity for some sort of biological proofreading mechanism during translation termination [14].

Once nascent polypeptide chain termination occurs, RF3 assists in the dissociation of the class I release factor from the ribosomal A-site [138]. Using smFRET, Sternberg and coworkers showed that RF1 stably binds the A-site of the ribosome, blocking the GS1 $\rightarrow$ GS2 transition, locking the ribosome into the non-rotated state with a classical P/P tRNA [161]. However, RF3 bound to the non-hydrolyzable GTP analogue GDPNP, in order to mimic the GTP-bound RF3 (RF3(GTP)) conformation, dissociates RF1 from the ribosome, and locks the ribosome into the rotated state with a hybrid P/E tRNA by blocking the GS2 $\rightarrow$ GS1 transition [161, 162]. Mechanistically, RF3(GTP) binds both the pre- and post-hydrolysis ribosome, but only after hydrolysis of the nascent polypeptide chain has been catalyzed by a class I release factor will RF3(GTP) catalyze dissociation from the ribosome [139, 140], presumably by biasing the ribosome towards GS2, which seems incompatible with class I release factor binding [161, 162].

1.4.1 Structure and Function of Release Factor 1

In this thesis, I investigate how the dynamics of RF1 and the ribosome regulate RF1 binding and codon discrimination (see Chapter 5). RF1 is a 360 amino acid protein encoded by the prfA gene (Fig. 1.5). The prfA gene is located in the *hemA*-operon at 27 min in *E. coli*, where it is under the control of two promoters, and has a poor translation initiation region – mutations to which cause an anti-amber supressor phenotype, because of the resulting increase in RF1 concentration [163]. RF1 expression levels depend upon growth rate, with about 5,000 molecules per cell for a 25 minute doubling time; notably, there are about 25,000 molecules of RF2 per cell under the same conditions [164]. In most bacteria, prfA is a non-essential gene, presumably because only a few essential genes end with the UAG stop-codon ($\sim$7), and therefore RF2 can compensate for the absence of RF1; though, RF1 is essential in K-12 strains, however, where a point mutation in RF2 (A246T) has made RF2 five times less active at UAA stop codons, and presumably less able to compensate for the absence of RF1 [165]. Interestingly, in pull-down experiments using RF1 as a bait protein, RF1 is found to interact with both DnaK and GroL [166], which are both central players in the bacterial chaperone network [167]; this is not the case for RF2, and suggests that RF1 has specific folding requirements. Finally, RF1 and RF2 are both post-translationally methylated at a universally conserved glutamine residue (Q235; *E. coli* numbering) by the N^5-glutamine methyltransferase encoded by prmC, which is downstream of prfA in the *hemA*-operon; this methylation is required for efficient catalysis during peptide hydrolysis, and a knockout induces translation termination defects [163, 168–171].

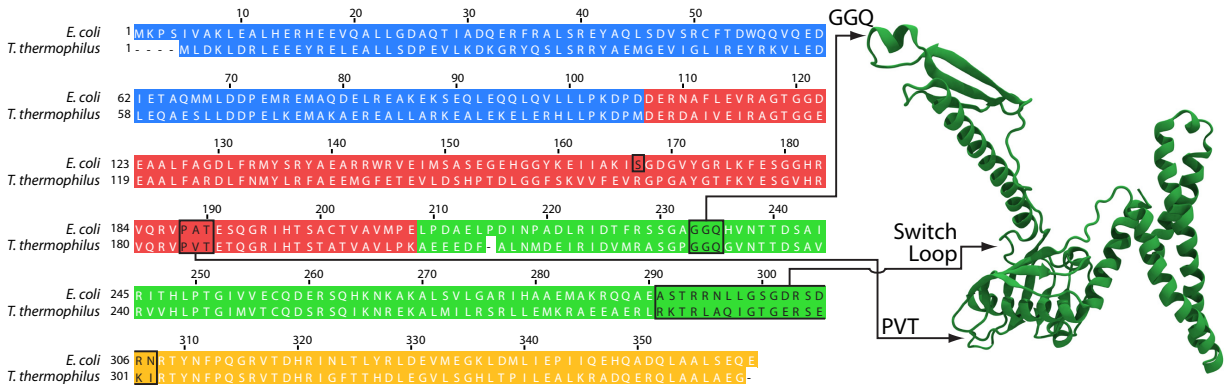

Figure 1.5: Structure of Release Factor 1. (Left) Primary Structure of *E. coli* RF1 aligned with *T. thermophilus* RF1 (PDB ID: 4V7P). Domains I through IV are denoted with blue, red, green, and yellow bars, respectively. The PxT motif, GGQ motif, and switch loop are marked, and their locations in the *T. thermophilus*, ribosome-bound RF1 X-ray crystallographic structure is shown (right). Additionally, the residue 167, which is used for acceptor-fluorophore labeling in smFRET experiments [161], is also marked.

RF1 is composed of four domains (Fig. 1.5). The N-terminus of RF1 comprises domain I, which interacts

with the rProtein L11 of the ribosomal GTPase-associated center [131, 132, 135]. *In vitro*, L11 is required for accurate stop-codon recognition ($k_{\mathrm{cat}}/\mathrm{K_M}$) of both UAA and UAG by RF1, but not for UAA and UGA by RF2 [172]. Also, *in vivo*, UAG read-through is increased in an L11 knockout strain [173]. Domain I of RF1 also interacts with RF3 [174], but it is not essential for peptide hydrolysis [175], does not lock RF1 into the GS1 state [161], and it is not required for RF3-catalyzed dissociation [161, 175].

Domains II and IV form a core domain which contacts the mRNA, rRNA, and rProteins in the DC [131, 132, 135]. Domain II contains the universally conserved, 'tripeptide anticodon' proline-any amino acid-threonine (PxT) motif, which directly interrogates the A-site codon [157] (Fig. 1.5). Only the threonine (T190) of the PxT motif directly contacts the mRNA, where it donates a hydrogen bond to the U1 base, and accepts a hydrogen bond from the A2 base; another residue, T198, interrogates the Hoogsteen edge of third base of the codon [31, 131, 132, 135, 176]. Chemical probing of the DC rRNA with a RF1 single-cysteine mutant (H156C) in domain II using Fe-BABE chemistry yields no hits when RF1 is bound at a stop-codon, presumably because the binding of RF1 to the ribosome in this area is too tight for the ascorbic acid to enter [177]. Notably, Wilson and coworkers describe the creation of several single-cysteine mutants in this domain that are competent for peptide hydrolysis [178]. Of these, Sternberg and coworkers utilized the S167C mutant as part of a smFRET signal between RF1 and the P-site aa-tRNA, which reports on RF1 binding [161]; this is the mutant used in Chapter 5.

Domain III directly contacts the PTC, $\sim$75 Å from the DC [131, 132, 135]. There, a universally conserved glycine-glycine-glutamine (GGQ) motif is directly involved in catalyzing the chain termination reaction [175, 179–181]. While the exact mechanism is still unclear, either a water or a hydroxide ion acts as a nucleophile [137, 181–183] which is coordinated by the N^5-methylated glutamine (Q235) of this GGQ motif to position the nucleophile and/or stabilize the transition state for the hydrolysis reaction of the nascent polypeptide chain from the 3' CCA end of the P-site tRNA [66, 131, 132, 134–136]. Between domains III and IV lies the 'switch loop' (A291 to N307), which is so disordered such that it could not be found in X-ray crystallography structures of the RF1 alone [184], or RF1 in complex with its N^5-glutamine methyltransferase [185], but which undergoes a conformational rearrangement and is relatively stabilized when RF1 is bound to the stop-codon programmed ribosomes [132, 135]. In the latter structures, the universally conserved 16S rRNA residues A1492 and A1493 in helix h44 form part of the switch loop binding pocket [132, 135], however, their conformations are similar to those in paromomycin bound-30S subunits [151], which are over-extended and not dynamic as in the analogous NMR structures, and therefore might be cryogenic or crystal induced artifacts [152]. Though, paromomycin was found to differentially affect RF1 catalyzed peptide hydrolysis at

stop- and near-stop codons [150]. Mutations in the RF1 switch loop coupled with deletion of helix H69 of the 23S rRNA, which forms part of intersubunit bridge B2A with helix h44 of the 16S rRNA, result in peptide hydrolysis defects at the UAA stop-codon [135]. Additionally, RF1 switch loop mutants G301S and R303H were found to decrease the efficiency of peptide hydrolysis at UAG stop-codons but not UAA stop-codons, suggesting that the switch loop participates in the codon recognition process [186].

There is considerable debate about the conformation of RF1 in solution. When bound to stop-codon programmed ribosomes, RF1 adopts an extended 'open' conformation where domains II and IV are located near the DC, while domain III has extended and is located in the PTC [131, 132, 135, 158] (Fig. 1.5). However, in X-ray crystallography structures of RF1 alone [184], or in complex with its N^5-glutamine methyltransferase [185], RF1 adopts a compact 'closed' conformation where domain III is folded down upon domains II and IV. One attractive hypothesis is that RF1 is in the closed conformation in solution, and upon binding and recognizing a stop-codon, it undergoes a conformational change to the open conformation where the GGQ motif is positioned in the PTC for peptide hydrolysis (reviewed in Ref. 31). However, such a conformational change has never been observed, and small-angle X-ray scattering experiments do not agree about the solution structure of release factors [187, 188]. Though, there is chemical probing evidence that RF1 undergoes different conformational changes when binding at stop- or near-stop codons, and that part of this change involves rearrangements of the rRNA near the switch loop and possibly orienting domain III into the PTC [177]. Regardless, it is unclear how the structure and dynamics of RF1 and the ribosome modulate RF1 binding affinity and codon discrimination in order to maintain the high fidelity of translation termination observed both *in vitro* and *in vivo* (see Chapter 5).

1.5 Single-molecule Fluorescence Resonance Energy Transfer Studies of Ribosomes

The advent of single-molecule biophysical techniques, particularly single-molecule fluorescence approaches, has provided a powerful complement to traditional ensemble biochemical, bioinformatics, and biophysical techniques for studying the ribosome. Single-molecule techniques allow stochastic transitions to rarely and transiently sampled states to be directly observed, thereby allowing those states that were difficult to detect in ensemble biophysical experiments to be sensitively detected and comprehensively characterized [189]. Moreover, the observation of individual transitions to and from a state of interest allows heterogeneities in the physical properties of that state to be identified, sorted, and analyzed [190–192]. However, despite the

tremendous promise that single-molecule fluorescence approaches hold for studies of the ribosome, these approaches are often limited by (i) data collection throughput that can be too low for sufficient statistical analysis, (ii) time resolutions that can be too slow to capture exceedingly transient states, (iii) limitations in spatial resolution that can make it difficult to determine when a ligand is co-localized to a biomolecule, and (iv) the 'concentration barrier' generated by the background fluorescence from the relatively high concentrations of fluorophore-labeled ligands that are required to observe ligand binding events within the experimental observation time, which compromises the high signal-to-background ratio that is required to sensitively observe the fluorescence from a single molecule [193–195]. Notably, single-molecule fluorescence resonance energy transfer (smFRET) approaches are particularly powerful for studies of ligand-binding reactions [196–198], as they ameliorate several of these limitations, and are able to directly report on both the spatial localization as well as the conformational dynamics of biomolecules and/or ligands [199].

For over a decade, ribosomes have been the subject of numerous smFRET studies [43, 95]. Several aspects of the experimental systems available for studying ribosomes provide important advantages for designing, implementing, and interpreting smFRET experiments. These advantages include: (i) the involvement of interactions between often only two molecular sub-complexes, the ribosome and a translation factor; (ii) the availability of a reconstituted in vitro translation system composed of a full set of purified components that allows individual components and steps of the reaction to be manipulated (reviewed in Ref. 200); (iii) the existence of a large number of small-molecule inhibitors that enable inhibition of specific and well-defined steps of translation [147, 148, 201]; and, (iv) the availability of a series of cryogenic electron microscopy (cryo-EM) and X-ray crystallography structures of conformations, including intermediates, relevant to translation initiation, elongation, and termination (reviewed in Refs. 69, 64, and 31, respectively). In this thesis, because of these many benefits, I utilize smFRET to study the ligand-binding dynamics of the ribosome during translation termination. Therefore, in the following sections, I will describe the benefits of single-molecule studies, the physical process underlying smFRET, and finally the experimental platform used to perform smFRET studies.

1.5.1 Single-molecule *versus* Ensemble Studies

Biophysicists are mostly concerned with understanding the general mechanisms that underlie biological, physico-chemical systems. In order to test whether their hypotheses are generally accurate, a large number of measurements must be considered, as it is necessary to obtain a high degree of precision when evaluating the predictions of these hypotheses. One approach to obtaining this precision is to use ensemble studies, in

which multiple, different molecules are measured at the same time, and the resulting measurement provides an average description of their behavior; this is an ensemble measurement. A fundamentally different approach is to measure each molecule separately, one at a time, and then determine the underlying behavior from the distribution of the different molecules; this is a singe-molecule measurement. However, this latter approach can be very time consuming, and often it is not possible to interrogate each subject separately. For example, imagine trying to measure the flocking behavior of birds. It is fairly easy to follow the flight path of the entire flock (the ensemble measurement). However, it is much more difficult to even identify an individual bird in the flock, let alone follow its flight path while inside the flock and repeat this measure for multiple other birds (the "single-molecule" measurement).

Regardless of the difficulties of performing single-molecule measurements, the benefits of single-molecule studies of biomolecules are immense [189]. For instance, single-molecule studies allow the observation of states that might otherwise be insufficiently populated, such as a rarely-populated intermediate, or intermediate states that are too transiently populated to be observed in an ensemble measurement, such as the codon-recognition intermediate state present in the aa-tRNA selection pathway [43, 95]. Additionally, they allow the observation of time-dependent fluctuations in biomolecular behavior, such as the memory effects observed in the substrate turn-over rate of the β-galactosidase enzyme [192]. Single-molecule studies also allow for the specific mechanistic pathways between states to be directly observed, such as the different spliceosome assembly pathways taken due to the competition between different small nuclear ribonucleoprotein particles assembling at a splice-site [202].

Fortunately, single-molecule fluorescence techniques are very sensitive methods for detecting individual biomolecules [203]. This is because the Stokes shift efficiently separates the fluorescence signal from the illumination background. Indeed, from the first hint that optical methods could observe a single-molecule [204] to the first single-molecule imaging of biomolecular reaction in solution [205], fluorescence has been a powerful single-molecule technique. By site-specifically labeling a biomolecule of interest with a single fluorophore, the fluorescence photons can be detected and attributed to the presence of the fluorophore-labeled biomolecule. Furthermore, if the fluorophore-labeled biomolecule can be spatially separated from from other fluorophore-labeled biomolecules by a distance greater than the diffraction limit, then the position of the individual biomolecule can be tracked through time. Finally, the fluorescence from multiple, spatially-separated, fluorophore-labeled biomolecules can be measured in parallel with wide-field detectors. With all of this in mind, single-molecule fluorescence techniques are powerful methods by which to perform single-molecule studies of biomolecules.

1.5.2 Förster Resonance Energy Transfer

The physical process underlying the ability of smFRET to report upon the conformational dynamics of a biomolecule is resonance energy transfer [206]. Resonance energy transfer is a process in which energy is transfered between a donor and an acceptor molecule through the electromagnetic interaction of their transition dipole moments. First observed in 1922 as sensitized fluorescence in atomic vapor spectroscopy [207–209], early explanations for resonance energy transfer employed classical electromagnetism to describe energy transfer between dipoles, but these yielded the wrong distance dependence of the energy transfer rate (R^{-3} not R^{-6}) [210]. The theory behind this process was correctly explained by Förster in 1948, and has since carried the name Förster Resonance Energy Transfer (FRET) [211]. Förster's insights were to use a quantum mechanics based approach with Fermi's golden rule, and to quantitatively account for the probability that the donor and acceptor transition dipole moments have the same energy through connections to experimentally measured absorption and emission spectra (reviewed in Ref. [212]). In this section, I will demonstrate how to briefly derive an expression for the efficiency of FRET that accounts for the donor-acceptor dipole-dipole interaction, and then give Förster's quantitative expression for this efficiency.

The general approach to calculating the FRET efficiency is to first calculate the rate of FRET transfer using Fermi's golden rule with a perturbation Hamiltonian obtained from a multipole-expansion of the Coulomb potential. First, consider the electric potential, Φ, from a collection of net-neutral charges. In Cartesian coordinates,

$$\begin{aligned}\Phi &= \sum_i \frac{q_i e}{\epsilon}\left((x-x_i))^2 + (y-y_i))^2 + (z-z_i))^2\right)^{-1/2} \\ &\equiv \sum_i \frac{q_i e}{\epsilon}\frac{1}{\vec{r}_i},\end{aligned} \tag{1.1}$$

where $q_i e$ is the charge of the i^{th} atom, and ϵ is the permittivity. A Taylor expansion around the origin in x_i, y_i, and z_i truncated after the first order term yields the dipole approximation

$$\begin{aligned}\Phi &= \frac{1}{\epsilon}\left(\sum_i \frac{q_i e}{\vec{r}} - mu_x \cdot \frac{\partial \frac{1}{\vec{r}}}{\partial x} - mu_y \cdot \frac{\partial \frac{1}{\vec{r}}}{\partial y} - mu_z \cdot \frac{\partial \frac{1}{\vec{r}}}{\partial z} + ...\right) \\ &= \frac{1}{\epsilon r^3}\left(\mu_x x + \mu_y y + \mu_z z\right) \\ &= \frac{1}{\epsilon r^3}\vec{\mu}\cdot\vec{r},\end{aligned} \tag{1.2}$$

where $\mu_x = \sum_i q_i e x_i$, and the sum in Φ is zero because it is assumed that there are an equal number of opposite charges. The electric field at $\vec{r}$ from the dipole $\vec{\mu}$ is then

$$\begin{aligned}
\vec{E} &= -\nabla\Phi \\
&= -\left((\vec{\mu}\cdot\nabla)\frac{\vec{r}}{\epsilon r^3} + (\vec{r}\cdot\nabla)\frac{\vec{\mu}}{\epsilon r^3} + \frac{\vec{\mu}}{\epsilon r^3}\times(\nabla\times\vec{r}) + \frac{\vec{r}}{\epsilon r^3}\times(\nabla\times\vec{\mu}) \right) \\
&= \frac{-1}{\epsilon}\left(\frac{\vec{\mu}}{r^3} - \frac{3(\vec{r}\cdot\vec{\mu})\cdot\vec{r}}{r^5} + 0 + 0 \right), \text{ since } \frac{\partial 1/r}{\partial x} = \frac{-1}{r^2}\frac{\partial r}{\partial x} = \frac{-1}{r^2}\hat{x} = \frac{-x}{r^3} \\
&= \frac{1}{\epsilon r^3}\left(3\hat{r}\cdot\vec{\mu}\cdot\hat{r} - \vec{\mu}\right). \qquad (1.3)
\end{aligned}$$

In order to obtain the rate of FRET, k_{FRET}, time-dependent perturbation theory and Fermi's golden rule are used. Here, our system is the defined by the ground state (S_0) or excited state (S_1) of the donor (D) or accepter (A) molecules. The corresponding wavefunction is

$$\Psi = \psi_a + \psi_b = a\cdot\langle D(S_1), A(S_0)| + b\cdot|D(S_0), A(S_1)\rangle\,, \qquad (1.4)$$

where $a + b = 1$, which is a linear combination of the cases when only the donor is the excited state and when only the acceptor is in the excited state. When the donor is in the excited state, the transition dipole moment between the excited and ground states creates an electric field perturbation that can effect the transition dipole moment of the acceptor molecule. The corresponding perturbation Hamiltonian for the acceptor-dipole donor-dipole electric field interaction is

$$\mathcal{H} = \mathcal{H}^0 + \mathcal{H}^1 = \mathcal{H}^0 - \vec{\mu_{\mathrm{A}}}\cdot\vec{E_{\mathrm{D}}}. \qquad (1.5)$$

From the time-dependent Schrödinger equation, if $a = 1$ and $b = 0$ at time $t = 0$, Fermi's golden rule yields

$$\begin{aligned}
k_{\mathrm{FRET}} = W_{\psi_a\to\psi_b} &= \frac{\partial}{\partial t}\langle\psi_b|\psi_b\rangle = \frac{\partial}{\partial t}|b(t)|^2 \\
&\propto |\langle D(S_0), A(S_1)|\mathcal{H}^1|D(S_1), A(S_0)\rangle|^2 \\
&\propto |-\vec{\mu}\cdot\vec{E}|^2 \\
&\propto \left|\frac{-1}{\epsilon r^3}\left(3(\vec{\mu}_A\cdot\hat{r})(\vec{\mu}_D\cdot\hat{r}) - \vec{\mu}_A\cdot\vec{\mu}_D\right)\right|^2 \\
&\propto \frac{\mu^2\kappa^2}{n^4 r^6}, \qquad (1.6)
\end{aligned}$$

where $n = \sqrt{\epsilon}$ is the index of refraction, and the substitution in the last line defines κ the orientation factor. Now, the FRET efficiency, E_{FRET} , is defined as

$$\begin{aligned} \mathrm{E_{FRET}} &\equiv \frac{k_{\mathrm{FRET}}}{k_{\mathrm{FRET}} + \frac{1}{\tau_e}} \\ &= \frac{1}{1 + \frac{n^4 r^6}{C\mu^2\kappa^2\tau_e}}, \end{aligned} \tag{1.7}$$

where C is some proportionality constant required to make Equation 1.6 an equivalency, and τ_e is the average, donor excited-state lifetime in the absence of acceptor. As an efficiency, E_{FRET} is bounded between 0 and 1, corresponding to no donor, excited states relaxing through FRET (*i.e.,*, all relaxation occurs through fluorescence), and all donor, excited states relaxing through FRET.

Two of the terms present in Equation 1.7 highlight very interesting aspects of FRET. First, E_{FRET} depends upon r^6, where r is the distance between the donor and acceptor transition dipole moments. Because of this responsive distance dependence, if E_{FRET} can be measured, it can be used as a 'spectroscopic-ruler', in which high levels of E_{FRET} correspond to a small donor-acceptor distance, while low levels of E_{FRET} correspond to a large donor-acceptor distance [213]. Second, the κ^2 term provides a dipole-dipole orientation contribution to E_{FRET} . If the donor and acceptor dipole moments were completely orthogonal to each other, no FRET would be expected to occur, even if the donor and accepter molecules were relatively close together. However, if all relative orientations are sampled equally, the average value is κ^2 = 2/3. Similarly, with isotropic rotation of the donor and acceptor, in a bulk sample the ensemble averaged $\langle\kappa^2\rangle$ = 2/3 also, since all relative orientations are sampled within the ensemble. However, for a single-pair of donor and acceptor molecules, each FRET event samples only one relative orientation, which will not necessarily be κ^2 = 2/3. Therefore, in order to approximate the average $\bar{\kappa^2}$ = 2/3 for a single-pair of isotropically rotating donor and acceptor molecules, multiple FRET events must occur. In practice, however, for freely rotating molecules only about 15 photons (*i.e.*, FRET events) must be observed before $\bar{\kappa^2} \approx$ 2/3 to within 10% relative error [214].

Further than Equations 1.6 and 1.7, a quantitative expression can be written down for k_{FRET} and E_{FRET} . Förster showed that E_{FRET} can be written as

$$\mathrm{E_{FRET}} = \frac{1}{1 + \frac{\mathrm{r}^6}{\mathrm{R}_0^6}}, \tag{1.8}$$

where R_0 is known as the Förster radius, which is the distance between donor and acceptor molecules

at which E_{FRET} = 0.5 [211, 212]. While R_0 is a function of the donor and acceptor molecules, as well as their environment, and can therefore be calibrated by measuring E_{FRET} for known separations, Förster also developed a quantitive expression for calculating R_0's, which uses parameters that can be measured from the donor and acceptor molecules separately, without any FRET. This expression is

$$R_0^6 = \frac{9000 \ln(10)}{128\pi^5} \cdot \frac{\kappa^2 \Phi_D}{n^4 N_A} \cdot J(\nu), \tag{1.9}$$

where κ is the orientation factor, Φ_D is the fluorescence quantum yield of the donor molecule in the absence of FRET, n is refractive index, N_A is Avagadro's number, and $J(\nu)$ is the spectral overlap integral where ν is wavenumbers [211, 212]. The spectral overlap integral quantifies the probability that the donor and acceptor transition dipole moments will have the same oscillator strength when FRET occurs. This can be calculated from the normalized emission spectrum of the donor molecule, $f^D(\nu)$, and the normalized absorption spectrum of the acceptor molecule, $\epsilon^A(\nu)$, as

$$J(\nu) = \int_0^\infty \frac{f^D(\nu) \cdot \epsilon^A(\nu) d\nu}{\nu^4}. \tag{1.10}$$

Förster's theory generally provides very accurate descriptions of experimental FRET data for r values that are approximately 0.5-10 nm; notably, below this range, where the dipole expansion is not valid, Dexter energy transfer occurs [215]. For many pairs of donor and acceptor molecules, such as fluorophores in the visible range of the electromagnetic spectrum, in a particular environment such as aqueous buffer, R_0 values have been tabulated, and can be used to reasonable accuracy in other experiments. In this work, I utilize the fluorophore pair Cy3 and Cy5, where Cy3 is the donor molecule, and Cy5 is the acceptor molecule, which have an R_0 = 55 Å (Fig. 1.6).

1.5.3 smFRET Microscopy Experimental Platform

By combining single-molecule measurements, the distance dependence of E_{FRET} , and site-specific fluorophore labeling strategies for biomolecules, smFRET allows one to observe the time dependent, conformational dynamics of individual biomolecules. Here, I will describe our experimental platform for making such smFRET measurements of ligand binding events for hundreds of individual ribosomes in parallel, in the presence of high concentrations fluorophore-labeled ligand in solution, by using wide-field microscopy (reviewed in Ref. 200).

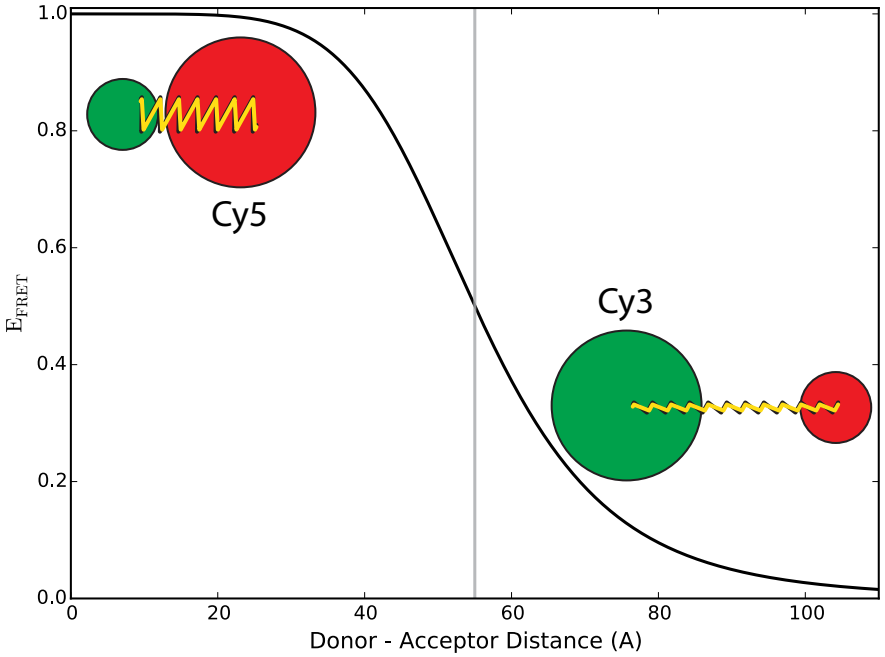

Figure 1.6: Efficiency of Förster Resonance Energy Transfer for Cy3-Cy5. E_{FRET} is plotted versus the distance between the donor fluorophore (Cy3) and the acceptor fluorophore (Cy5); R_0 = 55 Å . FRET is cartooned as yellow zig-zags.

Since E_{FRET} depends monotonically on the distance between the donor and acceptor fluorophores [211, 212, 215, 216], it can interpreted as a "spectroscopic-ruler" that reports on the distance between the two fluorophores [213]. smFRET utilizes this distance dependence by monitoring the E_{FRET} between individual donor- and acceptor-fluorophore labeled biomolecules, and using that E_{FRET} as a proxy for the distance between the location of the donor and acceptor-fluorophore labeled biomolecules [206, 217, 218]. Assuming that the only relaxation pathways available to an excited donor fluorophore are fluorescence or FRET, and that the only relaxation pathway available to an excited acceptor fluorophore is fluorescence, the E_{FRET} of a resonance energy-transfer event can be quantified by directly exciting the donor fluorophore, and then measuring the ratio of a large number of photons emitted through donor and acceptor fluorescence. This ratio is

$$\mathrm{E_{FRET}} = \frac{\mathrm{k_{RET}}}{\mathrm{k_{RET}} + \mathrm{k_{fl}^{D}}} = \lim_{\mathrm{n_A}+\mathrm{n_D}\to\infty} \frac{\mathrm{n_A}}{\mathrm{n_A} + \mathrm{n_D}} \approx \frac{\mathrm{I_A}}{\mathrm{I_A} + \mathrm{I_D}}, \tag{1.11}$$

where n_A and n_D are the number of acceptor and donor photons, and I_A and I_D are fluorescence intensities of the acceptor and the donor, respectively, as recorded by a detector such as an electron multiplying charge coupled device (EMCCD). Perhaps the most common experimental approach for measuring I_A

and I_D for ligand-binding reactions is diagrammed in Fig. 1.7. This approach involves tethering a biotin-derivatized, donor-labeled biomolecule of interest to the surface of a microfluidic, observation flow-cell via a biotin-streptavidin-biotin bridge formed with biotin-derivatized polyethylene glycol (PEG) that has been used to functionalize the surface of the flow-cell [43, 219]. Introduction of an acceptor-labeled ligand into the imaging buffer within the flow-cell and imaging using wide-field total internal reflection fluorescence (TIRF) microscopy [220–222] enables direct excitation of the surface-localized donors and simultaneous observation of both the donor- and acceptor fluorescence intensities originating from hundreds of individual biomolecules [43, 218]. In such experiments, an anti-correlated donor- and acceptor fluorescence intensities versus time trajectory that exhibits single-step fluorophore photobleaching originating from diffraction-limited donor and acceptor spots serves as unambiguous evidence for resonance energy transfer arising from the encounter of a single, acceptor-labeled ligand with a single, surface-tethered, donor-labeled biomolecule. Importantly, the acceptor fluorophore will only fluoresce when it is within tens of Å of the donor (*e.g.*, when a ligand is bound to the biomolecule). Thus, smFRET enables the detection of biomolecule-ligand encounters in the presence of relatively higher, tens to one hundred nM, concentrations of fluorophore-labeled ligand in the imaging buffer than is possible with other single-molecule fluorescence microscopy approaches – thereby partially alleviating the limitations imposed by the concentration barrier described earlier [195]. In addition to reporting on ligand-biomolecule encounters, smFRET can also report on conformational changes of the biomolecule and/or ligand that result in changes in the distance between the positions of the donor and acceptor. In summary, because it is able to simultaneously and sensitively report on hundreds of single biomolecule-ligand encounters at fluorophore-labeled ligand concentrations that are higher than is possible with other single-molecule fluorescence approaches, as well as because it can report on conformational changes that take place during the ligand-binding reaction, the TIRF-based smFRET experimental platform described here has been particularly successful for investigating the kinetic and thermodynamic properties of ligand-binding reactions [189].

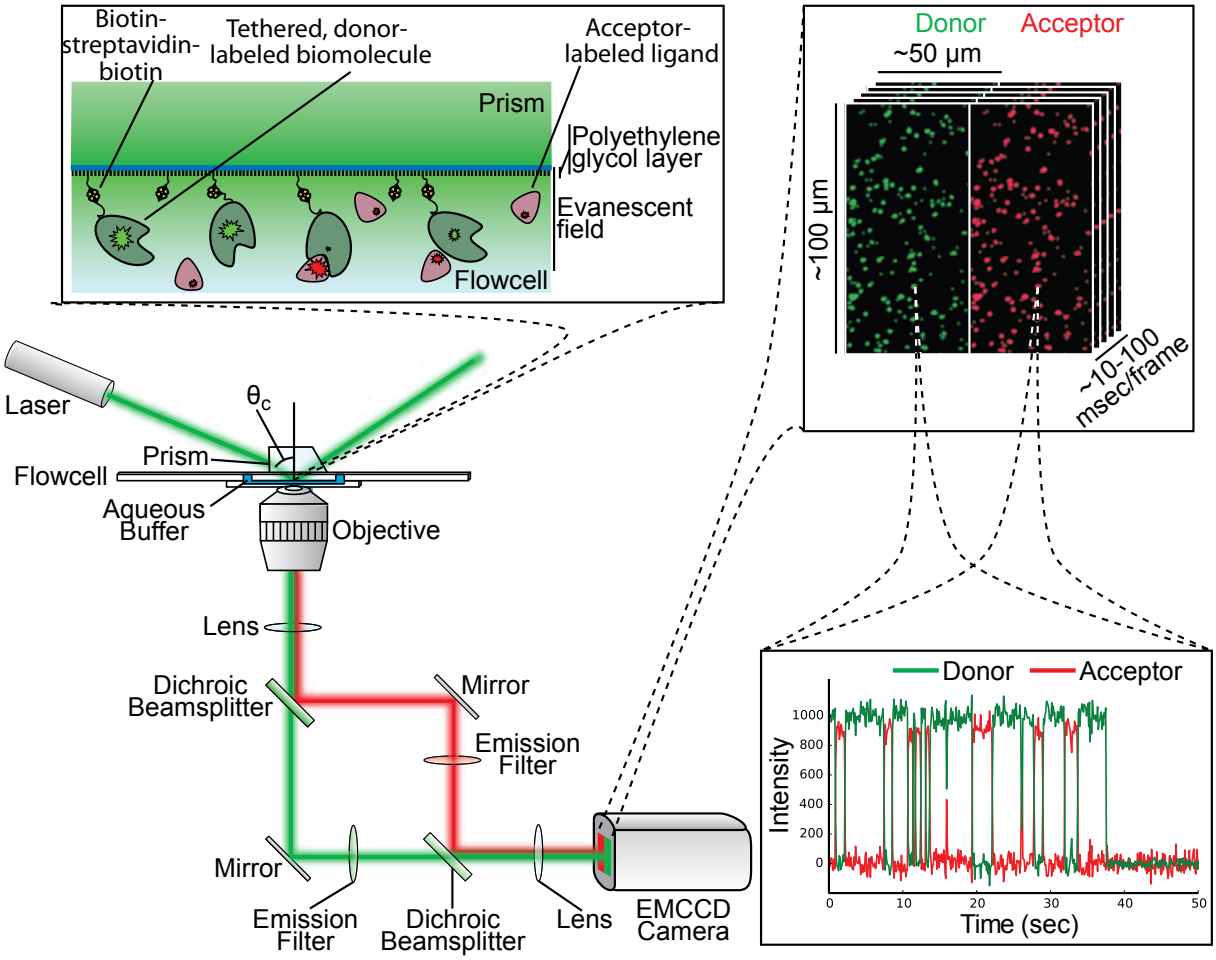

Figure 1.7: Schematic of smFRET experimental platform for Studying Ligand-binding Reactions. A laser excitation source is totally internally reflected (TIR) at the interface formed between the surface of the quartz, microfluidic, observation flowcell to which the donor-labeled biomolecules are tethered and the aqueous imaging buffer in the flowcell. The evanescent field that is generated by TIR propagates into the imaging buffer and decays exponentially as a function of increasing distance from the quartz-buffer interface, thereby selectively exciting only those donors that are localized within ∼300 nm of the quartz-buffer interface (top left inset). Donor and acceptor fluorescence emission is collected by an objective, separated by wavelength using dichroic mirrors, and detected using an electron-multiplying charge-coupled device (EMCCD) camera (bottom left inset). The separated donor and acceptor fluorescence intensities reporting on the binding of acceptor-labeled ligands to individual, spatially resolved, donor- labeled biomolecules can then be quantified and plotted as a function of time (bottom right inset). Figure and caption adapted from Ref. 223

1.5.4 Emerging experimental advances for smFRET studies of ligand-binding reactions

Very low affinity ligand binding reactions often have (i) very slow association rates, which makes binding events very rare, and therefore it is difficult to observe a significant number of binding events in a smFRET experiment; (ii) very fast dissociation rates, which makes the bound-state lifetimes very transient, and therefore it is very difficult to observe the binding process in a smFRET experiment; (iii) high equilibrium dissociation constants ($\mathrm{K_D}$), which makes it necessary to have a high concentrations of fluorophore-bound ligand in solution, and therefore compromises the signal-to-background ratio (SBR) in an smFRET experiment. In order to address these limitations to studying low-affinity ligand binding reactions using smFRET there are a number of emerging experimental advances beyond the principal experimental platform described in Section 1.5.3, which I describe below.

Microscope-based developments

As described in Section 1.5, one of the current limitations in single-molecule fluorescence studies of ligand-binding reactions is the concentration barrier that is created by the relatively high concentrations of fluorophore-labeled ligands that are required to be present in the imaging buffer in order to detect ligand-binding events within the experimental observation time [194]. Because an acceptor fluorophore will only be efficiently excited via FRET when it is within tens of Å of a donor fluorophore that is being directly excited by an excitation light source (*e.g.*, a laser), smFRET experiments in which a donor-labeled biomolecule of interest is tethered to the surface of the microfluidic, observation flowcell and an acceptor-labeled ligand is supplied in the imaging buffer greatly ameliorate the limitations imposed by the concentration barrier [224]. Despite this advantage of smFRET experiments, acceptor fluorophores can still be directly, albeit inefficiently, excited by the laser that is used to directly excite the donor fluorophore – a low probability event known as excitation crosstalk [189]. At high enough concentrations of acceptor-labeled ligands in the imaging buffer (typically well below the equilibrium dissociation constants of weakly interacting biomolecule-ligand complexes), excitation crosstalk becomes a major source of background fluorescence noise in such experiments, a situation that can ultimately prevent the detection of fluorescence from individual molecules [194].

By preventing the excitation source from penetrating deeper than $\sim$200-300 nm into the imaging buffer in the flowcell, wide-field TIRF microscopy-based smFRET experiments minimize the total amount of noise from excitation crosstalk of acceptor-labeled ligands in the imaging buffer. Unfortunately, the major disad-

vantage of this approach is that the electron-multiplying charge-coupled device (EMCCD) cameras that are used as detectors in TIRF microscopy-based smFRET experiments operate at time resolutions that are typically too slow to capture the weakest and most transient intermediate states (typically limited to $\sim$30 ms per a full-resolution frame [225]). This tradeoff between lower background noise and faster time resolution is a fundamental limitation of using wide-field smFRET to study encounter complexes. Using a confocal fluorescence microscope with an avalanche photodiode (APD) or a single-photon avalanche diode (SPAD) detector, rather than a TIRF microscope with an EMCCD camera detector, affords significant increases in the sensitivity, SBR, and time resolution (typically $\sim$ 1 ms per data point) of smFRET experiments [189, 226]. Unfortunately, however, this increase in the time resolution comes at the cost of a significant decrease in throughput, as the confocal fluorescence microscopy-based approach can only image one biomolecule at a time, whereas the TIRF microscopy-based approach can simultaneously image hundreds of surface-tethered biomolecules at a time. Such low throughput can be particularly difficult to overcome when attempting to detect weakly interacting, transient biomolecule-ligand complexes and/or conformational states that are rarely-populated. In the ideal setup, one would be able to couple a TIRF microscope with an APD- or SPAD-array detector [225, 227]. Such an approach would allow wide-field detection with APDs or SPADs, thereby allowing the high throughput associated with using a conventional TIRF microscope, but with the high time resolution that is associated with using a confocal fluorescence microscope. Towards this goal, smFRET was recently demonstrated using linear arrays of 8 SPADs [228], and so, with further development, commercially available products composed of tens of thousands of monolithically integrated SPADs will likely constitute an important future development in the field of single-molecule fluorescence microscopy [227]. Towards this goal, I have succeeded in integrating a 2,048 SPAD-array with our TIRF microscope, and successfully utilized this setup to observe pretranslocation ribosomes with smFRET (see Appendix A.2).

An alternative approach to minimizing the excitation crosstalk of acceptor-labeled ligands and improving the SBR of conventional TIRF microscopy-based smFRET experiments without compromising throughput is to remove excess ligand from the excitation volume without changing the local concentration of the ligand in the vicinity of the biomolecule. This can be achieved by locally confining ligands and the biomolecule of interest into spatially resolved regions of the flowcell that are smaller than the excitation volume associated with the detection of single, diffraction-limited spots in the field-of-view. Notable approaches include electrostatic and physical traps [229, 230] as well as encapsulation of the biomolecules and ligands in aqueous drops in oil [231]. Perhaps the most promising and physiologically compatible method of local confinement for smFRET experiments is to encapsulate ligands and individual biomolecules in surface-tethered lipo-

somes [232]. These lipid vesicles not only encapsulate ligands with individual biomolecules and confine them to spatially resolved regions on the surface of the microfluidic flowcell, but, because the biomolecule is free to diffuse within the surface-tethered liposome, they also serve to further reduce the interactions that a biomolecule of interest might have with the PEG-passivated surface of the microfluidic flowcell [233]. In addition, since their introduction in 2001 [234], liposomes have been optimized to allow for the free diffusion of small molecules into and out of the liposome [235].

Fluorophore developments

Although much work has gone into developing robust donor and acceptor pairs with well-behaved photophysics for smFRET studies [236], photobleaching and 'blinking' of fluorophores continues to impose limitations on smFRET studies of biomolecular systems [237, 238]. The irreversible photobleaching of a fluorophore effectively brings the smFRET experiment, at least for that fluorophore, to an end, thereby limiting the experimental throughput. The reversible blinking of a fluorophore similarly limits the throughput of smFRET experiments and, in addition, can convolute the subsequent analysis of the data (*e.g.*, it is often difficult to distinguish whether changes in E_{FRET} that are observed arise from *bona fide* changes in the distance between the donor and acceptor or from blinking of the donor or acceptor) [237, 238]. The limitations imposed by photobleaching and blinking are particularly challenging for smFRET studies of ligand-binding reactions. Rare events such as the formation of a long-lived complex formed by the generally weak, transient binding of an acceptor-labeled ligand to a surface-tethered, donor-labeled biomolecule of interest, for example, are exceedingly difficult to observe within the limited observation times that are imposed by photobleaching and blinking of the donor fluorophore. The continued development of longer-lived, brighter, and more stable fluorophores, such as Cy3B [239], is therefore an important area of research, as these enhanced fluorophores enable longer and more stable observation times in smFRET experiments [236, 240].

Complementing the development of longer-lived, brighter, and more stable fluorophores, considerable work has also gone into optimizing buffer conditions that minimize photobleaching and blinking [237]. Often, photobleaching is mediated via a photochemical reaction between a fluorophore in its electronically excited state and molecular oxygen. As a consequence, oxygen scavenging systems, including mixtures of glucose, glucose oxidase, and catalase or of protocatechuic acid (PCA) and protocatechuate-3,4-dioxygenase (PCD), are often added to experimental buffer systems to extend fluorophore survival times [241, 242]. As early as 1988, β-mercaptoethanol (BME) was being added to buffer systems for single-molecule fluorescence microscopy experiments in order to suppress the blinking of tetramethyl-rhodamine, presumably

by quenching a dark triplet-state of the fluorophore [241]. Since then, triplet-state quenchers, including mixtures of cyclooctatetraene (COT) [43], 4-nitrobenzyl alcohol (NBA) [43], and/or Trolox (a water-soluble analog of vitamin E) [243] have become standard additives in smFRET experiments [244]. Recently, direct, covalent conjugation of triplet-state quenchers such as COT, NBA, or Trolox to the entire class of cyanine dye fluorophores, including Cy3 and Cy5, was shown to significantly enhance the photostability of these fluorophores [245, 246]. Although these new fluorophore-quencher conjugates represent a breakthrough in fluorophore development, caution should be exercised in their use, as these conjugates are necessarily larger than the traditional fluorophores on which they are based and are therefore more likely to sterically perturb the biomolecules or ligands to which they are attached.

In addition to the development of fluorophores and/or buffer conditions that resist photobleaching and blinking, the development of fluorophores that enable selective excitation is another promising area of current research [238]. Indeed, one of the major advantages of smFRET is that the acceptor is excited selectively (*i.e.*, only when it is within tens of Å of a donor). Similarly, highly enzyme-specific fluorogenic substrates that fluoresce exclusively upon being enzymatically modified can be used to conduct single-molecule fluorescence microscopy studies of the mechanisms of action of enzymes for which fluorogenic substrates have been developed [192, 240]. This strategy, however, is difficult to generalize to all, or even many, enzymes and even more so to non-enzymatic biomolecules. In addition, to our knowledge, fluorogenic strategies have not yet been utilized in smFRET studies. As an alternative to the use of fluorogenic substrates, selective excitation can be achieved through the use of photoswitchable fluorophores [238]. For example, in the PhADE technique, an imaging buffer containing high concentrations of a photoswitchable fluorophore-labeled ligand is activated with a laser pulse and these are allowed to diffuse out of the imaging volume. This effectively removes all of the photoactivated ligands other than those that are bound to the surface-tethered biomolecules from the flowcell, thereby removing much of the background fluorescence that would otherwise arise from photoactivated ligands in solution [247]. Unfortunately, this strategy works only for relatively long-lived, weakly interacting biomolecule-ligand complexes, and, to our knowledge, has not yet been used in smFRET experiments.

Perhaps the most promising recent development in smFRET studies of biomolecular systems is the use of FRET-based quenchers as acceptors for smFRET experiments [248]. By returning from the excited state to the ground state without emitting a photon, quenchers not only free the optical spectrum for the use of additional fluorophores (*e.g.*, for colocalization experiments), but, of particular importance for smFRET studies of ligand-binding reactions, also allow smFRET experiments to be performed under conditions in which very

high concentrations of quencher-labeled ligands can be included in the imaging buffer without significantly increasing the fluorescence background. This is because a quencher that is excited by excitation crosstalk will not fluoresce, thereby minimizing the contribution of excitation crosstalk to the fluorescence background. Notably, a FRET-based quencher was used to report on the large-scale conformational dynamics of the ribosome during translation elongation [249, 250], although this approach was used to free the optical spectrum and expand the number of fluorophores that could be simultaneously detected, rather than to use FRET-based quencher-labeled ligands to overcome the concentration barrier. Notably, in our own efforts to utilize quenchers in order to observe the binding and dissociation of RF1 from near-stop codons, the increased hydrophobicity of the quenchers BHQ2 and QSY9 relative to more traditional fluorophores, such as Cy3, induced RF1 aggregation. However, future development of quenchers to include hydrophilic groups such as sulfonates could help to overcome this limitation.

Photonic developments

A general strategy for increasing the sensitivity of single-molecule fluorescence experiments is to enhance the fluorescence signal from the molecules of interest. Perhaps the most widely used approach for accomplishing this is plasmon-mediated enhancement, which has traditionally been used to enhance optical spectroscopies (*e.g.*, surface plasmon-enhanced Raman spectroscopy) and, in single-molecule fluorescence applications, can be used to decrease illumination volumes and increase fluorescence signals [251]. Nanofabricated plasmonic bowties [252] and plasmonic nanoantennas [253–255] have both been shown to dramatically increase signal-to-background ratios in single-molecule fluorescence applications. Of particular importance for single-molecule fluorescence studies of ligand-binding reactions, the 'antenna-in-a-box' platform has used plasmon-based enhancement to permit the observation of single-molecule fluorescence in the presence of micromolar concentrations of fluorophore in the imaging buffer [256]. Unfortunately, few of these platforms have thus far demonstrated their applicability with smFRET [253] and, while these technologies show promise, the fact that highly specialized facilities, equipment, and technical skills are required to nanofabricate the plasmon-producing structures (*e.g.*, the bowties and antennas) onto the surface of the microfluidic flow-cells that are used in single-molecule fluorescence experiments has thus far hindered widespread adoption by the community.

Another general strategy for attaining higher signal-to-background ratios in single-molecule fluorescence studies is to decrease the fluorescence background by reducing the excitation volume. For this purpose, arrays of nanoapertures – nanoscopic wells that have been fabricated into a thin metallic layer that has been

deposited onto the surface of a microfluidic flowcell (see Chapter 2) – have proven themselves very useful. Popularized as "zero-mode waveguides" (ZMW), the geometric confinement of light in the nanoaperture leads to a zeptoliter excitation volume at the very bottom of the nanoaperture. This permits the sensitive observation of a single, fluorophore-labeled molecule that has been tethered or otherwise localized to the bottom of the nanoaperture in a background of micromolar concentrations of fluorophore-labeled molecules in the imaging buffer [257]. Currently, nanoaperture arrays are used in Pacific Biosciences' next-generation, single-molecule real-time (SMRT) DNA sequencing technology. There, a single DNA polymerase is tethered to the bottom of each nanoaperture, and a nucleotide-specific, single-molecule fluorescence signal is observed as the polymerase incorporates each nucleotide into the nascent DNA molecule that is being synthesized [258].

Unfortunately, in the decade since nanoaperture arrays were first introduced for single-molecule fluorescence applications [257], only a handful of biomolecular systems beyond DNA replication have been investigated using this technology [107, 250, 259–261]. Notably, studies of all but one of these biomolecular systems have been investigated in collaboration with Pacific Biosciences and only one of these studies has used nanoaperture arrays in an smFRET application. As is the case with plasmon-based enhancement technologies, widespread adoption of nanoaperture-array technology has likely been hindered by the limited availability of the resources required for nanofabrication. In addition, the non-specific adsorption of biomolecules to the metallic and glass surfaces of the nanoapertures has likely limited their applications. Non-specific adsorption not only alters working concentrations by sequestering molecules from solution, but it also compromises single-molecule resolution by non-specifically localizing multiple, fluorophore-labeled molecules to the surfaces near the bottom of the nanoapertures. In the case of SMRT sequencing via DNA replication, a very successful polyvinyl phosphonic acid-based, nanoaperture-surface passivation scheme was developed that minimizes the non-specific adsorption of small, negatively charged nucleotide triphosphates to the surface of aluminum-based nanoapertures [262]. Given the negatively charged nature of the polyvinyl phosphonic acid passivation layer, however, it is likely that this passivation scheme will not be generally applicable to other biomolecules (*e.g.*, positively charged globular proteins).

As part of this thesis work, an alternative passivation scheme was developed that uses thiol-based self-assembled monolayers (SAMs) of PEG to robustly passivate gold-based nanoapertures (see Chapter 2) [224]. Given the widespread success of PEG-based passivation schemes in single-molecule fluorescence studies, I anticipate that PEG-passivated nanoaperture arrays will provide a more general solution to the problem of non-specific adsorption of biomolecules to the nanoaperture surfaces and will thereby enable

studies of a wide-range of biomolecular systems using nanoaperture arrays. Perhaps most excitingly, the use of these gold-based nanoapertures such as the PEG-passivated, gold based nanoapertures also allows the surface-plasmon of the gold to be used for plasmon-based enhancement of the fluorescence signals at the bottom of the nanoapertures, an approach that has been recently demonstrated [263–265], and that could potentially facilitate the use of nanoaperture arrays for smFRET applications.

Computational developments

In order to discuss recent improvements in the analysis of smFRET data, it is useful to begin with a brief, general review of how kinetic information is obtained from single-molecule experiments and, more specifically, smFRET experiments. Traditional treatments of chemical kinetics are insufficient to describe systems where fluctuations from the average behavior are important, such as systems with small numbers of molecules, such as is the case for single-molecules [266]. Instead, alternative approaches that treat molecules as independent, stochastic entities are used [267]. To interpret the kinetic data yielded by a single molecule from a smFRET experiment, a framework is used which describes the distribution of times that the molecule spends dwelling in a particular state [268]. To extract these lengths of the dwell-times from single-molecule E_{FRET} versus time trajectories, methods such as hidden Markov models (HMM) [269], change point analysis or wavelet analysis [270], which detect the time points at which switching between distinct E_{FRET} states occurs. It is worth noting that a complete theoretical description has been developed that allows high-resolution information about the conformational dynamics of a biomolecular system to be extracted from the joint distribution of E_{FRET} values and fluorescence lifetimes determined from experimentally observed photon bursts [271, 272]. Although the joint distribution provides more information regarding the conformational dynamics of the biomolecular system than the distribution of E_{FRET} values does alone, this approach is not generally applicable in traditional, wide-field, TIRF microscopy-based smFRET experiments, as these experiments do not monitor photon arrival times.

One of the most popular approaches to analyzing single-molecule E_{FRET} trajectories has been the application of HMMs, which yield the probabilities of transitioning between states as well the 'idealized' path between states. HMMs were first used in biology for analyzing conductance time series in single channel ion recordings [273], and first suggested for use with smFRET data in 2003 [269]. Since then, a number of software packages for HMM analysis of smFRET data have been published [274–280]. Although HMM analysis of smFRET data can provide the desired information necessary to develop mechanistic models of biomolecular function, HMM analysis approaches and software packages that use maximum likelihood

methods to estimate HMM parameters (*e.g.*, HaMMy [275], QuB [274], and SMART [278]) can result in overfitting of the smFRET data (*i.e.*, overestimating the number of states that can be confidently ascribed to the data), as the value of the likelihood function that is being maximized will always increase with the number of states included in the analysis. This overfitting problem has been recently addressed by HMM analysis approaches and software packages that use Bayesian inference methods to estimate HMMs (*e.g.*, vbFRET [276, 277] and ebFRET [279, 280]). Despite the success of HMMs for the analysis of smFRET data, however, it is important to note that HMMs are ill-suited for modeling E_{FRET} versus time trajectories containing rapid transitions, such as those into and out of energetically unstable, transiently sampled states (*e.g.*, weakly interacting biomolecule-ligand complexes with rapid rates of association and dissociation) and are inappropriate for modeling non-Markovian data (*e.g.*, data exhibiting 'dynamic disorder' where the probability of a transition changes with time) [276, 277, 279, 280].

Perhaps the most promising computational approaches for analyzing smFRET data from weakly interacting, transient biomolecule-ligand complexes employ Bayesian inference, and in doing so they are able to 'learn' from the data [281]. Bayesian inference approaches have been used in the analysis of smFRET data by removing noise from E_{FRET} versus time trajectories [282] and by analyzing photon bursts [283, 284]. Perhaps the most powerful applications, however, are those that use Bayesian inference methods to estimate the HMM parameters that are used to analyze E_{FRET} versus time trajectories [276, 277, 279, 280]. The vbFRET software package, for example, uses Bayesian inference on an HMM to select the number of states and rates of transitions between states (*i.e.*, the kinetic model) that best describes each individual E_{FRET} trajectory. This minimizes the overfitting problem faced by approaches and software packages that use maximum likelihood methods to estimate HMMs and subsequently rely on the user, or on an *ad hoc* metric, such as the Bayesian- or Akaike information criteria, in order to select the kinetic model that best describes the data [276, 281]. Moreover, vbFRET shows promise at detecting transiently sampled intermediate states that maximum likelihood approaches for estimating HMMs might otherwise miss, an extremely useful ability when studying weakly interacting, transient biomolecule-ligand complexes using smFRET [276].

One of the major limitations of most HMM-based approaches for smFRET data analysis is that these approaches provide a separate and unique kinetic model for each individual E_{FRET} trajectory. Ultimately, it is left up to the user to determine which single, consensus kinetic model best describes the entire population of kinetic models provided by the HMM-based analysis of the ensemble of E_{FRET} versus time trajectories that are typically obtained from a single smFRET experiment. Recently, van de Meent and coworkers have addressed this problem by expanding on the vbFRET framework and creating ebFRET. ebFRET uses a

Bayesian inference method to estimate HMM parameters that is analogous to the one that vbFRET uses, but does so on entire populations of E_{FRET} versus time trajectories rather than on individual E_{FRET} trajectories, thereby using all of the E_{FRET} versus time trajectories to learn the single, consensus kinetic model that best describes all of the available data [279, 280]. Because it uses all of the E_{FRET} versus time trajectories, ebFRET has an enhanced and statistically robust ability to detect rarely and transiently sampled states that might appear in some, but not all, of the E_{FRET} versus time trajectories associated with a single smFRET experiment. In addition, ebFRET can distinguish between states that have the same E_{FRET} but differ in their lifetimes. For example, ebFRET has been used to distinguish between two structurally similar conformations of the ribosome that yield the same E_{FRET} , but that differ in the lifetime of that state due to the presence or absence of a EF-G [280] – a development that dramatically extends the resolution of smFRET experiments.

One of the major challenges for HMM analyses of data collected from smFRET studies of weakly interacting, transient biomolecule-ligand complexes is that the encounter complexes and many of the intermediate states that are sampled are simply too energetically unstable, and too transiently sampled to be reliably modeled with a HMM. This is only likely to become more challenging as ever more sensitive, and higher time resolution single-molecule biophysical techniques are developed that increasingly render the most weakly interacting and transient biomolecule-ligand complexes accessible to study. To address this limitation, as part of this thesis work, a novel non-HMM-based Bayesian inference approach was developed which allows for temporal super-resolution beyond the acquisition rate of a detector (i.e., an EMCCD for smFRET). This approach is called **B**ayesian **I**nference for the **A**nalysis of **S**ub-temporal-resolution **D**ata (BIASD), and it is able to learn both the E_{FRET} values and rates of transitions into and out of transiently sampled states (*e.g.*, the rates of ligand association and dissociation from a biomolecule) even when the rates of these transitions are faster than the experimental time resolution (see Chapter 4). This method should therefore prove extremely useful in ongoing efforts to study weakly interacting, transient biomolecule-ligand complexes using smFRET.

1.6 Dissertation Overview and Motivation

Like other large molecular machines [13], the ribosome must harness random fluctuations induced by thermal energy in order to perform work relevant to cellular survival [21]. Under these conditions, it seems that the ribosome has evolved to exploit naturally occurring dynamics inherent to its architecture, as well as to induce cooperative conformational changes, in order to maximize the efficiency of translation in the cellular

environment [2, 21, 22, 118]. In this thesis, I work towards an understanding of how molecular interactions can be utilized to exploit such conformational dynamics during translation in order to maintain accurate protein synthesis (Section 1.1.1). In order to study these molecular interactions and conformational dynamics of the translational machinery, I primarily utilized smFRET since this approach is well-suited to studying the low affinity ligand binding reactions and transient conformational changes that are associated with non-cognate interactions (Section 1.5).

However, the transient nature of extremely low affinity interactions, such as RF1 interacting with a near-stop codon in the A-site, makes them difficult to study with smFRET TIRF microscopy. Therefore, part of this thesis work is devoted to developing smFRET methods to access the regimes where these interactions can be probed by smFRET. In Chapter 2, the development of a novel, robustly-passivated, gold-based nanoaperture array technology is discussed, which increases that allowed fluorophore-labeled ligand concentrations in solution for smFRET experiments by at least two orders of magnitude – breaking the concentration barrier. However, often such high concentrations of ligand in solution results in ligand binding reactions with dynamics that are too fast and/or transient to accurately observe. Therefore, in Chapter 3, the precision of quantifying such kinetics within the context of rare and/or transient events in E_{FRET} versus time trajectories is considered. Additionally, the accuracy of, and possible correction schemes for, calculating rate constants given the resulting missing ligand-binding events is investigated. Finally, I also investigate what occurs when the kinetic events are too fast to observe at all in an E_{FRET} versus time trajectory.

In order to address these shortcomings, I have developed a technique that achieves temporal super-resolution in smFRET experiments by learning the transition rate constants and E_{FRET} states from an E_{FRET} versus time trajectory using a Bayesian inference based approach called BIASD (Chapter 4). This method can accurately, and precisely obtain rate constants, which are several orders of magnitude faster than the acquisition rate of the detector. Furthermore, I show that this approach can be extended to account for the heterogeneities present in an experimental ensemble of single-molecules, and therefore learn those rate constants and E_{FRET} states for the various sub-populations present, in addition to determining the number of sub-populations in the sample, all in a united manner (Chapter 4). Finally, I computationally investigate how perturbations to conformational dynamics, which cannot be observed using smFRET, might allosterically contribute to regulating biomolecular interactions through the analysis of the coevolution of residues within RF1, as well as by applying network analysis to molecular dynamics simulations (Chapter 5). The development of these methods described above enabled several different studies of the regulation of translational processes through biomolecular interactions.

During the pretranslocation step of translation elongation, the ribosome uses thermal fluctuations to break and accurately reform an extensive number of non-covalent interactions in order to transition between GS1 and GS2 in a dynamic equilibrium. In Chapter 3, I investigate the presence of unstable intermediates in the PRE GS1 $\rightleftharpoons$ GS2 dynamic equilibrium by uniting smFRET measurements of PRE^{-A} complexes with cryo-electron microscopy structures of analogous PRE^{-A} complexes. In doing so, the dynamics of transient on-pathway intermediates that might be responsible for mediating the large scale conformational changes that the ribosome rectifies to perform work are explored. These dynamics, much like the multistep binding kinetics of an encounter complex, could act as points of regulation for nature to tune the efficiency of translation. Along these lines, in Chapter 4, I investigate the energetic contributions that various tRNA structural elements make to the GS1 $\rightleftharpoons$ GS2 energy landscape, which determines the energy differences and transition state energy barriers of the PRE complex, by using temperature-dependent smFRET studies of PRE complexes. Finally, during translation termination, class I release factors, such as RF1, must efficiently and accurately discriminate between stop-codons, and near-stop or sense codons in the A-site of the ribosome before catalyzing hydrolysis of the nascent polypeptide chain. In Chapter 5, I investigate the contribution that the conformational dynamics of RF1 and the ribosome make to modulating the RF1 binding affinity, and codon discrimination ability during translation termination. Together, these studies promote the view that the conformational dynamics of the translational machinery are modulated in order to regulate the overall process of protein synthesis.

1.7 References

1. Fersht, A. R. *Structure and mechanism in protein science. A guide to enzyme catalysis and protein folding.* 293–400 (W.H. Freeman and Co., New York, 1999).

2. McCammon, J. A. Protein dynamics. *Reports Prog. Phys.* **47,** 1–46 (1984).

3. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. *Molecular Biology of the Cell* 4th (Garland Science, New York, 2002).

4. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3,** 318–356 (1961).

5. Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190,** 576–581 (1961).

6. Crick, F. On protein synthesis. *Symp. Soc. Exp. Biol.* **12,** 138–63 (1958).

7. Berg, J. M., Tymoczko, J. L. & Stryer, L. *Biochemistry* 6th (W.H. Freeman, New York, 2006).

8. Johansson, M., Lovmar, M. & Ehrenberg, M. Rate and accuracy of bacterial protein synthesis revisited. *Curr. Opin. Microbiol.* **11,** 141–7 (2008).

9. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74,** 129–77 (2005).

10. Zaher, H. S. & Green, R. Fidelity at the molecular level: lessons from protein synthesis. *Cell* **136,** 746–62 (2009).

11. Rodnina, M. V. *Quality control of mRNA decoding on the bacterial ribosome.* 1st ed., 95–128 (Elsevier Inc., 2012).

12. Pauling, L. Nature of Forces between Large Molecules of Biological Interest. *Nature* **161,** 707–709 (1948).

13. Astumian, R. D. Thermodynamics and Kinetics of a Brownian Motor. *Science* **276,** 917–922 (1997).

14. Sartori, P. & Pigolotti, S. Kinetic versus energetic discrimination in biological copying. *Phys. Rev. Lett.* **110,** 1–5 (2013).

15. Hopfield, J. J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proc. Natl. Acad. Sci.* **71,** 4135–4139 (1974).

16. Ninio, J. Kinetic amplification of enzyme discrimination. *Biochimie* **57,** 587–595 (1975).

17. Ehrenberg, M. & Blomberg, C. Thermodynamic Constraints on Kinetic Proofreading in Biosynthetic Pathways. *Biophys. J.* **31,** 333–358 (1980).

18. Palade, G. E. A small particulate component of the cytoplasm. *J. Biophys. Biochem. Cytol.* **1,** 59–68 (1955).

19. Finka, A. & Goloubinoff, P. Proteomic data from human cell cultures refine mechanisms of chaperone-mediated protein homeostasis. *Cell Stress Chaperones* **18,** 591–605 (2013).

20. Lewin, B. *Genes VIII* 135 (Pearson Education, 2004).

21. Frank, J. & Gonzalez, R. L. Structure and dynamics of a processive Brownian motor: the translating ribosome. *Annu. Rev. Biochem.* **79,** 381–412 (2010).

22. Munro, J. B., Sanbonmatsu, K. Y., Spahn, C. M. T. & Blanchard, S. C. Navigating the ribosome's metastable energy landscape. *Trends Biochem. Sci.* **34,** 390–400 (2009).

23. Tissières, A., Watson, J., Schlessinger, D. & Hollingworth, B. Ribonucleoprotein particles from Escherichia coli. *J. Mol. Biol.* **1,** 221–233 (1959).

24. Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147,** 1462–1465 (1965).

25. Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General Nature of the Genetic Code for Proteins. *Nature* **192,** 1227–1232 (1961).

26. Matthaei, J. H., Jones, O. W., Martin, R. G. & Nirenberg, M. W. Characteristics and composition of RNA coding units. *Proc. Natl. Acad. Sci.* **48,** 666–77 (1962).

27. Ibba, M., Becker, H. D., Stathopoulos, C., Tumbula, D. L. & Söll, D. The adaptor hypothesis revisited. *Trends Biochem. Sci.* **25,** 311–316 (2000).

28. Fersht, A. R. & Kaethner, M. M. Mechanism of aminoacylation of tRNA. Proof of the aminoacyl adenylate pathway for the isoleucyl- and tyrosyl-tRNA synthetases from Escherichia coli K12. *Biochemistry* **15,** 818–23 (1976).

29. Ling, J., Reynolds, N. & Ibba, M. Aminoacyl-tRNA synthesis and translational quality control. *Annu. Rev. Microbiol.* **63,** 61–78 (2009).

30. Capecchi, M. R. Polypeptide Chain Termination In Vitro: Isolation of a Release Factor. *Proc. Natl. Acad. Sci.* **58,** 1144–1151 (1967).

31. Korostelev, A. a. Structural aspects of translation termination on the ribosome. *RNA* **17,** 1409–1421 (2011).

32. Jorgensen, F., Adamski, F. M., Tate, W. P. & Kurland, C. Release Factor-dependent False Stops are Infrequent in Escherichia coli. *J. Mol. Biol.* **230,** 41–50 (1993).

33. Bennett, C. H. Dissipation-error tradeoff in proofreading. *BioSystems* **11,** 85–91 (1979).

34. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5,** 789–96 (2009).

35. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.* **44,** 98–104 (1958).

36. Monod, J., Changeux, J.-P. & Jacob, F. Allosteric proteins and cellular control systems. *J. Mol. Biol.* **6,** 306–329 (1963).

37. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).

38. Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H. & Gunsalus, I. C. Dynamics of ligand binding to myoglobin. *Biochemistry* **14,** 5355–5373 (1975).

39. Frauenfelder, H., Parak, F. & Young, R. D. Conformational substates in proteins. *Annu. Rev. Biophys. Biophys. Chem.* **17,** 451–479 (1988).

40. Boehr, D. D., Dyson, H. J. & Wright, P. E. An NMR perspective on enzyme dynamics. *Chem. Rev.* **106,** 3055–3079 (2006).

41. Schnell, J. R. Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase. **33,** 119–140 (2004).

42. Williams, J. C. & McDermott, A. E. Dynamics of the flexible loop of triosephosphate isomerase: the loop motion is not ligand gated. *Biochemistry* **34,** 8309–8319 (1995).

43. Blanchard, S. C., Kim, H. D., Gonzalez, R. L., Puglisi, J. D. & Chu, S. tRNA dynamics on the ribosome during translation. *Proc. Natl. Acad. Sci.* **101,** 12893–8 (2004).

44. Tsai, C. J., Ma, B. & Nussinov, R. Folding and binding cascades: shifts in energy landscapes. *Proc. Natl. Acad. Sci.* **96,** 9970–9972 (1999).

45. Ma, B., Kumar, S., Tsai, C. J. & Nussinov, R. Folding funnels and binding mechanisms. *Protein Eng.* **12,** 713–720 (1999).

46. Ubbink, M. The courtship of proteins: understanding the encounter complex. *FEBS Lett.* **583,** 1060–6 (2009).

47. Schreiber, G. Kinetic studies of protein-protein interactions. *Curr. Opin. Struct. Biol.* **12,** 41–7 (2002).

48. Sheinerman, F. B., Norel, R. & Honig, B. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10,** 153–9 (2000).

49. Fersht, A. R. *Structure and mechanism in protein science. A guide to enzyme catalysis and protein folding* 349–400 (W.H. Freeman and Co., New York, 1999).

50. Cooper, A. & Dryden, D. T. F. Allostery without conformational change - A plausible model. *Eur. Biophys. J.* **11,** 103–109 (1984).

51. Parrondo, J. M. R., Horowitz, J. M. & Sagawa, T. Thermodynamics of information. *Nat. Phys.* **11,** 131–139 (2015).

52. Thompson, R. C. & Stone, P. J. Proofreading of the codon-anticodon interaction on ribosomes. *Proc. Natl. Acad. Sci.* **74,** 198–202 (1977).

53. Volkov, A. N., Worrall, J. a. R., Holtzmann, E. & Ubbink, M. Solution structure and dynamics of the complex between cytochrome c and cytochrome c peroxidase determined by paramagnetic NMR. *Proc. Natl. Acad. Sci.* **103,** 18945–50 (2006).

54. Iwahara, J. & Clore, G. M. Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature* **440,** 1227–30 (2006).

55. Tang, C., Iwahara, J. & Clore, G. M. Visualization of transient encounter complexes in protein-protein association. *Nature* **444,** 383–6 (2006).

56. Kaczanowska, M. & Ryden-Aulin, M. Ribosome Biogenesis and the Translation Process in Escherichia coli. *Microbiol. Mol. Biol. Rev.* **71,** 477–494 (2007).

57. Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G. & Yusupov, M. One core, two shells: bacterial and eukaryotic ribosomes. *Nat. Struct. Mol. Biol.* **19,** 560–567 (2012).

58. Lecompte, O., Ripp, R., Thierry, J. C., Moras, D. & Poch, O. Comparative analysis of ribosomal proteins in complete genomes: An example of reductive evolution at the domain scale. *Nucleic Acids Res.* **30,** 5382–5390 (2002).

59. Yusupov, M. M. *et al.* Crystal structure of the ribosome at 5.5 A resolution. *Science* **292,** 883–896 (2001).

60. Schuwirth, B. S. *et al.* Structures of the Bacterial Ribosome at 3.5 A Resolution. *Science* **310,** 827–834 (2005).

61. Bokov, K. & Steinberg, S. V. A hierarchical model for evolution of 23S ribosomal RNA. *Nature* **457,** 977–980 (2009).

62. Petrov, A. S. *et al.* Evolution of the ribosome at atomic resolution. *Proc. Natl. Acad. Sci.* **111,** 10251–10256 (2014).

63. Petrov, A. S. *et al.* History of the ribosome and the origin of translation. *Proc. Natl. Acad. Sci.* **112,** 15396–15401 (2015).

64. Voorhees, R. M. & Ramakrishnan, V. Structural basis of the translational elongation cycle. *Annu. Rev. Biochem.* **82,** 203–36 (2013).

65. Beringer, M. & Rodnina, M. V. The ribosomal peptidyl transferase. *Mol. Cell* **26,** 311–21 (2007).

66. Rodnina, M. V. The ribosome as a versatile catalyst: Reactions at the peptidyl transferase center. *Curr. Opin. Struct. Biol.* **23,** 595–602 (2013).

67. Schmeing, T. M. & Ramakrishnan, V. What recent ribosome structures have revealed about the mechanism of translation. *Nature* **461,** 1234–42 (2009).

68. Korostelev, A., Ermolenko, D. N. & Noller, H. F. Structural Dynamics of the Ribosome. *Curr. Opin. Chem. Biol.* **12,** 674–683 (2014).

69. Simonetti, A. *et al.* A structural view of translation initiation in bacteria. *Cell. Mol. Life Sci.* **66,** 423–436 (2009).

70. Shine, J. & Dalgarno, L. Determinant of cistron specificity in bacterial ribosomes. *Nature* **254,** 34–38 (1975).

71. Laursen, B. & Sørensen, H. Initiation of protein synthesis in bacteria. *Microbiol. ...* **236,** 747–771 (2005).

72. Fei, J. *Coupling of Ribosome and tRNA Dynamics during Protein Synthesis* PhD thesis (Columbia University: New York, 2010).

73. Rodnina, M. V. in *Ribos. Struct. Funct. Dyn.* (eds Rodnina, M. V., Wintermeyer, W. & Green, R.) 199–212 (Springer-Verlag, 2011).

74. Simonović, M. & Steitz, T. A. A structural view on the mechanism of the ribosome-catalyzed peptide bond formation. *Biochim. Biophys. Acta* **1789,** 612–23 (2009).

75. Moazed, D. & Noller, H. F. Intermediate states in the movement of transfer RNA in the ribosome. *Nature* **342,** 142–8 (1989).

76. Chen, J., Tsai, A., O'Leary, S. E., Petrov, A. & Puglisi, J. D. Unraveling the dynamics of ribosome translocation. *Curr. Opin. Struct. Biol.* **22,** 804–14 (2012).

77. Rodnina, M. V. & Wintermeyer, W. The ribosome as a molecular machine: the mechanism of tRNA-mRNA movement in translocation. *Biochem. Soc. Trans.* **39,** 658–62 (2011).

78. Cooperman, B. S. *et al.* in *Ribos. Struct. Funct. Dyn.* (eds Rodnina, M. V., Wintermeyer, W. & Green, R.) 339–348 (Springer-Verlag, 2011).

79. Noller, H. F. *et al.* in *Ribos. Struct. Funct. Dyn.* (eds Rodnina, M. V., Wintermeyer, W. & Green, R.) 349–360 (Springer-Verlag, 2011).

80. Munro, J. B., Altman, R. B., Tung, C.-S., Cate, J. H. D., Sanbonmatsu, K. Y. & Blanchard, S. C. Spontaneous formation of the unlocked state of the ribosome is a multistep process. *Proc. Natl. Acad. Sci.* **107,** 709–14 (2010).

81. Shoji, S., Walker, S. E. & Fredrick, K. Ribosomal translocation: one step closer to the molecular mechanism. *ACS Chem. Biol.* **4,** 93–107 (2009).

82. Gavrilova, L. P. *et al.* Factor-free ("non-enzymic") and factor-dependent systems of translation of polyuridylic acid by Escherichia coli ribosomes. *J. Mol. Biol.* **101,** 537–52 (1976).

83. Goodenbour, J. M. & Pan, T. Diversity of tRNA genes in eukaryotes. *Nucleic Acids Res.* **34,** 6137–46 (2006).

84. Bouadloun, F., Donner, D. & Kurland, C. G. Codon-specific missense errors in vivo. *EMBO J.* **2,** 1351–6 (1983).

85. Edelmann, P. & Gallant, J. Mistranslation in E. coli. *Cell* **10,** 131–7 (1977).

86. Kramer, E. B. & Farabaugh, P. J. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA* **13,** 87–96 (2007).

87. Laughrea, M., Latulippe, J., Filion, A.-M. & Boulet, L. Mistranslation in twelve Escherichia coli ribosomal proteins. Cysteine misincorporation at neutral amino acid residues other than tryptophan. *Eur. J. Biochem.* **169,** 59–64 (1987).

88. Grosjean, H. J., de Henau, S. & Crothers, D. M. On the physical basis for ambiguity in genetic coding interactions. *Proc. Natl. Acad. Sci.* **75,** 610–4 (1978).

89. Kierzek, R., Burkard, M. E. & Turner, D. H. Thermodynamics of single mismatches in RNA duplexes. *Biochemistry* **38,** 14214–23 (1999).

90. Rodnina, M. V. & Wintermeyer, W. Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu. Rev. Biochem.* **70,** 415–35 (2001).

91. Pape, T., Wintermeyer, W. & Rodnina, M. V. Complete kinetic mechanism of elongation factor Tu-dependent binding of aminoacyl-tRNA to the A site of the E. coli ribosome. *EMBO J.* **17,** 7490–7 (1998).

92. Rodnina, M. V., Pape, T., Fricke, R., Kuhn, L. & Wintermeyer, W. Initial Binding of the Elongation Factor Tu-GTP-Aminoacyl-tRNA Complex Preceding Codon Recognition on the Ribosome. *J. Biol. Chem.* **271,** 646–652 (1996).

93. Ruusala, T., Ehrenberg, M. & Kurland, C. G. Is there proofreading during polypeptide synthesis? *EMBO J.* **1,** 741–5 (1982).

94. Pape, T., Wintermeyer, W. & Rodnina, M. Induced fit in initial selection and proofreading of aminoacyl-tRNA on the ribosome. *EMBO J.* **18,** 3800–7 (1999).

95. Blanchard, S. C., Gonzalez, R. L., Kim, H. D., Chu, S. & Puglisi, J. D. tRNA selection and kinetic proofreading in translation. *Nat. Struct. Mol. Biol.* **11,** 1008–14 (2004).

96. Wang, Y. *et al.* Single-molecule structural dynamics of EF-G–ribosome interaction during translocation. *Biochemistry* **46,** 10767–75 (2007).

97. Lee, T.-H., Blanchard, S. C., Kim, H. D., Puglisi, J. D. & Chu, S. The role of fluctuations in tRNA selection by the ribosome. *Proc. Natl. Acad. Sci.* **104,** 13661–5 (2007).

98. Effraim, P. R. *et al.* Natural amino acids do not require their native tRNAs for efficient selection by the ribosome. *Nat. Chem. Biol.* **5,** 947–53 (2009).

99. Geggier, P. *et al.* Conformational sampling of aminoacyl-tRNA during selection on the bacterial ribosome. *J. Mol. Biol.* **399,** 576–95 (2010).

100. Carbon, J. & David, H. Studies on the thionucleotides in transfer ribonucleic acid. Addition of N-ethylmaleimide and formation of mixed disulfides with thiol compounds. *Biochemistry* **7,** 3851–8 (1968).

101. Seo, H.-S., Abedin, S., Kamp, D., Wilson, D. N., Nierhaus, K. H. & Cooperman, B. S. EF-G-dependent GTPase on the ribosome. conformational change and fusidic acid inhibition. *Biochemistry* **45,** 2504–14 (2006).

102. Stark, H., Rodnina, M. V., Wieden, H.-J., Zemlin, F., Wintermeyer, W. & van Heel, M. Ribosome interactions of aminoacyl-tRNA and elongation factor Tu in the codon-recognition complex. *Nat. Struct. Biol.* **9,** 849–54 (2002).

103. Valle, M. *et al.* Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy. *Nat. Struct. Biol.* **10,** 899–906 (2003).

104. Skogerson, L. & Moldave, K. Evidence for aminoacyl-tRNA binding, peptide bond synthesis, and translocase activities in the aminoacyl transfer reaction. *Arch. Biochem. Biophys.* **125,** 497–505 (1968).

105. Cool, R. H. & Parmeggiani, a. Substitution of histidine-84 and the GTPase mechanism of elongation factor Tu. *Biochemistry* **30,** 362–6 (1991).

106. Wolf, H., Chinali, G. & Parmeggiani, a. Mechanism of the inhibition of protein synthesis by kirromycin. Role of elongation factor Tu and ribosomes. *Eur. J. Biochem.* **75,** 67–75 (1977).

107. Uemura, S., Aitken, C. E., Korlach, J., Flusberg, B. a., Turner, S. W. & Puglisi, J. D. Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature* **464,** 1012–1017 (2010).

108. Frank, J. & Agrawal, R. K. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature* **406,** 318–22 (2000).

109. Frank, J. Intermediate states during mRNA-tRNA translocation. *Curr. Opin. Struct. Biol.* **22,** 778–85 (2012).

110. Fei, J., Kosuri, P., MacDougall, D. D. & Gonzalez, R. L. Coupling of ribosomal L1 stalk and tRNA dynamics during translation elongation. *Mol. Cell* **30,** 348–59 (2008).

111. Fei, J., Bronson, J. E., Hofman, J. M., Srinivas, R. L., Wiggins, C. H. & Gonzalez, R. L. Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation. *Proc. Natl. Acad. Sci.* **106,** 15702–7 (2009).

112. Cornish, P. V., Ermolenko, D. N., Noller, H. F. & Ha, T. Spontaneous intersubunit rotation in single ribosomes. *Mol. Cell* **30,** 578–88 (2008).

113. Cornish, P. V. *et al.* Following movement of the L1 stalk between three functional states in single ribosomes. *Proc. Natl. Acad. Sci.* **106,** 2571–6 (2009).

114. Chen, C. *et al.* Single-molecule fluorescence measurements of ribosomal translocation dynamics. *Mol. Cell* **42,** 367–77 (2011).

115. Kim, H. D., Puglisi, J. D. & Chu, S. Fluctuations of transfer RNAs between classical and hybrid states. *Biophys. J.* **93,** 3575–82 (2007).

116. Munro, J. B., Altman, R. B., O'Connor, N. & Blanchard, S. C. Identification of two distinct hybrid state intermediates on the ribosome. *Mol. Cell* **25,** 505–17 (2007).

117. Wang, B., Ho, J., Fei, J., Gonzalez, R. L. & Lin, Q. A microfluidic approach for investigating the temperature dependence of biomolecular activity with single-molecule resolution. *Lab Chip* **11,** 274–81 (2011).

118. Ning, W., Fei, J. & Gonzalez, R. L. The ribosome uses cooperative conformational changes to maximize and regulate the efficiency of translation. *Proc. Natl. Acad. Sci.* **111,** 12073–8 (2014).

119. Fei, J., Richard, A. C., Bronson, J. E. & Gonzalez, R. L. Transfer RNA-mediated regulation of ribosome dynamics during protein synthesis. *Nat. Struct. Mol. Biol.* **18,** 1043–51 (2011).

120. Wang, L. *et al.* Allosteric control of the ribosome by small-molecule antibiotics. *Nat. Struct. Mol. Biol.* **19,** 957–63 (2012).

121. Valle, M., Zavialov, A., Sengupta, J., Rawat, U., Ehrenberg, M. & Frank, J. Locking and unlocking of ribosomal motions. *Cell* **114,** 123–34 (2003).

122. Frank, J., Gao, H., Sengupta, J., Gao, N. & Taylor, D. J. The process of mRNA-tRNA translocation. *Proc. Natl. Acad. Sci.* **104,** 19671–8 (2007).

123. Agirrezabala, X., Lei, J., Brunelle, J. L., Ortiz-Meoz, R. F., Green, R. & Frank, J. Visualization of the hybrid state of tRNA binding promoted by spontaneous ratcheting of the ribosome. *Mol. Cell* **32,** 190–7 (2008).

124. Trabuco, L. G. *et al.* The role of L1 stalk-tRNA interaction in the ribosome elongation cycle. *J. Mol. Biol.* **402,** 741–60 (2010).

125. Agirrezabala, X. *et al.* Structural characterization of mRNA-tRNA translocation intermediates. *Proc. Natl. Acad. Sci.* **109,** 6094–9 (2012).

126. Sarabhai, A. S., Stretton, A. O., Brenner, S. & Bolle, A. Co-linearity of the gene with the polypeptide chain. *Nature* **201,** 13–17 (1964).

127. Brenner, S., Stretton, A. O. W. & Kaplan, S. Genetic Code: The Nonsense Triplets for Chain Termination and their Suppression. *Nature* **206,** 994–998 (1965).

128. Weigert, M. G. & Garen, A. Base Composition of Nonsense Condons in E. coli: Evidence from Amino-Acid Substitutions at a Tryptophan Site in Alkaline Phosphatase. *Nature* **206,** 992–994 (1965).

129. Sambrook, J. F., Fan, D. P. & Brenner, S. A Strong Suppressor Specific for UGA. *Nature* **214,** 452–453 (1967).

130. Scolnick, E., Tompkins, R., Caskey, T. & Nirenberg, M. Release factors differing in specificity for terminator codons. *Proc. Natl. Acad. Sci.* **61,** 768–74 (1968).

131. Petry, S. *et al.* Crystal structures of the ribosome in complex with release factors RF1 and RF2 bound to a cognate stop codon. *Cell* **123,** 1255–1266 (2005).

132. Laurberg, M., Asahara, H., Korostelev, A., Zhu, J., Trakhanov, S. & Noller, H. F. Structural basis for translation termination on the 70S ribosome. *Nature* **454,** 852–7 (2008).

133. Korostelev, A. *et al.* Crystal structure of a translation termination complex formed with release factor RF2. *Proc. Natl. Acad. Sci.* **105,** 19684–19689 (2008).

134. Weixlbaumer, A. *et al.* Insights into Translational Termination from the Structure of RF2 Bound to the Ribosome. *Science* **322,** 953–956 (2008).

135. Korostelev, A., Zhu, J., Asahara, H. & Noller, H. F. Recognition of the amber UAG stop codon by release factor RF1. *EMBO J.* **29,** 2577–85 (2010).

136. Youngman, E. M., McDonald, M. E. & Green, R. Peptide release on the ribosome: mechanism and implications for translational control. *Annu. Rev. Microbiol.* **62,** 353–73 (2008).

137. Indrisiunaite, G., Pavlov, M. Y., Heurgué-Hamard, V. & Ehrenberg, M. On the pH Dependence of Class-1 RF-Dependent Termination of mRNA Translation. *J. Mol. Biol.* **427,** 1848–1860 (2015).

138. Freistroffer, D. V., Pavlov, M. Y., MacDougall, J., Buckingham, R. H. & Ehrenberg, M. Release factor RF3 in E.coli accelerates the dissociation of release factors RF1 and RF2 from the ribosome in a GTP-dependent manner. *EMBO J.* **16,** 4126–33 (1997).

139. Peske, F., Kuhlenkoetter, S., Rodnina, M. V. & Wintermeyer, W. Timing of GTP binding and hydrolysis by translation termination factor RF3. *Nucleic Acids Res.* **42,** 1812–1820 (2014).

140. Koutmou, K. S., McDonald, M. E., Brunelle, J. L. & Green, R. RF3:GTP promotes rapid dissociation of the class 1 termination factor. *RNA* **20,** 609–620 (2014).

141. Hirokawa, G., Nijman, R. M., Raj, V. S., Kaji, H., Igarashi, K. & Kaji, A. The role of ribosome recycling factor in dissociation of 70S ribosomes into subunits. *RNA* **11,** 1317–1328 (2005).

142. Zavialov, A. V., Hauryliuk, V. V. & Ehrenberg, M. Splitting of the posttermination ribosome into subunits by the concerted action of RRF and EF-G. *Mol. Cell* **18,** 675–686 (2005).

143. Peske, F., Rodnina, M. V. & Wintermeyer, W. Sequence of steps in ribosome recycling as defined by kinetic analysis. *Mol. Cell* **18,** 403–412 (2005).

144. Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & Krogh, A. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* **17,** 425–428 (2001).

145. Freistroffer, D. V., Kwiatkowski, M., Buckingham, R. H. & Ehrenberg, M. The accuracy of codon recognition by polypeptide release factors. *Proc. Natl. Acad. Sci.* **97,** 2046–51 (2000).

146. Sund, J., Andér, M. & Aqvist, J. Principles of stop-codon reading on the ribosome. *Nature* **465,** 947–950 (2010).

147. Pestka, S. Inhibitors of ribosome functions. *Annu. Rev. Microbiol.* **25,** 487–562 (1971).

148. Blanchard, S. C., Cooperman, B. S. & Wilson, D. N. Probing translation with small-molecule inhibitors. *Chem. Biol.* **17,** 633–45 (2010).

149. Brown, C. M., McCaughan, K. K. & Tate, W. P. Two regions of the Escherichia coli 16S ribosomal RNA are important for decoding stop signals in polypeptide chain termination. *Nucleic Acids Res.* **21,** 2109–2115 (1993).

150. Youngman, E. M., He, S. L., Nikstad, L. J. & Green, R. Stop Codon Recognition by Release Factors Induces Structural Rearrangement of the Ribosomal Decoding Center that Is Productive for Peptide Release. *Mol. Cell* **28,** 533–543 (2007).

151. Ogle, J. M., Brodersen, D. E., Clemons, W. M., Tarry, M. J., Carter, A. P. & Ramakrishnan, V. Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science* **292,** 897–902 (2001).

152. Lynch, S. R., Gonzalez, R. L. & Puglisi, J. D. Comparison of X-ray crystal structure of the 30S subunit-antibiotic complex with NMR structure of decoding site oligonucleotide-paromomycin complex. *Structure* **11,** 43–53 (2003).

153. Tate, W., Poole, E., Dalphin, M., Major, L., Crawford, D. & Mannering, S. The translational stop signal: Codon with a context, or extended factor recognition element? *Biochimie* **78,** 945–952 (1996).

154. Pavlov, M. Y., Freistroffer, D. V., Dincbas, V., MacDougall, J., Buckingham, R. H. & Ehrenberg, M. A direct estimation of the context effect on the efficiency of termination. *J. Mol. Biol.* **284,** 579–90 (1998).

155. Ito, K., Uno, M. & Nakamura, Y. Single amino acid substitution in prokaryote polypeptide release factor 2 permits it to terminate translation at all three stop codons. *Proc. Natl. Acad. Sci.* **95,** 8165–9 (1998).

156. Uno, M., Ito, K. & Nakamura, Y. Polypeptide release at sense and noncognate stop codons by localized charge-exchange alterations in translational release factors. *Proc. Natl. Acad. Sci.* **99,** 1819–1824 (2002).

157. Ito, K., Uno, M. & Nakamura, Y. A tripeptide 'anticodon' deciphers stop codons in messenger RNA. *Nature* **403,** 680–684 (2000).

158. Rawat, U., Gao, H., Zavialov, A., Gursky, R., Ehrenberg, M. & Frank, J. Interactions of the release factor RF1 with the ribosome as revealed by Cryo-EM. *J. Mol. Biol.* **357,** 1144–1153 (2006).

159. Rawat, U. B. S. *et al.* A cryo-electron microscopic study of ribosome-bound termination factor RF2. *Nature* **421,** 87–90 (2003).

160. Klaholz, B. P. *et al.* Structure of the Escherichia coli ribosomal termination complex with release factor 2. *Nature* **421,** 90–94 (2003).

161. Sternberg, S. H., Fei, J., Prywes, N., McGrath, K. a. & Gonzalez, R. L. Translation factors direct intrinsic ribosome dynamics during translation termination and ribosome recycling. *Nat. Struct. Mol. Biol.* **16,** 861–868 (2009).

162. Gao, H. *et al.* RF3 induces ribosomal conformational changes responsible for dissociation of class I release factors. *Cell* **129,** 929–41 (2007).

163. Dahlgren, A. & Rydén-Aulin, M. Effects of two cis-acting mutations on the regulation and expression of release factor one in Escherichia coli. *Biochimie* **86,** 431–438 (2004).

164. Adamski, F. M., McCaughan, K. K., Jørgensen, F., Kurland, C. G. & Tate, W. P. The Concentration of Polypeptide Chain Release Factors 1 and 2 at Different Growth Rates of Escherichia coli. *J. Mol. Biol.* **238,** 302–308 (1994).

165. Johnson, D. B. F. *et al.* Release factor one is nonessential in Escherichia coli. *ACS Chem. Biol.* **7,** 1337–44 (2012).

166. Arifuzzaman, M. Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res.* **16,** 686–691 (2006).

167. Calloni, G. *et al.* DnaK Functions as a Central Hub in the E. coli Chaperone Network. *Cell Rep.* **1,** 251–264 (2012).

168. Dinçbas-Renqvist, V., Engström, Å., Mora, L., Heurgué-Hamard, V., Buckingham, R. & Ehrenberg, M. A post-translational modification in the GGQ motif of RF2 from Escherichia coli stimulates termination of translation. *EMBO J.* **19,** 6900–6907 (2000).

169. Heurgué-Hamard, V., Champ, S., Engström, Å., Ehrenberg, M. & Buckingham, R. H. The hemK gene in Escherichia coli encodes the N5-glutamine methyltransferase that modifies peptide release factors. *EMBO J.* **21,** 769–778 (2002).

170. Nakahigashi, K. *et al.* HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and hemK knockout induces defects in translational termination. *Proc. Natl. Acad. Sci.* **99,** 1473–1478 (2002).

171. Mora, L., Heurgué-Hamard, V., De Zamaroczy, M., Kervestin, S. & Buckingham, R. H. Methylation of bacterial release factors RF1 and RF2 is required for normal translation termination in vivo. *J. Biol. Chem.* **282,** 35638–35645 (2007).

172. Bouakaz, L., Bouakaz, E., Murgola, E. J., Ehrenberg, M. & Sanyal, S. The role of ribosomal protein L11 in class I release factor-mediated translation termination and translational accuracy. *J. Biol. Chem.* **281,** 4548–56 (2006).

173. Van Dyke, N., Xu, W. & Murgola, E. J. Limitation of ribosomal protein L11 availability in vivo affects translation termination. *J. Mol. Biol.* **319,** 329–339 (2002).

174. Pallesen, J. *et al.* Cryo-EM visualization of the ribosome in termination complex with apo-RF3 and RF1. *Elife* **2,** e00411 (2013).

175. Mora, L., Zavialov, A., Ehrenberg, M. & Buckingham, R. H. Stop codon recognition and interactions with peptide release factor RF3 of truncated and chimeric RF1 and RF2 from Escherichia coli. *Mol. Microbiol.* **50,** 1467–1476 (2003).

176. Zhou, J., Korostelev, A., Lancaster, L. & Noller, H. F. Crystal structures of 70S ribosomes bound to release factors RF1, RF2 and RF3. *Curr. Opin. Struct. Biol.* **22,** 733–42 (2012).

177. He, S. L. & Green, R. Visualization of codon-dependent conformational rearrangements during translation termination. *Nat. Struct. Mol. Biol.* **17,** 465–470 (2010).

178. Wilson, K. S., Ito, K., Noller, H. F. & Nakamura, Y. Functional sites of interaction between release factor RF1 and the ribosome. *Nat. Struct. Biol.* **7,** 866–870 (2000).

179. Frolova, L. Y. *et al.* Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis. *RNA* **5,** 1014–20 (1999).

180. Zavialov, A. V., Mora, L., Buckingham, R. H. & Ehrenberg, M. Release of peptide promoted by the GGQ motif of class 1 release factors regulates the GTPase activity of RF3. *Mol. Cell* **10,** 789–798 (2002).

181. Shaw, J. J. & Green, R. Two Distinct Components of Release Factor Function Uncovered by Nucleophile Partitioning Analysis. *Mol. Cell* **28,** 458–467 (2007).

182. Trobro, S. & Åqvist, J. Mechanism of the translation termination reaction on the ribosome. *Biochemistry* **48,** 11296–11303 (2009).

183. Kuhlenkoetter, S., Wintermeyer, W. & Rodnina, M. V. Different substrate-dependent transition states in the active site of the ribosome. *Nature* **476,** 351–354 (2011).

184. Shin, D. H., Brandsen, J., Jancarik, J., Yokota, H., Kim, R. & Kim, S. H. Structural analyses of peptide release factor 1 from Thermotoga maritima reveal domain flexibility required for its interaction with the ribosome. *J. Mol. Biol.* **341,** 227–239 (2004).

185. Graille, M. *et al.* Molecular basis for bacterial class I release factor methylation by PrmC. *Mol. Cell* **20,** 917–927 (2005).

186. Diago-Navarro, E., Mora, L., Buckingham, R. H., Díaz-Orejas, R. & Lemonnier, M. Novel Escherichia coli RF1 mutants with decreased translation termination activity and increased sensitivity to the cytotoxic effect of the bacterial toxins Kid and RelE. *Mol. Microbiol.* **71,** 66–78 (2009).

187. Vestergaard, B. *et al.* The SAXS solution structure of RF1 differs from its crystal structure and is similar to its ribosome bound cryo-EM structure. *Mol. Cell* **20,** 929–38 (2005).

188. Zoldák, G. *et al.* Release factors 2 from Escherichia coli and Thermus thermophilus: Structural, spectroscopic and microcalorimetric studies. *Nucleic Acids Res.* **35,** 1343–1353 (2007).

189. Tinoco, I. & Gonzalez, R. L. Biological mechanisms, one molecule at a time. *Genes Dev.* **25,** 1205–1231 (2011).

190. Xie, S. Single-Molecule Approach to Enzymology. *Single Mol.* **2,** 229–236 (2001).

191. Solomatin, S. V., Greenfeld, M., Chu, S. & Herschlag, D. Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature* **463,** 681–4 (2010).

192. English, B. P. *et al.* Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat. Chem. Biol.* **2,** 87–94 (2006).

193. Moerner, W. E. & Fromm, D. P. Methods of single-molecule fluorescence spectroscopy and microscopy. *Rev. Sci. Instrum.* **74,** 3597 (2003).

194. Holzmeister, P., Acuna, G. P., Grohmann, D. & Tinnefeld, P. Breaking the concentration limit of optical single-molecule detection. *Chem. Soc. Rev.* (2013).

195. Kinz-Thompson, C. D. & Gonzalez, R. L. smFRET studies of the 'encounter' complexes and subsequent intermediate states that regulate the selectivity of ligand binding. *FEBS Lett.* **588,** 3526–38 (2014).

196. Gambin, Y., VanDelinder, V., Ferreon, A. C. M., Lemke, E. A., Groisman, A. & Deniz, A. A. Visualizing a one-way protein encounter complex by ultrafast single-molecule mixing. *Nat. Methods* **8,** 239–41 (2011).

197. Sivasankar, S., Zhang, Y., Nelson, W. J. & Chu, S. Characterizing the initial encounter complex in cadherin adhesion. *Structure* **17,** 1075–81 (2009).

198. Ragunathan, K., Liu, C. & Ha, T. RecA filament sliding on DNA facilitates homology search. *Elife* **1,** e00067 (2012).

199. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5,** 507–16 (2008).

200. Fei, J. *et al. A highly purified, fluorescently labeled in vitro translation system for single-molecule studies of protein synthesis.* 1st ed. **10,** 221–59 (Elsevier Inc., 2010).

201. Vázquez, D. *Inhibitors of Protein Biosynthesis* 52–185 (Springer-Verlag, Berlin, 1979).

202. Shcherbakova, I. *et al.* Alternative spliceosome assembly pathways revealed by single-molecule fluorescence microscopy. *Cell Rep.* **5,** 151–65 (2013).

203. Weiss, S. Fluorescence Spectroscopy of Single Biomolecules. *Science* **283,** 1676–1683 (1999).

204. Hirschfeld, T. Optical microscopic observation of single small molecules. *Appl. Opt.* **15,** 2965–6 (1976).

205. Funatsu, T., Harada, Y., Tokunaga, M., Saito, K. & Yanagida, T. Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution. *Nature* **374,** 555–559 (1995).

206. Ha, T., Zhuang, X., Kim, H. D., Orr, J. W., Williamson, J. R. & Chu, S. Ligand-induced conformational changes observed in single RNA molecules. *Proc. Natl. Acad. Sci.* **96,** 9077–9082 (1999).

207. Cario, G. Uber Entstehung wahrer Lichtabsorption und scheinbare Koppelung von Quantensprungen. *Zeitschrift fur Phys.* **10,** 185–199 (1922).

208. Cario, G. & Franck, J. Uber Zerlegung von Wasserstoffmolekulen durch angeregte Quecksilberatome. *Zeitschrift fur Phys.* **11,** 161–166 (1922).

209. Franck, J. Bemerkung uber Anregungs- und Ionisierungsspannung des Heliums. *Zeitschrift fur Phys.* **11,** 155–160 (1922).

210. Perrin, J. Fluorescence et induction moleculaire par resonance. *C. R. Hebd. Seances Acad. Sci.* **184,** 1097–1100. (1927).

211. Förster, T. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.* **437,** 55–75 (1948).

212. Clegg, R. M. The history of FRET : From conception through the labors of birth. *Rev. Fluoresc. 2006,* 1–45 (2006).

213. Stryer, L. & Haugland, R. Energy Transfer: A Spectroscopic Ruler. *Proc. Natl. Acad. Sci.* **58,** 719 (1967).

214. Yang, H. The Orientation Factor in Single-Molecule Forster-Type Resonance Energy Transfer, with Examples for Conformational Transitions in Proteins. *Isr. J. Chem.* **49,** 313–321 (2009).

215. Clegg, R. M. *Forster resonance energy transfer-FRET what is it, why do it, and how it's done* 1st ed. **08,** 1–57 (Elesvier B.V., 2009).

216. Cantor, C. R. & Schimmel, P. R. *Biophysical Chemistry* 433–466 (W.H. Freeman and Co., New York, 1980).

217. Ha, T., Enderle, T., Ogletree, D. F., Chemla, D. S., Selvin, P. R. & Weiss, S. Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci.* **93,** 6264–6268 (1996).

218. Schütz, G., Trabesinger, W. & Schmidt, T. Direct observation of ligand colocalization on individual receptor molecules. *Biophys. J.* **28,** 2223–2226 (1998).

219. Ha, T. *et al.* Initiation and re-initiation of DNA unwinding by the Escherichia coli Rep helicase. *Nature* **419,** 638–41 (2002).

220. Axelrod, D. Cell-substrate contacts illuminated by total internal reflection fluorescence. *J. Cell Biol.* **89,** 141–5 (1981).

221. Axelrod, D. Chapter 9 Total Internal Reflection Fluorescence Microscopy. *Methods Cell Biol.* **30,** 245–270 (1989).

222. Axelrod, D. Chapter 7: Total internal reflection fluorescence microscopy. *Methods Cell Biol.* **89,** 169–221 (2008).

223. Perez, C. E. & Gonzalez, R. L. In vitro and in vivo single-molecule fluorescence imaging of ribosome-catalyzed protein synthesis. *Curr. Opin. Chem. Biol.* **15,** 853–63 (2011).

224. Kinz-Thompson, C. D. *et al.* Robustly Passivated, Gold Nanoaperture Arrays for Single-Molecule Fluorescence Microscopy. *ACS Nano* **7,** 8159–8166 (2013).

225. Michalet, X., Siegmund, O. H. W., Vallerga, J. V., Jelinsky, P., Millaud, J. E. & Weiss, S. Detectors for single-molecule fluorescence imaging and spectroscopy. *J. Mod. Opt.* **54,** 239 (2007).

226. Tan, Y.-w., Hanson, J. A., Chu, J.-w. & Yang, H. in *Protein Dyn. Methods Protoc.* (ed Livesay, D. R.) 51–62 (Humana Press, New York, 2014).

227. Michalet, X. *et al.* Development of new photon-counting detectors for single-molecule fluorescence microscopy. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20120035 (2013).

228. Ingargiola, A. *et al.* 8-spot smFRET analysis using two 8-pixel SPAD arrays. *Proc. SPIE* **8590** (eds Enderlein, J., Gregor, I., Gryczynski, Z. K., Erdmann, R. & Koberling, F.) pages (2013).

229. Cohen, A. E. & Moerner, W. E. Method for trapping and manipulating nanoscale objects in solution. *Appl. Phys. Lett.* **86,** 093109 (2005).

230. Shon, M. J. & Cohen, A. E. Mass action at the single-molecule level. *J. Am. Chem. Soc.* **134,** 14618–23 (2012).

231. Reiner, J. E., Crawford, a. M., Kishore, R. B., Goldner, L. S., Helmerson, K. & Gilson, M. K. Optically trapped aqueous droplets for single molecule studies. *Appl. Phys. Lett.* **89,** 013904 (2006).

232. Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C. & Ha, T. Advances in single-molecule fluorescence methods for molecular biology. *Annu. Rev. Biochem.* **77,** 51–76 (2008).

233. Rasnik, I., McKinney, S. a. & Ha, T. Surfaces and orientations: much to FRET about? *Acc. Chem. Res.* **38,** 542–8 (2005).

234. Boukobza, E., Sonnenfeld, A. & Haran, G. Immobilization in Surface-Tethered Lipid Vesicles as a New Tool for Single Biomolecule Spectroscopy. *J. Phys. Chem. B* **105,** 12165–12170 (2001).

235. Cisse, I., Okumus, B., Joo, C. & Ha, T. Fueling protein DNA interactions inside porous nanocontainers. *Proc. Natl. Acad. Sci.* **104,** 12646–50 (2007).

236. Zheng, Q. *et al.* Ultra-stable organic fluorophores for single-molecule research. *Chem. Soc. Rev.* **43,** 1044–56 (2014).

237. Ha, T. & Tinnefeld, P. Photophysics of fluorescent probes for single-molecule biophysics and super-resolution imaging. *Annu. Rev. Phys. Chem.* **63,** 595–617 (2012).

238. Stennett, E. M. S., Ciuba, M. a. & Levitus, M. Photophysical processes in single molecule organic fluorescent probes. *Chem. Soc. Rev.* **43,** 1057–75 (2014).

239. Cooper, M. *et al.* Cy3B: improving the performance of cyanine dyes. *J. Fluoresc.* **14,** 145–50 (2004).

240. Wysocki, L. M. & Lavis, L. D. Advances in the chemistry of small molecule fluorescent probes. *Curr. Opin. Chem. Biol.* **15,** 752–9 (2011).

241. Kishino, A. & Yanagida, T. Force measurements by micromanipulation of a single actin filament by glass needles. *Nature* **334,** 74–6 (1988).

242. Aitken, C. E., Marshall, R. A. & Puglisi, J. D. An oxygen scavenging system for improvement of dye stability in single-molecule fluorescence experiments. *Biophys. J.* **94,** 1826–35 (2008).

243. Rasnik, I., McKinney, S. A. & Ha, T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat. Methods* **3,** 891–3 (2006).

244. Dave, R., Terry, D. S., Munro, J. B. & Blanchard, S. C. Mitigating unwanted photophysical processes for improved single-molecule fluorescence imaging. *Biophys. J.* **96,** 2371–81 (2009).

245. Altman, R. B. *et al.* Cyanine fluorophore derivatives with enhanced photostability. *Nat. Methods* **9,** 68–71 (2012).

246. Altman, R. B., Zheng, Q., Zhou, Z., Terry, D. S., Warren, J. D. & Blanchard, S. C. Enhanced photostability of cyanine fluorophores across the visible spectrum. *Nat. Methods* **9,** 428–9 (2012).

247. Loveland, A. B., Habuchi, S., Walter, J. C. & van Oijen, A. M. A general approach to break the concentration barrier in single-molecule imaging. *Nat. Methods* **9,** 987–92 (2012).

248. Schwartz, J. J. & Quake, S. R. Single molecule measurement of the "speed limit" of DNA polymerase. *Proc. Natl. Acad. Sci.* **106,** 20294–9 (2009).

249. Chen, J., Tsai, A., Petrov, A. & Puglisi, J. D. Nonfluorescent quenchers to correlate single-molecule conformational and compositional dynamics. *J. Am. Chem. Soc.* **134,** 5734–7 (2012).

250. Chen, J., Petrov, A., Tsai, A., O'Leary, S. E. & Puglisi, J. D. Coordinated conformational and compositional dynamics drive ribosome translocation. *Nat. Struct. Mol. Biol.* **20,** 718–27 (2013).

251. Novotny, L. & van Hulst, N. Antennas for light. *Nat. Photonics* **5,** 83–90 (2011).

252. Kinkhabwala, A., Yu, Z., Fan, S., Avlasevich, Y., Müllen, K. & Moerner, W. E. Large single-molecule fluorescence enhancements produced by a bowtie nanoantenna. *Nat. Photonics* **3,** 654–657 (2009).

253. Acuna, G. P., Möller, F. M., Holzmeister, P., Beater, S., Lalkens, B. & Tinnefeld, P. Fluorescence enhancement at docking sites of DNA-directed self-assembled nanoantennas. *Science* **338,** 506–10 (2012).

254. Yuan, H., Khatua, S., Zijlstra, P., Yorulmaz, M. & Orrit, M. Thousand-fold enhancement of single-molecule fluorescence near a single gold nanorod. *Angew. Chem. Int. Ed. Engl.* **52,** 1217–21 (2013).

255. Estrada, L. C., Aramendía, P. F. & Martínez, O. E. 10000 Times Volume Reduction for Fluorescence Correlation Spectroscopy Using Nano-Antennas. *Opt. Express* **16,** 20597–602 (2008).

256. Punj, D. *et al.* A plasmonic 'antenna-in-box' platform for enhanced single-molecule analysis at micromolar concentrations. *Nat. Nanotechnol.* **8,** 512–6 (2013).

257. Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G. & Webb, W. W. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).

258. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

259. Tsai, A., Petrov, A., Marshall, R. A., Korlach, J., Uemura, S. & Puglisi, J. D. Heterogeneous pathways and timing of factor departure during translation initiation. *Nature* **487,** 390–3 (2012).

260. Tsai, A. *et al.* The Impact of Aminoglycosides on the Dynamics of Translation Elongation. *Cell Rep.* **3,** 497–508 (2013).

261. Sameshima, T. *et al.* Single-molecule study on the decay process of the football-shaped GroEL-GroES complex using zero-mode waveguides. *J. Biol. Chem.* **285,** 23159–23164 (2010).

262. Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci.* **105,** 1176–1181 (2008).

263. Wenger, J. *et al.* Emission and excitation contributions to enhanced single molecule fluorescence by gold nanometric apertures. *Opt. Express* **16,** 3008 (2008).

264. Gérard, D. *et al.* Nanoaperture-enhanced fluorescence: Towards higher detection rates with plasmonic metals. *Phys. Rev. B* **77,** 045413 (2008).

265. De Torres, J., Ghenuche, P., Moparthi, S. B., Grigoriev, V. & Wenger, J. FRET enhancement in aluminum zero-mode waveguides. *ChemPhysChem* **16,** 782–788 (2015).

266. McQuarrie, D. A. Stochastic Approach to Chemical Kinetics. *J. Appl. Probab.* **4,** 413–478 (1967).

267. Bartholomay, A. F. Enzymatic Reaction-Rate Theory: A Stochastic Approach. *Ann. N. Y. Acad. Sci.* **96,** 897–912 (1962).

268. Schnitzer, M. J. & Block, S. M. Statistical kinetics of processive enzymes. *Cold Spring Harb. Symp. Quant. Biol.* **60,** 793–802 (1995).

269. Andrec, M., Levy, R. M. & Talaga, D. S. Direct Determination of Kinetic Rates from Single-Molecule Photon Arrival Trajectories Using Hidden Markov Models. *J. Phys. Chem. A* **107,** 7454–7464 (2003).

270. Yang, H. in *Single-Molecule Biophys. Exp. Theory, Vol. 146* (eds Komatsuzaki, T., Kawakami, M., Takahashi, S., Yang, H. & Silbey, R. J.) 219–243 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2011).

271. Gopich, I. V. & Szabo, A. in *Single-Molecule Biophys. Exp. Theory, Vol. 146* (eds Komatsuzaki, T., Kawakami, M., Takahashi, S., Yang, H. & Silbey, R. J.) 245–297 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2011).

272. Gopich, I. V. & Szabo, A. Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *Proc. Natl. Acad. Sci.* **109,** 7747–52 (2012).

273. Chung, S. H., Moore, J. B., Xia, L. G., Premkumar, L. S. & Gage, P. W. Characterization of single channel currents using digital signal processing techniques based on Hidden Markov Models. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **329,** 265–85 (1990).

274. Qin, F., Auerbach, A. & Sachs, F. Maximum likelihood estimation of aggregated Markov processes. *Proc. Biol. Sci.* **264,** 375–83 (1997).

275. McKinney, S. a., Joo, C. & Ha, T. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* **91,** 1941–51 (2006).

276. Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez, R. L. & Wiggins, C. H. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* **97,** 3196–205 (2009).

277. Bronson, J. E., Hofman, J. M., Fei, J., Gonzalez, R. L. & Wiggins, C. H. Graphical models for inferring single molecule dynamics. *BMC Bioinformatics* **11,** S2 (2010).

278. Greenfeld, M., Pavlichin, D. S., Mabuchi, H. & Herschlag, D. Single Molecule Analysis Research Tool (SMART): an integrated approach for analyzing single molecule data. *PLoS One* **7,** e30024 (2012).

279. Van De Meent, J.-W., Bronson, J. E., Wood, F., Gonzalez Jr., R. L. & Wiggins, C. H. Hierarchically-coupled hidden Markov models for learning kinetic rates from single-molecule data. *Proc. 30th Int. Conf. Mach. Learn.* (2013).

280. Van de Meent, J.-W., Bronson, J. E., Wiggins, C. H. & Gonzalez, R. L. Empirical Bayes Methods Enable Advanced Population-Level Analyses of Single-Molecule FRET Experiments. *Biophys. J.* **106,** 1327–1337 (2014).

281. Bishop, C. M. *Pattern Recognition and Machine Learning* 461–652 (Springer, New York, 2006).

282. Taylor, J. N., Makarov, D. E. & Landes, C. F. Denoising single-molecule FRET trajectories with wavelets and Bayesian inference. *Biophys. J.* **98,** 164–73 (2010).

283. Devore, M. S., Gull, S. F. & Johnson, C. K. Reconstruction of Calmodulin Single-Molecule FRET States, Dye-Interactions, and CaMKII Peptide Binding by MultiNest and Classic Maximum Entropy. *Chem. Phys.* **422,** 238–245 (2013).

284. Okamoto, K. & Sako, Y. Variational Bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories. *Biophys. J.* **103,** 1315–24 (2012).

Part I

Methods Developments

Chapter 2

Defeating the Concentration Barrier*

2.1 Introduction

Nature often exploits weak intermolecular interactions to permit the reversible assembly of macromolecular complexes, achieve high binding specificities, and facilitate functionally important biomolecular dynamics [1]. Although, in principle, single-molecule fluorescence (smF) microscopies provide powerful tools for dissecting the mechanisms of the fundamentally important biological processes that involve weakly interacting target and ligand biomolecules [2], in practice, these investigations remain challenging. This is because the observation of these weak intermolecular interactions on a timescale that is experimentally accessible to conventional smF microscopies requires prohibitively high background concentrations of ligand biomolecules in solution – a condition that results in high background noise and, consequently, compromises the quality of smF data [3].

Many strategies have been devised to reduce the noise arising from high background concentrations of ligand biomolecules in smF microscopy experiments. Notably, strategies that specifically activate fluorophores of interest (*e.g.*, fluorescence resonance energy transfer (FRET) [4], photoactivatable fluorophores [5, 6], and photogenic fluorophores [7]) or confine the excitation field (*e.g.*, total internal reflection fluorescence (TIRF) microscopy [8], and confocal microscopy [9]) have been used to selectively excite only those ligand biomolecules that are bound to a target biomolecule. Perhaps the most effective strategy described to date is the use of metal-based, sub-wavelength nanoapertures, often called zero-mode waveguides (ZMWs), to generate a confined excitation field near the silica bottom of a nanoaperture structure, thereby selectively exciting only those ligand biomolecules that are bound to a target biomolecule that is localized within the

* Adapted with permission from Kinz-Thompson, C.D., Palma, M., Pulukkunat, D.K., Chenet, D., Hone, J., Wind, S.J., Gonzalez, Jr., R.L. Robustly Passivated, Gold Nanoaperture Arrays for Single-Molecule Fluorescence Microscopy *ACS Nano* **2013** 7(9), 8158-8166. Copyright 2013 American Chemical Society.

confined excitation volume [3]. While nanoaperture fluorescence microscopy is currently the only optical confinement-based microscopy that permits smF measurements to be made at physiologically-relevant, μM background concentrations of ligand biomolecules (Fig. 2.1), the non-specific adsorption of these ligand biomolecules to the metallic and silica nanoaperture surfaces often compromises the signal-to-background ratio (SBR) of the smF data. In order to overcome this problem, an orthogonal surface chemistry process was developed to selectively passivate the metallic cladding and silica bottoms of nanoapertures. Unfortunately, this process is quite limited, as it uses negatively charged poly(vinyl) phosphonic acid (PVPA) to minimize the non-specific adsorption of negatively charged, fluorophore-labeled deoxynucleotides for single-molecule DNA sequencing applications [10], and it is not generally applicable to other biomolecular systems. The lack of significant progress in the development of more generalizable surface passivation chemistries has thus far restricted the use of nanoaperture fluorescence microscopy to only a handful of biological systems [11–13] over the decade since nanoaperture fluorescence microscopy of biological systems was first introduced [3].

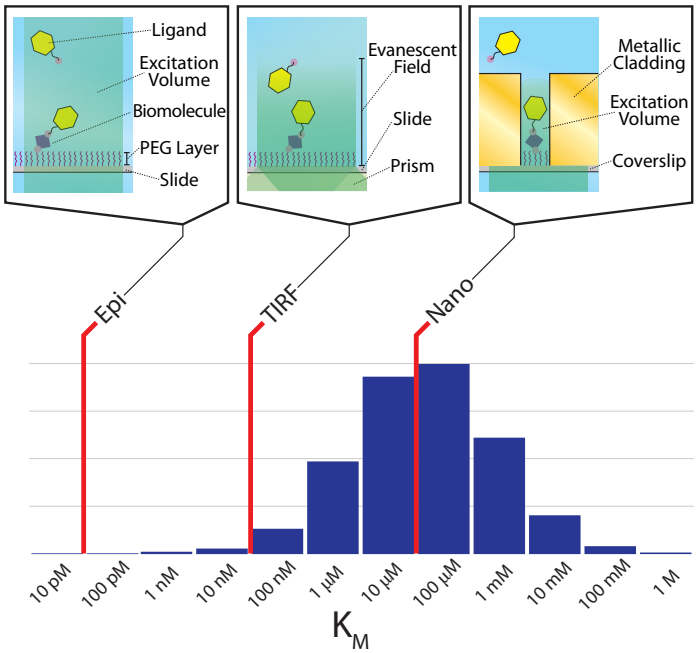

Figure 2.1: Diagram of concentration ranges accessible by various microscopy techniques. Red lines represent the upper limit of the background concentration of ligand biomolecules that can be employed in smF experiments using epi-fluorescence microscopy (Epi), TIRF microscopy (TIRF), and nanoaperture fluorescence microscopy (Nano). The microscope schematics connected to each red line provide molecular-level diagrams corresponding to each technique (upper panel). The histogram shows the distribution of Michaelis constants (K_M), a characterization of the interactions between enzymes and their corresponding substrates, of all eukaryotic enzymes in the BRENDA enzyme database [14]. This distribution is analogous to the distribution of background concentrations of ligand biomolecules required to observe interactions with a target biomolecule on an experimentally accessible timescale using smF microscopies (lower panel).

Here, we report a cost-effective and widely accessible method for the fabrication and selective functional-

ization of microfluidic devices containing gold nanoaperture arrays for smF microscopy investigations. A key advantage of our approach is the robust passivation of the nanoaperture surfaces against the non-specific adsorption of biomolecules. For the gold cladding, this is provided by the formation of self-assembled monolayers (SAMs) on the gold surface using methoxy-terminated, thiol-derivatized PEG. SAMs have traditionally been employed to mitigate non-specific adsorption, and expansively ordered, thiolate SAMs readily form on gold surfaces [15, 16]. Notably, SAMs with terminal PEG blocks have been shown to provide particularly non-specific adsorption-resistant, biocompatible backgrounds [16–18], as PEG chains are extensively hydrated, and the non-specific adsorption of biomolecules induces enthalpically unfavorable desolvation and entropic penalties from compression of a monolayer [19]. By combining this approach for passivating the gold cladding with a fully orthogonal approach for functionalizing the silica bottom of the nanoapertures using a binary mixture of methoxy- and biotin-terminated, silane-derivatized PEG, we have developed a scheme that enables specific tethering of biotinylated target biomolecules to the silica bottom of the nanoapertures while robustly blocking the non-specific adsorption of ligand biomolecules to both the gold cladding and the silica bottom of the nanoapertures. As a proof-of-principle, we demonstrate that the time-resolved fluorescence signals from an individual biotinylated, surface-tethered, FRET donor- and acceptor-fluorophore labeled protein, bacterial peptide chain release factor 1 (RF1)–a protein that has never been shown to be compatible with PVPA-passivated aluminum-based nanoapertures–can be easily monitored in the presence of more than 1 μM background concentrations of FRET acceptor-labeled RF1 [20] without detectable deterioration of the SBR from excitation cross-talk.

2.2 Experimental Methods

Nanoaperture Array Fabrication. Borosilicate coverslips (No. 1.5, VWR) were degreased in piranha solution (3:1 H_2SO_4:30% H_2O_2) for 15 min, rinsed with Milli-Q ultrapure water, sonicated for 15 min in ethanol, and then sonicated for 15 min in Milli-Q ultrapure water. Degreased coverslips were exposed to an O_2 plasma for 2 minutes, and, immediately thereafter, a thin layer of Ma-N 2403 (Micro Resist Technology GmbH), was deposited with a spin-coater. Coverslip substrates were then prebaked at 90 °C for 1 minute. Next, 2.5% poly(2,3-dihydrothieno-1,4-dioxin)-poly(styrenesulfonate) (PEDOT:PSS) (Sigma-Aldrich) was filtered through a 0.2 μm Acrodisc syringe filter (VWR), deposited with a spin-coater as a conductive layer, and then prebaked at 90 °C for 5 minutes. Arrays of circles were patterned on the substrate using an FEI Sirion SEM with a 30 kV electron beam adapted with the Nanometer Pattern Generation System (JC Nabity Lithography

Systems). Patterns were developed by immersion in 8.74 M acetic acid for 5 minutes, followed by mild agitation in Milli-Q ultrapure water and mild agitation in ethanol. The emergent, $\sim$500 nm cylindrical columns were metalized with $\sim$100 nm of gold (Kurt J. Lesker, Co.) with a 50 Å thick, optically transparent, titanium (Kurt J. Lesker, Co.) adhesion layer using electron beam deposition with an Angstrom EvoVac Deposition System. Liftoff was performed by sonication for 2 minutes in 1M KOH, yielding nanoapertures in the relief of the columns. Nanoapertures were characterized with an Agilent 8500 FE-SEM and a Digital Instruments AFM, and analyzed with ImageJ [21].

Nanoaperture Functionalization. Nanofabricated substrates were cleaned with 1.5 hour-aged piranha solution (3:1 H_2SO_4:30% H_2O_2) and an O_2 plasma. Substrates were then incubated for 12 hours in 1 mM mPEG-SH (MW = 350 g mol^{-1}) (Nanocs Inc.)in anhydrous ethanol (Sigma), and then rinsed with ethanol and dried with N_2. Substrates were then silanized for 24 hours in 100 µM solutions of biotin-PEG-Si (MW = 2000 g mol^{-1}) and mPEG-Si (MW = 3400 g mol^{-1}) (Laysan Bio Inc.) in anhydrous toluene (Sigma) with 10 mM glacial acetic acid (Sigma), rinsed thoroughly with ethanol, Milli-Q ultrapure water, and isopropyl alcohol, and then dried with N_2.

Recombinant Gene Construction. The *Escherichia coli* (*E. coli*) prfA gene encoding wildtype polypeptide chain termination factor 1 (RF1) was previously cloned into the pPROEX-HTb expression vector (Life Technologies, Inc.) downstream of an N-terminal hexa-histidine affinity purification tag and a Tev protease cleavage site; additionally, all native cysteines were removed and a single cysteine was introduced using at amino-acid position 167 with a serine-to-cysteine mutation ($RF1_{S167C}$) using site-directed mutagenesis [20]. Starting with this $RF1_{S167C}$ construct, an N-terminal enzymatic biotinylation tag (GLNDIFEAQKIEWHE) was subcloned downstream of the Tev protease cleavage site, and site-directed mutagenesis was used to introduce a glutamic acid-to-cysteine mutation at amino-acid position 256 (biotag-$RF1_{S167C,E256C}$).

Protein Expression. RF1 proteins were overexpressed in *E. coli* cells, purified using affinity chromatography, Tev protease treated, and purified away from the cleaved hexa-histidine tags and Tev protease using previously published protocols [20]. Briefly, electro-competent BL21(DE3) *E. coli* cells cotransfected with pPROEX-HTb expression vectors carrying either $RF1_{S167C}$ or biotag-$RF1_{S167C,E256C}$, and pET-26b(+) expression vectors carrying the gene prmC–a methyltransferase that modifies RF1–were grown under ampicillin and kanamycin selection, and RF1 protein overexpression was induced by addition of 1 mM isopropyl β-D-1-thiogalactopyranoside. Cells were lysed with a French Press, and RF1 proteins were purified using

affinity chromatography with a Ni-NTA agarose bead column. Purified RF1 proteins were subsequently incubated with hexa-histidine-tagged Tev protease overnight, and the proteolyzed RF1 proteins were purified from the cleaved hexa-histidine tags and the hexa-histidine-tagged Tev protease using a second passage through a Ni-NTA agarose bead column. biotag-$RF1_{S167C,E256C}$ was biotinylated (bio-$RF1_{S167C,E256C}$) using recombinant *E. coli* BirA biotin-ligase (plasmid obtained from AddGene) that was overexpressed, purified, and used following a previously published protocol [22].

Protein Labeling. Cysteine 167 in $RF1_{S167C}$ was reduced by incubation with 1 mM tris (2-carboxyethyl) phosphine hydrochloride at room temperature and labeled by reaction with 15x molar excess of maleimide-derivatized Cy5 (G.E. Lifesciences) using a previously published protocol [20]. bio-$RF1_{S167C,E256C}$ was reduced and labeled following a similar protocol, but using equivalent concentrations of both maleimide-derivatized Cy3 and Cy5. Labeled RF1 proteins were purified from unreacted dyes using size-exclusion column chromatography with a HiLoad 16/600 Superdex 75 pg chromatography column (GE Lifesciences), and were subsequently purified from unlabeled RF1 proteins using hydrophobic interaction column chromatography with a TSKgel Phenyl-5PW column (Tosoh Biosciences) using previously published protocols [20]. The purification yielded pure, 100% Cy5-labeled $RF1_{S167C}$ ($RF1_{Cy5}$) and pure, 100%, 1:1 stoichiometrically Cy3- and Cy5-labeled bio-$RF1_{S167C,E256C}$ (bio-$RF1_{Cy3,Cy5}$).

Fluorescence Microscope. Nanoaperture arrays were illuminated by epi-fluorescence through a Nikon, water-immersion 60x NA=1.2 PlanApo objective using a 50 mW, 532 nm diode-pumped solid-state laser (CrystaLaser) through a 552 nm, single-edge dichroic beamsplitter (Semrock) and a downstream 533 nm (FWHM=17 nm) notch filter (Thorlabs) using a Nikon Ti-U inverted microscope. To increase incident illumination intensity, only a quarter of the field of view was illuminated; this area contained $\sim$1000 to 4500 nanoapertures depending on the spacing employed between the nanoapertures. Fluorescence emissions were collected through the objective, and imaged through a Photometrics DV2 wavelength splitter containing a 630dcxr dichroic beamsplitter, and HQ575/40m and HQ680/50m emission filters (for Cy3 and Cy5, respectively) onto a 512 x 512 pixel Andor iXon3 897E electron-multiplying charge-coupled-device camera. Movies were recorded at a 10 Hz acquisition rate without binning using Metamorph software (Molecular Devices).

Microscopy Buffers. Tris-polymix buffer (50 mM Tris-acetate ($pH_{25\,°C\,=\,7.0}$), 100 mM KCl, 5 mM ammonium acetate, 500 µM calcium acetate, 100 µM EDTA, 10 mM 2-mercaptoethanol, 5 mM putrescine, 1 mM

spermidine, 15 mM magnesium acetate, and 1% (w/v) β-D-glucose). Imaging buffer (Tris-polymix buffer supplemented with 300 mg mL^{--1} glucose oxidase (Sigma-Aldrich), 40 mg mL-1 catalase (Sigma-Aldrich), 1 mM 1,3,5,7-cyclooctatetraene (Sigma-Aldrich), and 1 mM *p*-nitrobenzyl alcohol (Fluka))

Nanoaperture Fluorescence Microscopy. Flow cells were prepared for imaging by incubation with ultrapure bovine serum albumin (Ambion) and a 50-nucleotide oligomer of random-sequence duplex DNA (IDT), and, when specified, a subsequent incubation with 1 µM streptavidin (Invitrogen) using a previously published protocol [20]. bio-$RF1_{Cy3,Cy5}$ was diluted to the specified concentration in Tris-polymix buffer, loaded into the flow cells, incubated for 5 min at room temperature, and untethered bio-$RF1_{Cy3,Cy5}$ was removed by washing the flow cells with 200 µL of Tris-polymix buffer. Immediately prior to imaging, flow cells were washed and filled with Imaging buffer, and, when specified, various concentrations of $RF1_{Cy5}$.

Data Processing. All analysis was performed with Python using NumPy and SciPy [23], Matplotlib [24], and the Python Imaging Library. Individual, Cy3 or Cy5 intensity versus time trajectories were constructed by selecting those pixels on the Cy3 half of the initial frame with intensities exceeding three standard deviations of the mean pixel intensity, clustering neighboring pixels into regions, calculating the center of mass (COM) of each region, and mapping those COM coordinates onto the Cy5 half of the frame when applicable. The intensity of a region in each frame of a movie was then obtained by summing the four pixels neighboring the region COM, linearly scaled by distance to the COM, such that the total pixel area employed in the sum was one pixel. The x- and y-coordinates for each region COM were drift-corrected in each frame of each movie by the drift of the COM of the entire illumination profile in that movie.

2.3 Nanofabrication of Sub-wavelength Gold Nanoapertures

Gold nanoaperture arrays were fabricated on borosilicate coverslips using an electron-beam lithography process in which nanoapertures are formed in the relief of metal-embedded pillars of cross-linked polymer (Fig. 2.2). During the development of the electron-beam lithography steps of this process, we found that the poly(3,4-ethylenedioxythiophene) poly(styrenesulfonate) conductive layer employed (applied to the Ma-N 2403 negative-tone resist in order to prevent charging during electron beam writing) failed to be completely removed by water washing following electron-beam writing. It is possible that this layer, which has a pH of 1.5 - 2.5 at room temperature, crosslinks to the Ma-N 2403 negative-tone resist layer, which is composed of a phenolic resin with a bisazide photoactive compound. Therefore, instead of washing with water, treatment

with acetic acid, which would be expected to reverse the crosslinking equilibrium, was used to produce well-defined pillars. Typically, this process yields ∼80% of the intended nanoapertures, with poor pillar adhesion prior to metallization or incomplete lift-off being the most common causes of defects. Despite these issues, this process yields nanoapertures with an average diameter of 177 nm, with a relatively narrow diameter distribution (Fig. 2.2), and a spacing of 1 - 5 µm . Within these dimensions, gold nanoapertures exhibit both the strong electromagnetic confinement responsible for the reduction in excitation volume as well as the gold surface plasmon-mediated fluorescence enhancement that is characteristic of smaller diameter gold nanoapertures [25, 26].

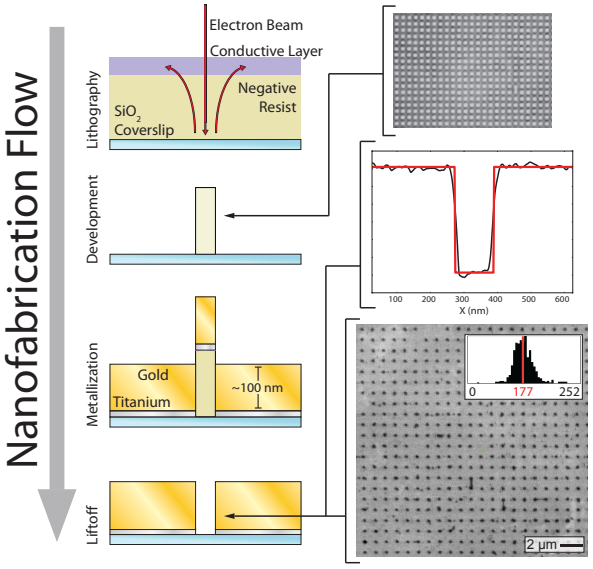

Figure 2.2: A schematic diagram of the nanoaperture fabrication process. Negative-tone electron-beam lithography crosslinks patterns in the negative resist on a silica coverslip; excess electrons are removed to a ground via a conductive layer. Pillars of patterned, cross-linked resist remain following substrate exposure to aqueous developer. The top panel shows a wide-field, optical microscope image of a pre-metallization pillar array. An optically transparent, adhesion layer of titanium is then deposited with electron-beam evaporation. Subsequently, a layer of gold is deposited with electron-beam evaporation, such that the cross-linked resist remains solvent exposed. Nanoapertures are formed in the relief of the pillars following solvent-based liftoff. The middle panel shows an atomic force microscope image cross-section of a typical nanoaperture; the red line is a boxcar function fit with a 115 nm step length. The bottom panel shows a scanning electron microscope image of a nanoaperture array, post-fabrication; the average diameter of the nanoapertures in this array is 177 $\pm$ 16 nm (1σ, n = 499).

2.4 Orthogonal Functionalization and Passivation of Gold Nanoapertures

Following nanofabrication, the gold nanoaperture arrays were passivated with a SAM formed using methoxy-terminated, thiol-derivatized PEG (mPEG-SH) and a SAM formed using a binary mixture of methoxy-terminated,

triethoxy-functionalized, silane-derivatized PEG (mPEG-Si) and biotin-terminated, triethoxy-functionalized, silane-derivatized PEG (biotin-PEG-Si) to create biomolecule adsorption-resistant backgrounds on the gold cladding and the silica nanoaperture bottoms, respectively (Fig. 2.3). As silanes have a propensity to covalently bond to gold [27], thiolation of the gold cladding was performed prior to silanization of the silica nanoaperture bottoms to yield more homogeneously passivated nanoapertures. Following extensive substrate cleaning, the passivation process begins with a 12 hour incubation in a solution prepared by dissolving mPEG-SH in anhydrous ethanol to a final PEG concentration of 1 mM – thereby ensuring sufficient time for SAM formation [16]. The substrates were then silanized for 24 hours in a solution of 100 µM biotin-PEG-Si and mPEG-Si in anhydrous toluene at the specified molar ratio of biotin-PEG-Si to mPEG-Si.

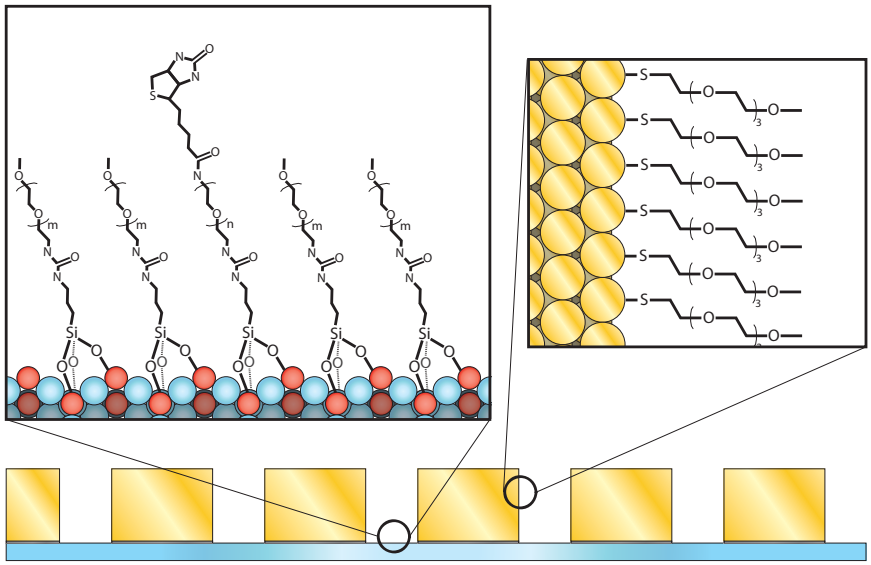

Figure 2.3: Molecular-level schematic diagram of thiol and silane passivated surfaces. The gold surfaces of the nanoaperture arrays were passivated with a SAM formed using mPEG-SH, and the borosilicate surfaces of the nanoaperture arrays were passivated with a SAM formed using a binary mixture of biotin-PEG-Si and mPEG-Si.

2.5 Tunable, Chemical Control Over Nanoaperture Occupation

Microfluidic devices were then assembled by mounting substrates comprising fully passivated, gold nanoaperture arrays onto quartz microscope slides that had been drilled to form sets of inlet/outlet ports, cleaned and passivated with mPEG-Si using a previously published procedure [28], and divided into multiple, separate reaction flow cells using adhesive spacers (Fig. 2.4) [29]. Once the substrates had been mounted onto the adhesive spacers, epoxy was used to seal the sides of the flow cells. This multiple flow cell geometry allows

for several independent experiments and controls to be performed in the same microfluidic device, thereby eliminating device fabrication and processing as a source of experimental variation.

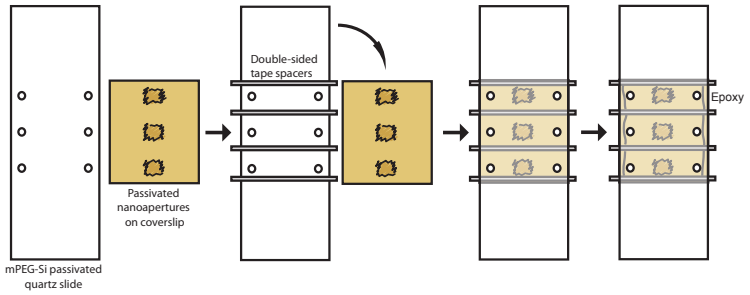

Figure 2.4: Schematic diagram of microfluidic device construction. Thin, adhesive spacers are carefully placed between the inlet/outlet ports of an mPEG-Si passivated, quartz microscope slide. The borosilicate coverslip substrate containing the passivated nanoaperture arrays is placed on the adhesive spacers, and aligned such that the nanoaperture arrays are positioned between the spacers, facing the interior of the soon-to-be flow cells. This alignment is facilitated by placing the slide on an illuminated surface, and then using the diffraction from the nanoaperture arrays as a guide. Once the coverslip is affixed to the slide, the flow cells are sealed with fast-drying epoxy. Superfluous spacer and epoxy can be removed from the sides of the microfluidic device with a razor blade once the epoxy is cured.

2.5.1 Presence of Biotin-streptavidin-biotin Bridge is Obligatory for Nanoaperture Occupation

To assess the robustness of our gold nanoaperture passivation scheme, we performed several nanoaperture fluorescence microscopy experiments using a biotinylated, Cy3 FRET donor- and Cy5 FRET acceptor-labeled, double-cysteine mutant of the *E. coli* RF1 (bio-RF1$_{Cy3,Cy5}$). The ability to specifically localize a biotinylated target biomolecule from solution to the silica bottom of a nanoaperture is dependent upon the formation of a biotin-streptavidin-biotin bridge between the biotin-PEG-Si on the silica bottom of the nanoaperture, streptavidin, and the biotinylated target biomolecule – in this case, bio-RF1$_{Cy3,Cy5}$. After incubation of a flow cell with 1 μM streptavidin for 5 minutes at room temperature, washing with Tris-polymix buffer, incubation with 1 pM bio-RF1$_{Cy3,Cy5}$ for 5 minutes at room temperature, and washing with Imaging buffer, wide-field fluorescence imaging of the flow cell yielded fluorescence emission from bio-RF1$_{Cy3,Cy5}$ in a well-defined pattern corresponding to the nanoaperture array (Fig. 2.5A, top panel). When performing the same experiment in a neighboring flow cell, but in the absence of streptavidin, no fluorescence emission was detected from the nanoaperture array. Even imaging of regions of bulk silica just proximal to the nanoaperture array revealed only minimal fluorescence emission from spatially localized bio-RF1$_{Cy3,Cy5}$, which we attributed to the non-specific adsorption of bio-RF1$_{Cy3,Cy5}$ to defects in the borosilicate surface of the coverslip substrate

(Fig. 2.5A, bottom panel). Thus, the passivating SAMs resisted the non-specific adsorption of bio-RF1$_{Cy3,Cy5}$ to both the gold cladding and the silica bottoms of the gold nanoaperture arrays, while the presence of streptavidin at the bottom of the nanoaperture arrays allowed the specific localization of bio-RF1$_{Cy3,Cy5}$ to the nanoaperture arrays. In addition to demonstrating the robustness of our SAM-based passivation scheme, these results show that the presence of neither the gold cladding nor the mPEG-SH SAMs interfered with passivation of the silica regions of the substrates with the biotin-PEG-Si and mPEG-Si SAMs.

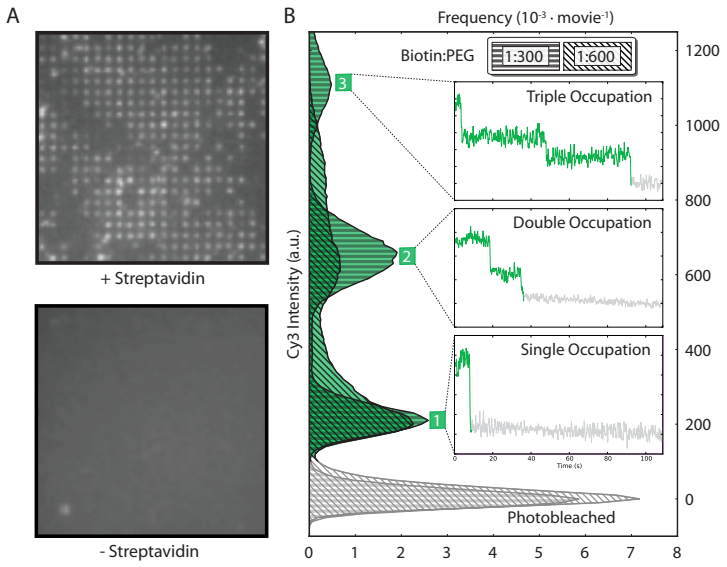

Figure 2.5: (A) Streptavidin-dependence of nanoaperture occupation. After incubating bio-RF1$_{Cy3,Cy5}$ in a flow cell with streptavidin, Cy3 fluorescence was observed in a pattern corresponding to that of the nanoaperture arrays, demonstrating that bio-RF1$_{Cy3,Cy5}$ was specifically localized to the silica bottoms of the nanoapertures (upper panel); the nanofabrication defects described in the text result in the lack of fluorescence in some of the regions where nanoapertures should be located. After incubating bio-RF1$_{Cy3,Cy5}$ in a second flowcell without streptavidin, Cy3 fluorescence was not observed from the nanoaperture array under the same imaging conditions; moreover, only minimal Cy3 fluorescence emission was observed from regions of bulk silica just proximal to the nanoaperture array (lower panel). (B) Tunable nanoaperture occupation. Histograms show the distributions of Cy3 fluorescence intensities observed over 100 seconds from nanoapertures in flow cells that had been passivated with a 1:300 or a 1:600 ratio of biotin-PEG-Si:mPEG-Si and then incubated with both bio-RF1$_{Cy3,Cy5}$ and streptavidin. Insets show discrete photobleaching events in Cy3 fluorescence intensity versus time trajectories that contribute to the histogram peaks and correspond to the occupancy of bio-RF1$_{Cy3,Cy5}$ in individual nanoapertures.

2.5.2 Biotin-PEG:PEG Ratio Modulates Nanopaerture Occupation

We expect that the biotin-PEG-Si that is responsible for localizing streptavidin, and therefore bio-RF1$_{Cy3,Cy5}$, to the silica nanoaperture bottoms via a biotin-streptavidin-biotin bridge should be Poisson-distributed throughout the nanoapertures; thus, whereas individual nanoapertures may contain zero, one, two, three, or more surface-tethered bio-RF1$_{Cy3,Cy5}$ molecules, the exact distribution that is observed is dependent upon the ratio of biotin-PEG-Si to mPEG-Si employed during passivation–the larger the ratio, the greater the proba-

bility of observing multiple surface-tethered bio-RF1$_{Cy3,Cy5}$ molecules per nanoaperture. To demonstrate this tunable control over nanoaperture occupation, we characterized the Cy3 fluorescence emission from single nanoapertures containing surface-tethered bio-RF1$_{Cy3,Cy5}$ in flow cells that have been passivated using our passivation scheme with particular ratios of biotin-PEG-Si to mPEG-Si. Normalized histograms were constructed of the intensity of Cy3 fluorescence emitted from individual nanoapertures in flow cells passivated with 1:600 (n = 173, where n is the number of nanoapertures that were characterized) and 1:300 (n = 195) biotin-PEG-Si:mPEG-Si that were both incubated with 1 μM streptavidin for 5 minutes at room temperature, washed with Tris-polymix buffer, incubated with 100 nM bio-RF1$_{Cy3,Cy5}$ for 5 minutes at room temperature, and washed with Imaging buffer to remove all unbound bio-RF1$_{Cy3,Cy5}$ (Fig. 2.5B). As expected, the distributions of the average Cy3 fluorescence intensity per nanoaperture are resolved into discrete peaks that correspond to discrete numbers of bio-RF1$_{Cy3,Cy5}$ molecules per nanoaperture. By correlating the number of individual Cy3 photobleaching events observed in the Cy3 fluorescence intensity versus time trajectories to the Cy3 fluoresence intensity observed per nanoaperture, we determined the absolute number of bio-RF1$_{Cy3,Cy5}$ molecules that corresponded to each peak in the histogram; the peaks in the histogram of Cy3 intensities correspond to nanoapertures that contain either one, two, or three biotin-streptavidin-biotin-tethered, fluorescing bio-RF1$_{Cy3,Cy5}$ molecules (Fig. 2.5B, insets). As expected, doubling the biotin-PEG-Si:mPEG-Si ratio from 1:600 to 1:300 increases the populations of Cy3 fluorescence intensities corresponding to higher numbers of bio-RF1$_{Cy3,Cy5}$ molecules per nanoaperture. In addition, control over the distribution of bio-RF1$_{Cy3,Cy5}$ molecules per nanoaperture can be achieved by altering the concentrations and incubation times of streptavidin and bio-RF1$_{Cy3,Cy5}$. For the nanoaperture arrays of the dimensions used here, we find that a 1:1000 ratio of biotin-PEG-Si:mPEG-Si with a 5 minute incubation at room temperature in 1 μM streptavidin followed by a 5 minute incubation at room temperature in 10 - 100 pM bio-RF1$_{Cy3,Cy5}$ yields a maximal population of nanoapertures that contain a single bio-RF1$_{Cy3,Cy5}$ molecule. Taken together with the streptavidin dependence shown in Figure 2.5A, our ability to predictably control the distribution of bio-RF1$_{Cy3,Cy5}$ molecules per nanoaperture by altering the ratio of biotin-PEG-Si:mPEG-Si demonstrates that the tethering of bio-RF1$_{Cy3,Cy5}$ molecules to the silica nanoaperture bottoms is specifically dependent on the presence of a biotin-streptavidin-biotin bridge. Moreover, the chemistries employed are general enough to produce similar results in different nanoaperture geometries (*e.g.*, squares) and with other metallic claddings that can form thiol SAMs (*e.g.*, silver).

2.6 Fluorescence Resonance Energy Transfer in Gold Nanoapertures

High background concentrations of ligand biomolecules and/or the non-specific adsorption of ligand biomolecules cause deterioration of the SBR in FRET-based smF experiments used to characterize weak biomolecular interactions–a limitation of FRET-based smF experiments that we expect our passivated, gold-based nanoapertures to overcome. In order to simulate the conditions of such an experiment without the usual complications to the FRET efficiency versus time trajectories that are introduced by the binding and dissociation of FRET acceptor-labeled ligand biomolecules from solution to individual, surface-tethered, FRET donor-labeled target biomolecules, we performed the FRET-based titration experiment diagrammed in Figure 2.6A. Briefly, we used a biotin-streptavidin-biotin bridge to specifically tether bio-$RF1_{Cy3,Cy5}$ to the silica nanoaperture bottoms of a fully passivated (1:1000 biotin-PEG-Si:mPEG-Si) nanoaperture array and imaged a single flow cell at successively higher background solution concentrations of a Cy5 FRET acceptor-labeled variant of RF1 ($RF1_{Cy5}$). To observe FRET signals arising from individual bio-$RF1_{Cy3,Cy5}$ molecules tethered to the silica nanoapertures bottoms, the Cy3 and Cy5 fluorescence intensities from individual, single bio-$RF1_{Cy3,Cy5}$ occupancy nanoapertures (I_{Cy3} and I_{Cy5}, respectively) were converted into FRET efficiency, $E_{FRET} = I_{Cy5}/(I_{Cy3}+I_{Cy5})$, which is a ratiometric measure of acceptor-fluorophore fluorescence. Although we were able to observe FRET signals arising from nanoapertures containing single bio-$RF1_{Cy3,Cy5}$ molecules, the resulting E_{FRET} versus time trajectories were characterized by a very low SBR (Fig. 2.6B). It is possible that this reduction in the SBR of E_{FRET} versus time trajectories is caused by quenching of Cy3 and/or Cy5 that arises from close proximity of these fluorophores to the gold surface, by a red-shift of the fluorescence emission of Cy5 past our filter bandwidth, or that our bio-$RF1_{Cy3,Cy5}$ construct exists mostly in a conformation where Cy3 and Cy5 are too far away from each other so as to undergo appreciable energy transfer–an unlikely explanation, as we have observed energy transfer between Cy3 and Cy5 in TIRF-based FRET experiments using this same bio-$RF1_{Cy3,Cy5}$ construct (Pulukkunat, D.K.; Gonzalez Jr., R.L. unpublished results).

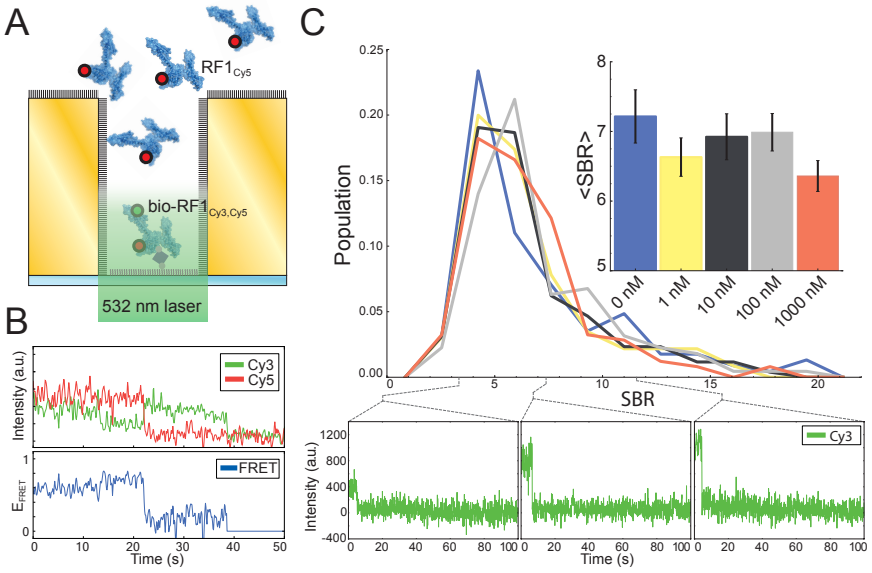

Figure 2.6: (A) Schematic diagram of a nanoaperture fluorescence microscopy experiment designed to simulate the effect of increasing background concentrations of a FRET acceptor-labeled ligand biomolecule on a FRET-based nanoaperture fluorescence microscopy experiment. (B) A typical FRET signal from a single bio-$RF1_{Cy3,Cy5}$ in a nanoaperture. Cy3 and Cy5 fluorescence intensities versus time trajectories are plotted on top, and the corresponding EFRET versus time trajectory is plotted below. (C) Distributions of Cy3 SBRs (calculated as the change in Cy3 fluorescence intensity due to photobleaching divided by the standard deviation of pure background fluorescence) imaged with 0 nM (n=136), 1 nM (n=137), 10 nM (n=155), 100 nM (n=131), and 1000 nM (n=146) background concentrations of $RF1_{Cy5}$. Representative Cy3 fluorescence intensity versus time trajectories at specific SBRs from the 1000 nM $RF1_{Cy5}$ data set are shown below. The inset shows the average SBRs with bootstrapped, 1σ error bars (n=1000).

2.7 Passivated Nanoapertures Maintain Signal-to-Background Ratio

In order to quantify the ability of our nanoapertures to resist the SBR deterioration caused by high ligand biomolecule concentrations in solution and/or non-specific adsorption of ligand biomolecules to the nanoaperture surfaces, we calculated the SBR distributions of Cy3 fluorescence intensity observed during this experiment. As a reference, the SBR of the analogous TIRF microscopy experiment drops prohibitively low with greater than $\sim$50 nM of $RF1_{Cy5}$ in solution. After nonlinear least squares fitting a step function to each Cy3 fluorescence intensity versus time trajectory observed in single bio-$RF1_{Cy3,Cy5}$ occupancy nanoapertures, we calculated the SBR as the Cy3 fluorescence intensity difference due to photobleaching divided by the standard deviation of pure background fluorescence (measured using the last 50 photobleached timepoints of each Cy3 fluorescence intensity versus time trajectory), and plotted the SBR distribution of $\sim$130 - 150 single bio-$RF1_{Cy3,Cy5}$ molecules per background concentration of $RF1_{Cy5}$ that was tested (Fig. 2.6C). We observed an insignificant decrease in the average SBR at the highest back-

ground concentration of $RF1_{Cy5}$ we tested (1 μM), and presumably the fully passivated, gold nanoaperture arrays reported here would continue to enable detection of single bio-$RF1_{Cy3,Cy5}$ molecules with adequate SBR at concentrations of $RF1_{Cy5}$ much greater than 1 μM . Notably, this robust passivation would also be particularly effective in fluorescence colocalization-type smF experiments in which fluorophore-labeled ligand biomolecules in solution are directly excited by a laser excitation source and are observed to interact with either unlabeled or differentially fluorophore labeled target biomolecules tethered to the bottoms of the nanoapertures.

2.8 Conclusion

The significant SBR deterioration caused by the non-specific adsorption of ligand biomolecules onto the metallic and silica surfaces of nanoaperture arrays, as well as difficulties in the nanofabrication, passivation, and microscopy of nanoaperture arrays, have been major limiting factors for the widespread adoption of nanoaperture fluorescence microscopy for smF studies of biological systems. By combining a facile and effective procedure for fabricating and functionalizing gold nanoaperture arrays, we have generated a microfluidic device that can enable powerful new smF experiments of weakly interacting biological systems that were previously impracticable with microscopy techniques such as TIRF. We have demonstrated that our passivated nanoaperture arrays can be used to probe biological interactions under physiological concentrations of ligand biomolecules with a protein of interest, RF1; notably, straightforward extensions of these experiments enable the structural dynamics of RF1 to be characterized as it weakly interacts with ribosomes programmed with non-stop messenger RNA codons–previously impracticable smF experiments that will allow us to investigate the mechanisms governing the fidelity with which RF1 recognizes stop codons and terminates protein synthesis. Moreover, the well-documented, robust nature of the non-specific adsorption-resistance of PEG SAMs [15–18] suggests that the passivation scheme presented here should be resistant to the non-specific adsorption of many other types of biomolecules. Furthermore, the plasmon-mediated fluorescence enhancements that are attainable with a variety of gold nanoaperture geometries [25] are fully compatible with the surface functionalization approaches we present here–a benefit that we will attempt to exploit in future developments of the nanoaperture array design reported here.

2.9 References

1. Fersht, A. R. *Structure and mechanism in protein science. A guide to enzyme catalysis and protein folding.* 293–400 (W.H. Freeman and Co., New York, 1999).

2. Tinoco, I. & Gonzalez, R. L. Biological mechanisms, one molecule at a time. *Genes Dev.* **25,** 1205–1231 (2011).

3. Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G. & Webb, W. W. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).

4. Ha, T., Enderle, T., Ogletree, D. F., Chemla, D. S., Selvin, P. R. & Weiss, S. Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci.* **93,** 6264–6268 (1996).

5. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313,** 1642–1645 (2006).

6. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3,** 793–795 (2006).

7. English, B. P. *et al.* Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat. Chem. Biol.* **2,** 87–94 (2006).

8. Funatsu, T., Harada, Y., Tokunaga, M., Saito, K. & Yanagida, T. Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution. *Nature* **374,** 555–559 (1995).

9. Nie, S., Chiu, D. & Zare, R. Probing individual molecules with confocal fluorescence microscopy. *Science* **266,** 1018–1021 (1994).

10. Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci.* **105,** 1176–1181 (2008).

11. Uemura, S., Aitken, C. E., Korlach, J., Flusberg, B. a., Turner, S. W. & Puglisi, J. D. Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature* **464,** 1012–1017 (2010).

12. Sameshima, T. *et al.* Single-molecule study on the decay process of the football-shaped GroEL-GroES complex using zero-mode waveguides. *J. Biol. Chem.* **285,** 23159–23164 (2010).

13. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

14. Schomburg, I. *et al.* BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* **41,** D764–D772 (2013).

15. Mrksich, M. & Whitesides, G. M. Using self-assembled monolayers to understand the interactions of man-made surfaces with proteins and cells. *Annu. Rev. Biophys. Biomol. Struct.* **25,** 55–78 (1996).

16. Love, J. C., Estroff, L. A., Kriebel, J. K., Nuzzo, R. G. & Whitesides, G. M. Self-assembled monolayers of thiolates on metals as a form of nanotechnology. *Chem. Rev.* **105,** 1103–1169 (2005).

17. Palma, M. *et al.* Controlled Confinement of DNA at the Nanoscale: Nanofabrication and Surface Bio-Functionalization. *Methods Mol. Biol. Methods in Molecular Biology* **749** (eds Zuccheri, G. & Samorì, B.) 169–185 (2011).

18. Palma, M. *et al.* Selective biomolecular nanoarrays for parallel single-molecule investigations. *J. Am. Chem. Soc.* **133,** 7656–7659 (2011).

19. Mrksich, M. & Whitesides, G. in *Poly(ethylene glycol)* 23–361 (American Chemical Society, 1997).

20. Sternberg, S. H., Fei, J., Prywes, N., McGrath, K. a. & Gonzalez, R. L. Translation factors direct intrinsic ribosome dynamics during translation termination and ribosome recycling. *Nat. Struct. Mol. Biol.* **16,** 861–868 (2009).

21. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat Meth* **9,** 671–675 (2012).

22. Howarth, M. & Ting, A. Y. Imaging proteins in live mammalian cells with biotin ligase and monovalent streptavidin. *Nat. Protoc.* **3,** 534–45 (2008).

23. Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **9,** 10–20 (2007).

24. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9,** 90–95 (2007).

25. Wenger, J. *et al.* Emission and excitation contributions to enhanced single molecule fluorescence by gold nanometric apertures. *Opt. Express* **16,** 3008 (2008).

26. Gérard, D. *et al.* Nanoaperture-enhanced fluorescence: Towards higher detection rates with plasmonic metals. *Phys. Rev. B* **77,** 045413 (2008).

27. Owens, T. M., Nicholson, K. T., Banaszak Holl, M. M. & Süzer, S. Formation of alkylsilane-based monolayers on gold. *J. Am. Chem. Soc.* **124,** 6800–1 (2002).

28. Ha, T. *et al.* Initiation and re-initiation of DNA unwinding by the Escherichia coli Rep helicase. *Nature* **419,** 638–41 (2002).

29. Blanchard, S. C., Kim, H. D., Gonzalez, R. L., Puglisi, J. D. & Chu, S. tRNA dynamics on the ribosome during translation. *Proc. Natl. Acad. Sci.* **101,** 12893–8 (2004).

Chapter 3

Accurately and Precisely Inferring Single-Molecule Rate Constants*

3.1 Introduction

Dynamic single-molecule biophysical techniques provide sequential measurements of a signal originating from an individual biomolecule that is, or can be turned into, a proxy for the underlying molecular state of the system. For instance, the fluorescence resonance energy transfer (FRET) efficiency (E_{FRET}) of an individual biomolecule that is measured in a single-molecule FRET (smFRET) experiment is a one-dimensional measure of the distance between two fluorophore-labeled parts of the system, or the distance between two microbeads in laser traps during a single-molecule force spectroscopy experiment can be a proxy for the state of folding of an individual biomolecule that is attached to the microbeads [1]. Using such single-molecule approaches to investigate biomolecular dynamics avoids the ensemble averaging that can often obscure biologically relevant, transient, or rare kinetic events in bulk experiments. However, in order to perform such experiments, such dynamic single-molecule signals and experimental setups must be sensitive enough to unambiguously yield a descriptive time-resolved trajectory through the molecular states sampled during the experiment.

In order to obtain relevant kinetic information about a biomolecular mechanism from such dynamic single-molecule experiments, the signal versus time trajectories must first be transformed into state versus time

* Adapted from Kinz-Thompson, C.D., Bailey, N.A., Gonzalez, Jr., R.L. Accurately and Precisely Inferring Single-Molecule Rate Constants, *In Preparation*.
Additionally, adapted with permission from Kinz-Thompson, C.D., Sharma, A.K., Frank, J., Gonzalez, Jr., R.L., Chowdhury, D Quantitative Connection Between Ensemble Thermodynamics and Single-Molecule Kinetics: A Case Study Using Cryo-EM and smFRET Investigations of the Ribosome *Journal of Physical Chemistry B*, **2015**, 119(34), 10888-10901. Copyright 2015 American Chemical Society.

trajectories. This transformation process is not trivial, as limitations in signal- and temporal-resolution can easily obscure the relevant molecular states. With sufficient resolution, the molecular states can be determined manually–a scientist can choose the data point where the molecule transitions to a new state. Unfortunately, this process is subjective, time consuming, and often the data is not sufficiently resolved for this approach to yield reasonable results. A second, more systematic method involves setting a signal threshold that denotes a transition between states and indicates a new state once crossed, but this is only reasonable when a few well-resolved states are present, and is still subjective because the threshold must be set manually. A third, more rigorous and widely adopted method uses hidden Markov models (HMMs) to transform inherently noisy signal versus time trajectories into state versus time trajectories by estimating the underlying 'hidden' state responsible for producing the signal during each measurement period in the signal versus time trajectory [2, 3]. An advantage of using HMMs for this transformation is that they can manage many states simultaneously, and that methods have been developed to select the correct number of states present in the trajectory [4–7]. Regardless of whichever of these three, or another, methods is used to transform a signal versus time trajectory into the corresponding state versus time trajectory, the relevant kinetic parameters must then be calculated in order to obtain practical kinetic information about the observed system.

Traditional chemical kinetics approaches to modeling reactions employ phenomenological reaction rate equations to describe their time evolution [8]. Changes in concentration per unit time are proportional to the reaction rate constants, and the reaction rate constants can be determined by measuring the time-dependent changes in concentration and fitting reaction rate equations. Unfortunately, these traditional methods do not account for the stochastic fluctuations that are present in small systems such as in a cell or in the case of a single-molecule [9]; so, methods like the chemical master equation and the stochastic simulation algorithm were developed to model the time evolution of a such a reaction [10–15]. These stochastic methods model the probability of a single molecule reacting in a time period with a probability distribution described by the stochastic rate constant (as opposed to a reaction rate constant), and this approach allows the kinetics of individual molecules to be analyzed. While both the reaction rate constant and the stochastic rate constant can be used to characterize the kinetics of biomolecular systems, errors are introduced in both cases if the discretized state versus time trajectory to be classified misrepresents the behavior of the molecule(s) in state space. Fortunately, as we will show, these errors can be accounted for in a statistically rigorous manner.

Herein, we begin by clarifying the basis for several methods to calculate reaction rate constants and stochastic rate constants from single-molecule state versus time trajectories. Additionally, we introduce

Bayesian statistics-based approaches to calculating rate constants, which affords a natural method to account for the precision of the rate constants, given the finite length of the state versus time trajectories obtained from dynamic single-molecule experiments. Then, after discussing the types of missed events which result in misrepresentation of the state versus time behavior and therefore introduce misestimations into these rate constant calculation methods, we introduce methods that can be used to correct for these missed events.

3.2 Calculating Rate Constants from Single-Molecule Data

In order to calculate relevant kinetic parameters (*e.g.*, rate constants) from state versus time trajectories, it is helpful to quantify the trajectories in a manner which can be used by the various rate constant calculation methods. The state versus time trajectories described above are a series of sequential, discretized datapoints, where each datapoint indicates the state occupied by the single-molecule during a measurement period of length τ; it is worth noting that this state was inferred from a time-averaged signal collected during the measurement period. From these sequential datapoints, we can obtain a dwell time list, $\boldsymbol{n}_{ij}$, which is a list of the number of contiguous frames spent in a state, i, before transitioning to a second state, j. This has the form: $\boldsymbol{n}_{ij}$ = [5, 13, 12, 7, ...]. Additionally, we can construct a counting matrix, $\boldsymbol{M}$, where its elements, $\boldsymbol{M}_{ij}$, represent the number of times that the state versus time trajectories began in state i at measurement n (*i.e.*, at time t) and ended in state j at measurement n+1 (*i.e.*, at time $t + \tau$). $\boldsymbol{M}$ is related to the $\boldsymbol{n}_{ij}$ such that the off-diagonal elements, M_{ij}, are the number of entries in the corresponding $\boldsymbol{n}_{ij}$, and the on-diagonal elements, M_{ii}, are

$$M_{ii} = \sum_{j \neq i} \left((\sum n_{ij}) - M_{ij} \right), \tag{3.1}$$

where $\sum n_{ij}$ is the sum of the entries in n_{ij}. Note that $\sum\limits_{j \neq i} (\sum n_{ij}) = \sum\limits_{j} M_{ij}$. $\boldsymbol{M}$ may be row normalized, such that each element in a row (*i.e.*, with the same i) is divided by the sum of that row to yield the transition matrix, $\boldsymbol{P}$. The off-diagonal elements of the transition matrix, $\boldsymbol{P}_{ij}$, give the probability that a single-molecule in state i has transitioned to state j at the next measurement period. Below, we detail several methods to explain how the reaction rate constants and stochastic rate constants that characterize kinetic processes may be obtained from the calculated dwell time list, $\boldsymbol{n}_{ij}$, counting matrix, $\boldsymbol{M}$, or transition matrix, $\boldsymbol{P}$.

3.2.1 Ensemble Relaxtion Analysis

In a chemical system that was initially prepared experimentally to be away from equilibrium, the relaxation of a concentration, C, to its equilibrium value can be observed by many types of bulk experimental techniques. With such techniques, the large number of molecules typically present in the ensemble yields well-defined, ensemble-averaged relaxation behaviors. These time-dependent changes in C as the system relaxes to equilibrium are then typically described with phenomenological rate equations that are derived from postulated differential equations[14]. Notably, these rate equations are (i) deterministic in that an initial set of concentrations determines the subsequent values of the concentrations, and (ii) continuous in that individual numbers of molecules are not seen to undergo reactions, but rather the reaction occurs in terms of changes in concentrations. By fitting such experimentally observed relaxations to equilibrium to these phenomenological rate equations, one can determine the reaction rate constants that characterize the dynamics of the system [8].

For example, the reaction coordinate of a biological system along which a conformational change, such as protein folding, occurs. Due to the multiplicity of interactions present in such a complex system, the folding and unfolding events are can be considered as separate, random, irreversible, and unimolecular reactions [16, 17]. Thus, the dynamics of these types of biological reactions are often well described phenomenologically by the integrated first-order rate equation,

$$C\left(t\right) = C\left(t=0\right) \cdot e^{-k \cdot t}, \tag{3.2}$$

[8]. Flexibility in the description of the underlying dynamics can be obtained by varying the particular mathematical model employed (*e.g.*, using a double exponential decay), and can easily be derived from mass action considerations [18]. Regardless of the particular functional form, it is unfortunately often difficult to prepare biomolecular systems in a synchronized state from which one can monitor the these relaxations to equilibrium and then fit the corresponding rate equation. This is easily avoided in single-molecule experiments.

Single-molecule biophysical techniques can monitor the dynamics of an individual molecules as it progresses over time, and, from such data, particular behaviors of interest can be identified. While many behaviors might be occurring simultaneously (though asynchronously) in the bulk ensemble, they can be separated by at the single-molecule level. By isolating the particular instances of a behavior and then synchronizing

them such that they occur at the same time, a subsystem composed of only the process of interested can be created. This process is called "post-synchronization." The resulting post-synchronized ensemble of single-molecule data (composed of only the sub-system of interest) can be ensemble averaged, and then analyzed using the rate equations discussed above. The result is a description of the dynamics for the isolated behavior chosen to compose the post-synchronized sub-system (*i.e.*, reaction rate constants). To be more concrete, we will provide an example of this method, which we call ensemble relaxation analysis, below. Consider an effectively unimolecular reaction of interest in which a transitions occurs from an initial state, i, to a second state, j. We would first isolate the events of interest from the single-molecule state versus time trajectories by enumerating the number of consecutive measurements that the single-molecules spent in state i before transitioning to state j into $\boldsymbol{n}_{ij}$, as described above in Section 3.2. This entries in this list are then converted into lifetimes, t_{ij}, by multiplying each n_{ij} by τ. At this point, we perform an ensemble average and construct a population survival function, $S_{ij}(t)$, which is the fraction of the events of interest (as measured by t_{ij}) that have not undergone the transition from state i to state j by time, t. This can be thought of as analogous to a measurement of the normalized concentration of state i relaxing to equilibrium through a transition to state j. Assuming that the lifetimes in state i are exponentially distributed, $S_{ij}(t)$ is then fit with an integrated rate equation in Equation (3.2) to yield the reaction rate constant, k_{ij}, that describes the rate of transition from state i to state j. It is worth noting that, since $S_{ij}(t)$ is constructed from discrete measurements of lifetimes in units of measurement periods, fitting it with a continuous function of time is an approximation. However, this is not necessarily a bad approximation, as $S_{ij}(t)$ is approximately continuous when (i) $\boldsymbol{n}_{ij}$ contains a sufficiently large number of dwell times (more than 100), so that the data more closely resemble a continuous function, and (ii) the length of the measurement period is significantly shorter than the average lifetime t_{ij} (since lifetimes that are shorter than the measurement period τ would not be seen). If either condition does not hold, then fitting the data to an continuous function can result in misestimation of the reaction rate constant.

Interestingly, the population survival function, $S_{ij}(t)$ is closely related to the cumulative distribution function (CDF); the CDF describes the population that has already transitioned from state i to state j at time t, while the survival function describes the population that is left to transition from state i to state j. Since, a molecule can only have transitioned or have not transitioned by some time, conservation of probability imposes that the CDF = 1 - $S_{ij}(t)$. Therefore, since ensemble relaxation to equilibrium implies a particular the survival function, it also implies a particular CDF. Similarly, a particular CDF implies a connection to a probability distribution function (PDF), which is the derivative of the CDF. Onsager's regression hypothesis [15, 19],

which connects macroscopic relaxation to equilibrium to the underlying microscopic relaxation processes involved in the reaction, asserts that we can interpret this PDF as the distribution of times that individual single-molecules spend on average undergoing the reaction. In other words, the bulk ensemble reaction rate constants should be related to the average stochastic rate constants of the individual single-molecules.

However, this approach (effectively 'moment-matching') requires several assumptions about the observed single-molecule data, including that a sufficient amount was observed to accurately represent the ensemble average, and that the subpopulations present sample were equilibrated and would not have changed over time. These assumptions are difficult to confirm. A slightly more reasonable method of obtaining kinetic parameters for the ensemble is to first obtain stochastic rate constants for each single-molecule, as this yields information about the subpopulations that are present, instead of treating all the single-molecules equivalently. Unfortunately, such a calculation lacks precision due to the small amounts of information present for state versus time trajectory from an individual single-molecule.

3.2.2 Dwell-Time Analysis

Rather than calculate the reaction rate constant of an ensemble decaying to equilibrium, a different approach is to calculate the stochastic rate constants governing the individual molecules of the ensemble. This approach is more in-line with the motivation behind single-molecule experiments. One method to calculate stochastic rate constants from single-molecule state versus time trajectories is by analyzing the distribution of dwell-times (*i.e.*, number of consecutive frames occupied by a single-molecule before it transitions to another state).

A state versus time trajectory can be thought of as a sequence of discrete measurements that report on whether a transition has occurred between two measurements (Figure 3.1). These 'transition trials' are reminiscent of a series of repeated Bernoulli trials from probability theory [20], which are events where the outcome is either a success with probability p, or a failure with probability $1-p$. In this analogy, a successful Bernoulli trial would be when the single-molecule transitions from state i at time t to state j at time t+1; whereas a Bernoulli trial failure is when, instead, the single-molecule remains in state i at time t+1.

The number of repeated, failed trials before a success (transition) occurs is distributed according to the geometric distribution (see Appendix B) [20]. Therefore, the probability mass function (PMF) of the number of measurement periods until a transition occurs in a Markovian state versus time trajectory occurs can be modeled using the geometric distribution. From the geometric distribution, we expect that the mean number

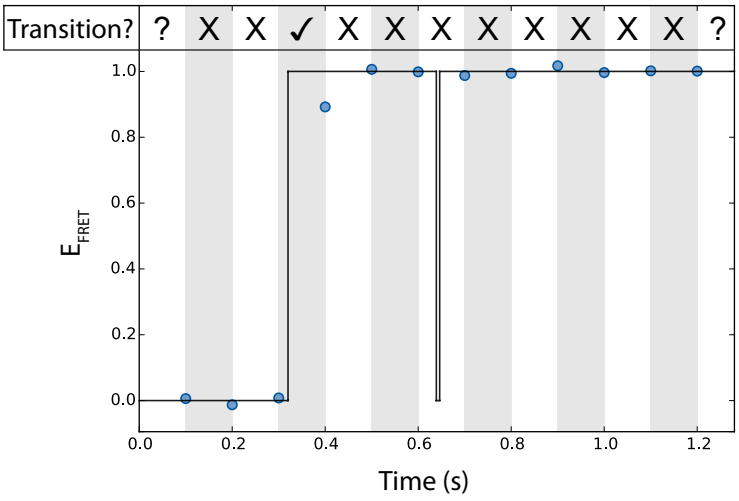

Figure 3.1: Schematic of Transition Probability. The black path represent the exact signal versus time trajectory of a simulated single-molecule. The blue scatter plot shows the discrete and noisy, observed signal versus time data. Alternating white and grey stripes denote the periods during which the presence of one transition is determined from the datapoints at the beginning and end of the period. The observed presence or absence of a transition is shown above.

of successive measurements periods in state i, n_i, until a transition out of state i occurs is,

$$\langle n_i \rangle = \frac{1 - P_i}{P_i}, \tag{3.3}$$

where, P_i is the probability of a successful transition out of state i to any other state, j. Before proceeding, we note that the geometric distribution is the negative binomial distribution with $r = 1$ (Appendix B), and so this derivation can be recast using the negative binomial distribution. That formulation will be used in Section 3.3.

Given a particular state versus time trajectory that is Markovian, an estimate of the mean number of measurement periods before a transition out of state i occurs would then allow the probability of a successful transition out of state i by solving Equation (3.3). The maximum-likelihood estimate of n_i is the total number of measurement periods observed to be in state i divided by total number of transitions out of state i, or rather, in terms of the quantities calculated in Section 3.2,

$$\langle n \rangle = \frac{\sum\limits_{j} (\sum n_{ij})}{\sum\limits_{j \neq i} M_{ij}}, \tag{3.4}$$

where $\sum n_{ij}$ is the sum of all entries in n_{ij}, and the M_{ij} are the total number of observed transitions from

state i to state j. Solving Equation 3.3 with Equation 3.4, yields,

$$P_i = \frac{\sum\limits_{j \neq i} M_{ij}}{\sum\limits_{j \neq i} M_{ij} + \sum\limits_{j} \left(\sum n_{ij}\right)}. \tag{3.5}$$

For an uncorrelated, Markovian system, the time lengths that a single-molecule will spend in a particular state before transitioning to a different state are distributed according the the exponential distribution (*c.f.*, Appendix B). For such a Markovian system, the probability of a transition between state i to state j during the observation period of one measurement in a signal versus time trajectory, P_{ij}, is therefore the integral of this PDF from $t = 0$ to τ, the period of the measurement, which is

$$P_{ij} = \int_0^{\tau} k_{ij} \cdot e^{-k_{ij} \cdot t} \cdot dt = \left(1 - e^{-k_{ij} \cdot \tau}\right). \tag{3.6}$$

Equation (3.6) implies that a rate constant can be calculated as

$$k_{ij} = \frac{-\ln\left(1 - P_{ij}\right)}{\tau}, \tag{3.7}$$

if the transition probability can be quantified. Regardless of the particular state j that is transitioned into, the distribution of dwell-times in state i is then exponentially distributed with according to a rate constant that is the sum of all the stochastic rate constant that make up the parallel reaction pathways out of state i. This relationship yields,

$$k_i = \sum_j k_{ij} = \frac{-\ln\left(1 - P_i\right)}{\tau} = \frac{-\ln\left(1 - \frac{\left(\sum\limits_{j \neq i} M_{ij}\right)}{\left(\sum\limits_{j \neq i} M_{ij}\right) + \left(\sum\limits_{j} \left(\sum n_{ij}\right)\right)}\right)}{\tau}. \tag{3.8}$$

As a consequence, expect for cases when there is only one state to transition to (*e.g.*, two-state systems), the stochastic rate constant obtained by considering the dwell times in a particular state will be a sum of multiple rate constants. Notably, analyzing only those lifetimes which transition from state i to a particular state j still yields the same sum of the rate constants, and not the associated k_{ij}. However, the benefit of analyzing the distribution of dwell-times is that deviations from Markovian behavior can be observed as non-exponential behavior, and then quantified.

Interestingly, Equation 3.4 reveals a difficulty in applying this dwell-time distribution analysis method. This

difficulty is the stipulation of the geometric distribution that requires the state versus time trajectories to have discrete dwell times that last {0,1,2,...} measurement periods before a transition occurs. Unfortunately, in a state versus time trajectory, the dwell times of zero measurement periods are never included in the n_{ij} lists, because a dwell time must be at least one measurement period long for it to be associated with a particular state. The result is an undercounting of M due to the exclusion of all zero measurement period-long dwell times, and a subsequent miscalculation of P_i; compounding this undercounting is the fact that, from the geometric distribution, the highest probability dwell-times are the zero measurement period-long ones. As a result, the rate constant calculated here by dwell-time distribution analysis is a misestimate, and more specifically, an underestimate of the true stochastic rate constant. However, this underestimate is easily accounted for by the process of conditioning the geometric PMF such that only dwell-times that are greater than zero frames in length are considered.

To account for these unclassified dwell times, we condition the geometric PMF so that it only considers $n > 0$, and denote these dwell-time lengths with $n^\star \in \{1, 2, ...\}$ to maintain clarity. From the law of conditional probability, we note that

$$\begin{aligned} P(n^\star|p, n > 0) &= \frac{P(n|p, n = 0 \cap n > 0)}{P(n|p, n = 0)} \\ P(n^\star|p) &= \frac{p(1-p)^{n^\star}}{1 - (1 - (1-p)^{0+1})} \\ P(n^\star|p) &= p(1-p)^{n^\star - 1} = \frac{1}{1-p} \cdot P(n|p). \end{aligned} \tag{3.9}$$

Therefore, the geometric PMF conditioned upon all dwell-times being greater than zero-measurements is equivalent to the regular geometric PMF divided by $(1 - p)$. Because $p(n^\star|p)$ is proportional to $p(n|p)$ in a manner that does not depend upon n, the expectation values of $p(n^\star|p)$ are also proportional to those of $p(n|p)$ in the same manner due to linearity. Therefore,

$$\langle n^* \rangle = \frac{1}{1-p} \langle n \rangle. \tag{3.10}$$

Since Equation (3.4) in practice would only include dwell-times where $n > 0$, it is effectively for $n^\star$, not n. We can then follow the same derivation of P_i above, but substitute the right hand side of Equation (3.10) in

place of $\langle n \rangle$. This yields,

$$\langle n_i^\star \rangle = \frac{\sum_j \sum n_{ij}}{\sum_{j \neq i} M_{ij}} = \frac{1}{1 - P_i} \cdot \frac{1 - P_i}{P_i} \tag{3.11}$$

$$P_i = \frac{\sum_{j \neq i} M_{ij}}{\sum_j \sum n_{ij}}. \tag{3.12}$$

Interestingly, this is the exact same result for the transition probability P_{ij} as is obtained with the transition probability expansion analysis described below in Section 3.2.3.

For more insight into this expression, consider that, from the Poisson distribution, the expected value for the number of transitions out of state i is $k_i \cdot T_i$, where T_i is the total time spent in state i. Then, from the expected value of $\langle n_i \rangle$, rather than the observed value of $\langle n_i \rangle$, we find that

$$P_i = \frac{\sum_{j \neq i} M_{ij}}{\sum_j \sum n_{ij}} \approx \frac{T_i \sum_j k_{ij}}{T_i / \tau} = \sum_j k_{ij} \tau = k_i \tau. \tag{3.13}$$

This expression for P_i is different from that in Equation 3.6. From the Taylor series

$$e^x = 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + ..., \tag{3.14}$$

we see that Equation 3.13 is the Taylor series expansion of Equation 3.6 truncated after the first-order term. With this in mind, the method of calculating rate constants using Equation 3.12 as the transition probability for Equation 3.7 is a first-order expansion of the transition probability. Notably, since this expression is conditioned upon only the observation of dwell-times that are longer than one frame long, it is insensitive to the types of missed dwells in state i that are less than one frame long. However, as we will show later in Section 3.3, it is sensitive to other types of missed events.

Finally, it is worth noting that stochastic rate constants for a particular reaction pathway out of state i, k_{ij},

can be calculated from k_i, by equating the splitting probability, p_{ij}^{split} and the observed branching ratios as

$$\begin{aligned} p_{ij}^{\text{split}} &= \frac{k_{ij}}{\sum\limits_{j} k_{ij}} \\ \frac{M_{ij}}{\sum\limits_{j \neq i} M_{ij}} &\approx \frac{k_{ij}}{k_i} \\ \therefore k_{ij} &= \frac{M_{ij}}{\sum\limits_{j \neq i} M_{ij}} \cdot k_i. \end{aligned} \tag{3.15}$$

Since we will not discuss this approach in Sec. 3.2.4, we note here that the calculation of k_{ij} given above can be recast with a Bayesian inference approach by utilizing a Dirichlet distribution as the conjugate prior and a multinomial distribution as the likelihood function. Regardless, while this dwell-time distribution approach to calculating individual stochastic rate constants is quite effective, and it has the benefit of allowing the dwell-times to be checked for non-Markovian behavior that would render the calculated rate constants much less meaningful, there are more straightforward methods to calculate the stochastic rate constants for each parallel reaction pathway of a single-molecule.

3.2.3 Transition Probability Expansion Analysis

Another method to calculate stochastic rate constants is to consider the observed frequency with which a single-molecule transitions from one state to another. For the discrete state versus time trajectories considered here, this is equivalent to determining whether the single-molecule in state *i* during a measurement period is in state *j* during the subsequent measurement period (see Fig. 3.1). As discussed above, since the data consists of multiple 'trials' of whether or not the transition has occurred, the transition probability can be modeled by the binomial distribution (*c.f.*, Appendix B). The binomial distribution is appropriate to model the number of successful trials (*i.e.*, transitions from state i to state j) from a certain number of performed trials (*i.e.*, the number of times the single-molecule was in state i in the state versus time trajectory) that can each succeed with a fixed probability (*i.e.*, P_{ij}). From the expectation values of the binomial distribution, we

will take a frequentist approach and assume that the maximum-likelihood solution is

$$\mathbb{E}[m] = np \tag{3.16}$$

$$M_{ij} \approx \left(\sum_j (\sum n_{ij}) \right) P_{ij}$$

$$= \left(\sum_j M_{ij} \right) P_{ij} \tag{3.17}$$

where m is the number of successful trials, n is the total number of trials, $p = P_{ij}$ is the probability of a successful trial, and, as described in Section 3.2, M_{ij} is the number of observed measurements associated with state i that transition to state j. In doing this, we have equated P_{ij} with the observed frequency of the transitions from state i to state j. However, in an experiment, only a finite number of transitions from state i to j are observed; as such, the equality will only be approximate. Regardless, by the central-limit theorem, with more and more measurements, M_{ij} should approach the expectation value; so, baring a small number of measurements, we might reasonably estimate that

$$P_{ij} = \frac{M_{ij}}{\sum_j M_{ij}}, \tag{3.18}$$

and from this expression, estimate k_{ij} using Equation (3.7).

Now, we will consider the accuracy of calculating a rate constant in this manner. Interestingly, given an amount of time spent in state i, T_i, the Poisson distribution indicates that

$$\langle M_{ij} \rangle = k_{ij} \cdot T_i \approx k_{ij} \cdot \left(\sum_j M_{ij} \tau \right), \tag{3.19}$$

where $\langle ... \rangle$ denotes the mean, and where the substitution for T_i is generally accurate excepting the types of missed events which we will discuss in Section 3.3.1. With this in mind, by substituting Equation 3.19 into Equation 3.18, we find that

$$P_{ij} = \frac{k_{ij} \cdot \sum_j M_{ij} \cdot \tau}{\sum_j M_{ij}} = k_{ij} \cdot \tau. \tag{3.20}$$

Therefore, this method is inherently (rather than being corrected to) a Taylor series expansion truncated at the first order term. As a result, this method is fairly accurate when $k_{ij}\tau$ is small (*i.e.*, much less than one)

where the higher-order terms are negligible. However, when $k_{ij}\tau$ begins to become large (*i.e.*, approaching and greater than 1), the possibility of experimentally recording measurements where more than one state is occupied during the measurement period becomes substantially probable (see Fig. 3.1). While under these conditions, the first-order expansion of the Taylor series is not well justified, neither is the process of idealizing a signal versus time trajectory into a state versus time trajectory; so other approaches beyond the scope of this chapter should be developed to handle these situations (one such method to achieve such 'temporal super-resolution' is developed in Chapter 4). Regardless, before moving on to discuss a more comprehensive method of analyzing state versus time trajectories to calculate stochastic rate constants from individual molecules, we would like to note here that the transition probability expansion analysis described in this section has the added benefit of being insensitive to certain types of missed events. These events are described in Section 3.3.1. Finally, we note that this type of analysis is analogous to using the transition matrix from an HMM for P_{ij}.

3.2.4 Statistically Rigorous Precision

While several methods of calculating rate constants from a collection of single-molecule state versus time trajectories were provided above, and even some allusions to their accuracy were made, the reliability of calculating stochastic rate constants has not been addressed. For instance, fitting errors such as those from the curve-fitting-based analysis in Section 3.2.1, do not provide information about the reliability of calculating ensemble parameters from a limited amount of single-molecule data. In fact, with a finite amount of data, this type of 'counting' error often dominates the calculation.

One simplistic attempt to account for this source of variability is to report statistical uncertainty in the context of 'bootstrapping' of the data. Bootstrapping is an attempt to simulate the data of future experiments from a set of observed data. From the bootstrapped, 'future' data, any variation in a subsequent analysis can attributed to the uncertainty present in the original dataset. For example, when calculating rate constants as described above in Section 3.2, the bootstrapping process involves creating a resampled data set, n'_{ij}, by randomly sampling from n_{ij} with replacement. The new transition probability, P'_{ij}, can then be calculated from n'_{ij}, and this yields new rate constants, k'_{ij}. The bootstrapping process is then repeated several times, and the reported rate constant k_{ij} is given as the mean of the set of bootstrapped k'_{ij}, with the error of the reported k_{ij} given as the standard deviation of the set of bootstrapped k'_{ij}. In this manner, bootstrapping is used to generate the uncertainty in calculated values, but this method inherently assumes that the collected data accurately represents the characteristics of an infinitely large ensemble. Consequentially, bootstrap-

ping artificially inflates the data set in a way that perpetuates any misrepresentations of the infinitely large ensemble that are present in the actual dataset. The smaller the collected dataset is, the more likely it is to misrepresent this infinitely large ensemble. Therefore, bootstrapping single-molecule results, where there are often only several hundreds of individual molecules in a dataset, often perpetuates these misrepresentations and leads to misrepresentation of the rates, all-the-while not providing a reasonable estimate of the statistical error present in the calculation.

Consider the following, extreme, hypothetical calculation where only one transition with a one measurement period-long dwell-time has been observed in a single-molecule experiment. Using conditioned dwell-time distribution analysis or transition probability expansion analysis, we find that P_{ij} is equal to 1.0, and that all of the bootstrapped P'_{ij} are also equal to 1.0. Thus, in this case, there is *no* uncertainty in the calculation of the transition probability, or, subsequently, in the rate constant, and that rate constant is infinitely large. However, we know intuitively that the rate constant is not infinity, and also that there most likely is uncertainty in this calculation that employs only one measurement. The uncertainty lies in the fact that the one transition we have observed simply cannot be representative of an entire ensemble or even a single-molecule. Likewise, we should suspect that P_{ij} is probably a poor estimate of the true transition probability. It is easy to imagine that after recording a few more measurements from that hypothetical single molecule, we might calculate a different value of P_{ij}, and that the extra data would give us a better sense of the uncertainty in P_{ij}. This extreme example illustrates how the mathematical analyses described in Section 3.2 are insufficient by themselves, even when supplemented by bootstrapping. Fortunately, in contrast to these analytical shortcomings, Bayesian inference provides a statistically rigorous manner with which to encode our intuition that the number of observations should change our knowledge about P_{ij}, and systematically address the uncertainty in our rate constant calculations.

Bayesian inference is a statistical method grounded in the Bayesian approach to probability (see Ref. 21 or Section 4.2.2 for an introduction). Given a model that is supposed to describe some data, rather than treating the parameters of that model as fixed values, they are treated as distributions with probabilities that reflect their consistency with the data. These probability distributions can then be updated if new data is acquired so as to be consistent with the new, and any previous, data–an approach that is very similar to the way that a scientific hypothesis is tested and then updated with each new laboratory experiment [21]. In the context of quantifying single-molecule state versus time trajectories, Bayesian inference allows us to formulate a hypothesis about the underlying kinetic rate constants of a system (*i.e.*, the probability of certain rate constants producing the observed state versus time trajectories), and then to update that hypothesis

as each transition, or lack thereof, is observed in the state versus time trajectory. In this way, we can use Bayesian inference to describe the probability distribution of a rate constant as each measurement period is analyzed from a state versus time trajectory.

Generally, the basis of Bayesian inference is Bayes' rule, which can be mathematically written as:

$$P\left(\Theta|D\right) \propto P\left(D|\Theta\right) \cdot P\left(\Theta\right), \tag{3.21}$$

where Θ represent the parameters of the model, and D represents the data values; the first, second, and third terms are referred to as the 'posterior', the 'likelihood', and the 'prior', respectively. Bayes' rule can be expressed verbally as: the probability of the model's parameter values after observing the data is proportional to the product of (i) the probability of observing the data given those particular parameter values, and (ii) the initial probability of those parameters. More succinctly, the posterior probability is proportional to the product of the likelihood and the prior probability.

With expressions for the likelihood and the prior probability distribution, we can calculate the posterior probability distribution and learn about the distribution of parameter values that are consistent with the data. Unfortunately, these calculations are often analytically and/or numerically difficult making their practical use relatively intractable. However, there are certain conditions that significantly simplify these calculations. For instance, certain pairs of likelihood functions and prior distributions yield posterior distributions that are the same algebraic form as the prior (*i.e.*, they have the same form of probability distribution); in such a case, the prior is called the conjugate prior for that particular likelihood function. The benefit of using a conjugate prior, given a particular likelihood function, is that simple updating rules can be applied to the parameters of the conjugate prior probability distribution to yield the resulting posterior probability distribution; this circumvents the need to enumerate the posterior over the entire probability space, which is computationally expensive. The Bayesian approach to calculating rate constants using dwell-time distribution analysis and transition probability expansion analysis that we describe below obeys these conditions–they will use likelihood functions and their associated conjugate priors–and therefore they are extremely tractable.

3.2.5 Bayesian Dwell-Time Distribution Analysis

To perform Bayesian inference using sequential transition trial analysis, we must first identify the likelihood function and its conjugate prior probability distribution. As described in Section 3.2.2, if we wish to model the probability of a successful transition from the observed lengths of dwell-times, the likelihood function

that describes the probability of observing sequential measurements in state i before transitioning to state j is a geometric distribution. The conjugate prior of the geometric distribution is the beta distribution (see Appendix B), which is often used to describe the probability of a probability (in this case, of a successful transition, p), because it is defined on the interval $[0, 1]$, and is a function of only two parameters, α and β, which have intuitive interpretations. Notably, when $\alpha = \beta = 1$, the beta distribution is flat, as all values of p have equal probabilities–in this case, the beta distribution mathematically expresses a lack of knowledge about p in a similar manner as the equal, *a priori* probability assumption of statistical mechanics. Along these lines, larger values of α and/or β yield more defined and peaked distributions, which expresses the increased knowledge about p. As we will discuss below, the process of performing Bayesian inference amounts to modifying the initial values of α and β in a data dependent manner to yield the posterior distribution. In this sense, Bayesian inference mathematically encodes a method to express the incremented knowledge that originates from new information.

By using the geometric distribution as the likelihood function, and the beta distribution for the conjugate prior, we can now calculate the posterior probability distribution of the transition probability P_{ij} from state versus time trajectories. We begin by assuming that all transition probabilities are initially equally probable. Therefore, the prior distribution is a beta distribution with $\alpha = \beta = 1$. The posterior probability distribution will be another beta distribution where α and β are interpreted as 1 + the number of Bernoulli trial successes, and 1 + the number of Bernoulli trial failures, respectively. Thus, for the transitions in a state versus time trajectory, sequential transition trial analysis yields a posterior probability distribution where $\alpha = 1 + \sum_{j \neq i} M_{ij}$, and $\beta = 1 + \sum_{j} (\sum n_{ij})$. Therefore, from the mean of the beta distribution (see Appendix B), the mean transition probability out of state i after having observed the single-molecule state versus time trajectories is

$$\langle P_i \rangle = \frac{1 + \sum_{j \neq i} M_{ij}}{1 + \sum_{j \neq i} M_{ij} + 1 + \sum j (\sum n_{ij})}. \tag{3.22}$$

This mean value of the transition probability converges to the maximum likelihood estimate of P_{ij} from Section 3.2.2 when $\sum_{j \neq i} M_{ij} \gg 1$ and $\sum_{j} \sum n_{ij} \gg 1$; we note that the mode of a beta distribution function is $(\alpha - 1) / (\alpha + \beta - 2)$, which is equivalent to the maximum likelihood estimate of P_i. The benefit of this approach is that the posterior probability distribution of P_i not only provides a mean value, but also speaks to the uncertainty inherent in P_i due to limited amounts of data. This uncertainty is expressed by a credible interval, which is similar to the frequentist idea of a confidence interval. A credible interval is the range in which a certain percentage of the probability density of resides; typically one uses a 95% credible interval

(*e.g.*, $\sim \pm 2\sigma$), but this choice is arbitrary. The upper- and lower boundaries of the credible interval can be found through the inverse of the CDF of the beta distribution (see Appendix B). Many standard computational programs come with a function to do this, which is called the inverse function of the regularized incomplete beta function, $I_x(\alpha, \beta)$, where α and β are the posterior parameters and x is the fraction of the boundary (*e.g.*, 0.025 for 2.5%).

Finally, let us consider the application of this Bayesian approach to actual observed data, where the length of a dwell-time must be greater than zero measurements. With the linearity of conditioning the geometric distribution upon $n > 0$, the posterior probability distribution conditioned upon $n > 0$ will contain a term of $\left(\frac{1}{1-p}\right)^{\sum_{j\neq i} M_{ij}}$, because the total likelihood function is the product of the likelihood function from each individual datapoint. This is equivalent to setting $\beta' = \beta - M_{ij}$, where β' is the parameter used in the beta distribution for the posterior probability distribution, and β is the parameter calculated above. Using α and β', the mean and the 95% credible interval for P_i are calculated to account for the unclassified dwell times of zero length. For instance,

$$\langle P_i \rangle = \frac{1 + \sum\limits_{j\neq i} M_{ij}}{1 + \sum\limits_{j\neq i} M_{ij} + 1 + \sum\limits_{j}(\sum n_{ij}) - \sum\limits_{j\neq i} Mij} = \frac{1 + \sum\limits_{j\neq i} M_{ij}}{2 + \sum\limits_{j}(\sum n_{ij})} \approx \frac{\sum\limits_{j\neq i} M_{ij}}{\sum\limits_{j}(\sum n_{ij})}. \tag{3.23}$$

With a sufficient number of measurements, this treatment yields the same mean transition probability as the maximum likelihood estimate of the transition probability expansion analysis, and so it is also insensitive to some types of missed dwells. These values can then be transformed into rate constants using Equation (3.7). Since k_i is a monotonic function of P_i, this Bayesian method also provides an intuitive, explicit expression for how the uncertainty in k_i diminishes with additional observations. Finally, we note that when no measurements have been made, the posterior distribution is equivalent to the prior distribution; all rate constants from 0 to ∞ are therefore equally probable. Thus, this analysis method is a very objective approach to analyzing transition probabilities from discrete signal versus time trajectories, and it is one that intrinsically encodes a statistically rigorous approach to the precision of such calculations.

To be concrete, we will use this Bayesian sequential dwell-time distribution analysis in the extreme, hypothetical case of the single observed transition introduced above in Section 3.2.4. The posterior probability distribution would be a beta distribution with $\alpha = (1 + 1) = 2$, and $\beta' = (1 + 1 - 1) = 1$. This yields $\langle P_i \rangle$ = 0.66, and a lower-bound of P_i = 0.16 and an upper-bound of P_i = 0.99 for the 95% credible interval. Notably, the mean value of the transition probability calculated using the Bayesian dwell-time distribution analysis is

not infinitely large as was the earlier estimate of P_i using the maximum-likelihood approach as prescribed in Section 3.2.2, and by the credible interval, this method inherently accounts for the large uncertainty in the transition probability that we intuitively expect.

3.2.6 Bayesian Transition Probability Expansion Analysis

Similar to the the extension of the dwell-time distribution analysis in Sec. 3.2.2 using Bayesian inference above in Sec. 3.2.5, we can also extend transition probability distribution analysis to account for precision in a statistically robust manner. Since we modeled the probability of undergoing a transition from state i to state j during a measurement period with the binomial distribution, the binomial distribution will be the likelihood function used to perform Bayesian inference. The conjugate prior to the binomial distribution is the beta distribution, as described in Sec. 3.2.5 and Appendix B. Without any fore-knowledge of the stochastic rate constants, we should use a flat, uninformative prior of $\alpha = \beta = 1$. From this, the resulting posterior probability distribution for P_{ij} is a beta distribution with $\alpha = 1 + M_{ij}$, and $\beta = 1 + \left(\sum_j M_{ij}\right) - M_{ij}$. For the extreme case example of a state versus time trajectory is one transition from a one-measurement-long dwell-time, the posterior probability distribution would then be $\alpha = (1 + 1 = 2)$, and $\beta = (1 + 1 - 1) = 1$. The mean and the credible interval can then be calculated as described above in Section 3.2.5, and then the stochastic rate constants related to these transition probabilities can be calculated with Equation (3.7).

Interestingly, a more precise mathematical description for the transition probability is given by considering all of the parallel reaction pathways out of state i at once. In this case, the multivariate generalization of the binomial distribution, the multinomial distribution, is more appropriate for the likelihood function as it models the probability of a Bernoulli trial where there are multiple possible successes (though only one is chosen at a time). The conjugate prior to the multinomial distribution is the Dirichlet distribution (Appendix B). The Dirichlet distribution is the multivariate-equivalent of the beta distribution; in fact, with only one dimension, they are equivalent. Analogously, we will use a flat, uninformative prior of $\alpha_{ij} = 1$, such that each element of $\boldsymbol{\alpha}$ is unity. Then, the posterior probability distribution is $\alpha_{ij} = 1 + M_{ij}$, or rather it is $\boldsymbol{M}$ plus the prior probability distribution. To analyze the transition probability of an individual reaction pathway out of state i, we can marginalize the posterior Dirichlet distribution, which then becomes a beta distribution with $\alpha = \alpha_{ij} = 1 + M_{ij}$, and $\beta = (\sum_j \alpha_{ij}) - \alpha_{ij} = (\sum_j 1) + (\sum_j M_{ij}) - M_{ij} - 1$. This is equivalent to the binomial result when there are only two states (*i.e.*, i and j), however, we see slight deviations due to the prior in higher dimensions. However, the most notable aspect of this treatment that when the $M_{ij} \gg 1$, we find that $\langle \boldsymbol{\alpha} \rangle$ is approximately $\boldsymbol{P}$, the transition matrix calculated in Section 3.2. This transition matrix is essentially equivalent to the

transition matrix that is calculated using an HMM. Finally, the credible interval for the marginalized result can be calculated as above.

Importantly, the Bayesian approach to transition probability expansion analysis enables the statistically robust analysis of trajectories where there are not only no transitions to a particular state, but also when there are no transitions at all during a state versus time trajectory. In these cases, the on diagonal elements of M_{ii} should reflect the measurements from the state versus time trajectory that were assigned to state i, even though it was unclear what the final state j is. In doing so, the prior probability distribution takes care of the numerical instability the would otherwise yield infinitely precise estimates of rate constants that are zero.

3.3 Correcting Rate Constants for Missed Events

Complicating Features of Single-Molecule Trajectories As discussed above, discretized, idealized state versus time trajectories are used to analyze the dynamics of single-molecules. Many factors complicate the quantification of these state versus time trajectories, and limit the amount of information that can be extracted from them. For instance, if the underlying single-molecule dynamics are faster than the time resolution (*i.e.*, integration time of the experimental measurement period) of the experimental technique used to record the signal versus time trajectories from which the state versus time trajectories originate, then the underlying single-molecule dynamics will not be well represented by an idealized, discretized time trajectory. Here we discuss some types of events that complicate the process of analyzing single-molecule data using state versus time trajectories, in order to later discuss correcting for the effects of these events.

3.3.1 Types of Missed Events

Finite-Length

Many factors limit the length of the signal versus time trajectories that can be collected from individual biomolecules using single-molecule techniques. Superficially, the patience of the experimenter, and practical data storage limitations of computers restrict this length. More practically, the stability of the biomolecular system can limit the length of an experiment–for instance, many *in vitro* reconstituted enzymes become inactive after a certain time spent at room temperature, or the buffering capacity of a buffer might saturate. More probably, the signal corresponding to the single-molecule can simply be lost, and this terminates a signal versus trajectory–for instance, by photobleaching of a fluorophore, or dissociation of a tether. Regardless

of the cause, signal versus time trajectories are finite in length and do not extend to t = ∞. Considering the ergodic hypothesis, data from a single-molecule necessarily does not contain enough information to completely characterize a system. In an extreme case, one can imagine a state versus time trajectory where no transitions occur before signal loss. Such a situation places a clear limitation on precision with which the dynamics of the single-molecule system can be quantified. This consideration applies to all state versus time trajectories, because all will have a finite length.

Missed Transitions

Consider a single-molecule that dwells in a particular state, A, for some length of time. Eventually, the single-molecule will transition to a new state, B. If the dwell-time in B is shorter than the measurement period, there is a chance that the single-molecule might transition back to A during the measurement period (Figure 3.2, orange boxes). This is more likely to occur with increasingly fast rate constants for the transition from B to A. In a state-versus time trajectory, such a transition from A to B would not be registered. Instead, the single-molecule would appear to have remained in A throughout this measurement period–not having transitioned to the new state; this event is called a missed transition, and they affect the M_{ij} and the n_{ij}. The direct consequence of the missed transition is that the number of transitions from A to B, M_{AB}, would be under-estimated, and this leads to an under-estimation of k_{AB}. Additionally, as a result of the missed transition, the initial dwell-time in A would be over-estimated, because it would be the combined length of the initial dwell-time and the following dwell-time in A. This leads to an over-estimation of M_{AA}, and therefore an under-estimation of the rate constant from A to B, k_{AB}. Similarly, in this example, the transition back from B to A is also missed, and this results in an under-estimation of M_{BA}, and therefore an under-estimation of the rate constant for that transition, k_{BA}.

Misclassified Transitions

A related occurrence is that of misclassified transitions, rather than of missed transitions. Similarly, a single-molecule beginning in A could transition to B, where it dwells a time that is less then the measurement period. Instead of transitioning back from B to A, as in the example above, the single-molecule could transition to a different state, C. In this case, from the state versus time trajectory, the initial dwell-time in A can be approximately correctly measured, but the transition from A to B will be misclassified as a transition from A to C, and the transition from B to C will be entirely missed (Figure 3.2, green boxes). As a result of the misclassification, M_{AB} will be under-estimated, while M_{AC} will be over-estimated. These mis-estimations

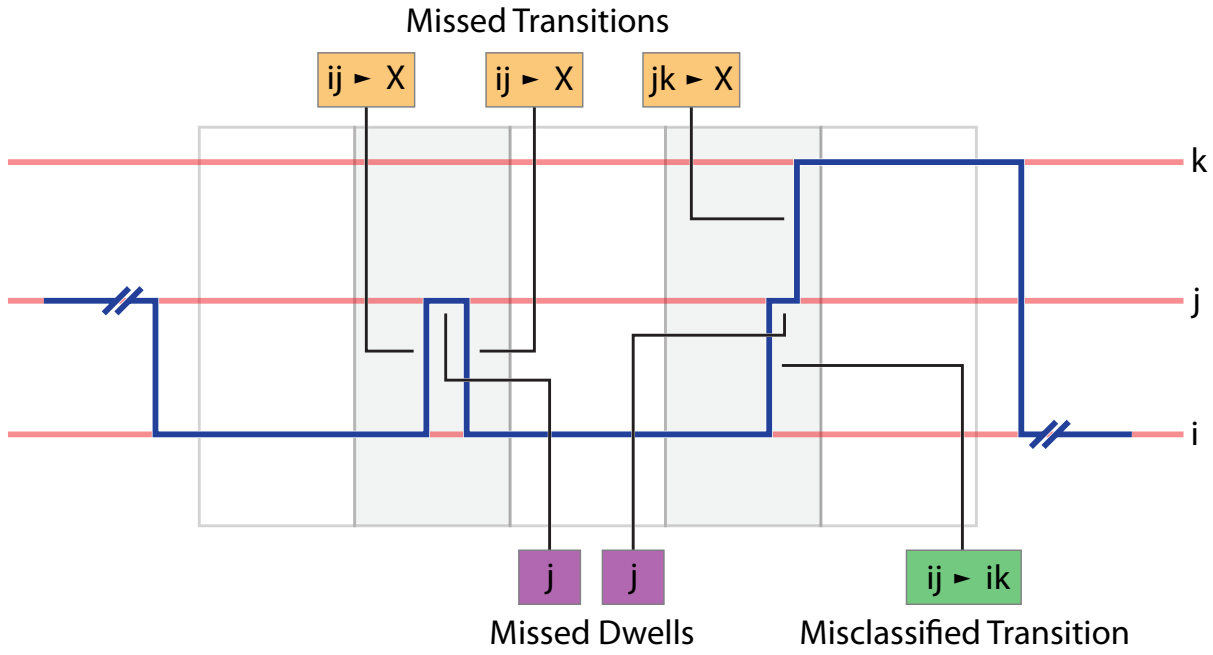

Figure 3.2: Types of Missed Events. Part of a single-molecule's path through state-space is shown in blue, which transitions between three states (i, j, and k) shown in red. Measurement periods over which signal time averaging is performed are shown as alternating white and grey boxes. Missed transitions are shown in orange, missed dwells are shown pink, and misclassified transitions are shown in green.

result in an under-estimation of k_{AB}, and an over-estimation of k_{AC}.

Missed Dwells

In the example of the missed transition from A to B given in above in Section 3.3.1, there was a sub-measurement period dwell-time in state B. This transient dwell induced the missed transition, because it was so short that the time spent in the B was not not registered in the state versus time trajectory (Figure 3.2, pink boxes). This failure to register the time spent in B in the state versus time trajectory is called a missed dwell. While the missed dwell is closely related to the missed transitions (it is causal), it and its effects are conceptually distinct from a missed transition. The missed dwell in B yields an under-estimation of M_{BB}, and consequentially an over-estimation of k_{BX}, where X stands for any state accessible from B. However, it also can provide drastic overestimates of the entries in n_{AX}, which as shown below, can seemingly distort otherwise normal Markovian behavior.

3.3.2 Correcting Rate Constants for Finite Length

Biomolecular systems may undergo transitions between states that are very long lived relative to the finite length of a signal versus time trajectory; reasons for this finite length were discussed in Section 3.3.1. For example, in an smFRET experiment, signal loss due to fluorophore photobleaching can occur before a transition occurs. In such a case, the entire state versus time trajectory is discarded, and not included in any subsequent analysis. This is because the arbitrary experimental end time of the signal versus time trajectory truncated the last (and only) dwell-time, and it is therefore unclear to which n_{ij} the dwell-time belongs. As a result, such long-lived dwell-times are systematically excluded from the analysis, and this can result in a mis-estimated counting matrix, $\boldsymbol{M}$, but, more often, it reduces the amount of data in $\boldsymbol{M}$ to a point where any subsequent calculation of a rate constant will be extremely imprecise (*c.f.*, Section 3.2.5).

Fortunately, there is a straight-forward correction that can be employed to correct for this loss of excluded data, which relies on a control experiment. Consider a smFRET experiment on a reversible two-state system, $A \rightleftharpoons B$. For the isolated system, the transitions between A and B, might be frequent enough that many transitions are observed before photobleaching of the fluorophore ends the signal versus time trajectory. However, in the presence of a certain ligand that significantly stabilizes A, the individual molecules might primarily begin and remain in A until photobleaching occurs; a rare few may sample B prior to photobleaching. As such, many of the measurement periods recorded in the presence of the ligand would not be analyzed, and M_{AA} would be significantly under-estimated. We address this under-estimation by calculating the number of signal versus time trajectories that photobleach before any transitions occur, N_{pb}, and by measuring the average photobleaching time of all of the signal versus time trajectories, T_{pb}. Assuming that all of the single-molecules in N_{pb} are capable of undergoing the transition, we augment $\boldsymbol{M}$ such that,

$$\boldsymbol{M}_{AA} = \boldsymbol{M}_{AA,\text{initial}} + N_{\mathrm{pb}} \cdot T_{\mathrm{pb}}, \tag{3.24}$$

where $M_{AA,initial}$ is the value of M_{AA} calculated initially as in Section 3.2. This correction should then provide a more reasonable estimate of M, and, therefore, of the rate constants. We note that this correction is easily modified to additionally address many considerations (*e.g.*, inactive sub-populations), as it simply entails augmenting the counting matrix by expected contributions as measured by control experiments.

3.3.3 Correcting Rate Constants Through Virtual States

One well-characterized method to correct for the effects of missed dwells, missed transitions, and misclassified transitions upon the calculation of rate constants is through the augmentation of the kinetic mechanism with 'virtual states' [22]. This method originated in the field of single-molecule conductance measurements on ion-channels, where workers such as Colquhoun and Hawkes pioneered the use of HMMs to analyze the stochastic kinetics of individual ion-channel opening and closing events [16]. The general approach of this method to correct rate constants for the missed events is to attempt to consider the number of expected missed dwells in a particular state. These expected missed dwells are then re-classifying into 'virtual states.' These virtual states then account for the missed dwells, and therefore do not contaminate the observed dwells. While this method was developed in Ref. 22, and reviewed in Refs. 16 and 23, we explore it here for completeness, as well as use it to discuss certain quantities that will be employed in Section 3.3.4.

Imagine that there is some 'cutoff time', τ_c, for which a dwell-time less than τ_c would become a missed dwell in a state-versus time trajectory. However, τ_c is more related to the threshold between two states in a signal-versus time trajectory, than to a particularly definitive dwell-time in a state. For instance, if one is assigning states in a state versus time trajectory based upon the crossing of a threshold, then τ_c is the amount of time in a state that yields a time-averaged signal (see Section 4) that crosses the threshold. Additionally, τ_c is affected by the noise and any particulars of the recording equipment. However, how to exactly determine τ_c remains an open question [22, 23]. For some insight into this difficultly, consider the asynchronicity of the stochastic transitions between states and the start of a measurement period. We note that for a dwell-time of length $t = \tau$, a single-molecule will occupy the state at least one half of and at most all of some measurement period (where the exact amount depends upon when the transition occurred and when the measurement began). Regardless, for a linear-signal with an evenly-spaced threshold, these dwell-times of length τ would then time average the signal past the threshold in some measurement period (either the one where the transition occurred, or the neighboring one), because they would all be at least half of the signal distance between states. However, even ignoring the feasibility of perfectly determining an appropriate threshold, for a dwell-time of length $\tau/2 < t < \tau$, only some of these would be registered due to the stochasticisity of the transition. Any static value of τ_c would then only exclude some of these dwell-times, whereas others would not produce missed events. Regardless, τ_c should be between $\tau_c \in (0, \tau)$.

To perform the rate constant correction, first, we will define notation for later use, then we will recast the two-state 'virtual state' correction in these terms, and finally generalize to multiple states. Consider a

single-molecule system where measurements are made with a time period, τ. For a particular observed dwell in one state, the following dwell in the next state will either be a missed dwell or an observed dwell if it is of length $t < \tau_c$ or $t > \tau_c$, respectively. For a Markovian system, the fraction of dwells in the first state that are less than τ_c, and greater than τ_c are

$$\begin{aligned} f_{ij}^{-} &= 1 - e^{-k_{ij}\tau_c}, \text{ and} \\ f_{ij}^{+} &= e^{-k_{ij}\tau_c}, \end{aligned} \tag{3.25}$$

respectively, where the subscript ij denotes the transition was from state i to state j, and $f^- + f^+ = 1$. Similarly, the fractions of dwells in the next state, regardless of whether the initial state dwell-time was an f_{ij}^- or f_{ij}^+ are

$$\begin{aligned} g_{jk}^{-} &= 1 - e^{-k_{jk}\tau_c}, \text{ and} \\ g_{jk}^{+} &= e^{-k_{jk}\tau_c}, \end{aligned} \tag{3.26}$$

where g is used instead of f to denote the fraction is for the successive dwell-time, and the subscript jk denotes the transition was from state j to state k. Again, $g^- + g^+ = 1$. For subsequent dwells, we switch back to f rather than use h, so that a series of dwells is described with alternating f and g. To demonstrate this notation, consider some transition between states A and B with a dwell-time that is long enough to be observed, followed by a transition from state B to state C with a dwell-time that is too short to be observed, followed by another transitions from state C to an unspecified state X with a dwell-time that is long enough to be observed. The probability of this series of events occurring can then be written $f_{AB}^{+} \cdot g_{BC}^{-} \cdot f_{CX}^{+}$.

Consider the two-state system, $1 \rightleftharpoons 2$, with forward and reverse rate constants k_{12} and k_{21} with some τ_c. These rate constants can be interpreted as the number of transitions that occur per time unit, so we will use rates constants and expected numbers of transitions interchangeably in this section; note, this is in line with the Poisson distribution. The true number of transitions can be split into those that are observed transitions, and those that are missed transitons. Therefore, the observed transition rate is then

$$\begin{aligned} k_{\text{corrected}} &= k_{\text{observed}} + k_{\text{virtual}} \\ &= k_{\text{observed}} + f_{\text{missed}} \cdot k_{\text{corrected}} \end{aligned} \tag{3.27}$$

where k_{virtual} is the contribution from the virtual state. That contribution is the fraction of missed transitions, f_{missed}, times the corrected number of transitions. For the two-state system, $f_{\text{missed}} = g^-$, because a missed dwell in the successive state causes a missed transition in an initial state. Therefore,

$$k_{12,\text{corrected}} = \frac{k_{12,\text{observed}}}{1 - g_{21}^-} = k_{12,\text{observed}} \cdot e^{k_{21}\tau_c}, \text{ and}$$
$$k_{21,\text{corrected}} = \frac{k_{21,\text{observed}}}{1 - g_{12}^-} = k_{21,\text{observed}} \cdot e^{k_{12}\tau_c}. \tag{3.28}$$

As pointed out by Stigler *et al.*, these equations are non-linear, so the corrected rate constants can be calculated easily by numerically minimizing the sum of squares of these equations [23]. To demonstrate the efficacy of this correction, we simulated a two-state system using the stochastic simulation algorithm [11, 12] across values of k_1 and k_2 that ranged between two decade less than and one decade greater than the acquisition rate τ^{-1}. At each pair of rate constants, 30 signal versus time trajectories were simulated, each containing approximately 1000 transitions. The observed rate constant was then taken to be the average of these rate constants as calculated using transition probability expansion analysis thresholded exactly half-way between the emission means of the two states. The observed rate constants relative to the exact, simulated values are shown in Figure 3.3A. These rate constants were then corrected using Equation (3.28) where $\tau_c = \tau/2$, and those are shown relative to the exact rate constant in Figure 3.3B. Notably, the observed rate constants begin to become inaccurate as the rate constants approach one-tenth of the acquisition rate, τ^{-1}. While the correction increases the region over which rate constants can be accurately calculated, many pairs of rate constants do not have solutions, this method is still inaccurate as the rate constants approach τ^{-1}, and as noted in Ref. 22, there are two solutions (fast and slow), so it can be difficult to pick the proper solution.

To generalize this virtual state approach to multiple-state systems, note that, as above, the corrected rate constant can be decomposed into contributions from the observed rate constant and the virtual state, but that, for systems with more than two states, the observed rate constant is overestimated by any misclassified transitions. Therefore,

$$k_{\text{corrected}} = (k_{\text{observed}} - k_{\text{misclassified}}) + k_{\text{virtual}}. \tag{3.29}$$

While $k_{\text{corrected}}$ is to be calculated by using Equation 3.29, and k_{observed} is obtained from experimental data by using the methods discussed in Section 3.2, expression must be derived for $k_{\text{misclassified}}$ and k_{virtual}. As

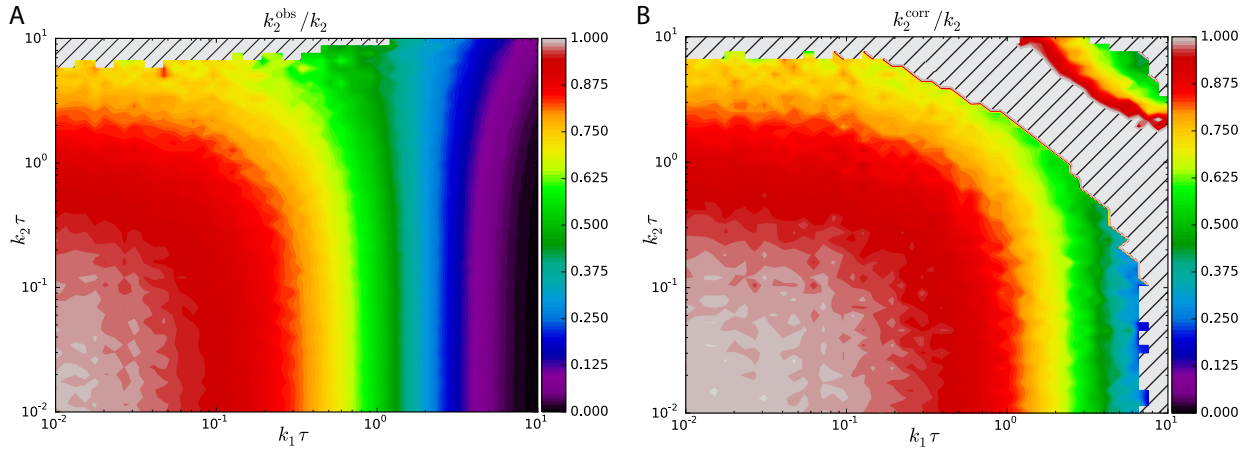

Figure 3.3: Correcting Rate Constants using Virtual States. (A) Observed rate constants relative to the exact rate constant calculated from simulated data for a two-state system. Hatched region contains undefined rate constants where no transitions could be observed in the idealized state versus time trajectory. (B) By applying the virtual state correction for a two-state system as given in Equation (3.28) using $\tau_c = \tau/2$, the corrected rate constants are closer to the exact rate constants over a larger area than was true for the observed rate constants in panel A. The hatched region has expanded to include cases where the solution to the equations used for the correction yield non-supported values.

described in Section 3.3.1, the most basic consideration of a misclassified transition is that in which only one missed dwell is considered. In this case, a missed dwell in an intermediate state makes the state versus time trajectory seem to transition between the previous and subsequent state. For example, in the case of $f_{AB}^{\pm} \cdot g_{BC}^{-}$, where the $\pm$ denotes either a lifetime either shorter or longer than τ_c, the transition from states A to B is misclassified as a transition from states A to C (the transition from states B to C is a missed transition). Note that $f_{AB}^{+} \cdot g_{BA}^{-}$ does not result in a misclassified transition and a missed transition, but rather two missed transitions. With this in mind, the number of misclassified transitions from state i to state j is

$$N_{ij,\text{misclassified}} = \sum_{l \neq i \neq j} N_{ilj}, \text{ where} \tag{3.30}$$

$$\begin{aligned} N_{ilj} &= \left(g_{lj}^{-} \cdot N_{ij,\text{corrected}}\right) \cdot P_{lj}^{\text{splitting}} \\ &= \left(g_{lj}^{-} \cdot N_{ij,\text{corrected}}\right) \cdot \left(\frac{k_{lj,\text{corrected}}}{\sum\limits_{j \neq l} k_{lj,\text{corrected}}}\right). \end{aligned} \tag{3.31}$$

From Equation 3.31, we see that the number of transitions that are misclassified from state i to state l to transitions from states i to j is equal to the product number of transitions from states i to j that can be misclassified regardless of the state they are misclassified into, and splitting probability, which is the probability that a single-molecule already in state l transitions to state j. Then, from Equation 3.30, the total number

of transitions misclassified as a transition from states i to j, is calculated as the sum of all of the possible misclassifications. As a result,

$$k_{ij,\text{misclassified}} = \sum_{l \neq i \neq j} \left(\left((1 - e^{-k_{lj}\tau_c}) \cdot k_{ij,\text{corrected}} \right) \cdot \left(\frac{k_{lj,\text{corrected}}}{\sum\limits_{j \neq l} k_{lj,\text{corrected}}} \right) \right). \tag{3.32}$$

Finally, for a system with more than two states, k_{virtual} is involves all the possible transitions that can occur from a particular state. Since the system is Markovian, despite the presence of parallel pathways, the number of transitions per unit time from state i to state j is k_{ij}. Of these transitions, the number of transitions that end up in a virtual state depend upon the length of the dwell time in the subsequent state j. As mentioned in Section 3.2.2, the stochastic rate constant that governs the dwell-times in state j is $k_j = \sum\limits_i k_{ji}$. Only those dwell-times in state j shorter than τ_c will induce a transition from i to a virtual state. Therefore

$$k_{ij,\text{virtual}} = g_j^- k_{ij,\text{corrected}} = (1 - e^{-k_j\tau_c}) k_{ij,\text{corrected}}. \tag{3.33}$$

Therefore, by substituting Equations 3.32 and 3.33 into equation 3.29, an expression for the corrected rate constant from states i to j can be found as

$$k_{ij,\text{corr}} = \left(k_{ij,\text{obs}} - \sum_{l \neq i \neq j} \left((1 - e^{-k_{lj}\tau_c}) \cdot k_{ij,\text{corr}} \frac{k_{lj,\text{corr}}}{\sum\limits_{j \neq l} k_{lj,\text{corr}}} \right) \right) + (1 - e^{-k_j\tau_c}) k_{ij,\text{corr}}, \tag{3.34}$$

and this can be repeated for all transitions a the kinetic scheme to yield a set of non-linear equations that can be sovled numerically, as mentioned above, to yield the corrected rate constants.

3.3.4 Seemingly Non-Markovian Behavior Induced by Missed Events

While we have described how to partially account for missed events when calculating rate constants from signal versus time trajectories, we note that the assumptions used to both calculate the observed rate constant and to correct the observed rate constant rely on the system being Markovian. Experimentally, many single-molecule systems seem to exhibit non-Markovian behavior [24, 25], and this is typically assessed, if at all, by checking to see that the dwell-times observed in a particular state are exponentially distributed. Non-single-exponential behavior is evidence for non-Markovian behavior. Again, all of the methods described above that directly address stochastic rate constants assume Markovian behavior, and should not be ap-

plied in the case of non-Markovian behavior. Additionally, it is worth noting that model selection for HMMs depends upon this assumption as well [4, 7]. With these limitations in mind, here, we demonstrate that one particularly nefarious consequence of missed events in an otherwise Markovian state versus time trajectory is the introduction of seemingly non-Markovian behavior.

To demonstrate the introduction of seemingly non-Markovian behavior into a Markovian system, consider a single-molecule experiment that is performed on a two-state, Markovian system. This system can be represented as $1 \rightleftharpoons 2$, with forward and reverse rate constants k_1 and k_2, respectively. If one rate constant is relatively fast compared to the acquisition rate, there will be many missed dwells in that state. To be concrete, one such system might be that where $k_1 = 0.5$ s^{-1}, $k_2 = 10$ s^{-1}, and $\tau = 0.1$ s; here k_2 is equal to the acquisition rate while k_1 is 20 times slower, and we expect that state 2 will have many missed dwells. The missed events can be readily observed in the signal versus time trajectory plotted in Figure 3.4.

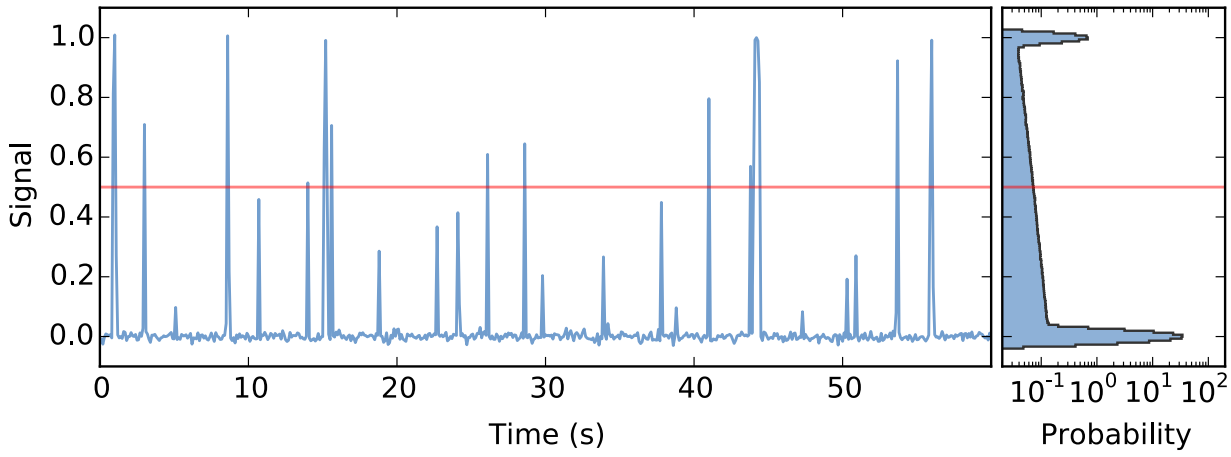

Figure 3.4: Signal from a Markovian, two-state system. The first 60 s of a signal versus time trajectory from a simulated two-state, Markovian system. This system was simulated for 2.5×10^6 s, then the signal was time averaged with signal means of 0 and 1 for states 1 and 2, respectively, and then negligible Gaussian noise was added for visibility in the histogram (right). The red line denotes the threshold used to idealize the data into a state versus time trajectory. Many dwells in the upper state are so transient that they result in missed dwells and missed transitions.

After idealizing this signal versus time trajectory into a state versus time trajectory, perhaps by using a threshold, the observed length of each dwell is used to calculate the rate constants. While the observed length of a dwell-time in the state versus time trajectory depends upon the true length of the dwell-time in question, but also upon the true lengths of previous and subsequent dwell-times. This is evident by considering the effect that a missed dwell has upon the state versus time trajectory. Consider a dwell-time in state 1 that is longer than the measurement period τ. Using the notation in Equation 3.25, the probability of this dwell-time occurring is $f^+ \equiv f_{12}^+$. The probability that the subsequent dwell-time in state 2 is also

longer than the measurement period τ is $g^+ \equiv g_{21}^+$, in which case there will be no missed dwell in state 2 to complicate the measurement of the observed dwell-time in state 1. Therefore, the probability of observing a dwell-time in state 1 without the complicating aspects of a missed dwell in state 2 is $f^+ \cdot g^+$. An example path through state-space of this case is shown in Figure 3.5A. However, this suggests that if there is a missed dwell in state 2, which occurs with probability $g^- \equiv g_{21}^-$, then the measurement of the observed dwell-time in state 1 will be more complicated. With each missed dwell in the transiently occupied state 2, the previous and subsequent dwell-times in state 1 are compounded together to create an overly long observed dwell-time in the state versus time trajectory. These overly long dwell-times can be composites of two, three, four, or higher integer-numbers of dwells in state 1; where the exact number is one more than the number of missed events in state 2. The case of one missed dwell, $f^+ \cdot g^- \cdot f^+$, is shown in Figure 3.5B, and two missed dwells, $f^+ \cdot g^- \cdot f^+ \cdot g^- \cdot f^+$, is shown in Figure 3.5C. Further compounding of dwell-times will occur until a sufficiently long dwell-time in state 2 occurs, g^+, at which point the end of the observed dwell in state 1, which is composed of the several compound dwell-times is indicated in the state versus time trajectory. In this system, each observed dwell in state 1 is concluded by a g^+ dwell (excepting so many g^- and f^- in one measurement that the time-averaged signal is past the threshold). Finally, note that all of this compounding-phenomenon also occurs for the dwell-times in state 2, however since k_1 is so slow, there are rarely ever any missed dwells in state 1 (*i.e.*, f^- is too small for those types of dwell-times to appreciably occur).

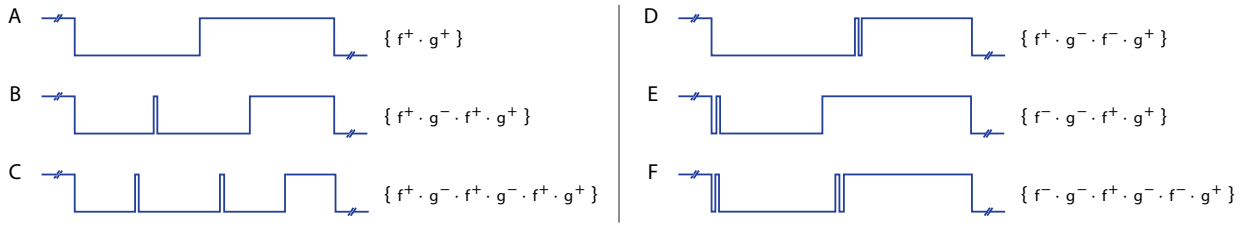

Figure 3.5: State-space Pathways that Lead to Seemingly Non-Markovian Behavior. (A-F) Schematic plots of Markovian pathways through a two-state state-space. Each sequence is also written in using the notation described in Equations 3.25 and 3.26, where f and g correspond to the first and second states, respectively, and $-$ and $+$ denote dwell-times that are less than or greater than τ_c, respectively.

Those observed dwell-times that are actually several compounded dwell-times introduce seemingly non-Markovian behavior into the state versus time trajectory. This is apparent when inspecting the histogram of the lengths of the observed dwells in the state versus time trajectory (Figure 3.6). If the system is Markovian, these discrete dwell-times should be distributed according to the geometric distribution as described in Section 3.2.2. However, the geometric distribution does not adequately describe the distribution of these observed dwell-times, especially for the dwell-times in state 1 (Figure 3.6, left). Notably, the geometric dis-

tribution is much more adequate for the observed-dwell times is state 2, which is expected because there many fewer compound dwell-times in this state. Furthermore, conditioning the geometric distribution such that only dwell-times greater than one frame long are considered describes the observed dwell-times in state 2 rather well, but not those in state 1 (Figure 3.6, center). Here, rather than condition upon all dwell-times greater than zero frames long as was done for Equation 3.12, conditioning upon all dwell-times greater than one frame long minimizes the effects of over- and under-estimation due to thresholding and discretization. Regardless, as is visible for state 1, the Markovian behavior used to simulate this two-state system yields behavior that is markedly non-Markovian. However, this effect can be accounted for by augmenting the model of the observed dwell-times to include those that are composed of several compounded dwell-times.

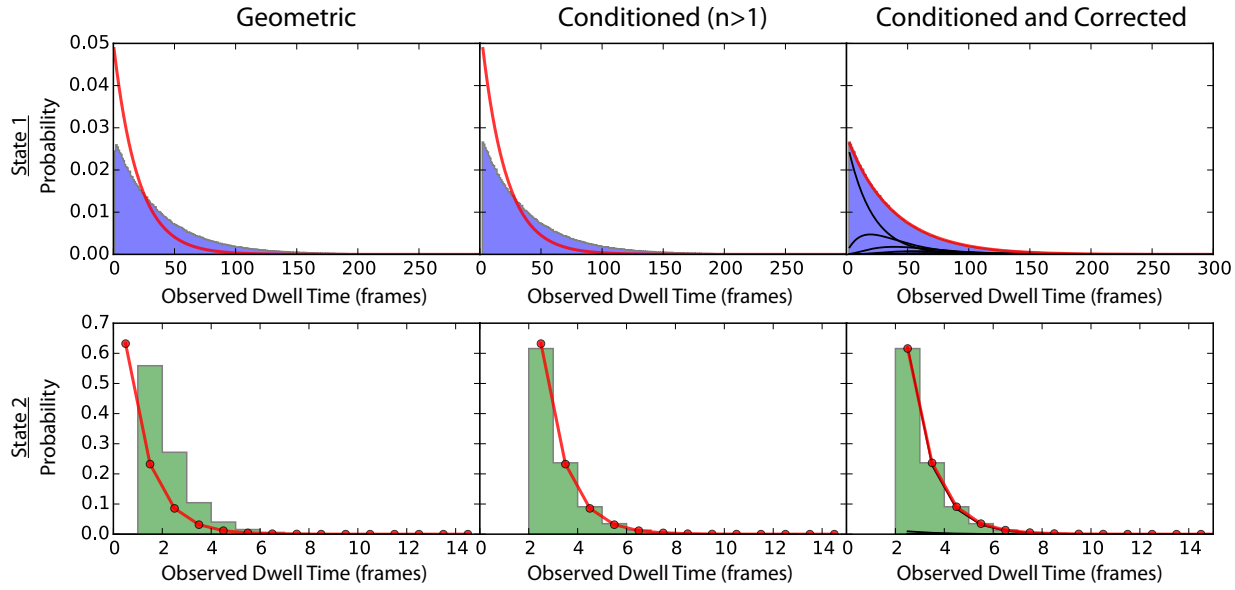

Figure 3.6: Dwell-time Correction for Missed Dwells. Histograms of the probabilities for the observed dwell-times from the simulated two-state system shown in Figure 3.4 for state 1 (top), and state 2 (bottom). The red-curves show various PMFs plotted using the exact, simulation parameters. The 'Geometric' column corresponds to the geometric distribution with probability from Equation 3.6 as is described in Section 3.2.2 for the PMF. The 'Conditioned $(n > 1)$' column corresponds to the geometric distribution conditioned upon all the number of trials being greater than one as derived in manner similar to that used for Equation 3.12 for the PMF. The 'Conditioned and Corrected' column corresponds to a PMF that is a mixture of negative binomial distributions all conditioned upon the number of trials being greater than one, and weighted according to the probability of the different contributing pathways with $\tau_c = 0.75\tau$.

To augment the PMF of the observed dwell-times for the effects of compounded dwell-times, we must have a probabilistic description of the distributions of the total lengths of the various, integer-number-compounded dwell-times. As described previously in Section 3.2.2, for one dwell, this probabilistic description of the total length is the geometric distribution. In the case of two compounded dwell-times, the probabilistic description will be the convolution of the two, identical geometric distribution (one for each dwell), because we are calculating the sum of the dwell-times, which are stochastic variables [20]. For three compounded dwell-times,

the probabilistic description will be the convolution of three geometric distributions, and so on. These convolutions yield a compound distribution for $X = X_1 + ... + X_r$, where X_i is the randomly distributed length of one dwell-time, and r is the number of dwell-times (or geometric distributions) being convoluted together. As described in Ref. 20, this convolution is easily performed through multiplication of the geometric distribution generating function,

$$P_{X_i}(s) = p/(1-(1-p)s), \tag{3.35}$$

which yields the compound distribution's generating function

$$P_X(s) = \prod_{i=1}^{r} P_{X_i}(s), \tag{3.36}$$

which can be expanded as a power series and rewritten using the binomial theorem to yield the compound distribution's PMF of

$$\mathrm{NB}(k|r,p) \equiv P[X=k] = (-1)^k \binom{-r}{k} p^r (1-p)^k, \tag{3.37}$$

where NB is the PMF of the negative binomial distribution, k is the number of frames in the observed dwell-time, r is the number compounded dwells $r \in \{1, 2, ...\}$, and p is the transition probability from Equation 3.6. Essentially, each of the observed dwell-times can be modeled using the appropriate negative binomial distribution for the total number of compounded dwell-times–even if there is only one dwell-time (*i.e.*, the geometric distribution is the negative binomial distribution with $r = 1$). From these expressions, and because a Markovian system is memoryless, we see that the PMF of all of the observed dwell-times is a mixture of negative binomial distributions each weighted by an appropriate factor as

$$\mathrm{PMF}_{12}^{\mathrm{corr}}(n^\star|k_1,k_2,\tau,\tau_c) = \sum_{r=1}^{\infty} p_r(k_1,k_2,\tau,\tau_c) \cdot \mathrm{NB}^\star(n^\star|r, 1-e^{-k_1\tau}, n^\star \in \{2,3,4,...\}), \tag{3.38}$$

where $p_r(k_1, k_2, \tau, \tau_c)$ is the probability of observing a dwell-time that is composed of r compounded dwells, and we have used $n^\star$ and $\mathrm{NB}^\star$ to denote the conditioning of n and NB upon $n > 1$, which as mentioned in Section 3.2.2, is a linear correction. However, this explicitly highlights the fact that we have been ignoring any contributions to the total length of the observed dwell-times from any missed dwells, because they are inherently negligible. Therefore, it is important to acknowledge that only those dwell-times that are longer

than τ_c (*i.e.*, f^+) will contribute to the length of a compounded dwell-time. Consequentially, to proceed with this augmented PMF and correct for the compound dwell-times induced by missed dwells, we must calculate the probability that an observed dwell-time will consist of r f^+'s.

To calculate the probability that an observed dwell-time will consist of r dwells, every possibly stochastic path through state-space must be tabulated in order to find how many of these paths yield the desired number of f^+ compoundings–this is impractical for anything more than a two-state system. To do so, consider the possible sequences of dwells following an initial dwell in state 1. These can be denoted using the notation introduced in Equations 3.25 and 3.26 as

$$\left\{f^{\pm} \cdot g^{\pm} \cdot f^{\pm} \cdot g^{\pm} \cdot ...\right\} = \left\{\prod_{i=0}^{\infty} f^{\pm} \cdot g^{\pm}\right\}, \tag{3.39}$$

where the length of each subsequent dwell time relative to τ_c is readily assessed, and additionally the probability of each sequence is simply the product of the terms in the sequence. For instance, as is the case in Figure 3.5A, a long dwell in state 1 followed by a long dwell in state 2 can be denoted $\{f^+ \cdot g^+\}$. Interestingly, for an initial dwell of $f^{\pm}$, the observed dwell-time will continue to compound dwells in both states until a g^+ occurs, at which point, the state versus time trajectory will indicate the end of that observed dwell. Therefore, for this two-state case, the notation on the right-hand side of Equation 3.39 is informative in that each dwell in state 1 is accompanied by another dwell that is in state 2; even more informatively, these sequences can be written

$$\left\{\left(\prod_{i=0}^{\infty} f^{\pm} \cdot g^{\pm}\right) \cdot f^{\pm} \cdot g^{+}\right\}, \tag{3.40}$$

where for our purposes, there must be at least one f^+ in order for the sequence to correspond to a particular negative binomial distribution. One way to think of the possible sequences is that they are composed of f^+'s with interspersed f^- and g^- continuing until the final g^+. This approach yields two categories of neighboring transitions that can be used as building blocks to extend each sequence until the final g^+: $f^- \cdot g^-$ (Figure 3.5D,E), and $f^+ \cdot g^-$ (Figure 3.5B,C). Multiple blocks can be combined in any order to create and extend unique stochastic pathways (Figure 3.5). Alternatively, either $f^+ \cdot g^+$ or $f^- \cdot g^+$ will effectively terminate an observed dwell. Therefore, to calculate the probability that there are r number of f^+ dwell-times in an observed dwell that begins in state 1 before the first g^+, we must write out the possible sequences, use this to calculate the probability for each sequence, and then apply the above rules to classify the number of

compounded f^+ dwells that comprise the observed dwell. Fortunately, since the values of f^- and g^- are relatively small, these sequences do not need to be tabulated to infinity, as they become vanishingly small with additional terms. Our approach is to choose an arbitrary length of sequences to calculate such that the inclusion of further dwells does not significantly alter the probability of observing r dwells in an observed dwell; here we used $i = 6$ (12 dwells). These are easily enumerated numerically using quaternary indexing for the possible pairs of $f^{\pm} \cdot g^{\pm}$ to yield $4^6 = 4096$ possible paths. Pathways without an f^+ or a g^+ are excluded, as are those where all f^+ present are located after the first g^+. Then, the remaining pathways are categorized according to the number of f^+ before the first g^+ occurs. The product of each sequence in a particular category is then summed to yield the probability of observing that number of compound dwell times as

$$p_r(k_1, k_2, \tau, \tau_c) = \sum_j \delta(r - \sum_{f^+ \in j} 1) \cdot \prod \left(\left\{ \left(\prod_{i=0}^{\infty} f^{\pm} \cdot g^{\pm} \right) f^{\pm} \cdot g^+ \right\}_j \right), \tag{3.41}$$

where δ is a Dirac-delta function used as an indicator function so that only pathways with the proper number of f^+ are included in p_r, and j is used to index the possible pathways. Python code to perform this calculation is provided in Appendix D.

By explicitly addressing the presence of the missed-dwell-induced compound dwell-times explicitly in the PMF of the observed dwell-times present in a state versus time trajectory by using Equations 3.38 and 3.41, we can accurately account the seemingly non-Markovian effects found in the state versus time trajectory of a Markovian system affected by missed events. This can be seen in the plots in Figure 3.6 on the right, where the PMF (red) matches the histogram of the simulated dwell times. This curve was not a fit to the data, but rather the PMF calculated using the exact simulated values. As such, the fast decay and long tail found in the histogram of the observed dwell times in state 1 are still Markovian features, despite such non-exponentiality typically being seemingly non-Markovian in nature. This explanation for the effects of the missed dwells therefore not only questions the accuracy of utilizing Markovian-based assumptions to calculate rate constants for a Markovian system where many events are missed, but also raises questions about the accuracy of the model selection performed using Bayesian HMMs which also use Markovian assumptions to determine which model has the most evidence. Regardless, despite the difficultly in determining an appropriate τ_c, and the unaccounted-for effects of the f^--type dwells in the PMF that render this method difficult to employ to calculate rate constants, it provides a simple and accurate measure for whether seemingly non-Markovian behavior in a state-versus time trajectory is actually Markovian or not. However, in order to

truly calculate such rate constants accurately, despite any number of missed events, another approach must be developed. Such an approach would effectively enable temporal super-resolution of the data collected from any single-molecule technique (*c.f.*, Section 4).

3.4 Quantitatively Connecting Ensemble Thermodynamics and Single-Molecule Kinetics

At equilibrium, thermodynamic and kinetic information can be extracted from biomolecular energy landscapes by many techniques. However, while static, ensemble techniques yield thermodynamic data, often only dynamic, single-molecule techniques can yield the kinetic data that describes transition-state energy barriers. Here we present a generalized framework based upon dwell-time distributions that can be used to connect such static, ensemble techniques with dynamic, single-molecule techniques, and thus characterize energy landscapes to greater resolutions. We demonstrate the utility of this framework by applying it to cryogenic electron microscopy and single-molecule fluorescence resonance energy transfer studies of the bacterial ribosomal pretranslocation complex. Among other benefits, application of this framework to these data explains why two transient, intermediate conformations of the pretranslocation complex, which are observed in a cryogenic electron microscopy study, may not be observed in several single-molecule fluorescence resonance energy transfer studies.

Biomolecular machines operate on energy landscapes with transition-state energy barriers which range from $\sim k_B T$ to the energy of covalent bonds [26–28]. Characterizing the wells and barriers which comprise these energy landscapes is important for understanding the thermodynamics and kinetics of biomolecular machines, and how these thermodynamics and kinetics can be modulated in order to regulate the activities of these machines [29–34]. However, due to the stochasticity inherent to these processes, as well as the transient and/or rare nature of states separated by low transition-state energy barriers, extremely sensitive techniques are often required to obtain the level of detail necessary to adequately describe such systems [1, 35]. Sufficiently sensitive ensemble techniques, such as cryo-electron microscopy (cryo-EM), can measure static, equilibrium-state populations, and this provides information on the relative energy differences between distinct states. However, because of the vanishingly small probability of observing a transition state, techniques such as cryo-EM are not able to characterize the transition-state energy barriers responsible for much of the regulation of biomolecular processes [36]. Fortunately, dynamic, time-dependent, single-molecule techniques, such as single-molecule fluorescence resonance energy transfer (smFRET),

can directly monitor the kinetics of these processes, and allow the characterization of the transition-state energy barriers with theories such as transition-state theory or Kramers' theory [15]. smFRET is a particularly powerful technique for connecting single-molecule kinetics to ensemble thermodynamics obtained from cryo-EM in that the FRET efficiency (E_{FRET}) obtained from the smFRET experiments can be correlated to structures obtained from the cryo-EM experiments. Despite this significant advantage over many other single-molecule techniques, like all techniques, smFRET approaches often suffer from limitations to spatial and temporal resolution, and also often require structural information to develop biologically informative signals [37]. Therefore, for any particular system, static, equilibrium-state, ensemble techniques and dynamic, time-dependent, single-molecule techniques provide complementary approaches for studying the underlying biological processes. Nonetheless, given the current limitations in their application, the pictures they provide may not always be congruous.

Here, we present such a framework based upon equilibrium-state probabilities and dwell-time distributions (see Refs. 38 and 39, and references therein). This framework is general, and can be applied to various other ensemble and single-molecule techniques; we use a linear kinetic model, but emphasize that the equations can also be derived for other models. As an illustrative case study, we apply this generalized framework to analyze the data obtained from the cryo-EM and smFRET studies of Agirrezabala *et al.* and Fei *et al.*, respectively. In doing so, we connect the distribution of the MS I, MS II, IS1, and IS2 states of the PRE complex observed by cryo-EM to the transition rates observed between the GS1 and GS2 states observed by smFRET.

3.4.1 Dwell-Time Distribution Framework

Consider two distinct chemical states, 1 and N, of the system connected linearly by a number of on-pathway intermediate states, 2 through N-1, with transitions from state i to state j occurring at rate α_{ij},

$$1 \underset{\alpha_{21}}{\overset{\alpha_{12}}{\rightleftharpoons}} 2 \underset{\alpha_{32}}{\overset{\alpha_{23}}{\rightleftharpoons}} \cdots \underset{\alpha_{N-1,N-2}}{\overset{\alpha_{N-2,N-1}}{\rightleftharpoons}} N-1 \underset{\alpha_{N,N-1}}{\overset{\alpha_{N-1,N}}{\rightleftharpoons}} N. \tag{3.42}$$

If $P_\mu(t)$ is the probability of finding the system in chemical state μ at time t, then the time evolution of these probabilities are governed by a set of coupled master equations. The steady-state of this set of equations, corresponding to the the constraints $\partial P_\mu / \partial t = 0$, yields the equilibrium-state occupation probabilities P_μ^{eq} of populating each state. The distribution of times taken by the system to reach one terminus from the other can be calculated by modifying the original kinetic scheme imposing an absorbing boundary at the destination

state. If the destination state is N, then the corresponding modified kinetic scheme would be,

$$1 \underset{\alpha_{21}}{\overset{\alpha_{12}}{\rightleftharpoons}} 2 \underset{\alpha_{32}}{\overset{\alpha_{23}}{\rightleftharpoons}} \cdots \underset{\alpha_{N-1,N-2}}{\overset{\alpha_{N-2,N-1}}{\rightleftharpoons}} N-1 \xrightarrow{\alpha_{N-1,N}} N. \tag{3.43}$$

By writing master equations for this new scheme, and adopting the method of Laplace transform, we analytically calculated the probability, $f^p(t)dt$, that the system, initially at state 1, reaches state N in the time interval between t and $t + dt$. By evaluating the first moment of this distribution, the mean time for this transition, $\langle t_p \rangle$, can be calculated. An analogous process can be performed to calculate the probability, $f^r(t)dt$, that the system, initially in state N, reaches state 1 in the time interval between t and $t + dt$, and hence obtain the mean time for this transition, $\langle t_r \rangle$.

The two expressions for the mean transition times between the termini, $\langle t_p \rangle$ and $\langle t_r \rangle$, form a system of equations with all 2N-2 α_{ij} as variables. Ratios of the P_μ^{eq} define relationships between the rate constants α_{ij}; so, if the equilibrium-state probabilities are known, substitution of these ratios into the expressions for $\langle t_p \rangle$ and $\langle t_r \rangle$ reduces the number of degrees of freedom in the system of equations. With an experimental measure of the mean transition time between the terminal states, the system of equations can be solved for the α_{ij} rate constants. The four-state model (two intermediate states) is solved in Appendix A, as is the derivation of the expression for the variances of t_p and t_r. Equivalent expressions for the three-state model (one intermediate state) are in Appendix B.

Because of their complexity, biomolecular systems are often investigated with multiple techniques – each with their individual strengths and weaknesses. However, different techniques occasionally yield disparate mechanistic pictures that must ultimately be resolved. One situation in which this problem manifests itself is when a static, equilibrium-state technique such as cryo-EM detects on-pathway intermediates, but a dynamic, time-dependent technique such as smFRET does not. This situation could arise if the dynamic technique is not sensitive enough to distinguish the intermediate state from other states of the biomolecular system. In order to reconcile such contrasting measurements, we need to estimate the lifetime of the transient intermediates if these, indeed, exist. In an effort to get these estimates we consider a linear kinetic pathway with on-pathway intermediates, such as in equation 3.42, though the framework presented here can easily be extended to include off-pathway intermediates. Note that for an N-state linear kinetic scheme there are 2N-2 rate constants α_{ij}. Therefore, in principle, the numerical values of all the individual α_{ij} could be obtained if 2N-2 independent algebraic equations satisfied by these rate constants were available. As we argue now, except for some small values of N, the rate constants α_{ij} are usually underdetermined by the

available experimental information.

Time-dependent smFRET experiments are typically analyzed with a hidden Markov model (HMM) [4–7, 37]. Among other things, such an analysis yields HMM-idealized state versus time trajectories from which a distribution of lifetimes in a particular state can be calculated. If sufficiently transient, on-pathway intermediates between the initial and final state exist, the distribution of idealized lifetimes will not appear significantly different from what it would be in the absence of the intermediate states (*e.g.,* an exponential distribution for a random transition with a time-independent probability of occurrence). In such a case, the simplest model that the smFRET data supports is that of a transition with no intermediate states; so, assuming Markovian transitions, the mean lifetimes obtained from the HMM-idealized trajectories would be taken to be the inverses of the effective rate constants for the transitions between the initial and final states. In contrast, the analytical expressions for the mean lifetimes spent traveling between the terminal states, via intermediate states, of a N-state kinetic scheme contain the rate constants α_{ij} that describe the direct transitions between the intermediate states (see Appendices A and B). Therefore, equating the mean lifetimes for the forward and reverse transitions between the terminal states that were inferred from dynamic, single-molecule experiments with the corresponding theoretically calculated mean lifetimes yields two algebraic equations that involve 2N-2 rate constants α_{ij}.

Thus, in practice, except for the trivial case of N=2, the information available in the form of the effective rates of forward and reverse transitions between the two terminal states would be inadequate to determine all the 2N-2 rate constants α_{ij} that describe the full kinetic mechanism. Obviously, for larger values of N, the number of degrees of freedom must be reduced further by acquiring additional experimental information. This extra information comes from the equilibrium-state experiments. Including information about the equilibrium-state probabilities for the N states provides N-1 additional independent equations (the constraint of normalization of the probabilities, i.e., their sum must be equal to unity, reduces the number from N to N-1). So, for N=3 one would have just enough information to write down four independent equations satisfied by the four rate constants in the three-state model. However, for N=4, we have fewer equations than the number of unknowns and, therefore, in the absence of any other information, one of the rate constants would remain a free parameter. Any one of the six rate constants can be selected as the free parameter. Then, as we will show later in this section, varying the selected free parameter allows one to enumerate all the solutions which are consistent with the data, and thereby impose lower and/or upper bounds on the magnitudes of the rates. In case of higher values of N, the analytical expressions for the variance of t_p and t_r, reported in the Appendix, can be utilized for further reduction of the number of degrees of freedom if the

corresponding experimental data becomes available in the future.

3.4.2 Bacterial Pretranslocation Complexes as a Model System

Agirrezabala and coworkers collected cryo-EM data on PRE complexes containing tRNA^{fMet} in the P site, and fMet-Trp-tRNA^{Trp} in the A site [40]. Using ML3D, a maximum-likelihood based classification method, particles from this data set were more recently classified into six classes [41]. The conclusions of this study strongly suggest that three of the classes represent MS I and MS II–MS II being comprised of two structurally similar classes. Additionally, the authors propose that two of the other classes represent on-pathway intermediate states (IS1 and IS2, respectively) between MS I and MS II. As the remaining class represents PRE complexes that are missing a tRNA in the A site, it is therefore ignored. Thus, the model of PRE dynamics proposed by this study is,

$$1 \underset{\alpha_{21}}{\overset{\alpha_{12}}{\rightleftharpoons}} 2 \underset{\alpha_{32}}{\overset{\alpha_{23}}{\rightleftharpoons}} 3 \underset{\alpha_{43}}{\overset{\alpha_{34}}{\rightleftharpoons}} 4, \tag{3.44}$$

where states 1, 2, 3, and 4 represent MS I, IS1, IS2, and MS II, respectively, and are distributed as shown in table 3.1.

State Index (μ)	State	Class	P_μ^{eq}
1	MS I	2	0.231
2	IS1	4A	0.131
3	IS2	4B	0.140
4	MS II	5/6	0.498

Table 3.1: A summary of the ribosomal PRE complexes observed by Agirrezabala and coworkers [41].

Similarly, smFRET experiments performed by Fei and coworkers monitored PRE complexes as they transitioned, driven by thermal energy, between two global conformational states, GS1 and GS2, which correspond structurally to MS I and MS II, respectively [42, 43]. By monitoring the relative change in the distance between the P-site tRNA and the ribosomal protein L1 within the L1 stalk of the 50S subunit, the smFRET signal developed by Fei and coworkers probably reports upon the tRNA motions along the pathway proposed by Agirrezabala and coworkers. However, these smFRET measurements were performed for several PRE complexes of variable composition. Of these complexes, perhaps the most relevant complex to the work performed by Agirrezabala and coworkers is the PRE$_{fM/F}$ complex carrying a tRNA^{fMet} in the P site and a fMet-Phe-tRNA^{Phe}, rather than a fMet-Trp-tRNA^{Trp}, in the A site. Since the identity of the A-site

dipeptidyl-tRNA in the two experiments differs, this could potentially lead to tRNA-dependent differences in the populations and lifetimes of the various states and, consequently, the rates of transitions between these states. Indeed, smFRET studies have shown that the lifetimes of GS1 and GS2 do depend on the presence [42, 43] and acylation status (*i.e.*, deacylated $tRNA^{Phe}$ versus Phe-$tRNA^{Phe}$ versus fMet-Phe-$tRNA^{Phe}$) of the A-site dipeptdyl-tRNA [44, 45]. It should be noted, however, that the effect of the identity of the A-site dipeptidyl-tRNA itself (*i.e.*, tRNAs other than $tRNA^{Phe}$) has not yet been tested by smFRET. In addition to the difference in the identity of the A-site dipeptidyl-tRNA in the cryo-EM and smFRET studies, the Mg^{2+} concentrations employed in the two studies differ. The cryo-EM studies were performed at $\left[Mg^{2+}\right] = 3.5$ mM and the smFRET studies were performed at $\left[Mg^{2+}\right] = 15$ mM. Previously, smFRET studies have demonstrated that changes to the Mg^{2+} concentration over this range affects the populations and lifetimes of the GS1 and GS2 states [45]. With this in mind, it is likely that the equilibrium-state populations observed in the cryo-EM experiments and the corresponding state occupancies in the smFRET experiments are disparate. Nonetheless, despite their experimental differences, these cryo-EM and smFRET studies are the most experimentally similar cryo-EM and smFRET studies of wild-type bacterial PRE complexes that have been reported in the literature. Therefore, as a case study, we have chosen to quantitatively compare these two particular studies in order to demonstrate the application of the general framework developed in Section 3.4.1.

The transition rates between GS1 and GS2 reported using the $PRE_{fM/F}$ complex for the L1-tRNA donor-acceptor labeling scheme were $k_{GS1 \rightarrow GS2} = 2.8 \pm 0.2\ s^{-1}$ and $k_{GS2 \rightarrow GS1} = 3.0 \pm 0.4\ s^{-1}$ [43]. Given that no evidence of intermediate states was observed, this suggests that any intermediates states, if they exist, might be very transient relative to the time resolution with which the smFRET data was acquired. Indeed, there is a limitation to the time resolution with which smFRET data can be acquired with the electron-multiplying charge-coupled device (EMCCD) cameras that are typically used as detectors in TIRF microscopy-based smFRET experiments. Transitions that are faster than the EMCCD camera's acquisition rate (20 s^{-1} in Fei, et al.) [43] result in time averaging of the E_{FRET} and the recording of a single, artifactual data point that appears at the time-averaged value of the E_{FRET} between the states involved in the rapid fluctuations. This is a well-documented feature of smFRET data analysis, which we term "blurring" [4]. This effect is further compounded by the fact that current state-of-the-art computational methods used to analyze the smFRET data cannot distinguish between artificial, short-lived (*i.e.*, one data point) "states" resulting from blurring and actual, short-lived (*i.e.*, one data point) states resulting from the sampling of true intermediate states [4]. With such an analysis, the true molecular states become hidden among the "blurred" states.

Reanalysis of the original $PRE_{fM/F}$ data using the software-package ebFRET–a state-of-the-art, HMM-based analysis method for smFRET data [6, 7]–yields a better estimate of the transition rates. This is because ebFRET uniquely enables analysis of the entire ensemble of individual E_{FRET} versus time trajectories, instead of analyzing them in the traditional, isolated, one-by-one manner. The two-state rates inferred by means of ebFRET are similar to, though perhaps more accurate than, those reported originally by Fei and coworkers: $k_{GS1\rightarrow GS2} = 2.0 \pm 0.2\ s^{-1}$, and $k_{GS2\rightarrow GS1} = 2.8 \pm 0.1\ s^{-1}$ (see Table 3.2). Interestingly, application of ebFRET reveals that the smFRET $PRE_{fM/F}$ data are best described by a five-state model. However, further analysis indicates that the three additional states are probably artifacts of blurring, because they are negligibly populated, have extremely transient lifetimes, and occur at an E_{FRET} that is in between the E_{FRET} of the two well-defined states.

$k_{GS1\rightarrow GS2}$ (s^{-1})	$k_{GS2\rightarrow GS1}$ (s^{-1})	Reference
2.8 ± 0.2	3.0 ± 0.4	43
2.0 ± 0.2	2.8 ± 0.1	This work

Table 3.2: A summary of the rates of transition between $GS1$ and $GS2$ observed by Fei and coworkers [43].

3.4.3 Four-state Model of Pretranslocation Complex Dynamics

The dynamics of PRE complexes were analyzed using the general framework presented in section 3.4.1 where the experimental data summarized in Tables 3.1 and 3.2 were used as constraints for the linear, four-state kinetic scheme shown in Equation 3.44. For this kinetic scheme, the mean time needed for the forward transition from the terminal state 1 to the terminal state 4, $< t_p >$, is given by (see Appendix A for the full derivation)

$$< t_p >= \frac{1}{\alpha_{12}}\left[1 + \frac{\alpha_{21}}{\alpha_{23}} + \frac{\alpha_{21}\alpha_{32}}{\alpha_{23}\alpha_{34}}\right] + \frac{1}{\alpha_{23}}\left[1 + \frac{\alpha_{32}}{\alpha_{34}}\right] + \frac{1}{\alpha_{34}}. \tag{3.45}$$

The corresponding mean time for the reverse transition from the state 4 to the state 1, $< t_r >$, is given by (see Appendix A for the full derivation)

$$< t_r >= \frac{1}{\alpha_{43}}\left[1 + \frac{\alpha_{34}}{\alpha_{32}} + \frac{\alpha_{34}\alpha_{23}}{\alpha_{32}\alpha_{21}}\right] + \frac{1}{\alpha_{32}}\left[1 + \frac{\alpha_{23}}{\alpha_{21}}\right] + \frac{1}{\alpha_{21}}. \tag{3.46}$$

The probabilities for the occupation of the four states at equilibrium are given by (see Appendix A for the full derivation)

$$P_1^{eq} = \frac{\alpha_{43}\alpha_{32}\alpha_{21}}{\alpha_{43}\alpha_{32}\alpha_{21} + \alpha_{12}\alpha_{43}\alpha_{32} + \alpha_{12}\alpha_{23}\alpha_{43} + \alpha_{12}\alpha_{23}\alpha_{34}}, \tag{3.47}$$

$$P_2^{eq} = \frac{\alpha_{12}\alpha_{43}\alpha_{32}}{\alpha_{43}\alpha_{32}\alpha_{21} + \alpha_{12}\alpha_{43}\alpha_{32} + \alpha_{12}\alpha_{23}\alpha_{43} + \alpha_{12}\alpha_{23}\alpha_{34}}, \tag{3.48}$$

$$P_3^{eq} = \frac{\alpha_{12}\alpha_{23}\alpha_{43}}{\alpha_{43}\alpha_{32}\alpha_{21} + \alpha_{12}\alpha_{43}\alpha_{32} + \alpha_{12}\alpha_{23}\alpha_{43} + \alpha_{12}\alpha_{23}\alpha_{34}}, \text{ and} \tag{3.49}$$

$$P_4^{eq} = \frac{\alpha_{12}\alpha_{23}\alpha_{34}}{\alpha_{43}\alpha_{32}\alpha_{21} + \alpha_{12}\alpha_{43}\alpha_{32} + \alpha_{12}\alpha_{23}\alpha_{43} + \alpha_{12}\alpha_{23}\alpha_{34}}, \tag{3.50}$$

with the normalization condition $\sum_{\mu=1}^{4} P_\mu^{eq} = 1$. Using the equations for $< t_p >$ and $< t_r >$ (Equations 3.45 and 3.46, respectively), and the equations for P_μ^{eq} (Equations 3.47-3.50), a plot was generated of all the rate constants α_{ij} as functions of an independent α_{43} (Fig. 3.9).

Notably, for some values of the independent rate constant, solutions for the dependent rate constants are negative. While this is a consistent solution of the model, only the values where all rate constants are positive are physically-relevant solutions. The boundaries to this region where all rate constants are positive therefore represent the upper or lower bounds on the rate constants for the four-state model that are consistent with both the cryo-EM and the smFRET studies.

Interestingly, Fig. 3.7 depicts the upper and lower bounds for α_{34} and α_{43}, but only lower-bound cutoffs for the other α's. These other rate constants all asymptotically converge to positive infinity; α_{12} and α_{21} increase with increasing α_{43}, while α_{23} and α_{32} compensate by decreasing to their lower-bound. Plotting these results as a function of an independent α_{12} or α_{23} yields the same bounds–there is only a narrow window where all α's are consistent with the cryo-EM and smFRET data. With such boundaries on the individual rate constants, one can estimate the EMCCD camera acquisition rate that would be needed to distinctly observe the transient states of interest.

The theoretical results reported here are based on similar approaches followed in Refs. 38 and 39 for analytical calculations of the distribution of the dwell-times of a ribosome.

We use the kinetic scheme

$$1 \underset{\alpha_{21}}{\overset{\alpha_{12}}{\rightleftharpoons}} 2 \underset{\alpha_{32}}{\overset{\alpha_{23}}{\rightleftharpoons}} 3 \underset{\alpha_{43}}{\overset{\alpha_{34}}{\rightleftharpoons}} 4, \tag{3.51}$$

where integer indices 1, 2, 3 and 4 represent a discrete chemical state and α_{ij} denotes the transition probability per unit time (*i.e.,* rate constant) for the $i \to j$ transition. If $P_\mu(t)$ is the probability of finding the system in chemical state μ at time t, then the time evolution of these probabilities are governed by the following

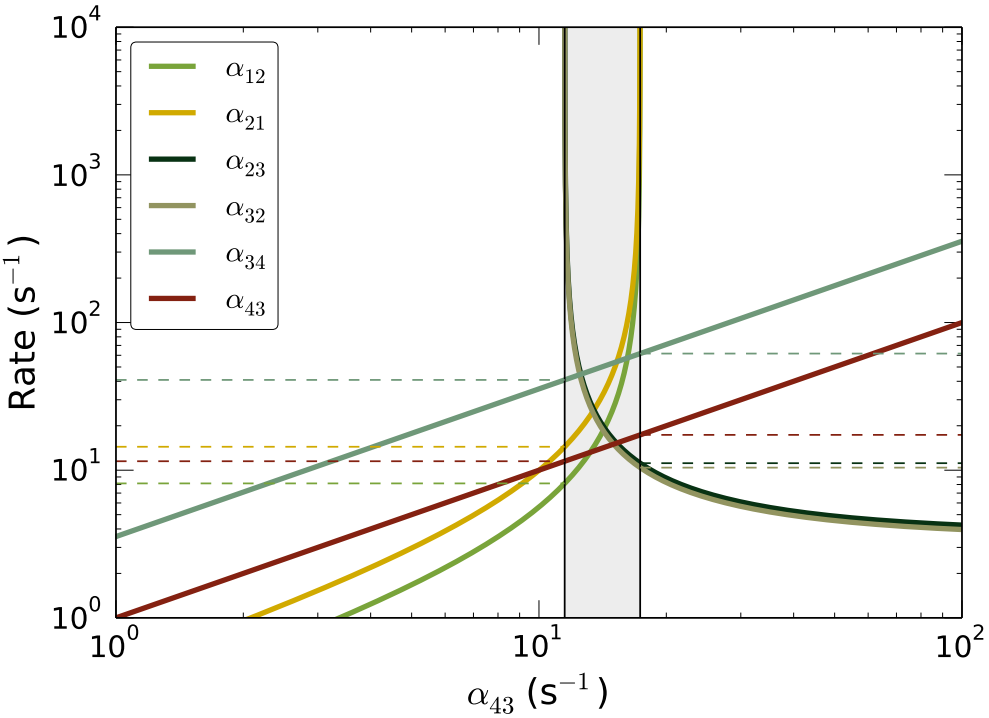

Figure 3.7: Rate constants for the four-state model as a function of α_{43}. The gray region contains the solutions where all rate constants are positive. Both axes are log scaled; so, any incompatible rates (negative valued) are not shown. The points where rate constants switch from being negative to positive valued are denoted with a black, vertical line. Horizontal, dashed lines denote upper or lower bounds for particular rate constants

master equations:

$$\frac{dP_1(t)}{dt} = -\alpha_{12}P_1(t) + \alpha_{21}P_2(t), \tag{3.52}$$

$$\frac{dP_2(t)}{dt} = \alpha_{12}P_1(t) - (\alpha_{23} + \alpha_{21})P_2(t) + \alpha_{32}P_3(t), \tag{3.53}$$

$$\frac{dP_3(t)}{dt} = \alpha_{23}P_2(t) - (\alpha_{32} + \alpha_{34})P_3(t) + \alpha_{43}P_4(t), \text{ and} \tag{3.54}$$

$$\frac{dP_4(t)}{dt} = \alpha_{34}P_3(t) - \alpha_{43}P_4(t). \tag{3.55}$$

Now we calculate the time-independent occupation probability, P_μ^{eq}, of each of these states by finding the

equilibrium-state solutions of equation 3.52-3.55,

$$P_1^{eq} = \frac{\alpha_{43}\alpha_{32}\alpha_{21}}{\alpha_{43}\alpha_{32}\alpha_{21} + \alpha_{12}\alpha_{43}\alpha_{32} + \alpha_{12}\alpha_{23}\alpha_{43} + \alpha_{12}\alpha_{23}\alpha_{34}}, \tag{3.56}$$

$$P_2^{eq} = \frac{\alpha_{12}\alpha_{43}\alpha_{32}}{\alpha_{43}\alpha_{32}\alpha_{21} + \alpha_{12}\alpha_{43}\alpha_{32} + \alpha_{12}\alpha_{23}\alpha_{43} + \alpha_{12}\alpha_{23}\alpha_{34}}, \tag{3.57}$$

$$P_3^{eq} = \frac{\alpha_{12}\alpha_{23}\alpha_{43}}{\alpha_{43}\alpha_{32}\alpha_{21} + \alpha_{12}\alpha_{43}\alpha_{32} + \alpha_{12}\alpha_{23}\alpha_{43} + \alpha_{12}\alpha_{23}\alpha_{34}}, \text{ and} \tag{3.58}$$

$$P_4^{eq} = \frac{\alpha_{12}\alpha_{23}\alpha_{34}}{\alpha_{43}\alpha_{32}\alpha_{21} + \alpha_{12}\alpha_{43}\alpha_{32} + \alpha_{12}\alpha_{23}\alpha_{43} + \alpha_{12}\alpha_{23}\alpha_{34}}. \tag{3.59}$$

We also calculate the distribution of the time spent transitioning from chemical states 1 to 4 for the first time by modifying the original kinetic scheme into:

$$1 \underset{\alpha_{21}}{\overset{\alpha_{12}}{\rightleftharpoons}} 2 \underset{\alpha_{32}}{\overset{\alpha_{23}}{\rightleftharpoons}} 3 \overset{\alpha_{34}}{\rightarrow} 4$$

and writing the master equations according to this new scheme:

$$\frac{dP_1(t)}{dt} = -\alpha_{12}P_1(t) + \alpha_{21}P_2(t), \tag{3.60}$$

$$\frac{dP_2(t)}{dt} = \alpha_{12}P_1(t) - (\alpha_{21} + \alpha_{23})P_2(t) + \alpha_{32}P_3(t), \tag{3.61}$$

$$\frac{dP_3(t)}{dt} = \alpha_{23}P_2(t) - (\alpha_{32} + \alpha_{34})P_3(t), \text{ and} \tag{3.62}$$

$$\frac{dP_4(t)}{dt} = \alpha_{34}P_3(t). \tag{3.63}$$

These equations can be re-written in terms of the following matrix notations:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{M}\mathbf{P}(t), \tag{3.64}$$

where $P(t)$ is a column matrix whose elements are $P_1(t)$, $P_2(t)$ and $P_3(t)$, and

$$\mathbf{M} = \begin{bmatrix} -\alpha_{12} & \alpha_{21} & 0 \\ \alpha_{12} & -(\alpha_{21} + \alpha_{23}) & \alpha_{32} \\ 0 & \alpha_{23} & -(\alpha_{34} + \alpha_{32}) \end{bmatrix}. \tag{3.65}$$

Now, by introducing the Laplace transform of the probability of kinetic states,

$$\tilde{P}_\mu(s) = \int_0^\infty P_\mu(t) e^{-st} dt, \tag{3.66}$$

the solution of equation (3.64) in Laplace space is

$$\tilde{\mathbf{P}}(s) = (s\mathbf{I} - \mathbf{M})^{-1}\mathbf{P}(0). \tag{3.67}$$

The determinant of matrix $(s\mathbf{I} - \mathbf{M})$ is the third-order polynomial

$$(s\mathbf{I} - \mathbf{M})^{-1} = a_3 s^3 + a_2 s^2 + a_1 s + a_0, \tag{3.68}$$

where

$$a_3 = 1, \tag{3.69}$$

$$a_2 = \alpha_{12} + \alpha_{21} + \alpha_{23} + \alpha_{32} + \alpha_{34}, \tag{3.70}$$

$$a_1 = \alpha_{21}\alpha_{34} + \alpha_{21}\alpha_{32} + \alpha_{23}\alpha_{34} + \alpha_{12}\alpha_{23} + \alpha_{12}\alpha_{34} + \alpha_{12}\alpha_{32}, \text{ and} \tag{3.71}$$

$$a_0 = \alpha_{12}\alpha_{23}\alpha_{34}. \tag{3.72}$$

We can solve equation 3.67 by using the initial condition

$$P_1(0) = 1 \text{ and } P_2(0) = P_3(0) = P_4(0) = 0. \tag{3.73}$$

Suppose that the probability of transitioning from chemical state 1 to 4 in the time interval of t and $t + \Delta t$ is $f^p(t)\Delta t$. Then,

$$f^p(t)\Delta t = \Delta P_4(t). \tag{3.74}$$

Therefore, we find

$$f^p(t) = \frac{dP_4(t)}{dt} = \omega_{34} P_3(t). \tag{3.75}$$

Taking the Laplace transform of equation 3.75 gives

$$\tilde{f}^p(s) = \omega_{34}\tilde{P}_3(s). \tag{3.76}$$

Now we can calculate an analytical expression for $\tilde{P}_3(s)$ by solving Equation 3.67. Inserting this expression into equation 3.76 yields,

$$\tilde{f}^p(s) = \frac{\alpha_{12}\alpha_{23}\alpha_{34}}{(s+\omega_1)(s+\omega_2)(s+\omega_3)}, \tag{3.77}$$

where ω_1, ω_2 and ω_3 are the solution of the equation

$$\omega^3 - a_2\omega^2 + a_1\omega - a_0 = 0. \tag{3.78}$$

Taking inverse Laplace transform of Equation 3.77 gives

$$f^p(t) = \frac{\alpha_{12}\alpha_{23}\alpha_{34}}{(\omega_1-\omega_2)(\omega_1-\omega_3)}e^{-\omega_1 t} + \frac{\alpha_{12}\alpha_{23}\alpha_{34}}{(\omega_2-\omega_1)(\omega_2-\omega_3)}e^{-\omega_2 t} + \frac{\alpha_{12}\alpha_{23}\alpha_{34}}{(\omega_3-\omega_1)(\omega_3-\omega_2)}e^{-\omega_3 t}. \tag{3.79}$$

Now, solving for the first moment of this distribution,

$$< t_p >= \int_0^\infty t f^p(t)dt, \tag{3.80}$$

gives

$$< t_p >= \frac{1}{\alpha_{12}}\left[1 + \frac{\alpha_{21}}{\alpha_{23}} + \frac{\alpha_{21}\alpha_{32}}{\alpha_{23}\alpha_{34}}\right] + \frac{1}{\alpha_{23}}\left[1 + \frac{\alpha_{32}}{\alpha_{34}}\right] + \frac{1}{\alpha_{34}}. \tag{3.81}$$

Similarly, the second moment is

$$< t_p^2 >= \int_0^\infty t^2 f^p(t)dt = \frac{2(a_1^2 - a_0 a_2)}{a_0^2}. \tag{3.82}$$

Analogously, one can also obtain the exact formula for the distribution of the time spent transitioning from chemical states 4 to 1:

$$f^r(t) = \frac{\alpha_{43}\alpha_{32}\alpha_{21}}{(\Omega_1-\Omega_2)(\Omega_1-\Omega_3)}e^{-\Omega_1 t} + \frac{\alpha_{43}\alpha_{32}\alpha_{21}}{(\Omega_2-\Omega_1)(\Omega_2-\Omega_3)}e^{-\Omega_2 t} + \frac{\alpha_{43}\alpha_{32}\alpha_{21}}{(\Omega_3-\Omega_1)(\Omega_3-\Omega_2)}e^{-\Omega_3 t}. \tag{3.83}$$

Here, Ω_1, Ω_2 and Ω_3 are the solution of the equation

$$\Omega^3 - \Omega^2 b_2 + \Omega b_1 - b_0 = 0, \tag{3.84}$$

where

$$b_0 = \alpha_{43}\alpha_{32}\alpha_{21}, \tag{3.85}$$

$$b_1 = \alpha_{34}\alpha_{21} + \alpha_{34}\alpha_{23} + \alpha_{32}\alpha_{21} + \alpha_{43}\alpha_{32} + \alpha_{43}\alpha_{21} + \alpha_{43}\alpha_{23}, \text{ and} \tag{3.86}$$

$$b_2 = \alpha_{43} + \alpha_{34} + \alpha_{32} + \alpha_{21} + \alpha_{23}. \tag{3.87}$$

Now, solving for the first moment of this distribution gives:

$$< t_r >= \int_0^\infty t f^r(t)dt = \frac{b_1}{b_0} = \frac{1}{\alpha_{43}}\left[1 + \frac{\alpha_{34}}{\alpha_{32}} + \frac{\alpha_{34}\alpha_{23}}{\alpha_{32}\alpha_{21}}\right] + \frac{1}{\alpha_{32}}\left[1 + \frac{\alpha_{23}}{\alpha_{21}}\right] + \frac{1}{\alpha_{21}}. \tag{3.88}$$

and solving for the second moment gives:

$$< t_r^2 >= \int_0^\infty t^2 f^r(t)dt = \frac{2(b_1^2 - b_0 b_2)}{b_0^2}. \tag{3.89}$$

Assuming that the experimentally observed, fractional population of chemical state i, χ_i, represents the equilibrium-state solutions, P_μ^{eq}, ratios of χ can be used to write relationships between several α,

$$\alpha_{21} = \frac{\chi_1}{\chi_2}\alpha_{12} \qquad \alpha_{32} = \frac{\chi_2}{\chi_3}\alpha_{23} \qquad \alpha_{34} = \frac{\chi_4}{\chi_3}\alpha_{43}. \tag{3.90}$$

The expectation values for the time spent transitioning from chemical states 1 to 4 ($\langle t_p \rangle$), and transitioning from chemical states 4 to 1 ($\langle t_r \rangle$) are assumed to be equivalent to the inverses of the experimentally observed transition rates between the two states of a two-state model (k_{12}, and k_{21}). Using these expressions and the experimentally observed two-state rates, making substitutions with equations (3.90) and rearranging yields the following system of equations,

$$\frac{1}{k_{12}} = 1 \cdot \frac{1}{\alpha_{12}} + C_1 \cdot \frac{1}{\alpha_{23}} + C_2 \cdot \frac{1}{\alpha_{43}}; \qquad C_1 = \left(\frac{\chi_1}{\chi_2} + 1\right) \qquad C_2 = \left(\frac{\chi_1 + \chi_2 + \chi_3}{\chi_4}\right) \tag{3.91}$$

$$\frac{1}{k_{21}} = C_3 \cdot \frac{1}{\alpha_{12}} + C_4 \cdot \frac{1}{\alpha_{23}} + 1 \cdot \frac{1}{\alpha_{43}}; \qquad C_3 = \left(\frac{\chi_4 + \chi_3 + \chi_2}{\chi_1}\right) \qquad C_4 = \left(\frac{\chi_4 + \chi_3}{\chi_2}\right), \tag{3.92}$$

which has fewer constraints than degrees of freedom. To proceed, we solve the system of equations keeping one degree of freedom independent, moving that term to the left-hand side of the equation, and treating it as part of the constraints.

Independent α_{12}

We solve the system of equations for an independent α_{12} by matrix inversion:

$$B = AX \rightarrow \begin{bmatrix} \left(\frac{1}{k_{12}} - 1 \cdot \frac{1}{\alpha_{12}}\right) \\ \left(\frac{1}{k_{21}} - C_3 \cdot \frac{1}{\alpha_{12}}\right) \end{bmatrix} = \begin{bmatrix} C_1 & C_2 \\ C_4 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\alpha_{23}} \\ \frac{1}{\alpha_{43}} \end{bmatrix} \Rightarrow X = A^{-1}B, \tag{3.93}$$

$$\text{where } A^{-1} = \frac{1}{(C_1 \cdot 1 - C_2 C_4)} \cdot \begin{bmatrix} 1 & -C_2 \\ -C_4 & C_1 \end{bmatrix},$$

which yields the following rate constants,

$$\begin{aligned}
\alpha_{12} &= Independent, \\
\alpha_{21} &= \left(\frac{\chi_1}{\chi_2}\right)\alpha_{12}, \\
\alpha_{23} &= \frac{(C_1 - C_2C_4)(k_{12}k_{21}\alpha_{12})}{(1 \cdot (k_{21}\alpha_{12} - k_{12}k_{21}) - C_2 \cdot (k_{12}\alpha_{12} - C_3 k_{12}k_{21}))}, \\
\alpha_{32} &= \left(\frac{\chi_2}{\chi_3}\right)\alpha_{23}, \\
\alpha_{34} &= \left(\frac{\chi_4}{\chi_3}\right)\alpha_{43}, \text{ and} \\
\alpha_{43} &= \frac{(C_1 - C_2C_4)(k_{12}k_{21}\alpha_{12})}{(-C_4 \cdot (k_{21}\alpha_{12} - k_{12}k_{21}) + C_1 \cdot (k_{12}\alpha_{12} - C_3 k_{12}k_{21}))}.
\end{aligned}$$

Independent α_{23}

Similarly, for an independent α_{23},

$$\begin{aligned}
\alpha_{12} &= \frac{(1 - C_2C_3)(k_{12}k_{21}\alpha_{23})}{(1 \cdot (k_{21}\alpha_{23} - C_1k_{12}k_{21}) - C_2 \cdot (k_{12}\alpha_{23} - C_4k_{12}k_{21}))}, \\
\alpha_{21} &= \left(\frac{\chi_1}{\chi_2}\right)\alpha_{12}, \\
\alpha_{23} &= Independent, \\
\alpha_{32} &= \left(\frac{\chi_2}{\chi_3}\right)\alpha_{23}, \\
\alpha_{34} &= \left(\frac{\chi_4}{\chi_3}\right)\alpha_{43}, \text{ and} \\
\alpha_{43} &= \frac{(1 - C_2C_3)(k_{12}k_{21}\alpha_{23})}{(-C_3 \cdot (k_{21}\alpha_{23} - C_1k_{12}k_{21}) + 1 \cdot (k_{12}\alpha_{23} - C_4k_{12}k_{21}))}.
\end{aligned}$$

Independent α_{43}

Finally, for an independent α_{43},

$$\begin{aligned}
\alpha_{12} &= \frac{(C_4 - C_1C_3)(k_{12}k_{21}\alpha_{43})}{(C_4 \cdot (k_{21}\alpha_{43} - C_2k_{12}k_{21}) - C_1 \cdot (k_{12}\alpha_{43} - k_{12}k_{21}))}, \\
\alpha_{21} &= \left(\frac{\chi_1}{\chi_2}\right)\alpha_{12}, \\
\alpha_{23} &= \frac{(C_4 - C_1C_3)(k_{12}k_{21}\alpha_{43})}{(-C_3 \cdot (k_{21}\alpha_{43} - C_2k_{12}k_{21}) + 1 \cdot (k_{12}\alpha_{43} - k_{12}k_{21}))}, \\
\alpha_{32} &= \left(\frac{\chi_2}{\chi_3}\right)\alpha_{23}, \\
\alpha_{34} &= \left(\frac{\chi_4}{\chi_3}\right)\alpha_{43}, \text{ and} \\
\alpha_{43} &= Independent.
\end{aligned}$$

3.4.4 Three-State Model of Pretranslocation Complex Dynamics

Since the number of equations available in our four-state model is five whereas the number of unknown rate constants is six, we could only express five rate constants in terms of the sixth one. In contrast, because we can reduce the number of states from four to three (*vide infra*), in this subsection we use the four corresponding independent equations to extract the absolute values of the four rate constants associated with

the three-state model,

$$1 \underset{\alpha_{21}}{\overset{\alpha_{12}}{\rightleftharpoons}} 2 \underset{\alpha_{32}}{\overset{\alpha_{23}}{\rightleftharpoons}} 3. \tag{3.94}$$

As explained in Appendix C, structural analysis strongly suggests that the L1-tRNA distance in IS1 is insufficiently different from that of MS I so as to result in an E_{FRET} that is significantly different than that of MS I. Thus, MS I and IS1 can be combined into a single state, state 1, thereby reducing the four-state model into a three-state kinetic scheme of Eqn. 3.94, where the states 2 and 3 correspond to IS2 and MS II, respectively.

The expressions for $< t_p >$, $< t_r >$, and P_μ ($\mu = 1, 2, 3$) are given by (see Appendix B for detailed derivations),

$$< t_p >= \frac{1}{\alpha_{12}}\left[1 + \frac{\alpha_{21}}{\alpha_{23}}\right] + \frac{1}{\alpha_{23}}, \tag{3.95}$$

$$< t_r >= \frac{1}{\alpha_{32}}\left[1 + \frac{\alpha_{23}}{\alpha_{21}}\right] + \frac{1}{\alpha_{21}}, \tag{3.96}$$

$$P_1^{eq} = \frac{\alpha_{21}\alpha_{32}}{\alpha_{21}\alpha_{32} + \alpha_{12}\alpha_{32} + \alpha_{23}\alpha_{12}}, \tag{3.97}$$

$$P_2^{eq} = \frac{\alpha_{12}\alpha_{32}}{\alpha_{21}\alpha_{32} + \alpha_{12}\alpha_{32} + \alpha_{23}\alpha_{12}}, \text{ and} \tag{3.98}$$

$$P_3^{eq} = \frac{\alpha_{23}\alpha_{12}}{\alpha_{21}\alpha_{32} + \alpha_{12}\alpha_{32} + \alpha_{23}\alpha_{12}}, \tag{3.99}$$

where the normalization condition is $\sum_{\mu=1}^{eq} P_\mu = 1$. These expressions, together with the corresponding experimental data, are utilized to write down four independent equations. The rate constants computed by solving those four equations are shown in Table 3.3.

Interestingly, this calculation suggests that the rate limiting steps for the forward and reverse reactions are IS2 $\rightarrow$ MS II and MS II $\rightarrow$ IS2, respectively, while interconversion between MS I/IS1 and IS2 occur relatively rapidly. These rates for MS I/IS1 to IS2 interconversion are approximately the same or faster than the 20 s^{-1} EMMCD camera acquisition rate of the original smFRET data from Fei *et al.*, and, additionally, the change in the L1-tRNA distance between the MS I/IS1 and IS2 states is relatively small (~ 80 Å to 64 Å), resulting in a correspondingly small difference in E_{FRET} (~ 0.15 to 0.40); so, any separation of MS I/IS1 and IS2 that might have been observed in the smFRET data would likely have been obscured in the HMM analysis process by camera blurring arising from interconversion rates that are similar to the acquisition rate. The fast interconversion and small expected changes in E_{FRET} suggest that MS I, IS1, and IS2 might have

originally been interpreted as a 'single', averaged state in the analysis of the smFRET data.

α	Max $k_{L1\text{-}tRNA}$ (s^{-1})	Mean $k_{L1\text{-}tRNA} \pm 1\sigma$ (s^{-1})
α_{12}	18.1	23.3 ± 22.7
α_{21}	46.8	52.5 ± 33.6
α_{23}	5.90	5.89 ± 0.42
α_{32}	1.66	1.66 ± 0.12

Table 3.3: Rate constants for $PRE_{fM/F}$ using a linear, three-state kinetic scheme where MS I and IS1 have been combined into the first state. Error from the ebFRET-estimated smFRET-determined rate constants, and the counting error from the cryo-EM study were propagated into distributions of the α. These distributions are strictly not normal distributions, although α_{23} and α_{32} are approximately normal.

Given the following linear, three-state model,

$$1 \underset{\alpha_{21}}{\overset{\alpha_{12}}{\rightleftharpoons}} 2 \underset{\alpha_{32}}{\overset{\alpha_{23}}{\rightleftharpoons}} 3,$$

the time evolution of the probability, $P_\mu(t)$, will be governed by

$$\frac{dP_1(t)}{dt} = -\alpha_{12}P_1(t) + \alpha_{21}P_2(t), \tag{3.100}$$

$$\frac{dP_2(t)}{dt} = \alpha_{12}P_1(t) - (\alpha_{21} + \alpha_{23})P_2(t) + \alpha_{32}P_3(t), \text{ and} \tag{3.101}$$

$$\frac{dP_3(t)}{dt} = \alpha_{23}P_2(t) - \alpha_{32}P_3(t). \tag{3.102}$$

We calculate the equilibrium-state probabilities

$$P_1^{eq} = \frac{\alpha_{21}\alpha_{32}}{\alpha_{21}\alpha_{32} + \alpha_{12}\alpha_{32} + \alpha_{23}\alpha_{12}} \tag{3.103}$$

$$P_2^{eq} = \frac{\alpha_{12}\alpha_{32}}{\alpha_{21}\alpha_{32} + \alpha_{12}\alpha_{32} + \alpha_{23}\alpha_{12}} \tag{3.104}$$

$$P_3^{eq} = \frac{\alpha_{23}\alpha_{12}}{\alpha_{21}\alpha_{32} + \alpha_{12}\alpha_{32} + \alpha_{23}\alpha_{12}} \tag{3.105}$$

Then, we also calculate the distribution of the time spent by a molecule transitioning from chemical states 1 to 3. For this purpose, we modify the scheme into

$$1 \underset{\alpha_{21}}{\overset{\alpha_{12}}{\rightleftharpoons}} 2 \overset{\alpha_{23}}{\rightarrow} 3,$$

and write the master equations according to this new scheme

$$\frac{dP_1(t)}{dt} = -\alpha_{12}P_1(t) + \alpha_{21}P_2(t), \tag{3.106}$$

$$\frac{dP_2(t)}{dt} = \alpha_{12}P_1(t) - (\alpha_{21} + \alpha_{23})P_2(t), \text{ and} \tag{3.107}$$

$$\frac{dP_3(t)}{dt} = \alpha_{23}P_2(t). \tag{3.108}$$

As in the four-state result, we can solve these equations with the Laplace transform method to yield,

$$f^p(t) = \frac{\alpha_{12}\alpha_{23}}{\omega_2 - \omega_1}e^{-\omega_1 t} + \frac{\alpha_{12}\alpha_{23}}{\omega_1 - \omega_2}e^{-\omega_2 t}, \tag{3.109}$$

where ω_1 and ω_2 are the solution of equation

$$\omega^2 - \omega(\alpha_{12} + \alpha_{21} + \alpha_{23}) + \alpha_{12}\alpha_{23} = 0. \tag{3.110}$$

Now, solving for the first moment,

$$< t_p >= \frac{1}{\alpha_{12}}\left[1 + \frac{\alpha_{21}}{\alpha_{23}}\right] + \frac{1}{\alpha_{23}}, \tag{3.111}$$

and the second moment,

$$< t_p^2 >= \frac{2(c_1^2 - c_0)}{c_0^2}, \tag{3.112}$$

where

$$c_0 = \alpha_{12}\alpha_{23}, \text{ and} \tag{3.113}$$

$$c_1 = \alpha_{12} + \alpha_{21} + \alpha_{23}. \tag{3.114}$$

Similarly, we can also calculate the distribution of the time spent transitioning from chemical states 3 to 1,

$$f^r(t) = \frac{\alpha_{21}\alpha_{32}}{\Omega_2 - \Omega_1}e^{-\Omega_1 t} + \frac{\alpha_{21}\alpha_{32}}{\Omega_1 - \Omega_2}e^{-\Omega_2 t}, \tag{3.115}$$

where Ω_1 and Ω_2 are the solution of the equation

$$\Omega^2 - \Omega(\alpha_{32} + \alpha_{23} + \alpha_{21}) + \alpha_{32}\alpha_{21} = 0. \tag{3.116}$$

In this case, we calculate the first moment,

$$< t_r >= \frac{1}{\alpha_{32}}\left[1 + \frac{\alpha_{23}}{\alpha_{21}}\right] + \frac{1}{\alpha_{21}}, \tag{3.117}$$

and the second moment,

$$< t_r^2 >= \frac{2(d_1^2 - d_0)}{d_0^2}, \tag{3.118}$$

where

$$d_0 = \alpha_{32}\alpha_{21}, \text{ and} \tag{3.119}$$

$$d_1 = \alpha_{32} + \alpha_{23} + \alpha_{21}. \tag{3.120}$$

Assuming that the experimentally observed, fractional population of state i, χ_i, represents the equilibrium-state solutions, P_μ^{eq}, ratios of χ can be used to write relationships between several α,

$$\alpha_{21} = \frac{\chi_1}{\chi_2}\alpha_{12}, \text{ and} \qquad \alpha_{23} = \frac{\chi_3}{\chi_2}\alpha_{32}. \tag{3.121}$$

The expectation values for the time spent transitioning from chemical states 1 to 3 ($\langle t_p \rangle$), and transitioning from chemical states 3 to 1 ($\langle t_r \rangle$) are assumed to be equivalent to the inverses of the experimentally observed transition rates between the two final states of a two-state model (k_{12}, and k_{21}). Using these expressions and experimentally observed rates, substituting Eqn. (3.121), and then rearranging yields the following system of equations,

$$\langle t_p \rangle = \frac{1}{k_{12}} = 1 \cdot \frac{1}{\alpha_{12}} + C_1 \cdot \frac{1}{\alpha_{32}}; \qquad C_1 \equiv \left(\frac{\chi_1}{\chi_3} + 1\right) \tag{3.122}$$

$$\langle t_r \rangle = \frac{1}{k_{21}} = C_2 \cdot \frac{1}{\alpha_{12}} + 1 \cdot \frac{1}{\alpha_{32}}; \qquad C_2 \equiv \left(\frac{\chi_3 + \chi_2}{\chi_1}\right), \tag{3.123}$$

which can be solved as for the four-state model,

$$B = AX \rightarrow \begin{bmatrix} \frac{1}{k_{12}} \\ \frac{1}{k_{21}} \end{bmatrix} = \begin{bmatrix} 1 & C_1 \\ C_2 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\alpha_{12}} \\ \frac{1}{\alpha_{32}} \end{bmatrix} \Rightarrow X = \begin{bmatrix} \frac{1}{k_{12}} \\ \frac{1}{k_{21}} \end{bmatrix} = \frac{1}{(1 - C_1 C_2)} \cdot \begin{bmatrix} 1 \cdot \frac{1}{k_{12}} - C_2 \cdot \frac{1}{k_{21}} \\ -C_2 \cdot \frac{1}{k_{12}} + 1 \cdot \frac{1}{k_{21}} \end{bmatrix}, \tag{3.124}$$

to yield the three-state model rate constants,

$$\alpha_{12} = \frac{(1 - C_1C_2)(k_{12}k_{21})}{(1 \cdot k_{21} - C_1 \cdot k_{12})},$$
$$\alpha_{21} = \left(\frac{\chi_1}{\chi_2}\right)\alpha_{12},$$
$$\alpha_{23} = \left(\frac{\chi_3}{\chi_2}\right)\alpha_{32}, \text{ and}$$
$$\alpha_{32} = \frac{(1 - C_1C_2)(k_{12}k_{21})}{(-C_2 \cdot k_{21} - 1 \cdot k_{12})}.$$

3.4.5 Simulated Pretranslocation Complex Dynamics

We simulated 100 E_{FRET} versus time trajectories with a linear, three-state kinetic scheme. Dwell-times prior to transitions to other states were exponentially distributed according to the appropriate rate constants. In each E_{FRET} versus time trajectory, the value of E_{FRET} corresponding to each state was randomized by choosing r from a normal distribution, and calculating $E_{FRET} = (1 + (r/R_0)^6)^{-1}$, where R_0 is the Förster radius. For each E_{FRET} versus time trajectory, R_0 was also randomly chosen from a normal distribution. Noise reflecting a reasonable SBR for the total-internal reflection fluorescence (TIRF) microscope used in the smFRET experiments (*i.e.*, $\sigma = 0.05$) was also added to each E_{FRET} versus time trajectory. More details can be found in Appendix C.

In order to simulate PRE complex dynamics with transient intermediates, we estimated the values of E_{FRET} for each PRE complex conformational state. We estimated E_{FRET} for the L1-tRNA labeling scheme by measuring the distances between the β-carbon of the threonine at position 202 of the L1 protein of the 50S ribosomal subunit and the sulfur of the thiouridine at position 8 of $tRNA^{fMet}$ from the atomic-resolution, molecular dynamics flexible fitting of the classes of PRE complexes deposited in the Protein Data Bank by Agirrezabala and coworkers (Table 3.4) [41]. While the absolute accuracy of these estimates is likely imprecise, it is reasonable to interpret the relative distances as an informative measure of the relative E_{FRET} values. As the distances measured ignore any foreshortening due to the space occupied by the fluorophore and its hydrocarbon linker, subsequent analysis compensated by overestimating R_0 = 60 Å. Notably, all the classes measured yielded distinct values of E_{FRET} except for classes 2 and 4A (MS I and IS1, respectively). Since the distances between the labeling sites on the L1 protein and the P-site tRNA are 78 Å ($E_{FRET} \approx 0.17$) and 81 Å ($E_{FRET} \approx 0.14$) for classes 2 and 4A, respectively, MS I and IS1 are most likely indistinguishable given the SBR of the TIRF-based smFRET measurements used by Fei and coworkers [42, 43]. As such, we

chose to group MS I and IS1 together into state 1, while IS2 corresponded to state 2, and MS II corresponded to state 3.

Class	r(L1 $\rightarrow$ tRNA)$\left(\text{Å}\right)$	E_{FRET}, R_0 = 60 Å
2	78	0.17
4A	81	0.14
4B	64	0.41
5	43	0.89
6	55	0.62

Table 3.4: Distances and approximate FRET efficiencies of $\text{PRE}_{\text{fM/F}}$ ribosomes from cryo-EM structures.

From these E_{FRET} estimates, Markovian transitions along a linear, three-state kinetic scheme were then simulated for 100 state versus time trajectories. Each state versus time trajectory was 50 sec in length, and they were eventually transformed into discrete, E_{FRET} versus time trajectories, where each data point is the mean E_{FRET} value during a 50 msec time period. For each state versus time trajectory, the distances between the donor- and acceptor fluorophores in the i^{th} state, r_i, were randomized for each time series with a normal distribution, $\mathcal{N}\left(\mu = r_i, \sigma = 2\text{Å}\right)$. The E_{FRET} of each state was then calculated as $E_{\text{FRET}} = (1 + (r/R_0)^6)^{-1}$, where R_0 is the Förster radius (a parameter dependent upon the identity of donor- and acceptor fluorophores, as well their local environments). R_0 was randomized for each time series within a reasonable range for the Cy3-Cy5 FRET donor-acceptor pair with a normal distribution, $\mathcal{N}\left(\mu = 60\text{Å}, \sigma = 2\text{Å}\right)$. The E_{FRET} versus time trajectory was then discretized by calculating the average value of E_{FRET} during sequential 50 msec long time periods. Noise was added to the E_{FRET} versus time trajectories that was normally distributed at each data point with a standard deviation of 0.05 – a reasonable SBR for data collected on the TIRF microscope used in the smFRET experiments.

In order to model the histogram of the simulated E_{FRET} versus time trajectories, we used a Gaussian mixture model where each state is modeled to contribute as a normal distribution centered at the respective mean E_{FRET} value for that state, and is weighted by the equilibrium-state probability for that state (see Appendix B). To account for the simulated heterogeneity in the ensemble of synthetic E_{FRET} versus time trajectories, the mean E_{FRET} value of each state was marginalized out by integrating over the joint-probability distribution of the normal distribution of E_{FRET} and a beta distribution of the mean E_{FRET} observed in that state (Fig. 3.9C) with parameters determined by a maximum likelihood estimate from the exact simulated E_{FRET} means. For the "Avg. State" distribution (*c.f.,* Fig. 3.9B), the distribution of E_{FRET} means that was employed was beta distributed with a linear combination of the parameters of the mean E_{FRET} distributions

from states 1 and 2 (Fig. 3.9C). The standard deviation used for the normal distribution of E_{FRET} values for each state was taken exactly as the standard deviation used to add noise to the synthetic E_{FRET} versus time trajectories.

As described in Section 3.4.5, this model of the observed E_{FRET} value histograms is for a temporally-resolved histogram without any blurring present. Performing the same simulation described above, but with a 0.5 msec acquisition time period yields an accurately modeled set of E_{FRET} versus time trajectories (Fig. 3.10). Furthermore, analyzing this data with ebFRET yields an accurately estimated number of states, rate constants, distribution of E_{FRET} means, and noise parameter.

Analysis of the 50 msec time resolution data with ebFRET found the most evidence for a five-state kinetic model–all of which were significantly populated. Since the synthetic data was simulated with a three-state kinetic model, the ebFRET analysis is not consistent with the original simulation. Additionally, the rate constants inferred by ebFRET for a three-state model from the simulated data do not match the original simulation parameters. The rate constants inferred by ebFRET for the two-state model were $k_{GS1\rightarrow GS2} = 1.41 \pm 0.05 s^{-1}$, and $k_{GS2\rightarrow GS1} = 1.41 \pm 0.05 s^{-1}$, but these differ from those learned from the data of Fei and coworkers (*c.f.,* Section 1.3.2). This suggests that the original simulation parameters are not consistent with the experimental data from the smFRET study of Fei and coworkers [42, 43]. Most likely, the discrepancy is due to experimental differences between the smFRET and cryo-EM studies that were compared using the general framework presented here.

To investigate how ebFRET would treat this 'single', averaged smFRET state, synthetic time series simulating a three-state $PRE_{fM/F}$ complex were constructed guided by the analysis above. Estimates for E_{FRET} were based upon the cryo-EM structures of Agirrezabala and coworkers (see Appendix C), and the kinetic scheme and associated rate constants employed are those in Section 3.4.4. Since the rate constants for the transition between states MS I/IS1 and IS2 are of the same order of magnitude as the frame rate of this simulation, traditional smFRET data analysis of this synthetic data provides insight into whether blurring could have obscured any transient, intermediate states in the data of Fei and coworkers. Typically, such obfuscation begins to manifest when dwell-times in a state of interest approach the same order of magnitude as the EMCCD camera acquisition time, because of errors in estimating the lengths of the dwell-times [4].

The synthetic smFRET dataset was constructed by carrying out simulations of E_{FRET} versus time trajectories where each 'single-ribosome' had randomized simulation parameters as described in Appendix C. This probabilistic approach accounts for experimental variation (*e.g.,* uneven illumination in the field-of-view), as well as ensemble variations (*e.g.* static disorder from a small sub-population of ribosomes lacking an A-

site dipeptidyl-tRNA). An example of a synthetic E_{FRET} vs. time trajectory is shown in Fig. 3.8, where the ensemble mean values of E_{FRET} were $\sim$ 0.16, 0.40, and 0.74 for states MS I/IS1, IS2, and MS II, respectively.

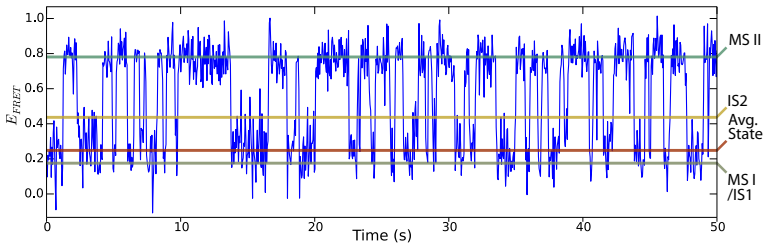

Figure 3.8: Example synthetic E_{FRET} versus time trajectory. The gray, yellow, and blue horizontal lines denote the E_{FRET} means used to simulate MS I/IS1, IS2, and MS II for this time series, respectively. The red horizontal line is the equilibrium-state-weighted average of the means of the MS I/IS1 and IS2 states.

With regard to the distribution of E_{FRET} values observed from any ensemble of E_{FRET} versus time trajectories, blurring would result in a shift of some of the density of the equilibrium-state occupancy probability distribution to an intermediate, averaged value between the blurred states. Deviation of a histogram of the observed, simulated E_{FRET} in the synthetic data set from the distribution predicted by the equilibrium-state occupancies of the linear, three-state model therefore can be used to characterize the amount of blurring present in the synthetic data. We modeled the normalized histogram of the synthetic ensemble with normal distributions weighted by their respective equilibrium-state probability, P_{μ}^{eq} (Fig. 3.9A). The mean of each state was distributed according to the distribution of static E_{FRET} for that state in each of the synthetic time series (Fig. 3.9C). This approach accurately reflects a non-blurred histogram of E_{FRET} (Fig. 3.10). Deviations that occur are therefore due to blurring, or, if they had been simulated in this synthetic dataset, could have been due to the presence of unaccounted-for states. Notably, a large portion of the MS I/IS1 density in Fig. 3.9A is relocated into the region between MS I/IS1 and IS2. This is a direct manifestation of blurring. By collapsing MS I/IS1 and IS2 into one averaged state (Fig. 3.9B and 4C), we find that the data are much better described by only two states (Fig. 3.9D). In this case, the artifactual, blurred, averaged state overwhelms any distinction between the MS I/IS1 and IS2 states, whereas when the simulation is performed with an acquisition rate that is significantly faster than the transitions of interest (2000 s^{-1}), these states are well-resolved (Fig. 3.10).

This blurred, synthetic, three-state ensemble of E_{FRET} versus time trajectories was then analyzed with ebFRET. Interestingly, ebFRET overestimated the true number of kinetic states. Most likely this is due to the fact that the dwell-times in states MS I/IS1 and IS2 are too transient relative to the simulation's 'acquisition

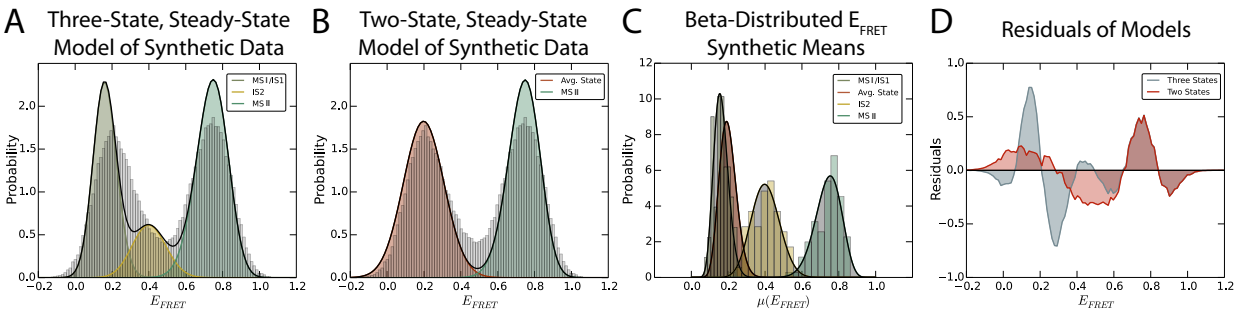

Figure 3.9: (A) Histogram of synthetic PRE complex data (gray) modeled with three, non-blurred states. (B) Histogram of synthetic PRE complex data (gray) modeled with two, non-blurred states. The "Avg. State" is a weighted combination of MS I/IS1 and IS2. (C) Histograms and probability distributions of the static E_{FRET} value means generated for the synthetic PRE complex dataset. (D) Plots of the model probability densities minus the normalized histograms from panels A and B. The sum of the squares of these residuals are 10.03, and 4.02 for the three-state and two-state models, respectively, suggesting that the blurred, synthetic, PRE complex data are better represented by a two-state model.

rate', because ebFRET is able to accurately infer the model parameters from the same data with a more appropriate acquisition rate of 2000 s^{-1} (Fig. 3.10). As a result, the blurred, synthetic data points are modeled by ebFRET as distinct 'states' – even though these states do not actually exist on the 'energy landscape' of the simulated $PRE_{fM/F}$ complex.

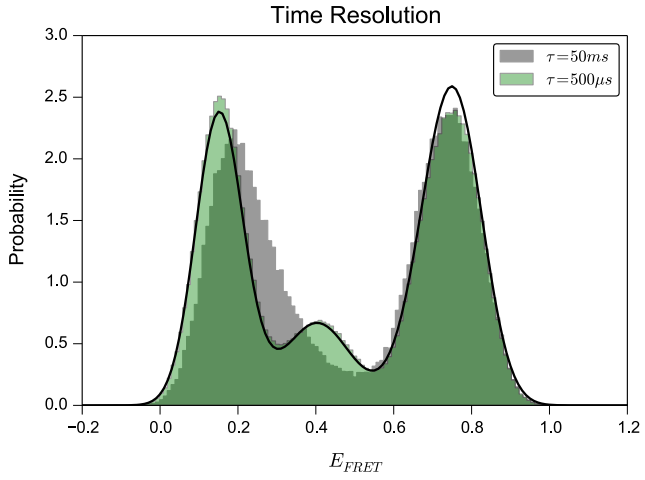

Figure 3.10: Histograms of the same ensemble of synthetic E_{FRET} vs. time trajectories rendered with different time resolutions. The left-most peak in the grey histogram, which was used in Fig. 3.9, is well-resolved into two separate peaks when the synthetic data is rendered with a 100x faster acquisition rate (green histogram). The three-state model distribution (black curve) is that from Fig. 3.9, which was modeled using only the initial simulation parameters. Analysis of the faster time resolution time series (green histogram) using ebFRET accurately estimated all four transition rates ($k_{12} = 17.8 \pm 0.6$ s^{-1}, $k_{21} = 47.2 \pm 1.6$ s^{-1}, $k_{23} = 6.4 \pm 0.6$ s^{-1}, $k_{32} = 1.82 \pm 0.17$ s^{-1}), as well as accurately estimated the distribution of E_{FRET} means ($\mu_1 = 0.16$, $\sigma_1 = 0.035$, $\mu_2 = 0.41$, $\sigma_2 = 0.067$, $\mu_3 = 0.74$, $\sigma_3 = 0.058$) and the noise parameter ($\sigma_{noise} = 0.050$).

3.5 Conclusion

We have established a general, theoretical framework that integrates equilibrium-state cryo-EM observations with dynamic, time-dependent smFRET observations. This framework is not limited exclusively to cryo-EM and smFRET; any technique that provides static, equilibrium-state populations can be integrated with any other technique that provides dynamic, time-dependent transition rates. The analysis reported here suggests that states IS1 and IS2, previously identified by cryo-EM, are not detected by smFRET because their short lifetimes result in transitions that are too fast to be characterized, given the acquisition rate of the EM-CCD camera and the limitations of the manner in which hidden Markov models are implemented. Moreover, structural analyses indicate that the L1-tRNA smFRET signal is unable to yield distinguishable signals for MS I and IS1, given the typical SBR of TIRF microscope-based smFRET measurements. The quantitative analysis of the experimental data, based on the analytical theory presented here, provides a possible explanation for why the PRE complex intermediates observed by cryo-EM (*i.e.*, IS1 and IS2) escaped detection by the smFRET studies of Fei *et al.* [42, 43], and, possibly, other groups [46–48]. Conversely, application of this framework to smFRET data in which intermediate states have been detected, but have been difficult to assign to specific PRE complex structures [49, 50], should allow researchers to determine whether and how such intermediates correspond to IS1 and/or IS2. Perhaps more importantly, we can now predict the

lifetimes and corresponding rates of transitions into and out of IS1 and IS2 that are crucial for designing future kinetic experiments. This work therefore resolves a discrepancy in the field and opens a path for performing and analyzing future experiments. Furthermore, we hope to use this theoretical framework to make predictions regarding future experiments in which the cryo-EM and smFRET data would be collected under more comparable conditions. This theoretical framework will also be useful in guiding future experimental explorations which include, for example, stabilization of IS1 or IS2 (or any other intermediates that may be ultimately identified) using different tRNAs or mutant ribosomes.

3.6 References

1. Tinoco, I. & Gonzalez, R. L. Biological mechanisms, one molecule at a time. *Genes Dev.* **25,** 1205–1231 (2011).
2. Colquhoun, D. & Hawkes, A. G. Relaxation and fluctuations of membrane currents that flow through drug-operated channels. *Proc. R. Soc. Lond. B. Biol. Sci.* **199,** 231–262 (1977).
3. Colquhoun, D. & Hawkes, A. G. On the stochastic properties of single ion channels. *Proc. R. Soc. London. Ser. B* **211,** 205–235 (1981).
4. Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez, R. L. & Wiggins, C. H. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* **97,** 3196–205 (2009).
5. Bronson, J. E., Hofman, J. M., Fei, J., Gonzalez, R. L. & Wiggins, C. H. Graphical models for inferring single molecule dynamics. *BMC Bioinformatics* **11,** S2 (2010).
6. Van De Meent, J.-W., Bronson, J. E., Wood, F., Gonzalez Jr., R. L. & Wiggins, C. H. Hierarchically-coupled hidden Markov models for learning kinetic rates from single-molecule data. *Proc. 30th Int. Conf. Mach. Learn.* (2013).
7. Van de Meent, J.-W., Bronson, J. E., Wiggins, C. H. & Gonzalez, R. L. Empirical Bayes Methods Enable Advanced Population-Level Analyses of Single-Molecule FRET Experiments. *Biophys. J.* **106,** 1327–1337 (2014).
8. Zhou, H.-X. *Rate theories for biologists.* **2,** 219–93 (2010).
9. McQuarrie, D. A. Kinetics of Small Systems. I. *J. Chem. Phys.* **38,** 433 (1963).
10. McQuarrie, D. A. Stochastic Approach to Chemical Kinetics. *J. Appl. Probab.* **4,** 413–478 (1967).
11. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22,** 403–434 (1976).
12. Gillespie, D. T. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **81,** 2340–2361 (1977).
13. Gillespie, D. T. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58,** 35–55 (2007).

14. Van Kampen, N. *StochasticProcesses in Physics and Chemistry* 3rd ed., 1–464 (North Holland, Amsterdam, 2007).

15. Zwanzig, R. *Nonequilibrium Statistical Mechanics* 1–222 (Oxford University Press, Oxford, 2001).

16. Colquhoun, D. & Hawkes, A. G. in *Single-Channel Rec.* 397–482 (Springer US, Boston, MA, 1995).

17. Zwanzig, R. Two-state models of protein folding kinetics. *Proc. Natl. Acad. Sci.* **94,** 148–50 (1997).

18. Espenson, J. H. *Chemical Kinetics and Reaction Mechanisms* 2nd (McGraw-Hill, New York, 1995).

19. Onsager, L. Reciprocal Relations in Irreversible Processes. I. *Phys. Rev.* **37,** 405–426 (1931).

20. Resnick, S. I. *Adventures in Stochastic Processes* (Birkhauser Verlag, Basel, Switzerland, Switzerland, 1992).

21. Sivia, D. S. & Skilling, J. *Data Analysis: A Bayesian Tutorial* 1–259 (Oxford University Press, Oxford, 2006).

22. Crouzy, S. C. & Sigworth, F. J. Yet another approach to the dwell-time omission problem of single-channel analysis. *Biophys. J.* **58,** 731–743 (1990).

23. Stigler, J. & Rief, M. Hidden Markov Analysis of Trajectories in Single-Molecule Experiments and the Effects of Missed Events. *ChemPhysChem* **13,** 1079–1086 (2012).

24. English, B. P. *et al.* Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat. Chem. Biol.* **2,** 87–94 (2006).

25. Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H. & Gunsalus, I. C. Dynamics of ligand binding to myoglobin. *Biochemistry* **14,** 5355–5373 (1975).

26. Astumian, R. D. Thermodynamics and Kinetics of a Brownian Motor. *Science* **276,** 917–922 (1997).

27. Alberts, B. The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. *Cell* **92,** 291–4 (1998).

28. Peskin, C. S., Odell, G. M. & Oster, G. F. Cellular Motions and Thermal Fluctuations: The Brownian Ratchet. *Biophys. J.* **65,** 316–324 (1993).

29. Frank, J. in *Mol. Mach. Biol. Cell* (ed Frank, J.) 1–3 (Cambridge University Press, New York, 2011).

30. Bustamante, C. In singulo biochemistry: when less is more. *Annu. Rev. Biochem.* **77,** 45–50 (2008).

31. Chowdhury, D. in *Mol. Mach. Biol. Cell* (ed Frank, J.) 38–58 (Cambridge University Press, New York, 2011).

32. Chowdhury, D. Modeling stochastic kinetics of molecular machines at multiple levels: from molecules to modules. *Biophys. J.* **104,** 2331–41 (2013).

33. Kolomeisky, A. B. Motor proteins and molecular motors: how to operate machines at the nanoscale. *J. Physics. Condens. Matter* **25,** 463101 (2013).

34. Tinoco, I. & Wen, J.-D. Simulation and analysis of single-ribosome translation. *Phys. Biol.* **6,** 025006 (2009).

35. Xie, S. Single-Molecule Approach to Enzymology. *Single Mol.* **2,** 229–236 (2001).

36. Frank, J. Story in a sample-the potential (and limitations) of cryo-electron microscopy applied to molecular machines. *Biopolymers* **99,** 832–6 (2013).

37. Kinz-Thompson, C. D. & Gonzalez, R. L. smFRET studies of the 'encounter' complexes and subsequent intermediate states that regulate the selectivity of ligand binding. *FEBS Lett.* **588,** 3526–38 (2014).

38. Garai, A., Chowdhury, D., Chowdhury, D. & Ramakrishnan, T. Stochastic kinetics of ribosomes: Single motor properties and collective behavior. *Phys. Rev. E* **80,** 011908 (2009).

39. Sharma, A. K. & Chowdhury, D. Distribution of dwell times of a ribosome: effects of infidelity, kinetic proofreading and ribosome crowding. *Phys. Biol.* **8,** 026005 (2011).

40. Agirrezabala, X., Lei, J., Brunelle, J. L., Ortiz-Meoz, R. F., Green, R. & Frank, J. Visualization of the hybrid state of tRNA binding promoted by spontaneous ratcheting of the ribosome. *Mol. Cell* **32,** 190–7 (2008).

41. Agirrezabala, X. *et al.* Structural characterization of mRNA-tRNA translocation intermediates. *Proc. Natl. Acad. Sci.* **109,** 6094–9 (2012).

42. Fei, J., Kosuri, P., MacDougall, D. D. & Gonzalez, R. L. Coupling of ribosomal L1 stalk and tRNA dynamics during translation elongation. *Mol. Cell* **30,** 348–59 (2008).

43. Fei, J., Bronson, J. E., Hofman, J. M., Srinivas, R. L., Wiggins, C. H. & Gonzalez, R. L. Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation. *Proc. Natl. Acad. Sci.* **106,** 15702–7 (2009).

44. Blanchard, S. C., Kim, H. D., Gonzalez, R. L., Puglisi, J. D. & Chu, S. tRNA dynamics on the ribosome during translation. *Proc. Natl. Acad. Sci.* **101,** 12893–8 (2004).

45. Kim, H. D., Puglisi, J. D. & Chu, S. Fluctuations of transfer RNAs between classical and hybrid states. *Biophys. J.* **93,** 3575–82 (2007).

46. Cornish, P. V., Ermolenko, D. N., Noller, H. F. & Ha, T. Spontaneous intersubunit rotation in single ribosomes. *Mol. Cell* **30,** 578–88 (2008).

47. Cornish, P. V. *et al.* Following movement of the L1 stalk between three functional states in single ribosomes. *Proc. Natl. Acad. Sci.* **106,** 2571–6 (2009).

48. Chen, C. *et al.* Single-molecule fluorescence measurements of ribosomal translocation dynamics. *Mol. Cell* **42,** 367–77 (2011).

49. Munro, J. B., Altman, R. B., O'Connor, N. & Blanchard, S. C. Identification of two distinct hybrid state intermediates on the ribosome. *Mol. Cell* **25,** 505–17 (2007).

50. Munro, J. B., Altman, R. B., Tung, C.-S., Cate, J. H. D., Sanbonmatsu, K. Y. & Blanchard, S. C. Spontaneous formation of the unlocked state of the ribosome is a multistep process. *Proc. Natl. Acad. Sci.* **107,** 709–14 (2010).

Chapter 4

Temporal Super-Resolution*

4.1 Bayesian Inference for the Analysis of Sub-temporal Resolution Data (BIASD)

Given their inherent ability to avoid ensemble averaging, time-resolved single-molecule techniques have revolutionized the study of biological mechanisms by enabling distributions of molecular properties to be observed, transiently sampled reaction intermediates to be characterized, and stochastic fluctuations from equilibrium to be investigated [1]. Despite their impact, however, these techniques continue to be significantly limited by the maximum time resolutions that can be achieved while still maintaining acceptable signal-to-noise ratios (SNR) and sufficient experimental throughput (i.e., observation of a statistically significant number of molecules given a feasible experimental effort) [2]. For example, single-molecule wide-field fluorescence microscopy- [3, 4], force spectroscopy- [5], and tethered particle motion [6] approaches are typically limited to time resolutions of milliseconds to hundreds of milliseconds per data point. Consequently, these single-molecule methods often fail to detect or properly characterize biomolecular processes, such as early steps in ligand binding and/or dissociation, structural domain rearrangements, or local folding and unfolding events that occur on the microsecond to millisecond timescale [7, 8]. Although the time resolution of some techniques can be improved so that they might be able to report on these processes, such as by performing single-molecule fluorescence microscopy experiments with a confocal, rather than a wide-field, fluorescence microscope, this often comes at the cost of a significant decrease in the SNR and/or a several orders of magnitude decrease in the experimental throughput, either of which can be as powerful a limitation as the lower time resolution [2].

* Adapted from Kinz-Thompson, C.D., Gonzalez, Jr., R.L. Temporal Super-Resolution, *in preparation*.
Additionally, adapted Kinz-Thompson, C.D.*, Fei, J.*, Gonzalez, Jr., R.L. Temperature-dependent, Single-molecule FRET Studies of Macromolecular Dynamics, *in preparation*.

To push beyond the time-resolution limitations of time-resolved single-molecule techniques without altering their SNRs or experimental throughputs, we have developed a computational approach, which we call Bayesian Inference for the Analysis of Sub-temporal-resolution Dynamics (BIASD), that can infer the rates of transitions between multiple ligand-binding or conformational states (hereafter referred to as “states") of a single molecule from the analysis of any time-resolved, single-molecule experimental signal–even if those rates are faster than the time resolution of the recorded experimental signal. By using Bayesian inference, BIASD employs a natural framework with which to describe the precision that the amount of data collected during the single-molecule experiment will lend to the determination of the parameters governing the single-molecule dynamics. Primarily due to recent developments in computational tractability, Bayesian inference has become a powerful method for the analysis of biophysical data, such as determining the phases of X-ray reflections in X-ray crystallographic studies [9], performing simultaneous phylogenetic analysis of nucleotide and protein datasets [10], elucidating the number of structural classes present in cryogenic electron microscopy images [11], and ascertaining the number of states and the rates of transitions between those states present in single-molecule fluorescence resonance energy transfer (smFRET) efficiency (E_{FRET}) versus time trajectories [12–14]. For a practical introduction to Bayesian inference, see Ref. 15 and Sec. 4.2.2.

Here, we describe the Bayesian inference-based framework underlying BIASD. We then validate BIASD by accurately recovering the known rates of transitions between states and signal values corresponding to each state that were used to generate computer-simulated signal versus time trajectories. Finally, we apply BIASD to experimental data by using it to infer the unknown rates of transitions between states and signal values corresponding to each state from experimentally observed E_{FRET} versus time trajectories–these particular E_{FRET} versus time trajectories have thus far remained challenging to accurately analyze due to the presence of transitions that are fast relative to the time resolution of the E_{FRET} versus time trajectories [16]. Remarkably, the results of our studies demonstrate that, even when the rates of transitions between states are orders of magnitude faster than the time resolution of the signal versus time trajectories, BIASD permits accurate inference of the rates of transitions between these states and the signal values corresponding to these states.

4.2 Bayesian Inference-based Framework Underlying BIASD

In biomolecular systems, functional motions–such as those involved in ligand binding and dissociation processes, or large-scale conformational rearrangements–very often involve the simultaneous formation and/or

disruption of numerous, non-covalent interactions. The relatively low probability of simultaneously forming and/or disrupting these numerous interactions can therefore result in large, entropically dominated, transition-state energy barriers for such functional motions [17, 18]. Consequently, individual biomolecules are generally expected to exhibit effectively discrete and instantaneous transitions between relatively long-lived states [8], an expectation that is consistent with the step-like transitions that are generally observed in time-resolved single-molecule experiments [19]. By definition, such time-resolved single-molecule techniques record the time evolution of an experimentally observable signal originating from an individual molecule (*i.e.*, a signal versus time trajectory) that, ideally, conveys information about the time evolution of the underlying state of that molecule. Correspondingly, the analysis of a single-molecule signal versus time trajectory frequently involves thresholding the trajectory at particular signal values or modeling the trajectory (*e.g.*, using a hidden Markov model (HMM)) such that each data point in the signal versus time trajectory is assigned to a single, specific state of the molecule [20, 21]. As a result, these methods 'idealize' the original signal versus time trajectory into a state versus time trajectory. From the idealized state versus time trajectory, the distribution of dwell times spent in a particular state before undergoing a transition to another state can be used to determine the rates of the transitions between states, and the distribution of observed signal values originating from a particular state can be used to determine the signal value corresponding to that state [1, 22].

An important consideration when idealizing signal versus time trajectories is that whenever an individual molecule undergoes a transition from one state to another, the transition occurs stochastically during the time period, τ, over which the detector collects and integrates the signal to record a data point in the signal versus time trajectory. Thus, the probability that a transition will coincide exactly with the beginning or end of the τ in which it takes place is essentially zero. As a result, when a transition takes place, the signal value that is recorded during that τ does not solely represent either of the states involved in that transition. Instead, it represents the average of the signal values corresponding to the states that are sampled during τ, weighted by the time spent in each of those states. This time averaging makes it imprudent, if not incorrect, to idealize the signal value recorded during such a τ by assigning it to any one particular state, because the molecule will have occupied multiple states during that τ. Notably, when the rates of transitions between states become comparable to or greater than $1/\tau$, there is a large probability that the τs of a signal versus time trajectory will contain one or more transitions, and that, consequently, many of the signal values of the signal versus time trajectory will exhibit this time averaging. Idealization of such signal versus time trajectories, therefore, introduces significant errors into the resulting dwell-time- and signal-value distributions as well as into the

rates of transitions between states and signal values corresponding to those states that are determined from these distributions.

In order to overcome the potential errors associated with determining rates of transitions and signal values from the analysis of idealized state versus time trajectories, BIASD instead determines the rates of transitions and signal values by analyzing the fraction of time that a molecule spends in each state during the τ corresponding to each signal value in a signal versus time trajectory. To illustrate this approach, we consider the case of a single molecule that undergoes stochastic and uncorrelated (*i.e.*, Markovian) and reversible transitions between two states, denoted 1 and 2, (*i.e.*, $1 \rightleftharpoons 2$, with forward and reverse rate constants of k_1 and k_2, respectively) that have unique signal values of ϵ_1 and ϵ_2. If the fraction of time that the molecule spends in state 1 during a particular τ is f, then, because of the two-state nature of the system, the fraction of time that the molecule spends in state 2 during that τ is 1-f. It is important to note that, although the molecule is at equilibrium between states 1 and 2, the value of f for any particular τ will not necessarily be the equilibrium value of $f = (1 + (k_1/k_2))^{-1}$, because τ might not be long enough for sufficient time averaging to occur (*i.e.*, to invoke ergodicity). Instead, each τ will exhibit a different, time-averaged value of f.

The exact value of f for a particular τ will depend upon the molecule's stochastic path through state-space during τ. As such, a probabilistic description of f, which accounts for all possible paths through state-space, is needed to calculate the likelihood of observing a particular value of f during a τ. In particular, for the reversible, two-state system considered here, such a description, which has roots in nuclear magnetic resonance chemical exchange experiments [23] and sojourn-time distributions [24], was first given by Dobrushin [25]; in its use here, this expression depends upon k_1, k_2, and τ, and is derived in Sec. 4.2.1. Experimentally, if the exact values of f, ϵ_1, and ϵ_2 during each τ were known, one would be able to calculate the expected value of the corresponding time-averaged signal, μ, for each τ because it would be the linear combination $\mu = (\epsilon_1 \cdot f) + (\epsilon_2 \cdot (1 - f))$. Unfortunately, the analysis of time-resolved single-molecule experiments deals with the opposite problem–observing a signal, D, during each τ and trying to infer f, ϵ_1, and ϵ_2–with additional uncertainty that is due, in part, to noise in the measurement of D.

A conservative, yet generally applicable, approach to analyzing the value of D recorded during each τ is to treat it as a noisy measurement of μ. By assuming that detection noise (*e.g.*, readout noise) dominates over other possible sources of noise (*e.g.*, fluctuations in laser power) and that such detection noise is effectively uncorrelated, the observed values of D will have a probability that is distributed according to a normal (*i.e.*, Gaussian) distribution with a mean, μ, and a standard deviation, σ, corresponding to the amount of noise in

D. However, since μ depends upon f, which is not an experimental observable, we have no way of knowing the exact value of μ during a measurement, information that would ordinarily be required to calculate the probability of observing a particular value of D. To circumvent this experimental limitation, this dependence upon f can be removed by marginalizing f out of the expression for the stochastic probability distribution of D that was described above. This marginalized probability distribution of D then describes the likelihood of experimentally observing a particular value of D during a τ as a function of the rates of transitions between the states (k_1, and k_2), the signal values corresponding to those states (ϵ_1, and ϵ_2), and the amount of noise in D (σ), regardless of the exact value of f (Fig. 4.1) (See Sec. 4.2.2 for a full derivation of the expression describing the marginalized probability distribution of D for a two-state system).

With such an expression describing the marginalized probability distribution of D, we can then use Bayesian inference to estimate the model parameters governing the kinetics of the single-molecule system (*i.e.*, k_1, k_2, ϵ_1, ϵ_2, and σ) from the series of D that comprise each of the signal versus time trajectories. Unfortunately, Bayesian inference on a multi-parameter system, such as the one described here, results in a multi-dimensional, joint-probability distribution of the model parameters, known as a posterior probability distribution, that is difficult to evaluate [15]. In order to overcome this difficulty, we evaluate the posterior distribution of the model parameters by numerically sampling it using a Markov chain Monte Carlo (MCMC) sampling method [26, 27]. Although alternative methods that approximate the posterior distribution of the model parameters, such as the Laplace approximation or variational inference, might be more computationally tractable, MCMC sampling is advantageous in that, unlike such approximation methods, it can provide an exact result [27]. Regardless of the choice of method, however, the most important aspect of the approach described here is that we can evaluate or estimate the posterior distribution of the model parameters from the series of D that comprise each of the single-molecule signal versus time trajectories in a manner that is completely irrespective of the time resolution of the trajectories.

4.2.1 Distributions of Fractional Occupancies

Consider a single-molecule whose dynamics are governed by the stochastic, two-state system,

$$1 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} 2. \tag{4.1}$$

For a memoryless (*i.e.,* Markovian, or lacking dynamic disorder) two-state system, the individual lifetimes that the single-molecule dwells in states 1 and 2 are exponentially distributed with rate constants k_1 and k_2,

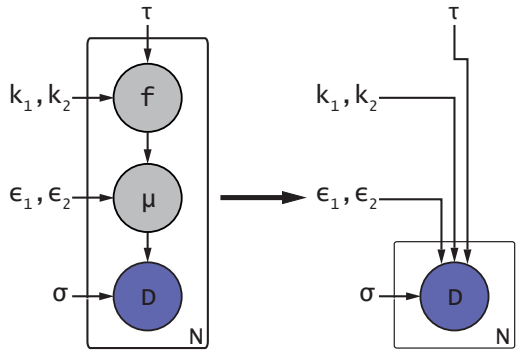

Figure 4.1: Graphical model for two-state BIASD. In BIASD, the dependence of the observed, D, upon the fractional occupancy, f, (Left) is marginalized to yield the graph on the right.

respectively. In this case, the equilibrium probabilities of finding states 1 and 2 occupied are,

$$p_1 = \frac{k_2}{k_1 + k_2}, \text{ and} \tag{4.2}$$

$$p_2 = 1 - p_1 = \frac{k_1}{k_2 + k_2}, \tag{4.3}$$

respectively. However, during a period of observation, $t = 0$ to τ, a single-molecule will begin in one of the two states, and then will processively transition between the two states a random number of times, which is governed by k_1, k_2, and τ. Along these lines, during the observation period, the time spent in the ith state (*i.e.*, the sojourn time), T_i, is also governed k_1, k_2, and τ. For T_1 and $T_2 > 0$ sec, the probability distribution of T_1 is known to be [24, 25],

$$\begin{aligned} P^{\dagger}\left(T_1 | k_1, k_2, \tau\right) = e^{-k_1 T_1 - k_2(\tau - T_1)} \cdot \Bigg[& \left(p_1 k_1 + p_2 k_2\right) \cdot I_0\left(2(k_1 k_2 T_1(\tau - T_1))^{1/2}\right) \\ & + \left(k_1 k_2\right)^{1/2} \left(p_1 \left(\frac{T_1}{\tau - T_1}\right)^{1/2} + p_2 \left(\frac{\tau - T_1}{T_1}\right)^{1/2} \right) \cdot I_1\left(2(k_1 k_2 T_1(\tau - T_1))^{1/2}\right) \Bigg], \end{aligned} \tag{4.4}$$

where p_1 and p_2 are the probability of finding the single-molecule is states 1 and 2, respectively, (*i.e.,* at equilibrium, $p_1 = \frac{k_2}{k_1+k_2}$ and $p_2 = \frac{k_1}{k_1+k_2}$), and I_0 and I_1 are modified Bessel functions of the first kind. This distribution is not normalized because it lacks density at $T_1 = 0$ and $T_1 = \tau$. These conditions represent the situations when the single-molecule is exclusively in state 1 or in state state 2, respectively, for the entire duration of the observation period. *Ad hoc*, at equilibrium, the contributions to the probability distribution at these points should be the equilibrium population-weighted probability that the single-molecule does not undergo a transition out of the ith during the observation period; this probability is the survival probability (*i.e.,*

1 - the cumulative distribution function) of the exponential distribution, which is $e^{-k_i\tau}$. Therefore, considering these contributions, the entire probability distribution of time spent in state 1 during the observation period is,

$$P\left(T_1|k_1,k_2,\tau\right) = P^{\dagger}\left(T_1|k_1,k_2,\tau\right) + p_1 e^{-k_1\tau}\cdot\delta\left(\tau - T_1\right) + p_2 e^{-k_2\tau}\cdot\delta\left(T_1\right), \tag{4.5}$$

where δ is the Dirac delta function, which ensures that these terms only contribute when the single-molecule is exclusively in one of the two states. This expression also provides the probability distribution of T_2 during an observation period by interchanging the rate constants. If the observation period, τ, is known, then the probability distribution of the total time spent in state 1 during an observation period given in Eqn. (4.5), can be transformed into the probability distribution of the fraction of time spent in state 1 during the observation period, f. This transformation is,

$$P\left(f|k_1,k_2,\tau\right) = P(T_1 = f\cdot\tau|k_1,k_2,\tau)\cdot|J|^{-1}; \text{ where } f \equiv \frac{T_1}{\tau}, \text{ so } \frac{\partial f}{\partial T_1} = \frac{1}{\tau} \text{ and } |J|^{-1} = \tau, \tag{4.6}$$

where J is the Jacobian. Therefore, by plugging Equation (4.5) into Equation (4.6) and making the substitution $T_1 = f\tau$,

$$\begin{aligned} P\left(f|k_1,k_2,\tau\right) = {} & \frac{k_2}{k_1+k_2}\tau e^{-k_1\tau}\cdot\delta\left(\tau-(f\tau)\right) + \frac{k_1}{k_1+k_2}\tau e^{-k_2\tau}\cdot\delta\left((f\tau)\right) \\ & + 2\frac{k_1k_2}{k_1+k_2}\tau e^{-k_1(f\tau)-k_2(\tau-(f\tau))}\cdot\left[I_0\left(y\right) + \frac{k_2\left(f\tau\right)+k_1\left(\tau-(f\tau)\right)}{y}\cdot I_1\left(y\right)\right], \end{aligned} \tag{4.7}$$

where $y \equiv 2\sqrt{k_1k_2\left(f\tau\right)\left(\tau-(f\tau)\right)} = 2\tau\sqrt{k_1k_2f\left(1-f\right)}$.
Noting the identity,

$$\delta\left(ax\right) = \frac{1}{|a|}\cdot\delta\left(x\right), \tag{4.8}$$

we can simplify Equation (4.7) to yield,

$$\begin{aligned} P\left(f|k_1,k_2,\tau\right) = {} & \frac{k_2}{k_1+k_2}e^{-k_1\tau}\cdot\delta\left(1-f\right) + \frac{k_1}{k_1+k_2}e^{-k_2\tau}\cdot\delta\left(f\right) \\ & + 2\frac{k_1k_2}{k_1+k_2}\tau e^{-(k_1f+k_2(1-f))\tau}\cdot\left[I_0\left(y\right) + \frac{\left(k_2f+k_1(1-f)\right)\tau}{y}\cdot I_1\left(y\right)\right]. \end{aligned} \tag{4.9}$$

Above, we have chosen to simplify Equation (4.9) so that it is apparent that it is equivalent to the expression of Berezhkovskii and coworkers, and that used by Gopich and coworkers [23, 28].

Though, we note several typos in those publications. In Berezhkovskii *et al.*, in Equation (3.16), the term $\exp\left(\frac{k_1 k_2 \tau}{2}\right)$ should read $\exp\left(\frac{k_1 k_2 \tau}{s}\right)$, and in Equation (3.18), after the $\delta\left(\tau - T\right)$, there should be a plus sign [23]. In Gopich *et al.*, in Equation (S47) the arguments in the Dirac delta functions in the first two terms should be interchanged [28]. We mention these typos, because below we replicate and elaborate upon the derivation of $P\left(f|k_1, k_2, \tau\right)$ from Berezhkovskii *et al.* [23], using the variables defined earlier instead of the definitions from that work. This derivation shall be useful for Section 4.2.1.

Following Berezhkovskii and coworkers [23], we shall derive $P\left(f|k_1, k_2, \tau\right)$ for the memoryless, two-state system described above .

$$P\left(f|k_1, k_2, \tau\right) = p_1 \cdot P\left(f|k_1, k_2, \tau, 1\right) + p_2 \cdot P\left(f|k_1, k_2, \tau, 2\right), \tag{4.10}$$

where $P\left(f|k_1, k_2, \tau, i\right)$ is the probability distribution of f given that the single-molecule begins in the ith state. As before, $P\left(f|k_1, k_2, \tau, i\right)$ is calculated by transforming $P\left(T_1|k_1, k_2, \tau, i\right)$ using

$$P\left(f|k_1, k_2, \tau, i\right) = P\left(T_1 = f\tau|k_1, k_2, \tau, i\right) \cdot |J|^{-1}. \tag{4.11}$$

To calculate $P\left(T_1|k_1, k_2, \tau, 1\right)$, we define an indicator function,

$$\chi\left(t\right) = \begin{cases} 1 & \text{if the single-molecule is in state 1 at time t} \\ 0 & \text{otherwise} \end{cases}, \tag{4.12}$$

which means that,

$$T_1 = \int_0^\tau dt \cdot \chi\left(t\right). \tag{4.13}$$

Now, consider the Fourier transform of $P\left(T_1|k_1, k_2, \tau, 1\right)$,

$$\hat{P}\left(\omega|k_1, k_2, \tau, 1\right) = \int_{-\infty}^{\infty} dT_1 \cdot e^{i\omega T_1} \cdot P\left(T_1|k_1, k_2, \tau, 1\right), \tag{4.14}$$

and note that this is the expectation value $\langle e^{i\omega T_1} \rangle$ over all trajectories starting in state 1; so, we can write

$$\hat{P}\left(\omega|k_1,k_2,\tau,1\right) = \langle e^{i\omega \int_0^\tau dt \cdot \chi(t)} \rangle. \tag{4.15}$$

These types of averages can be written as a sum two terms,

$$\hat{P}\left(\omega|k_1,k_2,\tau,1\right) = c_1\left(\omega|k_1,k_2,\tau,1\right) + c_2\left(\omega|k_1,k_2,\tau,1\right), \tag{4.16}$$

where $\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ satisfies $\frac{d\mathbf{c}}{d\tau} = \mathbf{R} \cdot \mathbf{c}$ solved with the condition $\mathbf{c}\left(\tau = 0\right) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and

$$\mathbf{R} = \mathbf{K} + \begin{pmatrix} i\omega \\ 0 \end{pmatrix} = \begin{pmatrix} i\omega - k_1 & k_2 \\ k_1 & -k_2 \end{pmatrix}.$$

This can be solved using the Green's function approach in Laplace space,

$$\mathbf{c}(\tau) = \mathbf{G}(\tau) \cdot \mathbf{c}(\tau = 0), \tag{4.17}$$

by solving

$$\mathbf{G}(\tau) = \mathcal{L}^{-1}\left(\tilde{\mathbf{G}}(s)\right), \tag{4.18}$$

where $\mathcal{L}^{-1}$ is the inverse Laplace transform, and

$$\tilde{\mathbf{G}}(s) = \left(s \cdot \mathbf{I} - \mathbf{R}\right)^{-1} = \begin{pmatrix} s + k_1 - i\omega & -k_2 \\ -k_1 & s + k_2 \end{pmatrix}^{-1}, \tag{4.19}$$

where $\mathbf{I}$ is the identity matrix. Now, for a 2x2 matrix,

$$\left(\mathbf{A}\right)^{-1} = \frac{1}{|\mathbf{A}|} \cdot \begin{pmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{pmatrix}; \tag{4.20}$$

so, from Equation (4.19),

$$\tilde{\mathbf{G}}(s) = \frac{1}{(s+k_1-i\omega)(s+k_2)-k_1k_2} \cdot \begin{pmatrix} s+k_2 & k_2 \\ k_1 & s+k_1-i\omega \end{pmatrix}, \tag{4.21}$$

but, because of the initial conditions of $\mathbf{c}(\tau=0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and the linearity of the Laplace transform, we only need to consider the first column of the matrix, *i.e.*,

$$\check{\mathbf{c}}(s) = \frac{1}{(s+k_1-i\omega)(s+k_2)-k_1k_2} \cdot \begin{pmatrix} s+k_2 \\ k_1 \end{pmatrix}, \tag{4.22}$$

therefore, the solution of Equation (4.16) in Laplace space is,

$$\check{P}(\omega|k_1,k_2,s,1) = \check{c}_1 + \check{c}_2 = \begin{pmatrix} 1 & 1 \end{pmatrix} \cdot \check{\mathbf{c}}(s) = \frac{(s+k_2)+k_1}{(s+k_1-i\omega)(s+k_2)-k_1k_2}. \tag{4.23}$$

Expanding the denominator, and then simplifying yields,

$$\begin{aligned} \check{P}(\omega|k_1,k_2,s,1) &= \frac{(s+k_1+k_2)}{s(s+k_1+k_2)-i\omega(s+k_2)} \\ &= \frac{1}{s-i\omega\left(\frac{s+k_2}{s+k_1+k_2}\right)}. \end{aligned} \tag{4.24}$$

The probability distribution in Equation (4.24) is the Fourier transform and Laplace transform of our initially desired probability distribution, $P(T_1|k_1,k_2,\tau,1)$. To obtain our desired distribution, we must perform an inverse Fourier transform, $\mathcal{F}^{-1}$, and an inverse Laplace transform, $\mathcal{L}^{-1}$. Starting with the inverse Fourier transform, we write

$$\tilde{P}(T_1|k_1,k_2,s,1) = \mathcal{F}^{-1}\left(\check{P}(\omega|k_1,k_2,s,1)\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \cdot e^{-i\omega T_1} \cdot \check{P}(\omega|k_1,k_2,s,1). \tag{4.25}$$

Noting the identity that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \cdot e^{-i\omega T_1}\frac{1}{a-i\omega b} = \frac{1}{b}\cdot e^{-\frac{aT_1}{b}}, \tag{4.26}$$

if $a = s$ and $b = \frac{s+k_2}{s+k_1+k_2}$, then

$$\begin{aligned}
\tilde{P}(T_1|k_1,k_2,s,1) &= \left(\frac{s+k_1+k_2}{s+k_2}\right)\cdot e^{-T_1\left(\frac{s(s+k_1+k_2)}{(s+k_2)}\right)} \\
&= \left(1+\frac{k_1}{s+k_2}\right)\cdot e^{-T_1\left(\frac{s^2+(k_1+k_2)s+(k_1k_2-k_1k_2)}{(s+k_2)}\right)} \\
&= \left(1+\frac{k_1}{s+k_2}\right)\cdot e^{-T_1\left(\frac{(s+k_1)(s+k_2)}{(s+k_2)}-\frac{k_1k_2}{s+k_2}\right)} \\
&= \left(1+\frac{k_1}{s+k_2}\right)\cdot e^{-T_1\left(s+k_1-\frac{k_1k_2}{s+k_2}\right)}. \qquad (4.27)
\end{aligned}$$

Finally, the inverse Laplace transform will yield the desired probability distribution,

$$\begin{aligned}
P(T_1|k_1,k_2,\tau,1) &= \mathcal{L}^{-1}\left(\tilde{P}(T_1|k_1,k_2,s,1)\right) \\
&= \mathcal{L}^{-1}\left(\left(1+\frac{k_1}{s+k_2}\right)\cdot e^{-T_1\left(s+k_1-\frac{k_1k_2}{s+k_2}\right)}\right) \\
&= e^{-k_1T_1}\cdot\mathcal{L}^{-1}\left(\left(1+\frac{k_1}{s+k_2}\right)\cdot e^{-T_1\left(s-\frac{k_1k_2}{s+k_2}\right)}\right). \qquad (4.28)
\end{aligned}$$

This can be simplified more by using the identity,

$$\mathcal{L}^{-1}\left(\tilde{F}(s+a)\right) = e^{-a\tau}\mathcal{L}^{-1}\left(\tilde{F}(s)\right) \qquad (4.29)$$

and setting $a = k_2$ to yield,

$$\begin{aligned}
P(T_1|k_1,k_2,\tau,1) &= e^{-k_1T_1}e^{-k_2\tau}\mathcal{L}^{-1}\left(\left(1+\frac{k_1}{s}\right)\cdot e^{-T_1\left(s-k_2-\frac{k_1k_2}{s}\right)}\right) \\
&= e^{-k_1T_1-k_2\tau+k_2T_1}\mathcal{L}^{-1}\left(\left(1+\frac{k_1}{s}\right)\cdot e^{-T_1\left(s-\frac{k_1k_2}{s}\right)}\right). \qquad (4.30)
\end{aligned}$$

Then, use the identity

$$\mathcal{L}^{-1}\left(e^{-sT_1}\cdot\tilde{F}(s)\right) = F(\tau-T_1)\cdot H(\tau-T_1), \qquad (4.31)$$

where H is the Heaviside step function, in order to change the τ to $\tau - T_1$ for

$$P(T_1|k_1,k_2,\tau,1) = e^{-(k_1T_1-k_2(\tau-T_1))}H(\tau-T_1)\mathcal{L}^{-1}\left(\left(1+\frac{k_1}{s}\right)\cdot e^{\frac{T_1k_1k_2}{s}}\right), \qquad (4.32)$$

where from now the inverse Laplace transform is over $\tau - T_1$ instead of τ; this follows for the remainder of the derivation. Knowing the following identities,

$$\mathcal{L}\left(I_0\left(2\sqrt{at}\right)\right) = \frac{1}{s}\cdot e^{a/s}, \text{ and} \tag{4.33}$$

$$\mathcal{L}\left(\frac{1}{\sqrt{t}}\cdot I_1\left(2\sqrt{at}\right)\right) = \frac{1}{\sqrt{a}}\cdot\left(e^{a/s}-1\right), \tag{4.34}$$

we can rearrange the previous expression in order to find these forms present in the Laplace transformed expression. Noting that, since the Laplace transform is an integral transform, it has linearity,
i.e., $\mathcal{L}\left(aF(t)+bG(t)\right) = a\tilde{F}(s)+b\tilde{G}(s)$, we write

$$\begin{aligned}
P(T_1|k_1,k_2,\tau,1) &= e^{-(k_1T_1-k_2(\tau-T_1))}\cdot H(\tau-T_1)\cdot\mathcal{L}^{-1}\left(e^{k_1k_2T_1/s}-1+1+\frac{k_1}{s}e^{k_1k_2T_1/s}\right)\\
&= e^{-(k_1T_1-k_2(\tau-T_1))}\cdot H(\tau-T_1)\cdot\left[\mathcal{L}^{-1}\left(1\right)+\mathcal{L}^{-1}\left(e^{k_1k_2T_1/s}-1\right)+\mathcal{L}^{-1}\left(\frac{k_1}{s}e^{k_1k_2T_1/s}\right)\right]\\
&= e^{-(k_1T_1-k_2(\tau-T_1))}\cdot H(\tau-T_1)\cdot\left[\delta\left(\tau-T_1\right)+\sqrt{\frac{k_1k_2T_1}{\tau-T_1}}\cdot I_1\left(2\sqrt{k_1k_2T_1(\tau-T_1)}\right)\right.\\
&\left.+\sqrt{\frac{k_1k_2T_1}{\tau-T_1}}k_1\cdot I_0\left(2\sqrt{k_1k_2T_1(\tau-T_1)}\right)\right]\\
&= e^{-k_1\tau}\cdot\delta\left(\tau-T_1\right)+e^{-(k_1T_1-k_2(\tau-T_1))}\cdot H(\tau-T_1)\cdot\left[\sqrt{\frac{k_1k_2T_1}{\tau-T_1}}\cdot I_1\left(2\sqrt{k_1k_2T_1(\tau-T_1)}\right)\right.\\
&\left.+\,k_1\cdot I_0\left(2\sqrt{k_1k_2T_1(\tau-T_1)}\right)\right].
\end{aligned} \tag{4.35}$$

This expression can then be transformed into our desired probability distribution, $P(f|k_1,k_2,\tau,1)$, using Equation (4.11) to yield,

$$\begin{aligned}
P(f|k_1,k_2,\tau,1) &= \tau\cdot e^{-k_1\tau}\cdot\delta(\tau-f\tau)+\tau\cdot e^{-(k_1f\tau-k_2(\tau-f\tau))}\cdot H(\tau-f\tau)\left[\sqrt{\frac{k_1k_2f\tau}{\tau-f\tau}}\cdot I_1(y)+k_1\cdot I_0(y)\right]\\
&= e^{-k_1\tau}\cdot\delta(1-f)+\tau\cdot e^{-\tau(k_1f-k_2(1-f))}\cdot\left[\sqrt{\frac{k_1k_2f}{1-f}}\cdot I_1(y)+k_1\cdot I_0(y)\right],
\end{aligned} \tag{4.36}$$

where $y\equiv 2\tau\sqrt{k_1k_2f(1-f)}$, the Dirac delta function identity in Equation (4.8) was used to cancel the τ in the first term, and the Heaviside step function was removed because it was redundant with the support of f.

In order to derive $P(f|k_1,k_2,\tau,2)$, a nearly identical approach is used. Though, an *ad hoc* derivation can be made by beginning with Equation (4.35), relabeling T_1 to T_2, interchanging k_1 and k_2, and performing the

transformation to $P(f|k_1, k_2, \tau, 2)$ using the definition $T_2 = (1-f)\tau$. Finally, using those results, we revisit Equation (4.10) and plug in to get

$$P(f|k_1,k_2,\tau) = \frac{k_2}{k_1+k_2}\left(e^{-k_1\tau}\delta(1-f) + k_1\tau\left(I_0(y) + \sqrt{\frac{k_2 f}{k_1(1-f)}}\cdot I_1(y)\right)\cdot e^{-\tau z}\right) + \frac{k_1}{k_1+k_2}\left(e^{-k_2\tau}\delta(f) + k_2\tau\left(I_0(y) + \sqrt{\frac{k_1(1-f)}{k_2 f}}\cdot I_1(y)\right)\cdot e^{-\tau z}\right), \tag{4.37}$$

where $z \equiv k_1 f + k_2(1-f)$. This is what was to be demonstrated. This expression is equivalent to that derived earlier in Equation (4.9).

N-state and Non-Markovian Systems Much like the two-state system described above, the probability distributions for the fractional occupancies in an N-state kinetic scheme (*i.e.*, one where a single-molecule can exist in N different states and transition between states i and j with rate constant k_{ij}) can be also be derived from a transformation of the probability distribution of the time spent in a state during a measurement period. Such a distribution of the total sojourn time spent in a particular state has been discussed by several authors [29, 30]. However, in order to use such a distribution for the likelihood function described in Section 4.2.2, this distribution must be the joint-probability distribution of the total sojourn times for the N - 1 independent states; there are N-1 independent states because $\sum_{i=1}^{N} T_i = \tau$.

Thus far, we have only described kinetic schemes where the lifetimes of a single-molecule in a particular state are assumed to be exponentially distributed. This is not always the case, as, experimentally, single-molecule systems have been observed that have power-law distributed lifetimes, or rather have memory effects or exhibit dynamic disorder (*i.e.*, the lifetimes in a state are time-dependent). Much like the N-state system, such effects can be accounted for with BIASD by deriving the expected probability distribution of the fractional occupancies for the desired kinetic model. For the two-state system, Berezhkovskii and coworkers describe a method to easily incorporate non-exponential lifetimes into their derivation, which was replicated above [23]. Briefly, the form for the desired probability distribution given the particular type of lifetimes desired is derived in joint Laplace-Fourier space. It seems that the inverse Fourier transform may be relatively simple to compute analytically, while the inverse Laplace transform can be more difficult. In such a case, the Laplace transform can be numerically inverted [31]. For instance, the Talbot algorithm as described by Abate and

Whitt,

$$F(t,M) = \frac{2}{5t}\sum_{k=0}^{M-1} Re\left(\gamma_k \cdot \tilde{F}\left(\frac{\delta_k}{t}\right)\right), \tag{4.38}$$

where M is an integer related to the precision of the numerical inversion, and

$$\delta_0 = \frac{2M}{5},\ \delta_k = \frac{2k\pi}{5}(\cot(k\pi/M) + i),\ \text{for } 0 < k < M,\ \text{and} \tag{4.39}$$

$$\gamma_0 = \frac{1}{2}e^{\delta_0},\ \gamma_k = \left[1 + i(k\pi/M)(1 + \cot^2(k\pi/M)) - i\cot(k\pi/M)\right] \cdot e^{\delta_k},\ \text{for } 0 < k < M. \tag{4.40}$$

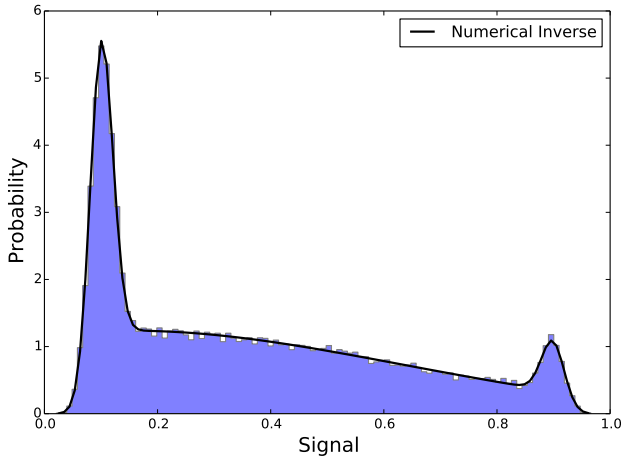

Figure 4.2: BIASD Likelihood function calculated by numerical inverse of Laplace transform. A signal verse time trajectory was simulated for 50000 sec with $\tau = 1$ s, $\epsilon_1 = 0.1$, $\epsilon_2 = 0.9$, $\sigma = 0.02$, $k_1 = 1\ \text{s}^{-1}$, $k_2 = 2\ \text{s}^{-1}$. The histogram of this trajectory is shown in blue. The BIASD likelihood function described in Section 4.2.2 was calculated using a probability of fractional occupancy of state 1 that was calculated by numerically inverting the Laplace transform, where M = 64. This result is exact to within numerical error.

Calculating the numerical inversion for the memoryless, two-state system described above, yields the exact result of Equation (4.9). This expression for the numerically inverted probability distribution can be used to calculate the likelihood function for BIASD as described in Section 4.2.2. This likelihood function fits simulated two-state data exactly (Fig. 4.2).

Notably, Gopich and Szabo have shown how to approximate an N-state system using Gaussians [32]. This would allow for fast and analytical calculations of the occupation probabilities, however, it is specialized to photon-counting fluorescence resonance energy transfer experiments (*i.e.*, experiments with poisson counting statistics), and assumes that during a single measurement period, transitions only occur between two states.

4.2.2 Bayesian Inference Overview

Bayesian inference allows for the parameters describing an initial hypothesis, θ, to be modified to account for new data, D. Mathematically, this process can be written using Bayes' rule,

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p\,(D)} = \frac{p(D|\theta) \cdot p(\theta)}{\sum \left(p(D|\theta) \cdot p(\theta)\right)} \propto p(D|\theta) \cdot p(\theta) \tag{4.41}$$

which is analogous to saying that the probability of the hypothesis after having seen the data (the posterior probability, $p(\theta|D)$) is proportional to the product of the probability of the data given the hypothesis (the likelihood, $p(D|\theta)$) and the initial probability of the hypothesis itself (the prior probability, $p(\theta)$). For a highly recommended introduction to Bayesian inference written for scientists such as chemists and physicists, see Ref. 15. To calculate the probability of the hypothesis given a series of single-molecule observations, one can use Bayes' rule, and a model for the data, which provides an expression for likelihood. Unfortunately, direct enumeration of the posterior distribution can be computational expensive, especially with a large number of parameters. As an alternative, methods such as Markov chain Monte Carlo (MCMC) explore the expansive hyper-volume of the posterior distribution in a much more economical manner. Sufficient sampling of the posterior probability distribution then provides random samples, which can be used to statistically describe the posterior probability distribution (see below).

Likelihood Function A simple model for the signal emitted from a two-state single-molecule is that each datapoint is normally distributed about a mean that is a linear combination of the mean emissions from each of the two states. For instance, if a molecule is in state 1 for the entire i$^{\text{th}}$ period of observation, the observed emission signal, D_i, would be centered at ϵ_1. Similarly, pure state 2 emission would be centered at ϵ_2. An observation period with a fractional occupation of these states would then have an emission signal centered at a linear combination of these ϵ (*i.e.,* $\epsilon_1 f + \epsilon_2\,(1-f)$). To account for factors such as detection noise, the probability of observing a particular emission value given the fractional occupation during that observation period can be taken as a normal distribution centered at $\mu = \epsilon_1 f + \epsilon_2\,(1-f)$ with a standard deviation, σ,

$$P\left(D_i|\epsilon_1, \epsilon_2, \sigma, f\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-1}{2\sigma^2}\left(D_i - (\epsilon_1 f + \epsilon_2(1-f))\right)^2}. \tag{4.42}$$

Unfortunately, the exact fractional occupation during the observation period is not a parameter that is observed during an experiment; so, Equation (4.42) is useless. Instead, the probability distribution of the

fractional occupation can be derived; the exact form for the memoryless two state model, $P(f|k_1, k_2, \tau)$, was given in Section 4.2.1. Knowing this probability distribution allows the dependence upon f to be removed from Equation (4.42); this process is known as marginalization. The resulting, marginalized probability distribution using the exact probability distribution given in Equation (4.9) is,

$$P(D_i|\epsilon_1, \epsilon_2, \sigma, k_1, k_2, \tau) = \int_0^1 df \cdot P\left(D_i|\epsilon_1, \epsilon_2, \sigma, f\right) \cdot P\left(f|k_1, k_2, \tau\right) \tag{4.43}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \left[\frac{k_2}{k_1+k_2} e^{-\frac{1}{2}\left(\frac{D_i-\epsilon_2}{\sigma}\right)^2 - k_1\tau} + \frac{k_1}{k_1+k_2} e^{-\frac{1}{2}\left(\frac{D_i-\epsilon_1}{\sigma}\right)^2 - k_2\tau} \right.$$
$$\left. + \frac{2k_1k_2\tau}{k_1+k_2} \int_0^1 df \cdot e^{-\frac{1}{2}\left(\frac{D_i-(\epsilon_1 f+\epsilon_2(1-f))}{\sigma}\right)^2 - (k_1 f + k_2(1-f))\tau} \cdot \left(I_0(y) + \frac{(k_2 f + k_1(1-f))\tau}{y} \cdot I_1(y) \right) \right] \tag{4.44}$$

In this form, the integral can be computed numerically, for instance, with a Gaussian quadrature method. Having experimentally observed a signal versus time trajectory, $D = [D_1, \cdots, D_N]$, we can calculate the likelihood of observing this particular trajectory, $L \equiv P\left(D|\Theta\right)$; this expression is the product of the probability of the observed signal during each measurement period, $P\left(D_i|\epsilon_1, \epsilon_2, \sigma, k_1, k_2, \tau\right)$, which is

$$L = P\left(D|\Theta\right) = P\left(D|\epsilon_1, \epsilon_2, \sigma, k_1, k_2, \tau\right) = \prod_{i=1}^{N} P\left(D_i|\epsilon_1, \epsilon_2, \sigma, k_1, k_2, \tau, \right) \tag{4.45}$$

This assumes that the datapoints are independent and identically distributed. That assumption is reasonable for the case of fast dynamics (*i.e.*, $k_1\tau$ or $k_2\tau > 1$), where subsequent signal measurements are made during a time when the single-molecule occupies several different states. For the case of slow dynamics, if there are a sufficient number of measurements in the signal versus time trajectory such that the single-molecule can be considered at equilibrium (*i.e.*, $\bar{f} \approx k_2/(k_1+k_2)$), this assumption amounts to a first-order expansion of the fractional occupation probability – thus, we believe it to be a reasonable approximation. Moreover, as we show in the main text, under this assumption, the likelihood function performs well in this regime.

Finally, given a large number of measurements in D, L can easily underflow on a computer; so, it is better computed on a log scale, since this is monotonic. The resulting log-likelihood, $\ln(L)$, is

$$\ln(L) = \ln\left(P\left(D|\Theta\right)\right) = \sum_{i}^{N} \ln\left(P\left(D_i|\epsilon_1, \epsilon_2, \sigma, k_1, k_2, \tau\right)\right). \tag{4.46}$$

Unfortunately, in the case of the exact probability distribution for the memoryless, two-state system, be-

cause of the four terms in the expression for $P\left(D_i|\epsilon_1, \epsilon_2, \sigma, k_1, k_2, \tau\right)$, the summation in the log-likelihood in Equation (4.46) does not simplify this expression.

Prior Distributions To some, the idea of a prior distribution is concerning, and even seems subjective. However, it is not a cause for concern. Probabilities are always conditional, and setting a prior probability is simply equivalent to encoding these initial conditions. For instance, if a particular parameter describes the magnitude of something, it must be a greater than or equal to zero. Already, in defining this particular parameter as a magnitude, we have begun to describe our initial knowledge of the system (*i.e.*, to set the prior). In this sense, since even the act of defining a parameter is 'subjective', setting a prior should not bother anyone much more than does defining a model system. Objectivity is ingrained in this process since, given the same initial conditioning, two rational individuals should produce the same prior distribution [15]. Thus, a prior probability distribution should just be considered a manner with which to formally and mathematically define a system.

Certain types of parameters are more conveniently represented by certain types of probability distributions. Just considering the allowed values of particular parameters and the support of particular distributions, probabilities and ratios are conveniently represented with beta distributions, magnitudes and rates are conveniently represented with gamma distributions, and positions and signals are conveniently represented by normal distributions. These distributions are given in Appendix B.

For the prior distributions used for BIASD when analyzing single-molecule fluorescence resonance energy transfer (smFRET) experiments, we typically employ beta distributions for ϵ_1 and ϵ_2, because they are ratios of distances or photon counts. We use a gamma distribution for σ, because it represents the magnitude of the noise. Finally, for k_1 and k_2, we also use gamma distributions, because these rate constants represent a number per time and therefore cannot be negative.

Consider a model parameter and how to go about setting the corresponding prior probability distribution. Given no additional previous knowledge of a system, it seems reasonable that the prior probability distribution should be uniform for every possibility (*e.g.*, the idea of equal, *a priori* probability from statistical mechanics). This is sometimes known as an uninformative prior, since no possibility is favored over the others; though, this is slightly misleading, because the support for the distribution must be chosen beforehand, and as such, prior knowledge has already been encoding into the prior. On the other hand, if one were absolutely certain about the value of a parameter, a Dirac delta function could be chosen as the prior probability distribution– this choice is a very strong statement about the system. This is not recommended, because it is equivalent

to saying that there is no possible way that the parameter value is different than specified; even the values of physical constants change in time as measurements become more precise.

Most prior probability distributions should be somewhere between the uniform distribution and a delta function. A weakly informative prior would be more similar to a uniform distribution, while a strongly informative prior would be more similar to a delta function; such similarities can be quantified with, for instance, the Kullback-Leibler divergence, but, intuitively, less peaked distributions (*i.e.*, large variance) are less informative.

To find the parameters that define the desired prior probability distribution, one can use the expressions for the mean and variance of the distribution, set the variance to the desired magnitude, and, using the mean as a constraint, solve for the parameter values. For beta distributions, smaller parameters (α and β) yield less informative distributions; $\alpha = 1$, and $\beta = 1$ is flat, and thus an uninformative prior. For gamma distributions, smaller values of α yield less informative distributions; $\alpha = 1$ is an exponential distribution. For normal distributions, larger values of σ^2 yield less informative distributions.

Markov Chain Monte Carlo Sampling Markov chain Monte Carlo (MCMC) allows for the efficient sampling of high dimensional space (such as a posterior probability distribution) by taking a random walk along the distribution (see Ref. 27 for an introduction). To do so, random steps are drawn from a proposal distribution, and these steps are either accepted or rejected based upon an acceptance criteria. The sequence of steps forms a Markov chain that is ideally able to traverse the relevant (*i.e.*, higher probability) hyper-volume of the N-dimensional space; thus eliminating the need to enumerate the entire N-dimensional space. To do so effectively, the acceptance rate (AR) of steps must not be too large, because then the Markov chain will not diffuse far from its initial location. Likewise, the AR must not be too small, because then the Markov chain will not diffuse along the probability distribution at all.

One common MCMC method is the Metropolis algorithm (generalized by the Metropolis-Hastings (MH) algorithm) [26]. Here, a proposed step is created by drawing random numbers from a distribution (*e.g.*, normal), and the N variables describing the initial location in N-dimensional space, $\mathbf{x}$, are incremented by these values to generate a proposed location for the next step, $\mathbf{x}^*$. The acceptance criteria for making this step is based upon the relative probabilities of the proposed and the initial locations. If the proposed location has a higher probability, the step will be accepted; otherwise, the step might still be accepted, but with certain

probability. Explicitly, steps are accepted with probability

$$A = \text{min}\left(1, \frac{P(x^*)}{P(x)}\right), \tag{4.47}$$

where A is the acceptance probability. After calculating A, a random number is then drawn from a uniform distribution between 0 and 1, and the proposed step is accepted if the random number is greater than A, and rejected if the random number is less than A.

Often, a proposed step is chosen from a normal distribution. As such,

$$x^* \sim x + \mathcal{N}\left(\mu = 0, \sigma_{\text{MH}}\right), \tag{4.48}$$

where σ_{MH} is the standard deviation. However, this is not entirely appropriate for certain parameters which are not supported from $(-\infty, \infty)$. In such a case, certain proposed moves might not exist in the parameter's space, and this makes the MCMC sampling less efficient. To circumvent such unsupported proposals, when performing MCMC in this work, one can transform the parameter into a space with support from $(-\infty, \infty)$, generate a proposed step using a normal distribution, and then transformed this proposed, transformed parameter back into the original parameter space. For parameters between 0 and 1 (*i.e.*, ϵ_i for smFRET), which might have beta distributed priors, the following (logit) scale can be used,

$$\begin{aligned} y &\equiv \ln(\frac{x}{1-x}), \text{ and} \\ x &= \frac{e^y}{1+e^y}, \end{aligned} \tag{4.49}$$

where y is the transformed parameter, and x is the initial parameter. Similarly, for parameters which are positive (*e.g.*, σ, k_i), which might have gamma distributed priors, the following (logarithmic) scale can be used,

$$\begin{aligned} y &\equiv \text{log}_{10}(x), \text{ and} \\ x &= 10^y. \end{aligned} \tag{4.50}$$

Additionally, when performing MCMC in this work, we adaptively tuned the AR of the Markov chains. This adaptive tuning allows the Markov chain to optimally explore the N-dimensional space. The intuition behind this process is that if steps are rejected too often, then the distance between steps is too large and should

be decreased; if steps are accepted too often, then the distance between steps is too small and should be increased. For an N-dimensional space, the optimal AR is $\sim$0.25 [33]. To tune the Markov chain, the AR is calculated for some number of previous steps; we chose five. Subsequently, the variance of the normal distribution used to generate proposed moves is modified by setting

$$\sigma_{\text{MH,new}} = \sigma_{\text{MH,old}} \cdot \left(\sqrt{\frac{2}{.25}} \cdot \text{AR} \right), \tag{4.51}$$

where σ_{MH} is the standard deviation of the normal distribution in Equation (4.48).

Another variant of MCMC is component-wise MH. This approach is similar to the MH algorithm, but instead of varying all N variables for each proposed step, only one variable is allowed to move at a time. Thus, the proposed location differs from the initial location by the chosen variable. This proposal is accepted in the same manner as in the Metropolis algorithm (see Equation (4.47)). However, once a proposed step is accepted, the process is repeated with a different variable, until the Markov chain has moved through all N-dimensions. The advantage of this approach is that it allows for separate adaptive tuning of the individual dimensions.

4.2.3 Analysis Using BIASD

Here, we will summarize the process employed to analyze the data presented in the main text using BIASD. Where necessary, all signal versus time trajectories were pre-processed to ensure two-state behavior. Effectively, this means truncating the trajectories at the first photophysical anomaly, such as photobleaching. To begin analysis, prior probability distributions, as described in Sec. 4.2.2, were chosen for the five parameters in the model, ϵ_1, ϵ_2, σ, k_1, and k_2. Unless otherwise specified, the prior probability distributions for the BIASD parameters were chosen to be:

$$\begin{aligned}
\epsilon_1 &\sim \text{Beta}\left(\alpha_{\epsilon_1}, \beta_{\epsilon_1}\right) \\
\epsilon_2 &\sim \text{Beta}\left(\alpha_{\epsilon_2}, \beta_{\epsilon_2}\right) \\
\sigma &\sim \text{Gamma}\left(\alpha_{\sigma}, \beta_{\sigma}\right) \\
k_1 &\sim \text{Gamma}\left(\alpha_{k_1}, \beta_{k_1}\right) \\
k_2 &\sim \text{Gamma}\left(\alpha_{k_2}, \beta_{k_2}\right)
\end{aligned}$$

ϵ_1 and ϵ_2 were chosen to be Beta-distributed to demonstrate the utility of BIASD for smFRET experiments. The observable in smFRET experiments is the FRET efficiency, E_{FRET}, which is approximated by

$$E_{\text{FRET}} \approx \frac{n_{\text{acceptor}}}{n_{\text{donor}} + n_{\text{acceptor}}} \approx \frac{I_{\text{acceptor}}}{I_{\text{donor}} + I_{\text{acceptor}}}, \tag{4.52}$$

where n is the number of photons observed from the specified fluorophore, and I is the detector registered intensity for the wavelength range of the specified fluorophore. Since the observable is supposed to the ratio of photons, it is reasonable to model this as the probability of successful emission of a photon from the donor fluorophore, which is ostensibly Beta-distributed if the probability is constant. The remaining three parameters were chosen to be Gamma-distributed because of they are supported between 0 and ∞.

The hyper-parameters describing these prior probability distributions were chosen so that, in the case of the synthetic titration data, the mean of distribution corresponded to the synthetic value of the parameter. For the pretranslocation complex data, the means were chosen from global, non-linear least-squares fits to Equation (4.44) – individual signal versus time trajectories often show significant deviation from the global fits. Once the desired variance in the distribution was decided upon (they varied depending upon the desired analysis), the hyper-parameters were calculated by moment-matching the distributions to the first and second moments, $E[X]$ and $E[X^2]$. These calculations were:

Distribution	α	β
Beta	$E[X] \cdot \left(\frac{E[X](1-E[X])}{(E[X^2]-E[X]^2)} - 1 \right)$	$(1 - E[X]) \cdot \left(\frac{E[X](1-E[X])}{(E[X^2]-E[X]^2)} - 1 \right)$
Gamma	$\frac{E[X]^2}{(E[X^2]-E[X]^2)}$	$\frac{E[X]}{(E[X^2]-E[X]^2)}$

The posterior probability distributions for each individual signal versus time trajectory were calculated using the log-likelihood function in Equation (4.46), using the exact form of the marginalized probability distribution for each datapoint that is shown in Equation (4.44). The integral in the latter equation is computed numerically using a Gaussian quadrature method. Again, this likelihood function assumes independent and identically distributed datapoints.

To quantify the posterior probability distributions, a MCMC sampler was written in C. This sampler initialized each Markov chain at the mean of the prior probability distribution, and took component-wise steps using the MH algorithm. Additionally, the step-sizes (σ_{MH}) for each component were adaptively tuned every five steps. Only one Markov chain was run for each signal versus time trajectory, and it consisted of at least 2000 steps for each component; since the chains were initialized at the prior mean, and the prior mean was often set to the value used for the simulations, no burn-in phase was used. These choices were

made because of the relatively long computational time required to calculate the log-likehood function for an entire signal versus time trajectory, which is required to be performed multiple times each MH step; on a 3.6 GHz AMD FX-8150 processor, choosing a successful sample for a single component took $\sim$ 0.2 sec, and therefore each complete MCMC step took $\sim$ 1 sec.

Because the computational expense of obtaining samples constrained the total number of samples we obtained, sufficient mixing of the Markov chain was ensured by checking autocorrelation functions of the components of the Markov chains sampled by MCMC. These autocorrelation functions should indicate sufficient lack of correlation to promote adequate sampling of the posterior probability distribution by MCMC. In most cases, the autocorrelation length was much shorter than the number of MH samples in the Markov chains performed on one of the sets of simulated signal versus time trajectories in the main text. However, for the cases of only the very slowest and very fastest dynamics (*i.e.*, lowest and highest concentration of synthetic titrant, respectively), the autocorrelation lengths were slightly longer. However, they still became uncorrelated in an adequate time, and even running a longer Markov chain (5000 steps) for these cases did not affect the posterior probability distribution.

To quantify the MCMC sampled posterior probability distributions, we calculated the expectation values, $E[X]$ and $E[X^2]$, where X is one of the five BIASD parameters, from the MCMC samples. This was done by weighting the samples by their posterior probability, which had first been normalized per Markov chain, and then normalized so that the ensemble (*e.g.*, all of the signal versus time series for a particular condition) summed to one. More explicitly, we calculated

$$E[X] = \sum_{i,j} X_{i,j} \cdot \frac{1}{\sum_j 1} \cdot \frac{P_{i,j}\left(\Theta_j | D_j\right)}{\sum_i P_{i,j}\left(\Theta_j | D_j\right)}, \tag{4.53}$$

where i indexes the MCMC samples, j indexes the signal versus time trajectories, and $X_{i,j}$ and $P_{i,j}$ are the sample values and the posterior probabilities of the i$^{\text{th}}$ sample from the Markov chain of the j$^{\text{th}}$ signal versus time trajectory. Calculation of $E[X^2]$ used the analogous weighting. These first and second moments were then used to calculate the parameters of the appropriate probability distribution. The relevant parameters for the prior distribution were determined by moment matching as described in the table above. Credible intervals were then calculated from these ensemble, marginalized, MCMC-sampled posterior probability distributions.

4.2.4 Experimental Methods

Simulating Sub-temporal-resolution Trajectories An initial state was selected randomly, and then sequential random lifetimes were drawn from two exponential distributions with the appropriate rate constants. A random starting point during the first lifetime was selected from a uniform distribution for the measurement initiation time (t = 0 sec). The fractional occupancies of each state during each sequential τ were then calculated from the sequence of lifetimes. These fractional occupation versus time trajectories were turned into signal versus time trajectories by computing μ, and then adding normally distributed noise with standard deviation, σ.

Idealizing Trajectories Idealized signal versus time trajectories were analyzed with a two-state maximum likelihood HMM (see Ref. 27), and used to analyze individual signal versus time trajectories; the convergence threshold employed was 1x10^{-3}. Rate constants from state i to j were then calculated from the off diagonals of the transition matrix, p_{ij}, as $k_{ij} = -\ln(1 - p_{ij})/\tau$. Signal versus time trajectories were also idealized by thresholding any measurement period with signal less than $(\epsilon_2 - \epsilon_1)/2 + \epsilon_1$ into state one, and otherwise into state two. Rate constants from state i to j were then calculated from the maximum-likelihood estimate of the transition probability.

BIASD Calculations A Gaussian quadrature method for numerical integration and modified Bessel function evaluations, both from the GNU Scientific Library 1.16 [34], were used to numerically integrate the BIASD likelihood function. The posterior probability distribution was sampled using a component-wise MCMC method using the Metropolis-Hastings algorithm. The step sizes for the sampling method were adaptively tuned to produce an acceptance ratio of 0.25. Additionally, based upon the choice of prior distributions, parameters that were beta distributed were sampled by diffusion in a logit scaled space, and parameters that were gamma distributed were sampled by diffusion in a log-10 scaled space. Ensemble-normalized, posterior-probability-weighted first and second moments, E[X] and E[X^2], of the samples were used to calculate the corresponding estimates of the posterior probability distribution by moment-matching to the initial choice of prior probability distribution.

Processing PRE^{-A} E_{FRET} Data Cy3 and Cy5 fluorescence intensity, I_{Cy3} and I_{Cy5}, versus time trajectories of the PRE^{-A} complexes from the study by Wang and coworkers [16] were transformed into E_{FRET} versus time trajectories by calculating $E_{FRET} = I_{Cy5}/(I_{Cy3} + I_{Cy5})$ at each τ. Outliers where E_{FRET} < -0.4 or E_{FRET} >

1.4 were then removed. The first photobleaching event in a time trajectory was detected by thresholding $I = (I_{\text{Cy3}} + I_{\text{Cy5}})$ at $1/e$ of the mean value of the first 20 data-points per trajectory, and this time was manually adjusted were necessary to ensure the two-state behavior of the E_{FRET} versus time trajectories. The number of single-molecules retained in the 22, 25, 28, 31, 34, and 37 °C datasets were 66, 102, 77, 84, 65, and 104, respectively. Histograms of each dataset were fit to the BIASD likelihood function using non-linear least-squares fitting, and used to set the means of the prior distributions employed for ϵ_{GS1}, ϵ_{GS2}, and σ (beta distributed, beta distributed, and gamma distributed, respectively). The prior probability distributions for k_{GS1} and k_{GS2} were both taken to be gamma distributions with a mean determined from the histogram fitting and α = 1 to ensure a weakly-informative prior.

4.3 Analysis of Computer-Simulated Single-Molecule Signal Versus Time Trajectories Reporting on the Kinetics of a Ligand Binding and Dissociation Process

To validate BIASD, we simulated single-molecule signal versus time trajectories that report on the dynamics of binding and dissociation of a ligand to its target biomolecule–a receptor. These dynamics were simulated according to the two-state, reversible, Markovian kinetic scheme that was introduced in the previous section. In this example, k_1 and k_2 represent the pseudo-first-order rate constant of ligand binding to the receptor and the first-order rate constant of ligand dissociation from the receptor, respectively. Correspondingly, ϵ_1 and ϵ_2 represent the signal values of the receptor in the ligand-free state and the ligand-bound state, respectively. As such, k_1 is treated as a variable whose value at a particular ligand concentration is given by $k_1 = k_1^\star \cdot [L]$, where $k_1^\star$ is the second-order rate constant for binding of the ligand to the receptor and [L] is the ligand concentration, and k_2 is treated as a constant whose value is not dependent on [L]. To begin, we simulated a set of 10 individual receptors for 60 sec, where k_1 and k_2 were both set to 8 sec^{-1}, using the stochastic simulation algorithm [35]. The magnitudes of k_1 and k_2 were chosen so as to be comparable to 1 / τ (*i.e.*, 10 sec^{-1}), because, in this regime, the process of idealizing signal versus time trajectories into state versus time trajectories begins to introduce errors into the analysis of the transition rate constants and signal distributions of particular states (Sec. 3.3). Moreover, since k_1 = k_2 for these trajectories, arbitrarily choosing the value of $k_1^\star$ to be 3 μM^{-1} sec^{-1} (*i.e.*, [L] = $\sim$2.67 μM) yielded an equilibrium dissociation constant, K_D, in the μM range for this simulated system, which is typical of many naturally occurring, weakly

interacting, ligand-receptor systems [2]. Notably, the finite length of the simulated signal versus time trajectories (60 sec) is consistent with a typical length of time that single-molecule E_{FRET} versus time trajectories can be recorded before fluorophore photobleaching terminates data collection; this creates a limitation to the amount of kinetic information contained in each simulation. Time-averaged signal versus time trajectories were then generated from these simulations by using values of ϵ_1 = 0.1, ϵ_2 = 0.9, σ = 0.04, and τ = 0.1 sec. These values are typical of experimental E_{FRET} versus time trajectories recorded using a wide-field, total internal reflection fluorescence (TIRF) microscope. Expanding from this first set of simulated signal versus time trajectories, we then repeated this simulation process while systematically increasing and decreasing the [L] by two orders of magnitude relative to the initial value of [L] = K_D = $\sim$2.67 μM in order to emulate a titration experiment–generating 16 additional sets of simulated signal versus time trajectories. Finally, all of these sets of simulated signal versus time trajectories were then analyzed using both threshold- and HMM-based idealization methods as well as using BIASD as described in Sec. 4.2.3 in order to obtain estimates of underlying simulation parameters. As shown in Figure 4.3, the values of k_1 and k_2 obtained using idealization-based methods are highly inaccurate. The rate constants were systematically underestimated across the entire range of [L]s that were simulated, and this worsens with increasing [L]. However, it is notable that the values of k_1 and k_2 obtained using these methods are highly precise, exhibiting standard error of the means (SEMs) that are exceedingly narrow across all [L]s–a misleading consequence of using these methods. Somewhat surprisingly, k_1 was underestimated at [L]s corresponding to a regime where k_1 much less than 1 / τ. This underestimation arises from the fact that k_2 is a constant that is comparable to 1 / τ across all [L]s in this simulation. As such, the transitions into the ligand-bound state of the receptor that are missed by idealization of the signal versus time trajectories due to the rapid transitions back to the ligand-free state, result in an over-estimation of the dwell-times in the ligand-free state and, consequently, an underestimation of k_1 (*c.f.*, Sec. 3.3).

With regard to the values of ϵ_1 and ϵ_2 obtained using idealization-based methods, Fig. 4.5 demonstrates that, while these methods can accurately determine the value of ϵ_1 if the receptor preferentially occupies the ligand-free state (low [L]) or ϵ_2 if the receptor preferentially occupies the ligand-bound state (high [L]), the time averaging caused by the large, simulated values of k_1 shift the inferred value of ϵ_1, sometimes quite significantly, toward the simulated value of ϵ_2, and vice versa.

In contrast to the idealization-based methods, the values of k_1 and k_2 obtained using BIASD are highly accurate (Fig. 4.3). The simulated values of k_1 and k_2 are well encompassed by the posterior probability distribution across the entire range of [L]s that were simulated. In addition, these results are remarkably

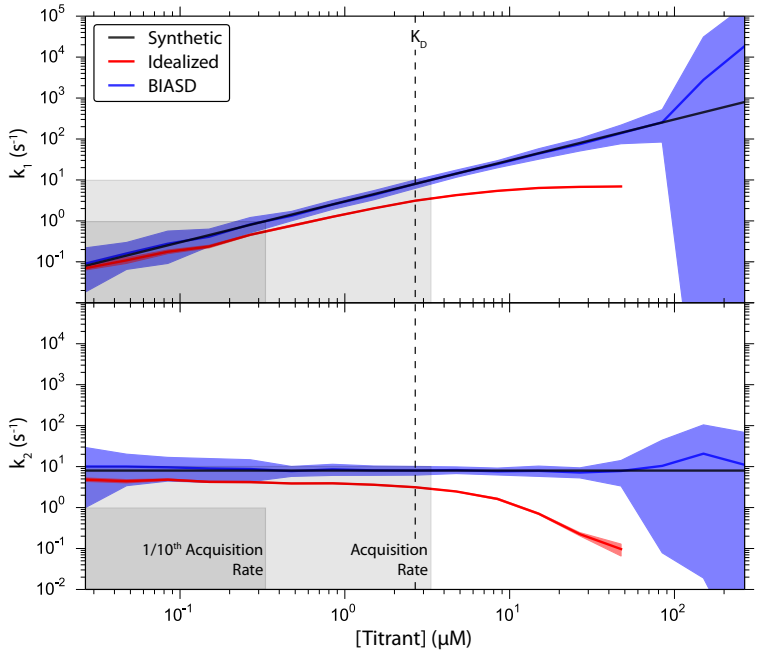

Figure 4.3: Analysis of k_1 and k_2 using BIASD (blue) and idealization-based (red) methods for a computer-simulated titration of a ligand to a receptor. Concentration of the titrant was varied two orders of magnitude above and below the concentration where the equilibrium occupation probability of both states is 0.5 (K_D). The regions where the rate constants are less than 1/10th of the acquisition rate, 1 / τ, is shown in dark grey; the regions where the rate constants are less than the acquisition rate are shown in light grey. The simulated rate constants are plotted as the black lines. The red area denotes the region containing one standard error of the mean rate constants calculated by idealizing the signal versus time trajectory. The blue area denotes the average 95% credible interval (CI) of the posterior probability distributions from analysis with BIASD using weakly informative priors (Fig. 4.4).

precise, as the posterior probability distributions are strikingly narrow over a range of [L] s that corresponds to k_1 being over an order of magnitude slower to over an order of magnitude faster than 1 / τ. Importantly, the results are insensitive to the type of prior probability distributions (Fig. 4.6) that BIASD uses for the analysis (*i.e.*, the initial knowledge of k_1, k_2, ϵ_1, ϵ_2, and σ) (Figs. 4.6 and 4.7). At the lower [L]s, the broadening of the posterior probability distribution, and the implied limitations to the precisions for estimating k_1 and k_2 arises from the finite amount of information regarding k_2 and, to a lesser extent, ϵ_2 that is contained in signal versus time trajectories that exhibit very little occupation of the ligand-bound state of the receptor. Consistent with this interpretation, these posterior probability distributions at the lower [L]s closely approximate the theoretical, maximum precision for inferring k_1 and k_2 given the 60 sec length of each signal versus time trajectory (Fig. 4.8). Likewise, at the higher [L]s, the broadening of the posterior probability distribution and the implied limitations to the precision for estimating k_1 and k_2 that is observed arises from the finite amount of information regarding $k_1$and ϵ_1 that is contained in signal versus time trajectories that exhibit very little occupation of the ligand-free state of the receptor. As a consequence of this finite amount of

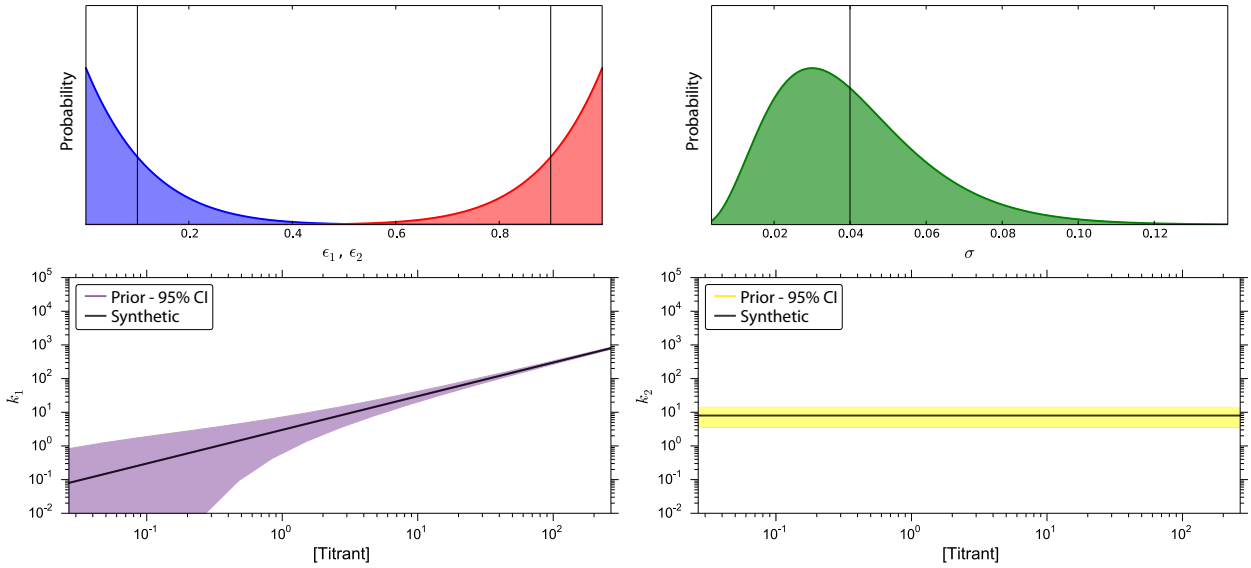

Figure 4.4: Plots of weakly informative prior probability distributions used with BIASD to analyze the synthetic titration. Black lines denote the simulated value of each parameter in the synthetic titration. (Top Left) The priors for the signal means were beta distributed– ϵ_1 (blue) with $\alpha = 1$ and $\beta = 9$, and ϵ_2 (red) with $\alpha = 9$ and $\beta = 1$. These parameters were kept constant for each concentration. (Top Right) The prior for the noise was gamma distributed, with $\alpha = 4$ and $\beta = 100$. (Bottom Left and Right) Prior probability distributions for k_1 and k_2 are shown as 95% CIs at each concentration point in the titration, because the simulated rate constants for k_1 was different at each concentration. For both rate constants, the prior probability distribution at each concentration point was a gamma distribution with $\alpha = k_i$ and $\beta = 1$.

information, many reciprocal pairs of k_1 and ϵ_1 values are, at most, congruous with the data (i.e., a faster k_1 and a smaller ϵ_1, or a slower k_1 and a larger ϵ_1). Consistent with this interpretation, reanalysis of the signal versus time trajectories simulated at the highest [L]s under conditions in which the prior probability distributions of ϵ_1 and ϵ_2 were strengthened results in much increased precisions for k_1 and k_2 (Fig. 4.9). In an experimental situation, strengthening of the prior probability distributions of ϵ values can be guided by the results of experiments performed under conditions in which one or the other state is preferentially occupied. In the current example of a ligand binding and dissociation process, this would correspond to setting a narrow prior probability distribution for ϵ_1 using the distribution of ϵ_1 values that are observed in experiments recorded in the absence of ligand or at the lowest ligand concentrations and setting a tight prior distribution for ϵ_2 using the distribution of ϵ_2 values that are observed in experiments recorded at highest ligand concentrations. In the case of large-scale conformational rearrangements, one could similarly use a ligand concentration, temperature, or mutation that preferentially stabilizes one or the other state, or, alternatively, one could use molecular modeling to estimate the distribution of ϵ values corresponding to one state and/or the other.

With regard to the values of ϵ_1 and ϵ_2 obtained using BIASD, Fig. 4.10 demonstrates that these values

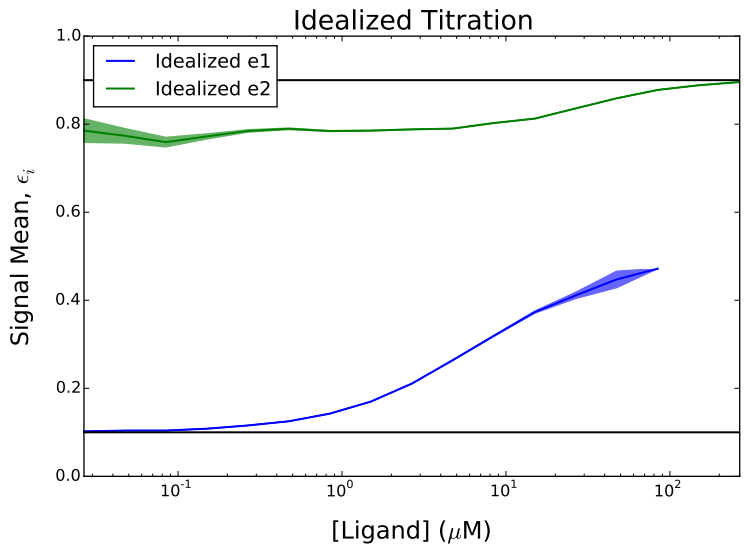

Figure 4.5: Plot of ϵ_1 and ϵ_2 as analyzed by idealizing the synthetic titration data. A threshold was set at 0.5, which is exactly midway $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.9$ (black lines) – the values used to simulate the synthetic titration data in the main text. In a given signal *versus* time trajectory, the mean and SEM of the data points below or above the threshold were calculated for the signal values of ϵ_1 (blue) and ϵ_2 (green), respectively. The plotted lines are the average of the means for the signal *versus* time trajectories at each [L]. The colored area represents the average of the means $\pm$ the average SEM at each [L].

were accurately inferred regardless of the value of [L], even when the idealization-based methods drastically misestimate these values (Fig. 4.5). Moreover, ϵ_1 and ϵ_2 were inferred with high precision across all values of [L] with two intuitive exceptions. Because the receptor primarily occupies the ligand-bound state under high [L] conditions, those signal versus time trajectories contain little information about the signal value corresponding to the ligand-free state (ϵ_1). As expected, the most uncertainty in the value of ϵ_1 occurs at high [L]. Similarly, the most uncertainty in the value of ϵ_2 occurs at low [L], because the ligand-free state of the receptor is primarily occupied, and therefore the signal versus time trajectories contain little information on the signal value corresponding to the ligand-bound state (ϵ_2). This lack of precision is not an artifact of BIASD, but rather a consequence of the amount of information contained in the finite-length signal versus time trajectories. Thankfully, because of its underlying Bayesian inference framework, BIASD provides a natural description of this uncertainty present in the collected data unlike the more traditional, maximum-likelihood methods. Finally, BIASD was also able to accurately and precisely infer σ from the simulated signal versus time trajectories (Fig. 4.10).

In summary, we were able to use BIASD to obtain accurate and precise posterior probability distributions for k_1, k_2, ϵ_1, ϵ_2, and σ across the entire range of [L]s that were simulated. Notably, BIASD was even successful when the rates of transitions between states in the simulated, single-molecule signal versus time

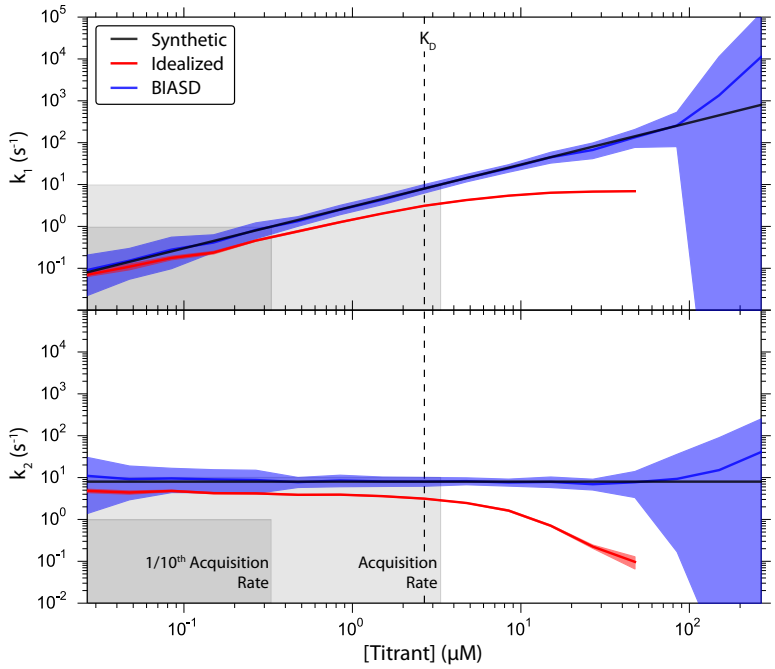

Figure 4.6: Analysis of k_1 and k_2 from a synthetic titration by BIASD and idealized methods using alternative priors. The weakly informative priors utilized for this plot are described in Figure 4.7. Concentration of the titrant was varied two decades above and below where equilibrium occupation probability of both states is 0.5 (K_D). The regions where the rate constants are less than 1/10th of the acquisition rate is shown in dark grey; the regions where the rate constants are less than the acquisition rate are shown in light grey. The synthetic rates from the simulation are plotted as the black lines. The red area denotes the region containing one standard error of the mean rate constants calculated by idealizing the signal versus time trajectory. The blue area denotes the 95% CI of the posterior distributions from analysis with BIASD using weakly informative priors. Notably, these posterior distributions are nearly identical to those obtained using the weakly informative priors described in Fig. 4.4 (*c.f.* Fig. 4.3).

trajectories were much slower then 1 / τ, although we note that, in this regime, the conventional analysis of idealizing the signal versus time trajectories is as accurate and much more computationally efficient. Most importantly, BIASD was able to accurately and precisely infer the rates of transitions between states and the signal values corresponding to those states for simulated, single-molecule signal versus time trajectories in which the rates of transitions between states were nearly two orders of magnitude faster than 1 / τ and about three orders of magnitude faster than that where conventional idealization of signal versus time trajectories begins to yield significant errors in the rates of transitions between states.

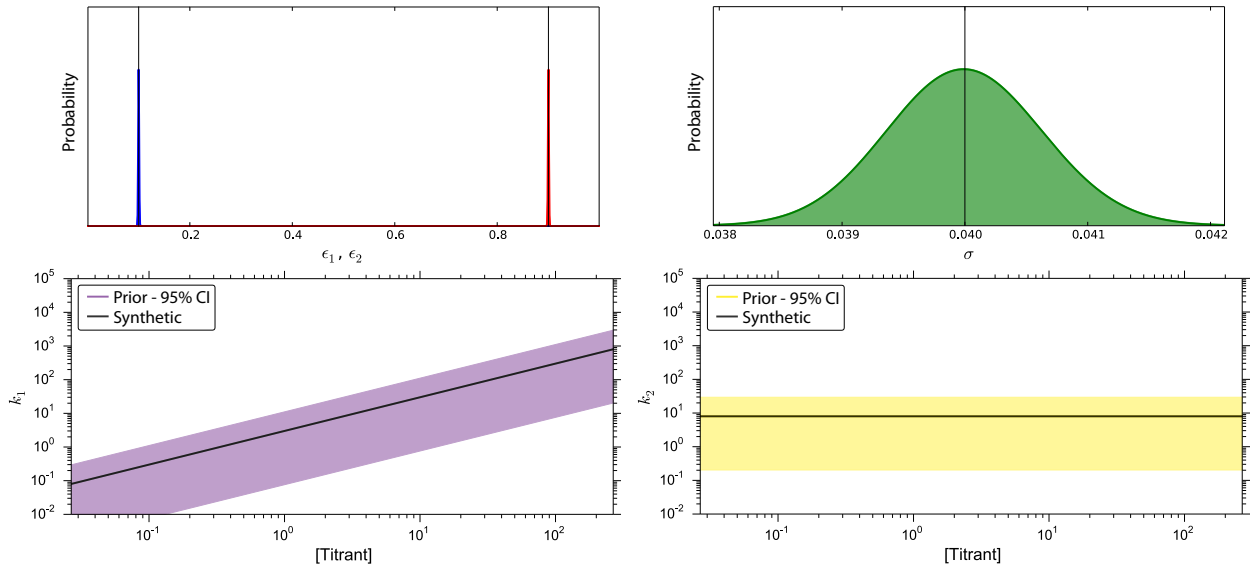

Figure 4.7: Alternative, weakly informative prior probability distributions used to analyze the synthetic titration. (Top Left) The priors for the signal means were beta distributed–ϵ_1 (blue) with $\alpha = 1 \times 10^4$ and $\beta = 9 \times 10^4$, and ϵ_2 (red) with $\alpha = 9 \times 10^4$ and $\beta = 1 \times 10^4$. The distributions appear narrow on this scale, but still have substantial density between 0 and 1. These parameters were kept constant for each concentration. (Top Right) The prior for the noise was gamma distributed, with $\alpha = 4 \times 10^3$ and $\beta = 1 \times 10^5$. (Bottom Left and Right) Prior probability distributions for k_1 and k_2 are shown as 95% CI at each concentration point in the titration, because the simulated rate constants for k_1 was different at each concentration. For both rate constants, the prior probability distribution at each concentration point was a gamma distribution with $\alpha = 1$ and $\beta = 1/k_i$; these forms are exponential distributions.

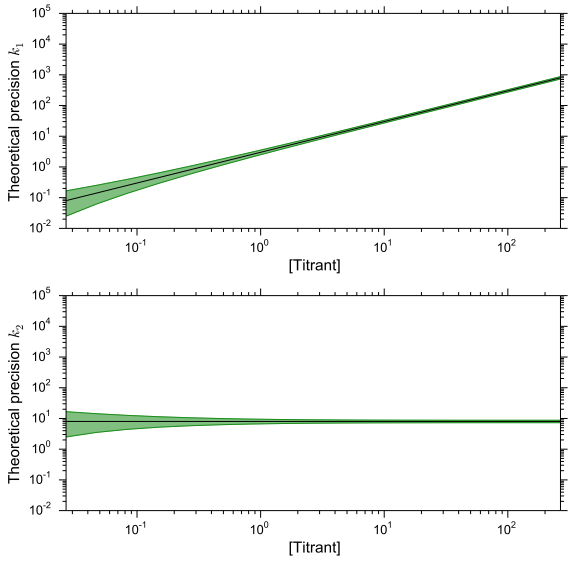

Figure 4.8: Theoretical precision of determining rate constants for synthetic titration given the average number of observed transitions. Given the rate constants of the two-state system (black), the finite observation length of each signal verse time trajectory might not allow sufficient number of transitions to occur before data collection ends in order to precisely determine the rate constants. Under the conditions of the synthetic titration, the green areas denote the 95% CI for the inferred rate constants.

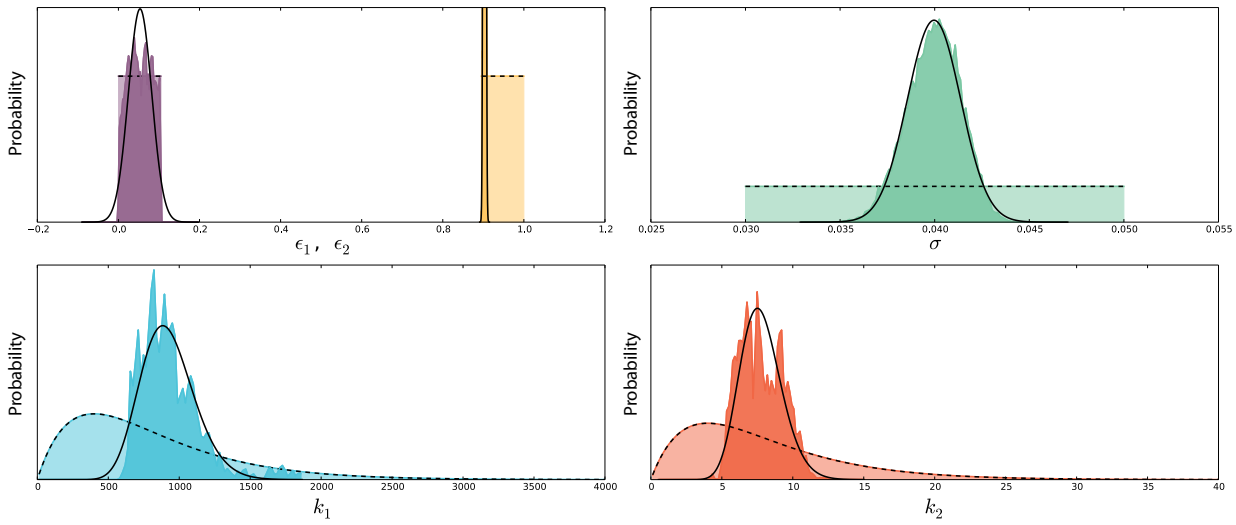

Figure 4.9: Posterior probability distributions for highest concentration in synthetic titration with strict priors. By using uniform priors such that $\epsilon_1 \in [0.0, 0.1]$, $\epsilon_2 \in [0.9, 1.0]$, and $\sigma \in [0.03, 0.05]$, which could have been deduced by first processing the lower concentration data, both the posterior distributions for k_1 and k_2 were tightly clustered around their respective synthetic values of 800 s^{-1} and 8 s^{-1}. Specifically, 95% CI and mean for k_1 are 550 to 1427 s^{-1} and 938 s^{-1}, and for k_2 are 5.1 to 11.0 s^{-1} and 7.8 s^{-1}. Notably, while these results are very accurate, they are also much more precise than those obtained by allowing ϵ_1 and ϵ_2 to freely vary; as such, BIASD can accurately and precisely determine the rate constants of dynamics approximately two orders of magnitude faster than the acquisition rate.

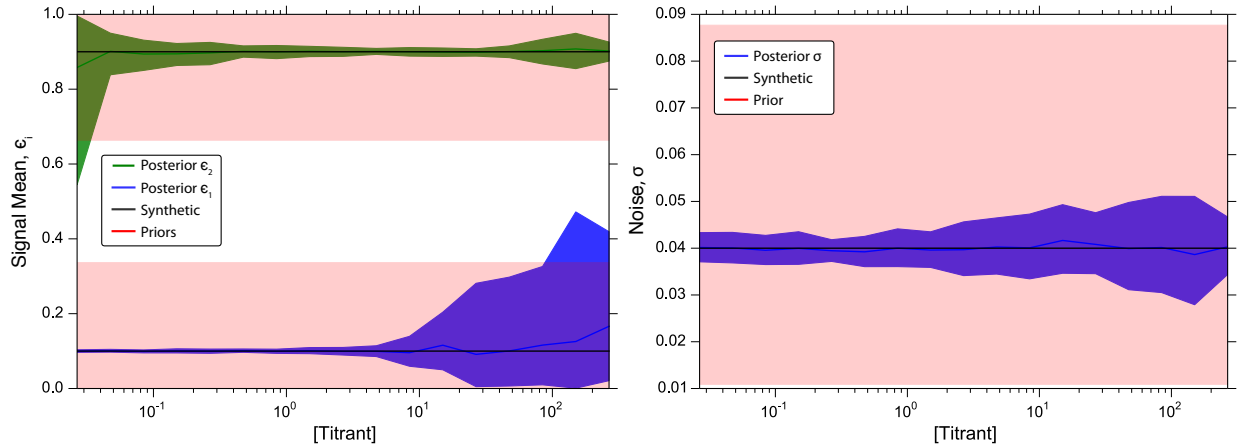

Figure 4.10: Posterior probability distributions of ϵ_1, ϵ_2, and σ for the synthetic titration. The weakly informative priors utilized for this plot are described in Figure 4.4. (Left) The 95% CI for ϵ_1 (blue area) and ϵ_2 (green area) as a function of titrant concentration. Solid blue and green lines denote the mean of the posterior distributions of ϵ_1 and ϵ_2, respectively. The synthetic values use to simulate the synthetic titration are shown in black. The red areas denote the 95% CI of the prior distributions. (Right) The blue area is the 95% CI for σ–the measurement noise–as a function of titrant concentration. Similarly, the black line is the synthetic value of σ, and the red area denotes the 95% CI of the prior probability distribution.

4.4 Analysis of Experimentally Observed Single-Molecule E_{FRET} Versus Time Trajectories Reporting on the Kinetics of a Large-Scale Conformational Rearrangement

To demonstrate the utility of BIASD in the analysis of experimental data, we chose to analyze experimentally observed, single-molecule E_{FRET} versus time trajectories reporting on a large-scale rearrangement of the ribosome–the essential, two-subunit, biomolecular machine that is universally responsible for translating messenger RNAs (mRNAs) into proteins in all living cells. During the elongation stage of protein synthesis (reviewed in Ref. 36), amino acids are added to the nascent polypeptide chain by aminoacyl-transfer RNAs (tRNAs) that deliver each subsequent amino acid to the ribosome as they bind in the ribosomal aminoacyl-tRNA binding (A) site, and then receive the nascent polypeptide chain from the peptidyl-tRNA currently in the peptidyl-tRNA binding (P) site. Following this transfer of the nascent polypeptide chain, the newly elongated peptidyl-tRNA in the A site and newly deacylated-tRNA in the P site must translocate through the ribosome to the P site and to the ribosomal tRNA exit (E) site, respectively, in order to prepare the ribosome for subsequent rounds of elongation. Prior to this translocation event, the ribosomal pre-translocation (PRE) complex undergoes stochastic, thermally driven fluctuations between two major, on-pathway conformational states that we refer to as global state 1 (GS1) and global state 2 (GS2), defining a dynamic equilibrium, GS1 $\rightleftharpoons$ GS2 (Fig. 4.11) (*c.f.*, Sec. 1.3.2). These transitions between GS1 and GS2 constitute large-scale rearrangements of the PRE complex that involve relative rotations of the ribosomal subunits, reconfigurations of the ribosome-bound tRNAs, and repositioning of a ribosomal structural domain known as the L1 stalk (Fig. 4.11) [37].

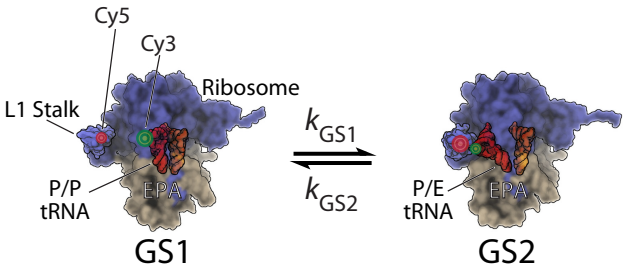

Figure 4.11: Schematic of PRE Complex GS1 $\rightleftharpoons$ GS2 equilibrium. Approximate positions of the donor and acceptor fluorophores for the L1-tRNA labeling scheme in the smFRET experiments of Wang and coworkers [16] are shown as green and red circles, respectively. The size of the fluorophores denotes the relative fluorescence in each state due to resonance energy transfer.

Recently, we have used a PRE complex analog lacking an A site-bound tRNA (*i.e.*, a PRE^{-A} complex) that contains a FRET donor (Cy3)-labeled P site-bound tRNA, and a FRET acceptor (Cy5)-labeled L1 stalk (Fig. 4.11) to conduct smFRET studies of the GS1 $\rightleftharpoons$ GS2 equilibrium. Using a wide-field, TIRF microscope, we were able to collect E_{FRET} versus time trajectories from individual PRE^{-A} complexes in a temperature-controlled, microfluidic observation flowcell at various temperatures [16]. These experiments were motivated by the fact that determination of the temperature dependence of the rate constants for the transitions between GS1 and GS2 would enable an analysis of the thermodynamic properties of the transition state energy barrier that controls the GS1 $\rightarrow$ GS2 and GS2 $\rightarrow$ GS1 conformational rearrangements. Unfortunately, due to the increase in thermal energy with increasing temperature, the rate constants for the transitions between GS1 and GS2 increased such that, at the highest temperatures, the E_{FRET} versus time trajectories contained a significant number of time-averaged data points. Accordingly, the E_{FRET} versus time trajectories at these higher temperatures could not be properly idealized, precluding the determination of the rate constants for the transitions between GS1 and GS2 at these higher temperatures and, correspondingly, the thermodynamic analysis of the transition states that control the GS1 $\rightarrow$ GS2 and GS2 $\rightarrow$ GS1 conformational rearrangements [16].

To overcome these limitations, here we have used BIASD to analyze the sets of E_{FRET} versus time trajectories of PRE^{-A} complexes that were collected at 22, 25, 28, 31, 34, and 37 °C by Wang and coworkers [16]. In doing so, we assume that the GS1 $\rightleftharpoons$ GS2 equilibrium can be represented by the Markovian, reversible, two-state kinetic scheme discussed earlier. In this kinetic scheme, k_{GS1} and k_{GS2} represent the rate constants for the GS1 $\rightarrow$ GS2 and GS2 $\rightarrow$ GS1 conformational rearrangements, respectively. Correspondingly, ϵ_{GS1} and ϵ_{GS2} represent the E_{FRET} values of GS1 and GS2, respectively. Using this approach, all six sets of E_{FRET} versus time trajectories were analyzed using BIASD as described in Secs. 4.2.4 and 4.2.3, providing estimates of k_{GS1}, k_{GS2}, ϵ_{GS1}, ϵ_{GS2}, and σ for each E_{FRET} versus time trajectory at each temperature.

Unfortunately, we cannot speak to the accuracy of results obtained through the analysis of experimental, rather than simulated, data. However, in an attempt to validate the accuracy of the values of k_{GS1} and k_{GS2} obtained using BIASD (Fig. 4.12A), we also analyzed all six sets of E_{FRET} versus time trajectories using an HMM-based idealization approach as described in Sec. 4.2.4, providing values of k_{GS1}, and k_{GS2} (Fig. 4.13). At the lowest temperature (22 °C), the values of k_{GS1} and k_{GS2} obtained using BIASD (0.9 ± 0.1 s^{-1} and 2.1 ± 0.3 s^{-1}, respectively) are nearly equivalent to those obtained using the HMM-based idealization method (0.7 ± 0.1 s^{-1} and 1.5 ± 0.1 s^{-1}, respectively), suggesting that the values of k_{GS1} and k_{GS2} obtained using BIASD are at least as accurate as those obtained using conventional HMM-based idealization methods. At

the higher temperatures, where more time averaging occurs (*c.f.*, Fig. 4.14), the values of k_{GS1} and k_{GS2} obtained using BIASD, which increase monotonically, are systematically faster than those obtained using the HMM-based idealization method (Figs. 4.12 and 4.13). This observation is consistent with our analysis of the simulated data in the previous section. There, when the rate constants were greater than 1 / 10th of 1 / τ, the idealization-based methods systematically underestimated those rate constants. This increases our confidence that the values of k_{GS1} and k_{GS2} obtained using BIASD are accurate. Additionally, we note that the posterior probability distributions of ϵ_{GS1} and ϵ_{GS2} inferred using BIASD have means of 0.14 and 0.77, respectively, which are values of ϵ_{GS1} and ϵ_{GS2} that very closely match the values of the mean E_{FRET} of GS1 and GS2 reported in previous, room-temperature studies of the analogous PRE^{-A} complex (*e.g.*, 0.16 and 0.76, respectively, in Ref. 38) (Fig. 4.15). This similarity suggests that the values of ϵ_{GS1} and ϵ_{GS2} obtained using BIASD are also accurate, regardless of the time resolution of the smFRET TIRF experiment.

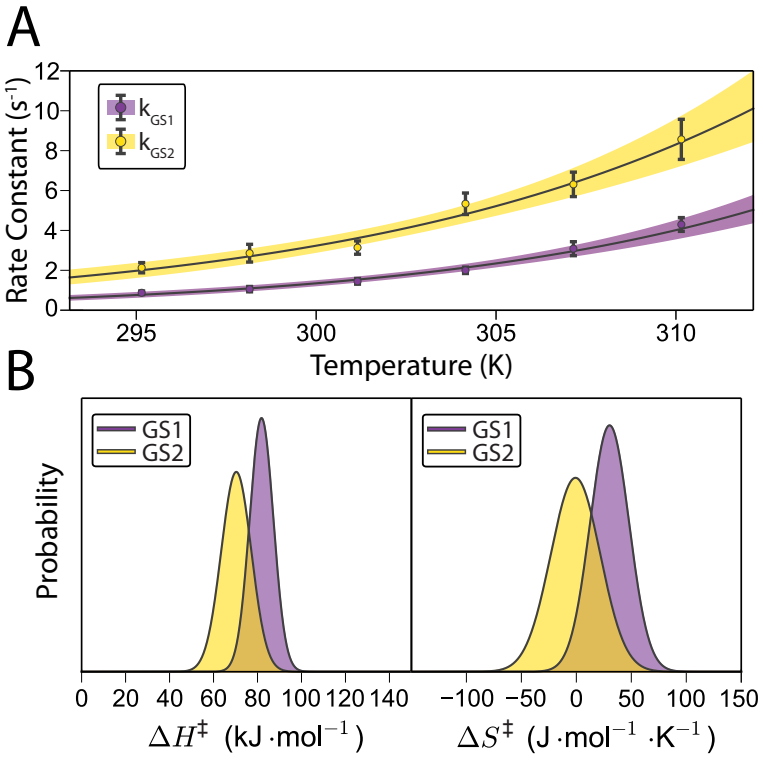

Figure 4.12: (A) temperature dependence of k_{GS1} and k_{GS2} for PRE^{-A} complexes using BIASD. The scatter plots show the expectation value of the posterior probability distributions of k_{GS1} and k_{GS2} and the error bars represent ± 1 SEM. The solid lines are the non-linear least-squares fit to transition-state theory, and the shaded regions represent fitting error ($\pm 1\sigma$). (B) Probability distribution of $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ from the fit to transition-state theory for the GS1 and GS2 transition state energy barrier.

With accurate measurements of k_{GS1} and k_{GS2} as a function of temperature, we then used transition-state theory (Reviewed in Refs. 39, 40, 41, and 42) to quantify the apparent transition-state energy barrier

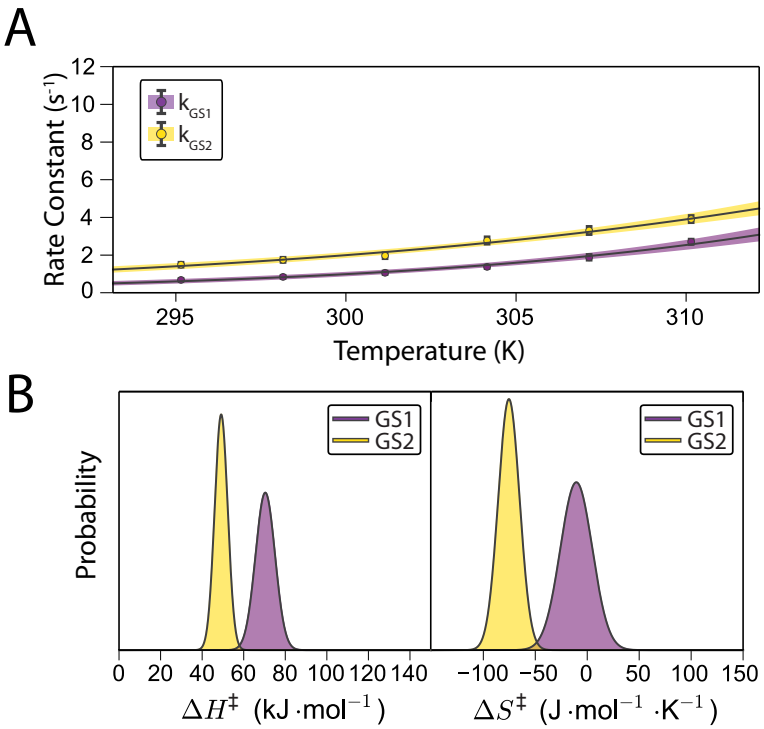

Figure 4.13: (A) Temperature dependence of k_{GS1} and k_{GS2} for PRE^{-A} complexes using a maximum likelihood HMM. The scatter plots show the mean value of k_{GS1} and k_{GS2} and the error bars represent ± 1 SEM. The solid lines are the non-linear least-squares fit to transition-state theory, and the shaded regions represent fitting error ($\pm 1\sigma$). (B) Probability distribution of $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ from the fit to transition-state theory for GS1 and GS2 to the transition state.

along the apparent GS1 $\rightleftharpoons$ GS2 reaction coordinate [43, 44]. Kramers' barrier-crossing theory [45], which was developed to analyze thermally activated, condensed-phase transitions of a Brownian particle [39] and is increasingly being used to analyze the conformational dynamics and folding of small, globular proteins [19], may ultimately provide a more exact analysis of the apparent transition-state energy barrier along the apparent GS1 $\rightleftharpoons$ GS2 reaction coordinate. However, its application requires knowledge regarding the viscosity of the aqueous buffer in which the PRE^{-A} complex is dissolved and the 'internal friction' of the PRE^{-A} complex that are unavailable in the current study. As such, we have opted to use transition-state theory, and regard the results as an upper limit of the apparent transition-state energy barrier along the apparent GS1 $\rightleftharpoons$ GS2 reaction coordinate, which does not account for internal friction or transition-state recrossings. To apply transition-state theory, we fit the mean rate constants at each temperature to the equation

$$k_{\text{TST}} = \left(\frac{\kappa k_B T}{\hbar}\right) e^{\frac{-1}{k_B T}(\Delta H^{\ddagger} - T\Delta S^{\ddagger})}, \tag{4.54}$$

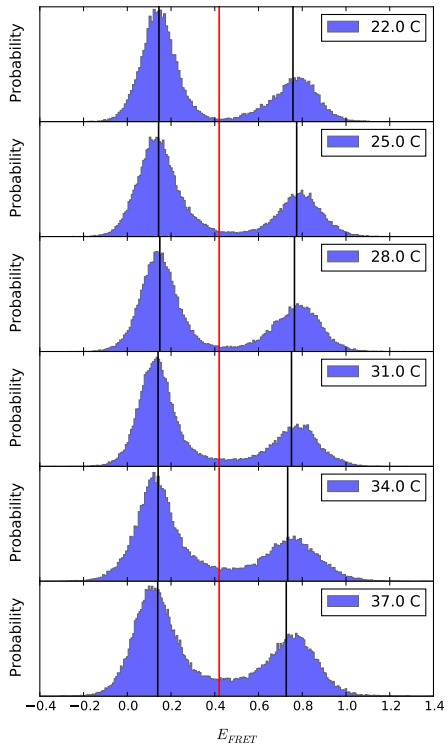

Figure 4.14: Temperature dependence of idealized E_{FRET} means for PRE-A complexes. The E_{FRET} values observed from the PRE-A are histogrammed, and were thresholded at a value of 0.42 (red). The black vertical lines denote the mean values of the subsequently idealized data. As time-averaging increases, the signal means trend towards region between ϵ_{GS1} and ϵ_{GS2}

rather than estimate the prefactor term, where κ is taken to be unity, k_B is the Boltzman constant, $\hbar$ is Planck's constant, $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ are the enthalpic and entropic differences between the transition and ground states, which are associated with the temperature-dependent- and temperature-independent contributions to the rate constants, respectively. The results of these fits provide $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ for the GS1 $\rightarrow$ GS2 transition of $\Delta H^{\ddagger}_{GS1}$ = 81.2 $\pm$ 5.4 kJ mol^{-1} ($\pm$ 1σ) and $\Delta S^{\ddagger}_{GS1}$ = 30 $\pm$ 18 J mol^{-1} K^{-1} ($\pm$1σ), and $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ for the GS2 $\rightarrow$ GS1 transition of $\Delta H^{\ddagger}_{GS2}$ = 70.3 $\pm$ 6.8 kJ mol^{-1} and $\Delta S^{\ddagger}_{GS2}$ = -1 $\pm$ 23 J mol^{-1} K^{-1}, values that are significantly different from those that are determined using the values of k_{GS1} and k_{GS2} that are obtained at each temperature using HMM-based idealization methods (Fig. 4.13). Unfortunately, structure-based interpretations of the absolute values of the $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ for the GS1 $\rightarrow$ GS2 and GS2 $\rightarrow$ GS1 transitions that are observed for the particular PRE-A complex studied here are significantly complicated by the complexity of the enthalpic and entropic changes that are associated with conformational rearrangements of large macromolecular complexes such as PRE-A complexes [46] and the inherent limitations of transition-state theory [8, 40]. Nonetheless, structure-based interpretations of the changes in the absolute values of

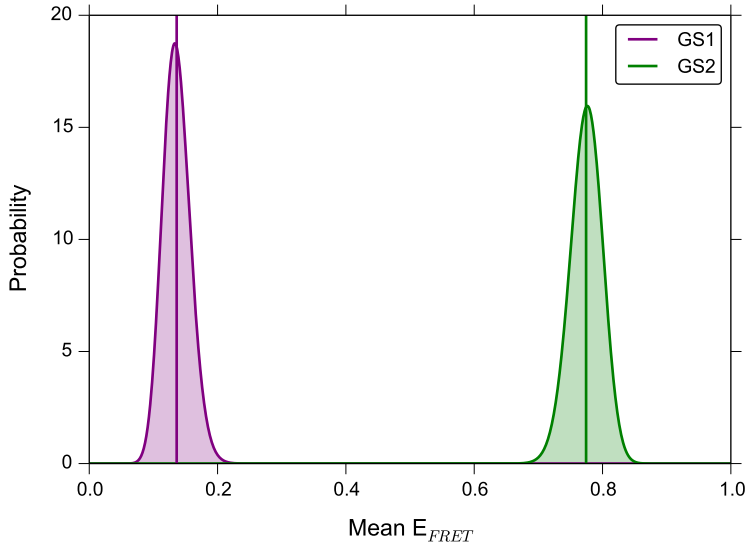

Figure 4.15: The mean E_{FRET} signal for GS1 and GS2 from the posterior probability distributions of ϵ_1 and ϵ_2. Using the posterior probability distributions at all temperatures from the L1-tRNA PRE-A smFRET data analyzed in the main text, beta distributions with $\alpha = \sum_{\text{temperatures}} \alpha_i$, and $\beta = \sum_{\text{temperatures}} \beta_i$ were plotted for GS1 (Purple) and GS2 (Green). Means of each distribution (0.14 and 0.77, respectively) are denoted with vertical lines; these agree very well with previously published values.

the $\Delta H^{\ddagger}$s and $\Delta S^{\ddagger}$s (*i.e.*, $\Delta\Delta H^{\ddagger}$ and $\Delta\Delta S^{\ddagger}$) for the GS1 $\rightarrow$ GS2 and GS2 $\rightarrow$ GS1 transitions that are observed between pairs of PRE-A complex (*e.g.*, containing different P-site tRNAs, wildtype versus mutant P-site tRNAs, wildtype versus mutant ribosomes, etc.) are much more straightforward and can reveal the thermodynamic contributions that particular structural features of PRE-A complexes make to the apparent transition-state energy barrier along the apparent GS1 $\rightleftharpoons$ GS2 reaction coordinate [40]. Combined with the experimental platform that we have previously described, the analytical framework presented in this section now enables the collection, analysis, and interpretation of such data.

4.5 Evaluation of BIASD

By analyzing the fraction of time that a single-molecule spends in each state of a particular kinetic scheme as well as the signal value of each of those states during each τ in a signal versus time trajectory, BIASD adopts a fundamentally different approach to the analysis of time-resolved single-molecule experiments than has been traditionally utilized by methods that idealize the trajectories (*e.g.*, thresholding, HMMs). Using computer-simulated and experimentally observed data, we have demonstrated that with this approach BIASD is able to accurately and precisely infer the rates of transitions between the two states of a two-state kinetic scheme as well as the signal values corresponding to these two states, even when the rates

of transitions between states are orders of magnitude faster than the time resolution of the signal versus time trajectories. When applied to the experimentally observed PRE-A complex smFRET data of Wang and coworkers [16], this ability allowed us to infer otherwise inaccessible thermodynamic parameters of the transition state energy barriers for GS1 and GS2. BIASD can be applied to any experimentally observed signal versus time trajectory that exhibits stochastic transitions between distinct states, regardless of the nature or the origin of the signal. Thus, BIASD can be used to analyze data collected using virtually any time-resolved single-molecule experimental method, including single-molecule wide-field fluorescence microscopy, force spectroscopy, and tethered particle motion methods. Moreover, although we have developed BIASD to analyze single-molecule signal versus time trajectories, as described here, BIASD does not consider the temporal ordering of the data. Consequently, in addition to analyzing individual single-molecule signal versus time trajectories, BIASD can also be used to analyze the distribution of fractional occupancies seen across an entire ensemble of single-molecules during a given τ. This could allow non-equilibrium phenomenon to be monitored across an ensemble of single-molecules (*e.g.*, stopped-flow delivery of an antibiotic to PRE^{-A} complexes). In addition, BIASD can be expanded to include the time evolution of the state occupation probabilities (*c.f.*, Eq. 4.9), or to incorporate time dependence into the model parameters k_i, ϵ_i, and σ (*e.g.*, changing ϵ_i in particle tracking).

Regarding the performance of BIASD on experimental data, we note that the rate constants and signal values of a system can be more precisely inferred from experiments that collect higher SNR data, because then there is less uncertainty in the time-averaged fractional occupancies of the signal versus time trajectories. Therefore, somewhat counterintuitively, sub-temporal-resolution dynamics can, to some degree, be more precisely inferred from signal versus time trajectories collected with a higher SNR but a longer τ (*e.g.*, due to better photon conversion efficiencies on an electron-multiplying charge-coupled device), than those collected with a lower SNR but a shorter τ. Additionally, although we have focused the current work on the most widely applicable case of a Markovian, two-state system in which the noise of the signal can be modeled using a Gaussian distribution, the Bayesian inference-based framework underlying BIASD can be readily extended to non-Markovian dynamics [23, 31], N-state kinetic schemes [29, 30], or systems in which the noise of the signal can be modeled using distributions other than a Gaussian distribution [28, 32]. To facilitate the analysis of single-molecule data using BIASD, as well as to enable the future extension of BIASD along the lines described here, we have made a graphical user interface and the BIASD source code, both implemented in Python, available at `http://www.columbia.edu/cu/chemistry/groups/gonzalez/`.

4.6 Reconstructing Mesoscopic Ensembles from Single-molecule Dynamics

A single-molecule experiment probes only one molecule from an ensemble of molecules. However, ergodicity suggests that, given an infinite amount of observation, that single-molecule should be representative of the equilibrium ensemble. Interestingly, phenomenon such as inhomogeneous broadening within an experimentally observed bulk spectrum suggests that there can be a range of behaviors within an ensemble, and that certain molecules may not have the same experience as others. For a single-molecule experiment, this is better highlighted by the presence of static heterogeneity within the ensemble (*c.f.*, Ref. [47]). For instance, post-transcriptional modifications of tRNAs can influence elongation efficiency and codon-context effects [48], because the modifications might produce different dynamics during translation as the tRNAs interact with ribosome. Similarly, varied stoichiometric composition of multimeric protein complexes (*e.g.*, the sub-stoichiometric occupation of the S1 protein in ribosomes [49]), different isoforms of the same gene (*e.g.*, the seven different rRNA operons in *Escherichia Coli* [50]), and post-translational modifications (*e.g.*, methylation of class I release factors [51]) are all possible causes of static heterogeneity within a single-molecule ensemble of biomolecules. Also confounding the connection between the behavior of a single-molecule to that of the ensemble is the presence of non-equilibrium dynamic heterogeneity or disorder, in which the dynamics of a single-molecule will change over time [52, 53]. If these dynamics change slowly enough relative to the length of the experiment, the dynamic disorder can even be confused with static heterogeneity. For instance, the ligand-binding ability of a protein, such as oxygen binding to myoglobin [54], can change in time depending upon conformation of the protein [8, 55]. Because of ensembles of biomolecules often contain these sources of static and dynamic heterogeneity, single-molecule experiments are often designed to observe multiple single-molecules simultaneously (*c.f.*, Sec. 3.3.1). A 'mesoscopic ensemble' must then be constructed from these single-molecule observations in order to describe the entire ensemble of molecules.

One approach to constructing a mesoscopic ensemble is to take the average behavior (*e.g.*, stochastic rate constants) of the observed single-molecules as the expected behavior of the ensemble. The distribution of behaviors contained within the ensemble might then be described by the standard deviation of individual single-molecules' behaviors. Unfortunately, this approach ignores the contribution of any dynamic disorder from a single-molecule to the ensemble. Additionally, it is unclear how to account for the precision with which a single-molecule's behavior might be determined (one previously discussed method is to treat all of

the dwell-times within the mesoscopic ensemble of state versus time trajectories as contributing equally to the ensemble transition matrix (*c.f.*, Sec. 3.3.2)). Notably, over 40 years ago, Frauenfelder's group observed distributions of activation energies for ligand binding (*i.e.*, rates of ligand binding) that are attributable to an equilibrium occupation of a distribution of conformations within an ensemble of myoglobin [54]. In a corresponding single-molecule experiment, constructing a mesoscopic ensemble in which all single-molecules observed contribute to the same 'average', ensemble-representative behavior is clearly an improper approach. In order to accurately construct a mesoscopic ensemble from single-molecule signal versus time trajectories, one must somehow be able to identify and classify different types of single-molecule behaviors, and then, accounting for the precision of such a classification, quantify the different behaviors within the mesoscopic ensemble in order to approximate the macroscopic ensemble.

4.6.1 Variational Mixture Models

Consider an attempt to classify pieces of data (*e.g.*, single-molecule signal versus time trajectories) which belong distinctly to J classes of data. A non-probabilistic approach, such as K-means clustering, assigns each piece of data to the closest estimate of one of K classes–where K is a haphazardly chosen number of classes that does not necessarily correspond to J. Unfortunately, when some of the J, authentic classes are ill-resolved or overlapping, estimates of the K classes, and classification of the data into those K classes, are not robust. Fortunately, classification problems such as the one mentioned above can be approached in a statistically robust manner by using a mixture model [27]. In a mixture model, each piece of data instead has a probability of belonging to each of K classes in the model–much like a wavefunction which can be delocalized over many sites of a lattice such as a DNA oligomer [56–58]. This approach not only yields better estimates of each class, but also provides a description of the precision with which each piece of data can be classified, which is especially important when classes overlap. Because of this consideration of precision, mixture models are naturally integrated into the Bayesian inference framework (*c.f.*, Sec. 4.2.2 and Ref. 15). Unfortunately, as with most Bayesian inference approaches, it can be difficult to calculate the exact posterior probability distribution in order to learn the parameters of the mixture model after having observed the data.

Recently, to analyze smFRET signal versus time trajectories using Bayesian inference, the Gonzalez group has employed variational approximations to the posterior probability distributions of an HMM (vbFRET [12, 59]) and of an ensemble of HMMs controlled by consensus HMM parameters (ebFRET [13, 14]). Such 'variational Bayes' approaches find the optimal, analytically tractable approximation to the posterior probability distribution by assuming only that the distribution factorizes into separate, more simple parts. Fac-

toring the distribution in this manner is equivalent to the mean-field approximation in physics–it considers the average behavior of each factor to override any fluctuations that lead to coupling between terms. Additionally, as described below in Sec. 4.6.4, these variational approximations also provide a natural method with which to perform model selection and decide the most parsimonious number of K states present in the data; this method is notably robust against the type of overfitting which plagues maximum-likelihood and non-probabilistic methods.

In the following sub-sections, we demonstrate the use of a variational mixture model to construct mesoscopic ensembles of single-molecules behavior determined by BIASD. This mixtures model accounts for the precision with which each single-molecule is determined (*i.e.*, the entire posterior probability distribution of each signal versus time trajectory). Additionally, it automatically incorporates model selection in order to determine the different types of single-molecule behaviors present in the mesoscopic ensemble. These different behaviors are technically only for static heterogeneity present within the ensemble, however, with a sufficient number of single-molecules observed for a particular class displaying defined dynamic disorder, it should sample this behavior and represent the time-independent spread of the disorder.

4.6.2 Accounting for Sample Uncertainty with a Variational Gaussian Mixture Model

One established, and tractable variational mixture model is the variational gaussian mixture model. In this model, each class is considered a multivariate normal distribution with D number of variables in the data. For instance, with the results of a BIASD calculation, the posterior probability distribution is in a five-dimensional space (*i.e.*, ϵ_1, ϵ_2, σ, k_1, and k_2), so $D = 5$. Unfortunately, as typically derived, this model does not account for uncertainty within each sample (single-molecule) that is being classified [27]. Here, following the work of Bishop [27, 60], but making modifications in a fashion similar to work done to include 'noisy' data [61], we give the update equations for the posterior probability distribution of the variational gaussian mixture model using a hypothetical group of single-molecule signal versus time trajectories that were processed using BIASD to be more concrete.

Consider the n^{th} of N single-molecule signal versus time trajectories processed by BIASD. The posterior probability distribution for this molecule, has a mean $X_n = \mathbb{E}[\Theta_n]$, and a covariance of $\Sigma_n = \mathbb{E}[\Theta_n^2] - \mathbb{E}[\Theta_n]^2$. The entire group of posterior probability distributions is then $\boldsymbol{X} = \{X_1 ... X_N\}$ and $\boldsymbol{\Sigma} = \{\Sigma_1 ... \Sigma_N\}$. These values are shown in the graphical model shown in Figure 4.16; this graph demonstrates how these observed

parameters depend upon the parameters of the model. The dependences shown in the graphical model can be written mathematically as

$$p(\boldsymbol{X}, \boldsymbol{Z}, \pi, \mu, \Lambda | \boldsymbol{\Sigma}) = p(\boldsymbol{X}|\mu, \Lambda, \boldsymbol{\Sigma}) \cdot p(\boldsymbol{Z}|\pi) \cdot p(\pi) \cdot p(\mu|\Lambda) \cdot p(\Lambda), \tag{4.55}$$

where $\boldsymbol{Z} = \{z_1 ... z_N\}$ and each z_n is a vector of length K that is 1 only when the n^{th} single-molecule is of state i, where π is the probability of state i being found in the ensemble, where μ is the mean of state i, and where Λ is the precision (*i.e.*, inverse of covariance) of state i. For our model, these factors are

$$p(\boldsymbol{X}|\boldsymbol{Z}, \mu, \Lambda, \Sigma) = \prod_{n=1}^{N} \prod_{i=1}^{K} \mathcal{N}(X_n | (\Sigma_n^{-1} + \Lambda_i)^{-1} \cdot (\Sigma_n^{-1} \cdot X_n + \Lambda_i \cdot m_i), (\Sigma_n^{-1} + \Lambda_i)^{-1}), \tag{4.56}$$

$$p(\boldsymbol{Z}|\pi) = \prod_{n=1}^{N} \prod_{i=1}^{K} \pi_i^{z_{nk}}, \tag{4.57}$$

$$p(\pi) = \text{Dir}(\pi|\alpha), \tag{4.58}$$

$$p(\mu, \Lambda) = p(\mu|\Lambda)p(\Lambda) = \prod_{i=1}^{K} \mathcal{N}(\mu_i | m_i, (\beta_i \lambda_i)^{-1}) \cdot \mathcal{W}(\Lambda_i | W_i, \nu_i), \tag{4.59}$$

where $\mathcal{N}$ is the multivariate Normal distribution, $\mathcal{W}$ is the Wishart distribution, and Dir is the Dirichlet distribution, which are all parameterized with the hyperparameters specified above for the i^{th} class (*c.f.*, Appendix B). With this model, the $\boldsymbol{X}$ are assumed to be normally distributed because of the joint Normal-Wishart distribution above, where the variance of the $\boldsymbol{X}$ is controlled by the Wishart distribution over Λ.

For the variational approximation of this model, we will assume only that the the $\boldsymbol{Z}$ components factor from the posterior probability distribution and write

$$q(\boldsymbol{Z}, \pi, \mu, \Lambda) = q(\boldsymbol{Z}) \cdot q(\pi, \mu, \Lambda). \tag{4.60}$$

From this variational approximation, we derive update equations for the posterior probability distribution

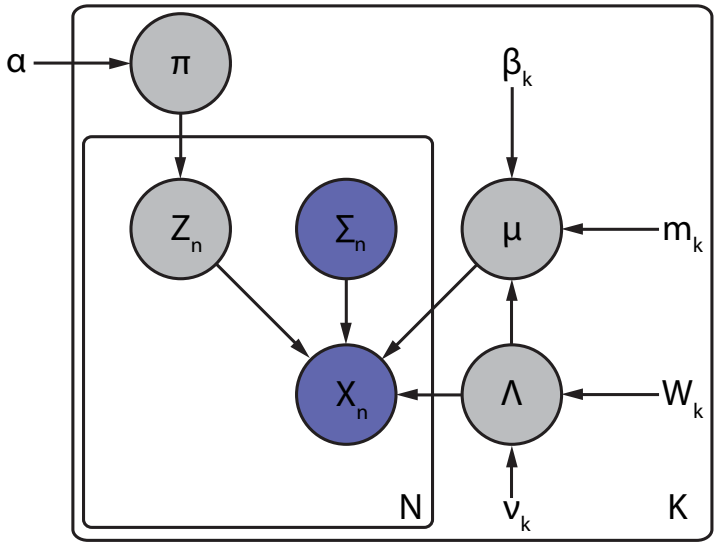

Figure 4.16: Graphical model of variational gaussian mixture model that accounts for uncertainty. The parameters in the blue circles are observed experimentally. Parameters in the grey circles are unobserved, but still stochastic parameters of the model. Parameters without a circle are learned by optimizing the factors of the variational model.

which optimize the variational model. These update equations are

$$\nu_i = \nu_0 + N_i, \tag{4.61}$$

$$\beta_i = \beta_0 + N_i \tag{4.62}$$

$$m_i = \frac{1}{\beta_k}(\beta_0 m_0 + N_i \bar{x_i}), \tag{4.63}$$

$$W_i = (W_0^{-1} + N_i \cdot S_i + \frac{\beta_0 N_i}{\beta_0 + N_i}(\bar{x_i} - m_0) \cdot (\bar{x_i} - m_0)^T), \tag{4.64}$$

$$\alpha_i = \alpha_0 + N_i. \tag{4.65}$$

where, from the statistics of the data,

$$N_i = \sum_{n=1}^{N} r_{ni}, \tag{4.66}$$

$$\bar{x_i} = \frac{1}{N_i} \sum_{n=1}^{N} r_{ni} \cdot (\Sigma_n^{-1} + \Lambda_i)^{-1} \cdot (\Sigma_n^{-1} \cdot X_n + \Lambda_i \cdot m_i), \tag{4.67}$$

$$S_i = \frac{1}{N_i} \sum_{n=1}^{N} r_{ni} \cdot ((\Sigma_n^{-1} + \Lambda_i)^{-1} \cdot (\Sigma_n^{-1} \cdot X_n + \Lambda_i \cdot m_i) - \bar{x_i}) \cdot ((\Sigma_n^{-1} + \Lambda_i)^{-1} \cdot (\Sigma_n^{-1} \cdot X_n + \Lambda_i \cdot m_i) - \bar{x_i})^T, \tag{4.68}$$

where $\Lambda_i = \nu_i W_i$. Except for the expressions for $\bar{x_i}$ and S_i, these expressions correspond to those in Ref. 27. The subscript-naught variables correspond to the prior probability distributions for the hyperparameters. Typically, one can use very uninformative parameters for each respective distribution, such as

$$\nu_0 = D + 1, \tag{4.69}$$

$$\beta_0 = 0, \tag{4.70}$$

$$m_0 = 0, \tag{4.71}$$

$$W_0 = \mathbb{I}, \tag{4.72}$$

$$\alpha_0 = 1, \tag{4.73}$$

and r_{ni} is from a K-means classification or chosen as random-variates from a Dirichlet distribution over α_0.

Following this optimization step, certain expectation values of the data given the optimized model parameters can be calculated which can then, in turn, be used to re-optimize the model parameters. These expectation values are

$$\mathbb{E}[\ln|\Lambda_i|] = \sum_{j=1}^{D} \psi\left(\frac{\nu_i + 1 - j}{2}\right) + D\ln(2) + \ln|W_i|, \tag{4.74}$$

$$\mathbb{E}[\ln\pi_i] = \psi(\alpha_i) - \psi(\sum_i \alpha_i), \tag{4.75}$$

$$\begin{aligned}\mathbb{E}_{\mu_i,\Lambda_i}[(X_n - \mu_i)^T \Lambda_i (X_n - \mu_i)] = D/\beta_i + ((\Sigma_n^{-1} + \Lambda_i)^{-1} \cdot (\Sigma_n^{-1} \cdot X_n + \Lambda_i \cdot m_i) - m_i)^T \\ \cdot \Lambda_i \cdot ((\Sigma_n^{-1} + \Lambda_i) \cdot (\Sigma_n^{-1} \cdot X_n + \Lambda_i \cdot m_i) - m_i),\end{aligned} \tag{4.76}$$

from which

$$r_{ni} = \frac{\rho_{ni}}{\sum_{i=1}^{K} \rho_{ni}}, \text{ where} \tag{4.77}$$

$$\ln(\rho_{ni}) = \mathbb{E}[\ln\pi_i] + \frac{1}{2}\mathbb{E}[\ln|\Lambda_i|] - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_i,\Lambda_i}[(X_n - \mu_i)^T(\Sigma_n^{-1} + \Lambda_i)^{-1}(X_n - \mu_i)]. \tag{4.78}$$

These optimization, and expectation steps are then iterated until convergence upon the ideal solution of the variational model for the posterior probability distribution.

4.6.3 Laplace Approximation of BIASD Posterior Probability Distribution

While the variational gaussian mixture model described above accounts for the uncertainty of each single-molecule, it only considers the first and second moments of the posterior probability distributions–effectively casting them into a symmetrical gaussian form. This choice ignores any asymmetry in the individual BIASD posterior probability distributions. While this assumption about the individual BISAD posterior probability distributions must be made to use the variational mixture model in order to build a mesoscopic ensemble of single-molecules, we can take advantage of it to simplify the calculation of each individual BIASD posterior probability distribution by using the Laplace approximation. As described below, the Laplace approximation makes BIASD much more computationally tractable than does MCMC, and the resulting multivariate normal posterior probability distributions fits naturally with the variational clustering algorithm.

The Laplace approximation approximates a probability distribution by expanding the logarithm of the distribution in a Taylor series at the maximum of the distribution and equating this to the logarithm of the normal distribution. Upon finding the maximum, x_0 of the logarithm of a probability distribution using numerical maximization, the second partial derivatives of the function (the Hessian matrix, $\boldsymbol{H}$) can be calculated at the maximum using finite difference formulas (*c.f.*, Eqn. 25.3.23 and 25.3.26 of Ref. 62). With $\boldsymbol{x}_0$ and $\boldsymbol{H}$ calculated, these are compared to a Taylor series expansion the logarithm of the multivariate normal distribution (*c.f.*, Appendix B). The multivariate Taylor series of $f(\boldsymbol{x}$ at $\boldsymbol{x} = \boldsymbol{a}$ is

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \frac{1}{1}((\boldsymbol{x} - \boldsymbol{a}) \cdot \nabla f(\boldsymbol{a})) + \frac{1}{2}((\boldsymbol{x} - \boldsymbol{a}) \cdot H(\boldsymbol{a}) \cdot (\boldsymbol{x} - \boldsymbol{a})) + ... \tag{4.79}$$

While the logarithm of the the multivariate normal distribution is

$$\ln(\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \left(\frac{-k}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\sigma}|)\right) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}). \tag{4.80}$$

By inspection, Eqns. 4.79 and 4.80 are equivalent if the gradient of f evaluated at $\boldsymbol{\alpha}$ is zero, and the Taylor series is truncated after the second-order term. At its maximum, the probability distribution that we wish to approximated has a gradient of zero. Therefore, a multivariate probability distribution, such as the posterior probability distribution from BIASD, can be approximated by a multivariate normal distribution by finding the maximum, calculating the H of the logarithm of the distribution at this point, and then calculating the

parameters by matching the equations above as

$$\boldsymbol{\mu} = \boldsymbol{x_0}, \tag{4.81}$$

$$\boldsymbol{\Sigma} = (-H_{\ln})^{-1}, \tag{4.82}$$

where the subscript ln is a reminder that it is the Hessian of the logarithm of the distribution. Since it is much easier to find the maximum and calculate the Hessian of a distribution, than to sample all of state-space, the Laplace approximation very computationally efficient. Moreover, when used with BIASD, the resulting mean and covariance of the posterior probability distribution can easily be used in the variational gaussian mixture model to account for the uncertainty in each single-molecule.

4.6.4 Model Selection

The use of the variational approximation to the Bayesian treatment of the gaussian mixture model allows a quantity termed the lowerbound to be calculated [27]. During the course of the optimization of the variational, approximating distribution, the lowerbound is maximized. The optimal distribution is found when the lowerbound is at its maximum, because the lowerbound is one part of a decomposition of the evidence of the model, $p(X)$ (*i.e.*, the denominator in Bayes' Rule (Eqn. 4.41)). By maximizing the lowerbound, it becomes closer and closer to the model evidence, and as a consequence, the approximating distribution becomes closer and closer to the actual distribution. With this in mind, the lowerbound can be used to perform 'model selection'. By varying the number of classes of gaussians included in the gaussian mixture model, the best model will be that with the largest lowerbound. Interestingly, including more classes does not necessarily increase the lowerbound such as occurs with maximum likelihood methods (*i.e.*, by fitting 10 points with a 10-degree polynomial). Given sufficient statistical evidence (*i.e.*, enough data points), the lowerbound will peak at the correct number of states present in the data, all while avoiding overfitting. Additionally, the variational method is also robust against overfitting in the sense that additional superfluous classes added to the mixture will be minimally populated. In fact, this approach can be utilized for active model section, in which a mixture model is initialized with many states, and these states are allowed to depopulate and then removed, until the lowerbound is maximal. This process is demonstrated in Fig. 4.17 for synthetic 2D multivariate gaussian data. Note that the best model as judged by the lower bound is the correct, three-state model, and that the mixture model converges to the correct solution by reestimating the optimal parameters as described in Sec. 4.6.2. The lowerbound equations for the variational gaussian mixture model can be

found in [27].

The selection of the correct number of classes in a variational mixture model is more complicated when including the uncertainty in each sample, as we will do with the the BIASD posterior probability distributions, because the magnitude of the uncertainty can overcome otherwise clear separation between the classes. We demonstrate this effect as a function of increasing magnitude of the covariances of the data samples in Fig. 4.18. As uncertainty in the samples increases, the spread of the sample centers becomes overwhelmed by uncertainty. Eventually, the classes begin to appear to overlap, even though, the samples belong to a distinct underlying class. With low uncertainty, the variational approach is able to discriminate between classes, and estimate the spread of each class. This estimate of the spread of each class even remains mostly the same size with increasing uncertainty, showing that the method is robust to uncertainty. Unfortunately, at high uncertainty, the seemingly overlapped classes forces the most parsimonious model choice to be that where the ill-resolved states are combined into one averaged class. Even so, the mixture model is able to account for the precision of classifying data that remains in the overlapped area of the other classes. Finally, we note that the lowerbound peaks at the simulated value of three classes for most of the cases tried. This suggests that the method can correctly discriminate the correct number of states given a sufficient number of samples. However, as shown, uncertainty easily overwhelms these distinctions. At higher uncertainty, the highest lowerbound is from the two-class model. This model can be described as being more parsimonious with the entirety of the data, than the three state model. Collecting more samples of data might better resolve the states, however, without a distinction between two classes (due to large uncertainty), there is effectively little difference between the classes, and the variational mixture model creates an averaged state governed by the collective uncertainty in each separate state. While these examples have been for two-dimensional data, extending the algorithm to higher dimensions is trivial but for the additional computational power, which is minimal because the method employs conjugate priors.

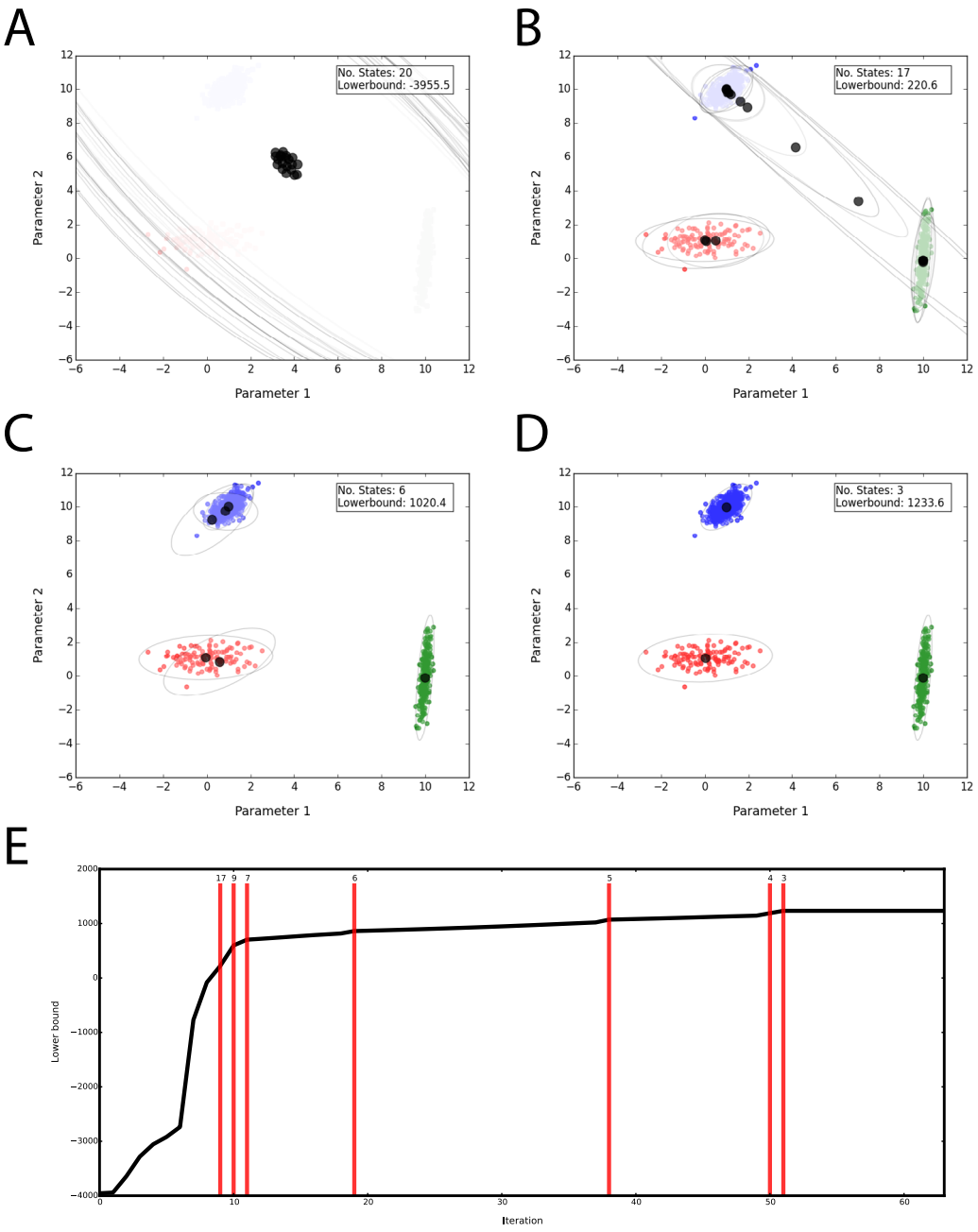

Figure 4.17: Dynamic model selection with a variational mixture model. (A - D) Scatter plots of classes of 2D synthetic data (blue, green, red) with the gaussian mixture model classes overlaid. The center of the mixture classes is represented by a black circle, and the variance in each class is shown as the ellipse at a 3σ threshold. (E) The lowerbound for this variational mixture model as a function of iterations of the variational algorithm. Note the ever increasing lowerbound, and the large increases provided by eliminating classes below an occupancy threshold of $\alpha_i < 0.001 \cdot \alpha_0$.

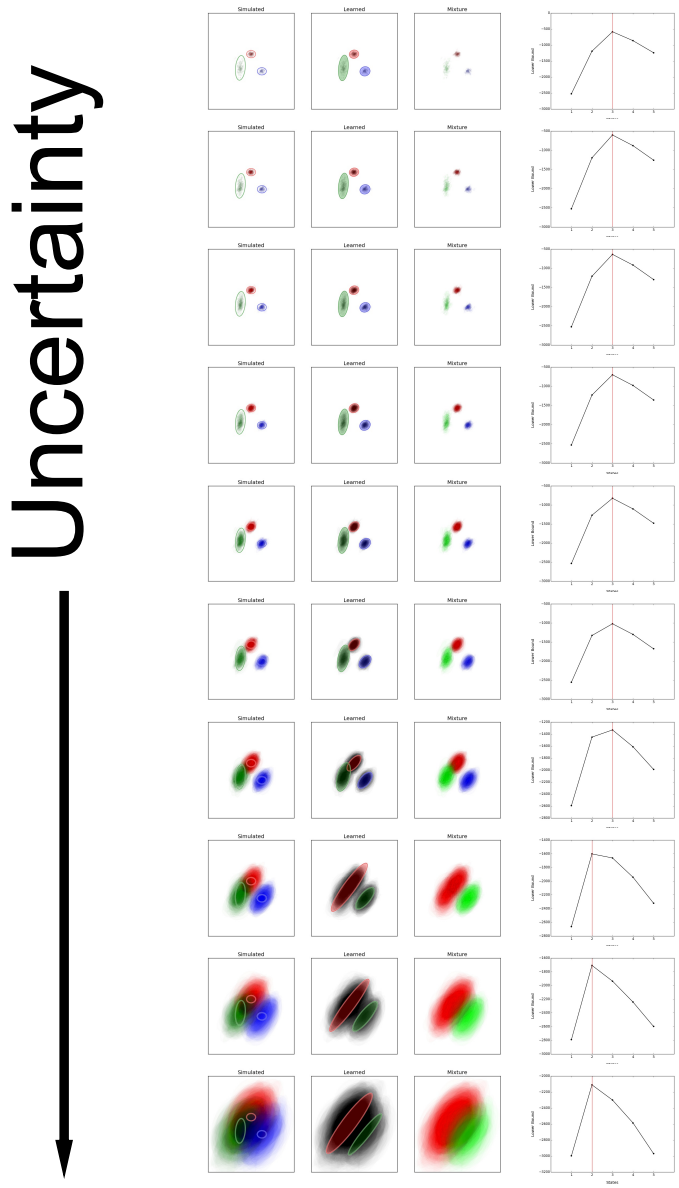

Figure 4.18: Effect of sample uncertainty on model selection with a variational mixture model. The first column shows a scatter point for each data point of the classes (delineated by different colors), with solid elipses at the 3σ cutoff of the spread of each class, and transparent elipses at the 3σ cutoff for the uncertainty for each data point to show the spread of the uncertainty. At low uncertainty, the clusters are easily resolved, while at high uncertainty, the significantly overlap. The second column shows the uncolored data, plotted as in the previous column, which is analyzed using the variational gaussian mixture model. The colored, transparent elipses show the 3σ thresholded covariance of each class. These number of states was chosen from the highest lowerbound. The third column represent the reconstructed ensemble of data in the simulated (first) column from the parameters that were learned by the variational gaussian mixture model. The fourth column shows the lowerbound calculated for 1–5 classes in the mixture. The results for the first three columns are shown for the mixture with the highest lowerbound. The simulated number of classes was three.

4.7 Structural Contributions of Transfer RNA to the Pretranslocation Thermodynamic Landscape

As described in Sec. 1.1, the ribosome is the universally conserved RNA-protein complex that is responsible for translating the mRNA into their encoded proteins. During the elongation stage of translation (*c.f.*, Sec. 1.3), ternary complexes deliver the different aa-tRNA to the ribosome, which then transfers the amino acid from the tRNA into the nascent polypeptide chain. Consequentially, during elongation, the A, P, and E sites of the ribosome are occupied by many permutations of the different types of tRNAs. These tRNAs differ in more than their anti-codon stemloops, and these differences play important roles in the process of elongation. For instance, the Hirsh supressor tRNA is a $tRNA^{Trp}$ with a G24A mutation that is distal from the anti-codon stemloop [63, 64]. This tRNA is most likely more easily distortable than the wildtype $tRNA^{Trp}$, and this results in an increased rate of EF-Tu GTPase activation during tRNA selection (*c.f.*, Sec. 1.3.1), which allows it to recognize the UGA non-sense codon [65]. Because the conformational dynamics of the ribosome and its associated factors are generally rate-limiting relative to chemical reactions like GTP hydrolysis, the effects of varied tRNA occupation of the ribosome is expected to be a significant determinant of the dynamics of translation elongation.

Generally, these ribosomal dynamics during elongation can be considered as a Brownian motor, where motion along the reaction coordinate between thermally-accessible, meta-stable, conformational states is rectified through events such as factor binding or chemical reactions [66–69]. The energy landscape over which elongation occurs is composed of contributions from many factors, of which the contribution from the tRNA occupying the ribosome during any particular step can be large. For instance, the dynamic fluctuations of a PRE^{-A} complex between GS1 and GS2 depend drastically upon the identity of the tRNA in the P site [70]. Notably, the structure of initiator $tRNA^{fMet}$ differs from the of elongator tRNAs in the aminoacyl acceptor stem, the D stem, and the anticodon stem. The differences stabilize GS1 by increasing k_{GS2} relative to elongator tRNAs. By introducing those structural differences into elongator tRNAs the effect that $tRNA^{fMet}$ has upon the GS1 $\rightleftharpoons$ GS2 dynamic equilibrium can be recapitulated [70]. Most likely the structural differences between different tRNAs have been selected for under evolutionary pressure in order to fine-tune the overall process of translation. However, how these structural elements contribute to the thermodynamics of the energy landscape the regulates the kinetics of ribosomal dynamics is unclear.

To investigate the thermodynamic contributions of tRNA structure to the energy landscape over which the

translating ribosome functions, we have performed temperature-dependent smFRET experiments, similar to those in described Ref. [16] and in Sec. 4.4, on PRE complexes containing different tRNAs. By investigating the temperature-dependence of the rate constants involved in the GS1 $\rightleftharpoons$ GS2 equilibrium as a function of tRNA identity, nascent polypeptide chain presence, and A and P site occupation, relative thermodynamic contributions of the different structural elements can be quantified. Unfortunately, this investigation is complicated by rate constants which approach the time resolution limitations of smFRET TIRF experiments, especially with the increased temperatures used in these experiments. Additionally, it is complicated by the heterogeneity within the ensemble of ribosomes that is created when some of the enzymatically-prepared ribosomal complexes fail to undergo, or undergo additional rounds of translation. To overcome these complications, the experiments are analyzed using BIASD and the variational mixture model described above, which enables the fast dynamics of the heterogenous populations to be accurately and precisely quantified. This then allows the contributions of the structural difference between the various tRNA to the underlying energy landscape to be determined. However, in order to connect these tRNA-structure-dependent rate constants to the energy landscape, we must utilize a rate theory.

4.7.1 Rate Theories

Inspired by Van't Hoff, Arrhenius' analysis of the rates of chemical reactions as a function of temperature [71] lead to the well-known Arrhenius equation,

$$k = Ae^{-E/RT}, \tag{4.83}$$

where k is a rate constant, R is the gas constant, E is an activation energy, and A is a the preexponential 'frequency' factor. However, at that point in time, it was unclear how E would affect an individual molecule participating in the reaction; though, as early as 1867 Pfaundler had suggested that a reacting molecule had to have at least some critical energy in order to react [72]. Additionally, it was even more unclear exactly what the preexponential frequency factor represented, and how it could be exactly calculated [73].

After Arrhenius, the next major development in rate theory came with the 1935 with the development of transitions state theory (TST) by Eyring [74], and Evans and Polanyi [75], among the developments of many others, especially Wigner [76]. Previous work on rate theories with thermodynamics, kinetic theory of gasses, and statistical mechanics (for a historical account see Ref. 73) had culminated in a method to calculate an absolute value of the preexponential factor, and how to interpret it (for a scientific account of

TST see Refs. 39, 77, 41, 44, and 42).

TST makes two assumptions about a system. The first assumption in TST is that, along a single generalized coordinate (*i.e.*, the reaction coordinate), there exists a saddle point in the potential energy surface in phase space. Orthogonal to this saddle point lies a surface (*i.e.*, in the non-reaction-coordinate coordinates) that defines the 'activated complex' which is the species that are capable of undergoing the reaction. As such, the reaction rate depends on the concentration of this activated complex, and not of the reactant molecules. The second assumption in TST is that the system is at equilibrium, and therefore the concentration of the activated complex does not change as product is formed from reactant. This is sometimes known as a 'quasi-equilibrium' between the reactants and the activated complex, though it does not mean that the reactants and activated complex are in a actual equilibrium, but that the reactants and products are in full equilibrium and the activated complex is part of this equilibrium [41, 77]. From these assumptions, the absolute reaction rate can be calculated by using statistical mechanics to calculate the concentration of molecules present as activated complexes, multiplying this by their velocity in state-space along the reaction coordinate, and then integrating across all particles moving towards the products [74]. This process yields the famous TST equations

$$k = \kappa \frac{k_b T}{h} \frac{Q_{\ddagger}}{Q_A} e^{\frac{-E}{RT}}, \text{and} \tag{4.84}$$

$$k = \kappa \frac{k_b T}{h} e^{\frac{\Delta S^{\ddagger}}{R}} e^{\frac{-\Delta H^{\ddagger}}{RT}}, \tag{4.85}$$

where Q is a partition function, A and ‡ denote reactants and the activated complex, respectively, κ is the transmission coefficient, h is Planck's constant, k_b is the Boltzmann constant, and $\Delta S^{\ddagger}$ and $\Delta H^{\ddagger}$ are the change is entropy and enthalpy, respectively, between the activated complex and the reactants. It should be noted that κ was added *ad hoc* by Eyring, though he suggests that one can estimate it as "the average number of crossings required for each complex that reacts. It will generally be about unity" [74]. Since then, Chandler has shown how to calculate this factor using correlation functions, though noting that in condensed phase when the molecule couples strongly to the bath, the contributions to the rate constant made by this transmission coefficient would not simply be multiplicative [44]. Regardless, a rate constant calculated using TST represents an upperlimit to the true rate constant [41, 44, 74], with it being equivalent when there are no recrossing of the transition state barrier within the time-scale of the amount of time it takes an average reactant molecule to reach the transition state along the reaction coordinate (*i.e.*, recrossings are uncorrelated) [39]. This idea is Wigner's "fundamental assumption" for TST and is the true assumption

underlying transition state theory [44]. It can be restated as each region representing a local state reaches thermodynamic equilibrium quickly before a barrier crossing occurs [41]. Violating this condition, as is likely to happen in condensed phase environments, results in an overestimation of rate constants.

After TST, the next major development in rate theory occurred when Kramers treated a molecule undergoing a chemical reaction as a Brownian particle undergoing motion on a potential surface [45]. This particle diffuses along the potential surface, driven by microscopic thermal forces, which, through fluctuation-dissipation theory [78], are connected to the macroscopic temperature and friction of the 'bath' [39]. This bath can include molecular degrees of freedom not strongly coupled to the reaction coordinate along which the particle diffuses – so called 'internal friction'. As a result, Kramers' theory provides expressions for the rate constant in several limiting cases of weak, moderate-strong, and strong-overdamped friction. All of these expressions have the same shape as the Arrhenius equation, though with different prefactor values. The difficulty in knowing the friction along the reaction coordinate, and solving for the rate constant with intermediate values renders Kramers' theory difficult to implement in some cases. However, it does provide intuitive predictions for the effects of solvent viscosity, and local side-chain induced drag upon the rate constant. Additionally, the Kramers' theory expressions contains the curvatures of the potential surface at the local reactant, and transition states, as well as for the friction of the reaction. Interestingly, both TST and Kramers' theory are subsets of multidimensional transition-state theory in which the full phase-space containing all degrees of freedom in the system is considered [39]. Though, because of the difficulty in measuring and interpreting the friction and types of noise for a biomolecule, and the additional degrees of freedom produced by the increased number of variables in the Kramers' theory expressions relative to the TST expression, we opt to use TST to analyze biomolecular systems.

Despite being conceptually more straightforward, applying TST to biomolecular systems is still difficult [40]. Non-equilibrium effects like frictional dissipation of energy along the potential surface (the case treated by Kramers) make the absolute accuracy of the thermodynamic parameters extracted from kinetic measurements suspect. Moreover, because the enthalpy of the transition state energy barrier might have distinct temperature dependence, care should be taken when interpreting $\Delta H^{\ddagger}$s, and $\Delta S^{\ddagger}$s obtained from temperature-dependent kinetic studies in terms of the local microscopic dynamics of the biomolecule [8]. However, it is much easier and more sound to interpet local contributions of the biomolecular structure to relative changes in these parameters (*i.e.*, $\Delta\Delta H^{\ddagger}$s, and $\Delta\Delta S^{\ddagger}$s), because the prefactor terms will cancel and the results will be independent of terms like κ (inasmuch as these terms are constant) [79]. This approach is the basis of Phi value analysis, where local contributions to the transition state are interrogated by intro-

ducing a mutation a specific point in the biomolecule to alter an important interaction, and the contribution of the resulting change relative to the wild-type biomolecule attributed to $\Delta G^{\ddagger}$ or to ΔG [80]. Regardless, even if the transition state energy barrier thermodynamics are accurately measured, the possibility of coupling between local and collective motions of a biomolecule makes it difficult to interpret any localized structural contributions these parameters [8].

4.7.2 Temperature Dependence of Pretranslocation Dynamics

We investigated the temperature dependence of PRE dynamics by monitoring a smFRET signal that reports on the GS1 $\rightleftharpoons$ GS2 equilibrium (Fig. 4.11). As described in Sec. 4.4, PRE complexes were tethered to the surface of temperature-controlled microfluidic flow-cells [16]. These PRE complexes were labeled with a donor fluorophore (Cy3) on the P-site tRNA and with an acceptor fluorophore (Cy5) on the L1 stalk (23S rRNA helices H76-78 and ribosomal protein L1), so that the L1 stalk, P/E-state tRNA interaction that occurs in GS2 results in an E_{FRET} state of about 0.78 that is much higher than that in GS1 of about 0.15 [3]. By utilizing a labeling scheme that depends upon the presence of a specifically labeled tRNA in the P-site, we avoid introducing some heterogeneity by including unreactive PRE complexes with the incorrect tRNA in the P-site, as could occur by using the previously developed L1 stalk-L9 E_{FRET} signal which does not depend upon the presence of the correct tRNA [4]. The PRE complexes with the L1-tRNA signal were then monitored over a range of strictly controlled temperatures, and the resulting E_{FRET} versus time trajectories were individually processed using BIASD (Sec. 4.2) with the Laplace approximation to the posterior probability distribution to make the calculations tractable (Sec. 4.6.3). With posterior distributions for each PRE complex, mesoscopic ensembles were constructed using the variational mixture model described in Sec. 4.6.1 to learn the distribution of E_{FRET} and rate constants for GS1 and GS2 for each of the sub-populations within the mesoscopic ensembles at each temperature; this process is demonstrated for an L1-L9 PRE complex at one temperature in Fig. 4.19. In all cases, the mesoscopic ensemble was augmented by including an additional 100 random variates of the BIASD parameters from each posterior probability distribution in order to increase the statistics of each dataset in a process similar to bootstrapping. With the temperature dependence of the rate constants of the mesoscopic ensemble, transition state theory can invoked to quantify the relative thermodynamic contributions of the structural differences between the complexes to the PRE energy landscape.

We examined two PRE complex analogs, PRE^{-A}_{fMet} and PRE^{-A}_{Phe}, which contain deacylated, donor-fluorophore labeled $tRNA^{fMet}$ and $tRNA^{Phe}$ in the P site, respectively, and have vacant A sites. The PRE^{-A}_{fMet} com-

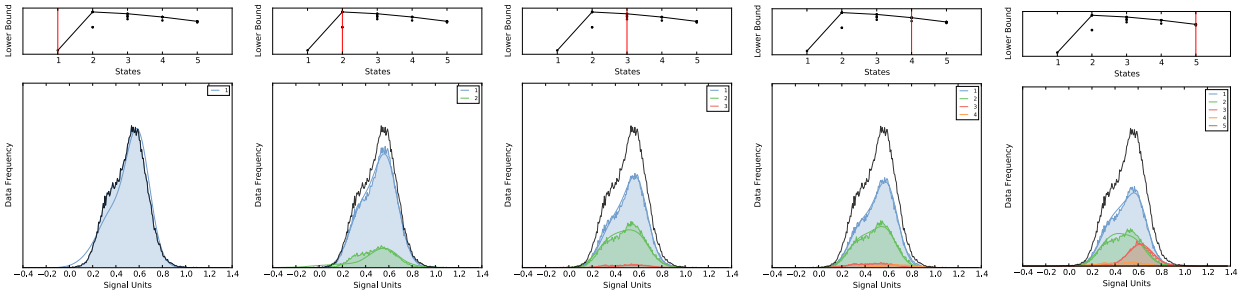

Figure 4.19: Hierarchical Selection of PRE Complex Sub-populations. smFRET measurements of deacylated PRE complexes containing a $tRNA^{Phe}$ in the P site using the L1-L9 FRET signal at 25 °C were analyzed using BIASD, and these results were clustered using a variational gaussian mixture model where each posterior was sampled 100 times to increase the statistics. Each plot shows a histogram of the data in the entire ensemble (black), and histograms of the data after it has been clustered into different classes (colors). Smooth curves represent the marginalized, predictive posterior distribution of the two-state BIASD model for each class. Significant deviation between the predictive posterior distribution and the histogram suggest the presence of additional classes. Above each plot is the lowerbound of the evidence; this suggests that the data is most parsimoniously represented by two classes, even though the residuals are lower when more classes are included.

plex is an analog to an initiation complex, while the PRE^{-A}_{Phe} complex is an analog to an elongation complex. Since the P site tRNA were site-specifically labeled with donor-fluorophores, any complexes where $tRNA^{Phe}$ failed to be delivered or undergo peptide-bond formation would not contribute any observable FRET. smFRET signal versus time trajectories were recorded from these complexes at 25, 28, 31, 34, and 37 °C . Additionally, we also examined an authentic PRE complex, $PRE_{Phe/Lys}$, where a deacylated donor-fluorophore labeled $tRNA^{Phe}$ occupies the P site, and a tripeptidyl fMet-Phe-Lys-$tRNA^{Lys}$ occupies the A site. The dynamics of the $PRE_{Phe/Lys}$ complexes are expected to be the most heterogeneous of the PRE complexes investigated here, because the acylation statuses of the tRNA involved in the complex formation are not uniform due to spontaneous hydrolysis of the peptide bond, and additional because the experiment can not distinguish between the presence or absence of $tRNA^{Lys}$ in the A site. Regardless, the dynamics of the $PRE_{Phe/Lys}$ complex were measured using the L1-tRNA signal at 22 (room-temperature), 25, 28, 31, and 37 °C . The data acquired at 31 and 37 °C were obtained using a shorter exposure time of τ = 33 ms in order to better resolve GS1 and GS2, though this does result in a lower average SNR of the E_{FRET} versus time trajectories. As in previous publications, at all temperatures measured, all three of these PRE complexes exhibited dynamic fluctuations between two E_{FRET} states that correspond to the open and closed L1 stalk-tRNA interaction that occurs in GS1 and GS2, respectively.

Across the range of temperatures investigated, the PRE^{-A}_{fMet} complexes favored GS1 (Fig. 4.20), which is consistent with previously published results [70], and the idea that $tRNA^{fMet}$ is less easily distorted than elongator tRNAs which must distort in order to promote the GTPase activation of EF-Tu during elongation. Interestingly, there are most likely two sub-populations within this mesoscopic ensemble of PRE^{-A}_{fMet} , and

the relative populations seem be temperature-invariant (∼3:1). While the PRE^{-A}_{Phe} results (Fig. 4.21) favor GS1 less than PRE^{-A}_{fMet} complexes due mostly to a slower k_{GS2}, this mesoscopic ensemble also seems most likely to contain two sub-populations with approximately the same, temperature-invariant ratio of ∼3:1. The origin of this heterogeneity is unclear. It is probably not due to incomplete deacylation by puromycin, as both subpopulations undergo fluctuations between GS1 and GS2, and the presence of an aa-tRNA in the P site blocks this transition. One possibility is that the two classes represent ribosomes with different ribosomal-protein compositions; for instance, S1 is a particularly weakly-bound protein with sub-stoichiometric occupation of the ribosome [49]. Overall, the less-populated state has both faster rate constants and slightly higher E_{FRET} values, so it is possible that either the different compositional or conformational types of the ribosome present expose the fluorophores to slightly different environments and have a different energy landscape along the GS1 ⇌ GS2 reaction coordinate, or that this is an artifact of the the mixture model – possibly because the posterior distribution is not symmetric like a normal distribution, and the only way for a mixture of normal distributions to compensate is by adding additional classes to fit the asymmetry.

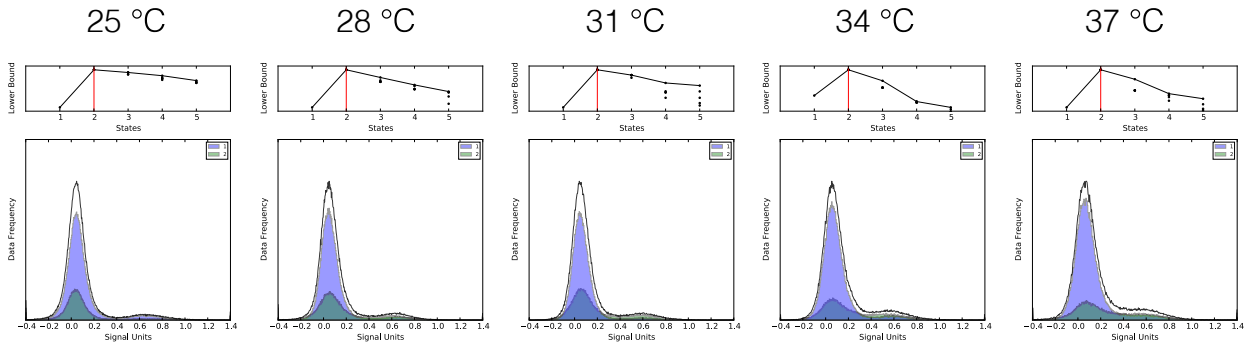

Figure 4.20: Temperature Dependence of Mesoscopic Ensembles of PRE^{-A}_{fMet} complexes. Each plot shows the histogram of experimentally observed E_{FRET} values (black) at a particular temperature. The variational lowerbounds are shown above each histogram. The data was classified according to the number of states with the largest value of the lowerbound. Histograms of the data classified to these states are shown in color.

The authentic $PRE_{Phe/Lys}$ complexes showed the fastest dynamics (Fig. 4.22). Consequentially, for the highest temperatures tested, we increased the acquisition time period to τ = 33 ms, and there was still a significant amount of blurring as evidenced by the density in between the E_{FRET} peaks corresponding to GS1 and GS2. Interestingly, for all but the room temperature T = 22 °C data, the lowerbound of the variational gaussian mixture model suggests that there are three sub-populations in the mesoscopic ensemble. The two major classes correspond to one class of $PRE_{Phe/Lys}$ complexes (∼30%) that favors GS1, and another class (∼55%) that favors GS2. The remaining class (∼15%) only slightly favors GS2. While it is unknown if these classes are static or whether they interconvert, the divergent dynamics of the major classes suggests

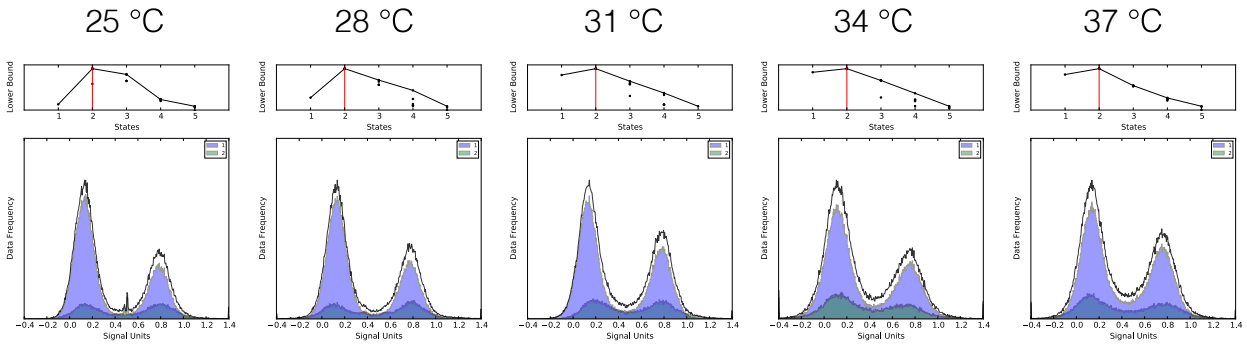

Figure 4.21: Temperature Dependence of Mesoscopic Ensembles of PRE^{-A}_{Phe} complexes. Each plot shows the histogram of experimentally observed E_{FRET} values (black) at a particular temperature. The variational lowerbounds are shown above each histogram. The data was classified according to the number of states with the largest value of the lowerbound. Histograms of the data classified to these states are shown in color.

that the variational mixture model has found two distinct types of $PRE_{Phe/Lys}$ complexes rather than creating a pathological overfitting class. However, the structural basis of these classes remains unclear.

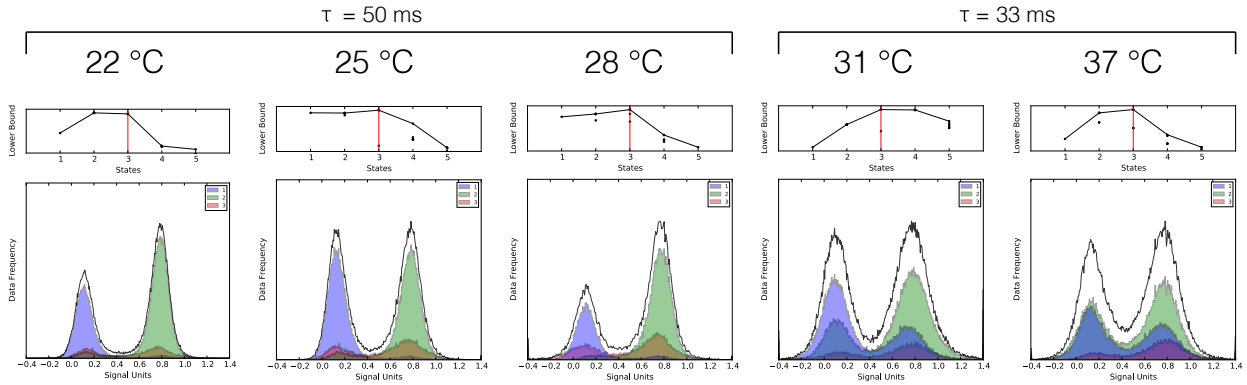

Figure 4.22: Temperature Dependence of Mesoscopic Ensembles of $PRE_{Phe/Lys}$ complexes. Each plot shows the histogram of experimentally observed E_{FRET} values (black) at a particular temperature. The variational lowerbounds are shown above each histogram. The data was classified according to the number of states with the largest value of the lowerbound. Histograms of the data classified to these states are shown in color.

Due to the lack of certainty in associating the various classes of PRE complexes across different temperatures, the ensemble averaged behavior of each PRE complex was utilized to quantify the temperature dependence of the rate constants and ΔG. The inferred rate constants and ΔG from a mesoscopic ensemble constructed using a single Gaussian in the mixture model were analyzed using maximum likelihood solutions to the equations for Gibbs free energy and transition state theory (Figure 4.23). The maximum

likelihood solution to $\Delta G = \Delta H - T\Delta S$ is

$$\mathcal{L} = P(D|\Theta) = \prod_i^N \mathcal{N}(\Delta G_i | \Delta H - T_i \Delta S, \Sigma)$$

$$\ln P \sim \sum_{i=1}^{N} \left(\Delta G_i - (\Delta H - T_i \Delta S)\right)^2 \tag{4.86}$$

$$\frac{\partial \ln P}{\partial \Delta S} = 0 = \sum_{i=1}^{N} \left(\Delta G_i - (\Delta H - T_i \Delta S) \cdot T_i\right) \implies \Delta S = \frac{\sum_{i=1}^{N} T_i (\Delta H - \Delta G)}{\sum_{i=1}^{N} T_i^2}, \tag{4.87}$$

$$\frac{\partial \ln P}{\partial \Delta H} = 0 = \sum_{i=1}^{N} \left(\Delta G_i - (\Delta H - T_i \Delta S)\right) \implies \Delta H = \frac{(\sum_{i=1}^{N} \Delta G_i) + \Delta S \sum_{i=1}^{N} T_i}{N}. \tag{4.88}$$

The dependence upon ΔS from Equation 4.88 can then be removed to yield the maximum-likelihood solution

$$\Delta H = \frac{\left(\sum_{i=1}^{N} \Delta G_i\right)\left(\sum_{i=1}^{N} T_i^2\right) - \left(\sum_{i=1}^{N} \Delta G_i T_i\right)\left(\sum_{i=1}^{N} T_i\right)}{\left(N \sum_{i=1}^{N} T_i^2\right) - \left(\sum_{i=1}^{N} T_i\right)^2}, \tag{4.89}$$

and then the maximum-likelihood solution for ΔS by solving Equation 4.87. The maximum-likelihood solution to the linearized form of transition state theory has a tractable solution that can be found as above for the Gibbs free energy case. For

$$y_i \equiv \ln\left(\frac{k_i h}{\kappa k_B T_i}\right) = \Delta S^\ddagger / R - \beta_i \Delta H^\ddagger, \tag{4.90}$$

where $\beta_i = (RT_i)^{-1}$, then

$$\ln P \sim \sum_{i=1}^{N} (y_i + \beta_i \Delta H^\ddagger - \Delta S^\ddagger / R)^2 \tag{4.91}$$

$$\frac{\partial \ln P}{\partial \Delta S^\ddagger} = 0 = \sum_{i=1}^{N} \left(y_i + \beta_i \Delta H^\ddagger - \Delta S^\ddagger / R\right) \cdot (-1/R) \implies \Delta S^\ddagger = \frac{R}{N}\left(\Delta H^\ddagger \left(\sum_{i=1}^{N} \beta_i\right) + \left(\sum_{i=1}^{N} y_i\right)\right) \tag{4.92}$$

$$\frac{\partial \ln P}{\partial \Delta H^\ddagger} = 0 = \sum_{i=1}^{N} \left(y_i + \beta_i \Delta H^\ddagger - \Delta S^\ddagger / R\right) \cdot \beta_i \implies \Delta H^\ddagger = \frac{\Delta S^\ddagger / R \cdot \left(\sum_{i=1}^{N} \beta_i\right) - \sum_{i=1}^{N} \beta_i y_i}{\sum_{i=1}^{N} \beta_i^2}. \tag{4.93}$$

Again, the dependence upon $\Delta S^\ddagger$ can be removed from Equation 4.93 to yield the maximum-likelihood

solution

$$\Delta H^{\ddagger} = \frac{\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} \beta_i\right) - N\sum_{i=1}^{N} \beta_i y_i}{N\left(\sum_{i=1}^{N} \beta_i^2\right) - \left(\sum_{i=1}^{N} \beta_i\right)^2}, \tag{4.94}$$

and then the maximum-likelihood solution for $\Delta S^{\ddagger}$ can be found with Equation 4.92. The resulting solutions are shown in Figure 4.23.

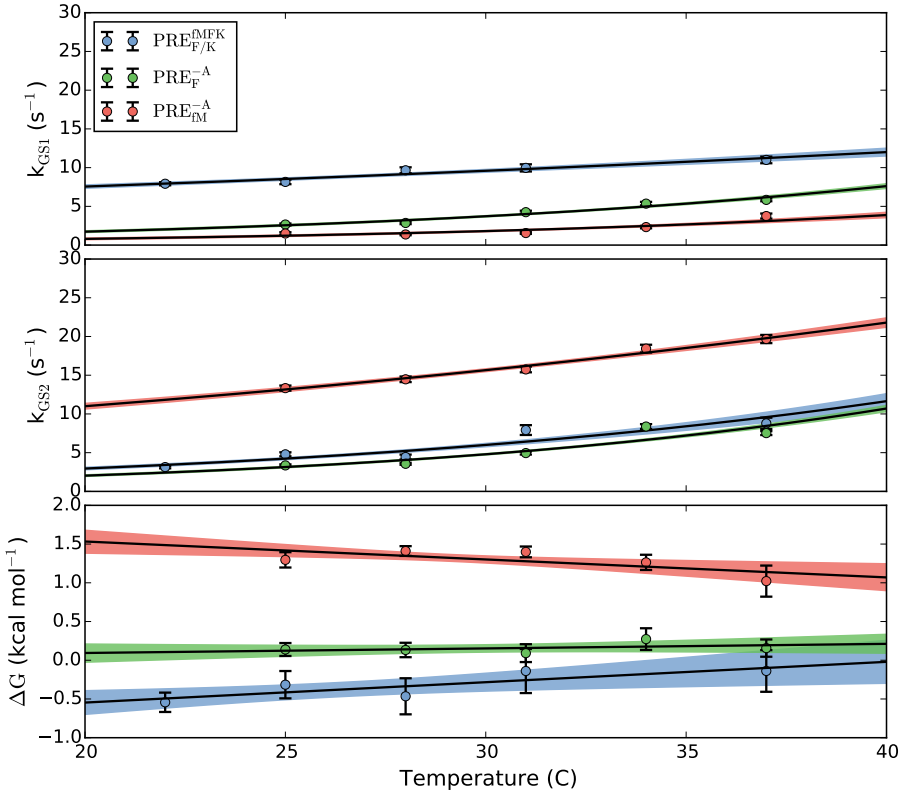

Figure 4.23: (A,B) Temperature dependence of PRE rate constants. Inferred mean rate constant ($\pm 1\sigma$) of entire ensemble of $PRE_{Phe/Lys}$, PRE^{-A}_{fMet} , and PRE^{-A}_{Phe} complexes. Maximum likelihood solutions to a linearized transition state theory with $\pm 1\sigma$ errorbars from bootstrapping resampled rate constants given the normal distribution about the inferred rate constants are shown as colored regions. (C) Temperature dependence of PRE Δ $G_{GS2-GS1}$. The mean Δ $G_{GS2-GS1}$ at each temperature is shown with $\pm 1\sigma$, and the maximum likelihood solution to $\Delta G = \Delta H - T\Delta S$, is shown with $\pm 1\sigma$ errorbars from bootstrapping as described above.

4.7.3 Transfer-RNA Contributions to the Pretranslocation Energy Landscape

By comparing the relative energetics of the PRE complexes reaction coordinates as analyzed using TST, the structural contributions that the various tRNA provide to the general PRE energy landscape can be inter-

preted. In order to facilitate these comparisons, both the PRE^{-A}_{fMet} and $PRE_{Phe/Lys}$ complexes are presented relative to the PRE^{-A}_{Phe} complex, such that the first comparison provides insight into the contribution of P-site tRNA structure, and the latter comparison provides insight into the contribution that A-site peptidyl-tRNA make to authentic PRE complexes. These relative thermodynamic contributions of the tRNA to the PRE energy landscape are presented as $\Delta\Delta G^{\ddagger}$ at 25 °C (Figure 4.24), though their interpretations do not change at 37 °C , or if $\Delta\Delta H^{\ddagger}$ or $\Delta\Delta S^{\ddagger}$ is considered instead.

The forward transition from GS1 to the transition state effectively requires an authentic PRE complex (*i.e.*, $PRE_{Phe/Lys}$), as $\Delta\Delta G^{\ddagger}$ is greater than k_BT relative to the PRE^{-A} complexes. The presence of the A-site peptidyl-tRNA accelerates the forward reaction, however this might also be an effect of having excess EF-Tu present in solution during the experiment. Interestingly, the A-site peptidyl-tRNA seems to have a negligible effect upon the reverse reaction, GS2 $\rightarrow$ GS1. Together, these suggest that the authentic PRE complexes predisposes the ribosome towards GS2, possibly to enhance the speed of translation elongation.

The P-site tRNA identity has a significant effect upon the reverse reaction as the $\Delta\Delta G^{\ddagger}_{GS2}$ for the PRE^{-A}_{fMet} complex is over one k_BT less that that of PRE^{-A}_{Phe} . This most likely reflects the decreased flexibility of $tRNA^{fMet}$ relative to that of $tRNA^{Phe}$, which renders it energetically unfavorable for $tRNA^{fMet}$ to maintain the distorted P/E conformation. Most likely, the GS2 state is destabilized by the structural differences between $tRNA^{fMet}$ and to elongator-tRNAs. These structural differences, such as the mismatched aminoacyl-acceptor stem basepair (C1-A72) or the purine-pyrimidine flipped basepair in the D-stem (A11-U24) in $tRNA^{fMet}$, have been well documented to affect PRE dynamics, and mutating these differences in $tRNA^{fMet}$ to their elongator-tRNA counterparts creates a more flexible $tRNA^{fMet}$ that restores elongator-tRNA-induced PRE complex dynamics [70]. As a complement to the destabilization of GS2 by $tRNA^{fMet}$, the forward transition from GS1 to the transition state is slowed by the presence of $tRNA^{fMet}$ relative to $tRNA^{Phe}$. This slowed rate of transition out of GS1 increases the likelihood that the PRE^{-A}_{fMet} complex will remain in the GS1 state, and is probably driven by the less-flexible $tRNA^{fMet}$ being energetically more unfavorable to distort than the more flexible $tRNA^{Phe}$ in order to reach the transition state from the stable P/P state.

Finally, we note that these observed PRE dynamics are essentially two-state, and do not exhibit extreme non-Arrhenius-like curvature over the range of temperatures observed here. As such, the above interpretations about the tRNA structure-dependent influences on PRE complex dynamics are established across a range of temperatures that a living organism might encounter in the world. Here, many of the distinctions between the dynamics of the various PRE complexes that were observed were induced by the structural differences of elongator and initiator tRNA. This might reflect the biological need for $tRNA^{fMet}$ to modulate

PRE dynamics during other steps of translation aside from elongation, such as initiation. As such, different sets of evolutionary pressures encouraged initiator and elongator tRNA to evolve to promote different modes of PRE complex dynamics. However, in both cases, distinct exploitation of PRE complex dynamics seems to be a general strategy used to modulate various steps of translation.

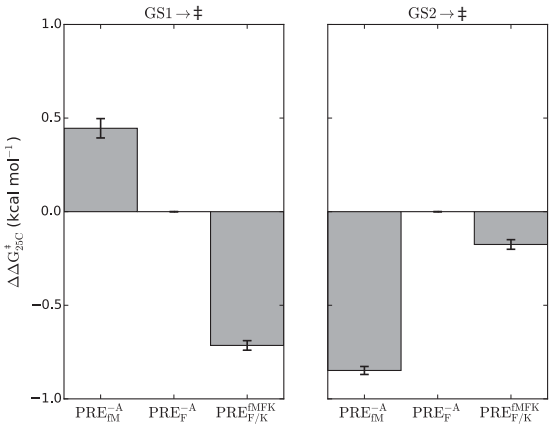

Figure 4.24: Relative values of $\Delta\Delta G^{\ddagger}_{25C}$ for different PRE complexes. The values of $\Delta\Delta G^{\ddagger}_{25C}$ relative to those of PRE^{-A}_{Phe} are shown for (left) GS1 to the transition state, and (right) GS2 to the transition state. Errorbars are $\pm 1\sigma$, and were derived by resampling and then calculating $\Delta\Delta G^{\ddagger}_{25C}$ from the maximum likelihood solutions to the linearized transition state theory expression.

4.8 References

1. Tinoco, I. & Gonzalez, R. L. Biological mechanisms, one molecule at a time. *Genes Dev.* **25,** 1205–1231 (2011).

2. Kinz-Thompson, C. D. & Gonzalez, R. L. smFRET studies of the 'encounter' complexes and subsequent intermediate states that regulate the selectivity of ligand binding. *FEBS Lett.* **588,** 3526–38 (2014).

3. Fei, J., Kosuri, P., MacDougall, D. D. & Gonzalez, R. L. Coupling of ribosomal L1 stalk and tRNA dynamics during translation elongation. *Mol. Cell* **30,** 348–59 (2008).

4. Fei, J., Bronson, J. E., Hofman, J. M., Srinivas, R. L., Wiggins, C. H. & Gonzalez, R. L. Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation. *Proc. Natl. Acad. Sci.* **106,** 15702–7 (2009).

5. Greenleaf, W. J., Frieda, K. L., Foster, D. A. N., Woodside, M. T. & Block, S. M. Direct Observation of Hierarchical Folding in Single Riboswitch Aptamers. *Science* **319,** 630–633 (2008).

6. Schafer, D. A., Gelles, J., Sheetz, M. P. & Landick, R. Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature* **352,** 444–448 (1991).

7. Boehr, D. D., Dyson, H. J. & Wright, P. E. An NMR perspective on enzyme dynamics. *Chem. Rev.* **106,** 3055–3079 (2006).

8. McCammon, J. A. Protein dynamics. *Reports Prog. Phys.* **47,** 1–46 (1984).

9. Gilmore, C. J. Maximum Entropy and Bayesian Statistics in Crystallography: a Review of Practical Applications. *Acta Crystallogr. Sect. A Found. Crystallogr.* **52,** 561–589 (1996).

10. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).

11. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).

12. Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez, R. L. & Wiggins, C. H. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* **97,** 3196–205 (2009).

13. Van De Meent, J.-W., Bronson, J. E., Wood, F., Gonzalez Jr., R. L. & Wiggins, C. H. Hierarchically-coupled hidden Markov models for learning kinetic rates from single-molecule data. *Proc. 30th Int. Conf. Mach. Learn.* (2013).

14. Van de Meent, J.-W., Bronson, J. E., Wiggins, C. H. & Gonzalez, R. L. Empirical Bayes Methods Enable Advanced Population-Level Analyses of Single-Molecule FRET Experiments. *Biophys. J.* **106,** 1327–1337 (2014).

15. Sivia, D. S. & Skilling, J. *Data Analysis: A Bayesian Tutorial* 1–259 (Oxford University Press, Oxford, 2006).

16. Wang, B., Ho, J., Fei, J., Gonzalez, R. L. & Lin, Q. A microfluidic approach for investigating the temperature dependence of biomolecular activity with single-molecule resolution. *Lab Chip* **11,** 274–81 (2011).

17. Fenimore, P. W., Frauenfelder, H., McMahon, B. H. & Parak, F. G. Slaving: solvent fluctuations dominate protein dynamics and functions. *Proc. Natl. Acad. Sci.* **99,** 16047–16051 (2002).

18. Lubchenko, V., Wolynes, P. G. & Frauenfelder, H. Mosaic energy landscapes of liquids and the control of protein conformational dynamics by glass-forming solvents. *J. Phys. Chem. B* **109,** 7488–7499 (2005).

19. Chung, H. S. & Eaton, W. a. Single-molecule fluorescence probes dynamics of barrier crossing. *Nature* **502,** 685–8 (2013).

20. Chung, S. H., Moore, J. B., Xia, L. G., Premkumar, L. S. & Gage, P. W. Characterization of single channel currents using digital signal processing techniques based on Hidden Markov Models. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **329,** 265–285 (1990).

21. Qin, F., Auerbach, a. & Sachs, F. A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophys. J.* **79,** 1915–1927 (2000).

22. Colquhoun, D. & Hawkes, A. G. On the stochastic properties of single ion channels. *Proc. R. Soc. London. Ser. B* **211,** 205–235 (1981).

23. Berezhkovskii, A. M., Szabo, A. & Weiss, G. H. Theory of single-molecule fluorescence spectroscopy of two-state systems. *J. Chem. Phys.* **110,** 9145 (1999).

24. Good, I. The Frequency Count of a Markov Chain and the Transition to Continuous Time. *Ann. Math. Stat.* **32,** 41–48 (1961).

25. Dobrushin, R. Limit theorems for a markov chain of two states. *Izv. Ross. Akad. Nauk. USSR Seriya Mat.* **17,** 291–330 (1953).

26. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21,** 1087 (1953).

27. Bishop, C. M. *Pattern Recognition and Machine Learning* 461–652 (Springer, New York, 2006).

28. Gopich, I. V. & Szabo, A. Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *Proc. Natl. Acad. Sci.* **109,** 7747–52 (2012).

29. Gibson, A. & Conolly, B. On a Three-State Sojourn Time Problem. *J. Appl. Probab.* **8,** 716–723 (1971).

30. Berezhkovskii, A. M., Szabo, A. & Weiss, G. H. Theory of the Fluorescence of Single Molecules Undergoing Multistate Conformational Dynamics. *J. Phys. Chem. B* **104,** 3776–3780 (2000).

31. Abate, J. & Whitt, W. A Unified Framework for Numerically Inverting Laplace Transforms. *INFORMS J. Comput.* **18,** 408–421 (2006).

32. Gopich, I. V. & Szabo, A. FRET efficiency distributions of multistate single molecules. *J. Phys. Chem. B* **114,** 15221–15226 (2010).

33. Roberts, G., Gelman, A. & Gilks, W. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7,** 110–120 (1997).

34. Galassi, M. *et al. GNU Scientific Library Reference Manual* Third (ed Gough, B.) (Network Theory Ltd., 2009).

35. Gillespie, D. T. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **81,** 2340–2361 (1977).

36. Voorhees, R. M. & Ramakrishnan, V. Structural basis of the translational elongation cycle. *Annu. Rev. Biochem.* **82,** 203–36 (2013).

37. Frank, J. Intermediate states during mRNA-tRNA translocation. *Curr. Opin. Struct. Biol.* **22,** 778–85 (2012).

38. Sternberg, S. H., Fei, J., Prywes, N., McGrath, K. a. & Gonzalez, R. L. Translation factors direct intrinsic ribosome dynamics during translation termination and ribosome recycling. *Nat. Struct. Mol. Biol.* **16,** 861–868 (2009).

39. Hänggi, P., Talkner, P. & Borkovec, M. Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.* **62,** 251–341 (1990).

40. Fersht, A. R. *Structure and mechanism in protein science. A guide to enzyme catalysis and protein folding.* 293–400 (W.H. Freeman and Co., New York, 1999).

41. Zwanzig, R. *Nonequilibrium Statistical Mechanics* 1–222 (Oxford University Press, Oxford, 2001).

42. Van Kampen, N. *StochasticProcesses in Physics and Chemistry* 3rd ed., 1–464 (North Holland, Amsterdam, 2007).

43. Wigner, E. Über das Überschreiten von Potentialschwellen bei chemischen Reaktionen. *Zeitschrift fur Phys. Chemie* **19,** 203–216 (1932).

44. Chandler, D. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.* **68,** 2959 (1978).

45. Kramers, H. A. Brownian Motion in a Field of Force and the Diffusion Model of Chemical Reactions. *Phys. VII,* 284–304 (1940).

46. Agirrezabala, X. *et al.* Structural characterization of mRNA-tRNA translocation intermediates. *Proc. Natl. Acad. Sci.* **109,** 6094–9 (2012).

47. Kuo, T.-L. *et al.* Probing static disorder in Arrhenius kinetics by single-molecule force spectroscopy. *Proc. Natl. Acad. Sci.* **107,** 11336–11340 (2010).

48. Björk, G. R. Genetic dissection of synthesis and function of modified nucleosides in bacterial transfer RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **50,** 263–338 (1995).

49. Van Knippenberg, P. H., Hooykaas, P. J. J. & Van Duin, J. The stoichiometry of E. coli 30S ribosomal protein S1 on in vivo and in vitro polyribosomes. *FEBS Lett.* **41,** 323–326 (1974).

50. Condon, C., Philips, J., Fu, Z. Y., Squires, C. & Squires, C. L. Comparison of the expression of the seven ribosomal RNA operons in Escherichia coli. *EMBO J.* **11,** 4175–4185 (1992).

51. Graille, M. *et al.* Molecular basis for bacterial class I release factor methylation by PrmC. *Mol. Cell* **20,** 917–927 (2005).

52. Zwanzig, R. Rate processes with dynamical disorder. *Acc. Chem. Res.* **23,** 148–152 (1990).

53. Xie, S. Single-Molecule Approach to Enzymology. *Single Mol.* **2,** 229–236 (2001).

54. Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H. & Gunsalus, I. C. Dynamics of ligand binding to myoglobin. *Biochemistry* **14,** 5355–5373 (1975).

55. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).

56. Kucherov, V. M., Kinz-Thompson, C. D. & Conwell, E. M. Polarons in DNA Oligomers. *J. Phys. Chem. C* **114,** 1663–1666 (2010).

57. Kinz-Thompson, C. & Conwell, E. Proton transfer in adenine-thymine radical cation embedded in B-form DNA. *J. Phys. Chem. Lett.* **1,** 1403–1407 (2010).

58. Kravec, S. M., Kinz-Thompson, C. D. & Conwell, E. M. Localization of a hole on an adenine-thymine radical cation in B-Form DNA in water. *J. Phys. Chem. B* **115,** 6166–6171 (2011).

59. Bronson, J. E., Hofman, J. M., Fei, J., Gonzalez, R. L. & Wiggins, C. H. Graphical models for inferring single molecule dynamics. *BMC Bioinformatics* **11,** S2 (2010).

60. Corduneanu, A. & Bishop, C. M. Variational Bayesian Model Selection for Mixture Distributions. *Artif. Intell.* **51,** 27–34 (2001).

61. Shaobo Hou & Galata, A. *Robust estimation of gaussian mixtures from noisy input data* in *2008 IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, 2008), 1–8.

62. Abramowitz, M. & Stegun, I. A. *Handbook of Mathematical Functions* 884 (Dover Publications, Mineola, 1965).

63. Hirsh, D. & Gold, L. Translation of the UGA triplet in vitro by tryptophan transfer RNA's. *J. Mol. Biol.* **58,** 459–468 (1971).

64. Hirsh, D. Tryptophan Transfer RNA as the UGA Suppressor. *J. Mol. Biol.* **58,** 439–458 (1971).

65. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74,** 129–177 (2005).

66. Astumian, R. D. Thermodynamics and Kinetics of a Brownian Motor. *Science* **276,** 917–922 (1997).

67. Frank, J. & Gonzalez, R. L. Structure and dynamics of a processive Brownian motor: the translating ribosome. *Annu. Rev. Biochem.* **79,** 381–412 (2010).

68. Garai, A., Chowdhury, D., Chowdhury, D. & Ramakrishnan, T. Stochastic kinetics of ribosomes: Single motor properties and collective behavior. *Phys. Rev. E* **80,** 011908 (2009).

69. Sharma, A. K. & Chowdhury, D. Distribution of dwell times of a ribosome: effects of infidelity, kinetic proofreading and ribosome crowding. *Phys. Biol.* **8,** 026005 (2011).

70. Fei, J., Richard, A. C., Bronson, J. E. & Gonzalez, R. L. Transfer RNA-mediated regulation of ribosome dynamics during protein synthesis. *Nat. Struct. Mol. Biol.* **18,** 1043–51 (2011).

71. Arrhenius, S. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift Phys. Chemie* **4,** 226–248 (1889).

72. Pfaundler, L. Beiträge zur chemischen Statik. *Ann. der Phys. und Chemie* **207,** 55–85 (1867).

73. Laidler, K. J. & King, M. C. The Development of Transition-State Theory. *J. Phys. Chem.* **87,** 2657–2664 (1983).

74. Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **3,** 107–115 (1935).

75. Evans, M. G. & Polanyi, M. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans. Faraday Soc.* **31,** 875 (1935).

76. Pelzer, H. & Wigner, E. The speed constansts of the exchange reactions. *ZEITSCHRIFT FUR Phys. CHEMIE-ABTEILUNG B-CHEMIE DER Elem. AUFBAU DER Mater.* **15,** 445–471 (1932).

77. Mahan, B. H. Activated complex theory of bimolecular reactions. *J. Chem. Educ.* **51,** 709 (1974).

78. Kubo, R. The fluctuation-dissipation theorem. *Rep. Prog. Phys.* **29,** 255 (1966).

79. Fersht, A. R. *Structure and mechanism in protein science. A guide to enzyme catalysis and protein folding* 349–400 (W.H. Freeman and Co., New York, 1999).

80. Matouschek, A., Kellis Jr., J. T., Serrano, L. & Fersht, A. R. Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340,** 122–126 (1989).

Part II

Experimental Studies

Chapter 5

Dynamics of Stop-Codon Discrimination by Release Factor 1

5.1 Introduction

The affinity and fidelity of biomolecular interactions are often described in solely in terms of enthalpic interactions such as hydrogen bonding [1, 2], or the monitoring of precise geometries such as Watson-Crick base pairing [3]. The importance of these types of enthalpic interactions to the affinity and fidelity of biomolecular interactions is readily over-promoted by techniques, such as X-ray crystallography, that present static pictures, often obtained using non-physiological conditions which can trap non-physiologically relevant interactions and damp out important conformational dynamics [4, 5]. Many biomolecules, such as proteins, RNA, and ribonucleoprotein complexes are inherently dynamic, and the accurate functioning of these biomolecules often depends upon these conformational dynamics [6–11]. For instance, conformational dynamics have been shown to be important for ensuring the fidelity of substrate selection; in fact, paromomycin, an aminoglycoside that promotes the misincorporation of aminoacyl-tRNAs at near-cognate codons does so by altering the conformational dynamics of 16S rRNA bases A1492 and A1493, which are, in part, responsible for enabling 'recognition' of the incorrect mRNA codon - tRNA anticodon pairing in the decoding center (DC) of the ribosome [5]. Similarly, tRNA mutations that alter the conformational dynamics of the tRNA, such as the Hirsh mutation found in the Hirsh supressor tRNA [12, 13] or disruption of the Schultz-Yarus base pair [14, 15], also promote misincorporation of aminoacyl-tRNAs at near-cognate codons [16]. It is quite possible that such modulation of conformational dynamics represents a general way for nature to regulate biomolecular function.

While bacterial class I peptidyl release factors, such as release factor 1 (RF1) (see Chapter 1), interact

with the ribosome in a manner similar to that of tRNA-EF-Tu(GTP) ternary complex during tRNA selection, the mechanism by which RF1 regulates its high-affinity binding and accurate recognition of stop codons is ill-understood (recently reviewed in Refs. 17, and 18). For instance, RF1 recognizes mRNA stop codons UAA and UAG in the DC, while discriminating against codons with a single base difference (near-stop codons) by a factor of 10^3 to 10^5 – a level seemingly unattainable from just hydrogen bonding differences [19] and in the absence of a traditional kinetic proofreading mechanism [20, 21] (see Section 1.1.1). Interestingly, mutations in RF1 that are distal to both the DC and the peptidyl transferase center (PTC) can alter the codon specificity of RF1 to additionally recognize the UGA stop codon (charge-swap mutations) [22], or to prevent recognition of the UAG stop codon while leaving UAA recognition intact (switch loop mutations) [23]. Because of their distance from the DC, presumably these mutations function by a mechanism other than disrupting or forming direct interactions with the mRNA codon. Similarly, the aminoglycoside antibiotic paromomycin has been observed to differentially affect stop-codon and near-stop codon recognition by RF1, despite it not interacting directly with the release factor [24–28].

Here, in an effort to speak to the contributions that conformational dynamics make to ligand binding and ligand discrimination, we investigated the binding affinity and codon recognition abilities of RF1. Binding affinity and codon recognition were monitored with single-molecule fluorescence resonance energy transfer (smFRET) total internal fluorescence reflection (TIRF) microscopy using a previously developed smFRET signal between RF1 and the peptidyl site peptidyl-tRNA that reports upon RF1 binding [29], and a biochemical peptide release assay [24, 30, 31]. Using these techniques, we investigated the effect of RF1 and DC dynamics by comparing wild type RF1 behavior to that of RF1 with a novel mutation, which is distal from the DC and the PTC, at stop and near-stop codons. These interpretations are additionally informed by complementary molecular dynamics simulations of RF1, as well as bioinformatics approaches. Finally, we investigated the effect of paromomycin on the binding kinetics of RF1. These studies show that the modulation of RF1 and/or ribosome dynamics within the context of the DC in the ribosomal aminoacyl (A) site is crucial for maintaining the binding affinity and codon discrimination ability of RF1, and therefore for the fidelity of translation termination.

5.2 Results

5.2.1 The switch loop of ribosome-bound RF1 aligns with mutations in the D-loop stem of pre-accommodated tRNA that control fidelity

The switch loop of RF1, located from residues 291 to 307 (in *Escherichia coli (E. coli)* numbering), spans the region between RF1 domains three and four. From X-ray crystallography structures, when RF1 is bound to the A-site of the ribosome at a stop-codon, the switch loop contacts residues which form the DC of ribosome [27, 28]. Specifically, this includes h44 of the 16S rRNA, and H69 of the 23S rRNA – in particular, A1492 and A1493 of the 16S rRNA [18]. These residues have been implicated in maintaining fidelity during the tRNA selection step of translation elongation [5, 32].

We aligned the phosphates of the 16S rRNA from X-ray crystallography structures of *Thermus thermophilus* (*T. thermophilus*) ribosomes containing kiromycin-stalled EF-Tu(GTP)-tRNA where the tRNA is in the pre-accommodated A/T state [33], and also RF1 bound at a stop-codon [28] using the program VMD [34] (Figure 5.1). This pre-accommodated tRNA, which is on-pathway during tRNA selection, overlaps significantly with the stop-codon bound RF1 as both are located in the A-site of the ribosome. In particular, the switch loop region of RF1 overlaps with the D-loop stem (D-arm) of the A/T tRNA. This overlap includes nucleotides in positions 24 and 27, for which mutations to reduce the fidelity of tRNA selection. The mutation of G24A is found in the Hirsh supressor tRNA, which enables $tRNA^{Trp}$ to misincorporate Trp at UGA stop-codons [12, 13], presumably by increasing the flexibility of the tRNA and allowing it to more readily undergo distortions necessary for tRNA accommodation [32]. Similarly, the same phenotype of in vivo, non-canonical recognition of the stop-codon UGA can be induced by mutations which disrupt the base pair (Schultz and Yarus Pair) formed by nucleotide 27 [14, 15], and is thought to be caused by the increased flexibility of the tRNA.

5.2.2 Residues G299 and G301 have the highest mutual information in the switch loop

Evolutionary pressures fight against mutations to genomic sequences in order to conserve molecular function. In this fight, compensatory mutations can be be made to rescue function (or develop new function) from a deleterious mutation; these compensatory mutations need not be strictly structural (*i.e.*, maintaining a hydrogren-bond), but can also be made to maintain dynamic, functional changes within a biomolecule.

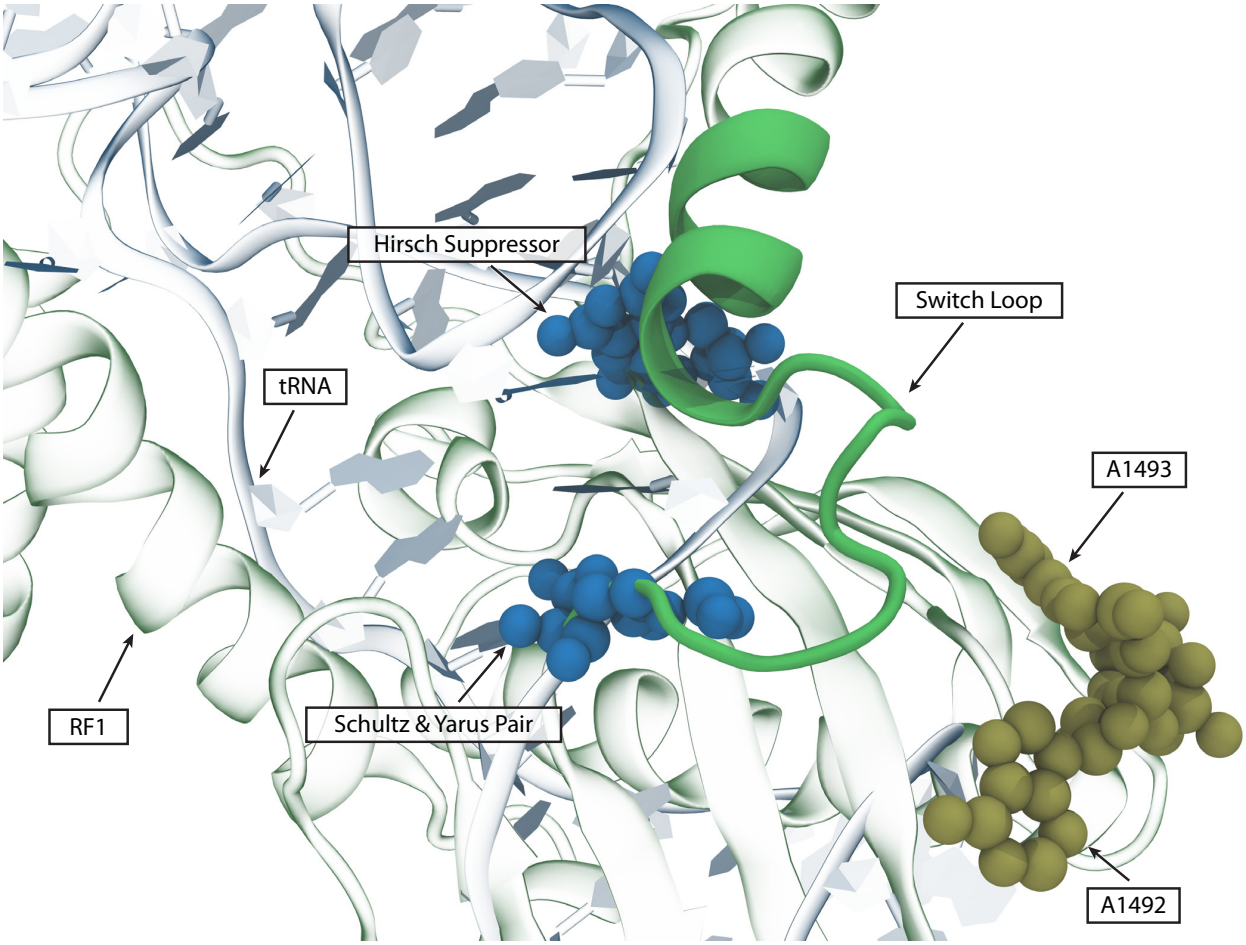

Figure 5.1: Alignment of A/T tRNA and RF1. Pre-accommodated A/T tRNA (PDB:4V5G) is shown aligned to stop codon bound RF1 (PDB:4V7P) by the 16S rRNA. The switch loop of RF1 is shown in green. Bases whose dynamics are associated with the fidelity of tRNA selection, such as the site of the Hirsh suppressor mutation (G24A) or part of the Schultz and Yarus pair (G27) are shown in blue. 16S rRNA residues A1492 and A1493, which comprise part of the decoding center, are shown in tan.

With these compensatory mutations in mind, the pairwise evolution of biomolecular sequences provides a living record of biomolecular functionality in the sequence conservation across species. In particular, this coevolution of residues (nucleotide or protein) can be used to infer functional relationships between distal inter- and intra-molecular regions of a biomolecule(s), and, in particular, it is capable of identifying functional- and specificity-determining residues [35]. Therefore, we sought to investigate RF1 switch loop function by analyzing its coevolution within the context of parts of the 16S and 23S rRNA, and also the rest of RF1. However, since regions of the rRNA near the switch loop are mostly immutable (*i.e.*, h44 and H69), here, we present only the intramolecular relationships within RF1.

To investigate the coevolution of RF1, we used mutual information as a proxy for coevolution [36]. Mutual information (MI) is a mathematical definition of the dependence of two variables, and can be thought of as an entropic measure of how well one variable can be used to describe the other. MI ranges between 0 and 1, representing no dependence and maximal dependence, respectively, and is calculated as,

$$MI(x,y) = \sum_{x}\sum_{y} p(x,y) \cdot \ln\left(\frac{p(x,y)}{p(x)\cdot p(y)}\right), \tag{5.1}$$

where p is the probability of an event. For the purposes of coevolution, x and y would be the different residues at two different positions within the primary sequence of RF1. Therefore the joint-probability $p(x,y)$ would be a 20 by 20 matrix representing the frequency of finding each pair of amino-acids at the two positions under consideration. The MI is then calculated from the pair-wise frequencies obtain from sequence alignments of a representative primary sequences. Here, we obtained the protein sequences of 9097 bacterial RF1 from NCBI ref-seq [37], specifically excluding plastidial and mitochrondrial RF1, and including only one strain per species, so as not to over-represent certain species (we did not control for over-represented genuses). These sequences were then aligned pairwise using Clustal Omega [38], and finally the MI was calculated using Equation 5.1.

The MI between all residues in bacterial RF1 are plotted in Figure 5.2, using the *E. coli* numbering and domain labels. MI is a symmetric quantity (*i.e.*, the same for residues A and B, as for residues B and A), so the MIs below the diagonal are the same as above. Notably, in the 'tripeptide anticodon' PXT motif found in domain two that is partially responsible for codon recognition in the decoding center [39], the proline and threonine are highly conserved, while the variable X position coevolves with domains three and four more than one or two. Similarly, the highly conserved GGQ motif that is important for catalyzing the hydrolysis of nascent polypeptide chains [40], is extremely invariant, suggesting that the sequences used here are

appropriately aligned. Interestingly, the domain structure can be seen in off-diagonal triangles (*c.f.* domain three) as the local residues within a domains probably coevolve with each other more than the average residue.

Many residues in the switch loop have a significant amount of MI between all other domains of RF1. This significant amount of MI is also true for the majority of residues in RF1, and makes it difficult to locate regions of interest. This complicating feature is probably, in part, due to the many interactions that RF1 must make *in vivo*, such as with the N^5-glutamine methyltransferase encoded by prmC that methylates the GGQ motif [41–43], or with RF3 when both are bound to the ribosome [44], in addition to its the ribosomal interactions (*c.f.*, Section 5.2.1 and Refs. 45, and 18) necessary for its role in the termination hydrolysis reaction [46].

Within the switch loop, there is also a significant amount of MI (Figure 5.2, inset). Of these positions within the switch loop, the pair of residues with the most MI is G299 and G301. Previously, a G301S mutation to RF1 was shown to suppress termination by RF1 at the UAG stop-codon by increasing read-through *in vivo*, and also rendered the *E. coli* cells more susceptible to mRNA cleavage by the toxin RelE, which cleaves A-site mRNA at UAG in a ribosome-dependent manner [23, 47]. From Figure 5.1, G299 and G301 do not directly interact with any rRNA, however, in this X-ray crystallography structure of the *T. thermophilus* RF1, the threonine residue in the middle at position T300 (serine in *E. coli*) does contact A1493; so it is possible that the dynamics of G299 and G301 affect this interaction.

5.2.3 RF1 switch loop mutant G299A, G301A releases significantly less dipeptide than wild type RF1 at near-stop codons, but not at stop codons

Previously, we had cloned the *E. coli* RF1 gene, prfA, and co-expressed this wildtype RF1 (wtRF1) with the release factor methyltransferase encoded by prmC in order to produce wtRF1 that is active in *in vitro* translation termination assays [29, 48]. Here, from this original wtRF1 construct, we generated the switch loop mutant G299A, G301A RF1 (mutRF1) using site-directed mutagenesis to mutate G896C and G902C in the wildtype prfA gene. These constructs (wtRF1 and mutRF1) were then purified using a nickle-affinity chromatography, and size exclusion chromatography.

We then investigated the ability of wtRF1 and and this newly constructed mutRF1 to release nascent polypeptide chains from the ribosome using a standard radioactive release assay [24, 30, 31]. Briefly, an *in vitro* transcribed mRNA message, fM-K-X where X is either the UAA stop-codon or UAU near-stop codon, is translated using formyl-$\left[^{35}S\right]$-Met-tRNA$^{\text{fMet}}$ and Phe-tRNA$^{\text{Phe}}$ to create ribosomal release complexes (RC_{UAA}

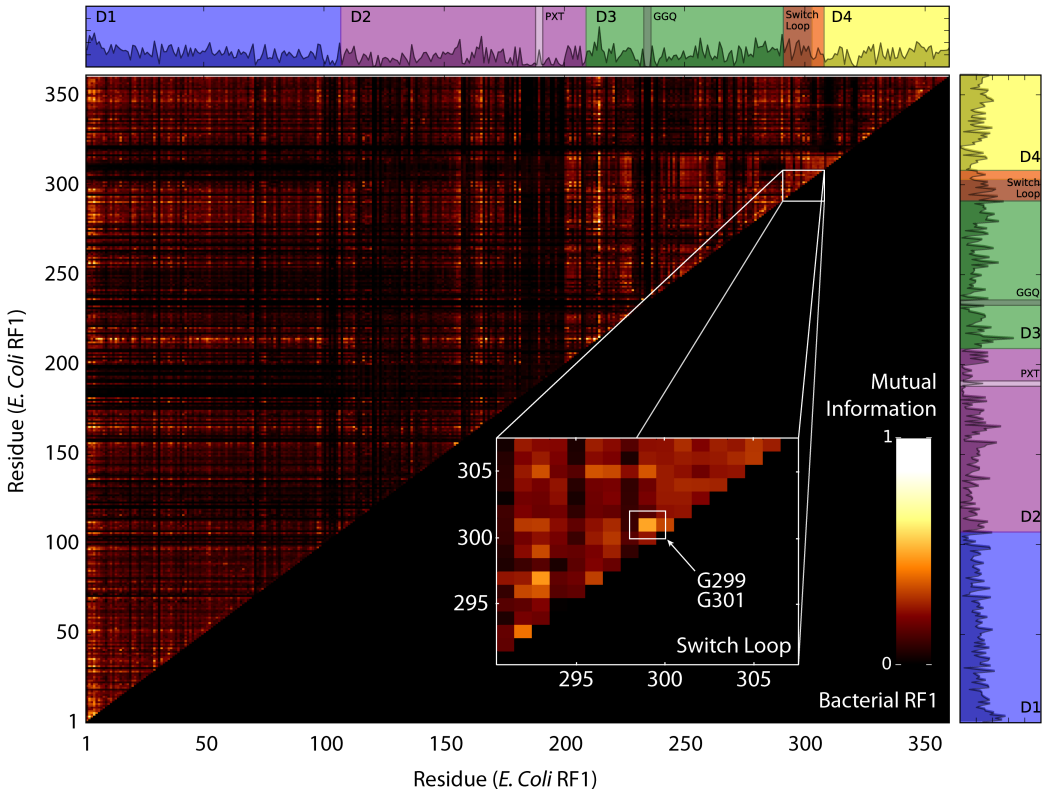

Figure 5.2: Mutual Information of Bacterial RF1 Primary Sequence. Mutual information is shown for n = 9097 bacterial RF1 sequences aligned with Clustal Omega, and was plotted using *E. coli* RF1 numbering, and ignoring any gaps. RF1 domains are annotated above and to the right. Inset shows the mutual information in the switch loop region of RF1.

and RC_{UAU}, respectively), which contain radiolabeled-dipeptide that is ready to be released by RF1. The radiolabeled dipeptide is then released from the RC by incubation with RF1, and then released and unreleased dipeptide are separated using electrophoretic thin-layer chromatography (eTLC). The eTLC plate is then exposed on a phosphorimaging screen, which is then scanned and quantified.

As shown in Figure 5.3, RC_{UAA} was incubated for 5 sec at room temperature with 25 $\mu\mathrm{M}$ (125-fold excess) of either wtRF1 or mutRF1. In both cases, 81% of the f-$\left[^{35}S\right]$-Met was released – presumably the experimental upper limit – while, in the absence of RF1, a negligible amount was released – presumably due to spontaneous, water-catalyzed hydrolysis. As measured by Freistroffer *et al.*, the K_M for wtRF1 release at both UAA and UAG is $\sim$8 nM [49]. Therefore, for wtRF1, Michaelis-Menten kinetics suggests that the reaction rate should >99.9% of the maximum rate. Most likely, the K_M is similar for mutRF1 catalyzed release on RC_{UAA}, as the amount of dipeptide released is the same.

When RC_{UAU} was incubated for 10 min at room temperature with 25 $\mu\mathrm{M}$ wtRF1, 80% of the dipeptide was released. The K_M for peptide release by wtRF1 at the near-stop codon UAU is $\sim$3.3 $\mu\mathrm{M}$ [19], so the saturating amount of wtRF1 that was used suggests that the reaction rate should be 88% of the maximum reaction rate, and most likely given the additional incubation time, the reaction proceeded to completion. However, when the release assay is performed by incubating RC_{UAU} for 10 min at room temperature with 25 $\mu\mathrm{M}$ mutRF1, there was no detectable release of dipeptide, and repeating this experiment at 37 °C yielded only a negligible amount of release (8%), which might be due to the increased basal level of hydrolysis by water. Regardless, while mutRF1 can catalyze release of dipeptide from the stop-codon containing RC_{UAA}, it does not seem to be able to catalyze release at a near-stop codon, which is unlike wtRF1.

5.2.4 Altering the dynamics of RF1 residues 299 and 301 biases the conformation of the switch loop, and can alter the dynamics of the entire release factor

Since the G299A, G301A mutations of mutRF1 differentially affect dipeptide release at stop and near-stop codons without directly forming or breaking new molecular interactions (*c.f.*, Section 5.2.3), we performed molecular dynamics simulations of wtRF1 and mutRF1 to provide insight into any effects the mutations have on the dynamics of RF1. As a starting point, we isolated the *T. Thermophilus*, wild type RF1 that was bound to the ribosome at a stop codon in the X-ray crystallography structure from Noller and coworkers from its surroundings [28]; this structure of wild type RF1 is in the 'open' conformation as it spanned the DC to the PTC. This wild type RF1 was then embedded in a rectangular $\sim 100^3$ Å^3 box of TIP3P water, such that, in

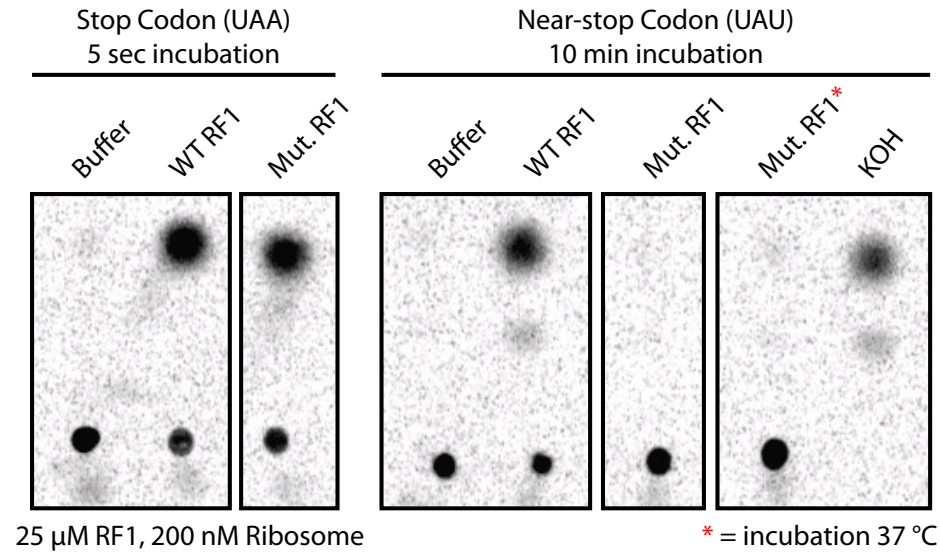

Figure 5.3: Wild Type and Mutant RF1 Dipeptide Release Assay. E-TLC separated products from a fMet-Phe dipeptide release assay after a 5 sec incubation of 25 μM wild type or mutant RF1 at room temperature with 200 nM termination complexes with the stop codon UAA in the A-site, or after a 10 min incubation of wild type or mutant RF1 with 200 nM termination complexes with the near-stop codon UAU in the A-site. Data from Dr. D. Pulukkunat.

each direction, the wild type RF1 was buffered by at least 10 Å of water molecules. The system was then neutralized, and brought to 150 mM salt concentration using Na^+ and Cl^- counterions. This system was then minimized and allowed to relax using Desmond [50] and the OPLS-AA force field [51], then a 1 ns long molecular dynamics simulation was run using an NPT ensemble at 300 K and atmospheric pressure. This short simulation was performed to remove any crystal-packing or ribosome-induced artifacts from the wild type RF1 structure, and the final structure was used as the starting point for the following molecular dynamics simulations.

From this common starting structure, the side chains of G294 and G296 were mutated to methyl groups for the *T. Thermophilus*-equivalent mutation of G299A, G301A. This mutant RF1 and the original wild type RF1 were then simulated for 11 ns, and the first was discarded as an equilibration time. The effect on the local dynamics of the switch loop can be monitored with the Φ and Ψ angles for T295, which is between the sites of the mutations. The time series of Φ_{T295} and Ψ_{T295} is shown for both the wild type and mutant RF1 in Figure 5.4. As more clearly seen in the Viterbi path from a 2D-variational Bayes HMM (Appendix D) of the Φ and Ψ trajectories, the switch loop of wild type RF1 is much more dynamic than that of mutant RF1. Additionally, the switch loop of wild type and mutant RF1 differentially sample areas of conformational space (Figure 5.5, left).

To investigate how the glycine to alanine mutations at positions 299 and 301 alter the dynamics of the

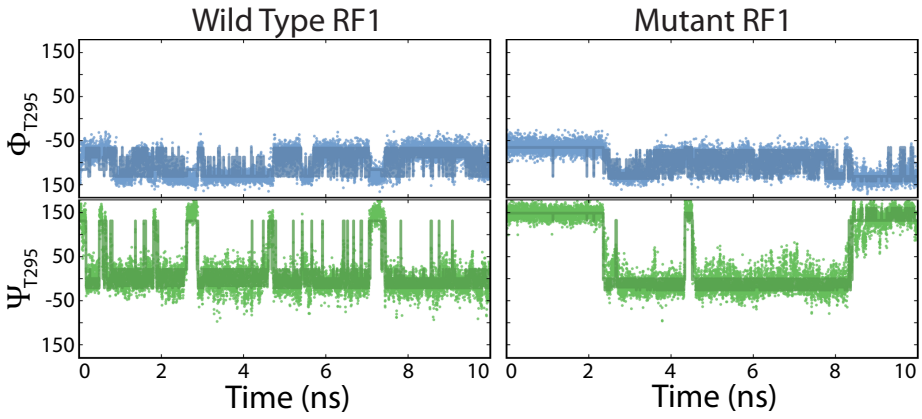

Figure 5.4: Mutant and Wild Type RF1 Molecular Dynamics Trajectories. Scatter plot of Φ and Ψ for residue T295 (*i.e.*, S300 in *E. coli*) in the context of the wild type RF1 (left) and mutant G294A, G296A RF1 (G299A, G301A in *E. coli*) (right). Overlapped is a the trajectory with the most likely states as determined by a 2-D variational HMM.

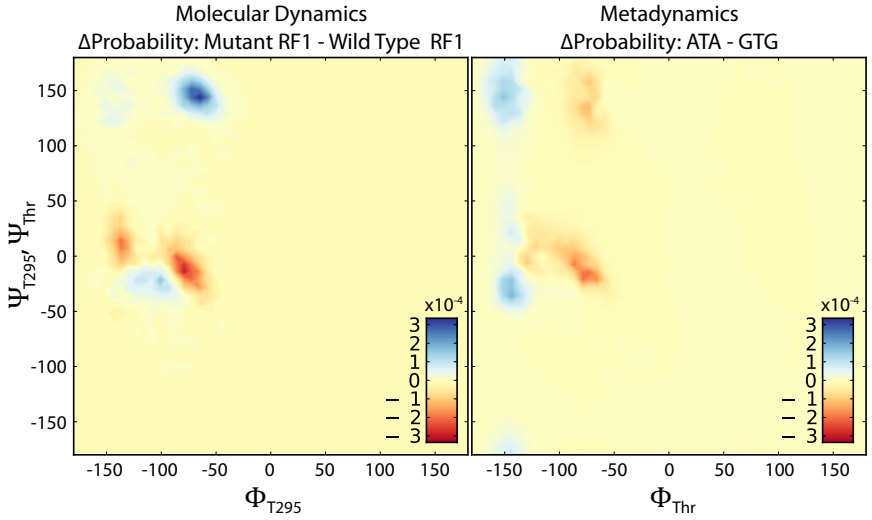

Figure 5.5: Simulation Population Differences between Mutant and Wild Type RF1 and Tripeptides. (Left) The difference in probability for Φ_{T295} and Ψ_{T295} (S300 in *E. coli*) as seen in the molecular dynamics simulation of wild type and mutant release factor. Positive values correspond to more occupancy in the simulation of the mutant RF1, while negative correspond to more occupancy of the wild type RF1. (Right) The difference in probability for Φ_{Thr} and Ψ_{Thr} as seen in the metadynamics simulation of GTG and ATA tripeptides. Probability was calculated as $P \sim e^{-\Delta G/RT}$. Positive values correspond to more expected occupancy of the ATA tripeptide, while negative correspond to more occupancy of the GTG tripeptide.

entire switch loop, rather than just the local dynamics of the intervening residue 300, we performed a metadynamics simulation [52] to calculate the free energy surface of the Φ and Ψ for the tripeptides GTG and ATA, corresponding to the wild type and mutant RF1 switch loop, respectively. This free energy surface should reflect the local, nearest neighbor contributions to the dynamics observed in Figure 5.4, and any differences should be due to the contribution of the rest of RF1. The free energy surfaces calculated with Desmond [50] are shown in Figure 5.6. From these free energies of Φ_{Thr} and Ψ_{Thr}, we can calculate the expected populations of the conformations at 300 K, the temperature of the molecular dynamics simulations of wild type and mutant RF1. The differences between the regions expected to be sampled by Φ_{Thr} and Ψ_{Thr} in the GTG and ATA tripeptides should reflect the local contributions of residues 294 and 296 to the dynamics of Φ_{T295} and Ψ_{T295}, and deviations from those distributions should reflect the contribution of wild type or mutant RF1 (Figure 5.5). The differences do not match those from the molecular dynamics simulation. This is primarily due to Φ_{T295} and Ψ_{T295} in the mutant RF1 over-occupying the (150°,-50°) region rather than (150°,-150°) or (0°,-150°) as is expected from the metadynamics free energy surface. This suggests that altering the dynamics of residues 299 and 300 by introducing the glycine to alanine mutations dynamically alters the conformation of the switch loop.

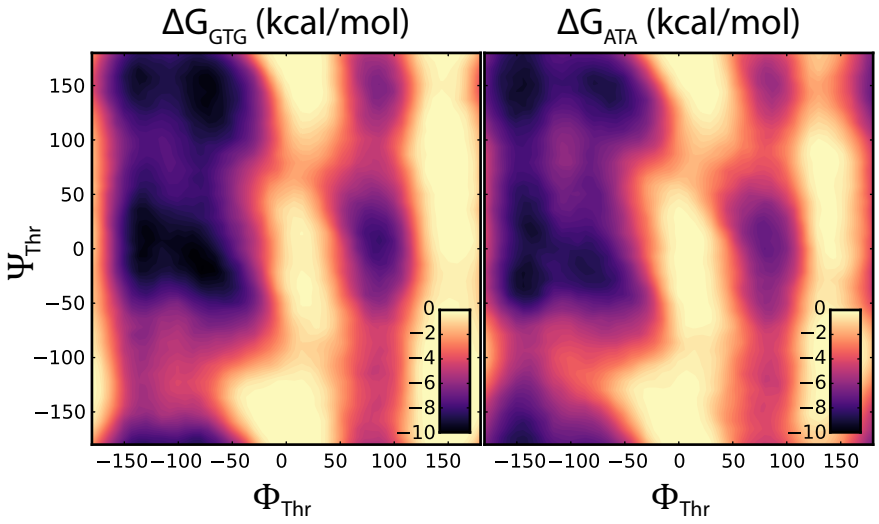

Figure 5.6: Tripeptide Free Energy Surfaces from Metadynamics Calculation. The Gibbs free energy as calculated from a metadynamics simulation for the tripeptide that is analogous to the wild type RF1 switch loop sequence, GTG (left), and for the tripeptide that is analogous to the mutant RF1 switch loop sequence, ATA (right).

With regard to the effect of these mutations on the dynamics of the rest RF1, we performed dynamical network analysis [53] on the wild type and mutant RF1 molecular dynamics simulation trajectories in order to investigate how the mutations affect the dynamics of RF1. In this approach, communities of residues within

the wild type and mutant RF1 are detected from the correlated motions of the residues in the molecular dynamics trajectories [53, 54]. These communities partition the molecule into clusters of highly connected and communicating residues, and information such as allosteric communications must flow from community to community through a limited set of residues [55]. Disruption of these communities can then result in a disruption of the communication between distal regions of the molecule in question.

Several of the communities found in wild type and mutant RF1 are shown in Figure 5.7. The importance of these particular communities is seen when looking at the sub-optimal paths through these communities between the PVT motif, which recognizes the stop-codon in the DC, and the GGQ motif, which catalyzes polypeptide hydrolysis in the PTC. For wild type RF1, these paths only pass through one intermediary community (green) between the PVT motif and the domain three alpha helix that supports the GGQ motif. However, for the mutant RF1, these paths must pass through an additional intermediary community (not shown). This disconnection between the originating (purple) and intermediary community (green) is a disruption of communication those communities, and is possibly represents a 'short-circuiting' of the flow of information from codon-recognition in the DC to the hydrolysis that subsequently occurs in the PTC. However, from this study, it is unclear how this communication and the communities would be affected in the presence of mRNA and the ribosome.

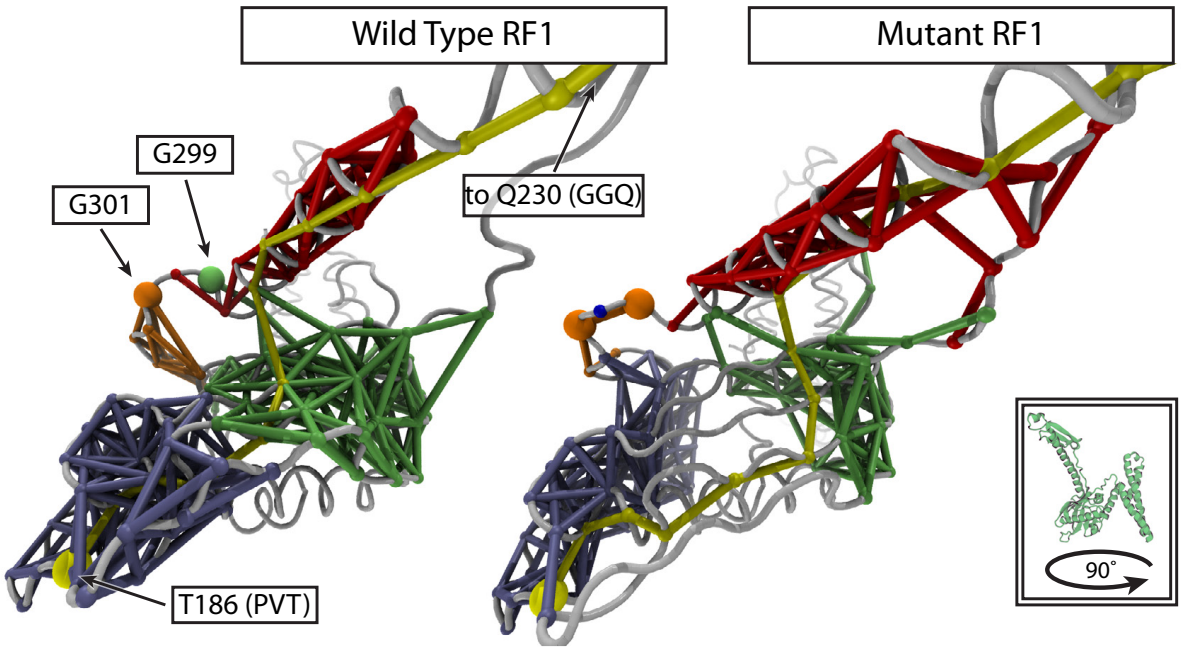

Figure 5.7: Network Analysis of Wild Type and Mutant RF1. The backbone trace of wild type (left) and mutant (right) RF1 is shown in gray. Several network-partitioning communities are shown (purple, green, orange, and red). Representative sub-optimal paths between Thr186 of the PVT motif, and Gln230 of the GGQ motif are shown in yellow.

5.2.5 Mutations to the switch loop of RF1 alter the binding affinity of RF1 for stop and near-stop codon programmed ribosomal release complexes

In order to further investigate how the G299A, G301A mutation in the switch loop of RF1 can differentially affect stop and near-stop codon recognition, we used smFRET to monitor the binding kinetics of wtRF1 and mutRF1 to various ribosomal release complexes (Figure 5.8). Briefly, we utilized a wtRF1 variant from

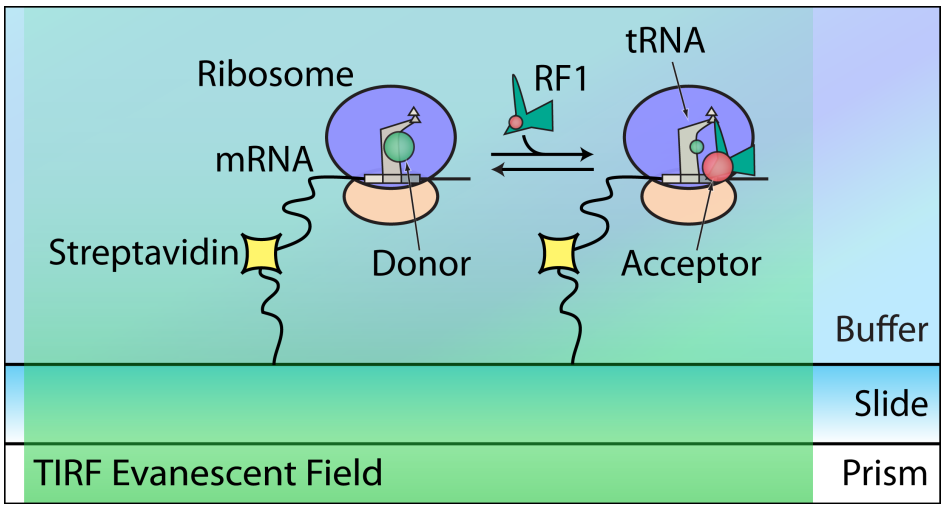

Figure 5.8: Cartoon of RF1 and RC smFRET Experiment. Ribosomes containing mRNA with either a stop-codon (RC_{UAA}) or a near-stop codon (RC_{UAU}) in the A-site are immobilized to the surface of a slide using a biotin-streptavidin-biotin bridge. These ribosomes contain a donor-fluorophore labeled $tRNA^{Phe}$ in the P-site, that will undergo FRET to an acceptor-fluorophore labeled RF1 when it is bound in the A-site. TIRF of a laser beam in a prism creates an evanescent field that illuminates the samples near the slide surface.

which we had previously [29] removed all wild type cysteines (C51S, C201S, C257S) and introduced a single cysteine at position 167 (S167C) – all mutations have been shown to retain wild-type activity [29, 56]). This wtRF1 variant was then labeled using a maleimide-derivatized acceptor fluorophore, Cy5, at position S167C using Michael-addition of the sulfhydryl group of the cysteine to the maleimide, and the labeled wtRF1 (wtRF1-(Cy5), herein) was then purified to 100% labeling efficiency using hydrophobic-interaction chromatography. Similarly, the G299A, G301A mutations were introduced into the S167C single-cysteine construct; this was labeled with Cy5, and also purified 100% labeling efficiency (mutRF1-(Cy5), herein). To create donor-fluorophore labeled, stop and near-stop codon programmed ribosomes, we constructed ribosomal release complexes that were labeled with NHS-ester derivatized Cy3 at position 47 of $tRNA^{Phe}$, which is a primary amine-containing 3-(3-amino-3-carboxypropyl)-uridine, through amide formation with the NHS-ester and primary amine, which were also purified to 100% labeling efficiency using hydrophoic-interaction chromatography (see Ref. 48 for a review). These Cy3-labeled release complexes programmed with stop

and near-stop codons in the A-site are herein referred to as RC_{UAA}-(Cy3) and RC_{UAU}-(Cy3), respectively. We then proceeded to image donor and acceptor fluorophore fluorescence from RC_{UAA}-(Cy3) or RC_{UAU}-(Cy3) with wtRF1-(Cy5) or mutRF1-(Cy5) in solution using TIRF microscopy in order to monitor RF1 binding kinetics *via* smFRET.

As previously demonstrated [29], wtRF1-(Cy5) binds stably to the stop-codon programmed RC_{UAA}-(Cy3) (Figure 5.9, top). In the example fluorescence intensity versus time trajectory of this reaction, loss of Cy5

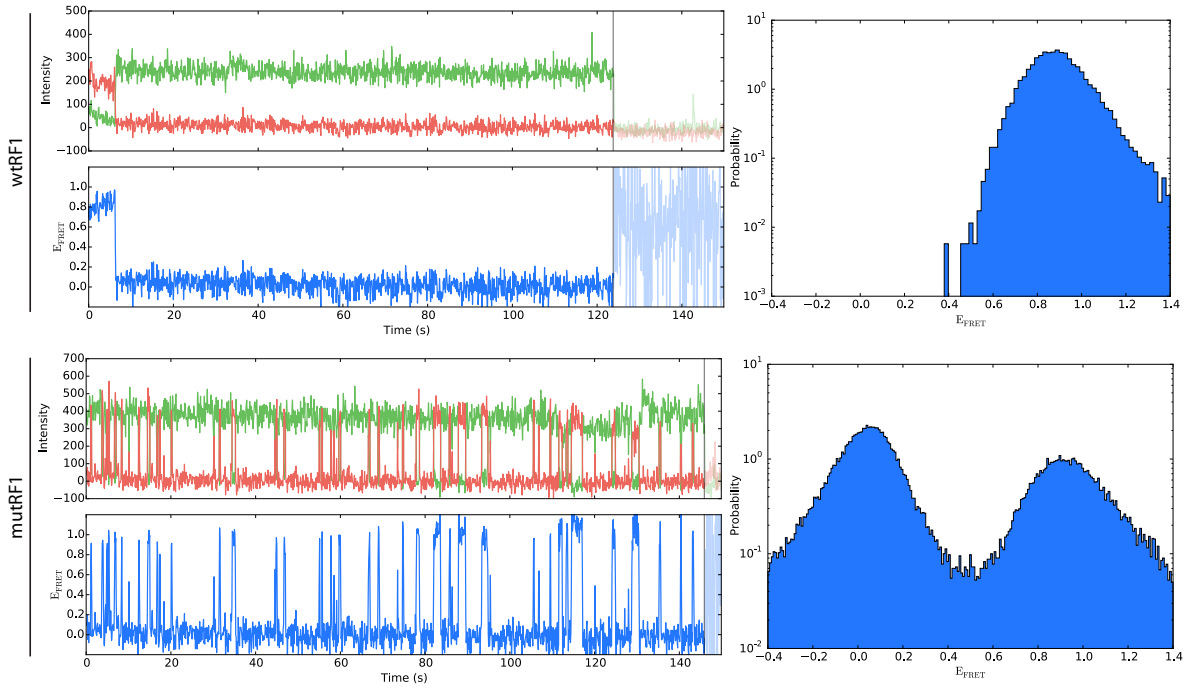

Figure 5.9: smFRET from 10 nM wtRF1 and mutRF1 Binding to Stop-codon Programmed RC_{UAA}. (Top) Surface-immobilized RC_{UAA}-(Cy3) was imaged with 10 nM wtRF1-(Cy5) in solution. Stable binding of one wtRF1-Cy5 occurs from the beginning of the trajectory, and continues past Cy5 photobleaching, until, at least, Cy3 photobleaching occurs. On the right, a $\log_{10}$-scaled histogram of E_{FRET} from n = 147 of these trajectories truncated at Cy5-photobleaching. (Bottom) Similarly, a surface-immobilized RC_{UAA}-(Cy3) is imaged with 10 nM mutRF1-(Cy5) in solution. The trajectory shows multiple binding and dissociation events before Cy3-photobleaching occurs. The $\log_{10}$-scaled histogram is composed of E_{FRET} trajectories from n = 364 of these complexes. However, unlike the wtRF1-(Cy5) case, the peak centered at ~ 0 primarily represents RC_{UAA} in the absence of mutRF1-(Cy5).

fluorescence intensity occurs after a few seconds of imaging. This loss of fluorescence intensity is due to photobleaching of the Cy5 fluorophore, rather than dissociation of wtRF1-(Cy5), as no other putative binding events occur after this point, despite there being 10 nM wtRF1-(Cy5) in solution. In the associated histogram of multiple E_{FRET} versus time trajectories, the peak at $\sim$0 is therefore due to photobleaching. Previously, we have demonstrated that this loss of fluorescence in laser-power dependent, and therefore due to photobleaching, rather than dissociation [29].

Unlike wtRF1-(Cy5), mutRF1-(Cy5) binds and dissociates very rapidly from RC_{UAA}-(Cy3) (Figure 5.9,

bottom). With 10 nM mutRF1-(Cy5) in solution, many binding events to a single RC_{UAA}-(Cy3) occur within a single E_{FRET} versus time trajectory before Cy3 photobleaching occurs. Some of these binding events are so transient that they do not even last a single time period. The log-scaled histogram of the E_{FRET} versus time trajectories, which presumably contains little to no photobleaching, shows peaks at $E_{FRET} \approx 0.05$ and $E_{FRET} \approx 0.93$. The former corresponds to the RF1 unbound state with no FRET occurring; the non-zero value being attributable to Cy3 fluorescence bleed-through into the Cy5 fluorescence intensity and also Cy5 autofluorescence from excitation-cross talk of mutRF1-(Cy5) in solution at relatively high concentrations. The latter corresponds to the RF1 bound-state, and the E_{FRET} matches that observed for wtRF1-(Cy5) binding to RC_{UAA}-(Cy3). Additionally, in the region of the histogram between the peaks corresponding to bound and unbound mutRF1-Cy5, there is a moderate amount of time-averaged E_{FRET} signal, though, for reasons discussed below, we do not apply BIASD (*c.f.*, Chapter 4).

As expected, the binding kinetics of mutRF1-(Cy5) to RC_{UAA}-(Cy3) are dependent upon the concentration of mutRF1-(Cy5) in solution. Lowering the concentration of mutRF1-(Cy5) in solution during imaging from 10 to 1 nM yields noticeably fewer binding events (Figure 5.10). Anecdotally, in the representative

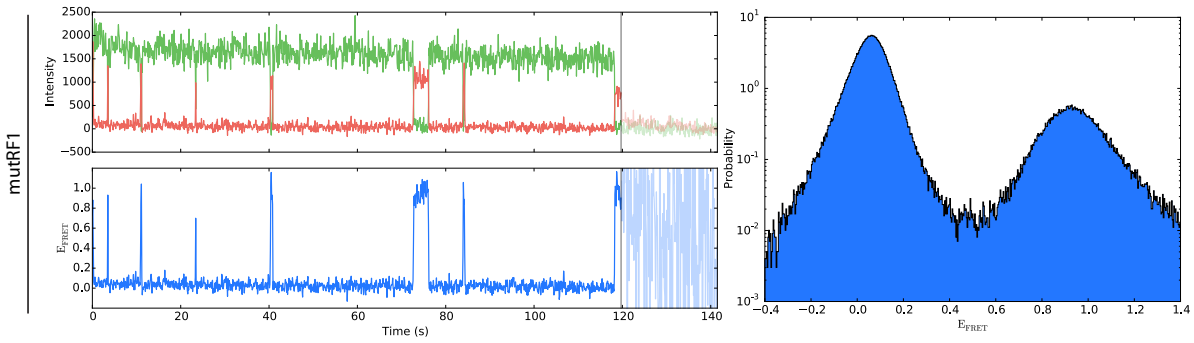

Figure 5.10: smFRET from 1 nM mutRF1-(Cy5) Binding to Stop-codon Programmed RC_{UAA}. As in Figure 5.9, an example smFRET trajectory and a histogram of n = 768 E_{FRET} versus time trajectories is shown for RC_{UAA}-(Cy3) imaged with 1 nM mutRF1-(Cy5) in solution.

E_{FRET} versus time trajectory shown here it seems that there might be multiple different timescales for dissociation of mutRF1-(Cy5) from RC_{UAA}-(Cy3) when comparing the multiple $\sim$1 frame-long dwell times at the beginning of the trajectory to the $\sim$30 frame-long dwell time in the middle of the trajectory. However, since this is a stochastic system, it could be a statistical coincidence. Looking back at the representative E_{FRET} versus time trajectory for 10 nM mutRF1-(Cy5) binding to RC_{UAA}-(Cy3) (Figure 5.9, bottom), there are similar situations with many transient dwell-times and several long-lived dwell-times. Regardless, comparing the histogram of all of the observed E_{FRET} versus time trajectories observed with 1 nM to that with 10

nM mutRF1-(Cy5) in solution, we find that the two maxima are located at approximately the same E_{FRET} , but that the peak associated with the mutRF1-(Cy5)-bound RC_{UAA}-(Cy3) complex ($E_{FRET} \approx 0.9$) is larger at the higher concentration, while the peak associated with the unbound complex is smaller at the higher concentration. Additionally, the widths of both peaks in the 1 nM mutRF1-(Cy5) histogram are narrower, because there is less excitation-crosstalk from the indirection excitation of the unbound mutRF1-(Cy5) in solution. Finally, there are fewer of the blurred, time-averaged E_{FRET} measurements between the two main peaks at 1 nM than at 10 nM mutRF1-(Cy5), which suggests that there is a slower rate constant at the lower concentration.

To investigate the kinetics of mutRF1-(Cy5) binding to RC_{UAA}-(Cy3), we thresholded the E_{FRET} versus time trajectories mid-way between the peaks in histograms in Figures 5.9 and 5.9 (E_{FRET} = 0.45), and used the resulting state versus time trajectories to calculate the survival function of dwell-times. The resulting curves are plotted in black in Figure 5.11. Fitting the survival curves to exponential decays conditioned upon

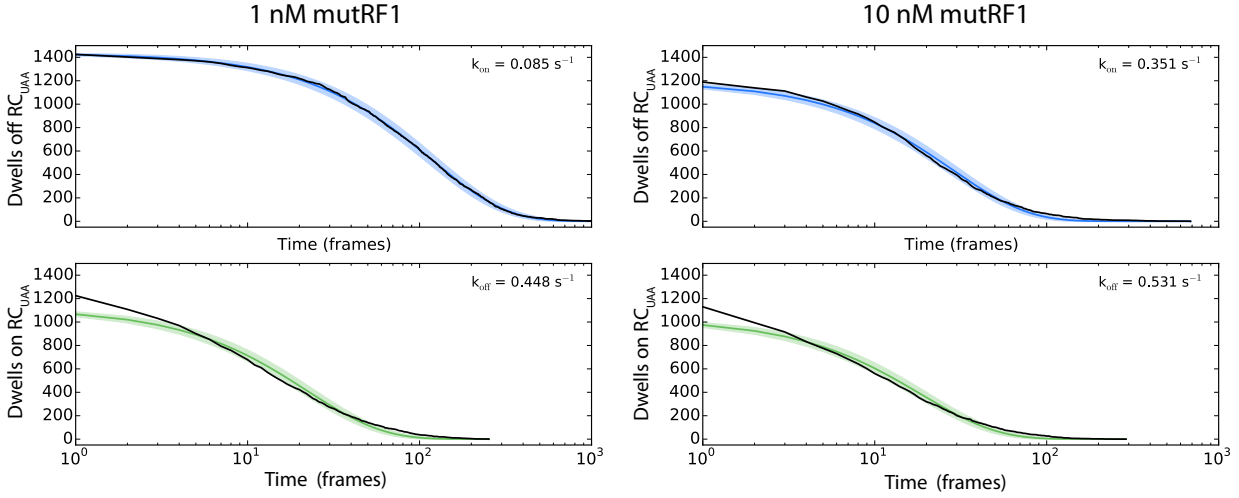

Figure 5.11: Concentration-dependence of the Kinetics of mutRF1 Interactions with RC_{UAA}. The number of dwell-times observed to be greater than a number of frames is plotted (black) for mutRF1 dissociated from (top) and bound to (bottom) surface-immobilized RC_{UAA} for 1 nM mutRF1-(Cy5) (left) and 10 nM mutRF1-(Cy5). Non-linear fits to single exponential decays are shown for $k_{\rm on}$ (blue) and $k_{\rm off}$ (green) with the colored region representing $\pm 3\sigma$ for the number of expected dwells given the stochastic decay [57]. The dwells on RC_{UAA} are non-single exponential.

all observed dwell-times being greater than one-measurement period long as described in Section 3.2.2 provides a rate constant that was can be used to determine that the decays are non-Markovian. While this rate constant predicts the average decay of the survival curve, since the binding and dissociation reactions are stochastic, the experimentally observed curves will only follow the average, and will deviate a certain amount that depends upon the total number of molecules considered. The time-dependent variance around

the mean for a stochastic unimolecular elementary reaction was derived by McQuarrie [57] and is

$$\sigma^2 = N_0 e^{-k \cdot t} \cdot (1 - e^{-k \cdot t}), \tag{5.2}$$

where σ^2 is the variance, k is the rate constant, t is time, and N_0 is the number of molecules at t = 0. Notably, the RC_{UAA}-(Cy3)-bound dwell-times ('on' dwells) seem ill-described by such a Markovian reaction mechanism, because of the substantial deviation from the $\pm 3\sigma$ region shown for the fitted-mean decay. In particular, there are hundreds of short-dwell times that are one, two, or three measurement periods long, which are unaccounted for by the Markovian model. This deviation is not accounted for by changing the threshold used to generate the state versus time trajectories. Additionally, this non-Markovian behavior does not seem to be induced by the compounding of dwell-times due to missed-events described in Section 3.3.4, because the mean decay rate constants are too slow to see such an effect. It is because of this apparent non-Markovian behavior that we have opted to utilize the thresholding approach rather than using an HMM or BIASD. Regardless, this fitted-rate constant is about the inverse of the average dwell-time, and as such is still useful to quantify the relative dwell-times for the various conditions. Notably, the apparent rate constant for mutRF1-(Cy5) association to RC_{UAA}-(Cy3) increase with increased concentration of mutRF1-(Cy5); though the increase is not the expected 10-fold even though the experiments were performed as serial dilutions from a common stock solution of mutRF1-(Cy5), though this could be partially due to the deficient method used to calculate these values. Additionally, the rate constants for the dissociation of the mutRF1-(Cy5) from RC_{UAA}-(Cy3) are relatively independent of mutRF11-(Cy5) concentration, as is expected for a unimolecular reaction.

As would be expected for a bimolecular reaction, Hetrick and coworkers showed that the second order association rate constant for RF1 binding to stop-codon programmed ribosomes differs little relative to near-stop or sense-codon programmed ribosomes by using bulk, stopped-flow fluorescence experiments [58]. Because their results seem to be a convolution of a biphasic approach to equilibrium [59], we estimated the upper-limited for a diffusion controlled association rate constant of a spherical RF1 to the A-site of a spherical ribosome. From the density of globular proteins, the minimal radius of a sphere of protein which could contain a certain mass of protein is

$$\mathrm{R_{min}} = 0.066\mathrm{M}^{1/3}, \tag{5.3}$$

where $R_{\min}$ is the radius of the sphere in nm, and M is the mass of the protein in Da [60]. For wtRF1-(Cy5), which is approximately 41.2 kDa, $R_{\min} \approx$ 2.3 nm. From X-ray crystallography structures, the diameter of the prokaryotic 70S ribosome is 210 Å [61], so the R_{ribosome} = 10.5 nm. We approximated the A-site as a circle with radius $R_{\text{A-site}}$ = 4.8 nm, from an average of measurements of the A-site diameter from an X-ray crystallography structure of RF1 bound to the *T. thermophilus* ribosome [28]. The diffusion current of these RF1 spheres to a tethered ribosome sphere that is accessible from all directions, I_0, is then

$$I_0 = 4\pi D_{\text{RF1}} \cdot R_{\text{ribosome}} \cdot [\text{RF1}]\,, \tag{5.4}$$

where D_{RF1} is the diffusion constant of RF1 in buffer, and $[\text{RF1}]$ is the concentration of RF1 in solution [62]. The diffusion constant can be calculated with the Stokes-Einstein equation

$$D_{\text{RF1}} = \frac{RT}{6\pi\eta R_{\min}}, \tag{5.5}$$

where R is the gas constant, T is the temperature, and η = .0091 Pa·s is the viscosity of water. For the RF1 sphere binding to ribosome spheres, we find that I_0 = 8.3 x 10^8 M^{-1} s^{-1}, however we are only interested in the diffusion current of RF1 to the A-site of the ribosome. We approximate the A-site as a single, disk-like absorber on the surface of the ribosome sphere. The diffusion current to the A-site, I, is then reduced from I_0 by

$$I = I_0 \frac{1}{1 + \frac{\pi R_{\text{ribosome}}}{n \cdot R_{\text{A-site}}}}, \tag{5.6}$$

where $n = 1$ is the number of disk-like absorbers [62]. In this case, we find that the second-order association rate constant is $k_{\text{association}} = I \approx 100$ μM^{-1} s^{-1}. This value is an estimate of the upper-limit of the diffusion-limited association rate constant, since RF1 was treated as a sphere of the minimum radius possible for a protein even though it is known that domains one and three of RF1 are relatively oblong for both the 'closed' [63, 64] and 'open' conformations [26–28]. Regardless, it is only about 2 to 3 times larger than that measured for wtRF1 binding to stop-codon programmed ribosomes [58, 59], as well as the rate constants measured using smFRET for mutRF1-(Cy5) binding to RC_{UAA}-(Cy3) (Figure 5.11).

We then investigated the binding kinetics of wtRF1-(Cy5) and mutRF1-(Cy5) to the near-stop codon programmed RC_{UAU}-(Cy3) using smFRET (Figure 5.12). The wtRF1-(Cy5) E_{FRET} versus time trajectories exhibit only rare, transient binding events. Presumably, there is at least the same rate of encounters between the

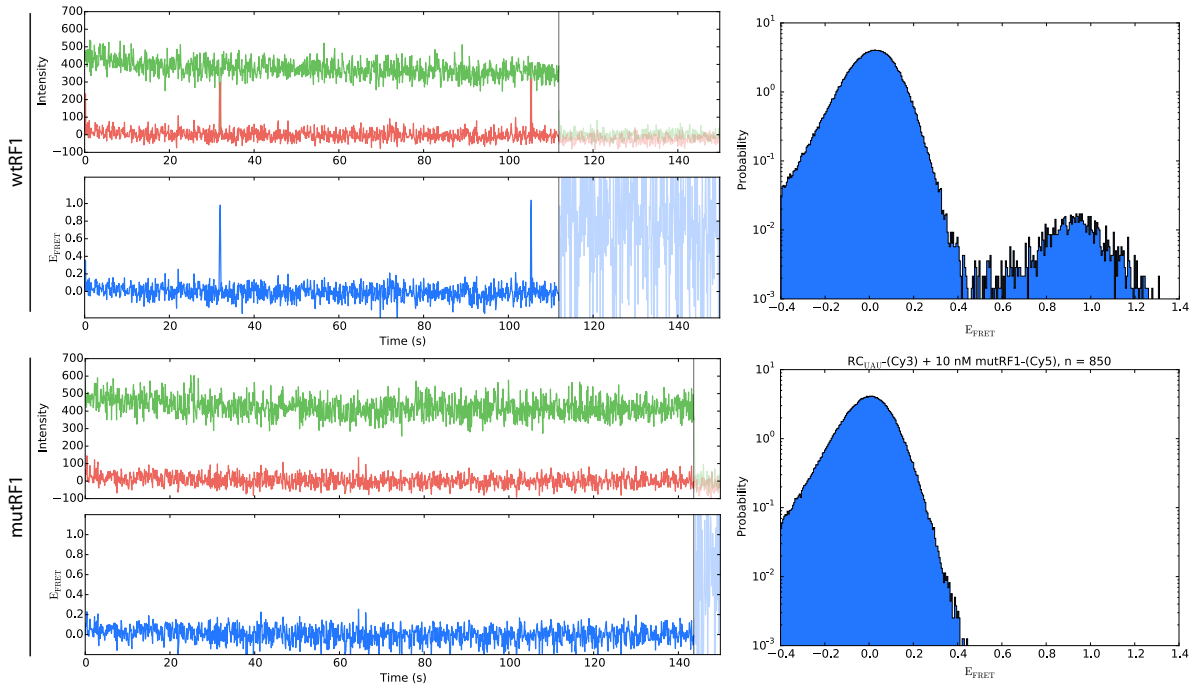

Figure 5.12: smFRET from wtRF1 and mutRF1 Binding to Near-stop Codon Programmed RC_{UAU}. As in Figure 5.9, an example smFRET trajectory and a histogram of n = 807 and n = 850 E_{FRET} versus time trajectories is shown for RC_{UAU}-(Cy3) imaged with both 10 nM wtRF1-(Cy5) and mutRF1-(Cy5) in solution, respectively.

RF1 and the ribosomal release complex as in the mutRF1-(Cy5) RC_{UAA}-(Cy3) experiment (Figure 5.9), because there is the same concentration of RF1 in solution. This suggests an encounter-complex type reaction. The histogram of the E_{FRET} versus time trajectories for wtRF1-(Cy5) binding to RC_{UAU}-(Cy3) has a distinctive high-E_{FRET} state peak at E_{FRET} = 0.95, which is the same as that of wtRF1-(Cy5) and mutRF1-(Cy5) binding to RC_{UAA}-(Cy3), however it is much smaller, because there are many fewer binding events, and these are generally much shorter. mutRF1-(Cy5) exhibits no detectable binding to RC_{UAU}-(Cy3) (Figure 5.12). In the histogram of the E_{FRET} versus time trajectories, there is no discernable mass outside of the peak centered at $E_{FRET} \approx 0$, despite there being 10 nM mutRF1-(Cy5) in solution. Again, we expect the same rate of RF1 encounters with the ribosome as were seen with mutRF1-(Cy5) binding to RC_{UAA}-Cy3 as the only difference is the third nucleotide of the A-site codon.

5.2.6 Paromomycin alters the affinity of RF1 for the ribosome

Paromomycin is an aminoglycoside antibotic that deceases translation elongation accuracy by inducing misreading of codons [65, 66]. This misreading occurs when paromomycin binds to h44 of the 16S rRNA and

increases the rate of GTPase activation of EF-TU and the rate of tRNA accommodation, as well as stabilizing the near-cognate codon-recognition complex [67]. In the context of RF1, Youngman and coworkers suggest that paromomycin acts as a competitive inhibitor of RF1 binding to stop-codon programmed ribosomes, while paromomycin and RF1 can bind simultaneously at a near-stop codon – instead resulting in a k_{cat} defect for nascent polypeptide hydrolysis [24]. However, the binding pocket of paromomycin and the location of a stop-codon bound RF1 on the ribosome do not overlap [5, 27, 28, 68]. These seemingly dissimilar behaviors could be explained either through conformational changes of the ribosome which occur or do not occur upon binding, or perhaps more generally, through ordinary or modulated ribosomal dynamics. Here, we investigated the binding of wtRF1-(Cy5) to both stop-codon programmed RC_{UAA}-(Cy3) and near-stop codon programmed RC_{UAU}-(Cy3) in the presence of 5 μM paromomycin.

As seen in the E_{FRET} versus time trajectory shown in Figure 5.13, wtRF1-(Cy5) binds and dissociates very rapidly from RC_{UAA}-(Cy3). We note here that having paromomycin in solution while imaging seemed

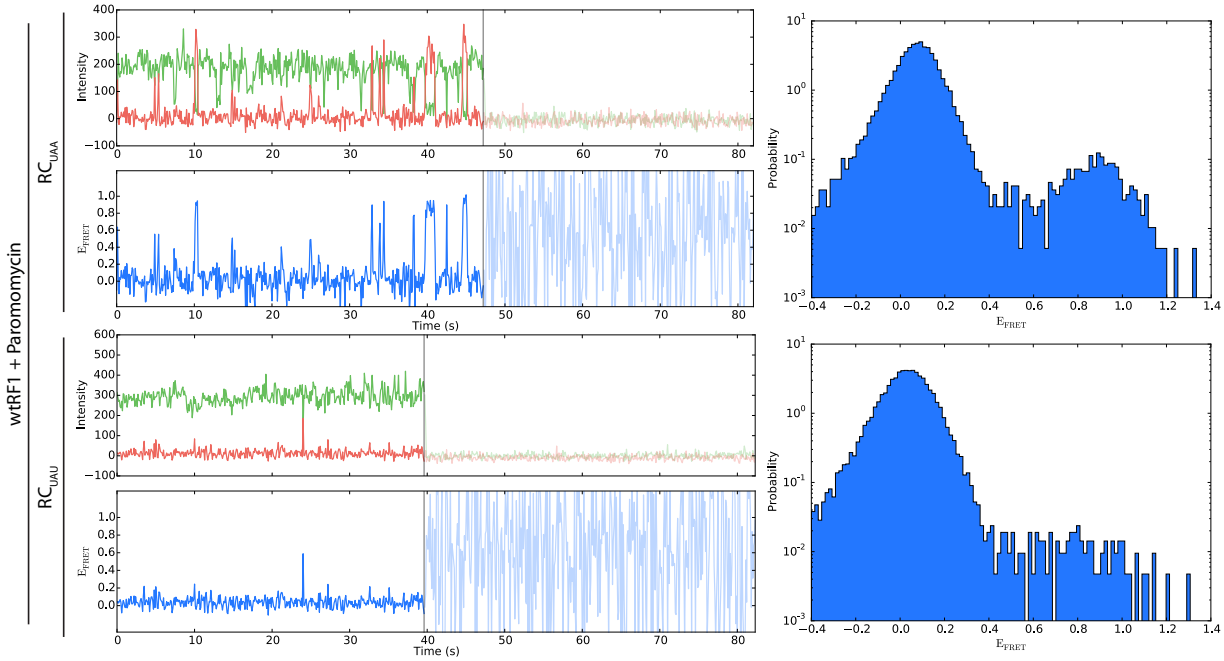

Figure 5.13: smFRET from wtRF1 Binding to Stop-Codon Programmed Ribosomes in the Presence of Paromomycin. (Top) As in Figure 5.9, an example smFRET trajectory and a histogram of n = 83 E_{FRET} versus time trajectories is shown for 10 nM wtRF1-(Cy5) binding to RC_{UAA}-(Cy3) in the presence of 5 μM paromomycin. (Bottom) Similarly, an example smFRET trajectory and a histogram of n = 127 E_{FRET} versus time trajectories is shown for 10 nM wtRF1-(Cy5) binding to RC_{UAU}-(Cy3) in the presence of 5 μM paromomycin.

to induce a substantial number of photophysical events similar to Cy3 blinking. However, this is just a complicating feature, and we do not believe that it influences the observed binding kinetics. Regardless, as seen before, the log-scaled histogram of E_{FRET} versus time trajectories of wtRF1-(Cy5) binding to RC_{UAA}-

(Cy3) in the presence of paromomycin shows a distinct high-E_{FRET} value peak centered at $E_{FRET} \approx 0.95$, which is similar to the previous cases in Section 5.2.5. Again, there is a moderate amount time-averaged data in the middle of the histogram.

These binding kinetics are in contrast to those of wtRF1-(Cy5) with the near-stop codon programmed RC_{UAU}-(Cy3) in the presence of paromomycin. We can use BIASD to quantify apparent, forcibly-Markovian association and dissociation rate constants (k_{on}, and k_{off}, respectively). For RC_{UAA}-(Cy3), this yields $k_{\mathrm{on}} = 0.10 \pm 0.01$ s^{-1}, and $k_{\mathrm{off}} = 3.2 \pm 0.5$ s^{-1}, while for RC_{UAU}-(Cy3) this yields $k_{\mathrm{on}} = 0.07 \pm 0.02$ s^{-1}, and $k_{\mathrm{off}} = 12.7 \pm 2.5$ s^{-1} The apparent association rate is slightly lower for RC_{UAU}-(Cy3) than for RC_{UAA}-(Cy3), as there are many fewer transitions to the high-E_{FRET} state. This high-E_{FRET} state can barely be seen in the histogram of wtRF1-(Cy5) binding to RC_{UAU}-(Cy3) in the presence of paromomycin, because the of the much faster dissociation rate constant, which is also faster than the acquisition rate of the data (10 s^{-1}). Since these experiments have the same concentrations of wtRF1-(Cy5) and paromomycin in solution while imaging took place, the differences originate from the codon in the ribosomal A-site.

5.3 Discussion

While the importance of dynamics to biomolecular function becomes increasingly apparent, the extent to which dynamics contribute to ligand binding and ligand discrimination remains a challenging problem in the biophysics of macromolecular biomolecules. Theoretically, several kJ/mol of energy can be obtained by small perturbations to dynamics ranging in scale from low-frequency collective modes to individual anharmonic atomic motions, and the energy from these changes in fluctuations could then be leveraged for biomolecular function [69]. For instance, it has been demonstrated that the conformational entropy of sub-ns side-chain motions is employed to tune the high-affinity ligand binding interactions that calmodulin makes with its target domains, of which there are over 300 distinct targets, rather than exclusively using enthalpic interactions [70]. Indeed, it seems implausible to achieve this breadth of highly-affinity ligand binding through specific, structural interactions within a single protein. Similarly, during translation the ribosome must interact with release factors and 41 unique isoforms of tRNA that are present in the *E. coli* [71], and these interactions must be high-affinity due to the requirement for rapid translation in the midst of this competition, as well as be specific, so as to avoid the deleterious effects of miscoding and premature termination. It is well known that the ribosome is a dynamic macromolecular machine, which operates by biasing large thermally-driven fluctuations in a cooperative and organized manner [10, 72]. Moreover, in some cases, these processes are regulated by the contributions of conformational dynamics, such as tRNA flexibility influencing codon-anticodon discrimination such as with the Hirsh suppressor tRNA [12, 13, 16], or influencing the GS1 $\rightleftharpoons$ GS2 fluctuations during pretranslocation, which are utilized for efficient translocation [73] (see Chapter 4); similarly, paromomycin alters the dynamics of 16S rRNA residue A1492, which in turn affects the affinity of amino-acyl tRNA-EF-Tu(GTP) ternary complex to the A-site, as well as affects codon-anticodon discrimination by increasing tRNA accommodation at non-cognate codons [5, 16].

The ability of RF1 to function during translation termination similarly depends upon its ability to control ligand binding and ligand discrimination at stop, near-stop, and sense codons. Moreover, calculations of the free energy of the direct structural interactions of RF1 at stop- (UAA) and near-stop (UAU) codons from molecular dynamics simulations [74] do not provide enough free energy to account for the fidelity [75] observed in *in vitro* translation termination experiments [19], and under-estimate this fidelity by a factor of $\sim$200. We hypothesized that this difference might, in part, be the contribution of conformational dynamics that are inherent to the release factor and the ribosome, and so below we discuss the extent to which the afore-mentioned experimental results address this question.

5.3.1 Switch loop dynamics modulate binding affinity

In an effort determine the mechanism through which RF1 dynamics participate in its binding affinity and codon discrimination ability, we investigated the switch loop. Nucleotides 24 and 27 of pre-accomodated tRNA located in the A-site of the ribosome overlap with the residues on the termini of the switch loop (Fig. 5.1). These nucleotides in the D-arm region of tRNA modulate the flexibility of the pre-accommodated tRNA which must transition from the A/T to A/A state by undergoing a distortion that places the aminoacylated 3'-CCA end of the tRNA into the PTC [16]. This suggests that one possible role of the switch loop might be to modulate or impart similar flexibility to RF1 in order to communicate between the DC and PTC. Moreover, we note that previous work has shown that mutations to the switch loop, and to the switch loop and H69 of the 23S rRNA which interacts with ribosome-bound RF1 in X-ray crystallography structures have shown reduced rates of hydrolysis at stop codons [28], and loss of recognition of the UAG stop codon for RF1 [23].

In X-ray crystallography structures of the 'closed' conformation of RF1, the solvent exposed switch loop cannot be located, because it is so dynamic [63, 64]. Similarly, when RF1 is bound to the ribosome (i.e., in the 'open' conformation), several residues within the switch loop have high B-factors as does A1492 of the 16S rRNA – suggesting that these residues are dynamic without exhibiting static enthalpic-based interactions [27, 28]. Additionally, the base stacking of universally conserved residues A1493 and A1913 (23S rRNA) occurs on different faces when the A-site stop codon is UAA or UAG, suggesting that interactions between the switch loop and A1493 of A1913 might not be strict requirements, so much as frozen-out conformations in the end-state presented by the X-ray crystallography structure [27, 28]. Perhaps, rather than it being important for the switch loop to make specific interactions with the ribosome for translation termination, it might be sufficient that the switch loop dock into this binding pocket, and in doing so, utilize the switch loop as a type of entropy sink in which the change in conformational entropy from the disordered solvent-exposed state to the bound state in order to regulate binding affinity and codon discrimination. This idea drove us to attempt to modulate these conformational dynamics to see if that would affect the binding affinity or codon discrimination abilities of RF1.

Rather than attempt to perturb a putative enthalpic interaction made by the switch loop, we chose to alter the dynamics of the switch loop by introducing a mutation that would have minimal direct enthalpic effects upon the switch loop, but that would perhaps disrupt some collective activity present in the switch loop. From the analysis of the mutual information of the primary structure of RF1 from different bacterial genuses, residues G299 and G301 were identified as having a high propensity to coevolve together, and therefore

perhaps conserve some function that does not require direct enthalpic interactions to the side chain since they are glycines (Figure 5.2). In choosing to mutate these glycine residues to alanines, we have introduced the minimal steric perturbation, and no difference in charge state or hydrogen bonding status. With minimal enthalpic changes, the biggest difference is that the region of Ramachandran space readily accessible by the residues is diminished. To investigate if these mutations would consequently alter the dynamics of the switch loop, we utilized molecular dynamics simulations.

From the metadynamics simulation of the GTG and ATA tripeptides corresponding to the wild type and mutant RF1 switch loops, respectively, we determined that the nearest-neighbor contributions of the mutations should produce a difference in the ϕ angles conformations sampled by the intermediate residue 300 (Figures 5.6, and 5.5). However, in the molecular dynamics simulations of the entire wtRF1 and mutRF1 in solution, we found that these residues sampled altered regions of Ramachandran space (Fig.5.5). Since it appears that in the context of the rest of RF1, these residues do not sample the free energy surface of the tripeptides, we interpret these results as a coupling of the dynamics of the mutations to that of the rest of the switch loop. This coupling in turn results in different conformational dynamics of the switch loop for wtRF1 and mutRF1, despite starting the simulation in the same conformation, suggesting that the mutations induce effects on the larger-scale collective behavior of RF1. Finally, the different community structures observed in the network analysis of the wtRF1 and mutRF1 molecular dynamics trajectories (Fig. 5.7), suggests that the mutation perturbs not only the more local dynamics of just the switch loop, but also the dynamics of the entire release factor. Since the different community structures include the core domain of RF1 which contact the mRNA and the 16S rRNA in the decoding center directly, it is possible that the mutations might perturb dynamics that are important for the binding affinity and/or codon discrimination ability of RF1.

From our steady-state smFRET experiments, we see that the switch loop mutation affects the binding affinity of RF1 to stop-codon programmed ribosomal release complexes (Figure 5.9). With 10 nM mutRF1-(Cy5) in solution, there are multiple, transient binding and dissociation events with RC_{UAA}-(Cy3) for a given E_{FRET} versus time trajectory. As expected for an association process, the frequency of these events depends upon the concentration of mutRF1-(Cy5) in solution, but because of the extremely transient nature of some of these dwell-times of mutRF1-(Cy5) on the ribosome as well as the possible non-Markovian behavior, we did not quantify the the binding affinity of mutRF1-(Cy5) to the stop-codon programmed RC_{UAA}-(Cy3). Regardless, the mutRF1-(Cy5) binding affinity is qualitatively very different than that of wtRF1-(Cy5) to RC_{UAA}-(Cy3), which has such a high affinity that it remains statically bound during the course of these and previous smFRET experiments [29]. Additionally, we note that pre-steady state bulk kinetic measurements of wtRF1 to

UAA-programmed ribosomal release complexes had such a high affinity that only an upper-limit for the K_D could be estimated [58, 59]. The markedly different binding affinity of mutRF1-(Cy5) at a stop codon suggests that the dynamics of the switch loop modulate release factor binding. This could occur if the ribosome and/or RF1 are unable to access or stabilize a particular conformation(s) in the absence of wild type switch loop dynamics. From hydroxyl radical probing experiments, it is known that more extensive conformational rearrangements of the rRNA comprising the switch loop binding pocket occur with RF1 binding stop-codon versus near-stop codon programmed ribosomal release complexes [76].

The binding affinity of both mutRF1-(Cy5) and wtRF1-(Cy5) is further reduced at the near-stop codon programmed RC_{UAU}-(Cy3) (Fig. 5.12). Occasional binding events could be observed for wtRF1-(Cy5) binding, however they were even more transient then those observed for mutRF1-(Cy5) binding to RC_{UAA}-(Cy3). Therefore the altered dynamics of the switch loop destabilize the binding affinity of mutRF1 to RC_{UAA}-(Cy3) to a lesser amount than a near-stop codon does for wtRF1-(Cy5) binding to RC_{UAU}-(Cy3). Moreover, no binding events were observed for mutRF1-(Cy5) to RC_{UAU}-(Cy3). It is clear that the mutation increased the $\mathrm{K_D}$ for mutRF1-(Cy5) binding to RC_{UAU}-(Cy3) far beyond what is accessible with smFRET on our TIRF microscope. However, for wtRF1-(Cy5) binding to RC_{UAU}-(Cy3), we observed $\mathrm{K_D} \approx 2\ \mu$M, and so we would estimate the lower-bound for mutRF1-(Cy5) binding to RC_{UAU}-(Cy3) as at least 10 times greater. Regardless, the dynamics of the switch loop can also modulate the binding affinity of RF1 to near-stop codon programmed ribosomal release complexes. Unfortunately, from these smFRET experiments, because we can't quantify the essentially static wtRF1-(Cy5) binding to RC_{UAA}-(Cy3), it is unclear whether the magnitude of the effects depend upon the codon in the A-site, and whether the contributions of the codon and the switch loop dynamics are independent or coupled.

Similar to result that the dynamics of the switch loop affect binding at both RC_{UAA}-(Cy3) and RC_{UAU}-(Cy3), we observed that the aminoglycoside antibiotic paromomycin affects the binding affinity of RF1 to both stop and near-stop codon programmed ribosomal release complexes (Fig. 5.13). With RC_{UAA}-(Cy3), repeated, transient wtRF1-(Cy5) binding and dissociation events could be observed, and even at RC_{UAU}-(Cy3), the binding and dissociation of wtRF1-(Cy5) was still readily observable – however, it was slightly less frequent. Using the association and dissociation rate constants we calculated assuming Markovian behavior, the $\mathrm{K_D}$ = 0.32 μM for RC_{UAA}-(Cy3), and $\mathrm{K_D}$ = 1.8 μM for RC_{UAU}-(Cy3), which is only about a 6-fold difference. This relative small difference is markedly different than the those seen for mutRF1-(Cy5) binding to RC_{UAA}-(Cy3) and RC_{UAU}-(Cy5) in the absence of paromomycin where the $\mathrm{K_D}$ changes by, at least, three orders of magnitude. Since we were still able to see binding and dissociation of wtRF1-(Cy5) at both RC_{UAA}-(Cy3) and

RC_{UAU}-(Cy3) in presence of paromomycin, unlike the case of mutRF1-(Cy5) in the absence of paromomycin, this suggests that perturbations induced in the DC by paromomycin and altered switch loops dynamics affect the binding affinity of RF1 through different mechanisms.

Paromomycin binds h44 of the 16S rRNA, where it changes the conformation and dynamics of A1492 and A1493 in DC, and pushes them toward the location of the minor groove of the mRNA/tRNA codon-anticodon [16]. Interestingly, in an X-ray crystallography structure of paromomycin bound to the 30S ribosomal subunit, A1493 can be found completely extended out of the aminoglycoside binding pocket in h44 [25] into a conformation believed sterically incompatible with the switch loop of RF1 bound at a stop-codon [27], and this was provided as structural evidence for the competitive inhibition by paromomycin of termination by RF1 at stop-codons [24]. However, the solution NMR structure of paromomycin bound to h44, which lacks crystal packing and cryo-induced artifacts, shows A1493 in a well-defined location that is still within the aminoglycoside binding pocket where it base pairs with A1408, and would not sterically occlude the switch loop [5]. Thus, when paromomycin is bound, A1493 probably does not sterically occlude RF1 binding, and it is possible that both RF1 and paromomcyin can both bind the ribosome, as is the case with RF1 bound at near-stop codons [24]. More likely, the altered location of A1493 and new dynamics of A1492 in the paromomycin-bound state [5] inhibit some conformational rearrangement of the ribosome and/or RF1 that is important for the high-affinity binding of RF1. This is consistent with our observation that paromomycin does not completely preclude wtRF1 binding at near-stop codons as can occur with only the altered switch loop dynamics of mutRF1. Such distinct behavior suggests that paromomycin and the conformational dynamics of the switch loop might function upon two independent molecular requirements of translation termination. This is similar to the observation that rearrangements of the decoding center and also of the tRNA ternary complex present two separate molecular requirements for tRNA selection to occur [77]. For tRNA selection, while paromomycin can overcome both requirements to induce tRNA selection in presence of DC mutations (A1492G or A1493G) or even to select non-cognate tRNA, the Hirsh supressor tRNA can only compensate for the non-cognate codon-anticodon interactions and cannot compensate for DC mutations [77]. Thus the dynamics of the DC and the tRNA contribute to differently to two conformational changes that are both important for tRNA selection. Analogously, rather than both inhibiting one common conformational change that governs RF1 binding affinity, the paromomycin-induced DC dynamics and the mutRF1 switch loop dynamics could contribute to different step, both of which are both required for the regulating the binding affinity of RF1 to the ribosome.

5.3.2 Tight-binding of RF1 is not necessary for codon discrimination

In our smFRET experiments, wtRF1-(Cy5) remained stably bound to RC_{UAA}-(Cy3) (Fig. 5.9) [29]. The low $\mathrm{K_D}$ of wtRF1 binding to stop codon programmed ribosomal release complexes originates in part from the slow dissociation rate of RF1 [58, 59, 78]. These RF1-ribosome complexes demonstrate such a tight binding that a GTPase, the class II release factor, RF3, is responsible for catalyzing the dissociation of RF1 [29, 49]. This is surprising, in part, because our results concerning the dynamics of the switch loop (Section 5.3.1) suggest small mutations in the switch loop can accelerate the dissociation rate by about two orders of magnitude, which is almost the same amount as RF3 achieves. Additionally, the tight-binding of RF1 to the ribosome is not required for accurate codon discrimination as discussed above, because mutRF1 catalyzes hydrolysis at stop-codons, but not near-stop codons (Fig. 5.3). Therefore, we find the tight-binding of RF1 at stop-codons is not expected.

Interestingly, in an *in vivo* screen, Diago-Navarro and coworkers identified several RF1 mutations in the switch loop that diminished recognition of the UAG stop codon; cells containing these prfA, which encodes RF1, mutants were more susceptible to the mRNA interferase, RelE [23]. RelE is the toxin of the relBE toxin-antitoxin system, and is known to cleave mRNA in the A-site with a sequence preference for UAA and UAG – the RF1 stop codons [47]. This mRNA cleavage generates a truncated message, which can serve as a substrate for tmRNA [79], and relBE has been implicated in the dormant persister phenotype [80]. Since those prfA mutations also render the cells susceptible to other toxin-antitoxin systems [23], perhaps, the role of the tight-binding of RF1 is to protect the mRNA in the A-site until ribosome recycling can occur, in order to avoid mRNA degradation at what might appear to be a stalled ribosome [81]. Additionally, tight-biding of RF1 after hydrolysis blocks the A-site from beginning translation again in a translation initiation-independent manner by misreading of the stop codon by a near-cognate tRNA. Such a mechanism would allow for novel regulation of translation in which proteins could begin at 'stop' codons. For instance, a downstream antitoxin gene could be used to relieve selective pressure from an over-expressed toxin in an *in vivo* screen for release factor mutations that do not exhibit this tight-binding property. Additionally, inducing transcription of miscoding tRNA could be used to out-compete weak-binding release factors, and induce translation of a gene of interest.

5.3.3 Switch loop dynamics modulate codon discrimination

Regarding the codon discrimination ability of RF1, the ability to discriminate stop versus near-stop codons can be characterized using the Michaelis-Menten formulation of enzyme kinetics as

$$A = \left(\frac{k_{\mathrm{cat}}}{K_M}\right)_{\mathrm{s}} / \left(\frac{k_{\mathrm{cat}}}{K_M}\right)_{\mathrm{ns}} ; \text{ where } K_M = \frac{k_{-1} + k_{\mathrm{cat}}}{k_1} = K_D + \frac{k_{\mathrm{cat}}}{k_1}, \tag{5.7}$$

A is the accuracy of recognizing a stop (s) codon versus a near-stop (ns) codon, k_1 and k_{-1} are association and dissociation rate constants, k_{cat} is the catalysis rate constant, $\mathrm{K_M}$ is the Michaelis constant, and $\mathrm{K_D}$ is the dissociation constant [82]. The changes in accuracy dependent upon codon identity have been extensively measured by Freistroffer and coworkers using this formalism [19]. In our smFRET assay, we did observe the effect that the switch loop dynamics had upon the affinity of RF1 for stop and near-stop codons, and Equation 5.7 demonstrates how these binding affinities affect the accuracy of codon discrimination. However, since we could not observe release of the nascent polypeptide chain in our smFRET assay, the net effect upon the accuracy of codon discrimination could not be quantified, because we did not have a measurements of k_{cat}. Fortunately, in our eTLC dipeptide release assay, we observed the release of the nascent polypeptide chain by wtRF1 and mutRF1 at stop and near-stop codon programmed ribosomal release complexes. Like wtRF1, a saturating amount of mutRF1 was able to catalyze release at a stop codon to completion within 5 sec; so to within the minimal time resolution of our manual-mixing, there did not seem to be much, if any, of a defect in k_{cat} at a stop codon. However, the altered switch loop dynamics of mutRF1 seem to have resulted in a hyper-accurate release factor, which releases a negligible amount of dipeptide from a near-stop codon ribosomal release factor when compared to the wtRF1. While the pronounced defect in binding affinity of mutRF1-(Cy5) to $\mathrm{RC_{UAU}}$-(Cy3) contributes significantly to this increase in accuracy, we did not measure k_{cat} for these reactions, so the absolute change in accuracy is unclear. Regardless, since mutRF1 is capable of catalyzing release at a stop codon, yet it releases much less than wtRF1 at a near-stop codon, it is clear that the dynamics of the switch loop modulate RF1 codon discrimination. This is consistent with *in vivo* translation experiments, which show that mutation of the switch loop can eliminate RF1 recognition of the UAG stop codon recognition [23], and with the network analysis of the molecular dynamics simulations of wtRF1 and mutRF1, which suggested that the altered switch loop dynamics of mutRF1 would influence codon recognition (Fig. 5.7).

In vivo, the ability of RF1 to discriminate between stop and near-stop or sense codons is complicated by

a tradeoff between accuracy and speed, because in order to achieving the maximal amount of discrimination between stop and near-stop codons, the rate of the termination reaction must approach zero [82, 83]. Practically, RF1 must interrogate the A-site codon at a certain rate in order to compete with all the ternary complexes in the cell, and presumably, the balance struck is just right so that this process occurs rapidly enough without diminishing the accuracy of termination to a deleterious level. Therefore, being able to optimize this ratio is important for regulating mistakes while maintaining competitiveness. Since, in our *in vitro* reconstituted dipeptide release assay, mutRF1 is more accurate than wtRF1, despite *E. coli* having opted for the wild type switch loop *in vivo*, our data here suggests that the dynamics of the switch loop play a large part in striking this balance.

5.3.4 RF1 binding and codon discrimination is a multiple step process

Considering the bimolecular association reaction of RF1 and the ribosome, the second-order rate constant for this event is determined primarily by diffusion of RF1 to the ribosomal A-site. Assuming that the G299A, G301A mutations negligibly change the diffusion constant of RF1, the association rate constants for RF1 to both RC_{UAA} and RC_{UAU} should be the same. Assuming that the ribosome does not intrinsically signal the presence of a stop-codon in the A-site, this is because if RF1 is to determine whether the codon in the A-site is a stop codon or not, it must first have at least partially bound the A-site. However, in the E_{FRET} versus time trajectories of wtRF1-(Cy5) and mutRF1-(Cy5) binding to RC_{UAA}-(Cy3) and RC_{UAU}-(Cy3) (Figs. 5.9, and 5.12, respectively) the apparent association rate constants are extremely varied. Notably, mutRF1-(Cy5) should encounter the ribosomal A-site of RC_{UAA}-(Cy3) at the same rate as for RC_{UAU}-(Cy3), however in the former case, binding occurs with an observed rate constant that is only about a factor of 3 less than the estimated diffusion-limited rate constant, while in the latter case, no binding events could be observed at all. Since mutRF1-(Cy5) must still be sampling the A-site, this is evidence of an encounter complex in the binding reaction (see Ref. 84 or Chapter 1 for a review). This encounter complex is too short lived and/or positioned such that the E_{FRET} is lower to register in the E_{FRET} versus time trajectories. Given the 100 ms acquisition period of this smFRET data, we estimate that the upper-bound of the lifetime of the encounter complex is on the order of $\sim$ 1 ms. Additional evidence for the encounter complex comes from the observation that wtRF1-(Cy5) associates with RC_{UAU}-(Cy3) less frequently than mutRF1-(Cy5) associates with RC_{UAA}-(Cy3), even when accounting for possible missed events. Ultimately, the encounter complex state seems to be an on-pathway intermediate between the unbound state and a state in which RF1 interrogates the codon.

After the encounter complex, there is considerable evidence for a conformational rearrangement of the

ribosome and/or RF1 upon stop codon recognition. From our smFRET experiments discussed above, we note that the differential effects of altered switch loop dynamics and paromomycin suggest that they function to affect RF1 binding affinity and codon recognition through different molecular requirements of termination. Additionally, the dwell-times of mutRF1-(Cy5) bound to RC_{UAA}-(Cy3) are not distributed according to a single-exponential distribution (Fig. 5.11), and this seemingly non-Markovian behavior is easily explained with the Markovian behavior of two 'bound' states, which can interconvert. The analytical solution of this probability distribution function of these 'bound' dwell-times can be determined using the Laplace transform method described in Section 3.4. However, the non-Markovian behavior could also be dynamic disorder within the dissociation rate constant, or static disorder in that the RF1 in solution might be heterogenous. Though similarly, Trappl and coworkers observed biphasic pre-steady state kinetics of RF1 binding to stop-codon programmed ribosomal release complexes [59], which could also be explained by an on-pathway intermediate. These behaviors are most likely related the discrete conformational changes observed by hydroxyl radical probing of the rRNA, which occur upon stop-codon recognition, but not near-stop codon recognition [76]. Moreover, small-angle X-ray scatting studies of release factors in solution are consistent with an equilibrium between compact and extended conformations [85, 86], and so there is significant speculation about the role that a conformational change between the open [26–28] and closed forms [63, 64] of RF1 might play in signaling proper codon recognition in the DC to the PTC, whereupon hydrolysis of the nascent polypeptide chain would occur [87]. Furthermore, recent bulk, transition metal FRET experiments have demonstrated that RF1 is in a more compact conformation in solution, and that it opens upon codon recognition at the ribosome [88]. Additional support for this equilibrium comes from smFRET experiments of dual-labeled RF1 in solution that primarily yield an E_{FRET} state that is consistent with a compact conformation, but do demonstrate transient fluctuations to a low-E_{FRET} state that might represent the 'open' state (Appendix A).

Through consideration of the evidence for an encounter complex in RF1 binding, as well as conformational change of RF1 and/or the ribosome that occurs after codon recognition, we conclude that RF1 binding and codon discrimination is a multistep process. The most parsimonious mechanism consistent with our data is that where initial binding occurs through an encounter complex; the encounter complex then transitions to a state from which codon recognition occurs; after codon recognition, a conformational change results in a fully accommodated RF1 from which hydrolysis can be catalyzed (Fig. 5.14). This proposed mechanism is analogous to the first half of that of tRNA selection during translation elongation; there initial binding, which is not seen by smFRET, is followed by codon recognition, and then GTPase activation, from which GTP hy-

drolysis occurs [16, 84]. Such ligand binding mechanism with multiple intermediate states allow for efficient ligand binding, and also, by extension, opportunities to regulate ligand discrimination [84]. Additionally, they provide opportunities for error correction mechanisms such kinetic proofreading [20, 21].

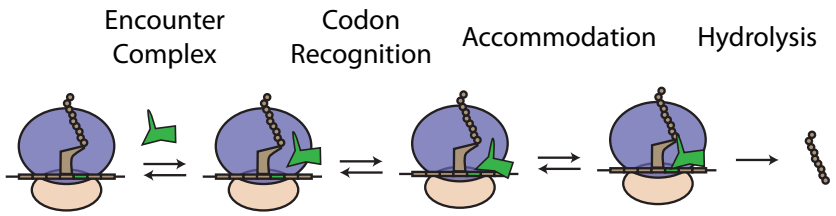

Figure 5.14: Proposed Mechanism of RF1 binding to Release Complexes. Initially weak binding of RF1 (green) to the ribosome (purple, tan) forms an encounter complex which registers negligible E_{FRET} in our smFRET experiments. Once codon recognition by RF1 occurs, this can be followed by a more strongly bound accommodated state.

Many other biological systems aside from the ribosome utilize such multistep mechanisms to for ligand binding and discrimination. For instance, DNA polymerase selects the correct dNTP to incorporate in a two-step reaction in which initial selection utilizes the ΔG difference from the different dNTPs hydrogen bonding to the template strand, and the second step consists of a conformational change that depends of correct Watson-Crick base-pairing geometry [89]. Complicated sequence searches, such as those performed by RNAP polymerase searching for a promoter sequence [90], or the CRISPR-Cas system searching for DNA complement to the guide RNA [91], also utilize multiple step binding mechanisms in order to facilitate the search process. Protein-protein interactions, such as those between enzyme I and HPr of the bacterial phosphotransferase system, often occur through on-pathway encounter complexes as these reduce the dimensionality of the search process, and speed up stereospecific complex formation [92]. Similarly, ligand binding by the aptamer domain of the guanine-sensing riboswitch occurs through a multiple step mechanism involving an encounter complex [93]. Even enzymes utilize multiple conformational changes for binding and catalysis of their substrates [8]. Given the ubiquity of such multistep approaches in nature, it is not surprising to find evidence for the involvement of multiple steps during RF1 binding and codon discrimination during translation termination.

5.4 Methods

RF1 Mutation, Expression, Labeling, and Purification The G299A, G301A mutations in mutRF1 and 167C-mutRF1 were introduced through site-directed mutagenesis (G896C, and G902C) of the previous constructs wtRF1 and 167C-wtRF1, respectively [29]. wtRF1, mutRF1, 167C-wtRF1, and 167C-mutRF1 were

all co-overexpressed with the methyltransferase encoded by prmC in BL21(DE3) *E. coli* cells, and purified using affinity chromatography with Ni-NTA derivatized agarose beads and tobacco etch virus protease, as previously described [29, 94]. 167C-wtRF1 and 167C-mutRF1 were then labeled with maleimide derivatized Cy5 (G.E. Lifesciences), and purified to 100% labeling using gel filtration and hydrophobic interaction chromatography as previously described [29, 94]. Concentrations were quantified using the Bradford protein assay reagent and absorption spectroscopy at λ = 650 nm.

Ribosomal Release Complex Formation RC_{UAA} and RC_{UAA}-(Cy3) were previously described [29]. Briefly, complexes were enzymatically prepared in tris polymix buffer containing 50 mM tris-acetate (pH=7.5), 100 mM KCl, 5 mM ammonium acetate, 0.5 mM calcium acetate, 0.1 mM EDTA, 10 mM 2-mercaptoethanol, 5 mM putrescine, 1 mM spermidine, and 5 mM magnesium acetate using formylated $\left[^{35}S\right]$-Met-tRNA^{fMet} and Phe-tRNA^{Phe}, or formylated Met-tRNA^{fMet} and Phe-tRNA^{Phe} labeled with an NHS ester derivatized Cy5 at the acp^{3}U47 position, respectively. The mRNA message was *in vitro* transcribed from a pUC119 vector containing a gene derived from the gene encoding gene product 32 from the T4 bacteriophage, such that the gene to be translated was AUG-UUU-UAA. Complexes were then purified from the excess translation factors with sucrose density gradient ultracentrifugation. The process for RC_{UAU} and RC_{UAU}-(Cy3) formation was equivalent, but used an mRNA message transcribed from a vector-derived template that had been mutagenized so the the gene to be translated was AUG-UUU-UAU. For RC_{UAA}-(Cy3) and RC_{UAU}-(Cy3), the mRNA was hybridized to complementary a biotinylated DNA oligo [29].

Mutual Information Calculation Bacterial RF1 sequences were obtained from the NCBI ref-seq database [37]. Over-represented sequences were culled using custom Python scripts. Sequences were pairwise aligned using Clustal Omega [38]. The mutual information of these sequences was calculated using Equation 5.1 with custom Python scripts.

Molecular Dynamics Simulations Molecular dynamics simulations were performed as described in Section 5.2.4 using Desmond [50] and the OPLS-AA force field [51] with TIP3P waters. The Ewald sum cutoff was 9 Å , and the RESPA integrator was 2 fs. The metadynamics simulations were 10 ns long with a height of 0.03 kcal mol^{-1} for the height of the repulsive Gaussians, which were added every 0.09 ps. Two-dimensional variational hidden Markov modeling was performed using custom Python scripts (Appendix D).

In Vitro Translation Assay Details concerning the radioactivity-based translation assay were previously described [24, 31]. Briefly, 200 nM release complexes containing formyl-$\left[^{35}S\right]$-Met-Phe dipeptide incubated with 25 μM RF1 at room temperature in the tris polymix buffer described above, and then quenched with 1:1 v/v 25% formic acid. Released dipeptide was then separated by eTLC as previously described [30]. eTLC plates were then exposed with PhosphorImager screen (GE Lifesciences) overnight, and then the screen was scanned with a Typhoon FLA 7000 PhosphorImager (GE Lifesciences), and quantified using ImageJ [95].

TIRF Microscopy Release complexes labeled with Cy3 were immobilized within microfluidic flow cells, and imaged with a TIRF microscope as previously described [29]. Briefly, samples were illuminated with a 532 nm diod-pumped solid-state laser (gem532, Laser Quantum) with prism-based TIRF on a Nikon Ti-U inverted microscope. Fluorescence was collected through a Nikon 60x PlanApo objective, and imaged through a Photometrics DV2 wavelength splitter onto a Andor iXon3 897E electron-multiplying charge-coupled-device camera at 10 Hz. Prior to imaging, flow-cells containing tethered release complexes were incubated with 10 nM Cy5-labeled RF1 in tris-polymix buffer containing 15 mM magnesium acetate, 1% (w/v) β-D-glucose, 300 mg mL^{-1} glucose oxidase (Sigma-Aldrich), 40 mg mL^{-1} catalase (Sigma-Aldrich), 1mM 1,3,5,7-cyclooctatetraene (Sigma- Aldrich), and 1 mM *p*-nitrobenzyl alcohol (Fluka).

5.5 References

1. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids. *Nature* **171,** 737–738 (1953).
2. Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147,** 1462–1465 (1965).
3. Westhof, E., Yusupov, M. & Yusupova, G. Recognition of Watson-Crick base pairs: constraints and limits due to geometric selection and tautomerism. *F1000Prime Rep.* **6,** 19 (2014).
4. Keedy, D. a. *et al.* Crystal cryocooling distorts conformational heterogeneity in a model michaelis complex of DHFR. *Structure* **22,** 899–910 (2014).
5. Lynch, S. R., Gonzalez, R. L. & Puglisi, J. D. Comparison of X-ray crystal structure of the 30S subunit-antibiotic complex with NMR structure of decoding site oligonucleotide-paromomycin complex. *Structure* **11,** 43–53 (2003).
6. Frauenfelder, H., Parak, F. & Young, R. D. Conformational substates in proteins. *Annu. Rev. Biophys. Biophys. Chem.* **17,** 451–479 (1988).
7. Jardetzky, O. Protein dynamics and conformational transitions in allosteric proteins. *Prog. Biophys. Mol. Biol.* **65,** 171–219 (1996).

8. Hammes, G. G. Current Topics Multiple Conformational Changes in Enzyme Catalysis. *Biochemistry* **41,** 8221–8228 (2002).

9. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5,** 789–96 (2009).

10. Frank, J. & Gonzalez, R. L. Structure and dynamics of a processive Brownian motor: the translating ribosome. *Annu. Rev. Biochem.* **79,** 381–412 (2010).

11. Wand, A. J. The dark energy of proteins comes to light: conformational entropy and its role in protein function revealed by NMR relaxation. *Curr. Opin. Struct. Biol.* **23,** 75–81 (2013).

12. Hirsh, D. & Gold, L. Translation of the UGA triplet in vitro by tryptophan transfer RNA's. *J. Mol. Biol.* **58,** 459–468 (1971).

13. Hirsh, D. Tryptophan Transfer RNA as the UGA Suppressor. *J. Mol. Biol.* **58,** 439–458 (1971).

14. Schultz, D. W. & Yarus, M. tRNA structure and ribosomal function. I. tRNA nucleotide 27-43 mutations enhance first position wobble. *J Mol Biol* **235,** 1381–1394 (1994).

15. Schultz, D. W. & Yarus, M. tRNA Structure and Ribosomal Function II. Interaction Between Anticodon Helix and other tRNA Mutations. *J. Mol. Biol.* **235,** 1395–1405 (1994).

16. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74,** 129–77 (2005).

17. Dunkle, J. A. & Cate, J. H. Ribosome Structure and Dynamics During Translocation and Termination. *Annu. Rev. Biophys.* **39,** 227–244 (2010).

18. Zhou, J., Korostelev, A., Lancaster, L. & Noller, H. F. Crystal structures of 70S ribosomes bound to release factors RF1, RF2 and RF3. *Curr. Opin. Struct. Biol.* **22,** 733–42 (2012).

19. Freistroffer, D. V., Kwiatkowski, M., Buckingham, R. H. & Ehrenberg, M. The accuracy of codon recognition by polypeptide release factors. *Proc. Natl. Acad. Sci.* **97,** 2046–51 (2000).

20. Hopfield, J. J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proc. Natl. Acad. Sci.* **71,** 4135–4139 (1974).

21. Ninio, J. Kinetic amplification of enzyme discrimination. *Biochimie* **57,** 587–595 (1975).

22. Uno, M., Ito, K. & Nakamura, Y. Polypeptide release at sense and noncognate stop codons by localized charge-exchange alterations in translational release factors. *Proc. Natl. Acad. Sci.* **99,** 1819–1824 (2002).

23. Diago-Navarro, E., Mora, L., Buckingham, R. H., Díaz-Orejas, R. & Lemonnier, M. Novel Escherichia coli RF1 mutants with decreased translation termination activity and increased sensitivity to the cytotoxic effect of the bacterial toxins Kid and RelE. *Mol. Microbiol.* **71,** 66–78 (2009).

24. Youngman, E. M., He, S. L., Nikstad, L. J. & Green, R. Stop Codon Recognition by Release Factors Induces Structural Rearrangement of the Ribosomal Decoding Center that Is Productive for Peptide Release. *Mol. Cell* **28,** 533–543 (2007).

25. Ogle, J. M., Brodersen, D. E., Clemons, W. M., Tarry, M. J., Carter, A. P. & Ramakrishnan, V. Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science* **292,** 897–902 (2001).

26. Petry, S. *et al.* Crystal structures of the ribosome in complex with release factors RF1 and RF2 bound to a cognate stop codon. *Cell* **123,** 1255–1266 (2005).

27. Laurberg, M., Asahara, H., Korostelev, A., Zhu, J., Trakhanov, S. & Noller, H. F. Structural basis for translation termination on the 70S ribosome. *Nature* **454,** 852–7 (2008).

28. Korostelev, A., Zhu, J., Asahara, H. & Noller, H. F. Recognition of the amber UAG stop codon by release factor RF1. *EMBO J.* **29,** 2577–85 (2010).

29. Sternberg, S. H., Fei, J., Prywes, N., McGrath, K. a. & Gonzalez, R. L. Translation factors direct intrinsic ribosome dynamics during translation termination and ribosome recycling. *Nat. Struct. Mol. Biol.* **16,** 861–868 (2009).

30. Youngman, E. M., Brunelle, J. L., Kochaniak, A. B. & Green, R. The Active Site of the Ribosome Is Composed of Two Layers of Conserved Nucleotides with Distinct Roles in Peptide Bond Formation and Peptide Release. *Cell* **117,** 589–599 (2004).

31. Englander, M. T. *et al.* The ribosome can discriminate the chirality of amino acids within its peptidyl-transferase center. *Proc. Natl. Acad. Sci.* **112,** 6038–6043 (2015).

32. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74,** 129–177 (2005).

33. Schmeing, T. M. *et al.* The Crystal Structure of the Ribosome Bound to EF-Tu and Aminoacyl-tRNA. *Science* **326,** 688–694 (2009).

34. Humphrey, W., Dalke, A. & Schulten, K. {VMD} – {V}isual {M}olecular {D}ynamics. *J. Mol. Graph.* **14,** 33–38 (1996).

35. Sandler, I., Zigdon, N., Levy, E. & Aharoni, A. The functional importance of co-evolving residues in proteins. *Cell. Mol. Life Sci.* **71,** 673–682 (2014).

36. Dickson, R. J. & Gloor, G. B. in *Homing Endonucleases* (ed Edgell, D. R.) 223–243 (Humana Press, Totowa, NJ, 2014).

37. Pruitt, K. D., Tatusova, T. A. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** 61–65 (2007).

38. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539–539 (2014).

39. Ito, K., Uno, M. & Nakamura, Y. A tripeptide 'anticodon' deciphers stop codons in messenger RNA. *Nature* **403,** 680–684 (2000).

40. Frolova, L. Y. *et al.* Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis. *RNA* **5,** 1014–20 (1999).

41. Dinçbas-Renqvist, V., Engström, Å., Mora, L., Heurgué-Hamard, V., Buckingham, R. & Ehrenberg, M. A post-translational modification in the GGQ motif of RF2 from Escherichia coli stimulates termination of translation. *EMBO J.* **19,** 6900–6907 (2000).

42. Heurgué-Hamard, V., Champ, S., Engström, Å., Ehrenberg, M. & Buckingham, R. H. The hemK gene in Escherichia coli encodes the N5-glutamine methyltransferase that modifies peptide release factors. *EMBO J.* **21,** 769–778 (2002).

43. Nakahigashi, K. *et al.* HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and hemK knockout induces defects in translational termination. *Proc. Natl. Acad. Sci.* **99,** 1473–1478 (2002).

44. Pallesen, J. *et al.* Cryo-EM visualization of the ribosome in termination complex with apo-RF3 and RF1. *Elife* **2,** e00411 (2013).

45. Klaholz, B. P. Molecular recognition and catalysis in translation termination complexes. *Trends Biochem. Sci.* **36,** 282–292 (2011).

46. Indrisiunaite, G., Pavlov, M. Y., Heurgué-Hamard, V. & Ehrenberg, M. On the pH Dependence of Class-1 RF-Dependent Termination of mRNA Translation. *J. Mol. Biol.* **427,** 1848–1860 (2015).

47. Pedersen, K., Zavialov, A. V., Pavlov, M. Y., Elf, J., Gerdes, K. & Ehrenberg, M. The bacterial toxin RelE displays codon-specific cleavage of mRNAs in the ribosomal A site. *Cell* **112,** 131–140 (2003).

48. Fei, J. *et al. A highly purified, fluorescently labeled in vitro translation system for single-molecule studies of protein synthesis.* 1st ed. **10,** 221–59 (Elsevier Inc., 2010).

49. Freistroffer, D. V., Pavlov, M. Y., MacDougall, J., Buckingham, R. H. & Ehrenberg, M. Release factor RF3 in E.coli accelerates the dissociation of release factors RF1 and RF2 from the ribosome in a GTP-dependent manner. *EMBO J.* **16,** 4126–33 (1997).

50. Bowers, K. J. *et al.* Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. 2006 ACM/IEEE Conf. Supercomput. - SC '06,* 84 (2006).

51. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and Testing of the OLPS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118,** 11225–11236 (1996).

52. Laio, A. & Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports Prog. Phys.* **71,** 126601 (2008).

53. Sethi, A., Eargle, J., Black, A. A. & Luthey-Schulten, Z. Dynamical networks in tRNA:protein complexes. *Proc. Natl. Acad. Sci.* **106,** 6620–6625 (2009).

54. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99,** 7821–7826 (2002).

55. Chennubhotla, C. & Bahar, I. Markov propagation of allosteric effects in biomolecular systems: application to GroEL{\textendash}GroES. *Mol. Syst. Biol.* **2** (2006).

56. Wilson, K. S., Ito, K., Noller, H. F. & Nakamura, Y. Functional sites of interaction between release factor RF1 and the ribosome. *Nat. Struct. Biol.* **7,** 866–870 (2000).

57. McQuarrie, D. A. Kinetics of Small Systems. I. *J. Chem. Phys.* **38,** 433 (1963).

58. Hetrick, B., Lee, K. & Joseph, S. Kinetics of Stop Codon Recognition by Release Factor 1. *Biochemistry* **48,** 11178–11184 (2009).

59. Trappl, K., Mathew, M. a. & Joseph, S. Thermodynamic and kinetic insights into stop codon recognition by release factor 1. *PLoS One* **9,** e94058 (2014).

60. Erickson, H. P. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biol. Proced. Online* **11,** 32–51 (2009).

61. Schuwirth, B. S. *et al.* Structures of the Bacterial Ribosome at 3.5 A Resolution. *Science* **310,** 827–834 (2005).

62. Berg, H. C. *Random Walks in Biology* 2nd, 17–36 (Princeton University Press, Princeton, NJ, 1993).

63. Shin, D. H., Brandsen, J., Jancarik, J., Yokota, H., Kim, R. & Kim, S. H. Structural analyses of peptide release factor 1 from Thermotoga maritima reveal domain flexibility required for its interaction with the ribosome. *J. Mol. Biol.* **341,** 227–239 (2004).

64. Graille, M. *et al.* Molecular basis for bacterial class I release factor methylation by PrmC. *Mol. Cell* **20,** 917–927 (2005).

65. Davies, J., Gorini, L. & Davis, B. D. Misreading of RNA codewords induced by aminoglycoside antibiotics. *Mol. Pharmacol.* **1,** 93–106 (1965).

66. Pestka, S. Inhibitors of ribosome functions. *Annu. Rev. Microbiol.* **25,** 487–562 (1971).

67. Pape, T., Wintermeyer, W. & Rodnina, M. V. Conformational switch in the decoding region of 16S rRNA during aminoacyl-tRNA selection on the ribosome. *Nat. Struct. Biol.* **7,** 104–107 (2000).

68. Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-warren, R. J., Wimberly, B. T. & Ramakrishnan, V. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. **407** (2000).

69. Cooper, A. & Dryden, D. T. F. Allostery without conformational change - A plausible model. *Eur. Biophys. J.* **11,** 103–109 (1984).

70. Frederick, K. K., Marlow, M. S., Valentine, K. G. & Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **448,** 325–329 (2007).

71. Laursen, B. & Sørensen, H. Initiation of protein synthesis in bacteria. *Microbiol. ...* **236,** 747–771 (2005).

72. Ning, W., Fei, J. & Gonzalez, R. L. The ribosome uses cooperative conformational changes to maximize and regulate the efficiency of translation. *Proc. Natl. Acad. Sci.* **111,** 12073–8 (2014).

73. Fei, J., Richard, A. C., Bronson, J. E. & Gonzalez, R. L. Transfer RNA-mediated regulation of ribosome dynamics during protein synthesis. *Nat. Struct. Mol. Biol.* **18,** 1043–51 (2011).

74. Sund, J., Andér, M. & Aqvist, J. Principles of stop-codon reading on the ribosome. *Nature* **465,** 947–950 (2010).

75. Fersht, A. R. *Structure and mechanism in protein science. A guide to enzyme catalysis and protein folding.* 293–400 (W.H. Freeman and Co., New York, 1999).

76. He, S. L. & Green, R. Visualization of codon-dependent conformational rearrangements during translation termination. *Nat. Struct. Mol. Biol.* **17,** 465–470 (2010).

77. Cochella, L., Brunelle, J. L. & Green, R. Mutational analysis reveals two independent molecular requirements during transfer RNA selection on the ribosome. *Nat Struct Mol Biol.* **14,** 30–36 (2007).

78. Koutmou, K. S., McDonald, M. E., Brunelle, J. L. & Green, R. RF3:GTP promotes rapid dissociation of the class 1 termination factor. *RNA* **20,** 609–620 (2014).

79. Janssen, B. D. & Hayes, C. S. *The tmRNA ribosome-rescue system.* 1st ed., 151–91 (Elsevier Inc., 2012).

80. Lewis, K. Persister cells, dormancy and infectious disease. *Nat. Rev. Microbiol.* **5,** 48–56 (2007).

81. Deana, A. & Belasco, J. G. Lost in translation: The influence of ribosomes on bacterial mRNA decay. *Genes Dev.* **19,** 2526–2533 (2005).

82. Fersht, A. R. *Structure and mechanism in protein science. A guide to enzyme catalysis and protein folding* 349–400 (W.H. Freeman and Co., New York, 1999).

83. Johansson, M., Lovmar, M. & Ehrenberg, M. Rate and accuracy of bacterial protein synthesis revisited. *Curr. Opin. Microbiol.* **11,** 141–7 (2008).

84. Kinz-Thompson, C. D. & Gonzalez, R. L. smFRET studies of the 'encounter' complexes and subsequent intermediate states that regulate the selectivity of ligand binding. *FEBS Lett.* **588,** 3526–38 (2014).

85. Vestergaard, B. *et al.* The SAXS solution structure of RF1 differs from its crystal structure and is similar to its ribosome bound cryo-EM structure. *Mol. Cell* **20,** 929–38 (2005).

86. Zoldák, G. *et al.* Release factors 2 from Escherichia coli and Thermus thermophilus: Structural, spectroscopic and microcalorimetric studies. *Nucleic Acids Res.* **35,** 1343–1353 (2007).

87. Korostelev, A. a. Structural aspects of translation termination on the ribosome. *RNA* **17,** 1409–1421 (2011).

88. Trappl, K. & Joseph, S. Ribosome Induces a Closed to Open Conformational Change in Release Factor 1. *J. Mol. Biol.* **318** (2016).

89. Johnson, K. A. Conformational coupling in DNA polymerase fidelity. *Annu. Rev. Biochem.* **62,** 685–713 (1993).

90. Wang, F., Redding, S., Finkelstein, I. J., Gorman, J., Reichman, D. R. & Greene, E. C. The promoter-search mechanism of Escherichia coli RNA polymerase is dominated by three-dimensional diffusion. *Nat. Struct. Mol. Biol.* **20,** 174–81 (2013).

91. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).

92. Tang, C., Iwahara, J. & Clore, G. M. Visualization of transient encounter complexes in protein-protein association. *Nature* **444,** 383–6 (2006).

93. Buck, J., Fürtig, B., Noeske, J., Wöhnert, J. & Schwalbe, H. Time-resolved NMR methods resolving ligand-induced RNA folding at atomic resolution. *Proc. Natl. Acad. Sci.* **104,** 15699–15704 (2007).

94. Kinz-Thompson, C. D. *et al.* Robustly Passivated, Gold Nanoaperture Arrays for Single-Molecule Fluorescence Microscopy. *ACS Nano* **7,** 8159–8166 (2013).

95. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat Meth* **9,** 671–675 (2012).

Part III

Appendices

Appendix A

Additional Projects

A.1 RF1 Solution Conformation

As a complement to the measurements of RF1 on the ribosome (Chapter 5), we additionally investigated the dynamics of surface-tethered wtRF1. By adding an N-terminal enzymatic biotinylation tag (GLNDIFE AQKIEWHE; AviTag) that is *in vitro* biotinylated using recombinant *E. coli* BirA biotin-ligase [1], and one cysteine mutation near the section of RF1 which contacts the DC (S192C) and one cysteine mutation near the section of RF1 which contacts the PTC (E256C) to the cysteine-less wtRF1 (C51S, C201S, C257S), we were able to create a stoichiometrically 1:1 Cy3/Cy5 labeled, biotinylated wtRF1 (wtRF1-(Biotin,Cy3,Cy5)) [2]. When surface-immobilized, and imaged with a TIRF microscope under our standard imaging conditions, wtRF1-(Biotin,Cy3,Cy5) is found to be in a high-E_{FRET} state centered at $E_{FRET} \approx 0.8$ (Figure A.1). This is consistent with reports that RF1 can enter into a 'closed', compact conformation [3, 4] as opposed to the 'open' conformation adopted on the ribosome [5–7], though inconsistent with a small-angle X-ray scattering solution study which found that their data is modeled well by assuming that RF1 is nearly entirely in the 'open' conformation in solution [8]. Repeating the smFRET experiment with an increased acquisition rate (10 msec time period), yields a few E_{FRET} versus time trajectories with transient transitions to a low E_{FRET} state centered perhaps at $E_{FRET} \approx 0.1$. The distance between residues 192 and 256 is r $\approx$ 66 Å in the 'open' conformation [7], and r $\approx$ 38 Å in the 'closed' conformation [3, 4]. For Cy3 and Cy5 FRET donor and acceptor fluorophores, with an estimated R_0 = 55 Å , this yields predicted E_{FRET} of $E_{FRET\ \mathrm{open}} \approx 0.25$, and $E_{FRET\ \mathrm{closed}} \approx 0.90$. While the absolute E_{FRET} do not make the predicted, these observations are consistent with wtRF1 being primarily in the closed conformation when in solution, and transiently fluctuating to the open conformation that is stabilized in the X-ray crystallography structures [5–7]. There are a number reason for which the predicted and the measured E_{FRET} do not match, such as perturbations to the quantum yield of a fluorophore induced by the local biomolecular environment, though perhaps the most reasonable is that here

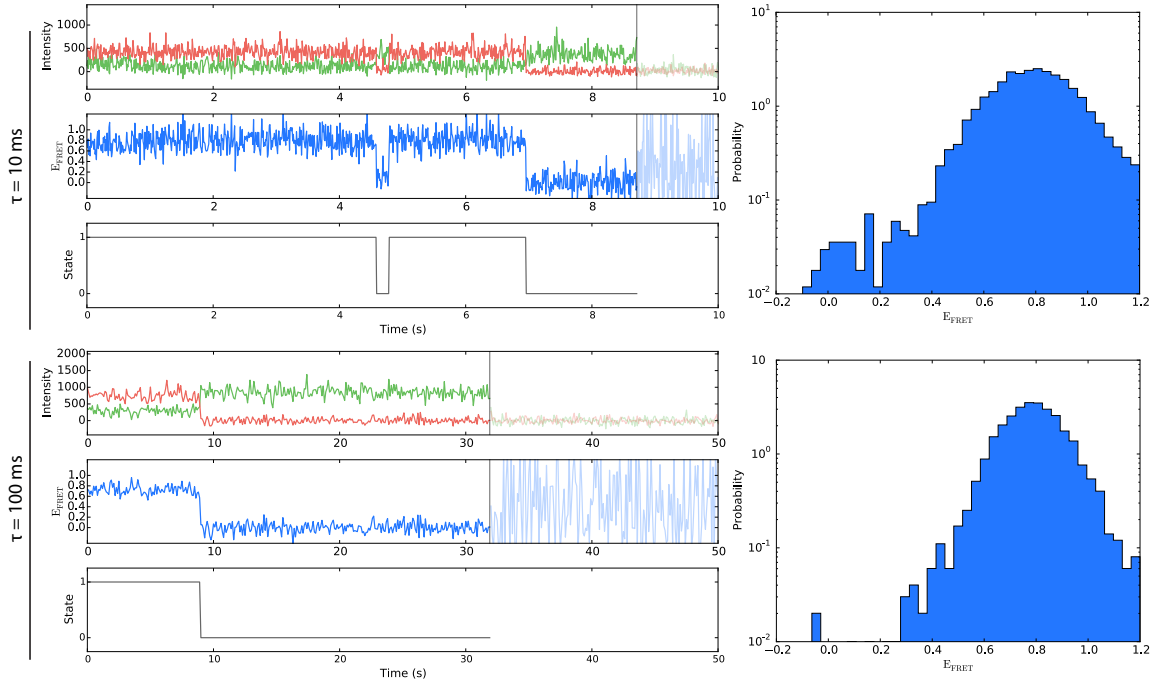

Figure A.1: smFRET from wtRF1-(biotin,Cy3,Cy5) in Solution. Fluorescence Intensity (green, Cy3; red, Cy5) and E_{FRET} (blue) versus time trajectories from individual, surface-immobilized wtRF1-(biotin,Cy3,Cy5) imaged in the absence of ribosomes are shown (left). The wtRF1-(biotin,Cy3,Cy5) mostly occupy a high-FRET state, suggesting a closed, compact form of RF1. Occasionally, at high frame rates of τ = 10 ms, fluctuations to a low-FRET state can be observed (top, n = 38), as shown in the Viterbi path from a two state HMM plotted in black, while at lower frame rates of τ = 100 ms these fluctuations are not observed (bottom, n = 54). Log-scaled histograms from multiple smFRET trajectories also show these effects (right).

we observe wtRF1-(Biotin,Cy3,Cy5) mostly in an ensemble of conformations that are compact and therefore represent a 'closed-like' conformation. Similarly, the low E_{FRET} state probably represents an ensemble of 'open-like' conformations, as the conformation observed bound to the ribosome in the X-ray crystallography structures, while accessable, is difficult to stabilize in the absence of the ribosome. Entropically, we would predict wtRF1 to favor a more compact conformation.

A.2 smFRET with Single-photon Avalanche Photodiode Arrays

In order to monitor Förster resonance energy transfer (FRET) with an extremely fast time resolution, but also with a high-throughput, wide-field approach, we integrated a 64x32 array of single-photon avalanche diodes (SPADs) into a total internal reflection fluorescence (TIRF) microscope. Single-photon Avalanche Diodes (SPAD) are semiconductor p-n junctions that are operated at a bias voltage beyond the breakdown voltage such that a single incident photon will initiate an avalanche of electrons-hole pair generation, which creates

a large, macroscopic electric current [9]. This avalanche current is detected by an active-quenching circuit (ACQ), which lowers the bias voltage, quenching the avalanche, and thus registering the detection of an event [10]. Because of this digital detection scheme, there is no added electronic noise in the measurement from analogue to digital converters or amplifiers, such as in an electron-multiplying charge-coupled device (EMCCD). Recently, CMOS-compatible SPAD fabrication approaches have enabled monolithic integration of multiple SPADs, each with separate AQCs and electronics, into a single device [11]. As a result, single-photon counting cameras composed of arrays of SPADs have been developed with 2,048 (64x32) SPADs, which can detect single photons at frame rates of up to 100,000 frames per second [12]. We have collaborated with Micro Photon Devices S.R.l. - Italy, a spin-off of Politecnico di Milano, to utilize this technology for wide-field detection of smFRET. Previously, Ingargiola and coworkers have utilized two 1x8 arrays of SPADs to simultaneous monitor smFRET from eight confocal spots [13]. Here, we utilize the SPC^3 SPAD array camera to monitor smFRET from Cy3-, and Cy5-labeled pretranslocation (PRE) complexes (see Chapter 4 for more details) with more than two orders of magnitude improvement in throughput.

After wavelength-separating the detected fluorescence using a commercial wavelength-splitter (Dv2, Photometrics) on our custom TIRF microscope (details discussed in Chapter 5), incident-photons recorded using the SPC^3 corresponding to Cy3 and Cy5 were aligned (Fig. A.2A). Then, surface-immobilized PRE^{-A} complexes in microfluidic flow-cells were illuminated via TIRF with a 532 nm laser and photon arrivals were counted with an SPAD integration time of 10.24 μs. The rate of incident photons from Cy3 or Cy5 was quantified using change-point analysis [14, 15] (Fig. A.2B). Anti-correlated photons-arrival rates could be observed from individual SPAD-pairs. Binning the photon record to 100 ms bins, an equivalent EMCCD integration time, reveals E_{FRET} versus time trajectories that are clearly characteristic of our smFRET measurements of PRE^{-A} complexes using an EMCCD (Fig. A.2C).

In this implementation, we were limited by our ability to aligned the two different wavelength images. In order to avoid cross-talk between neighboring SPADs in the array, the pitch between the 30 μm in diameter SPADs is 150 μm, a fill-factor of about 3% [12]. As a result, it is difficult to ensure that the active areas of a pair of wavelength-separated SPADs are exactly aligned. Integration of a microlens array into the camera will help with this limitation in the future. Additionally, the quantum efficiency of SPADs is $\sim$30%, and $\sim$20% at the wavelengths which Cy3 and Cy5 fluoresce [12]. Therefore, while we are able to detect and count individual photons at a very fast rate, the absolute number of detected photons is low, and this can make identifying fluorescence originating from a single-molecule more difficult to discern from hot pixels or spurious background autofluorescence. This can partially be alleviated by utilizing higher laser powers in order to

maximum photon flux from the fluorophores, however this results in faster photobleaching. Regardless, further application of this technology seems extremely promising for smFRET studies of biological systems.

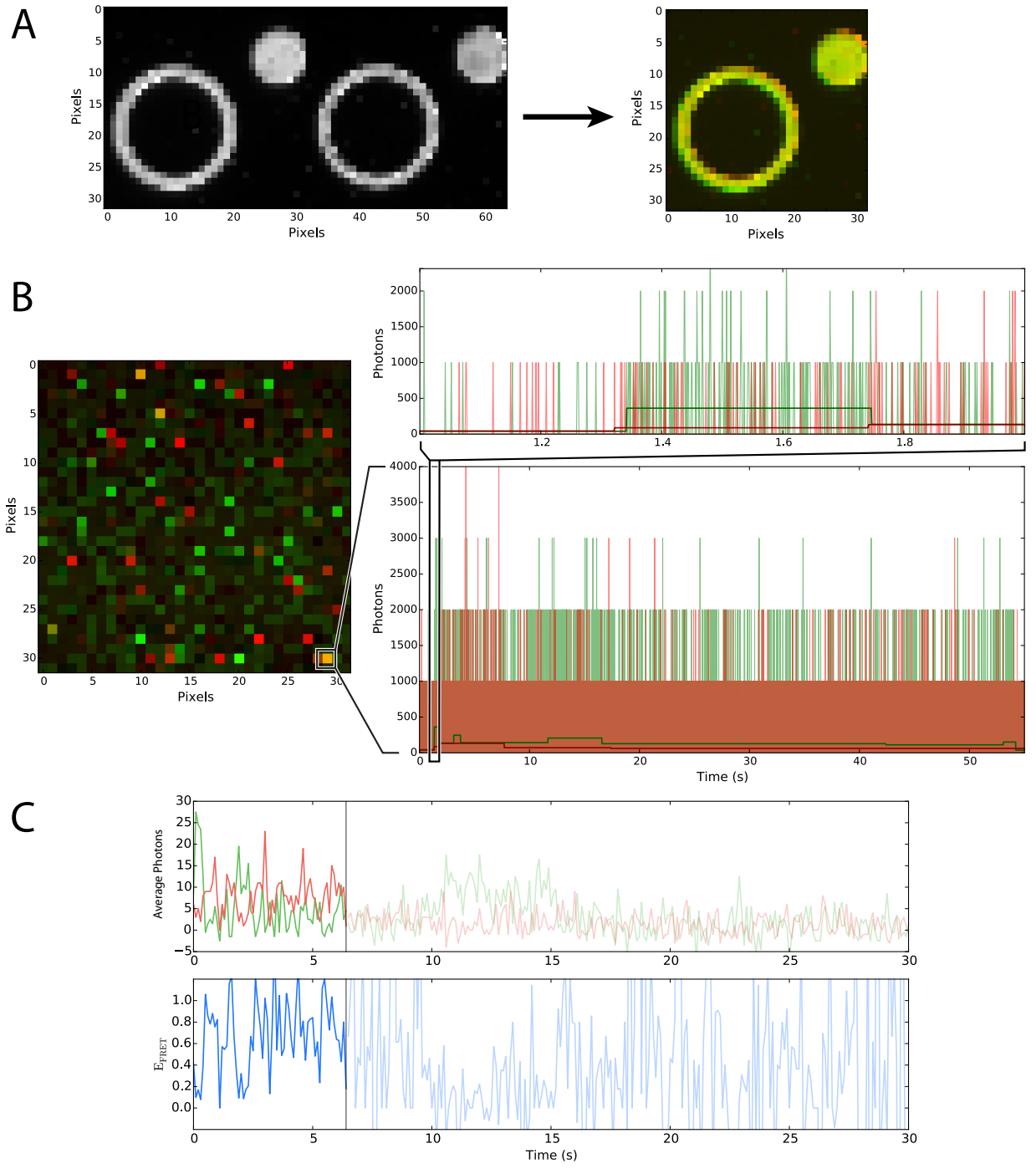

Figure A.2: smFRET Imaging of PRE^{-A} Complexes with SPC^3 SPAD Array. (A) Post-processing alignment of metallic-grid imaged with ambient light. (B) smFRET from PRE^{-A} complexes using TIRF microscopy. Cy3 and Cy5 wavelength-separated photons (green, and red, respectively) are counted from individual PRE^{-A} complexes with a hardware integration time period of 10.24 μs (binned here to τ = 1 ms). The Change-point analysis trajectories are shown for Cy3 (dark red), and Cy5 (dark green). (C) Binned photons from the photons versus time trajectory in panel B such that τ = 100 ms. Fluctuations between two E_{FRET} -levels consistent with GS1 and GS2 E_{FRET} can be seen in the E_{FRET} versus time trajectory.

A.3 Three-color smFRET from BtuCD-F

BtuCD is an ATP-binding cassette transporter that is responsible for importing vitamin B_{12}, which is delivered by the periplasmic binding protein BtuF, into *Escherichia coli* [16, 17]. Recently, Kim and coworkers developed a method to observe the conformational dynamics of nanodisk-reconstituted reconstituted $BtuC_2D_2$ in the presence and/or absence of B_{12}, ATP, the non-hydrolyzable ATP analogue AMPPNP, and BtuF using smFRET [18]. Here, in an effort to extend this smFRET method in order to simultaneously monitor both BtuCD conformational dynamics as well as with vitamin B_{12} transport, we have performed preliminary experiments demonstrating the feasibility of observing the conformational dynamics of $BtuC_2D_2$ as well as the status of BtuF binding by using multi-color smFRET. Here, we have surface-immobilized a nanodisk-reconstituted $BtuC_2D_2$ in which is a single-cysteine BtuC mutant (Q111C) has been labeled with Cy3 and Cy5 in equimolar quantities, such that $BtuC_2D_2$ can contain two Cy3, two Cy5, or one Cy3 and one Cy5; because this is a single-molecule approach, $BtuC_2D_2$s not exhibiting fluorescence from both Cy3 and Cy5 can be ignored. These complexes are then incubated with BtuF, which has been labeled with Alexa Fluor-488 (AF488) at residue 267C in the absence of both vitamin B_{12} and ATP. In the absence of vitamin B_{12} and ATP, we expect that these complexes will be BtuF bound ($BtuC_2D_2$-F) (Fig. A.3A,B), and will be in a static, non-fluctuating state [18]. Finally, these complexes are illuminated with a 488 nm laser on a TIRF microscope, and fluorescence is collected, wavelength-separated using a Quad-View (Photometrics) with Q510/20m, Q575/40m, and Q680/50m emission filters and 540 nm, 630 nm, and 720 nm bandpass filters (Chroma), and then imaged using an Andor iXon3 897 EMCCD.

We were able observe colocalized, anti-correlated fluorescence intensity from AF488, Cy3, and Cy5, corresponding to static, BtuF bound $BtuC_2D_2$-F (Fig. A.3C). In these trajectories, we always observed Cy5 photobleaching occur before AF488 and Cy3 photobleaching. In order to quantify the resonance energy transfer process, we created a histogram of relative fluorescence intensity from n = 38 of these fluorescence intensity versus time trajectories by dividing the observed AF488, Cy3, or Cy5 fluorescence intensity at each timepoint by the sum of the AF488, Cy3, and Cy5 fluorescence intensities at that time point, and histogramming these values (Fig. A.3D). Notably, the relative fluorescence histograms all have two maxima (Cy3's are overlapped), which correspond to the means pre- and post-Cy5 photobleaching. In order to interpret $BtuC_2D_2$-F structural information from these values, we modeled the kinetics of the FRET process. Since these mean relative fluorescence intensity values correspond to the fluorophore absorption, emission, and resonance energy transfer mechanism shown in Fig. A.3B, we modeled the kinetics of this process

with literature values for the fluorescence lifetimes, and relative absorption of 488 nm photons of the three fluorophores to estimate the RET lifetimes observed in from BtuC_2D_2-F complexes. For relative 488 nm photon absorption, we utilized arbitrarily fast 1 fs lifetimes weighted by the relative absorption probabilities of 0.71, 0.29,and 0.00 for AF488, Cy3, and Cy5 respectively, since electronic excitation is a very rapid process. Additionally, for the fluorescence lifetimes for the fluorophores we used 4.1 ns, 0.3 ns and 1.0 ns from measurements in water of AF488, Cy3, and Cy5, respectively. In order to obtain enough constraints to solve for the RET lifetimes, we utilized the full model and the mean relative fluorescence values pre-Cy5 photobleaching, and a modified model in which RET between AF488 and Cy5, and Cy3 and Cy5 were not allowed. Non-linear least-squares fitting these constraints to the steady-state values obtained by using numerical Laplace inversion of the corresponding Green's function, which is

$$P(t) = G(t) \cdot P(t = 0), \tag{A.1}$$

where

$$G(t) = (L)^{-1} \left(\tilde{G}(s) \right), \text{ and} \tag{A.2}$$

$$\tilde{G}(s) = (s \cdot \boldsymbol{I} - \boldsymbol{K})^{-1}, \tag{A.3}$$

where P is probability of a state, t is time, s is the Laplace space variable, $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{K}$ is the rate matrix, which obeys mass conservation. The resulting best-estimate RET lifetimes were 41.5 ns, 1.9 ns, and 0.4 ns for AF488 to Cy3, AF488 to Cy5, and Cy3 to Cy5, respectively (Fig. A.3B). While it is difficult to interpret the exact structural meaning these values in relation to BtuC_2D_2-F conformation, they are a function of the distance between the fluorophores, r, the R_0 value for the fluorophore pairs, which quantifies contributions such as spectral overlap and dipole moment orientation. If for some reason, the fluorophore dipole moment orientations were not isotropically averaged out (*e.g.*, steric occlusion), then these RET lifetimes could also depend upon the average dipole-dipole orientation $\langle \mu_{\text{donor}} \cdot \mu_{\text{acceptor}} \rangle$.

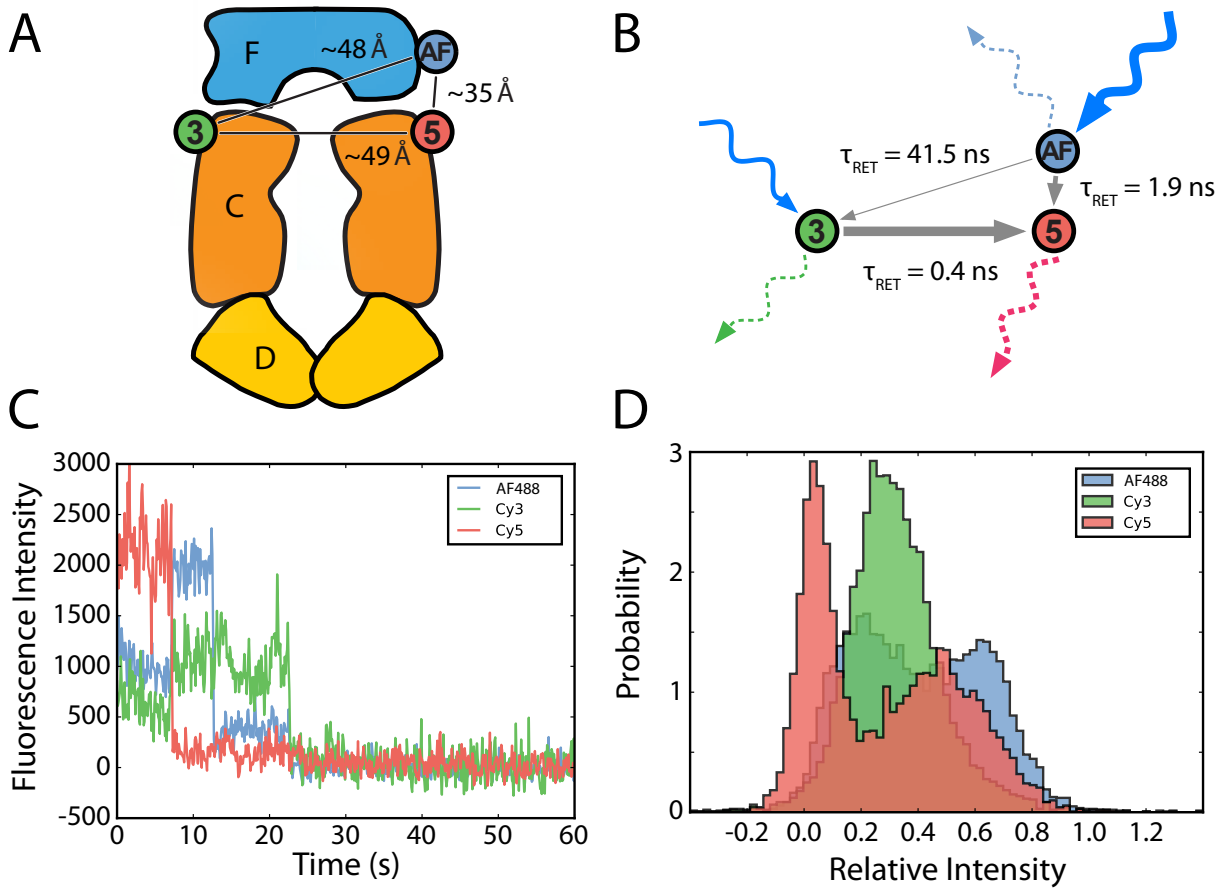

Figure A.3: Three-color smFRET from Nanodisc Embedded $BtuC_2D_2$-F. (A) Cartoon schematic of fluorophore labeled $BtuC_2D_2$-F. The cartoon and the distances between Alexa Fluor-488 (AF), Cy3 (3), and Cy5 (5) are based upon the $BtuC_2D_2$-F crystal structure (PDB:2Q19). (B) Schematic showing response of AF, Cy3, and Cy5 to illumination by 488 nm light. Solid and dashed curved lines represent absorbed and emitted photons, respectively. Grey arrows represent resonance energy transfer (RET). τ_{RET} is the average time to undergo RET as determined by kinetic modeling of the experimental smFRET results. (C) Representative three-color smFRET fluorescence intensity versus time trajectory from $BtuC_2D_2$ illuminated with via TIRF with a 488 nm laser. (D) Histogram of relative AF488, Cy3, and Cy5 fluorescence intensities at all time points prior to Cy3 photobleaching for n = 38 fluorescence intensity versus time trajectories.

A.4 Single-Molecule Fluorescence Spectroscopy

In this section, we detail an approach to learn about the motion of the dipole moment of a single fluorophore by analyzing the spectral lineshape function, $I(\omega)$. From Gordon [19, 20], in an isotropic fluid, this lineshape function is

$$I(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt e^{-i\omega t} \langle \boldsymbol{M}(0) \cdot \boldsymbol{M}(t) \rangle, \tag{A.4}$$

where $\boldsymbol{M}(t)$ is total electric dipole moment of the entire system. However, if the molecules are dilute or at the single-molecule level, then

$$\langle \boldsymbol{M}(0) \cdot \boldsymbol{M}(t) \rangle = N \langle \boldsymbol{\mu}(0) \cdot \boldsymbol{\mu}(t) \rangle, \tag{A.5}$$

where $\boldsymbol{\mu}$ is the transition dipole moment of a single-molecule. Therefore, Fourier inversion of the the spectral lineshape function equation yields the correlation function of the dipole moment, and speaks to its movement in solution. Here, we began to develop an attempt to measure fluorescence emission lineshape functions with wide-field optics so as to be able to observe the lineshapes of many single-molecules simultaneously. To do so, a pair of chromatically dispersing prisms was placed into the emission path of a TIRF microscope, so that the emitted fluorescence of the single-molecules would be chromatically dispersed and then collimated (Fig. A.4). Intensity along a strip of pixels on the EMCCD is then taken to be the fluorescence spectrum. An example of Cy3-labeled dsDNA that has been tethered to a coverslip using a biotin-streptavidin-biotin bridge is shown in Fig. A.5. In order to transform pixels to wavelengths, we used the dispersion formula for NSF11 glass

$$n(\lambda) = \left(1 + \frac{C_1 \lambda^2}{(\lambda^2 - C_2)} + \frac{C_3 \lambda^2}{(\lambda^2 - C_4)} + \frac{C_5 \lambda^2}{(\lambda^2 - C_6)}\right)^{(1/2)}, \tag{A.6}$$

where the c_i are tabulated constants. In order to calibrate the wavelengths, we placed an emission-filter into the fluorescence emission pathway, and utilized the known wavelength cutoffs to correspond to the disappearance of Cy3 fluorescence. An example is shown in Fig. A.5. Moreover, spectra can be collected from individual molecules in a time-dependent manner to observe if any large-scale conformational changes, which occur on a time-scale much slower than the movement of the dipole moment, are occurring.

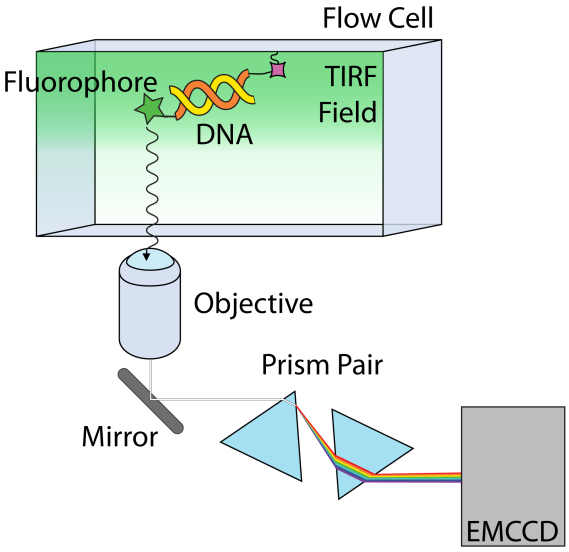

Figure A.4: Schematic Diagram of Single-Molecule Fluorescence Spectroscopy Instrumentation. Widefield fluorescence from multiple single molecules of DNA conjugated to Cy3 fluorophores is collected through and objective. A pair of chromatically dispersing prisms separates and collimates the incident fluorescence, which is then imaged with an EMCCD.

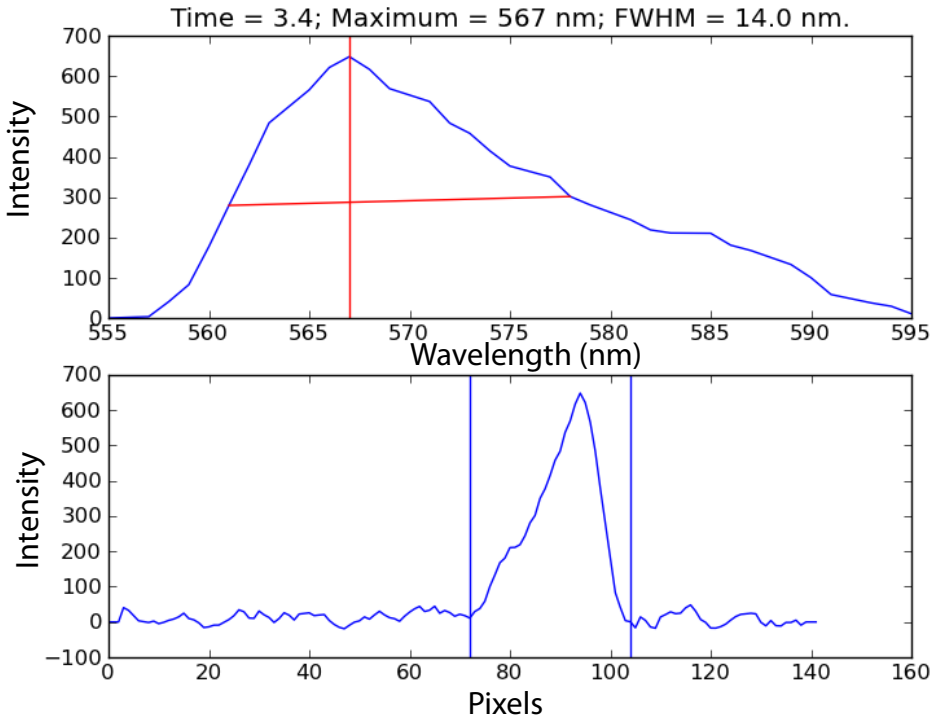

Figure A.5: Time-Dependent, Single-Molecule Cy3-DNA Fluorescence Spectrum. (Top) A timepoint from a series fluorescence intensity versus calibrated wavelength collected at 100 ms from Cy3-DNA illuminated with a 532 nm laser with a TIRF microscope. (Bottom) The fluorescence intensity versus pixel spectrum which is calibrated to yield the spectrum above. The first zero fluorescence intensity crossings from the global maximum are chosen to correspond to the emission filter (Q575/40m, Chroma) upper- and lower-wavelength bounds.

A.5 References

1. Howarth, M. & Ting, A. Y. Imaging proteins in live mammalian cells with biotin ligase and monovalent streptavidin. *Nat. Protoc.* **3,** 534–45 (2008).

2. Kinz-Thompson, C. D. *et al.* Robustly Passivated, Gold Nanoaperture Arrays for Single-Molecule Fluorescence Microscopy. *ACS Nano* **7,** 8159–8166 (2013).

3. Shin, D. H., Brandsen, J., Jancarik, J., Yokota, H., Kim, R. & Kim, S. H. Structural analyses of peptide release factor 1 from Thermotoga maritima reveal domain flexibility required for its interaction with the ribosome. *J. Mol. Biol.* **341,** 227–239 (2004).

4. Graille, M. *et al.* Molecular basis for bacterial class I release factor methylation by PrmC. *Mol. Cell* **20,** 917–927 (2005).

5. Petry, S. *et al.* Crystal structures of the ribosome in complex with release factors RF1 and RF2 bound to a cognate stop codon. *Cell* **123,** 1255–1266 (2005).

6. Laurberg, M., Asahara, H., Korostelev, A., Zhu, J., Trakhanov, S. & Noller, H. F. Structural basis for translation termination on the 70S ribosome. *Nature* **454,** 852–7 (2008).

7. Korostelev, A., Zhu, J., Asahara, H. & Noller, H. F. Recognition of the amber UAG stop codon by release factor RF1. *EMBO J.* **29,** 2577–85 (2010).

8. Vestergaard, B. *et al.* The SAXS solution structure of RF1 differs from its crystal structure and is similar to its ribosome bound cryo-EM structure. *Mol. Cell* **20,** 929–38 (2005).

9. Cova, S., Ghioni, M., Lotito, a., Rech, I. & Zappa, F. Evolution and prospects for single-photon avalanche diodes and quenching circuits. *J. Mod. Opt.* **51,** 1267–1288 (2004).

10. Cova, S., Ghioni, M., Lacaita, a., Samori, C. & Zappa, F. Avalanche photodiodes and quenching circuits for single-photon detection. *Appl. Opt.* **35,** 1956–1976 (1996).

11. Ghioni, M., Gulinatti, A., Rech, I., Zappa, F. & Cova, S. Progress in silicon single-photon avalanche diodes. *IEEE J. Sel. Top. Quantum Electron.* **13,** 852–862 (2007).

12. Bronzi, D. *et al.* 100 000 Frames/s 64 x 32 Single-Photon Detector Array for 2-D Imaging and 3-D Ranging. *IEEE J. Sel. Top. Quantum Electron.* **20,** 354–363 (2014).

13. Ingargiola, A. *et al.* 8-spot smFRET analysis using two 8-pixel SPAD arrays. *Proc. SPIE* **8590,** 1–11 (2013).

14. Watkins, L. P. & Yang, H. Detection of intensity change points in time-resolved single-molecule measurements. *J. Phys. Chem. B* **109,** 617–628 (2005).

15. Yang, H. in *Single-Molecule Biophys. Exp. Theory, Vol. 146* (eds Komatsuzaki, T., Kawakami, M., Takahashi, S., Yang, H. & Silbey, R. J.) 219–243 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2011).

16. Locher, K. P., Lee, A. T. & Rees, D. C. The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science* **296,** 1091–1098 (2002).

17. Hvorup, R. N., Goetz, B. A., Niederer, M., Hollenstein, K., Perozo, E. & Locher, K. P. Asymmetry in the Structure of the ABC Transporter-Binding Protein Complex BtuCD-BtuF. *Science* **317,** 1387–1390 (2007).

18. Kim, J., Ramos, J. E., Leng, K., Karpowich, N. K., Gonzalez Jr., R. L. & Hunt, J. F. A power-stroke coupled to ATP hydrolysis drives transport by the type II ABC importer BtuC2D2. *Prep.* (2016).

19. Gordon, R. G. Molecular Collisions and the Depolarization of Fluorescence in Gases. *J. Chem. Phys.* **45,** 1643–1648 (1966).

20. McQuarrie, D. A. *Statistical Mechanics* 470–476 (University Science Books, Sausalito, 2000).

Appendix B

Probability Distributions

Beta

Variable: x

Parameters: α, β

Support: $x \in (0, 1)$

PDF: $p(x|\alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha, \beta)$

CDF: $P(x|\alpha, \beta) = B(x; \alpha, \beta)/B(\alpha, \beta)$

Mean: $\mathbb{E}[x] = \alpha/(\alpha + \beta)$

Variance: $\mathbb{E}[x^2] - \mathbb{E}[x]^2 = \alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1))$

Binomial

Variable: k

Parameters: n, p

Support: $k \in 0, 1, ..., n$

PMF: $p(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$

CMF: $P(k|n, p) = B\left(1-p, n-k, 1+k\right)/B\left(n-k, 1+k\right)$

Mean: $\mathbb{E}[k] = np$

Variance: $\mathbb{E}[k^2] - \mathbb{E}[k]^2 = np(1-p)$

Dirichlet

Variable: $\boldsymbol{x}$

Parameters: $\boldsymbol{\alpha}$

Support: $x_i \in (0, 1)$

PDF: $p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{1}{\boldsymbol{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

Mean: $\mathbb{E}[x] = \alpha_i / \sum_i \alpha_i$

Variance: $\mathbb{E}[x^2] - \mathbb{E}[x]^2 = \frac{\alpha_i((\sum_i \alpha_i) - \alpha_i)}{(\sum_i \alpha_i)^2 (1 + \sum_i \alpha_i)}$

Exponential

Variable: x

Parameters: k

Support: $x \in [0, \infty)$

PDF: $p(x|k) = k \cdot e^{-kx}$

CDF: $P(x|k) = 1 - e^{-kx}$

Mean: $\mathbb{E}[x] = 1/k$

Variance: $\mathbb{E}[x^2] - \mathbb{E}[x]^2 = 1/k^2$

Gamma

Variable: x

Parameters: α, β

Support: $x \in (0, \infty)$

PDF: $p(x|\alpha, \beta) = \beta^\alpha x^{\alpha-1} e^{-\beta x} / \Gamma(\alpha)$

CDF: $P(x|\alpha, \beta) = \gamma(\alpha, \beta x) / \Gamma(\alpha$

Mean: $\mathbb{E}[x] = \alpha/\beta$

Variance: $\mathbb{E}[x^2] - \mathbb{E}[x]^2 = \alpha/\beta^2$

Geometric

Variable: k

Parameters: p

Support: $k \in 0, 1, ...$

PMF: $p(k|p) = p(1-p)^k$

CMF: $P(k|p) = 1 - (1-p)^{k-1}$

Mean: $\mathbb{E}[k] = (1-p)/p$

Variance: $\mathbb{E}[k^2] - \mathbb{E}[k]^2 = (1-p)/p^2$

Multinomial

Variable: $\boldsymbol{x}$

Parameters: $n, \boldsymbol{p}$

Support: $x_i \in 0, ..., n$

PMF: $p(\boldsymbol{x}|n, \boldsymbol{p}) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} p_i^{x_i}$

Mean: $\mathbb{E}[n_i] = n \cdot p_i$

Negative Binomial

Variable: k

Parameters: r, p

Support: $k \in 0, 1, ...$

PMF: $p(k|r, p) = (-1)^k \binom{-r}{k} p^r (1-p)^k$

CMF: $P(k|r, p) = 1 - B(1-p, k+1, r)/B(k+1, r)$

Mean: $\mathbb{E}[k] = r\frac{1-p}{p}$

Variance: $\mathbb{E}[k^2] - \mathbb{E}[k]^2 = \frac{r(1-p)}{p^2}$

Normal

Variable: x

Parameters: μ, σ^2

Support: $x \in (-\infty, \infty)$

PDF: $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$

CDF: $P(x|\mu, \sigma^2 = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$

Mean: $\mathbb{E}[x] = \mu$

Variance: $\mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$

Multivariate Normal

Variable: $\boldsymbol{x}$

Parameters: $\boldsymbol{\mu}, \boldsymbol{\Sigma}$

Support: $x_i \in (-\infty, \infty)$

PDF: $p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-k/2} \left|\boldsymbol{\sigma}\right|^{-1/2} e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$

Mean: $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$

Variance: $\mathbb{E}[\boldsymbol{x}^2] - \mathbb{E}[\boldsymbol{x}]^2 = \Sigma$

Uniform

Variable: x

Parameters: a, b

Support: $x \in [a, b]$

PDF: $p(x|a, b) = 1/(b-a)$

CDF: $P(x|a, b) = (x-a)/(b-a)$

Mean: $\mathbb{E}[x] = (a+b)/2$

Variance: $\mathbb{E}[x^2] - \mathbb{E}[x]^2 = (b-a)^2/12$

Wishart

Variable: $\boldsymbol{X}$

Parameters: $\boldsymbol{W}, \nu$

Support: $\boldsymbol{X}$ is positive definite

PDF: $p(\boldsymbol{X}|\boldsymbol{W}, \nu) = \frac{|\boldsymbol{W}|^{(\nu-d-1)/2} e^{-\mathrm{tr}(\boldsymbol{W}^{-1} \cdot X)/2}}{2^{\nu d/2} \boldsymbol{\Gamma}_d(\nu/2) |\boldsymbol{W}|^{\nu/2}}$

Mean: $\mathbb{E}[\boldsymbol{X}] = \nu \boldsymbol{W}$

Variance: $\mathbb{E}[\boldsymbol{X}^2] - \mathbb{E}[\boldsymbol{X}]^2 = \nu(\boldsymbol{W}_{ij}^2 + \boldsymbol{W}_{ii}\boldsymbol{W}_{jj})$

Functions

B(x,y) is the beta function of x and y, B(z;x,y) is the incomplete beta function of x and y evaluated at z, $\boldsymbol{B}$(x) is the multinomial beta function of x, Γ(x) is the gamma function of x, γ(x,y) is the incomplete gamma function of x and y, and $\boldsymbol{\Gamma}_d$(x) is the multivariate gamma function of x. erf is the error function. $\binom{x}{y}$ is the binomial coefficient of x and y, and $\binom{x}{y_1,\ldots,y_k}$ is the multinomial coefficient of x and y_i.

Appendix C

Release Factor 1

$(His)^6$ - TEV - 167C wtRF1

Primary Sequence: MSYYHHHHHHDYDIPTTENLYFQGAMKPSIVAKLEALHERHEEVQALLGDAQTI
ADQERFRALSREYAQLSDVSRSFTDWQQVQEDIETAQMMLDDPEMREMAQDE
LREAKEKSEQLEQQLQVLLLPKDPDDERNAFLEVRAGTGGDEAALFAGDLFRM
YSRYAEARRWRVEIMSASEGEHGGYKEIIAKISGDGVYGRLKFESGGHRVQRV
PATECQGRIHTSASTVAVMPELPDAELPDINPADLRIDTFRSSGAGGQHVNTTD
SAIRITHLPTGISVESQDERSQHKNKAKALSVLGARIHAAEMAKRQQAEASTRRN
LLGSGDRSDRNRTYNFPQGRVTDHRINLTLYRLDEVMEGKLDMLIEPIIQEHQA
DQLAALSEQE

MW: 43,569 Da

pI: 5.26

Number of Residues: 385

ϵ(280 nm, H_2O): 27,390 M^{-1} cm^{-1}

167C wtRF1

Primary Sequence: GAMKPSIVAKLEALHERHEEVQALLGDAQTIADQERFRALSREYAQLSDVSRSF
TDWQQVQEDIETAQMMLDDPEMREMAQDELREAKEKSEQLEQQLQVLLLPKD
PDDERNAFLEVRAGTGGDEAALFAGDLFRMYSRYAEARRWRVEIMSASEGEH
GGYKEIIAKISGDGVYGRLKFESGGHRVQRVPATECQGRIHTSASTVAVMPELP
DAELPDINPADLRIDTFRSSGAGGQHVNTTDSAIRITHLPTGISVESQDERSQHK

NKAKALSVLGARIHAAEMAKRQQAEASTRRNLLGSGDRSDRNRTYNFPQGRVT
DHRINLTLYRLDEVMEGKLDMLIEPIIQEHQADQLAALSEQE

MW: 40,601 Da

pI: 5.13

Number of Residues: 362

ϵ(280 nm, H_2O): 21,430 M^{-1} cm^{-1}

Appendix D

Computer Code

D.1 Compounded Dwell-time Probabilities

```
import numpy as np
from scipy import special

# Negative Binomial Distribution
def nb(k,r,p):
        return special.binom(k+r-1,k) * (1.-p)**r * p**k

# Compound Dwell-time Probabilities
def nbmix_probs(k1,k2,tau,tc):
        ### Inputs: k_1, k_2, τ, τ_c
        ### Returns: probability of the number of compounded dwells (1 to 6)

        # Probabilities of types of dwells
        fp = np.exp(-k1*tc)
        gp = np.exp(-k2*tc)
        fm = 1.-fp
        gm = 1.-gp

        # Pairwise probabilities
        ps = np.array((fp*gm,fm*gm,fp*gp,fm*gp))

        # Pathway probabilities for 12 dwells
        quad_index = np.mgrid[:4,:4,:4,:4,:4,:4].reshape((6,4**6)).T
        pathways = np.product(ps[quad_index],axis=1)

        # Calculate number of f^+ for each pathway
        n_dwells = np.zeros_like(pathways)
        n_fp_raw = (((quad_index == 0) + (quad_index == 2)) > 0).cumsum(1)
        first_gp = (((quad_index==2) + (quad_index == 3)) > 0).argmax(1)
        for i in range(first_gp.size):
                n_dwells[i] = n_fp_raw[i,first_gp[i]]

        ### Figure out which pathways to ignore, then remove them
        # If no g^+, remove
        cut1 = ((quad_index == 2) + (quad_index == 3)).sum(1) == 0
        # If no f^+, remove
        cut2 = ((quad_index == 0) + (quad_index == 2)).sum(1) == 0
        # if end with f^{-g^+}, without an f^+
        cut3 = quad_index[:,0] == 3
        pathways[(cut1 + cut2 + cut3) > 0] = 0.
        # If no f^+ before f^{-g^+} in ith position
        for i in range(1,6):
                cut = (quad_index[:,i] == 3)*np.all(quad_index[:,:i] != 2,axis=1)
```

```
                cut *= np.all(quad_index[:,:i]!=0,axis=1) > 0
                pathways[cut] = 0.

        # Sum compounded probabilities, and normalize
        ppp = np.array(())
        for i in range(6):
                ppp = np.append(ppp,pathways[np.nonzero(n_dwells == i+1)[0]].sum())
        return ppp / ppp.sum()
```

D.2 Single Molecule Stochastic Simulation Algorithm (SSA)

Generate Trajectories (.C)

```
#include <time.h>
#include <stdlib.h>
#include <math.h>
#include <stdio.h>

void sm_ssa(int steps, double tmax, int nstates, int initialstate, double *flat_rates, int *states, double
 ↪  *dwells, int *last);
void render_trace(int steps, int timesteps, int nstates, double *x, double *y, int *states, double *times,
 ↪  double *dwells, double *emissions);

void sm_ssa(int steps, double tmax, int nstates, int initialstate, double *flat_rates, int *states, double
 ↪  *dwells, int *last){

        int i,j;
        double r1,r2;
        double rates[nstates][nstates], cumulatives[nstates][nstates];
        double outrates[nstates];
        int currentstate = initialstate;
        double timetotal = 0.;

        // setup cutoffs and rates
        for(i=0;i<nstates;i++){
                outrates[i] = 0.;
                cumulatives[i][0] = 0.;
                for(j=0;j<nstates;j++){
                        rates[i][j] = flat_rates[nstates*i + j];
                        cumulatives[i][j] = flat_rates[nstates*i + j];
                        if (j > 0){
                                cumulatives[i][j] += cumulatives[i][j-1];
                        }
                }
                outrates[i] = cumulatives[i][nstates-1];
        }

        // seed random num generator
        srand(time(NULL));

        for(i=0;i<steps;i++){
                states[i] = currentstate;

                r1 = ((double)rand()) / ((double)RAND_MAX);
                r2 = ((double)rand()) / ((double)RAND_MAX);
                dwells[i] = 1./outrates[states[i]]* log(1./r1);

                if (i == 0){ // Start trace at random time.
                        dwells[i]*=(((double)rand()) / ((double)RAND_MAX));
                }
                timetotal += dwells[i];
```

```
                for(j=0;j<nstates;j++){
                        currentstate = j;
                        if (cumulatives[states[i]][j] > r2*outrates[states[i]]) {
                                break;
                        }
                }
                if (timetotal > tmax) {
                        last[0] = i+1;
                        break;
                }
        }
}

void render_trace(int steps, int timesteps, int nstates, double *x, double *y, int *states, double *times,
 ↪  double *dwells, double *emissions) {

        int i,j;
        double t0,t1;
        double tau = x[1] - x[0];
        int a=0, b=0,aflag = 1,bflag=1;

        for (i=0;i<steps;i++) {
                t1 = x[i];
                t0 = t1 - tau;
                aflag = 1;
                bflag=1;
                for (j=a;j<timesteps+1;j++){
                        if (times[j] > t0 && aflag) {
                                a = j-1;
                                aflag = 0;
                        }
                        if (times[j] > t1 && bflag) {
                                b = j;
                                bflag = 0;
                        }
                        if (!aflag && !bflag){
                                y[i] = 0.;
                                if (b-a == 1){
                                        y[i] += (t1-t0)/tau * emissions[states[a]];
                                }
                                else if (b-a > 1) {
                                        y[i] += (times[a+1]-t0)/tau * emissions[states[a]];
                                        for (j=a+1;j<b-1;j++) {
                                                y[i] += dwells[j]/tau * emissions[states[j]];
                                        }
                                        y[i]+= (t1 - times[b-1])/tau * emissions[states[b-1]];
                                }
                                break;
                        }
                }
        }
}

// To Compile
// gcc -shared -o ./sm_ssa-linux.so -fPIC -O3 ./sm_ssa.c
// gcc -shared -o ./sm_ssa-mac.so -fPIC -O3 ./sm_ssa.c
```

Python Wrapper

```
import numpy as np
import ctypes
```

```
from sys import platform
import os

path = os.path.dirname(__file__)

if platform == 'darwin':
        _sopath = path+'/sm_ssa-mac'
elif platform == 'linux' or platform == 'linux2':
        _sopath = path + '/sm_ssa-linux'
_lib = np.ctypeslib.load_library(_sopath, '.')

_lib.sm_ssa.argtypes = [ctypes.c_int, ctypes.c_double, ctypes.c_int, ctypes.c_int,
 ↪   np.ctypeslib.ndpointer(dtype = np.double), np.ctypeslib.ndpointer(dtype = np.int32),
 ↪   np.ctypeslib.ndpointer(dtype = np.double), np.ctypeslib.ndpointer(dtype=np.int32)]
_lib.sm_ssa.restype  = ctypes.c_void_p


_lib.render_trace.argtypes = [ctypes.c_int, ctypes.c_int, ctypes.c_int, np.ctypeslib.ndpointer(dtype =
 ↪   np.double),  np.ctypeslib.ndpointer(dtype = np.double), np.ctypeslib.ndpointer(dtype = np.int32),
 ↪   np.ctypeslib.ndpointer(dtype = np.double), np.ctypeslib.ndpointer(dtype=np.double),
 ↪   np.ctypeslib.ndpointer(dtype=np.double)]
_lib.render_trace.restype  = ctypes.c_void_p

class simtrace:
        @staticmethod
        def generate_states(rates,tlength):
                nstates = rates.shape[0]
                rates = rates.flatten()
                n = np.max((int(np.floor(tlength*rates.max())),1))*2
                initialstate = np.random.randint(nstates)
                states = np.zeros(n, dtype=np.int32)
                dwells = np.zeros(n, dtype=np.double)
                cut = np.array(0,dtype=np.int32)
                _lib.sm_ssa(n, (tlength), nstates, initialstate, rates, states, dwells,cut)
                states = states[:cut]
                dwells = dwells[:cut]
                return np.array((states,dwells))

        @staticmethod
        def render_trace(trace,steps,dt,emission):
                states = trace[0].astype(np.int32)
                dwells = trace[1]
                times = dwells.cumsum().astype(np.double)
                times = np.append(0,times)
                timesteps = states.shape[0]
                steps = int(steps)

                nstates = emission.size
                emissions = emission.astype(np.double)
                x = np.arange(steps,dtype=np.double)*dt + dt
                y  = np.zeros_like(x)
                _lib.render_trace(steps, timesteps, nstates, x, y, states, times, dwells, emissions)
                return x,y

        def __init__(self,rates,emissions,noise,frames,tau):
                self.k = rates
                self.emission = emissions
                self.a = self.generate_states(self.k,frames*tau)
                self.dt = tau
                self.sigma = noise
                self.x,self.y = self.render_trace(self.a,frames,self.dt,self.emission)
                self.y += np.random.normal(scale=self.sigma,size=self.y.size)
```

D.3 Multidimensional Machine Learning - Variational GMM and HMM

Forward-Backward Algorithm (.C)

```
// Adapted straight from JW van de Meent's forwardback.cpp for MatLab Mex-file from ebFRET.
// Turned it into C code instead of C++ for use with ctypes in python
// Also, redid the indexing

#include <math.h>
#include <stdlib.h>

void forward_backward(int T, int K, double *p_x_z, double *A, double *pi, double *g, double *xi, double
 ↪  *ln_z);

void forward_backward(int T, int K, double *p_x_z, double *A, double *pi, double *g, double *xi, double *ln_z)
 ↪  {
        double *a, *b, *c;
        a = malloc(T*K*sizeof(double));
        b = malloc(T*K*sizeof(double));
        c = malloc(T*sizeof(double));

        // initialize to zero
        int i;
        for (i=0; i<T*K; i++) {
                a[i] = 0;
                b[i] = 0;
        }
        for (i=0; i<T; i++) {
                c[i] = 0;
        }

        // Forward Sweep - Calculate
        //
        //   a(t, k)  =  sum_l p_x_z(t,k) A(l, k) alpha(t-1, l)
        //   c(t)         =  sum_k a(t, k)
        //
        // and normalize
        //
        //   a(t, k)  /=  c(t)

        // a(0, k)  =  p_x_z(0, k) pi(k)
        int k;
        for (k = 0; k < K; k++) {
                a[0*K + k] = pi[k] * p_x_z[0*K + k];
                c[0] += a[0*K + k];
        }

        // normalize a(0,k) by c(k)
        for (k = 0; k < K; k++) {
                a[0*K + k] /= c[0];
        }

        int t = 0;
        int l;
        for (t = 1; t < T; t++) {
                // a(t, k)  =  sum_l p_x_z(t,k) A(l, k) alpha(t-1, l)
                for (k = 0; k < K; k++) {
                        for (l = 0; l < K; l++) {
                                // a(t,k) +=  p_x_z(t,k) A(l, k) alpha(t-1, l)
                                a[t*K + k] += p_x_z[t*K + k] * A[l*K + k] * a[(t-1)*K + l];
                        }
```

```
                        // c(t) += a(t,k)
                        c[t] += a[t*K + k];
                }
                // normalize a(t,k) by c(t)
                for (k = 0; k < K; k++) {
                        a[t*K + k] /= c[t];
                }
        }

        // Back sweep - calculate
        //
        // b(t,k)  =  1/c(t+1) sum_l p_x_z(t+1, l) A(k, l) beta(t+1, l)

        // b(T-1,k) = 1
        for (k = 0; k < K; k++) {
                b[(T-1)*K + k] = 1;
        }

        // t = T-2:0
        for (t = T-2; t >= 0; t--) {
                // b(t, k)  =  sum_l p_x_z(t+1,l) A(k, l) beta(t+1, l)
                for (k = 0; k < K; k++) {
                        for (l = 0; l < K; l++) {
                                // b(t ,k) += p_x_z(t+1, l) A(k, l) betal(t+1, l)
                                b[t*K + k] += p_x_z[(t+1)*K + l] * A[k*K + l] * b[(t+1)*K + l];
                        }
                        // normalize b(t,k) by c(t+1)
                        b[t*K + k] /= c[t+1];
                }
        }

        // g(t,k) = a(t,k) * b(t,k)
        for (i=0; i<T*K; i++){
                g[i] = a[i] * b[i];
        }

        // xi(t, k, l) = alpha(t, k) A(k,l) p_x_z(t+1, l) beta(t+1, l) / c(t+1)
        for (t = 0; t < T-1; t++) {
                for (k = 0; k < K; k++) {
                        for (l = 0; l < K; l++) {
                                xi[t*K*K + k*K +l] = (a[t*K + k] \
                                                                * A[k*K + l] \
                                                                * p_x_z[(t+1)*K + l] \
                                                                * b[(t+1)*K + l]) / c[t+1];
                        }
                }
        }

        // ln_Z = sum_t log(c[t])
        ln_z[0] = 0;
        for (t=0; t<T; t++) {
                ln_z[0] += log(c[t]);
        }

        // delete memory allocated for a, b and c
        free(a);
        free(b);
        free(c);

        return;
}

// To compile:
// gcc -shared -o ./forward_backward-linux.so -fPIC -O3 ./forward_backward.c
```

```
// gcc -shared -o ./forward_backward-mac.so -fPIC -O3 ./forward_backward.c
```

Laplace Approximation, K-Means, Variational GMM, Variational HMM

```
import numpy as np
np.seterr(all='ignore')
eps = np.finfo(float).eps
from scipy import special
from ckt_utils import stats
import ctypes
from sys import platform
import os

path = os.path.dirname(__file__)

#Try to load forward backward
try:
        if platform == 'darwin':
                _sopath = path+'/forward_backward-mac'
        elif platform == 'linux' or platform == 'linux2':
                _sopath = path + '/forward_backward-linux'

        libflag = True
        _lib = np.ctypeslib.load_library(_sopath, '.')
except:
        libflag = False


def pdot(a,b):
        """
        Takes dot product along last two dimensions:
        i.e., dot((N,K,D,D),(N,K,D,D)) --> (N,K,1,1)
        """
        return np.einsum('...ij,...jk->...ik',a,b)

def calc_hessian(fxn,x,eps = np.sqrt(np.finfo(np.float64).eps)):
        """
        Finite difference approximation of the Hessian
        #Using Abramowitz & Stegun Eqn. 25.3.23 (on-diagonal), and 25.3.26 (off-diagonal)
        For Laplace Approximation
        """
        # xij is the position to evaluate the function at
        # if i or j = 0, it's the starting postion, 1 or m1 are x + 1.*eps and x - 1.*eps, respetively
        # yij is the function evaluated at xij

        h = np.zeros((x.size,x.size))
        y00 = fxn(x)

        for i in range(x.size):
                for j in range(x.size):
                        #Off-diagonals below the diagonal are the same as those above.
                        if j < i:
                                h[i,j] = h[j,i]
                        else:
                                #On-diagonals
                                if i == j:
                                        x10 = x.copy()
                                        xm10 = x.copy()

                                        x10[i] += eps
                                        xm10[i] -= eps
```

```
                                        y10 = fxn(x10)
                                        ym10 = fxn(xm10)

                                        h[i,j] = eps**(-2.) * (y10 - 2.*y00 + ym10)

                                #Off-diagonals above the diagonal
                                elif j > i:
                                        x11 = x.copy()
                                        x1m1 = x.copy()
                                        xm1m1 = x.copy()
                                        xm11 = x.copy()

                                        x11[i] += eps
                                        x11[j] += eps
                                        x1m1[i] += eps
                                        x1m1[j] -= eps
                                        xm1m1[i] -= eps
                                        xm1m1[j] -= eps
                                        xm11[i] -= eps
                                        xm11[j] += eps

                                        y11 = fxn(x11)
                                        y1m1 = fxn(x1m1)
                                        ym1m1 = fxn(xm1m1)
                                        ym11 = fxn(xm11)

                                        h[i,j] = 1./(4.*eps**2.) * (y11 - y1m1 - ym11 + ym1m1)
        return h

def kmeans(x,nstates,nrestarts=1):
        """
        K-means Clustering
        x is Nxd
        ----
        Returns pi_k, r_nk, mu_k, sig_k
        """

        if x.ndim == 1:
                x = x[:,None]


        jbest = np.inf
        mbest = None
        rbest = None
        for nr in range(nrestarts):
                mu_k = x[np.random.randint(0,x.shape[0],size=nstates)]
                j_last = np.inf
                for i in range(500):
                        dist = np.sqrt(np.sum(np.square(x[:,None,:] - mu_k[None,...]),axis=2))
                        r_nk = (dist == dist.min(1)[:,None]).astype('i')
                        j = (r_nk.astype('f') * dist).sum()
                        mu_k = (r_nk[:,:,None].astype('f') * x[:,None,:]).sum(0) /
                         ↪  (r_nk.astype('f').sum(0)[:,None] + 1e-16)
                        if np.abs(j - j_last)/j <= 1e-100:
                                if j < jbest:
                                        jbest = j
                                        mbest = mu_k
                                        rbest = r_nk
                                break
                        else:
                                j_last = j
        mu_k = mbest
        r_nk = rbest
        sig_k = np.empty((nstates,x.shape[1],x.shape[1]))
```

```
        for k in range(nstates):
                sig_k[k] = np.cov(x[r_nk[:,k]==1.].T)
        pi_k = (r_nk.sum(0)).astype('f')
        pi_k /= pi_k.sum()

        xsort = pi_k.argsort()[::-1]
        #pi_k is fraction, r_nk is responsibilities, mu_k is means, sig_k is variances
        return [pi_k[xsort],r_nk[:,xsort],mu_k[xsort],sig_k[xsort]]

def variational_gmm(x,nstates,maxiter=5000,lowerbound_threshold=1e-20,init_kmeans=0):
        """
        Based on variational Gaussian mixture model from C. Bishop - Chapter 10, Section 2,
        but accounts for covariances of x. x is an n by d array, where n is the number of
        points, and d is the dimensionality of the points. maxiter is the maximum number of
        rounds before stopping if the lowerbound_thershold is not met. This version is
        fully vectorized.
        """
        def calc_lowerbound(nstates, ndim, ntraces, alpha_0, beta_0, nu_0, W_0, Winv_0, m_0, r_nk, N_k,
         ↪  xbar_k, S_k, m_k, W_k, alpha_k, beta_k, nu_k, E_lam, E_pi, E_mulam):
                # Error somewhere?
                eq71 = 0.5 * np.sum( N_k * (E_lam - ndim/beta_k - nu_k *
                 ↪  np.trace(pdot(S_k,W_k),axis1=-2,axis2=-1)
                 ↪  -nu_k*pdot((xbar_k-m_k)[:,None,:],pdot(W_k,(xbar_k-m_k)[:,:,None]))[:,0,0]  -
                 ↪  ndim*np.log(2.*np.pi) ))
                eq72 = np.sum(np.sum(r_nk*E_pi[None,:],axis=1),axis=0)
                eq73 = special.gammaln(alpha_0*nstates) - nstates*special.gammaln(alpha_0) +
                 ↪  (alpha_0-1.)*np.sum(E_pi)
                eq74 = 0.5 * np.sum( ndim*np.log(beta_0/(2.*np.pi)) + E_lam - ndim*beta_0/beta_k - beta_0 *
                 ↪  nu_k * pdot((m_k-m_0)[:,None,:],pdot(W_k,(m_k-m_0)[:,:,None])))
                eq74 += nstates * stats.wishart_ln_B(W_0,np.array((nu_0))) + (nu_0 - ndim -1.)/2. *
                 ↪  np.sum(E_lam) - 0.5 * np.sum(nu_k *
                 ↪  np.trace(pdot(Winv_0[None,...],W_k),axis1=-2,axis2=-1))
                eq75 = np.sum(np.sum(r_nk*np.log(r_nk+1e-300),axis=1),axis=0)
                eq76 = np.sum((alpha_k-1.)*E_pi) + special.gammaln(alpha_k.sum()) -
                 ↪  np.sum(special.gammaln(alpha_k))
                eq77 = 0.5 *np.sum(E_lam + ndim*np.log(beta_k/(2.*np.pi)) - ndim -
                 ↪  2.*stats.wishart_entropy(W_k,nu_k))
                lowerbound =  eq71+eq72+eq73+eq74-eq75-eq76-eq77
                return lowerbound#,[eq71,eq72,eq73,eq74,eq75,eq76,eq77]

        #### Initialize
        ntraces = x.shape[0]
        ndim = x.shape[1]

        alpha_0 = .1
        beta_0 = 1e-20
        nu_0 = ndim + 1.
        W_0 = np.identity(ndim,dtype='f')
        Winv_0 = np.linalg.inv(W_0)
        m_0 = np.zeros((ndim),dtype='f')

        if init_kmeans:
                km = kmeans(x,nstates)
                r_nk = km[1] + .1
                r_nk /= r_nk.sum(1)[:,None]
        else:
                np.random.seed()
                r_nk = np.array([np.random.dirichlet(np.repeat(alpha_0,nstates)) for _ in range(ntraces)])


        N_k = np.zeros((nstates))
        m_k = np.repeat(m_0[None,:],nstates,axis=0)
        W_k = np.repeat(W_0[None,:,:],nstates,axis=0)
        alpha_k = np.repeat(alpha_0,nstates)
```

```
lb = None
state_log = None
finished_counter = 0

E_lam = np.zeros((nstates))
E_pi = np.zeros((nstates))
E_mulam = np.zeros((ntraces,nstates))


##################
### Iterations ###
##################

it = 0
while it < maxiter:

        ###############
        ### M-step ###
        ###############

        N_k = np.sum(r_nk,axis=0)
        xbar_k = np.sum(r_nk[:,:,None]*x[:,None,:],axis=0)/(N_k)[:,None]
        S_k = (np.sum(r_nk[:,:,None,None]*(pdot((x[:,None,:] -
         ↪  xbar_k[None,:,:])[:,:,:,None],(x[:,None,:] -
         ↪  xbar_k[None,:,:])[:,:,None,:])),axis=0)/(N_k)[:,None,None])

        nu_k = nu_0+ N_k
        beta_k = beta_0 + N_k
        alpha_k = alpha_0 + N_k
        m_k = (beta_0*m_0 + N_k[:,None]*xbar_k)/beta_k[:,None]
        Winv_k = Winv_0 + N_k[:,None,None]*S_k + (beta_0*N_k/(beta_0 + N_k))[:,None,None] *
         ↪  pdot((xbar_k - m_0)[:,:,None],(xbar_k - m_0)[:,None,:])
        W_k = np.linalg.inv(Winv_k)

        ###############
        ### E-Step ###
        ###############
        E_lam = np.sum(special.psi((nu_k[:,None] + 1. - np.linspace(1,ndim,ndim)[None,:])/2.),axis=1)
         ↪  + ndim*np.log(2.) + (np.log(np.linalg.det(W_k)))
        E_pi = special.psi(alpha_k) - special.psi(alpha_k.sum())
        E_mulam = ndim/beta_k[None,:] + nu_k[None,:] * pdot((x[:,None,None,:] -
         ↪  m_k[None,:,None,:]),pdot(W_k[None,:,:,:],(x[:,None,:,None] -
         ↪  m_k[None,:,:,None])))[:,:,0,0]

        rho_nk = E_pi[None,:] + .5 * E_lam[None,:] - ndim/2.*np.log(2.*np.pi) - .5*E_mulam
        rho_nk -= rho_nk.max(1)[:,None]
        rho_nk = np.exp(rho_nk)
        r_nk = rho_nk/np.sum(rho_nk,axis=1)[:,None]


        ##################
        ### Calc ELBO ###
        ##################
        elbo =  calc_lowerbound(nstates, ndim, ntraces, alpha_0, beta_0, nu_0, W_0, Winv_0, m_0, r_nk,
         ↪  N_k, xbar_k, S_k, m_k, W_k, alpha_k, beta_k, nu_k, E_lam, E_pi, E_mulam)
        if not np.ndim(lb):
                lb = np.array((it,elbo))[None,:]
        else:
                lb = np.append(lb,np.array((it,elbo))[None,:],axis=0)
        it += 1
        ##A few threshold options: equivalent, rel. change, abs. change
        # print it, lb[-1,1]
        # if it > 1 and lb[-2,1] == lb[-1,1]:
```

```
                if it > 1 and (np.abs((lb[-1,1]-lb[-2,1])/lb[-1,1]) < lowerbound_threshold):
                # if it > 1 and (np.abs((lb[-1,1]-lb[-2,1])) < lowerbound_threshold):
                        break


        xsort = alpha_k.argsort()[::-1]
        return [alpha_k[xsort], r_nk[:,xsort], m_k[xsort], beta_k[xsort], nu_k[xsort], W_k[xsort], lb]

def normalwishart_estep(y,beta,m,nu,W):
        ndim = m.shape[1]
        E_xLx = ndim/beta[None,:] + nu[None,:] * pdot((y[:,None,None,:] -
         ↪  m[None,:,None,:]),pdot(W[None,:,:,:],(y[:,None,:,None] - m[None,:,:,None])))[:,:,0,0]
        E_ln_det_lam = np.sum(special.psi((nu[:,None] + 1. - np.linspace(1,ndim,ndim)[None,:])/2.),axis=1) +
         ↪  ndim*np.log(2.) + (np.log(np.linalg.det(W)))

        E_ln_p_x_z = -ndim/2.*np.log(2.*np.pi) + .5*E_ln_det_lam - E_xLx
        return E_ln_p_x_z

def dirichlet_estep(alpha):
        E_ln_theta = special.psi(alpha) - special.psi(np.sum(alpha,axis=-1))
        return E_ln_theta

def normalwishart_mstep(y,gamma_tk,beta_0,m_0,nu_0,Winv_0):

                N_k = np.sum(gamma_tk,axis=0) + 1e-16
                Y_k = np.sum(gamma_tk[:,:,None]*y[:,None,:],axis=0)/(N_k)[:,None]
                Y2_k = (np.sum(gamma_tk[:,:,None,None]*(pdot((y[:,None,:] -
                 ↪  Y_k[None,:,:])[:,:,:,None],(y[:,None,:] -
                 ↪  Y_k[None,:,:])[:,:,None,:])),axis=0)/(N_k)[:,None,None])

                nu_k = nu_0+ N_k
                beta_k = beta_0 + N_k
                m_k = (beta_0*m_0 + N_k[:,None]*Y_k)/beta_k[:,None]
                Winv_k = Winv_0 + N_k[:,None,None]*Y2_k + (beta_0*N_k/(beta_0 + N_k))[:,None,None] * pdot((Y_k
                 ↪  - m_0)[:,:,None],(Y_k - m_0)[:,None,:])
                W_k = np.linalg.inv(Winv_k)

                return nu_k,W_k,m_k,beta_k

def forward_backward(p_x_z,A,ppi):
        ### p_x_z is TK
        ### A is KK
        ### pi is K

        nstates = A.shape[0]
        npoints = p_x_z.shape[0]

        a = np.zeros((npoints,nstates))
        b = np.zeros((npoints,nstates))
        c = np.zeros((npoints))

        #Forward Pass
        #t = 0
        a[0] = ppi[None,:] * p_x_z[0]
        c[0] = np.sum(a[0])
        a[0] /= c[0]

        # print ppi,p_x_z[0],p_x_z[1]
        # print a[0],c[0]

        #t > 0
        for t in range(1,npoints):
                a[t] = np.dot(a[t-1], A) * p_x_z[t,:]
                c[t] = np.sum(a[t])
```

```
                a[t] /= c[t]


        #Backward pass
        b[-1] = 1.
        for t in range(npoints-1)[::-1]:
                b[t] = np.dot((b[t+1] * p_x_z[t+1]), A.T) / c[t+1]

        #Posteriors
        #Prob for states
        gamma = a*b
        #Joint Prob for transition (for t > 0)
        xi = a[:-1,:,None] * p_x_z[1:,None,:] * A[None,:,:] * b[1:,None,:]  / c[1:,None,None]

        #Evidence
        ln_Z = np.sum(np.log(c))

        return gamma,xi,ln_Z



# int T, int K, double *p_x_z, double *A, double *pi, double *g, double *xi, double *ln_z
_lib.forward_backward.argtypes = [ctypes.c_int, ctypes.c_int, np.ctypeslib.ndpointer(dtype =
 ↪   np.double),np.ctypeslib.ndpointer(dtype = np.double),np.ctypeslib.ndpointer(dtype =
 ↪   np.double),np.ctypeslib.ndpointer(dtype = np.double),np.ctypeslib.ndpointer(dtype =
 ↪   np.double),np.ctypeslib.ndpointer(dtype = np.double)]
_lib.forward_backward.restype  = ctypes.c_void_p

def forward_backward_c(p_x_z,A,ppi,fbmin=1e-300):
        T = p_x_z.shape[0]
        K = A.shape[0]

        fpxz = p_x_z.flatten(order='C') + fbmin
        fA = A.flatten(order='C') + fbmin
        fppi = ppi.flatten(order='C') + fbmin

        gamma = np.empty(T*K,dtype='double')
        xi = np.empty((T-1)*K*K,dtype='double')
        ln_z = np.array((0.))

        _lib.forward_backward(T,K,fpxz,fA,fppi,gamma,xi,ln_z)
        gamma = gamma.reshape((T,K))
        xi = xi.reshape((T-1,K,K))
        ln_z = np.sum(ln_z)
        return gamma,xi,ln_z


def vb_hmm(y,nstates,init_kmeans=0,threshold = 1e-16):
        '''
        Data should be Points x Dimensionality
        ---------
        y = np.loadtxt(...)
        gam_tk,beta_k,m_k,nu_k,W_k,alpha_kl,rho_k,lb = mls.vb_hmm(y,2)

        from ckt_utils.fma import colors
        for j in range(y.shape[1]):
                plt.plot(m_k[gam_tk.argmax(1)][:,j],color = np.array(colors[j])/1.2,lw=1.5,alpha=1.)
                plt.plot(y[:,j],color = colors[j],alpha=.8)
        for xx in np.nonzero(np.roll(gam_tk.argmax(1),-1) - gam_tk.argmax(1))[0]:
                plt.axvline(x = xx,color='k',lw=1.2)
        plt.show()
        '''

        global libflag
```

```
if libflag:
        fb = forward_backward_c
else:
        fb = forward_backward

if y.ndim == 1:
        y = y[:,None]
npoints = y.shape[0]
ndim = y.shape[1]

#TKD
if init_kmeans:
        km = kmeans(y,nstates)
        gam_tk = km[1] + .1
        gam_tk /= gam_tk.sum(1)[:,None]
        rho_0 = km[0]
        rho_k = np.zeros((nstates)) + rho_0
        m_0 = km[2]

else:
        np.random.seed()

        m_0 = np.zeros((ndim),dtype='f') + y.mean(0)
        m_0 = np.repeat(m_0[None,:],nstates,axis=0)
        m_0 += np.random.rand(*m_0.shape)*1e-6

        rho_0 = np.ones(nstates)/nstates
        gam_tk = np.array([np.random.dirichlet(rho_0) for _ in range(npoints)])
        rho_k = rho_0.copy()

#vbFRET Priors
alpha_0 = np.ones((nstates,nstates))
nu_0 = ndim + 1.
nu_0 = np.repeat(nu_0,nstates)
W_0 = np.identity(ndim,dtype='f')
W_0 = np.repeat(W_0[None,:,:],nstates,axis=0)
Winv_0 = np.linalg.inv(W_0)
beta_0 = .25

m_k = m_0.copy()
W_k = W_0.copy()
nu_k = nu_0.copy()
beta_k = np.repeat(beta_0,nstates)
alpha_kl = np.zeros((nstates,nstates)) + alpha_0

lbs = [-np.inf]

#iteration loop
for i in range(1000):

        #Expectation
        E_ln_p_x_z = normalwishart_estep(y,beta_k,m_k,nu_k,W_k)
        E_ln_A_kl = dirichlet_estep(alpha_kl)
        E_ln_pi_k = dirichlet_estep(rho_k)

        # Forward-Backward
        # gam_tk, xi_tkl, ln_z = forward_backward(np.exp(E_ln_p_x_z), np.exp(E_ln_A_kl),
        ↪  np.exp(E_ln_pi_k))
        gam_tk, xi_tkl, ln_z = fb(np.exp(E_ln_p_x_z), np.exp(E_ln_A_kl), np.exp(E_ln_pi_k))

        #Maximization
        nu_k,W_k,m_k,beta_k = normalwishart_mstep(y,gam_tk,beta_0,m_0,nu_0,Winv_0)

        #xi_tkl is t-1 in t dim
```

```
                alpha_kl = alpha_0 + np.sum(xi_tkl,axis=0)
                rho_k = rho_0 + gam_tk[0]

                lowerbound = ln_z - stats.dirichlet_kl(alpha_kl,alpha_0) - stats.dirichlet_kl(rho_k,rho_0) -
                 ↪  stats.normalwishart_kl(beta_k,m_k,nu_k,W_k,beta_0,m_0,nu_0,W_0)
                lbs.append(lowerbound)
                # print i,lowerbound,np.abs((lowerbound - lbs[i])/lbs[i])
                if i > 1 and np.abs((lowerbound - lbs[i])) < threshold:
                        break

        return gam_tk,beta_k,m_k,nu_k,W_k,alpha_kl,rho_k,lbs
```

D.4 Green's Function Kinetics with Numerical Laplace Inversion

```
import numpy as np
np.seterr(all='ignore')

def gsquiggle(s,r):
        return np.linalg.inv(s[:,:,None,None]*np.identity(r.shape[0]) - r[None,None,:,:])

def talbot_inversion_vectorized(fxn,t,m):
        if np.ndim(t) != 1:
                raise Exception("T is the WRONG shape")

        i = np.arange(m).astype('Complex64')
        delta = np.empty(m,dtype='Complex64')
        gamma = np.empty(m,dtype='Complex64')

        delta[1:] = 2.*i[1:]*np.pi/5.*(1./np.tan(i[1:]*np.pi/m)+1.j)
        gamma[1:] = (1.+1.j * (i[1:]*np.pi/m) * (1.+(1./np.tan(i[1:] * np.pi/m))**2.) -
         ↪  1.j/np.tan(i[1:]*np.pi/m)) * np.exp(delta[1:])

        delta[0] = 2.*m/5.
        gamma[0] = .5*np.exp(delta[0])

        f_t = 2./(5.*t[:,None,None]) *
         ↪  np.sum(np.real(gamma[None,:,None,None]*fxn(delta[None,:]/t[:,None])),axis=1)
        return f_t

def greensfxn_kinetics(t,rate_matrix,m=64.):
        '''
        t is an array of times to calculate G_t
        rate_matrix must be square and dtype='float64'

        Returns G(t) for P(t) = G(t)*P(t=0)
        '''
        return talbot_inversion_vectorized(lambda s: gsquiggle(s,rate_matrix.T),t,m)
```

D.5 BIASD

BIASD Integrand .C

```
#include <math.h>

// Chebyshev Polynomial Evaluation and Modified Bessel Function Evaluations from GNU Science Library --
 ↪  http://www.gnu.org/software/gsl/
// Those have the GNU General Public License (GPL)
// From GSL's cheb_eval_e.c, bessel_I0.c, and bessel_I1.c with some hard-coded values
```

```
/* ----------------------------------------- */
/* ----------Chebyshev Polynomials---------- */
/* ----------------------------------------- */

typedef struct
{
  double * c;   /* coefficients  c[0] .. c[order] */
  int order;    /* order of expansion             */
  double a;     /* lower interval point           */
  double b;     /* upper interval point           */
} cheb_series;

static double bi0_data[12] = {
  -.07660547252839144951,
  1.92733795399380827000,
   .22826445869203013390,
   .01304891466707290428,
   .00043442709008164874,
   .00000942265768600193,
   .00000014340062895106,
   .00000000161384906966,
   .00000000001396650044,
   .00000000000009579451,
   .00000000000000053339,
   .00000000000000000245
};
static cheb_series bi0_cs = {bi0_data, 11, -1, 1};

static double ai0_data[21] = {
   .07575994494023796,
   .00759138081082334,
   .00041531313389237,
   .00001070076463439,
  -.00000790117997921,
  -.00000078261435014,
   .00000027838499429,
   .00000000825247260,
  -.00000001204463945,
   .00000000155964859,
   .00000000022925563,
  -.00000000011916228,
   .00000000001757854,
   .00000000000112822,
  -.00000000000114684,
   .00000000000027155,
  -.00000000000002415,
  -.00000000000000608,
   .00000000000000314,
  -.00000000000000071,
   .00000000000000007
};
static cheb_series ai0_cs = {ai0_data, 20, -1, 1};

static double ai02_data[22] = {
   .05449041101410882,
   .00336911647825569,
   .00006889758346918,
   .00000289137052082,
   .00000020489185893,
   .00000002266668991,
   .00000000339623203,
   .00000000049406022,
   .00000000001188914,
```

```
  -.00000000003149915,
  -.00000000001321580,
  -.00000000000179419,
   .00000000000071801,
   .00000000000038529,
   .00000000000001539,
  -.00000000000004151,
  -.00000000000000954,
   .00000000000000382,
   .00000000000000176,
  -.00000000000000034,
  -.00000000000000027,
   .00000000000000003
};
static cheb_series ai02_cs = {ai02_data, 21, -1, 1};

static double bi1_data[11] = {
  -0.001971713261099859,
   0.407348876675464810,
   0.034838994299959456,
   0.001545394556300123,
   0.000041888521098377,
   0.000000764902676483,
   0.000000010042493924,
   0.000000000099322077,
   0.000000000000766380,
   0.000000000000004741,
   0.000000000000000024
};
static cheb_series bi1_cs = {bi1_data, 10, -1, 1};

static double ai1_data[21] = {
  -0.02846744181881479,
  -0.01922953231443221,
  -0.00061151858579437,
  -0.00002069971253350,
   0.00000858561914581,
   0.00000104949824671,
  -0.00000029183389184,
  -0.00000001559378146,
   0.00000001318012367,
  -0.00000000144842341,
  -0.00000000029085122,
   0.00000000012663889,
  -0.00000000001664947,
  -0.00000000000166665,
   0.00000000000124260,
  -0.00000000000027315,
   0.00000000000002023,
   0.00000000000000730,
  -0.00000000000000333,
   0.00000000000000071,
  -0.00000000000000006
};
static cheb_series ai1_cs = { ai1_data, 20, -1, 1};

static double ai12_data[22] = {
   0.02857623501828014,
  -0.00976109749136147,
  -0.00011058893876263,
  -0.00000388256480887,
  -0.00000025122362377,
  -0.00000002631468847,
  -0.00000000383538039,
```

```
  -0.00000000055897433,
  -0.00000000001897495,
   0.00000000003252602,
   0.00000000001412580,
   0.00000000000203564,
  -0.00000000000071985,
  -0.00000000000040836,
  -0.00000000000002101,
   0.00000000000004273,
   0.00000000000001041,
  -0.00000000000000382,
  -0.00000000000000186,
   0.00000000000000033,
   0.00000000000000028,
  -0.00000000000000003
};
static cheb_series ai12_cs = {ai12_data, 21, -1, 1};


/* ----------------------------------------- */
/* ------------ Bessel Functions------------ */
/* ----------------------------------------- */

double cheb_eval(cheb_series * cs, double x) {
        int j;
        double d  = 0.0;
        double dd = 0.0;

        double y  = (2.0*x - cs->a - cs->b) / (cs->b - cs->a);
        double y2 = 2.0 * y;

        double e = 0.0;

        for(j = cs->order; j>=1; j--) {
                double temp = d;
                d = y2*d - dd + cs->c[j];
                e += fabs(y2*temp) + fabs(dd) + fabs(cs->c[j]);
                dd = temp;
        }

        double temp = d;
        d = y*d - dd + 0.5 * cs->c[0];
        e += fabs(y*temp) + fabs(dd) + 0.5 * fabs(cs->c[0]);
        return d;
}

double i0_scaled(double x) {
        double y = fabs(x);

        if(y < 2.0 * 1.490116e-08) {
                return 1.0 - y;
        }
        else if(y <= 3.0) {
                return exp(-y) * (2.75 + cheb_eval(&bi0_cs, y*y/4.5-1.0));
        }
        else if(y <= 8.0) {
                return (0.375 + cheb_eval(&ai0_cs, (48.0/y-11.0)/5.0))/ sqrt(y);
        }
        else {
                return (0.375 + cheb_eval(&ai02_cs, 16.0/y-1.0))/ sqrt(y);
        }
}

double i0(double x) {
```

```
        double y = fabs(x);

        if(y < 2.0 * 1.490116e-08) {
                return 1.0;
        }
        else if(y <= 3.0) {
                return 2.75 + cheb_eval(&bi0_cs, y*y/4.5-1.0);
        }
        else if(y < 7.097827e+02 - 1.0) {
                return exp(y) * i0_scaled(x);
        }
        else {
                return NAN;
        }
}

double i1_scaled(double x) {
        const double xmin    = 4.450148e-308;
        const double x_small =  4.214685e-08;
        double y = fabs(x);

        if(y == 0.0) {
                return 0.0;
        }
        else if(y < x_small) {
                return 0.5*x;
        }
        else if(y <= 3.0) {
                return x * exp(-y) * (0.875 + cheb_eval(&bi1_cs, y*y/4.5-1.0));
        }
        else if(y <= 8.0) {
                double b;
                double s;
                b = (0.375 + cheb_eval(&ai1_cs, (48.0/y-11.0)/5.0)) / sqrt(y);
                s = (x > 0.0 ? 1.0 : -1.0);
                return s * b;
        }
        else {
                double b;
                double s;
                b = (0.375 + cheb_eval(&ai12_cs, 16.0/y-1.0)) / sqrt(y);
                s = (x > 0.0 ? 1.0 : -1.0);
                return s * b;
        }
}

double i1(double x) {
        const double xmin    = 4.450148e-308;
        const double x_small =  4.214685e-08;
        double y = fabs(x);

        if(y == 0.0) {
            return 0.0;
        }
        else if(y < xmin) {
                return NAN;
        }
        else if(y < x_small) {
                return 0.5*x;
        }
        else if(y <= 3.0) {
                return x * (0.875 + cheb_eval(&bi1_cs, y*y/4.5-1.0));
        }
        else if(y < 7.097827e+02) {
```

```
                return exp(y) * i1_scaled(x);
        }
        else {
                return NAN;
        }
}



/*
#########################################################
        Integrand for BIASD
#########################################################
*/


double integrand(int n, double args[n]) {
        double f = args[0];
        double d = args[1];
        double ep1 = args[2];
        double ep2 = args[3];
        double sigma = args[4];
        double k1 = args[5];
        double k2 = args[6];
        double tau = args[7];

        double out;

        //Pre-calulate some parameters
        double k = k1 + k2;
        double p1 = k2/k;
        double p2 = k1/k;
        double y = 2.*k*tau * pow(p1*p2*f*(1.-f),.5);
        double z = p2*f + p1*(1.-f);

        //Enforce parameter support
        if (f < 0. || f > 1. || k1 <= 0. || k2 <= 0. || sigma <= 0. || tau <= 0. || ep1 >= ep2) {
                out = 0.;
        }
        else {
                //Time averaging part
                out = 2.*k*tau*p1*p2*(i0(y)+k*tau*(1.-z)*i1(y)/y)*exp(-z*k*tau);
                //Observation noise part
                out *= 1./sigma * M_2_SQRTPI/2. * M_SQRT1_2 *
                 ↪   exp(-.5/sigma/sigma*pow(d-(ep1*f+ep2*(1.-f)),2.));
        }
        return out;
}

// To Compile in Linux for CTYPES in Python:
// Note: This doesn't need GSL b/c it is the same fxns
// gcc -L/usr/local/lib -shared -o biasd_integrand-linux.so -fPIC -O3 biasd_integrand_gsl.c -lm
```

BIASD Likelihood and Python Wrappers for .C Integrand

```
################## Versions#########
# Anaconda                - 2.3.0
## Python                - 2.7.10
### Numpy                 - 1.9.2
### Scipy                 - 0.16.0
###################################
```

```
### Imports
import numpy as np
from scipy import special as _special
from scipy.integrate import quad as _quad

from os import path as _path
from sys import platform as _platform
import ctypes as _ctypes

### Python Integrand
def _python_integrand(x,d,e1,e2,sigma,k1,k2,tau):
        """
        Integrand for BIASD likelihood function
        """
        #Ensures proper support
        if x < 0. or x > 1. or k1 <= 0. or k2 <= 0. or sigma <= 0. or tau <= 0. or e1 >= e2:
                return 0.
        else:
                k = k1 + k2
                p1 = k2/k
                p2 = k1 /k
                y = 2.*k*tau * np.sqrt(p1*p2*x*(1.-x))
                z = p2*x + p1*(1.-x)
                pf = 2.*k*tau*p1*p2*(_special.i0(y)+k*tau*(1.-z)*_special.i1(y)/y)*np.exp(-z*k*tau)
                py = 1./np.sqrt(2.*np.pi*sigma**2.)*np.exp(-.5/sigma/sigma*(d-(e1*x+e2*(1.-x)))**2.) * pf
                return py

### Attempt to Load the .C Integrand
try:
        _libpath = _path.dirname(_path.abspath(__file__))
        if _platform == 'darwin':
                _sopath = _libpath+'/biasd_integrand-mac'
        elif _platform == 'linux' or _platform == 'linux2':
                _sopath = _libpath + '/biasd_integrand-linux'
        _lib = np.ctypeslib.load_library(_sopath, '.')

        #Setup the integrand function calls/returns
        _c_integrand = _lib.integrand
        _c_integrand.restype = _ctypes.c_double
        _c_integrand.argtypes = (_ctypes.c_int, _ctypes.c_double)

        integrand = _c_integrand

        _clibflag = True
        print "Loaded integrand written in C"
        print _sopath

### Failure: Load the Python Integrand
except:
        integrand = _python_integrand
        _clibflag = False
        print "Couldn't find the compiled library"
        print "Using integrand written in Python \n    This will be much slower!"

### Define important integral in likelihood function
def integral(d,e1,e2,sigma,k1,k2,tau,):
        """
        Use Gaussian quadrature to integrate the BIASD integrand across df between f = [0 ... 1]
        """

        return _quad(integrand,0.,1.,args=(d,e1,e2,sigma,k1,k2,tau),limit=1000)[0]
integral = np.vectorize(integral)

def pdf_gaussian(x,mu,std):
```

```
        return (1./(std*np.sqrt(2.*np.pi))*np.exp(-.5/std/std*(x-mu)**2.))

def log_likelihood(theta,d,tau):
        """
        Calculate the log of the BIASD likelihood function at theta
        using the data d
        given the time period of the d as tau
        """
        e1,e2,sigma,k1,k2 = theta
        p1 = k2/(k1+k2)
        p2 = 1.-p1
        out = integral(d,e1,e2,sigma,k1,k2,tau)
        peak1 = pdf_gaussian(d,e1,sigma)
        peak2 = pdf_gaussian(d,e2,sigma)
        out += p1*peak1*np.exp(-k1*tau)
        out += p2*peak2*np.exp(-k2*tau)

        # Avoid use of -infinity
        return np.log(out+1e-300)

def log_posterior(d,priors,tau,theta):
        """"
        Calculate the ln of the posterior probability distribution for data d,
        with priors as priors, time period of d as tau, at the point theta
        """

        return log_likelihood(theta,d,tau).sum()+priors.sum_log_pdf(theta)

def test_speed(n):
        """
        Test how fast the BIASD integral (python-based or C-based) runs.
        C-based should be ~30 us. Python-based is ~30x that.
        """

        from time import time
        d = np.linspace(-2,1.2,5000)
        t0 = time()
        for i in range(n):
                # _quad(integrand,0.,1.,args=(.1,0.,1.,.05,3.,8.,.1))[0]
                y = log_likelihood([0.,1.,.05,3.,8.],d,.1).sum()
        t1 = time()
        print "Total time for "+str(n)+" runs: ",np.around(t1-t0,4)," (s)"
        print 'Average speed: ', np.around((t1-t0)/n/d.size*1.e6,4),' (usec/datapoint)'
```

BIASD Classes

```
import numpy as np
np.set_printoptions(precision=4,linewidth=180)
np.seterr(all='ignore')
eps = np.finfo(float).eps
from scipy import special
from scipy.optimize import minimize
import cPickle as pickle

from ckt_utils import stats,mls,parallel
from biasd_likelihood import *

class dist:
        """
        Represents a probability distribution
        distribution is a string for the name of the distribution
        p1, and p2 are the parameters for that distribution
```

```
        """

        def __init__(self,distribution,p1,p2):
                self.type = distribution.lower()
                self.p1 = p1
                self.p2 = p2

                #Make distribution name types uniformily stored
                if self.type in ['uniform','u',0]:
                        self.type = 'uniform'
                elif self.type in ['normal','norm','n','gaussian','gauss',1]:
                        self.type = 'normal'
                elif self.type in ['beta','b',2]:
                        self.type = 'beta'
                elif self.type in ['gamma','g',3]:
                        self.type = 'gamma'

                #Check to see if this makes a valid distribution
                self.good = self.good_check()

        def good_check(self):
                """
                Check if parameters are withing support range and return 1 if so
                """
                if self.type == 'uniform':
                        if np.isfinite(self.p1) and np.isfinite(self.p2) and self.p1 < self.p2:
                                return 1
                elif self.type == 'normal':
                        if np.isfinite(self.p1) and self.p2 > 0.:
                                return 1
                elif self.type == 'beta' or self.type == 'gamma':
                        if self.p1 > 0. and self.p2 > 0.:
                                return 1
                return 0

        def pdf(self,x):
                """
                Return the probability distribution function
                x can be a vector
                """
                if self.good:
                        if self.type == "uniform":
                                pdffxn = stats.p_uniform(x,self.p1,self.p2)
                        elif self.type == "normal":
                                pdffxn = stats.p_gauss(x,self.p1,self.p2)
                        elif self.type == "beta":
                                pdffxn = stats.p_beta(x,self.p1,self.p2)
                        elif self.type == "gamma":
                                pdffxn = stats.p_gamma(x,self.p1,self.p2)
                        return pdffxn
                return 0

        def logpdf(self,x):
                """
                Return the log of the probability distribution function
                """
                return np.log(self.pdf(x))

        def mean(self):
                """
                Calculate E[x]
                """
                if self.type == "uniform":
                        mean = stats.mean_uniform(self.p1,self.p2)
```

```
                elif self.type == "normal":
                        mean = stats.mean_gauss(self.p1,self.p2)
                elif self.type == "beta":
                        mean = stats.mean_beta(self.p1,self.p2)
                elif self.type == "gamma":
                        mean = stats.mean_gamma(self.p1,self.p2)
                return mean

        def var(self):
                """
                Calculate E[x^2] - E[x]^2
                """
                if self.type == "uniform":
                        var = stats.var_uniform(self.p1,self.p2)
                elif self.type == "normal":
                        var = stats.var_gauss(self.p1,self.p2)
                elif self.type == "beta":
                        var = stats.var_beta(self.p1,self.p2)
                elif self.type == "gamma":
                        var = stats.var_gamma(self.p1,self.p2)
                return var

        def random(self,size_rvs):
                """
                Generate random numbers in shape of size_rvs
                """
                #At least correct for numpy 1.9.2
                np.random.seed()
                if self.type == "uniform":
                        rvs = np.random.uniform(self.p1,self.p2,size_rvs)
                elif self.type == "normal":
                        rvs = np.random.normal(self.p1,self.p2,size_rvs)
                elif self.type == "beta":
                        rvs = np.random.beta(self.p1,self.p2,size_rvs)
                elif self.type == "gamma":
                        rvs = np.random.gamma(shape=self.p1,scale=1./self.p2,size=size_rvs)
                return rvs

class biasddistribution:
        """
        Stores the five parameter probability distribution functions used for the BIASD \Theta.
        \Theta = [\epsilon_1, \epsilon_2, \sigma, \k_1, \k_2]
        This is used for both the prior and the posterior probability distributions
        """
        def __init__(self,e1,e2,sigma,k1,k2):
                self.names = ['e1','e2','sigma','k1','k2']
                self.e1 = e1
                self.e2 = e2
                self.sigma = sigma
                self.k1 = k1
                self.k2 = k2
                self.list = [self.e1,self.e2,self.sigma,self.k1,self.k2]

                #Ensure each distribution in \Theta is sound
                self.complete = self.test_distributions()
                if self.complete != 1:
                        print "Distributions are incomplete"

        def test_distributions(self):
                """
                If all distributions in \Theta are correct and return 1
                """
                good = 1
                for dists in self.list:
```

```
                        try:
                                if dists.good != 1:
                                        good = 0

                        except:
                                good = 0
                return good

        def get_dist_means(self):
                """
                Calculate means of \Theta
                """
                if self.complete == 1:
                        return np.array((self.e1.mean(), self.e2.mean(), self.sigma.mean(), self.k1.mean(),
                         ↪  self.k2.mean()))
                else:
                        print "Distributons are incomplete"
                        return np.repeat(np.nan,5)

        def get_dist_vars(self):
                """
                Calculate the variances of \Theta
                """
                if self.complete == 1:
                        return np.array((self.e1.var(), self.e2.var(), self.sigma.var(), self.k1.var(),
                         ↪  self.k2.var()))
                else:
                        print "Distributons are incomplete"
                        return np.repeat(np.nan,5)

        def sum_log_pdf(self,theta):
                """
                Returns \Sum_i \ln p\left( \theta_i \right) evaluated at theta (list of numpy array)
                """
                if self.complete == 1:
                        ll = 0
                        for theta_i,distribution in zip(theta,self.list):
                                ll += np.log(distribution.pdf(theta_i))
                        return ll
                else:
                        print "Distributions are incomplete"
                        return np.nan

        def which_bad(self):
                """
                Figure out which of the distributions is bad
                """
                if self.complete != 1:
                        print "Bad Distributions:"
                        baddists=[]
                        for i,j in zip([dists.good for dists in self.list],self.names):
                                if i != 1:
                                        print j
                                        baddists.append(j)
                        return baddists
                else:
                        print "All distributions seem complete"
                        return None

        def random_theta(self):
                """
                Generate a random [\epsilon_1, \epsion_2, \sigma, k_1, k_2] from the BIASD distributions
                """
                theta = np.repeat(np.nan,5)
```

```
                if self.complete == 1:
                        #Try a max of 100 times
                        for i in range(100):
                                for j,distribution in zip(range(5),self.list):
                                        theta[j] = distribution.random(1)
                                #Enforce conditions \epsilon_1 < \epsilon_2, and others are > 0
                                if theta[0] < theta[1] and theta[2] > 0. and theta[3] > 0. and theta[4] > 0.:
                                        break
                                theta = np.repeat(np.nan,5)
                return theta

class laplace_posterior:
        """
        Holds the results of a laplace approximation of the posterior probability distribution from BIASD
        """
        def __init__(self,means,covars):
                self.mu = means
                self.covar = covars

class trace:
        """
        A trace is a signal versus time trajectory loaded in from a dataset.
        It can be processed using the Laplace approximation.
        """

        def __init__(self, data, tau=None, prior=None,identity=0):
                #Signal versus time data as a numpy array
                self.data = data.flatten()
                #time period
                self.tau = float(tau)
                #Prior as a biasddistributions object
                self.prior = prior
                #Posterior will be a biasddistributions objet
                self.posterior = None
                #Identifying number
                self.identity = identity

        def log_likelihood(self,theta):
                '''
                Calculate the log of the BIASD likelihood function at theta
                for this trace
                '''
                return log_likelihood(theta,self.data,self.tau)

        def log_posterior(self,theta):
                '''
                Calculate the log of the BIASD posterior at theta for this trace
                '''
                return log_posterior(self.data,self.prior,self.tau,theta)

        def find_map(self,meth='nelder-mead',xx=None,nrestarts=2):
                '''
                Use numerical minimization to find the maximum a posteriori estimate of the posterior
                Provide xx to force first theta initialization at that theta
                meth is the method used by the minimizer - default to simplex
                nrestarts is the number of restarts for the MAP
                '''

                ylist = []

                #If no xx, start at the mean of the priors
                if type(xx).__name__ != 'ndarray':
                        xx = self.prior.get_dist_means()
```

```
                #Rounds to minimze the -log posterior
                ylist.append(minimize(lambda theta: -1.*self.log_posterior(theta),x0=xx,method=meth))
                for i in range(nrestarts):
                        #Try a random location consistent with the prior.
                        xx = self.prior.random_theta()
                        ylist.append(minimize(lambda theta: -1.*self.log_posterior(theta),x0=xx,method=meth))

                #Select the best MAP estimate
                ymin = np.inf
                for i in ylist:
                        if i['success']:
                                if i['fun'] < ymin:
                                        ymin = i['fun']
                                        y = i
                #If no MAP estimates, return None
                if ymin == np.inf:
                        y = None
                return y

        def laplace_approximation(self):
                '''
                Perform the Laplace approximation on the posterior probability distribution of this trace
                '''

                def sanitize_uniform(x,p):
                        '''
                        Make sure you can calculate the Hessian if \theta_i is at
                        the edge of a uniform distribution
                        '''
                        for i in range(5):
                                if p.list[i].type == 'uniform':
                                        if np.abs(p.list[i].p1 - x[i]) < 1e-5:
                                                x[i] += 1e-5
                                        elif np.abs(p.list[i].p2 - x[i]) < 1e-5:
                                                x[i] -= 1e-5
                        return x

                #Calculate the best MAP estimate
                mind = self.find_map()
                if not mind is None:
                        #Calculate the Hessian at MAP estimate
                        if mind['success']:
                                mu = sanitize_uniform(mind['x'],self.prior)
                                feps = np.sqrt(np.finfo(np.float).eps)
                                hessian = mls.calc_hessian(self.log_posterior,mu,eps=feps)

                                #Ensure that the hessian is positive semi-definite by checking that all
                                 ↪  eigenvalues are positive
                                #If not, expand the value of machine error in the hessian calculation and try
                                 ↪  again
                                try:
                                        #Check eigenvalues
                                        while np.any(np.linalg.eig(-hessian)[0] <= 0.):
                                                feps *= 2.
                                                #Calculate hessian
                                                hessian = mls.calc_hessian(self.log_posterior,mu,eps=feps)
                                        #Invert hessian to get the covariance matrix
                                        var = np.linalg.inv(-hessian)
                                        #Ensure symmetry of covariance matrix if witin machine error
                                        if np.allclose(var,var.T):
                                                var = np.tri(5,5,-1)*var+(np.tri(5,5)*var).T
                                                return (self.identity,laplace_posterior(mu,var))

                                #If this didn't work, return None
```

```
                                except np.linalg.LinAlgError:
                                        return (self.identity,None)
                return (self.identity,None)

class dataset:
        '''
        A dataset is a collection of traces(i.e., all of the signal
        vs. time trajectories from a particular experiment).
        '''
        def __init__(self,data_fname=None,fmt='2D-NxT',tau=None,temperature = 25., title = None, prior =
        ↪  None,analysis_fname=None):
                #Location of the file containing the signal vs. time trajectories
                self.data_fname = data_fname
                #Data format of the signal vs. time trajectories ('2D-NxT','2D-TxN','1D')
                self.fmt = fmt
                #Time period
                self.tau =tau
                #Temperature
                self.temperature = temperature
                #Random Notes from GUI are stored here
                self.title = title
                #prior for all traces
                self.prior = prior
                #List of traces to make accessing them easier
                self.traces = []
                #File name where this dataset will be/already is saved so that it can be loaded/saved later
                self.analysis_fname = analysis_fname
                #Best ensemble parameters for the ensemble of signal vs. time trajectories.
                #Calculated from the variational GMM of the Laplace approximated BIASD posteriors
                self.ensemble_result = None

        @staticmethod
        def _convert_2D_to_1D(trace_matrix):
                '''
                Convert NxT (with NaN\'s or inf\'s for no data) to 2x(N*T) with labels format
                '''
                if trace_matrix.ndim == 1:
                        trace_matrix = trace_matrix[None,:]
                identities = np.array([])
                traces = np.array([])
                for i in range(trace_matrix.shape[0]):
                        l = np.isfinite(trace_matrix[i]).sum()
                        identities = np.append(identities,np.repeat(i,l))
                        traces = np.append(traces,trace_matrix[i,:l])
                return np.array((identities,traces))

        @staticmethod
        def _convert_1D_to_2D(trace_matrix):
                '''
                Convert 2x(N*T) with labels format to NxT format (with NaN\'s for no data)
                '''

                ns = np.unique(trace_matrix[0])
                sizes = (trace_matrix[0][None,:] == ns[:,None]).sum(1)
                traces_out = np.ones((ns.size,sizes.max()))
                traces_out[:,:] = np.nan

                for i in range(ns.size):
                        d = trace_matrix[1][trace_matrix[0] == ns[i]]
                        traces_out[i,:sizes[i]] = d
                return traces_out


        def load_data(self):
```

```
                '''
                Creates traces from the signal vs. time trajectories in the file specified by data_fname.
                Stores the signal vs. time data in the '1D' format.
                '''

                if type(self.data_fname) == str:
                        if path.isfile(self.data_fname):
                                try:
                                        self.data = np.loadtxt(self.data_fname)
                                        if self.fmt == '2D-TxN':
                                                self.data = self.data.T
                                        if self.fmt != '1D':
                                                self.data = self._convert_2D_to_1D(self.data)
                                        if self.fmt == '1D':
                                                self.data =
                                                 ↪  self._convert_1D_to_2D(self._convert_2D_to_1D(self.data))
                                except:
                                        self.data = None
                                        print "Couldn't load "+self.data_fname
                                        return
                        else:
                                print self.data_fname + " isn't a file"
                                return

                        #Construct traces
                        ns = np.unique(self.data[0])
                        for i in range(ns.size):
                                self.traces.append(trace(self.data[1][self.data[0]==ns[i]], tau=self.tau,
                                 ↪  prior=self.prior, identity=i))
                        self.n_traces = len(self.traces)


        def save_analysis(self):
                '''
                Allows the entire dataset to be saved using pickle to the
                filename specified in analysis_fname.
                Everything including traces, laplace posteriors, and ensembles
                will be saved.
                '''
                if self.analysis_fname:
                        f = open(self.analysis_fname,'wb')
                        pickle.dump(self.__dict__,f,2)
                        f.close()

        def load_analysis(self):
                '''
                Loads the entire dataset that was save to the file specified by the
                filename in analysis_fname.
                Everything including traces, laplace posteriors, and ensembles
                will be loaded.
                '''
                if path.isfile(self.analysis_fname):
                        f = open(self.analysis_fname,'rb')
                        tmp_dict = pickle.load(f)
                        f.close()
                        self.__dict__.update(tmp_dict)
                else:
                        print "No file called ",self.analysis_fname

        def update(self):
                '''
                Changes to the priors or time period specified in the dataset will
                be updated to the traces in the dataset.
                '''
```

```
                for tracei in self.traces:
                        tracei.prior = self.prior
                        tracei.tau = self.tau

        def run_laplace(self,nproc=1):
                '''
                Calculates the Laplace approximation of the posterior probability
                distributions for the traces loaded into this dataset.
                If nproc is > 1, then this function will use multiple processors to
                perform this calculation on traces simultaneously.
                '''
                # if nproc > mp.cpu_count():
                #         print "Using max number of CPUs: "+str(mp.cpu_count())
                #         nproc = mp.cpu_count()

                j = 1
                print "----------------\nLaplace Approximations"
                # t0 = time.time()

                def laplace_wrapper(tracei):
                        item = tracei.laplace_approximation()
                        tracei.posterior = item[1]
                        return tracei

                laplace_result = parallel.embarrassingly(laplace_wrapper,self.traces,nproc =
                 ↪  nproc,batchsize=100)
                self.traces = laplace_result

#Example biasddistributions
prior_personal_distribution = biasddistribution(
dist('normal',0.15,.1),
dist('normal',.75,.1),
dist('gamma',70.,1000.),
dist('Gamma',2.,2./10.),
dist('Gamma',2.,2./10.))
```

Embarassingly Parallel Function Wrapper

```
import multiprocessing as _mp
# import numpy as np
import statusbar as _sb
from time import time as _time

def check_cpu(nproc):
        cpucount = _mp.cpu_count()
        if nproc > cpucount:
                print "Using max number of CPUs: "+str(cpucount)
                nproc = cpucount
        return nproc

class _worker(_mp.Process):
        """
        To parallel process a generic queue containing inputs given the provided function
        """
        def __init__(self,queue_in,queue_out,function):
                self.__queue_in = queue_in
                self.__queue_out = queue_out
                self.function = function
                _mp.Process.__init__(self)

        def run(self):
```

```
                while 1:
                        #Check the queue
                        item = self.__queue_in.get()

                        #If the queue is done
                        if item is None:
                                #add a None to watcher queue so it knows this is done.
                                self.__queue_out.put(item)
                                break

                        #Calculate the function of the item from the input queue and pass it to the watcher
                         ↪  queue
                        self.__queue_out.put([item[0],self.function(item[1])])

class _watcher(_mp.Process):
        """
        This allows the queue used to parallel process to be observed in real-time
        """
        def __init__(self,queue_in,queue_out,numworkers,numjobs,batchnum = None):
                self.__queue_out = queue_out
                self.__queue_in = queue_in
                self.numworkers = numworkers
                self.numjobs = numjobs
                self.batchnum = batchnum
                _mp.Process.__init__(self)

        def run(self):
                pillcount = 0
                self.resultcount = 0

                #Initialze the progress bar
                if self.batchnum is None:
                        message = "Processing"
                else:
                        message = "Batch "+str(self.batchnum+1)
                bar = _sb.percent(self.numjobs,message=message)

                while pillcount < self.numworkers:
                        item = self.__queue_in.get()
                        #If a worker is completely done
                        if item is None:
                                pillcount += 1
                        #If the worker provided a result
                        else:
                                #Return the result to be processed by something else
                                self.__queue_out.put(item)
                                self.resultcount += 1
                                #Update the progress bar
                                bar.next()

def embarrassingly(function, item_list,nproc=_mp.cpu_count(), batchsize=None,timer=True):
        '''
        Embarrassingly parallel function wrapper using python's multiprocessing.
        Inputs:
                * function - A function that acts on the elements in item_list and
                returns either a value or a class or something.
                * item_list - A list containing the argument/sets of arguments to be
                embarrassingly parallelly calculated by function e.g. [[.17, 'b'], [.42, 'a'], ...]
                * nproc - Number of processors to use. It is capped to cpu count.
                * batchsize - Approximate number of items from the list to calculate
                before reinitializing everything. Multiprocessing can sometimes eat too
                much memory if using classes for the calculation and it might freeze.
                If None, no batches will be used.
                * timer - If True, the total time the calculation took will be printed
```

```
                at the end.
        Output:
                * results - A list where each element corresponds to the output of the
                corresponding item in item_list when acted on by function.
        '''

        nproc = check_cpu(nproc)
        t0 = _time()

        listsize = len(item_list)
        results = [None]*listsize

        if not batchsize is None:
                while batchsize % nproc != 0:
                        batchsize -= 1
                batch_num = int((listsize-1)/batchsize) + 1
                batches = range(batch_num)
        else:
                batch_num = 1
                batches = [None]

        for batch_index in batches:
                proc_list = item_list[batch_index::batch_num]
                #Setup the queues for the multiprocessing workers
                queue_work = _mp.Queue(nproc)
                queue_out = _mp.Queue()
                queue_results = _mp.Queue()

                #Start the multiprocessing workers and the watcher
                workers = []
                for _ in range(nproc):
                        worker = _worker(queue_work,queue_out,function)
                        worker.start()
                        workers.append(worker)
                watcher = _watcher(queue_out,queue_results,nproc,len(proc_list),batch_index)
                watcher.start()

                #Begin by sending the traces to the queue
                idcounter = 0
                for item in proc_list:
                        queue_work.put([idcounter,item])
                        idcounter += 1

                #Add Poison Pills to the queue so that the workers know when they're done.
                for _ in workers:
                        queue_work.put(None)

                #Wait for the calculations to finish
                for worker in workers:
                        worker.join()

                #Collect the results from this batch
                batch_results = [None]*len(proc_list)
                for _ in range(len(proc_list)):
                        y = queue_results.get()
                        batch_results[y[0]] = y[1]
                results[batch_index::batch_num] = batch_results
                watcher.join()
        t1 = _time()
        if timer:
                print "Time: ",t1 - t0,"\n"
        return results
```

```
class _testclass:
        def __init__(self,x):
                self.x = x
        def flip(self):
                self.x = -self.x

def test(nproc):
        def function_regular(c):
                return -c
        def function_class(c):
                c.flip()
                return c

        regular_list = range(1000)
        class_list = []
        for n in range(len(regular_list)):
                class_list.append(_testclass(n))

        print "Function Approach Test"
        results_regular = embarrassingly(function_regular,regular_list,nproc = nproc,batchsize=None)
        print "Class Approach Test"
        results_class = embarrassingly(function_class,class_list,nproc = nproc,batchsize=200)

        correct_regular = 0
        correct_class = 0
        for i in range(len(results_class)):
                if results_class[i].x == - class_list[i].x:
                        correct_class += 1
                if results_regular[i] == - regular_list[i]:
                        correct_regular += 1

        print "Regular: ",correct_regular,"/",len(regular_list)
        print "Classes: ",correct_class,"/",len(class_list)
```