Title
Fluctuation Theories in Model Photosynthetic Systems

Permalink
https://escholarship.org/uc/item/8jw008kn

Author
Haberstroh, John Richard

Publication Date
2017

Peer reviewed|Thesis/dissertation
Fluctuation Theories in Model Photosynthetic Systems

by

John R Haberstroh

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

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Phillip Geissler, Co-chair
Associate Professor Oskar Hallatschek, Co-chair
Professor Eran Rabani
Professor Naomi Ginsberg

Fall 2017
Fluctuation Theories in Model Photosynthetic Systems

Copyright 2017
by
John R Haberstroh
Abstract

Fluctuation Theories in Model Photosynthetic Systems

by

John R Haberstroh

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Phillip Geissler, Co-chair

Associate Professor Oskar Hallatschek, Co-chair

Electronic energy transfer (EET) is the physical process responsible for directing and trapping photonic energy in the majority of autotrophic life. These autotrophs allow make life on earth possible by providing dense chemical energy sources for higher organisms. Over evolutionary time scales, autotrophs have evolved complex systems that robustly and resiliently convert molecular excitations into stored chemical energy. We will analyze one such system in depth: the Fenna-Matthews-Olson pigment protein complex of the green sulfur bacteria.

Although progress has been made towards accurate scientific models to describe EET, proteins are still treated only at the level of harmonic approximations. There are two main precedents for these approximations: First, it is well established that polymer energetic landscapes are not harmonic, but thanks to the central limit theorem, the collective motions of polymers can often be statistically approximated as a Gaussian-distributed liquid-like system. Second, folded proteins are often treated as energetically solid in analogy to crystalline solids by treating fluctuations as elastic motions about the native state. The fact that proteins fall in the intermediate regime between highly mobile polymers and elastic solids calls into question the harmonic approximations employed in EET modeling. We must validate any approximations made about the system as there is no a priori justification for these approximations in energetics or in a central limit theorem argument. In fact, a growing body of protein research suggests that anharmonic fluctuations and long range correlations are generic features of protein dynamics.

Anharmonic fluctuations in proteins are difficult to describe in a uniform way because they occur heterogeneously across time scales. To begin with, transitions occur over several picoseconds, but waiting times can be much greater than nanoseconds. Furthermore, proteins may experience periods of high activity followed by periods of no activity at all. This heterogeneity suggests a separation into what are called static fluctuations (comprising activity with duration longer than a few hundred picoseconds) and dynamic fluctuations (comprising activity with duration shorter than 100 ps at which scale the protein is well
approximated by elastic motion). Therefore, if we try to use equilibrium methods directly by averaging over long times, we fail to respect the unique ability of the fast fluctuations to respond to impulses. However, if we abandon equilibrium averaging, our results fail to acknowledge the conformational diversity inherent to proteins and the possibility for dynamic fluctuations to vary significantly as the conformational space is explored.

We propose a framework that reconciles these disparate limits. By analyzing chromophore electrochromic shifts in the Fenna-Matthews-Olson pigment protein complex on much longer time scales than reported in the literature, we can build a multiscale description of the pigment dynamics. The observed electrochromic shift dynamics are heterogeneous in time, and the protein itself is resistant to symmetric equilibration. Through a decomposition of time scales, we show that while fluctuations slower than 1ns are sluggish and heterogeneous in time, fluctuations faster than 1ns are unchanged by conformational exploration, validating the methods that have been published to estimate spectral densities from simulation. Fast fluctuations are protected from slow conformational heterogeneity by spatial proximity to the chromophore and non-collectivity of contributions from residues (except for a scant few exceptions). We assess the validity and extent to which fluctuations obey Gaussian statistics and linear response with a statistical methodology directly based on the use of a spectral density as a generative model. Slow fluctuations are notably heterogeneous, on the other hand. The conformational changes that correlate to slow electrochromic activity are widely spatially distributed around the protein, making sense of the breadth of time scales on which fluctuations occur.

In many areas of statistical mechanics, the central limit theorem is used as an explanation for the validity of harmonic assumptions. It is thus notable that proteins frequently fail to obey harmonic approximations as shown in experiments and simulations of increasing spatiotemporal resolution. There are likely many relationships between anharmonic protein fluctuations and biologically important observables yet to be discovered. In this vein, the analysis of electrochromic shifts fluctuations are a novel way to understand the coupling between anharmonic protein fluctuations and biological electronic energy transfer. With a better foundational understanding of protein’s impact on EET, we can carry on with greater certainty about the limits of approximate harmonic approaches. Furthermore, with improved protein modeling in EET, it may be possible for EET research to give back to protein research through the deployment and study of natural and artificial optical probes of protein environments.
Dedicated to all of the amazing students of Compass, to my mentors Joel Corbo and Dimitri Dounas-Fraser, and my cohort mates Hilary Jacks, Jesse Livezey, and Brad Barker. You are an endless and constant source of inspiration and motivation. Let’s keep doing our best to make the world a better place.
## Contents

**List of Figures**  
**List of Tables**

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Start Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Electronic Energy Transfer in Pigment Protein Complexes: EET in PPCs</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Modeling energy transfer in thermal environments</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1</td>
<td>EET Hamiltonian</td>
<td>4</td>
</tr>
<tr>
<td>1.1.2</td>
<td>Parameters for this Hamiltonian</td>
<td>5</td>
</tr>
<tr>
<td>1.1.3</td>
<td>Energy transfer dynamics</td>
<td>7</td>
</tr>
<tr>
<td>1.1.4</td>
<td>Summary of modeling energy transfer in thermal environments</td>
<td>9</td>
</tr>
<tr>
<td>1.2</td>
<td>Green Sulfur Bacteria and the Fenna Matthews Olson Pigment Protein Complex</td>
<td>9</td>
</tr>
<tr>
<td>1.2.1</td>
<td>FMO Crystal Structure</td>
<td>10</td>
</tr>
<tr>
<td>1.2.2</td>
<td>FMO in its natural habitat</td>
<td>12</td>
</tr>
<tr>
<td>1.2.3</td>
<td>Speciation and Mutation</td>
<td>17</td>
</tr>
<tr>
<td>1.3</td>
<td>Overview of text</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td><strong>Scientific Models, Scientific Theories</strong></td>
<td>19</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction to scientific models</td>
<td>20</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Models in practice</td>
<td>20</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Aside: Scientific modeling in the classroom</td>
<td>21</td>
</tr>
<tr>
<td>2.1.3</td>
<td>Developing models</td>
<td>22</td>
</tr>
<tr>
<td>2.1.4</td>
<td>Summary of models</td>
<td>23</td>
</tr>
<tr>
<td>2.2</td>
<td>Models of the protein</td>
<td>24</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Physical Structure: Crystal Structures</td>
<td>24</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Physical Structure: Molecular force fields</td>
<td>26</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Fluctuations: Free Energy</td>
<td>27</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Fluctuations: Linear response</td>
<td>30</td>
</tr>
<tr>
<td>2.2.5</td>
<td>Fluctuations: Modeling discrete photon absorption with linear response</td>
<td>31</td>
</tr>
<tr>
<td>2.2.6</td>
<td>Dynamics: Fluctuation-Dissipation</td>
<td>32</td>
</tr>
<tr>
<td>Section</td>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.2.7</td>
<td>Dynamics: Fluctuation-Dissipation by analogy</td>
<td>33</td>
</tr>
<tr>
<td>2.2.8</td>
<td>Dynamics: Models of protein landscapes</td>
<td>36</td>
</tr>
<tr>
<td>2.2.9</td>
<td>Dynamics: Markov State Models</td>
<td>36</td>
</tr>
<tr>
<td>2.3</td>
<td>Models of the energy gap</td>
<td>38</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Hartree-Fock and determinants</td>
<td>38</td>
</tr>
<tr>
<td>2.3.2</td>
<td>ZINDO</td>
<td>40</td>
</tr>
<tr>
<td>2.3.3</td>
<td>TD-DFT</td>
<td>41</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Point charge embedding</td>
<td>41</td>
</tr>
<tr>
<td>2.3.5</td>
<td>Intermolecular coupling</td>
<td>42</td>
</tr>
<tr>
<td>2.3.6</td>
<td>TDC, TrEsp, and CDC</td>
<td>43</td>
</tr>
<tr>
<td>2.4</td>
<td>Models of thermal quantum mechanics</td>
<td>44</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Linear optics</td>
<td>44</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Forster Resonant Energy Transfer</td>
<td>45</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Redfield Equation</td>
<td>46</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Semiclassics</td>
<td>48</td>
</tr>
<tr>
<td>2.4.5</td>
<td>Spectral density baths</td>
<td>49</td>
</tr>
<tr>
<td>2.5</td>
<td>Conclusion</td>
<td>49</td>
</tr>
</tbody>
</table>

3 Molecular dynamics of the FMO pigment protein complex 51

<table>
<thead>
<tr>
<th>Section</th>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Preparing FMO topology for sampling</td>
<td>53</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Forcefield parameters</td>
<td>53</td>
</tr>
<tr>
<td>3.1.2</td>
<td>CDC parameters</td>
<td>55</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Missing residues and heuristic modeling</td>
<td>55</td>
</tr>
<tr>
<td>3.1.4</td>
<td>Solvent and Ions</td>
<td>58</td>
</tr>
<tr>
<td>3.2</td>
<td>System stabilization</td>
<td>58</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Energy minimization</td>
<td>59</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Position restraints</td>
<td>60</td>
</tr>
<tr>
<td>3.3</td>
<td>Quantitative interpretations of FMO configurations</td>
<td>61</td>
</tr>
<tr>
<td>3.3.1</td>
<td>On sampling rates for molecular simulations</td>
<td>62</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Radius of gyration</td>
<td>64</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Potential energy</td>
<td>64</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Pair distances</td>
<td>65</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Radial distribution function</td>
<td>65</td>
</tr>
<tr>
<td>3.4</td>
<td>Qualitative interpretations</td>
<td>66</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Ion binding</td>
<td>67</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Solvent affinity</td>
<td>68</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Solvent penetration</td>
<td>69</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Interpreting the in-vivo analogous environment</td>
<td>70</td>
</tr>
<tr>
<td>3.5</td>
<td>Model iterations</td>
<td>72</td>
</tr>
<tr>
<td>3.5.1</td>
<td>4BCL structure</td>
<td>72</td>
</tr>
<tr>
<td>3.5.2</td>
<td>3EOJ zmshbsri</td>
<td>72</td>
</tr>
<tr>
<td>3.5.3</td>
<td>3EOJ zmshbsriPR37x</td>
<td>73</td>
</tr>
</tbody>
</table>
3.5.4 3EOJ svnrhP7y ........................................ 73
3.5.5 3EOJ svnrhdeiP7y ........................................ 74
3.6 Conclusion ..................................................... 74

4 FMO’s high frequency fluctuations are protected 75
4.1 Protein dynamics create unique challenges to sampling the gap ................. 76
  4.1.1 Biased simulations do not sufficiently accelerate protein sampling ........... 77
  4.1.2 The multi-well model of protein landscapes explains slow dynamics ......... 80
  4.1.3 Direct simulations sample conformational states slowly but reliably ......... 81
  4.1.4 Summary: Protein dynamics create unique challenges to sampling the gap . 82
4.2 Gap trajectories have multifaceted heterogeneity of physical origin .......... 82
  4.2.1 Warning about generalization of results ...................................... 83
  4.2.2 Basic gap descriptions show substantial heterogeneity across most major factors ........................................ 84
  4.2.3 All major heterogeneity is removed by eliminating low-frequency components ...................................... 84
  4.2.4 Consistency of high frequency covariance across monomers lends evidence to a physical basis for separating time scales ................. 91
  4.2.5 Summary: Gap trajectories have multifaceted heterogeneity of physical origin ...................................... 94
4.3 High-frequency gap distributions suggest some complexity beyond the direct spectral density ........................................ 96
  4.3.1 Spectral density in PPC bath modeling ...................................... 96
  4.3.2 Spectral density statistical overview ...................................... 96
  4.3.3 FMO Spectral Density .................................................. 97
  4.3.4 Spectral density estimation .................................................. 98
  4.3.5 Density of States, spectra, and response ...................................... 100
  4.3.6 Overview of published ZINDO calculations ...................................... 102
  4.3.7 Simple model for Density of States skew ...................................... 103
  4.3.8 Summary of gap overview .................................................. 104
4.4 Fast modes are independent and proximal while slow modes are collective and nonlocal .................................................. 105
  4.4.1 Principal component analysis defined ...................................... 105
  4.4.2 Practical challenges with PCA .................................................. 107
  4.4.3 Results .................................................. 108
4.5 Conclusion ..................................................... 112

Bibliography 114
## List of Figures

1.1 A cartoon representation of the green sulfur bacteria light harvesting system, reproduced from Scholes 2010 [1] [2]. The reaction center shown is from a related species of purple bacteria with similar overall features. .......................................................... 2

1.2 Linear absorption spectrum for FMO molecule from *C. tepidum* [6]. Note the three major peaks and the shoulder at 12650 cm\(^{-1}\). .......................................................... 7

1.3 Dorsal view of the front-most monomer, with the majority of the structure rendered transluscently. Top, helices are rendered in pink, and bottom, beta sheets are rendered in pink. Chromophores are colored with respect to their role in the energy transfer process in analogy to the colors associated with an “energy funnel”; blues are the entry points, greens are the intermediates, and red is the exit point. Water is rendered as a blue surface interspersed with the grey surface of the other two FMO monomers. .......................................................... 11

1.4 FMO trimer with one monomer left unrendered. The hydration network is visible thanks to the missing monomer. Dynamic bonds give depth and proximity information for the large quantity of water molecules present. Network connects the upper cone-shaped indentation to the loop 2 groove on the left, but not to the bottom flattened-spheroid indentation. Loop 2 of the right monomer protrudes in front of the water on the right of the figure, providing more spatial context about the location of loop 2 and its associated groove. .......................................................... 12

1.5 Cartoon of the parts of the green sulfur bacterial light harvesting antenna at a range of details, moving progressively closer to the FMO trimer. .......................................................... 13

1.6 A spatial structure of chlorosomes in the green sulfur bacteria cells from cryo-electron tomography. Image reproduced from Kudryasheva et al. [28] .......................................................... 15

1.7 Spatial construction of the CsmA molecule, showing hydrophobes in white and hydrophiles (basic, acidic, and polar) in purple. In blue cartoon transparent overlay, the variety of disordered structures resolved in the PDB file are rendered simultaneously. Image constructed from PDB file 2K37 in VMD. .......................................................... 16

1.8 Phylogenetic tree of a few species. Pairs in this sample are fairly evenly distributed among a rich diversity of species within the Chlorobiacea family. Phylogenetic information from Imhoff 2002 [36], 2003 [37]. .......................................................... 17
1.9 Partial aligned sequences of several Green Sulfur Bacterial species. Amino acids 1-10 show the good preservation of the hydrophobic tail, including the inclusion of an additional hydrophobe in Pld. phaeum. Sequence 149-187 included from Tronrud 2009 [38]. This sequence shows strong conservation between 148 and 164, as well as between 169 and 187.

2.1 An electron density mesh for small ligand within the 3eoj protein crystal structure as viewed by the JSMol molecular web viewer. While the crystal structure predicts distinct valence clouds for the protein and the ligand, there is an appreciable amount of valence interaction between the Mg+ atom on the ligand and the HIS110 group that can be observed directly in the crystal structure.

2.2 The difference in the free energy is much more practical to compute than the absolute free energy because the only region that contributes to the integral is the overlap regions, and contributions from “tail” regions of A and B are strongly attenuated. The overlap region is shared symmetrically.

2.3 The net effect of a linear bias to a Gaussian distribution can be seen clearly by considering the probability distribution in terms of its negative log, the free energy. The linear bias in this space serves to shift the parabola, but does not change its curvature.

2.4 A conventional RC Circuit

2.5 Time dependent applied voltage and response

3.1 Top: Configurations of 4BCL with three major groups highlighted. Bottom: Configurations of 4BCL zoomed on the pink group. Left: A configuration from the 4BCL crystal structure. Right: A configuration from the end of a low-quality 500ns monomer simulation. After 500ns, the protein begins to unfold. All three groups (green, orange and pink) move considerably over these 500ns. Most notably, pink begins to unravel from the folded structure, as shown in the zoom.

3.2 Spectral density from DFT vs spectral density from MD for ceccarelli shows that the forcefield implemented is able to reconstruct the vibrational density of states from DFT with substantial accuracy.

3.3 Residue 214 is missing in the 3EOJ crystal structure. Additionally, the adjacent residues have two possible altlocs. To complete the protein model, we must simultaneously select an altloc and complete the missing residue for that altloc.

3.4 HIS and BCL form a biologically important contact in systems with Mg+ ions. In fact, four of the seven BCL molecules in FMO are held in place by this conjugation. A contact between one such of these pairs, BCL375 (magenta) and HIS (cyan, blue, and red), is shown with the conjugation represented by a yellow stick. The protein scaffold is represented in silver and alternate BChl A molecules are in iceblue.
3.5 Ions play an important role in the equilibrium structure of 3eoj. Left: cyan colored Cl\textsuperscript{-} molecules accumulate in the top pore to interact with positively charged residues (blue). Right: blue colored Na\textsuperscript{+} molecules aggregate in the bottom pore to interact with negatively charged residues (red) as well as in one cluster of residues on the perimeter consisting of negatively charged residues and polar residues (green). This binding is an example of the solvent structurally screening long-ranged electrostatics.

3.6 A close up on the underside of the 3eoj protein showing violation of inter-monomer symmetry in the residue group 7-36-347. In the left monomer, circled residue 36 finds a relaxed position in which it is resting deeper in the bulk of the protein. In the middle and right monomers, residue 36’s position sits further outwards and with more contact to residue 347, which also causes interaction between residue 7’s charged oxygen groups and its charged nitrogen groups.

3.7 Demonstration of the effect of direct sampling vs filtered samples for a particular desired sample frequency. The red line is a mixed high and low frequency signal, while the blue thick line is the low frequency isolated. The red dots are the red signal subsampled by a factor of 10,000, and the blue dots are the red signal subsampled by a factor of 100 and then filtered with a window size of 100. By filtering from a higher frequency subsample, we are able to capture the trend in the blue signal that was obfuscated by the high-frequency noise.

3.8 Radius of gyration plot for svnrhcede simulation. The radius of gyration relaxes to a fairly consistent value after 600ns, indicating some degree of relaxation. Radius of gyration is sampled at 10ps and smoothed with a 20ns window. M1, M2, and M3 refer to monomers 1, 2, and 3, respectively.

3.9 Potential trajectory sampled at 1 ps smoothed with square window of 10 ns.

3.10 Trajectory of distances at different inter-monomer surfaces for residues 37 and 347. The trajectories show the breaking of symmetry even more clearly than figure 3.6. In fact, the two residues that looked quite similar from visual inspection actually show appreciable differences. Note that this is actually residue 37 and not the rendered residue 36 from 3.6, though the conclusion is the same due to the proximity of the two residues.

3.11 g(r) in time across monomers. The amplitude of g(r) communicates how stiffly bound an ion is, while the peak position communicates how closely bound the ion is. As time elapses, the g(r) across different monomers converge.

3.12 Visualization of the Na\textsuperscript{+} ion becoming embedded in the surface of the protein. Left: fully captured Na\textsuperscript{+} ion interacting stringly with neighboring O\textsuperscript{-} groups. Right: partially captured ion needs to wait for steric groups to move out of the way before it can become fully captured.
3.13 The upper surface of the FMO trimer with hydrophobicity used for coloring. White surfaces are hydrophobic, while purple ones are hydrophilic. While most proteins aim to bury their hydrophobic groups, this surface is mostly equally split with hydrophobic and hydrophilic groups. This suggests that the in-vivo environment may have some hydrophobic contacts to compensate.  

3.14 The lower surface of the protein shows what is much more expected of the exterior of a protein than FMO’s lower surface, seen in figure 3.13. There is one notable exception to this hydrophilic surface, which is the tail groups rendered as licorice in the center of the trimer. These tails are intrinsically disordered even in the crystal structure, and this disorder is likely driven by their hydrophobicity.  

3.15 The network of water molecules in the complex rendered with a cut in the protein. The frontmost residue is rendered as a glass cartoon for minor context. Water molecules are represented with dynamic bonds to emphasize the network structure present. The interior channels are a notable feature of this hydration network, which could allow for ion presence inside of the trimer and have important electrostatic implications on the structure.  

4.1 Despite significant variety in response mean, slope for response across all samples is consistent with the predicted slope of equation 4.10 as determined by the variance MLE.  

4.2 Left: a collection of 500ps biased trajectories where protein initial configurations were sampled from a 50ns trajectory uniformly. Right: a collection of 500ps biased trajectories where protein initial configurations were identical. The greater spread in the left figure proves that gap values depend strongly on subtle details of the protein conformation. Despite the differences in variance, though, the responses are not significantly different.  

4.3 The trajectory for BCL373 shows fluctuations on both slow and fast time scales. The separation of scales is essential because the fast fluctuations from any point in time span the range explored by the low-frequency variability. When considering the original signal, monomers are somewhat difficult to distinguish; when considering only means, the differences between monomers becomes more stark. Left: full-data trajectory, of high-frequency quantiles, demonstrating a convolution of fast and slow data. Right: low-frequency trajectory, making clear the case for distinguishability of monomer due to small sample size.  

4.4 A representation of the BCL pigments in FMO colored with their own identifying color that will be used throughout the chapter.  

4.5 A percentile timeseries for every chromophore in the trimer. The interquantile range is rendered in the interior, and the remaining range between the 5th and 95th percentiles is rendered in a lighter shade.  

4.6 The chromophores have densities of states that are tightly packed. Separations between means are on the scale of 1-3 standard deviations.
4.7 Inter-chromophore and inter-monomer correlations for low frequency signal (top) and for high frequency signal (bottom). Left panes show covariance matrix while the right panes show hexbin 2d histogram for the largest absolute correlation elements in the matrix; that is, for BCL376 monomer 1 against BCL376 monomer 3 and for BCL373 monomer 1 against BCL374 monomer 2 for the low and high frequency, respectively. The erratic and incomplete shape of the low frequency histogram shows clearly the poor sampling, while the clean Gaussian shape of the high frequency histogram shows thorough and high-quality sampling.

4.8 Inter-monomer q-q plots and empirical CDF comparisons for BCL373. From the ECDF, we can test for exact identicality of distributions. Because these monomers are identical, their p-scores should be uniformly drawn from the $U(0,1)$ distribution. Distributions are sparsified by a factor of 10 to eliminate as much correlation as possible to improve p-scores. Left: Before low-frequency filtering, p scores are significant to a massive degree. q-q plots show the qualitative difference between these distributions. Right: After mean removal, only the tails vary slightly, as seen in the q-q plot. The values of .4 and .2 for p-scores are quite acceptable scores that do not cause us to reject the hypothesis that these distributions are identical.

4.9 Histograms of the gap for BCL373 before low frequency filtering (left) and after mean removal (right). In both situations, the distributions are quite Gaussian.

4.10 Quantile scatterplot matrix for BCL373. Left: before low-frequency removal we see lots of correlation between the 5th 50th, and 95th percentile. Right: after low-frequency removal, nearly all of the correlation between these quantiles falls out, suggesting that the bulk of that correlation was simply from a consistent distribution with a continuously varying mean.

4.11 High frequency quantile trajectory featuring the 1st, 25th, 75th, and 99th percentiles of the trajectory. Once low frequency trending is removed, the remaining signal is quite well behaved in time.

4.12 Covariance matrix of a merged solvent+ion group and the monomer symmetrized union of all residues that were within the top-5 variance contributing to BCL 373 for any of monomers 1, 2 and 3 – thus the consistent choice of residue groups within each the top and bottom row. Covariance for monomer 1, 2, and three are in the left, middle, and right columns, respectively. Top: low-frequency covariance matrices have more differences and also show more difference between monomers for which residues are top contributors. Bottom: high frequency covariance matrices are almost identical across monomers.
4.13 Estimate of the spectral density for a short, high-resolution time series of the CDC method. Due to the trajectory’s heterogeneity, an average of sub-trajectories was necessary to reach a good estimate of the autocovariance function. Top left: quantile trajectory for the 1st, 25th, 75th, and 99th quantiles. Top right: Autocorrelation function for the full trajectory (thick red) shows greatly arrested decay after a brief period of rapid decay, while the ensemble of shorter trajectories (translucent red) shows much faster decay in the majority of samples. Bottom left: autocorrelations (translucent red) were "zeroed" at 1ps, and the residual was fit to a single exponential function and a collection of damped oscillators. Bottom right: parametrized values of exponential functions and damped oscillators allow for analytic evaluation of the spectral density.

4.14 Normality tests for various window sizes of T=1 ns, T=5 ns and T=20 ns on BCL373 Monomer 2. Skewness and kurtosis p-values are expected to be uniform for a Gaussian distributed process. This seems to largely be the case, though the skewness test for the 20 ns window indicates the emergence of detectable skew in a number of windows.

4.15 Example of a PCA gone wrong from an unused analysis of dihedral angles. A single eigenmode is represented here, and is formed from the sum of all of these features. The shown features are the individual trajectories for which the covariance matrix has been constructed. Alignment of trajectories that contain a single transition will create a large principal mode with no statistical meaning, and elimination of all modes with drift will lead to a severely underpowered analysis.

4.16 Basic summary of the largest modes of the principal component analysis for BCL373 for the high-frequency signal. Top row: pie charts for variance of each principal component, with the mode’s inverse participation ratio labelled on top. Middle row: residue against presence of residue within the top six principal components; As the sum of the squares of the eigenvector matrix sum to 1 in both the rowwise and the columnwise directions, this “presence” term is easy to compute directly from the eigenvectors. Bottom row: trajectories of the largest PC1 mode for each of the monomers.

4.17 The cumulative mean of the inverse participation ratio for PCA on fast and slow modes over different monomers on BCL373. Left: fast modes show a low mean IPR for monomers 1 and 2, but a spurious large IPR mode 2 for monomer 3 causes the mean to spike for the remainder of the figure. Right: slow modes show a higher mean IPR.

4.18 Histograms for BCL373 top 6 eigengaps for monomers 1, 2, and 3 for top, middle, and bottom locations, respectively. The color scheme for gaps in this figure matches that of 4.16 so that they can be cross referenced easily.
List of Tables

1.1 Significant contacts and hydrogen bonds with BChl A molecules in *P. aestu-arii* based on data from crystal structures. Residue numbers with an overbar (e.g. Ile137) are from a monomer adjacent to the referenced BChl’s monomer of convention. ................................................................. 14

4.1 Table of features of individual BCL molecules in the complex. Each environment has a different local environment that leads to a different spread in gap values. Each also has a few key adjacent residues that most strongly impact its high-frequency fluctuations. Residues with an asterisk (*) are located on the positive adjacent monomer (i.e. for a chromophore on monomer 2 ASP160* is located on monomer 3). ................................................................. 83

4.2 Table of values from quantum mechanical calculations [12] [16] benchmarked against CDC contributions from our simulations. CDC with screening falls about an order of magnitude short in scale to the contribution needed to explain the difference between vacuum and point-charge embedded DFT calculations, and CDC without screening still falls about a factor of two short of the smallest value possible to explain the difference in TDDFT variance estimates. .................. 102
Acknowledgments

Thank you first and foremost to all of the people and institutions working relentlessly and consistently to make intellectual growth like this possible; people like Anne Takizawa, Donna Sakima, Lisa Littlejohn, Leslie Dietrick, the Gilman Hall building management, the UC Berkeley janitorial staff, the UC Berkeley Physical plant stationary engineer who helped me get into my office over Christmas break, the employees of the Coffee Lab, Cafe Strada, Cafe Milano, and all the other coffee shops that fuel creativity and/or sleep deprivation, the NSF program managers, the IRS, the DOE, the Physics educational and lab staff like Terry Buehler, and countless others. I am grateful to be a part of a society where I can reasonably pursue such a tremendous intellectual endeavor for six years, and I am grateful to each of the individual people therein who have made my time at Berkeley that much easier.

Thank you to my advisor, Prof. Phillip Geissler, for taking me on and making sure I was supported financially throughout my entire PhD.

Thank you to the late Prof. David Chandler for his contributions to statistical mechanics and to the Pitzer Center at large – his vision and work impacted every facet of being a statistical mechanics researcher at UC Berkeley.

Thank you to my lab mates from my earlier years, especially Anna Schneider and Todd Gingrich, for informative tutorials, practical guidance, and sometimes just real talk about how difficult the uncertainty of research can be. Thank you to my lab mates from later years, especially Pratima Satish and Jaffar Hasnain, for working together to make our work environment work for us and for supporting each other by making sure we stay focused and and excited about science. Thank you to my contemporaneous Phillistines, Katie Klymko and Dayton Thorpe, for being exceedingly relatable and having my back when I needed it. Thank you to the rest of the researchers who have overlapped with me to any significant extent; I am very proud to be a part of the impressive and enthusiastic Geissler academic family.

Thank you to my parents for a safe and supportive start to my life. Thank you to my childhood role models, Sensei Juan and Christopher, for building in me the self-discipline, character, and perseverance I needed to accomplish a task like this. Thank you to the people who got me and kept me on the right path during my teen years, especially Tom Sharkey and Chie Helinski. Thank you to Ben Dreyfus, Dr. Esper, Gary Rubenstein, Mr. Fayad, Mike Zamansky, and all of the other public school science teachers laid the path for my academic direction. Thank you to Dr. Francesc Roig and Dr. Sathya Guruswamy for paving the way to graduate school with UC Santa Barbara’s College of Creative Studies Physics program. Thank you to Alex Sztein for taking me on as an undergraduate researcher, and thank you to the whole Nakamura Group for giving me a glimpse of what being a graduate student will look like.

Thank you to my amazing siblings who know me better than anyone else, even if we don’t spend that much time together. I’m very glad we’re in this “life” thing together.
Preface

We often use a model "because it works". My graduate research has involved snooping into many of the nooks and crannies of EET to ask: why do we *know* that this works? I have found many places where we can make good arguments for our particular approximations, but also several places where our approximations are probably reasonable, but fundamentally uncontrolled.

I hope this thesis can serve as a source of questions and answers for experimental researchers to acquire a greater appreciation of our good fortune to be able to employ our standard approximations. I also hope this thesis can serve as a roadmap for young theoretical researchers to acquire a greater intuition for why the standard approximations hold, and to get a better sense of where to look if they want systems that explore phenomenology outside of the bounds of our standard approximations. I believe that taking a fundamentally thermal-first approach to photosynthetic electronic energy transfer avoids many of the conceptual challenges of the abstract nature of quantum-first approaches by allowing us to reference the tangible, instantaneous, singleton state of the system at any particular time instead of immediately appealing to wave functions and density operators.

This framework will not introduce the interesting and important non-linear optics modeling that is necessary to analyze modern spectroscopic experiments, as these analyses require a quantum-first approach. Nor am I particularly qualified to speak on the topic, as I have only an understanding of it that I have found necessary to understand the essence of experimental results. The reader, particularly the experimentalist, should supplement this thesis with the appropriate sources on this topic from experts in non-linear spectroscopy. I have found "Mukamel for Dummies" to be a useful resource, and ostensibly the reader with a critical need to understand nonlinear spectroscopy in depth should consult Mukamel’s original text.
Chapter 1

Electronic Energy Transfer in Pigment Protein Complexes: EET in PPCs

The vast majority of energy generated by humans is generated by the same concept: spin a turbine and you can generate electricity. Whether it’s a river spinning a waterwheel or a carefully monitored uranium core boiling water in a nuclear plant, the energy is generated by converting mechanical work into electricity. One step more complex is the engine, with precise timing cycles that efficiently convert heat into work. Solar panels represent an epoch of complexity of man made electricity generation, using properties of semiconductor materials and an applied voltage to get the energy extracted from photoexcited electron-hole pairs.

In spite of all of this technology, the deftness with which plants harvest solar energy puts us to shame. With the precision and heterogeneity of an engine cycle and the quantum complexity of a solar panel plants are able to capture, transport, and extract energy from molecular photoexcitation robustly and reliably enough to sustain life. While these are not the most efficient systems in terms of work-to-heat ratios, their flexibility and ability to employ biological elements are worth careful study.

This dissertation is focused on the following broad problem: in what ways do the anharmonic fluctuations of structured proteins of natural pigment protein complexes interact with the electronic energy transfer? This question requires a thorough introduction before being addressed as it spans a wide range of complex and incompletely understood physical phenomenology.

In answering this core question, I address the question of why simple quadratic thermal baths are so incredibly ubiquitous and effective for modeling electronic energy transfer. Although the bath environment may not be strictly harmonic, we can still often make relevant and accurate predictions even when reducing the complexity of the bath to a harmonic level. I address the limits of this approximation, defining a more clear delineation between when we must consider the full atomistic nature of the protein environment and when it can be treated as a collection of thermalized springs.
Figure 1.1: A cartoon representation of the green sulfur bacteria light harvesting system, reproduced from Scholes 2010 [1] [2]. The reaction center shown is from a related species of purple bacteria with similar overall features.

The stages of energy transfer and energy extraction have many commonalities. The main conceptual continuity is the strong coupling of thermal and electronic degrees of freedom. Thermal coupling is vital for dissipating the energy from the quantum system so that the excitation can make progress along the structured and sequential segments of that system’s photosynthetic network. It also broadens pigment spectra such that energy transfer processes like FRET are accelerated and can be completed within the fluorescence lifetime.

Of the physical processes involved in biological photosynthesis, energy transfer has likely been the most studied. It is the process we too will focus on. This process lies at the intersection of quantum theory and thermal modeling. To begin understanding the system, we will first go through a brief overview of open quantum systems, the framework that forms the foundation of the energy transfer processes.

1.1 Modeling energy transfer in thermal environments

Here we introduce the open quantum system framework that allows for a formal treatment of electronic-thermal coupling.
The formalism of quantum mechanics hinges on a few essential symmetries: Hermiticity of operators, time-reversibility, and energy-conservation. Depending on the framework being used, these are not independent symmetries. In general, though, all three of these symmetries are present in quantum theories.

In contrast, thermalized dynamics are characterized by a decay towards equilibration instead of the coherent oscillations that quantum systems prefer. Time reversibility in thermalized theories is only present when all degrees of freedom are observed and the observation of each occurs with infinite precision. If degrees of freedom are unobserved, the standard practice is to average over all possible states of the unobserved system, thereby introducing non-Hermiticities into the time evolution operators of the reduced subsystem; If measurement happens with limited precision, any particular initial state will be modeled with a distribution, and that distribution will broaden both forward and backwards in time. In fact, the sequence of operations 1. measure; 2. forward-propagate; 3. measure; 4. backwards propagate will no longer be the identity operator. Though the mechanism through which thermalization occurs on a universal scale is debated, we can skirt the philosophy and just considering the logistics of modeling.

In the tradition of classical statistical mechanics, open quantum theories proceed through the inclusion of an infinite (usually harmonic) bath of abstract degrees of freedom that are “integrated out”. In this framework, not only are the initial conditions averaged over consistently, but the “ripples in time” created by particular initial bath conditions and fluctuations of the system are also treated properly, such that the dynamics of the reduced system is exactly equal to the ensemble-averaged dynamics of the complete system. These “ripples” are a form of memory, and one of the exciting things about open quantum systems is the emergence of non-Markovian and time non-local operators. Still, this bath integration helps us in two ways: first, we can develop our theory in an emergent way, rather than trying to write one down without symmetries to guide us; second, the bath gives us a huge space within which to tinker with parameters. In practice, oscillator baths are one of the few analytically solvable families of bath modes, so baths are frequently parametrized with respect to the “color”, the oscillator density of states, or spectral density. This spectral density is a crucial and debated parameter set, yet it is still fundamentally an approximation that asserts a strictly harmonic bath. We will explore this point throughout the text.

We can also rephrase these thermalized quantum mechanics in terms of spectroscopically observable features, rather than just focusing on the mathematical properties of the formalism. In terms of what we can spectroscopically observe, we want to be able to sink and source arbitrary amounts of energy from our subsystem of interest, and we want to introduce enough distinct elements that our subsystem has its oscillations dephased at a reasonable rate. We need the bath model we select to be generic enough that we can parametrize a variety of experimentally observed dephasing behaviors that arise from protein fluctuations (which, though they are not rigorously a bath of harmonic oscillators, are often effectively modelled as such). The spectral density bath excels in this regard, and is therefore a sensible choice in the absence of more atomistically detailed observations.
1.1.1 EET Hamiltonian

We can now begin to construct an EET Hamiltonian for a two chromophore system. This is a multi-wavefunction system, where $|\psi_i\rangle$ refers to the full wavefunction on the first chromophore and $|\psi_2\rangle$ refers to the full wavefunction on the second chromophore. Thus, we work in the product space basis $|\psi\rangle$. If we are concerned with modeling optical excitation, the simplest model we can construct that has the phenomenology we are trying to describe is one where the wave function $|\psi_i\rangle$ for either chromophore has only two states: $|i_g\rangle$ and $|i_e\rangle$, separated in energy by $G$. These states often refer to the HOMO and LUMO electronic states, respectively, and those reference states serve as a useful touchstone to understand the model Hamiltonian.

Schematically, a HOMO and LUMO state pair has a transition dipole moment, $\mu_T$, and they can couple via an electric field $E$. When this electric field is time varying with a frequency that corresponds to a quanta of energy that matches $G$, a resonance condition occurs that greatly amplifies the transition rate between the ground and excited state, providing a mechanism for optical excitation. A pair of pigments are then coupled together with a coupling constant $J$ that comes from the dipole-dipole interaction of each molecule’s transition dipole moment. The electronic Hamiltonian is then coupled to a “phonon” bath – though here the term phonon is being used out of convention as the bath may take any form – through some bath collective variable $x_i$. This collective variable does not need to be the same among chromophores, and does not need to be a linear function of the bath coordinates (though each site is usually taken to be a site-dependent linear combination of the bath coordinates). This linear combination of coefficients is defined by coupling strength $g_i$.

To summarize,

$$|\psi\rangle \equiv |\psi_1\rangle \otimes |\psi_2\rangle$$  (1.1)

$$|\psi\rangle \equiv c_1|1_g\rangle|2_g\rangle + c_2|1_e\rangle|2_g\rangle + c_3|1_g\rangle|2_e\rangle + c_4|1_e\rangle|2_e\rangle$$  (1.2)

$$H = \begin{bmatrix} 0 & \mu_T \cdot E & \mu_T \cdot E & (\mu_T \cdot E)^2 \\ \mu_T \cdot E & G & J & \mu_T \cdot E \\ \mu_T \cdot E & J & G & \mu_T \cdot E \\ (\mu_T \cdot E)^2 & \mu_T \cdot E & \mu_T \cdot E & 2G \end{bmatrix}$$  (1.3)

$$H_{el-ph} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & g_1 x_1 & 0 & 0 \\ 0 & 0 & g_2 x_2 & 0 \\ 0 & 0 & 0 & g_1 x_1 + g_2 x_2 \end{bmatrix}$$  (1.4)

The phonon bath is unspecified because it can be of any complexity. It is the research objective of this dissertation to quantify the validity of a spectral density bath by treating this bath with atomistic dynamics, and computing $x_i$ values from classical approximations of electrochromic shift. However, let’s look at the choices left to make when considering a
spectral density bath. In computing thermal quantum dynamics, the spectral density bath is often treated as a fully quantum mechanical, leading to an open-quantum framework as in the Hierarchy Equations of Motion [3]. Alternatively, it can be treated as a classical bath, which allows a more complicated form of bath landscape, but at the cost of relying on semiclassics for the electronic system’s quantum behavior.

The $\mu_T \cdot E$ terms are the coupling to the electric field. For EET, we usually take the limit of small applied electric field, in which the unexcited and doubly excited states are approximately time-independent in an interaction picture, while the central block will always have substantial quantum dynamics. In the bulk of light harvesting studies, we restrict our attention to this block, called the first excited state manifold. We rewrite our multi-electron product states as single-index kets, as in

$$|1\rangle = |1_e\rangle|2_g\rangle$$  \hspace{1cm} (1.5)
$$|2\rangle = |1_g\rangle|2_e\rangle.$$  \hspace{1cm} (1.6)

This basis restriction also has the benefit of making our representation independent of the applied electric field. Our system only comes into being upon the injection of an excitation, and from that point it persists indefinitely. This is a reasonable way to model the reduced dynamics of a system that would otherwise be growing in complexity at a combinatorial rate as a function of chromophores included. Because of the interaction picture, it is conventional to write the energies of the first excited state manifold with respect to the potential of the lowest energy pigment. For the bulk of analyses, the absolute values of the energy do not factor into the results.

For the FMO complex, there are a variety of values that exist for $E_i$, $J_{ij}$, and a few forms for $H_{ph}$. An exemplary matrix of values [4], with all energies in cm$^{-1}$, is

$$H_{el} = \begin{bmatrix}
240 & -121.8 & 7.1 & -7.5 & 8.1 & -16.8 & -9.6 \\
310 & 39.1 & 10.3 & 1.4 & 13.0 & 4.4 \\
0 & -73.4 & -2.2 & -11.7 & 7.3 \\
110 & -94.7 & -20.6 & -82.2 \\
340 & 95.3 & -2.8 \\
330 & 48.5 \\
260
\end{bmatrix}.$$  

1.1.2 Parameters for this Hamiltonian

The values in the FMO Hamiltonian come from a variety of experimental and theoretical sources.

The most direct way to estimate energy gap values is to measure a variety of linear optical spectra and use their fairly large parameter space to fit a simple spectroscopic model whose parameters are the values for the energy gaps [5]. This was attempted many times in the late 90’s for the FMO complex, leading to a variety of estimates and orders for the gap energies of the pigments. Of most importance was determining the lowest-energy pigment
in the complex, as that pigment more than any informs theories about functional aspects of the complex. The definitive answer of which chromophore was lowest in energy was not settled upon until the mid 2000’s [4]. This difficulty in estimation comes about because the optical spectrum is incredibly congested; with 7 (and later 8) identical chromophores in the optical range of 12000 to 12800 cm\(^{-1}\) and only four resolvable peaks at 6K, spectrum estimation is necessary to identify separately their respective spectral contributions. To make matters worse, excitonic states between chromophores complicate the estimations as these eigenstates are the true spectral signatures, and the extraction of site energies from exciton bands requires knowledge of the structure.

Thus, it is not possible to determine site-specificity of parameters from a single absorption measurement. However, with a variety of orientationally-sensitive spectroscopies, and with pigment relative orientations known from the crystal structure, it is possible to combine these sources of information to infer which site corresponds to which absorption peak. In fact, Wendling et al. [5] were able to improve the quality of their fits and their site specificity by taking four different linear spectra: linear absorption, circular dichroism, linear dichroism, and triplet-minus-singlet spectra. For circular dichroic and linear dichroic spectra, FMO molecules are orientationally fixed through biaxial pressing [6], allowing fits that leverage the orientations from the crystal structure. For triplet-minus-singlet spectra, coupling estimates of adjacent chromophores from structure allows for using excitonic information to localize the source of a particular excitonic band. This range of orientation and excitation specific spectroscopy was able to make a strong case for which energy corresponded to which chromophore.

To reduce the degree to which these parameters are unstable, as much outside information as possible is put into the lineshape estimation. These calculations depend a lot on the choices of spectral broadening, so it was possible for Adolphs & Renger (2006) [4] to improve estimates by using spectral densities measured by Fluorescent Line Narrowing experiments [7] in conjunction with advanced non-Markovian spectral theory. They also made the important connection between spectroscopic fit methods and direct structure-based electrochromic methods. Their electrochromic shift calculations were able to corroborate the story from the spectral fits with classical electrostatic coupling calculation to estimate gap shifts. These electrochromic estimates were optimized and developed further for a variety of pigment-protein complexes, and are now known as Charge Density Coupling (CDC, for shifts in Qy gap) [8] and Transition Couplings from Electrostatic Potential (TrEsp, for couplings between pigment pairs) [9].

After these computations interest in FMO renewed as it received press for being a candidate quantum-coherent biological system [10] – though this hypothesis never quite panned out. Still, this interest led to a variety of ab-initio methods being applied to the system to compute energy gap parameters and spectral densities [11] [12] [13] [14]. These researchers used the new atomistic resolution to search for unseen correlations between fluctuations in site energies, to integrate Schrodinger’s equation numerically across the empirical trajectories of gap values, and to estimate spectral densities from atomistic simulations.

These ab-initio simulations are the current state of the art, but they continue to be
CHAPTER 1. ELECTRONIC ENERGY TRANSFER IN PIGMENT PROTEIN COMPLEXES: EET IN PPCS

Figure 1.2: Linear absorption spectrum for FMO molecule from *C. tepidum* [6]. Note the three major peaks and the shoulder at 12650 cm\(^{-1}\).

Refined. TD-DFT is largely taking the place of previous semi-empirical methods, but the combination of force fields as point charges and electrostatic methods are still imprecise and uncontrolled. Efforts to characterize and improve upon this imprecision are in development and frequently improving the state of the field [15] [16], but new phenomenology for the system has not been unearthed since the putative quantum coherence [10].

1.1.3 Energy transfer dynamics

There are two major parameters that affect the modeling of energy transfer: the coupling and the reorganization energy. A unified framework is inherently very complicated, so it is common to break it down into two separate regimes that are treated with different limiting-case models: large reorganization and low coupling and large coupling with low reorganization. The reorganization energy plays the role of thermal dissipation in the excited-state system, so these are the limiting cases where excited state energy is very well conserved vs where excited state energy dissipates greatly after transfer.

Reorganization energy is best explained through a model of a single chromophore’s excited and ground state. Assuming both states have potential energy surfaces of identical curvature with just different phonon minima, we can write the full joint electronic-phononic Hamiltonian, using the state definitions from eqn 1.5,

\[
H = |0\rangle\langle 0| \left[ \frac{1}{2} \alpha^2 \omega^2 x^2 + \frac{1}{2} \alpha^2 x^2 \right] + |1\rangle\langle 1| \left[ \tilde{G} + \frac{1}{2} \alpha^2 \omega^2 (x - \Delta)^2 + \frac{1}{2} \alpha^2 x^2 - \lambda \right].
\] (1.7)

Here, \(\lambda\) refers to the reorganization energy, \(\tilde{G}\) refers to the mean optical gap, \(x\) is the phonon coordinate, \(\Delta\) is the phonon mean shift upon excitation, \(\alpha\) is the coupling strength, and \(\omega\) is the frequency of the phonon mode. The difference between the minimum energy in each state is \(\tilde{G} \equiv G - \lambda\). We can compute \(\lambda\) from the other parameters by considering
\[ G = 0. \] The difference in environmental contributions as a function of phonon coordinate, \( g(x) \), can be computed directly, and \( \lambda = g(0) \).

\[
g(x) = \left[ \frac{1}{2} \alpha^2 \omega^2 (x - \Delta)^2 + \frac{1}{2} \alpha^2 x^2 \right] - \left[ \frac{1}{2} \alpha^2 \omega^2 x^2 + \frac{1}{2} \alpha^2 x^2 \right] \tag{1.8}
\]

\[
g(x) = -\alpha^2 \omega^2 \Delta x + \frac{1}{2} \alpha^2 \omega^2 \Delta^2 \tag{1.9}
\]

\[
\lambda = \frac{1}{2} \alpha^2 \omega^2 \Delta^2 \tag{1.10}
\]

Thus, we can relate the amount of dissipated energy to the shift in phonon coordinates upon excitation, leaving us with a simple model of just three parameters: reorganization energy \( \lambda \), phonon frequency \( \omega \), and coupling strength \( \alpha \). This model predicts a Gaussian distribution of energy gaps.

The FMO complex is not in the Förster Resonant Energy Transfer (FRET) or Redfield regimes. It is in a regime where the reorganization energy is almost identical to the couplings. The initial framework for solving the dynamics given an initial condition looks very similar to the Redfield framework, except now the bath dynamics occur on a timescale similar to the excitation dynamics, breaking the Markovian assumptions that so simplified the Redfield limit\[17\]. The physics devolves into a cacophonous collection of oscillators and time-periodic self-interactions, with pigments rapidly transferring energy back and forth upon stochastic resonances, which will be discussed further in Section 2.4.3.

In the ensemble limit, these dynamics can be well computed by tracing out the phonon bath. In this way, ensemble dynamics for particular initial conditions can be computed, then averaged appropriately to reconstruct 2D nonlinear optical spectra \[18\]. These spectra are dense with information, and allow further refinement of model parameters, as the dynamics of 2D spectra are much more sensitive to the choices of model parameters than the linear spectra. The combination of these 2D spectra and dynamical calculations suggest a transfer rate of somewhere between 100fs - 4ps for an excitation to cross the FMO complex. This is consistent with previous findings, but provides an even more structurally sound basis for findings that had previously been hotly debated.

The reaction center of the FMO complex is very difficult to isolate, so conventional energy transfer dynamics studies in green sulfur bacteria were not performed early on. Just recently, developments in protein isolation have allowed for these studies \[19\], greatly enriching our understanding of the FMO pigment protein complex. What is found in experiments at 77K is that 66% of the extracted FMO molecules in the study “regardless of the fact that they were physically attached to the RCC, were not able to transfer exitation energy to the reaction center.” \[19\] In the same study, the reaction center was found to perform charge separation on a time scale of 30-38ps. Free FMO molecules were found to fluoresce with a time scale of 2600ps, while the reaction center-bound FMO molecules were found to fluoresce with a time scale more commensurate with the 700ps time scale of the reaction centers only, and could be fitted to the pair of individual fluorescence curves to compute that 77% of FMO molecules transferred their excitation to the reaction center. This relatively high rate of
inactive FMO molecules may be vital to understanding the high quantum efficiency of the complex. There is also a much wider range of time scales relevant to the FMO/grc complex than in the 2D spectroscopy studies of FMO alone. Both of these facts suggest a critical and phenomenological role for static disorder.

1.1.4 Summary of modeling energy transfer in thermal environments

There are a variety of approaches to simulating these complexes, and we have only scratched the surface with these – more complex approaches involving non-adiabatics [20] and non-linear spectroscopy [21] enrich the research landscape, but are beyond the scope of this thesis.

The lesson of this overview is that the energy transfer dynamics theories are very non-linear with respect to parameter choices; e.g. small changes to DFT parameters, local electrostatics, or amounts of fluctuation can lead to qualitative changes in the energy transfer pathways. If we want to parametrize FMO from simulation data, it depends critically on getting the calculation correct in the ways that matter most. With substantial degrees of static disorder possible in FMO, it is not clear that the sub-microsecond averages from previous studies that emphasize one particular configuration of the system are sufficient to sample the ensemble quantities in question.

Furthermore, dynamical results of the FMO/grc complex suggest phenomenology in longer timescales than were previously considered important to the FMO complex. In particular, it speaks to the potential importance of anharmonic fluctuations. We do not expect Gaussian distributed disorder to spontaneously cause the observed qualitative differences between active FMO-grc and inactive FMO-grc. While the specific information available about the states of these inactive dimers is limited, the observation of these qualitative states is novel for green sulfur bacteria and may have answers in simulation.

1.2 Green Sulfur Bacteria and the Fenna Matthews Olson Pigment Protein Complex

In biology, because of the diversity of features and genes generated by random genetic mutation and the lack of selectivity with respect to many of these features, it is common that the scientific community selects particular organisms and systems to study in extreme depth – “model organisms”. Model organisms allow for the explanation of related species’ features through analogical modeling. In the study of photosynthesis, the Fenna Matthews Olson pigment protein complex [22] from C. tepidum and P. aestuarii is exactly one of those so-called “model systems.” Both C. tepidum and P. aestuarii are focused on to identify some commonalities across genera without throwing the baby out with the bath water when it comes to organism specificity. Throughout the text, we will speak of the
two species interchangeably, possibly to the chagrin of biologists. However, in the light harvesting system, the differences are quite slight. I avoid quantitative comparison across the two species, but will be less careful when it comes to qualitative features. For simulations, I use a structure and sequence from *P. aestuarii*, as it is the structure used in the reference spectral density simulations.

1.2.1 FMO Crystal Structure

The FMO pigment protein complex is a biological trimer with a large amount of patterned secondary structure. The complex as a whole has a C3 symmetry along its z axis, with a wedge-like motif reminiscent of a cheese wheel cut in three, or the current logo for Google’s “Google Chrome” web browser. The top and bottom of the protein are defined by orthogonality to the C3 symmetric z-axis and by proximity to landmarks: the top is proximal to BChl 8 and the bottom is proximal to the N-terminus. The ventral face of the monomer is the interior of the wedge and the dorsal face of the monomer is the exterior of the wedge.

There is substantial tertiary structure to the trimer even in the crystal structure. On the ventral face of each monomer, there are seven primary alpha helices that form the bulk of the surface area, while on the outside, there are fifteen large beta sheets that encapsulate the BChl a cofactors. This has been humorously referred to beta-sheet taco. The assembly of adjacent monomers is strongly polar, in fact, 70% of the contacts between adjacent monomers are polar or charged [23]. The inter-monomer associations are particularly important at the loose loop 2 region (identified in Figure 1.3 as the glass-rendered loop connecting to the leftmost alpha helix). This loop associates more with its neighbor than with its own monomer by docking to the otherwise solvent exposed BChls 1 and 2 and hydrophobic residues present in the right lateral wedge of the monomer. The loop has very minimal interaction with its own monomer from a hydrogen bonding or hydrophobic perspective because a water solvated groove separates loop 2 sidechains from sidechains in the rest of its monomer. At the top of the trimer, the protein surface forms an cone-shaped indent along the symmetry axis with a peak radius about half the radius of the trimer, and which penetrates just less than halfway through the protein. On the bottom, a similar indent is present, except this one with a shape more like a spheroid flattened in the z-direction, penetrating only a quarter of the way through the protein. These two indentations are nearly (but not quite) connected by a single-molecule-wide channel of water molecules, however, this network does connect the top cone indent to the solvated groove near loop 2, making for a very solvent-rich assembly, as seen in figure 1.4.

BChl A molecules contain a Mg\(^{2+}\) ion that is never bare. These molecules coordinate mostly to Histidine residues, but there are three exceptions: BChls 2, 5, and 8. BChl 2 has coordination to a water molecule from a four molecule long water chain that slips beneath loop 2. BChl 5 is coordinated to an O atom on the protein backbone. BChl 8 is doubly coordinated; to a backbone oxygen from below and to a Serine residue from above. That leaves these sites as obvious candidates for qualitatively different fluctuation behavior. Details of direct bonding to important BChl are summarized in table 1.1.
Figure 1.3: Dorsal view of the front-most monomer, with the majority of the structure rendered transluscently. Top, helices are rendered in pink, and bottom, beta sheets are rendered in pink. Chromophores are colored with respect to their role in the energy transfer process in analogy to the colors associated with an “energy funnel”; blues are the entry points, greens are the intermediates, and red is the exit point. Water is rendered as a blue surface interspersed with the grey surface of the other two FMO monomers.
1.2.2 FMO in its natural habitat

The FMO monomer is just one of the units in the light harvesting antenna from the green sulfur bacteria. The other primary units are:

- **Chlorosome**: a large pouch densely packed with relatively blue BChl C chromophores that increase the total absorption cross section for the complex.
- **Baseplate**: the surface of the chlorosome that faces the inner membrane. The baseplate is made up of aggregated protein CsmA bound to BChl A molecules at a rate of 1 BChl A per CsmA [24].
- **FMO**: The connection between the baseplate and the inner membrane and the embedded reaction center.
- **grc**: The green reaction center, just recently molecularly isolated. Binds two FMO trimers opposite to one another [24].
- **inner membrane (IM)**: the bacterial cell membrane in which the grc is housed.

The FMO monomer lies in between the chlorosome baseplate and a membrane-bound reaction center, and much of the research into FMO proceeds with just this knowledge. However, technological developments in high-resolution microscopies from the last five years
Figure 1.5: Cartoon of the parts of the green sulfur bacterial light harvesting antenna at a range of details, moving progressively closer to the FMO trimer.
Table 1.1: Significant contacts and hydrogen bonds with BChl A molecules in *P. aestuarii* based on data from crystal structures. Residue numbers with an overbar (e.g. Ile137) are from a monomer adjacent to the referenced BChl’s monomer of convention.

<table>
<thead>
<tr>
<th>BChl</th>
<th>Important residue</th>
<th>type of contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>His110 (NE2)</td>
<td>Mg coordination</td>
</tr>
<tr>
<td>1</td>
<td>water (OH)</td>
<td>ring I hbond</td>
</tr>
<tr>
<td>2</td>
<td>water (OH)</td>
<td>Mg coordination</td>
</tr>
<tr>
<td>2</td>
<td>Ser72 (OG)</td>
<td>ring I hbond</td>
</tr>
<tr>
<td>2</td>
<td>Ile137 (N)</td>
<td>ring V hbond</td>
</tr>
<tr>
<td>2</td>
<td>Tyr138 (N)</td>
<td>ring V hbond</td>
</tr>
<tr>
<td>3</td>
<td>His298 (NE2)</td>
<td>Mg coordination</td>
</tr>
<tr>
<td>3</td>
<td>Tyr15 (OH)</td>
<td>ring I hbond</td>
</tr>
<tr>
<td>3</td>
<td>Ser40 (N)</td>
<td>ring V hbond</td>
</tr>
<tr>
<td>4</td>
<td>His290 (NE2)</td>
<td>Mg coordination</td>
</tr>
<tr>
<td>4</td>
<td>Tyr345 (OH)</td>
<td>ring V hbond</td>
</tr>
<tr>
<td>5</td>
<td>Leu242 (O)</td>
<td>Mg coordination</td>
</tr>
<tr>
<td>5</td>
<td>water (OH)</td>
<td>ring V hbond</td>
</tr>
<tr>
<td>6</td>
<td>His145 (NE2)</td>
<td>Mg coordination</td>
</tr>
<tr>
<td>6</td>
<td>Trp184 (NE1)</td>
<td>ring I hbond</td>
</tr>
<tr>
<td>6</td>
<td>Arg95 (NH2)</td>
<td>ring V hbond</td>
</tr>
<tr>
<td>7</td>
<td>His297 (ND1)</td>
<td>Mg coordination</td>
</tr>
<tr>
<td>7</td>
<td>water (OH)</td>
<td>ring I hbond</td>
</tr>
<tr>
<td>8</td>
<td>Tyr123(O)</td>
<td>Mg coordination (below)</td>
</tr>
<tr>
<td>8</td>
<td>Ser168 (OG)</td>
<td>Mg coordination (above)</td>
</tr>
<tr>
<td>8</td>
<td>Asp160</td>
<td>ring I adjacent</td>
</tr>
<tr>
<td>8</td>
<td>Arg125</td>
<td>ring I adjacent</td>
</tr>
<tr>
<td>8</td>
<td>Phe176 (O)</td>
<td>ring V hbond</td>
</tr>
</tbody>
</table>

have allowed us to learn a lot about the size, density, and structure of green sulfur bacterial light harvesting systems in vivo. Cells are complicated machines, and the exact role of each molecule of the cell is merely our best extrapolation of function from the facts we can measure. Thus, it is important to lay out the state of the art knowledge of structure from complexes adjacent to FMO.

**Chlorosome** Chlorosomes are cellular organs spread out against the inner membrane of the green sulfur bacteria, as shown in Figure 1.6. Chlorosomes range from 140nm to 220 nm in length and 30nm to 60nm in width [25]. Chlorosomes have large assemblies of BChl C molecules aggregated into collections of rod-like filaments with cylindrical symmetry. This has been demonstrated through the invariance of single-chlorosome 2d polarization spectra which imply a degree of pseudo-rotational invariance [26]. The only competing model for aggregate arrangement, undular lamellae, would show a strong dependence on orientation for
Figure 1.6: A spatial structure of chlorosomes in the green sulfur bacteria cells from cryo-electron tomography. Image reproduced from Kudryashev et al. [28]

polarization spectra [26]. They have a baseplate structure that functions as a significant part of their membrane. This baseplate is present only on one side – the side facing the cell’s inner membrane [27].

The mechanism of chlorosome self-assembly has been studied through a variety of knock-out experiments. Eliminate BChl C production and structure assembles as an “empty bag”; remove CsmA and the chlorosomes fail to form at all [29]. Pedersen et al. (2010) propose a model of chlorosome assembly that follows along this pathway that suggests FMO binds to the grc first, CsmA binds to FMO, ligands bind to CsmA, and finally BChl C pigments begin to aggregate and the membrane closes around them [29]. This implication of FMO in the assembly of the chlorosome suggests some potential avenues of study for protein binding specialists.

**Baseplate** A separate population of blue-shifted absorption is identified as the [29] baseplate of the chlorosome. The baseplate consists of a paracrystalline CsmA conjugated to BChl A molecules, beginning the energy funnel towards the FMO molecule [30]. The crystal structure of the CsmA protein is shown in Figure 1.7. This protein’s C-terminus has notable intrinsic disorder, yet is hypothesized to bind to the upper surface of the FMO trimer [30]. CsmA proteins dimerize strongly in opposing directions along first helix, and then oligimer-
Figures 1.7: spatial construction of the CsmA molecule, showing hydrophobes in white and hydrophiles (basic, acidic, and polar) in purple. In blue cartoon transparent overlay, the variety of disordered structures resolved in the PDB file are rendered simultaneously. Image constructed from PDB file 2K37 in VMD.

ize into larger paracrystalline structures [29]. This strong dimerization, para-crystallinity, and binding to FMO proteins places a strong restriction on the resulting geometries of the fully assembled structure.

**FMO** The FMO protein has a C3 symmetry axis that points perpendicular to the membrane upon which it rests [31]. The leading hypothesis is that this C3 symmetry axis points with BChl 3 towards the membrane embedded reaction center [32]. This is particularly consistent with the the eighth bacteriochlorophyll molecule attached to the top of the trimer – which would ostensibly be in contact with the baseplate – being blue shifted with respect to the others [33].

The FMO protein binds very tightly to the grc in a dimeric fashion. These dimers then dimerize again, forming FMO trimer-dimer-dimer squares in the electron microscopy images [24]. Center-to-center distances along both dimerized edges and inter-dimer edges are about 12nm, while inter-cluster center-to-center distances average to about 25nm [24]. The arrangement of the FMO trimer and the grc have been identified through scanning TEM, and they form a tightly bound complex with a tail-like feature adjoining the two at the bottom, and a system of PscB, C, and D proteins flanking the side of the FMO trimer from the center of the grc [34].

**green reaction center** The green reaction center is elusive in comparison to reaction centers from other photosynthetic bacteria like LH2. However, glimpses of it have made their
CHAPTER 1. ELECTRONIC ENERGY TRANSFER IN PIGMENT PROTEIN COMPLEXES: EET IN PPCS

Figure 1.8: Phylogenetic tree of a few species. Pairs in this sample are fairly evenly distributed among a rich diversity of species within the Chlorobiacea family. Phylogenetic information from Imhoff 2002 [36], 2003 [37].

way into the literature, most notably, the scanning TEM images of the FMO putatively bound to the reaction center from [34]. Recently, the reaction center has been reliably purified [35], opening the door for detailed spectroscopic studies to take place [19]. From these studies, we learn that the reaction center is made up of four proteins, PscA through PscD. Mass spectrometry can resolve many of the native bonds between the reaction center and the FMO complex, revealing information about the relative orientations without solving for a structure, reaffirming the docking motif suggested by scanning TEM [35]. As of now, there is not sufficient data to reconstruct the protein structure of the grc.

1.2.3 Speciation and Mutation

Speciation of the FMO protein and related complexes is a very broad topic. However, speciation can be used as a tool to augment atomistic theory by corroborating the importance of a feature through sequence conservation, or by invalidating the relevance of a feature through total non-conservation. Further, the range of allowed mutations can guide our understanding as well; mutations to similarly sized residues might suggest the importance of simple packing, while mutations to residues with similar charged groups might suggest the importance of direct interactions and hydrogen bonding.

Many experiments will compare different species of green sulfur bacteria to see how they compare, as the marginal cost of modifying the procedure to include another species is often worth the external validity that is gained by showing something to be true across species. Thus, there already exist many differences identified between C. tepidum and P. aestuarii, the two main model species.

Furthermore, it is pretty straight-forward to mutate a particular residue in an organism to induce a specific targeted change to test a hypothesis. Unfortunately, the search space for these single-replacements is very large, so these experiments will rarely be conducted randomly. Still, these modifications and their results provide an even broader base to form theories about the green sulfur bacteria’s light harvesting system.
CHAPTER 1. ELECTRONIC ENERGY TRANSFER IN PIGMENT PROTEIN COMPLEXES: EET IN PPCS

Figure 1.9: Partial aligned sequences of several Green Sulfur Bacterial species. Amino acids 1-10 show the good preservation of the hydrophobic tail, including the inclusion of an additional hydrophobe in Pld. phaeum. Sequence 149-187 included from Tronrud 2009 [38]. This sequence shows strong conservation between 148 and 164, as well as between 169 and 187.

1.3 Overview of text

In the following sections, we will use an atomistic bath combined with an empirical electrochromic electronic-bath coupling to assess the rigor of bath-integrated approaches to excitation dynamics and light harvesting theory in the FMO pigment protein complex. First, though, we must have a baseline of theory for understanding what aspects of modeling in the complex are incomplete. Thus, Chapter 2 is devoted to a thorough exploration of the variety of models used to analyze the FMO pigment protein complex. With that baseline, we spend chapter 3 homing in on the aspects of protein simulation that are necessary to appropriately gather robust statistics, especially on such a large complex. We then use these samples to compute statistics on the energy gap in Chapter 4 to analyze the validity of the spectral density approximation.

The energy gap is a difficult object of study because of the long time scales required to converge protein simulations and the complexity of its computation. However, we believe that this work constitutes a significant contribution to the understanding of the extent to which modeling assumptions that go into electronic energy transfer models are well grounded in physics.
Chapter 2

Scientific Models, Scientific Theories

The fact that photosynthesis bridges scientific disciplines (e.g. biology and physics) and methodologies (e.g. theory and experiment) makes it a scientifically rich frontier. In order to make headway in a field like this, we must be intentional in our scientific process, and aim to bridge the separation between theory and experiments. Here, we discuss modeling, a scientific framework that supersedes this divide, and allows us to unify our understanding of photosynthetic research across experimental and theoretical results.

If we take a moment to reflect on the job of the scientist, we see that it is not only to develop empirical knowledge and theories, but also to construct and define what constitutes the scientific practice. This section outlines my scientific approach and my beliefs about how science is performed, as I believe scientific beliefs constitute a large amount of the character of a practicing theoretician.

The goal of science is largely seen as a pursuit of truth, but more realistically, science is a system of reasoning that allows us to generalize predictive methods from one context to another. Abductive reasoning is the often overlooked partner to logic and reasoning that is uniquely essential to generating explanations from an observation by generalizing our prior experiences. Such a generalization requires some belief in an ontological significance of such descriptions. Models are the vehicle by which we perform generalizations and make tangible our ontological perspective.

But if we accept the premise that science is performed through abductive reasoning, we must also accept the conclusion that science is not strictly objective. In this acceptance, we must also accept that a discourse of our scientific practices is tantamount to reducing bias and realizing our potential for discovery.

In this chapter, we will begin by exploring framework which is scientific modeling. Terminology and viewpoints that aid in parsing theories will be presented. Then, for the remainder of the chapter, the focus will be on the three major categories of modeling involved in photosynthesis – protein, electronic, and thermal quantum mechanical. Each of these fields is multilayered, and must be broken into its constituent assumptions of we hope to synthesize a principled interdisciplinary model of photosynthetic phenomena. The topic will then be summarized, with a focus on how modeling perspectives can blend the constituent assump-
tions of these three fields into a model that assesses the extent to which photosynthetic phenomena depend on nonlinear or anharmonic features of the protein.

2.1 Introduction to scientific models

Scientific modeling neatly encapsulates what we think of as scientific discoveries, and will be the framework through which results in this dissertation will be presented. Theories like the Hierarchical Equations of Motion and experiments like Fluorescence Line Narrowing are quite distinct in terms of practice, but still aim to contribute to a model of the same physical practice. By extracting the assumptions made to construct each model, we can identify the extent to which the analyses are corroborate certain assumptions while excluding others. We can make headway into identifying points of improvement and weakness by analyzing results through a modeling lens.

A scientific modeling perspective makes it much easier to blend and combine scientific understanding across disciplines. This is of special interest for a fundamentally interdisciplinary research field like photosynthesis research. While I think all scientists should take the time to contemplate the utility of scientific modeling, explicit exposition on scientific models is of direct relevance to fundamental research in photosynthesis. Scientific modeling also discourages the partition between theory and experiments – either can provide justification and reasoning into the validity of a particular model.

Scientific modeling deserves a moment in the spotlight anyway, but explicitly addressing scientific modeling in this dissertation serves two purposes: First, to fully leverage the multidisciplinary nature of photosynthesis, and second, to underscore the importance that this explicit thought process had in my research process.

To discuss scientific models, we will address the following aspects of scientific modeling over the following sections: (a) what scientific models look like in practice; (b) the utility of scientific modeling in science education; and (c) model development practices.

2.1.1 Models in practice

The “scientific model” does not have a simple, standardized definition (which is to speak more of the quality of discourse surrounding scientific practice than of the utility of scientific modeling). While not everyone defines modeling consistently, it is likely that what is and what is not a scientific model is consistent across practitioners. Still, in the pursuit of clarity, we will use a working definition.

A decent summary of a scientific model is a set of assumptions that produce predictions about the physical world. We can define a model itself as a predicative set of assumptions. Under this definition, we can refer to both a mathematical equation and a procedure (such as a baking soda volcano) as a model. The predictions of a model can be quantitative or qualitative, too. The sophistication of models that we will encounter in this dissertation will
primarily be mathematical in nature, but there is no reason to treat the assumptions made in a mathematical model as any more sacred than those from the baking soda volcano. “Modeling” thereby consists of developing and validating models. Modeling is the scientific method in action. And just like the scientific method, the most identifiable feature of modeling is iteration.

The most confusing analog to a “model” is a “theory” – so much so that in practice the two words are often interchanged freely. Because a theory may exist independently of any modeled phenomenon, a theory is a separate entity from a model. A theory is simply a set of axioms that can be manipulated mathematically. A theory is a model just as much as a steam engine is a locomotive. Often, though, the lines between theory and model are blurred because a theory will be developed alongside a model by applying the theoretical axioms to that model. The model is then able to leverage our physical intuition to teach us about the implication of the axioms themselves. It is then sensible to make some distinction between a prolific theorist, who develops particularly novel theories, and a theoretician, who leverages a tremendous library of theories to build successful models. Realistically, the terms are more often used interchangeably than to signify difference.

In molecular simulation communities, the particular atomic topology, parameters and functional forms selected for a simulation are referred to collectively as the model. While this is fully consistent with our usage of a model, it is also unnecessarily limited to the simulation stage of the process. It is reasonable to say that the entire predictive process, from simulation through theory through analysis, constitutes a model.

In statistics communities, models refer to the assumptions made about the probability distributions from which the data is generated. The term shows up frequently, particularly in the context of model selection. Model selection refers to choosing some aspect of your procedure that cannot be wrapped into the statistics framework of estimation (e.g. choosing between a quadratic or cubic regression). This again fits nicely with the definition of model from above, and is actually quite inclusive of most modeling choices that are possible. However, it is limited in that it only makes sense in a statistical framework; a model like the ray model of light would be difficult to interpret as a statistical model. It is even further limited by the fact that its usual usage tends to be just with respect to choosing a value of a parameter for your analysis, e.g. selecting the number of clusters in a clustering analysis.

2.1.2 Aside: Scientific modeling in the classroom

Modeling is on the rise as a framework for teaching science in primary and secondary education [39] [40]. The changing curricular content of the common core push for science competencies that reflect the influence of modeling philosophy and practices, even if the framework itself is not directly referenced. Educational benefits of modeling are under research and scrutiny, as a shift towards a completely modeling-based curriculum would be a major academic overhaul, and would demand clear enough gains to justify the cost in teacher reeducation and curricular redevelopment.
Still, an alternate approach to science education forces us to answer the question: what is the objective of teaching science? Our current system excels transmitting factual information about the world. In this pursuit, it falls short of developing good scientific epistemology in students while simultaneously failing to promote the scientific career into an equitable meritocracy. There is significant evidence to believe that modeling makes strides towards these latter two objectives without sacrificing the first. Equity is a well understood concept, epistemology, less so – let’s take a moment to define it.

Epistemology is broadly the field of philosophy which asks how we define the substance of knowledge. While this metaphysical question is interesting, it is not particularly relevant to our science education. Rather, we care much more about applied epistemology, which is the study of how people make the decision of what constitutes knowledge and how it is distributed in the world. For example, the student belief that a textbook is the source of knowledge is a useful epistemology for a young student who benefits from a “monkey see, monkey do” approach to education which centers on a trust of authority. However, we expect fully developed scientists to approach new research projects with a qualitatively different epistemology than this first grade child or else they would not believe that it was even possible to succeed at their endeavor of creating new knowledge.

Modeling fails to exceed the expectations set by conventional teaching with regards to rate of learned information (though it frequently comes close), but it far surpasses conventional teaching with respect to developing productive epistemologies and equitability. One interpretation of why modeling is successful in these regards is because it fundamentally resembles science itself. If our classroom teach the same lessons of curiosity, humility, and adaptability that scientists must learn to succeed, students of all types are able to develop the belief that they too can be scientists, and can begin to perform that identity in other aspects of their lives.

2.1.3 Developing models

Our primary interest in modeling as practitioners is how we can leverage the understanding of modeling to encourage the productive generation of research through abductive reasoning. Our awareness of modeling provides clear utility either when we use key modeling motifs as intentional strategies, or also when we identify modeling motifs within the strategies we are already using. It is therefore useful to construct vocabulary around modeling motifs so we can use that terminology as either a launch pad for new strategies or a landing pad for processing and verbalizing a strategy already in use.

Many types of abductive model reasoning have been described in the literature, and likely many more than this are productively used in practice. N. Nersessian describes three types of modeling behaviors that have been demonstrated by her work to generate conceptual change in science. These are analogical modeling, visual modeling, and thought experimenting. These tools – analogies, visuals, and thought experiments – are explicit in educational settings, but they are not often identified by name in other contexts. In
CHAPTER 2. SCIENTIFIC MODELS, SCIENTIFIC THEORIES

a modeling picture, each can be thought of as a form of reasoning unto themselves, and therefore each deserves explicit recognition for their role in scientific reasoning.

Modeling behaviors are also codified in the scientific community through the naming of models. The types of models that are seen in physics have a range of nomenclatures, but the reason for selecting one word over another is rarely discussed. In particular, a non-exhaustive list of models is: theory, law, category, principle, rule, postulate, and criterion. An example of each is found in the following list of examples: linear response theory, Newton’s law, periodic table (i.e. categories), principle of superposition, Hund’s rule, postulates of quantum mechanics, and the Rayleigh criterion. Each of these identifiers says something different about the model.

- A theory is a broad framework intended to generalize semi-absolutely, under a small set of assumptions
- A law is an empirical relationship constructed for interpolation and extrapolation
- A category is an analogical model that identifies connections along a particular "axis"
- A principle is an abstract framework to be applied as a form of reasoning for certain approximations
- A rule is an approximation that gives insight into the underlying mechanisms without being strictly true
- A criterion is an approximation that can be used to analyze and understand data in real time

2.1.4 Summary of models

This is a shallow overview of modeling, but it suffices to: (a) introduce the framework to aid in understanding photosynthesis research’s multidisciplinary and multimodal frontier (b) make explicit the research approach taken in this dissertation. With this language of modeling, we can begin to dissect the assumptions and constituent layers present in the various models of photosynthesis, and bridge the disconnect between theory and experiment, as well as the disconnect between biochemistry, chemistry, and physics.

There is a lot more to talk about with regards to the proper ways of evaluating efficacy of curriculum, particularly with respect to abstract objectives like epistemology. For more on this and similar topics, consult the work of D. Hestenes [42], A. Elby [43], D. Hammer [44], and N. Finkelstein [45]. For example, an important development by D. Hammer in educational epistemology research posits that epistemologies are not fixed across time and context [44][46]. Hammer draws upon the concept of resources, which is a common framework applied to student approach to new concepts in physics, to build by analogy the concept of epistemological resources. This theory of epistemological resources states that epistemologies are contextually applied and present even when they are not explicitly and consciously developed. Thus, a student in a classroom, when confronted with information in a physics class that contradicts their lived experiences, may leverage an epistemology of abstract thinking that says something to the effect of “sure, this works in theory, but this isn’t
really how the world works.” Research in modeling and development of scientific identity in the context of the classroom is active and exciting.

With that, let’s continue to the exploration of models within three major frontiers of photosynthesis research: (a) proteins, (b) optically accessible excited states, and (c) quantum dynamics in thermal environments.

### 2.2 Models of the protein

Proteins are dynamically and structurally complex molecules that are ubiquitous in serving as scaffolds for light harvesting pigments. Despite their biochemical activity, they are fairly inert for EET – in EET, their main role is to hold the pigments in place while molecular energy occurs.

While the majority of protein research has gone into the sequence-structure problem (the “protein folding” problem), there is important research on the impact of fluctuations within well-folded proteins. This research connects closely to polymer physics. However, proteins, unlike polymers, have a blend of specific in addition to their non-specific interactions, which makes them much more difficult to analyze as a generic class of compounds. We depend on rigorous statistical inference to analyze their behavior and dynamics.

In this section, we investigate several of these techniques to understand proteins, primarily among the following major categories: physical structure, fluctuations and dynamics.

#### 2.2.1 Physical Structure: Crystal Structures

The majority of information we have about the chemistry of folded proteins comes from what are called crystal structures. This nomenclature comes from the typical methodology of extracting a purified protein by crystallizing and precipitating it out of a solution, at which point the crystal is imaged through x-ray diffraction. Because of the periodicity of the crystallized structure, it is possible to construct an electron density map from the diffraction pattern by determining the phase of the measured k-space Fourier amplitudes. With this electron density map, optimization routines can find structures that respect the molecular topology known from genetic sequencing, and thus, create a space-filling representation of the protein.

While the direct (and conservative) method of building crystal structures from electron densities is a tremendous start, additional prior information about protein structure can be added with advanced structure building methods. The first of these methods is comparative structure modeling. This is where solved structures that share genetically homologous segments with the protein of interest are used as reference structures around which to perform optimization searches [47]. The second of these is protein scoring. These systems use what are basically implicit solvent force field energy functions to assign a score to a structure. This score is optimized instead (or alongside) the electron density map [48]. Though these functions have fairly coarse resolution, they have been found to behave similarly to...
Figure 2.1: An electron density mesh for small ligand within the 3eoj protein crystal structure as viewed by the JSMol molecular web viewer. While the crystal structure predicts distinct valence clouds for the protein and the ligand, there is an appreciable amount of valence interaction between the Mg+ atom on the ligand and the HIS110 group that can be observed directly in the crystal structure.

direct molecular dynamics sampling with respect to the function's preference for entropically preferred, but non-optimal, alternative conformations [49].

Crystal structures, when rendered in a molecular visualization program like VMD, are a great example of visual modeling. There are many canonical visual representations (like secondary structure representations and color maps based on residue hydrophobicity) that identify structural motifs and allow for the formation of hypotheses about the mechanisms of the protein in question. These visual protein models are a great example of a model that is crucial to a broad range of research but also autonomous from the “hypothesis testing” class of models. Crystal structures have a strong track record of inspiring new scientific ideas, but they do not fit the simple hypothesis-test-conclusion version of the scientific method. Therefore, their ubiquity supplies more evidence of the importance of discussing modeling practices.

To get a handle of what crystal structures have been used to conclude and hypothesize, we can consider the FMO crystal structure as a case study. All of the structure publications identify essential features of the crystal structure. In particular, contacts to the BChl A molecules were identified, and the degrees of hydrophobicity on various surfaces were analyzed and visualized. More detail about this crystal structure in particular can be found in Chapter 1.2.1.
2.2.2 Physical Structure: Molecular force fields

With the crystal structure of FMO in hand, we still must have an energetics in order to perform the simulations. There are a variety of force fields available to simulate proteins, but none of the commonly used force fields are better than the others in a fundamental way. Instead, each is an approximation, a model designed to fit particular observed data. The quantification of the quality of these models is particularly challenging, as the sources of error are many, so the number of benchmarks required is high [50] [51] [52].

Usually, detailed information for each atom type comes from some type of quantum chemistry calculation. From there, the parameters are refined to capture properties of macroscopic observables, e.g. heat capacity. To test the quality of these parameterizations, a collection of exemplary systems are simulated, and some microscopic experimental observable is compared between the simulations and experiments. This forms a full loop – instead of parameterizing based on the microscopic information, we can use it to test against. We face a bit of a paradox in this regard: if we parameterize against all available data, we risk over fitting and we have no independent data to validate against. On the other hand, we face the effect of compounded errors if we opt out of using these real comparisons to restrain our parameterization.

Despite the care taken in building forcefields, force fields are typified by their approximate approach to a problem where the exact theory exists, but is intractably slow. Because we must sacrifice accuracy to accelerate calculations sufficiently, we need to make design choices in how we make these approximations. For example, protein force functions are not optimized to be interpolations of the bonded forces that arise from quantum mechanics. This is because we don’t necessarily seek to minimize the error between the quantum mechanical potential and the force field potential. In fact, we prefer a representation that captures the emergent properties of the statistics over long times over one that will optimize bond dynamics on femtosecond time scales.

A characteristic example of the power of approximation is in harmonic polymer models. A reasonable polymer force field can be constructed from a string of beads connected by “unphysical” bonds of zero length. We gain a huge benefit from this choice: the model has only a single parameter to be fit to the data of interest. For this incredibly crude model, many properties of the polymer (e.g. end-to-end distance statistics and scaling, radius of gyration, dissipation rates) can be computed, and thereby used to hone model parameters. In this light, we are in a bit of a struggle when it comes to molecular simulations; should we push towards a complex and atomistically precise model, or towards a simpler model that helps to identify trends and emergent behavior? We are fortunately, as a community, not bound to just one choice. We can expect to find important physics at the frontier of both approaches.

Protein force fields are of incredible complexity, and they are discussed in more detail in 3.1.1. However, for FMO in particular, we are fortunate that the approximations inherent in the most popular protein force fields have already been explored [16]. The result as to which of CHARMM or AMBER is preferred is inconclusive, but the research at the very
least provides a simple quantification of the force-field dependent variability, which serves as a good proxy for quantifying the external validity of atomistic results. Atomistic simulations are often so costly that comparisons of these types are rarely performed, particularly because though they provide vital information about the validity of the simulations, they do not provide any particularly riveting conclusions.

2.2.3 Fluctuations: Free Energy

Free energy is an indispensable generalization of energy that comes from statistical mechanics and which allows us to consider how our observations are impacted by entropy, and is such a useful and important generalization that it has even been adopted in fields like neuroscience and machine learning thanks to its similarity to concepts in information theory.

Historically, though, free energy was developed in the context of thermodynamics. In thermodynamics, free energy is used to identify reactions as spontaneous or non-spontaneous — thanks to entropy contributions — and to explain why some spontaneous processes absorbed heat and some released it. Free energies made it possible to factor in the impacts of other bulk effects, e.g. volume changes, on the spontaneity of a chemical process. Thermodynamic free energy $F$ is defined as

$$F = U - TS,$$  \hspace{1cm} (2.1)

where $U$ is the macroscopic potential energy of the system, $T$ is the macroscopic temperature of the system, and $S$ is the macroscopic entropy of the system.

Thermodynamics was later theoretically subsumed by statistical mechanics, and so the concept of free energy was generalized. In statistical mechanics, the concept of an ensemble is introduced to describe the quantities which are free to be interpreted probabilistically, e.g. volume. In turn, probabilities of various observations within these ensembles come to correspond to the free energy of the ensemble. In this generalization, we are also permitted to consider finer and finer features as characteristic of the “ensemble”, and thereby generalize free energy to microscopic degrees of freedom in a way that thermodynamics cannot. Thus, we simultaneously build statistical mechanics in analogy to the tools developed in thermodynamics while we fall back on thermodynamics as a set of rules: if our calculations from statistical mechanics disobey a thermodynamic rule, we know that we have failed to perform our calculation correctly. Statistical free energy (or more accurately, probabilistic free energy) is defined as

$$Z = \int d^{3N} x \ e^{-\beta U(x)} \hspace{1cm} (2.2)$$

$$F = -kT \ln(Z), \hspace{1cm} (2.3)$$

where $F$ is again the free energy, but now the statistical mechanics variant, $kT = 1/\beta$ is the thermal energy unit of the boltzmann constant times temperature, and $Z$ is the partition
function, defined by an integral over $N$ atomic degrees of freedom defined to interact with a potential energy $U$.

In statistical mechanics, an ensemble is the collection of states and their corresponding probabilities of occupation given a fixed set of external parameters. Ensemble thinking can be very difficult, as it is easy to fall back on reasoning that relates our system’s instantaneous configuration to an energy, as in classical mechanics. Of course each configuration does have an energy, but the free energy that is thermodynamically relevant is only a property of the entire ensemble, not simply the instantaneous configuration in question. Our choice of ensemble now determines the relevant free energy. The convention of practitioners is to assign a simple three-letter self-descriptive name to each ensemble: most commonly NVE, NVT, and NPT, which correspond to constant number/volume/energy, number/volume/temperature, and number/pressure/temperature, each of which then correspond to the energy, the Helmholtz free energy, and the Gibbs free energy, respectively. Thus, by allowing our extensive variable to fluctuate (among these cases either entropy or entropy and volume), we must include that canonical pair in the free energy.

In practice, all but the simplest cases of free energy calculations, it is impossible to absolute free energies, so instead, we compute differences in free energies. We sample the most populated microstates under a continuous set of external parameters, we can compute the small free energy difference between adjacent parameters with decent accuracy as the two ensembles will have relatively similar probability distributions, and thus we only need estimates of the probability in regions where the probability is large. In situations where one ensemble has substantial probability mass in regions of configuration space that the other has roughly no probability mass, we must unfortunately still have good estimates of the near-zero quantity as the log probability will be multiplied by the original probability to get the free energy change. Thus, to estimate free energy differences reliably, we must move sufficiently slowly such that our probability distributions do not change appreciably between adjacent ensembles.

The equation for difference in free energy $\Delta F_{AB}$ across two potential functions $U_A$ and $U_B$ (corresponding to ensemble $A$ and $B$) is defined as

$$\Delta F_{AB} = -kT \ln \left\langle e^{\beta(U_A - U_B)} \right\rangle_A,$$

where $\left\langle \cdot \right\rangle_A$ denotes an average of the quantity “.” sampled with the probabilities of ensemble $A$. Assuming $A$ and $B$ are sufficiently similar ensembles, we can notice a competition in sources of smallness. On one hand, the greatest contributions to the expectation are those where $U_B$ is greater than $U_A$; on the other hand, we will only sample regions of state space that are sufficiently populated in the $A$ ensemble. This overlap region brings a previously intractable absolute free-energy calculation into a quite tractable difference in free energies.

The free energy used most frequently in practice is one conditioning in which the free energy integral is conditioned upon a microscopic coordinate (a “collective variable,” or CV). Direct sampling will provide this free energy, but inefficiently. If we consider an extreme but theoretically well-defined potential parameterized by $\bar{q}$ such that $U(q) = -\delta(q - \bar{q})$, the
Figure 2.2: The difference in the free energy is much more practical to compute than the absolute free energy because the only region that contributes to the integral is the overlap regions, and contributions from “tail” regions of A and B are strongly attenuated. The overlap region is shared symmetrically.

The resulting free energy becomes

$$Z(q) = \int d^3N \, x \, e^{-U(x)} \, \delta q(x) - \bar{q}$$ (2.5)

$$F(q) = -kT \ln Z(q).$$ (2.6)

In this way, we have generalized free energy to be a function of an arbitrary collective variable. Noting that in conditioning the configuration integral on the value of the order parameter, we can factor out the impact of any biasing potential that depends on q, $U_B(q)$, by computing the biased ensemble free energy $F_B(q)$,

$$Z_B(q) = \int d^3N \, x \, e^{-U(x)+U_B(q)} \, \delta q(x) - \bar{q}$$ (2.7)

$$Z_B(q) = Z(q)$$ (2.8)

$$F_B(q) - U_B(q) = F(q),$$ (2.9)

and thereby we can utilize samples from the biased ensemble to reconstruct the CV-generalized free energy in the unbiased ensemble.

Throughout the text this CV-generalized free energy simply will be referred to simply as the “free energy.” With this tool for describing probability of microscopic states in hand, we can discuss one of the most frequent applications of this free energy.
2.2.4 Fluctuations: Linear response

The utility of this generalized free energy is that it allows us to succinctly describe any manner of dynamically non-linear systems by looking at their response. This approach comes with one condition: that we can find a CV of interest that has an \textit{energetic} coupling that is linear with the CV. If we are free to select the CV, as we often are in computation, this is not a strong constraint at all. In a physical scenario, though, the CV may be tightly specified by the coupling mechanism, in which case more care must be brought. Furthermore, even computationally, we are not presented with the optimal CV, and discovering a suitable one may be considerable effort. Despite these caveats, it is frequently possible to construct a satisfactory CV. Let’s proceed under the assumption that we have such an entity.

We can then consider the energetics of linearly biased ensembles parameterized by the coupling coefficient $k$ of order parameter $X$, where the unbiased Hamiltonian is $H_0$. This Hamiltonian has a free energy $F_0(X)$ and a probability distribution $p_0(X)$. In particular,

\begin{align}
    p_k(X) &\propto \int d^3x \ e^{-\beta H_0(\bar{x})+kX(\bar{x})}\delta(X(\bar{x}) - X) \\
    &\propto e^{kX(\bar{x})}p_0(X) \\
    F_k(X) &= -\ln(p_0(X)) + kX + \text{const.} \\
    &= F_0(X) + kX + \text{const.}
\end{align}

$F_k$ is the free energy in the new ensemble, and we see that it only depends on the free energy in the original ensemble, not on any additional averages over $\bar{x}$. Thus, measuring the free energy $F_0$ of the biased ensemble is a complete reduction of the system information with respect to the probability distribution in the biased ensemble. Thus, when you interact with macroscopic order parameter $X$, you cannot determine just from those interactions whether the forces in your system are a result of the entropy of the order parameter or as a result of the mean energy of the order parameter, and in some ways, you simply don’t care. This is the meaning of a “statistical spring,” it is an order parameter whose underlying microscopic energetics are irrelevant because we are able to know that the free energy is quadratic. In this way, the overwhelming majority of order parameters act like springs because their mean value is also often their minimum position, and the response of the mean to linear bias is linear.

Thus, we have the theory of linear response in statistical mechanics. For quadratic free energy $F_0$ with mean $\langle X \rangle = \bar{X}$ and curvature $c$, we have

\begin{equation}
    F_k(X) = \frac{1}{2}c \left( X - (\bar{X} + k/c) \right)^2 + \text{const.}
\end{equation}

For bias strength $k$, the shift in the mean is $k/c$, clarifying exactly what is meant by linear response. The other way to interpret linear response is with respect to the alternate name for the free energy: the potential of mean force (PMF). The quantity $\frac{\partial F_0}{\partial X}$ becomes an expectation value of $\frac{\partial H_0}{\partial X}$, which is the force acting on the collective variable $X$. Our new minima will be...
where the biasing force $kX$ is equal to the mean force because this is the only value of the order parameter (for a quadratic free energy) for which the mean force is zero.

### 2.2.5 Fluctuations: Modeling discrete photon absorption with linear response

While it is not unusual to have a system with a reasonable linear control, we can still use linear response when the control is discrete. This is the situation we are dealing with when we confront the environmental shift to the energy gap, $g$. This perturbation is always either on or off, thus we can write our Hamiltonian as

$$H' = H_0 + \lambda g,$$

(2.15)

where $\lambda$ is the external control parameter that takes either the value of 1 when an excitation is present or the value of 0 when an excitation is not present. This allows us to determine the range of precision we want for our sampling, and to describe the dynamical effects of photon absorption.

Looking at a free energy for a hypothetical zero-mean gaussian distributed energy gap, the form is

$$F_\lambda(g) = \frac{1}{2} \beta \langle g^2 \rangle g^2 + \lambda g.$$  

(2.16)

Under perturbation of $\lambda = 1$, the mean shift in the gap is $-\frac{\sigma_g}{kT} \sigma_g$. Thus, is is necessary to sample at least $\frac{\sigma_g}{kT}$ standard deviations from the mean in the negative direction, and
likely at least two more to account for natural fluctuations in the perturbed ensemble. This quantity is actually equal to the reorganization energy mentioned in section 1.1.3, creating a relationship between the reorganization energy $\lambda_R$, the spread of the distribution, and the temperature of $\lambda_R = (\beta \sigma_g) \sigma_g$.

Note that the direction of shift follows intuition despite very little information being supplied to the model. This simple model predicts that under optical excitation, energy is dissipated from the excited state by reducing the mean energy gap. This is slightly less shocking when we consider that these results depend on the validity of the linear approximation as well as the linear relationship between the ground and excited ensembles. Still, as long as these assumptions are not too imprecise for the system under consideration, this leading term in the mean shift under excitation will be negative.

2.2.6 Dynamics: Fluctuation-Dissipation

We have so far covered the energetics and determined that the energetics do not require explicit reference to the underlying microscopic variables. Is the same true for the dynamics? The general answer is: not necessarily. However, within a linear response regime, we can get an expansion that summarizes the mean response dynamics of the order parameter in terms of a dynamical function that only requires observation of the order parameter of interest in equilibrium that is eminently calculable.

The Fluctuation Dissipation Theorem as derived under Hamiltonian dynamics relates the observed equilibrium fluctuations to the unobserved decay rate of the mean. This is analogous to the result for fluctuation-dissipation under overdamped Langevin dynamics, which relates the dissipation parameter to the total amount of fluctuation observed in the system. Because protein simulations are usually performed with Hamiltonian dynamics, that formalism is more directly applicable and is thus the one we will proceed with.

Let’s return to the simple linear perturbation of equation 2.13. Consider a scenario where for all $t < 0$, $k = k$ and for all $t > 0$, $k = 0$. We know that the for all $t < 0$, $\langle X(t) \rangle = X_k$, that is, the equilibrium value for $X$ in ensemble $k$. We want to compute the average values of $X$ as it transitions from the two ensembles under the abrupt deactivation of bias $k$. Using the equilibrium time propagator that evolves a configuration from $t' = 0$ to $t = 0$ as $\hat{P}_t$, we can write this average as

\[
\langle X(t) \rangle = \frac{1}{Z_k} \int d^{3N} x \langle \hat{P}_t \hat{P}_0 \rangle e^{-\beta H_k} \quad (2.17)
\]

\[
= \frac{1}{Z_k} \int d^{3N} x \langle \hat{P}_t \hat{P}_0 \rangle e^{-\beta H_0 - \beta kX(x)} \quad (2.18)
\]

\[
= \frac{1}{Z_k} \int d^{3N} x \langle \hat{P}_t \hat{P}_0 \rangle e^{-\beta H_0} (1 - \beta kX(x) + O(k^2)) \quad (2.19)
\]

\[
= \frac{1}{Z_k} \left( \langle X \rangle_0 - \beta k \langle X(t)X(0) \rangle_0 \right). \quad (2.20)
\]
We can expand the partition function in the biased ensemble similarly, (noting that \( Z_k = Z_0 \) if the \( F_0(X) \) is perfectly harmonic), as

\[
Z_k = \int d^{3N}xe^{-\beta H_0 - \beta kX(x)}
\]

\( \vdash \)

\[
= \int d^{3N}xe^{-\beta H_0}(1 - \beta kX(x))
\]

\( \vdash \)

\[
= Z_0(1 - \beta k\langle X \rangle_0 + O(k^2)).
\]

Without loss of generality, we can look at these equations with \( hX_i \mid_0 = 0 \), and (to order \( O(k^2) \)) we get,

\[
\langle X(t) \rangle = -\beta k\langle X(t)X(0) \rangle_0.
\]

In this way, we have derived a response function of the order parameter \( X \) to perturbation \( k\eta(t) = -k\delta(t) \), which means we can compose an arbitrary number of these together using the Greens Function formalism, then integrating by parts,

\[
\langle X(t) \rangle = \int dt' \beta \frac{\partial}{\partial t'}k\eta(t')\langle X(t)X(t') \rangle_0
\]

\( \vdash \)

\[
= \int dt' \beta k(t') \frac{\partial}{\partial t'}\langle X(t)X(t') \rangle_0.
\]

\( \vdash \)

Therefore, the mean time evolution of the order parameter \( X \) is only dependent on its own expectation values in a direct way.

### 2.2.7 Dynamics: Fluctuation-Dissipation by analogy

One of the most confusing aspects of the Fluctuation-Dissipation Theorem to interpret is the fact that the response is related to the derivative of the correlation. Because nearly every physicist has taken elementary circuits and understands the topic in great depth, I am going to use an analogy in circuits to illustrate the relevance of this interpretation.

![A conventional RC Circuit](Figure 2.4: A conventional RC Circuit)
We will begin by setting up one of the first circuits introduced: the RC circuit. We use a switch to show that we have two states: full voltage $V_0$ applied or no voltage applied but still with finite resistance. Let us write down the differential equation governing the system,

$$IR + \frac{Q}{C} = V(t)$$

(2.28)

$$\dot{Q}R + \frac{Q}{C} = V(t).$$

(2.29)

We then solve for the homogeneous part of the equation with an exponential ansatz, and reach the homogeneous solution $Q(t) = Ae^{-t/RC}$. We can use the homogeneous solution any time that the differential equation has no applied force as long as we apply the correct boundary terms from the previous domain. Because of this, the simplest inhomogenous problem we can solve is the instantaneous removal of voltage, the function $V(t) = V_0\Theta(-t)$, where $\Theta(t)$ refers to the Heaviside step function that takes a value of 1 for all $t > 0$ and a value of 0 for all $t < 0$. Thus, all times $t > 0$ are governed by the homogeneous equation, and the initial condition is the converged charged value $Q(0) = CV_0$. Thus, the response equation is $Q(t) = CV_0e^{-t/RC}$.

This will all have been dreadfully obvious to someone with graduate-level preparation in the physical sciences, but please bear with me as I set up the formalism in a way which will lay bare the analogy to fluctuation dissipation.

While this step function solution is fine and well, we cannot get much more out of it than a fairly clumsy chain of solutions when our voltage takes several steps, and certainly not much at all when our voltage changes continuously. In a regular physics course, the next step might be to go to a Fourier decomposition to find the impedance as a function of frequency. This is not my goal. We can instead find a Green’s Function solution to this differential equation and thus write down a formal expression that holds for all applied voltages $V(t)$.

To solve for a Green’s Function, we set the inhomogeneous side of a differential equation equal to $\delta(t-t_0)$, where we conventionally set $t_0 = 0$. Thus, we want to solve the differential equation $\dot{Q}(t)R + \frac{Q(t)}{C} = \delta(t)$. With some fortitude of mind, we can try to interpret this equation, but it is not particularly physical in my opinion. It corresponds to applying an
infinite voltage for a very short amount of time, but one such that the integrated voltage
applied is equal to 1. We could relate this to the instantaneous charging of the capacitor,
but the physical means by which one could do that without removing the resistor from the
picture are dubious.

Fortunately, we can realize that we have already solved a much more physically relevant
Green’s Function if we consider the derivative of the Heaviside function $\Theta(t)$. This function’s
derivative is itself $\delta(t)$, which means that our original capacitor equation can be recast as a
Green’s function equation by taking a derivative.

\[\ddot{Q} R + \frac{\dot{Q}}{C} = \frac{d}{dt} \Theta(-t)\]  \hspace{1cm} (2.30)
\[\dot{I} R + \frac{I}{C} = -\delta(t)\]  \hspace{1cm} (2.31)

Differentiating our solution, $Q(t) = CV_0 e^{-t/RC}$ with $V_0 = -1$ (as it is for the delta function),
we arrive at our Green’s Function for $I$,

\[G(t - t') = \frac{1}{R} e^{-(t-t')/RC} \Theta(t - t'),\]

and consequently at our general solution for $I(t)$ and $Q(t)$,

\[I(t) = \int_{-\infty}^{t} dt' \frac{\dot{V}(t')}{R} e^{-(t-t')/RC},\]  \hspace{1cm} (2.32)
\[Q(t) = Q(t_0) + \int_{t_0}^{t} dt' I(t')\]  \hspace{1cm} (2.33)

Let’s enumerate the ways in which this result sheds insight into the nature of the
Fluctuation-Dissipation Theorem.

First, the set up of the system is such that the step function solution which can be much
more physically motivated is the inhomogeneous system that we solve. We do not use the
language of Green’s Functions when we solve this system, but we can relate our solution to
the full solution.

Second, the exponential decay process of the capacitor is easy to think about as a re-
 laxation or a decorrelation. It is a property inherent to the homogeneous system, much like
fluctuation is thought to be. Under the application of a perturbation, the response function
$G(t - t')$ is the derivative of the “decorrelation” function, and this response function acts
proportionally to the applied force which is no longer the potential itself, but the derivative
of the potential.

Third, the decay is exponential, and the derivative of the decay (i.e. the response) is also
exponential. While the idea of differentiating something can be somewhat confusing, it is
pleasant to know that a correlation which is a sum of exponential decays remains a sum of
exponential decays even when we differentiate it to acquire the response function.
2.2.8 Dynamics: Models of protein landscapes

Now that we have surveyed the theory of CV effective dynamics as a way to describe the dynamics of the energy gap, we can discuss the actual $d^{3N_x}$ degrees of freedom being integrated out by discussing the microscopic protein dynamics. In doing so, we can interpret the features of the microscopic space in the context of the CV to understand more clearly the validity of our approximations, including the poorly controlled linear response approximation.

Proteins have energy landscapes with tremendous numbers of locally stationary points. Natively folded proteins have densely packed interiors, with a variety of specific interactions, such as interactions between charged and polar groups, and aggregation of hydrophobic groups in the interior. However, this packing of side chains is not unique. The interior of the protein in fact fluctuates enough to identify it on the boundary between “liquid-like” and “solid-like” [53]. The variety of possible side chain packings contributes entropy that overpowers the native state’s “optimal” packing, and it is thus generic to observe transitions among these near optimal structures. We therefore cannot understand a protein if we do not understand it as an ensemble entity [54].

The protein’s densely packed structure makes these transitions kinetically difficult, though. Differences between disparate non-optimal structures will be non-locally distributed because of the high density; that is, if one side chain transitions rotameric states, the rest of the densely packed interior must nonlinearly respond. Equilibrium fluctuations have no causal nature, though, so it is more apropos to describe this relationship as long ranged correlations between side chain rotameric states across the protein structure [55] rather than as an “avalanche” reaction to a rotameric change.

The situation is exaggerated even further by the nearly-discrete nature of side chain dihedral angles [56]. Thus, these correlated motions do not happen smoothly. Instead, transitions in single rotamers occur only in the presence of an instantaneous vacancy of the target state at the same time as the discrete transition is attempted. These transitions “lock in” the instantaneous vacancy, which in turn can cause the protein to frequently explore states that would be incredibly rare in the initial state from this thought example. There is thus a discrete random walk occurring in the very high-dimensional dihedral rotamer space, and it is these discrete states that contribute to significant long-ranged correlations.

Proteins fluctuate on a variety of timescales. Fastest are “free” sidechain dihedral motions, which can transition with rates of 1 to 100ps. Next are densely packed interior dihedral motions, with a rate of roughly 1 to 100 nanoseconds. Next we have translations in secondary structure, which often require microseconds to occur [57]. Finally we have folding and unfolding motions, which span millisecond to second timescales [58].

2.2.9 Dynamics: Markov State Models

One mathematical framework for proteins with a good balance between theoretical tractability and complexity of representations is the Markov State Model. In this framework, we
define a mapping of configurations to meaningful discrete states (with the assistance of some form of machine learning algorithm), from which point we construct a matrix of transition rates between states [59]. We will explore why this assumption is reasonable.

If we assume that there are no dynamical variables that persist over timescales as long as the transition times between proteins, then we would say that the process has no memory. Markovity is a stochastic assumption which posits this assumption exactly: transition rates depend only on the instantaneously occupied state. In a Markovian system, transition rates to different states vary as a function of time, but this time dependence can be reduced to a function of the current configuration. This leads us to a first-order master equation. This is not to say that this reduces our system to a simplistic dynamics; rather, it just allows us to make progress on defining a moderate number of discrete states and their transition probabilities without needing to parameterize memory effects (e.g. coherent oscillations, in-state phonon relaxation).

Markov State Model studies of small proteins have identified a generic feature of hub-like network topology [60]. That is, there are a few “central” states with high connectivity to the rest of the network, and it is through these central hubs that the branches of the state network are explored. This observation simultaneously lends support to the consistency of the native state as statistically important, while also acknowledging the extent of unique exploration that spontaneously occurs in the network.

The Markov State description suggests several things about how we could conceive of the energy gap’s distribution. We would first want to consider whether the gap excitation is large enough to force a conversion between states. If it is, we might expect to see some nonlinearities in the mean gap as a function of \( \lambda \). This interpretation is slightly naive, though. The free energy curves are defined with respect to the full ensemble of explored states, so these non-linearities themselves must be random – both size of transition and barrier height. Thus, even if a particular configuration responds nonlinearily, the ensemble response may be perfectly linear. The dynamical response, however, can be more sensitive to anharmonic features of the free energy distribution than the mean response.

If, on the other hand, the gap excitation is not enough to force a transition between Markov states, we would expect to see a response much smaller than that which the full equilibrium distribution would suggest. If each state itself has a locally quadratic gap free energy, the response to the gap is quite more like the naive interpretation, but now with a response variance that depends on the particular state being occupied. Furthermore, the dynamics of small and fast events will be qualitatively different from large and slow ones, suggesting the introduction of a separation of time scales. This separation is analogous to the separation between static disorder and dynamic disorder.

We can thus leverage existing understanding of protein energetic landscapes to direct our formation of hypotheses about gap dynamics. When it comes to the practicalities of this procedure, we must generally know the natively folded structure of the protein, as determined experimentally. For this, we turn to crystal structures, most of which are reconstructed from x-ray scattering data.
2.3 Models of the energy gap

Although none of the original research here uses quantum chemistry calculations, a cursory understanding of the topic is necessary to appreciate the approximations being made to compute the energy gap.

Computation of the energy gap for a molecule from atomistic coordinates is not a particularly trivial task. It requires an accurate computation of both the ground state of the molecule and the first excited state of the molecule. Neither of these calculations are cheap or easy to perform. Though the ground state must always be computed en route to the excited state, methods for ground state properties do not always capture excited states well. We truly need to approach both problems to give the gap computation proper treatment.

Fortunately, in addition to full quantum mechanical approaches, we can use methods inspired by the physics involved in the quantum mechanics to approximate the full computation by a classical analog. I will give a brief overview of these methods here, emphasizing the approximate scaling behavior of the prominent methods and thereby the limitations of those methods.

2.3.1 Hartree-Fock and determinants

To explain the approximations made in the ZINDO method (a method used frequently in photosynthesis computations), it is necessary overview the basics of quantum chemistry, i.e. Hartree-Fock theory. This outline should be sufficient to explain the approximations made in the variety of methods used. Please consult Szabo and Ostlund [61] for a more proper and complete treatment of this topic. This treatment is intended to be mechanistic and not numerical. With this baseline in place, we can discuss the ZINDO method, which was developed to treat spectroscopic chemistry of pyrroles, a primary component of the porphyrin molecules of which Bacteriochlorophyll a is a member.

One of the biggest challenges to understanding quantum chemistry calculations are the number of moving parts. We are taught quantum mechanics in a single particle systems the majority of the time. We can select K basis states, and diagonalize the hamiltonian of matrix elements no matter what basis states we select. Of course, if we can select the eigenbasis, we have done incredibly well, and the matrix comes out diagonal. In reality, we approximate the eigenstates with basis states that are pragmatic in that their mutual matrix elements can be computed quickly, and that they span a range of configurational space that we reasonably expect to occupy.

The trouble comes when we have N electrons — quite simply, we have exploded our problem into around $2^K$ N-electron wave functions. Each of these N-electron wave functions are called determinants because they need not only be sums of the basis states, but anti-symmetrized sums. This anti-symmetrization is formed through the use of Slater determinants, and impacts the matrix elements that we compute. Note that Slater determinants are not the only way to satisfy anti-symmetrization, but they are both a complete basis
and a common one. In any case, we select a set of states such that every state respects the anti-symmetry, and such that the span of states is identical to the original span.

If we are gifted with a small enough problem that we can actually enumerate \( \binom{2K}{N} \) states and compute \( \left( \binom{2K}{N} \right)^2 \) matrix elements, we can diagonalize the determinant basis and get the exact result within the orbital basis we have selected. This approach is called Full-Configurational Integral, or Full CI for short. Obviously, though, for problems of substantive size, \( \binom{2K}{N} \) grows at such a rate as to make the problem completely intractable. To make progress, we aim to make as many matrix elements as possible disappear, and then completely ignore them. This amounts to a sparse treatment of the full CI matrix in which we hope that the elements we have set to zero do not contribute significantly. Our procedure for this follows this rough outline: first, use a smarter set of B basis elements through linear combinations. Second, use rigorous symmetry constraints to select which CI elements to always exclude. Third, use chemical intuition to select which additional CI elements to approximate as zero.

This first step of selecting a smarter one-electron basis of B states brings us to the Hartree-Fock approach. This introduces the linear combination of atomic orbitals (LCAO) to convert from an atomic orbital (AO) basis to a molecular orbital (MO) basis. Because we are trying to simplify our orbital basis, we do not want to be treating the full multi-electron problem yet; rather, we use a mean-field approach to write the Hartree-Fock equation from the full Schrödinger equation. We can then diagonalize our AOs into MOs which are LCAOs. This brings us quite a way towards our full solution, but we are systematically missing some interactions. All interactions missing from Hartree-Fock theory are referred to as the correlation energy.

In our second step, we consider which matrix elements could possibly be non-zero. With a second quantization approach, we can see that our Hamiltonian will only include either one-electron or two-electron interactions, where the Coulomb interaction between electrons constitutes the main two-electron term that is usually considered. In a second-quantization framework, both of these operators preserve the number of electrons, which is perfectly reasonable for a non-relativistic approach. In second quantization, a one-electron interaction destroys one electron and creates another while a two-electron interaction destroys two electrons and creates two more. From this understanding, we can see that only interactions between determinants that contain a difference of at most two MOs from one another are non-zero.

Here, we can group our determinants into how “far” a determinant is from the Hartree-Fock wavefunction by considering how many orbitals must be swapped to arrive at the determinant of interest. We call each orbital swap necessary to reach our determinant of interest as an excitation. Then these states are then categorized as Singles, Doubles, Triples, Quadruples, etc. to sort our states on the number of excitations present with respect to the Hartree-Fock wavefunction. There is one HF wavefunction, \( \binom{2K}{2} \) singles, \( \binom{2K}{4} \) doubles, \( \binom{2K}{6} \) triples, \( \binom{2K}{8} \) quadruples, and so on. For large K, we can think of each additional level as adding two powers to our exponent of K for the order of the number of states (e.g. to take up
CHAPTER 2. SCIENTIFIC MODELS, SCIENTIFIC THEORIES

40

to triples, we need $O(K^6)$ states. We also know that our Hamiltonian will not connect states with more than two excitation difference, so we know that all states of Triples and higher do not have any direct interaction with the Hartree-Fock wavefunction, though they have indirect interaction through the Singles and Doubles. We also have the tremendous benefit of knowing that because the Hartree-Fock wavefunction is the solution to the one-electron Hamiltonian, all of the matrix elements between the HF wavefunction and the Singles are zero, leaving only the doubles to offer interactions to higher excited states. The interactions of the Hartree-Fock wavefunction that occur indirectly are referred to as virtual interactions.

The third step is where we truly reduce the complexity of the problem as we systematically exclude huge numbers of matrix elements, but we must also employ physical reasoning to make our selections. These selections are where quantum chemistry becomes alphabet soup. To illustrate this, I will present several examples. First, we have the CI-singles (CIS) method. This method is very similar to a Hartree-Fock treatment (in fact, the ground state energy is the same because there is no coupling between the Hartree-Fock wavefunction and the singles), but the improvement which is made is that the singles (which in many systems are the low-lying optical states) are diagonalized with respect to one another, bringing us much closer to excited state accuracy with only $O(K^2)$ determinants. We also have coupled cluster which treats the $O(K^4)$ coefficients of the doubles as quantities to optimize, but which does not truncate the coefficients of quadruples and higher but instead sets their coefficients equal to the decomposed product of the doubles (i.e. a quadruples excitation of $ijkl \rightarrow abcd$ is the product of the doubles excitations $ij \rightarrow ab$ and $kl \rightarrow cd$). This method reduces the number of free parameters to $O(K^4)$ but does not suffer the pathologies of truncating CI at the doubles level. Many popular approximations bring the scaling down to around $O(K^3)$, which is somewhat poor but is within the realm in which substantial progress can be made.

2.3.2 ZINDO

Zerner Intermediate Neglect of Differential Orbitals (ZINDO) is a semi-empirical method that lives within the family “NDO” methods. These NDO methods are semi-empirical methods that use some parameterizations to reach a good result without the full computational effort. Each NDO method is specialized to treat different molecules, with ZINDO constructed to perform excited-state calculations on pyrroles. The simplest of this family is the Complete Neglect of Differential Orbitals (CNDO), and we will describe this in detail to outline the family of methods.

In the simplistic CNDO approximation, the truncation is performed not in the MO space like CIS, but on the AO matrix elements. These AO electron-electron matrix elements have the form of a product of four AO wavefunctions split into two pairs, where each pair has a real space coordinate associated to it. The distance in the Coulomb interaction is then the distance between the real space coordinates. While in general we can have any four AO wavefunctions in the product, the CNDO requires that we pair the orbitals with the same real space coordinate to be the same AO. This setup has a classical interpretation: because the square of a wavefunction is the probability density (and thus charge density),
this neglects Coulomb interactions from strictly quantum contributions where we consider
the charge density that comes from the overlap of nearby AOs. In practice, the remaining
integrals between s, p, d, f, etc. of all of the atoms are parameterized with semi-empirical
factors.

In practice, CNDO performs quite poorly. We can improve our estimates with the next
level of Intermediate Neglect of Differential Orbitals (INDO) by including one small addi-
tional set orbitals: differential orbitals that take place on the same atomic center. This
inclusion has a substantial e

fect on the difference between bonding and anti-bonding or-
bitals and comes at a relatively small cost. ZINDO further specifies a particular form for
the parameterization of the Coulomb integrals. Finally, the excited-state ZINDO/S selects
a set of about 10 each of HOMO and LUMO wavefunctions to perform full diagonalization
of after reaching the first level of results.

2.3.3 TD-DFT

DFT is a third methodology (in addition to wavefunction CI and Perturbation Theory)
to approach the problem of solving for energies, however, DFT is formally established only
for ground states. The foundational proof in DFT states that there does exist a unique func-
tional which can compute exactly the ground state energy that corresponds to a particular
electron density. Through the variational principle, the ground state is therefore the electron
density function which minimizes the energy functional. While the exact unique functional
for the Schrödinger equation is still not known, approximate methods which handle a variety
of exchange (present in Hartree-Fock but missing from some DFT functionals) and correla-
tion (missing from both Hartree-Fock and some DFT functionals) phenomena that must be
present in an exact treatment.

TD-DFT takes this theory to excited states through the application of a periodic electric
field at a resonant frequency. If the applied perturbation is weak, we can monitor the results
in the linear response regime, and sweep frequencies to find poles of the response function.
The analogy to photon absorption is somewhat obvious, as the weak periodic field can be
thought of as the electric field of an applied source of monochromatic light. This theory is less
fully developed than standard DFT, but benefits from having the capacity to be improved
with better functionals as well as from having desirable scaling to work with novel molecules
of interest that cannot be treated with a reasonable level of wavefunction-based theory.

2.3.4 Point charge embedding

Despite the substantial quantum theory presented up to this point, we are still funda-
mentally missing the environmental coupling so essential to the photosynthetic problem.
Fortunately the simplest acceptable approach is direct: to perform the quantum chemical
calculations in the applied field of the environment’s point charges.

It would be great if the story ended there, but we must be attentive to the complexities of
the long-ranged Coulomb force. For the coulomb coupling, it is easier to explain what being
long-ranged means in terms of potential. We know from Coulomb’s law that all atoms in space contribute to the potential proportional to their charge and inversely proportional to their distance. Long rangedness can be understood by considering the potential contribution from a gigantic (e.g. infinite) uniformly charged system that encloses our system of interest. Because the area of a sphere grows as \( r^2 \), where \( r \) is the distance from our internal point of interest, the potential contribution to this interior point grows with \( r \). Thus, for an infinitely large encapsulating body, the point at infinity contributes the most potential energy to a point in the interior. In practice, this usually comes down to a matter of boundary conditions; long rangedness means that our boundary, the shell at "infinity" will contribute just as much to the calculated potential as any of the internal features.

We therefore have two options: either we consider a neutrally charged periodically replicated system, or we embed a finite system in an infinite dielectric material. The prior approach is often preferrable because it can be treated with FFT-based Ewald summation, and can therefore be computed in \( \mathcal{O}(N \log N) \) time instead of \( \mathcal{O}(N^2) \) time. Unfortunately, Ewald summation requires a charge neutral system in a geometry that can tessellate space without leaving any unphysical gaps. If we instead want to consider a small subset of a larger system, embedding in a dielectric continuum is usually preferred. This approach approximates the rest of the solvent by treating it continuously and asymptotically.

Point charge embedding of various types has been ubiquitous in FMO research [62] [11] [16], but it is difficult to answer the extent to which this method produces accurate results. Some authors have improved upon the point charge embedding with polarizable force fields [63], but the dual questions of how many atoms to include in a QM model and how far out to compute coulomb interactions remain challenging to answer.

### 2.3.5 Intermolecular coupling

The quantum theory we have developed up until now draws inspiration from being computed for only a single molecule at a time. The theory works just as well for two (or more) molecules simultaneously, though, with the possibility for excitations to involve orbitals spread non-locally across the plurality of molecules. The cost for including a second set of orbitals corresponding to a second molecule unfortunately scales with the cost of our quantum chemistry method (often \( N^3 \) or larger). To our benefit, the noncovalent nature of adjacent molecules allows us to do much better with a perturbative approach than considering the quantum chemistry for such a large state space.

Because adjacent molecules are not covalently bound, and thus can be approximated as having no orbital overlap, the exchange and correlation energies can be ignored. If the orbitals are pressing up against one another because of geometry, empirical Van der Waals interactions can be used to capture the effect of exchange without resorting to full quantum chemistry. Thus, it is only left to describe the Coulomb interactions between monomers to summarize the full multi-molecule wavefunction space.

We will not treat the selection rules from the Wigner-Eckhart theorem, and thus our summary of intermolecular coupling is not particularly rigorous. However, it suffices to say
that the selection rules mandate that only certain operators, including the dipole operator \( \mu \), are allowed to induce transitions across states that differ in spin by 1 (as is the case for states that are excited by the spin 1 photon). Representing the two molecules by a joint wave function as \( |ge\rangle \) for first molecule in the ground state and second molecule in the excited state and \( |eg\rangle \) for the converse, the matrix element between these two states is determined by the coulomb matrix element \( \langle ge|\mu_1 \cdot \mu_2|eg\rangle \), which if we assume separability of the wavefunction \( |eg\rangle \) into \( |e_1\rangle|g_2\rangle \), becomes the coulomb interaction between the transition dipole moments \( \mu_T = \langle e|\mu|g\rangle \), and thus \[ \langle ge|\mu_1 \cdot \mu_2|eg\rangle \equiv \mu_{1T} \cdot \mu_{2T}. \]

More generally, this expression will involve the mutual spatial integration required to form \( \mu_1 \) and \( \mu_2 \) to occur simultaneously, and therefore, will be more similar to the interaction between two charge distributions than two point dipoles. However, at sufficient range where the coulomb interaction is dominated by the point dipole contribution, the point dipole is a sufficient approximation.

### 2.3.6 TDC, TrEsp, and CDC

Given our understanding of the way in which we couple the environment to our quantum system, we can turn around and compute several important quantities quite directly: the environmental contribution to the energy gap, and the coupling between chromophores.

The Transition Density Cube [64] was one of the first methods to carefully and rigorously use a classical electrostatic method to compute couplings between molecules. To compute coupling between molecules, the transition charge density (defined as the overlap integral between the ground and excited state) is computed for each molecule independently and interacted Coulombically. To actually evaluate this integral, it is necessary to discretize the volume that is occupied by transition charge, and as such the method acquires its name from the conventional use of a cubic grid. Unfortunately, these computations can be costly, as it is common to use upwards of 300,000 points to accurately represent the transition charge density. Fortunately, this method converges to exact accuracy as the grid spacing gets smaller.

Later, Madjet et al developed the transition charge from electrostatic potential (TrEsp) [9]. This method follows the CHELP-BOW [65] method of estimating point charges for force fields. The practical efforts to generate point charges from quantum chemical calculations do not use quantities direct from the quantum chemical calculations like the charge densities or the mulliken charges. Instead, the practice is to minimize the error in the electrostatic potential by optimizing charges subject to a constraint on the net charge, and possibly higher moments. CHELP-BOW specifies a particular choice of points sampled from the potential, an error function, and a set of constraints. In particular, CHELP-BOW samples points visited by a walker following molecular dynamics, fits the potential to a boltzmann-weighted sum of squared error, and constrains the total charge and dipole moment. Details of the methodology can be seen in greater detail in the original work [65]. Thus, the TrEsp
estimates transition charges using the field created by the transition density, reducing the number of interactions from around $300,000^2 \equiv 10^{11}$ to around $40^2 \equiv 10^3$.

CDC is parameterized as a sort of computational side-note in this research to compute the environmental contribution to the energy gap. Using the same CHELP-BOW method for parameterizing the electrostatic potential, the difference potential between the ground and excited state is computed. Both CDC and TrEsp are perturbative approaches because they cannot factor in the effect of environmental charges modifying the electron clouds of the molecule. The nature of these perturbations is qualitatively different, though. TrEsp does not, at its core, require the consideration of environmental charges, and so in vacuum, it can be considered a fit of the otherwise exact TDC. CDC, on the other hand, only acquires values from the presence of environmental charges. These charges interact with the ground state charge distribution, which is half of the difference interaction between the excited and ground state distributions, and we therefore might expect charges in the environment to substantially distort the electron cloud of our molecule when left to occupy their equilibrium positions.

### 2.4 Models of thermal quantum mechanics

With the groundwork laid out for atomistic protein dynamics and for coupling of those dynamics to the single-excited state manifold quantum subsystem, we can proceed with the variety of models developed to describe this combination in a solvable framework.

The phenomenology of thermal quantum systems become complicated quite quickly. It is useful, then, to discuss two predominant limits for reference before considering the general case: FRET and Redfield dynamics. FRET is used to model electronic energy transfer in a gargantuan number of physical systems, and is almost universally the first attempt at modeling. FRET is a great approximation for all but the most closely packed and structured pigment protein complexes. Redfield, on the other hand, is a useful model to understand the transition from a Markovian rate process to a Markovian quantum process. In the two cases, Markovity comes from different directions: in FRET, Markovity comes from the fast bath relaxation in response to the quantum process, whereas in Redfield, Markovity comes from the slowness of bath degrees of freedom, such that bath degrees of freedom cannot respond to (and therefore remember) the quantum dynamics. When we move to the general case, we approach from the Redfield direction, which justifies our treatment of that approach.

#### 2.4.1 Linear optics

Before computing dynamical quantities, we can look at the significant quantum observables as they are impacted by thermalization. There are a number of simpler observables to compute instead of dynamics. Dynamics are sufficient to compute any desired observables,
but are quite more intensive than necessary if our only interests are in linear optic quantities. A complicated and varied treatment of linear optics are presented in Adolphs et al. [4].

2.4.2 Forster Resonant Energy Transfer

In FRET theory, there are three factors that are important: the spectral overlap, the relative orientations, and the distances between donor and acceptor. The latter two are fairly similar across molecular species, i.e. if the pigments have known orientation distributions it is possible to compute the orientation factor irrespective of the molecules being used, and the distance proportionality of $1/R^6$ does not change from molecule to molecule. The spectral overlap therefore contains the bulk of the molecularly specific thermal physics. The spectral overlap is defined as

$$J = \int d\lambda F_D(\lambda) \epsilon_A(\lambda) \lambda^4 / \int d\lambda F_D(\lambda),$$

where $F_D(\lambda)$ is the wavelength-dependent donor fluorescence and $\epsilon_A(\lambda)$ is the wavelength-dependent acceptor molecular extinction coefficient. We can relate this to donor and acceptor probabilities by noting that

$$p_D(\lambda) = F_D(\lambda) / \int d\lambda F_D(\lambda)$$

and also noting that we can rewrite $\epsilon_A$ in terms of its gap probability distribution and a scattering factor $\sigma_A \propto \lambda^{-4}$ (via Rayleigh scattering),

$$\epsilon_A(\lambda) = p_A(\lambda) \sigma_A(\lambda)$$

$$\sigma_A(\lambda) = \bar{\sigma}_A / \lambda^4,$$

and therefore rewrite our overlap integral in terms of probability integral

$$J = \bar{\sigma}_A \int d\lambda p_D(\lambda) p_A(\lambda)$$

$$J = \bar{\sigma}_A \int d\lambda_D d\lambda_A \delta(\lambda_D - \lambda_A) p_D A(\lambda_D, \lambda_A)$$

$$J = \bar{\sigma}_A \langle \delta(\lambda_D - \lambda_A) \rangle.$$

Thus, dating back to Förster, the joint distribution of the instantaneous energy gap is the essential element in the phenomenon of excitation transfer.

In the FRET limit, asymmetric donors and acceptors are usually considered, such that the overlap integral of donor emission to acceptor absorption is much larger than the overlap integral of acceptor emission to donor absorption. Thus, FRET can be observed as a single event. The reorganization energy helps to decipher why this is so; for two pigments with $G$ values separated by $\lambda$, the donor will be in nearly constant resonance with the acceptor when the donor is excited, however, when the acceptor is excited, the reorganization energy will separate the two by and additional $\lambda$, pushing their means $2\lambda$ apart.
2.4.3 Redfield Equation

If we move to the other extreme, we can examine the Redfield model, where reorganization energy is extremely small in comparison to couplings. Using the interpretation of the reorganization energy as a dissipative term, it becomes clear that the system should act as the free quantum propagator with a small and continuous damping that eventually brings the system to equilibrium. The Markovity applies not in the Schrödinger picture, where oscillations between coherent quantum states can be quite rapid, but in the interaction picture, where coherent quantum oscillations are removed from the free propagator. The separation of time scales between the fast system degrees of freedom and the slow bath degrees of freedom means that the bath only feels the “average” system configuration, and thus Markovity is recovered in the interaction picture.

We will follow the superoperator and subspace projection formalisms as introduced in chapter 3.5 of “Charge and energy transfer dynamics in molecular systems” [66] with some slight changes in notation. In particular, we make time arguments \((t)\) implicit by using prime marks \(\prime\) and \(\prime\prime\) to identify operators in \(dt'\), \(dt''\) (and so forth), with reference to time of origin \(t_0\) (indicated by \(0\) when needed). We therefore use the following definitions of Louiville superoperator \(L\), pure space projection operator \(P\), coupled space projection operator \(Q\), bath equilibrium density matrix \(B_{eq}\), total Hamiltonian \(H\), system Hamiltonian \(H_S\), bath hamiltonian \(H_B\), system-bath coupling \(H_{SB}\), full space density operator \(\hat{W}\), and system and bath subspace density operators \(\hat{\rho}\) and \(\hat{B}\). Thus, the following relationships are defined among these defined variables,

\[
H = H_S + H_{SB} + H_B
\]

\[A^{(I)} = U^\dagger AU\]

\[H^{(I)} - E_0^B - E_0^S = H_{SB}^{(I)} = U^\dagger H_{SB}U\]

\[\mathcal{L} \cdot = -\frac{i}{\hbar} [H, \cdot]\]

\[\mathcal{P} \cdot = \hat{B}_{eq} \text{Tr}_B \{ \cdot \}\]

\[Q = 1 - \mathcal{P}\]

\[\text{Tr}_B \{ \hat{W} \} = \hat{\rho}\]

\[\mathcal{L}^{(I)} \cdot = -\frac{i}{\hbar} [H_{SB}^{(I)}, \cdot]\]

We will now drop the interaction picture \(^{(I)}\) demarcations with the implicit knowledge that \(\mathcal{L}\) now has time dependence from the interaction picture and \(\hat{W}\) is also in the interaction picture. We can then consider the exact coupled equations of motion that arise from splitting the full space density operator \(\hat{W}\) into the system and bath subspaces \(\hat{\rho}\) and \(\hat{B}\) by inserting the identity superoperator \((\mathcal{P} + Q)\) and performing the correct algebra,

\[
\partial_t \mathcal{P} \hat{W} = \mathcal{P} \mathcal{L} \mathcal{P} \hat{W} + \mathcal{P} \mathcal{L} Q \hat{W}
\]

\[
\partial_t Q \hat{W} = Q \mathcal{L} \hat{P} \hat{W} + Q \mathcal{L} Q \hat{W}.
\]
To begin interpreting and to understand the perturbation theory that we will generate with this pair of equations, consider the term $Q\hat{W}$. This is the projection of the full density matrix onto the coupled subspace. In equilibrium, i.e. when the state density matrix separates as $\hat{W} = \hat{B}_{eq}\hat{\rho}$, this term becomes zero, as evidenced by its definition: $Q\hat{W} = \hat{W} - \hat{B}_{eq}\text{Tr}_B\{\hat{W}\} = \hat{B}_{eq}\hat{\rho} - \hat{B}_{eq}\hat{\rho} = 0$. Thus, for our low coupling approximation, we can try to perturb around this quantity, which in other words expands $\hat{W}$ around its equilibrium value. This can also be thought of as a system and bath in mutual equilibrium, from $t = t_0$, but with dynamical feedback from system fluctuations from the bath.

With this interpretation, we can look at the $N$th order perturbative expansion by solving equation 2.49 self-consistently $N - 1$ times, with $Q\hat{W} = 0$ inserted for the first iteration (thus the reasoning behind the minus one). The first three orders of perturbation theory are

\begin{align*}
\text{PT1} & \quad \partial_t \mathcal{P}\hat{W} = \mathcal{P}\mathcal{L}\mathcal{P}\hat{W} \\
\text{PT2} & \quad Q\hat{W} = \int_{t'}^t dt' Q\mathcal{L}'\mathcal{P}\hat{W}' \\
& \quad \partial_t \mathcal{P}\hat{W} = \mathcal{P}\mathcal{L}\mathcal{P}\hat{W} + \int_{t'}^t dt' \mathcal{P}\mathcal{L}\mathcal{L}'\mathcal{P}\hat{W}' \\
\text{PT3} & \quad Q\hat{W} = \int_{t'}^t dt' Q\mathcal{L}'\mathcal{P}\hat{W}' \\
& \quad Q\hat{W} = \int_{t'}^t dt' Q\mathcal{L}'\mathcal{P}\hat{W}' + \int_{t''}^t dt'' Q\mathcal{L}'\mathcal{L}''\mathcal{P}\hat{W}'' \\
& \quad \partial_t \mathcal{P}\hat{W} = \mathcal{P}\mathcal{L}\mathcal{P}\hat{W} + \int_{t'}^t dt' \mathcal{P}\mathcal{L}\mathcal{L}'\mathcal{P}\hat{W}' + \int_{t''}^t dt'' \mathcal{P}\mathcal{L}\mathcal{L}'\mathcal{L}''\mathcal{P}\hat{W}''.
\end{align*}

For the Redfield equation, the second level of perturbation theory is used, as it is the first with nontrivial modifications to the dynamics, as first-order will only include an energetic shift from the bath. This level of perturbation theory with the $Q\hat{W}^0 = 0$ is also referred to as the Nakajima-Zwanzig equation.

From equation 2.52, we see that the effect of including bath fluctuations comes in the form of integral $\int dt' \mathcal{P}\mathcal{L}\mathcal{L}'\mathcal{P}\hat{W}'$. This can be split into two terms through the definition of $Q = 1 - \mathcal{P}$,

$$\int_{t'}^t dt' \mathcal{P}\mathcal{L}\mathcal{L}'\mathcal{P}\hat{W}' + \int_{t'}^t dt' \mathcal{P}\mathcal{L}\mathcal{P}\hat{W}' = \int_{t'}^t dt' \mathcal{P}\mathcal{L}\mathcal{L}'\mathcal{P}\hat{W}' + \int_{t'}^t dt' \mathcal{P}\mathcal{L}\mathcal{L}'\mathcal{P}\hat{W}' + \int_{t''}^t dt'' \mathcal{P}\mathcal{L}\mathcal{L}'\mathcal{L}''\mathcal{P}\hat{W}''.
$$

The second of these terms forces a projection onto the pure space between every application of the interaction picture Liouville superoperator, so the two superoperators have independent contributions. The first term, however, includes the direct action of two different-time superoperators sequentially before projecting back to the pure space. With a bilinear form between the $u$-th system order parameter $K_u$ and the $u$-th bath order parameter $\Phi_u$ of $H_{SB} = \sum_u K_u \Phi_u$, the resulting equation depends only on the bath time autocovariance $\text{Tr}_B\{\Phi_u(t)\Phi_u(t')\hat{B}_{eq}\}$. When this autocovariance is combined with the first term, the mean
values of $\Phi_v(t)$ are removed, and the combined expression reduces to the time-translationally invariant equilibrium autocorrelation between bath order parameters $u$ and $v$, $C_{uv}(\Delta t)$. Expanding the projection representation into a representation in terms of more parameterizable quantities, and reintroducing explicit interaction picture identifiers, we have

$$\partial_t \hat{\rho}^{(I)}(t) = -\frac{i}{\hbar} \sum_u \langle \Phi_u \rangle \left[ K^{(I)}_u(t), \hat{\rho}^{(I)}(t) \right]$$

$$- \sum_{uv} \int dt' \left( C_{uv}(t-t') \left[ K^{(I)}_u(t), K^{(I)}_v(t') \hat{\rho}^{(I)}(t') \right] - c.c. \right)$$

$$\partial_u \hat{\rho}^{(I)}(t) = -\frac{i}{\hbar} \sum_u \langle \Phi_u \rangle \left[ K^{(I)}_u(t), \hat{\rho}^{(I)}(t) \right]$$

$$+ \sum_{uv} \int dt \left( C_{uv}(t) \left[ K^{(I)}_u(t), K^{(I)}_v(t-\tau) \hat{\rho}^{(I)}(t-\tau) \right] - c.c. \right)$$

This is referred to as a quantum master equation. This particular equation is non-Markovian, as evidenced by the time-nonlocal $C_{uv}$. We can go one step further to bring this to the Markovian domain. If we assume $\hat{\rho}^{(I)}(t-\tau)$ is equal to $\hat{\rho}^{(I)}(t)$ for all times on which the correlation function $C_{uv}$ is nonzero, we can evaluate the $\tau$ integral directly as

$$\sum_v \int d\tau C_{uv}(\tau) K^{(I)}_v(t-\tau) = \Lambda^{(I)}_u(t),$$

and the final Markovian Redfield expression becomes

$$\partial_t \hat{\rho}^{(I)}(t) = -\frac{i}{\hbar} \sum_u \langle \Phi_u \rangle \left[ K^{(I)}_u(t), \hat{\rho}^{(I)}(t) \right] + \sum_u \left[ K^{(I)}_u(t), \Lambda^{(I)}_u(t) \hat{\rho}^{(I)}(t) \right] - c.c. \right).$$

Rewriting the final result in the Schrödinger picture is perhaps more illuminating, but can be found in reference [66] chapter 3.7. The main point is that by using a second order perturbation theory, the resulting system equations depend only on the correlation function of the bath order parameters, and that by assuming Markovity in the interaction picture – which requires the bath to relax much more slowly than oscillations occur – we can get a simple Markovian approximation that only depends on the integrated power from fluctuations and the coupling strength of bath order parameter $\Phi_u$ to system order parameter $K_u$. Despite a hefty bit of math, several key modeling assumptions had to be made to reach our final conclusion.

### 2.4.4 Semiclassics

An alternative to making both the bath and quantum system is to bring the quantum system into the classical domain. This sounds like a poor treatment of the quantum mechanics, but there is actually a long and important history of semiclassical quantum mechanics that recast a subset of quantum problems into a classical dynamics, with varying levels of quantum path integral phase cancellation behavior coming into play.

Semiclassics turn out to be a good approximation to the quantum dynamics in light harvesting complexes at room temperature, and they are amenable to baths of arbitrary
complexity. The results of open quantum solutions have been faithfully reproduced by semiclassics \[67\]. This suggests a further exploration towards using full atomistic simulations.

In fact, explicit simulations have been performed \[14\]. These simulations were affirming to the assumptions made by previous authors in that they did not find any new phenomenology. The limitation in this study is the high computational effort and complexity required to simulate, and thus, the ability to explore conformational diversity in the protein is quite limited.

A complementary method to open quantum dynamics is to use the empirical time-dependent hamiltonian generated by a classical MD trajectory as an input to a time-dependent Schrödinger equation, as in the NISE method \[68\]. Olbrich et al. report that the skewness in the epirically observed energy gap density of states has a substantial impact on the linear absorption spectrum, but does not continue to report any impacts on the dynamics \[12\].

### 2.4.5 Spectral density baths

The final subject to address is the spectral density bath in particular. The second order non-Markov equation 2.58 couples exclusively to the bath correlation function \(C_{uv}\). At this second order level of theory, there always exists a spectral density bath whose effect is identical to a bath belonging to an arbitrarily complicated family of bath models. Thus, it is sufficient to estimate a spectral density correlation function to capture the bath dynamics. Note that this level of theory is not guaranteed to suffice for our problem, but it serves as a reasonable starting point for parameterizing an effective dynamics.

Coincidentally, a harmonic oscillator bath itself forms a justification for the use of a theory of second order. In the Heirarchy Equations of Motion, this justification is used to generate a more exact theory for second order dynamics that includes an exact solution for a phonon bath, rather than a truncated perturbation expansion \[3\] \[69\]. The nonlocality inherent in the original equations can be replaced by auxiliary variables whose quantity grow as a function of the reorganization energy, thereby introducing a more gradual onset of intractibility as the reorganization energy grows and we enter a more FRET-like regime.

### 2.5 Conclusion

The models required to explain photosynthesis span multiple disciplines and modalities. For the theoretical research presented in this dissertation, an emphasis has been placed on models that form the foundation of theory, but certainly just as much could be written on models that are used primarily for explaining the experimental areas of research. For experiments, there is a wealth of mass spectrometry, spectroscopies across energy and time scales, and synthesis procedures that require robust models that fall outside of the scope of this dissertation.
All of that considered, an explicit approach to modeling gives us an ability to blend across these modes of thinking by borrowing assumptions and reducing the temptation of falling into the trap that there is one correct description of the photosynthetic process. For a multifaceted problem, a multifaceted approach is necessary, and a modeling framework provides a strong basis to explore a broader range those facets.
Chapter 3

Molecular dynamics of the FMO pigment protein complex

Proteins are complex entities of study that contain many constrained and correlated degrees of freedom. Their constituent atoms are connected by a topology of bonds, angles, and dihedrals stemming from valence interactions. Meanwhile, their charges add another layer of specific interaction between non-valence bonded atoms. With these constraints, proteins form complex secondary and tertiary structures that are able to fold and assemble robustly in biological environments.

MD must be performed properly for proteins if we hope to make any structural insights. The scale on which we would like to perform molecular dynamics varies based on our objectives, though. For example, we can learn about the dynamics of folded proteins with just a moderately sized simulation of the crystal structure, but such a simulation must be careful to sample the biologically relevant configurationals. If instead we would like to know about larger scale conformational activity, we will tend to prefer some sort of acceleration, in which case the details of the initial configuration are likely to be less important than selecting a robust collective variable to bias on. Finally, in the case of a folding simulation, we demand that we explore the majority of the configuration space, so considerable effort must be made to sample as broadly as possible.

Any proper MD procedure will involve refinements through iteration. Simulations can run into issues with both computational and physical origins. Slow physical dynamics may be real, but simulations can use acceleration to improve sampling. On the other hand, quenching methods may trap a configuration in an unlikely but sluggish trap; taking care with equilibration can mitigate this unphysical problem. An example of unphysical results of a simulation can be seen in figure 3.1.

This chapter serves two fairly distinct purposes. First, it serves as a case study. It should inform the reader about how to prepare a simulation and analyze the validity of a simulation – that is, that no simulation is correct on the first pass, and that preparing a simulation is a careful process of iteration. Second, it lays out functional aspects of the FMO protein. These purposes are achieved simultaneously, and the reader should acknowledge the possibility of
Figure 3.1: Top: Configurations of 4BCL with three major groups highlighted. Bottom: Configurations of 4BCL zoomed on the pink group. Left: A configuration from the 4BCL crystal structure. Right: A configuration from the end of a low-quality 500ns monomer simulation. After 500ns, the protein begins to unfold. All three groups (green, orange and pink) move considerably over these 500ns. Most notably, pink begins to unravel from the folded structure, as shown in the zoom.
these dual interpretations.

The chapter proceeds as follows. First, we discuss the preparation of the FMO configuration, enumerating the technical challenges that come along. Next, we interpret the FMO configurations and explore the structural analyses that lead to insights about equilibration and stability. Finally, we discuss the model iterations with an eye to identifying the ways in which model choices matter, despite occasionally being tedious.

### 3.1 Preparing FMO topology for sampling

While in theory MD is a rigorous and robust methodology, in practice careful preparation must be made to gather equilibrium samples. The theory states that each will be sampled in accordance to its Boltzmann factor $e^{-\beta U}$. However, this is only on long enough time scales that all states are visited multiple times. The initial configuration can be thought of as a bias, as it will also be sampled more than the other configurations.

Several choices are necessary to go from experimentally collected structure data to a computational model. In particular, we must make a choice of forcefield, protonation states, solvation, and restraints. Collectively these features describe the model.

Simulations are complicated even further by the presence of ligands. Ligands are ubiquitous in biology, but the parameters of forcefields are not optimized with respect to their presence. They require a degree of faith in the generalization of the forcefield. In this light, it is desirable to seek conclusions that are robust to slight changes in ligand parameters.

#### 3.1.1 Forcefield parameters

The success of molecular simulation depends fundamentally on force fields. The forcefield is the engine of the simulation. Like an engine, we cannot get very far if it doesn’t work, but there is also no singular correct implementation.

Ideally, our computed physical observables should be consistent across forcefields. The trajectories, of course, will be different. We therefore care about important emergent features that will remain consistent across simulations. For example, we expect the radius of gyration to have a similar distribution across different forcefields. Hopefully, properties of interest will obey this robustness to forcefield variation. We do not want to be too fussy about the choices of parameters if our goal is to find behavior that is generic.

Regardless of how distasteful a researcher might find it, sensible forcefield selection is an important step in the simulation process. We elaborate on the minimalist sensible choices made for modeling our FMO system.

#### Amber forcefield & methodology

Amber is a well-trusted forcefield for modeling proteins [70] [71] [72]. Amber is fundamentally physics-based. Its charges are based on a restrained electrostatic potential (RESP),
focused on optimizing the similarity between quantum chemical fourier coefficients. Forcefield dihedrals are fit to optimize similarity to a library of NMR data.

**Ligand forcefield**

All forcefields have areas of specialization. To capture emergent behavior, often a trade off between physics and emergent properities is made. This specialized optimization means that a forcefield optimized for one calculation may be quite poorly suited for another. It is because of this specialization that forcefield parameters for new molecules must be constructed with reference to a specific forcefield’s context.

After understanding the context of a forcefield, parameters of a force field are selected to best reconstruct some desired feature, and the ligand must cooperate with these choices. To compute ligand parameters, we generalize the forcefield model fitting procedure. The ligand can be further optimized for either more general or more specific usage; for example, the AMBER BChl A ligand parameters are fit to the vacuum vibrational density of states in order to align more closely with ab-initio physics as shown in figure 3.2, and therefore, to be more general.

In this way, we must be sensitive to both the reference forcefield and the ligand when selecting a forcefield from the literature.
CHAPTER 3. MOLECULAR DYNAMICS OF THE FMO PIGMENT PROTEIN COMPLEX

3.1.2 CDC parameters

In the case of an analysis that goes beyond the quantities that can be directly computed from MD, we must find further compatible analysis parameters. Our analysis, computing the gap, is perfectly well suited to a forcefield parametrization, which has been implemented in the charge density coupling (CDC) method [73]. The CDC parameters, like the AMBER parameters, are physics-based. They are a set of charges chosen to best reconstruct a potential. In the case of the CDC, they are selected to optimize the accuracy of coupling to atoms whose positions are empirically relevant, called the CHELPBOW method [74].

On top of the core parameter development, the CDC method also includes a factor that accounts for screening [62]. To compute this factor, the dielectric constant for the mixed media environment must be estimated. For FMO, this factor has been empirically estimated to be $1/3$ [62].

3.1.3 Missing residues and heuristic modeling

Once the forcefield is constructed, it is necessary to construct from experimental data a sensible system configuration. A good configuration will tend to be one which is incredibly boring. Boring is indicative that slow modes are not being explored and that the basin of configurations in a particular state of those slow modes is being explored extensively.

In refining a crystal structure, experimental groups will prefer data omission over false-positives. This makes sense from the perspective that an incomplete structure is still useful for many purposes, but also that any degree of incompleteness in a structure from a simulation standpoint is unacceptable. The role that these crystal structures fulfill is to contribute a strong baseline about the certainty of a structure, while leaving any decisions of how to handle the uncertainty to the practitioner who needs to have a complete structure.

Fortunately, we can grapple with uncertainty with well-tested and computationally efficient heuristics. One popular example of a program for this purpose is MODELLER [47]. This program implements comparative modeling, where the uncertain segments of a protein are compared against similar sequence segments in a different protein. On top of this comparative modeling, it implements a score system that approximates the free energy of the protein after integrating out the solvent, and further implements a Monte Carlo moveset to explore this configuration space with respect to this score. This combination of approaches makes it possible to get reasonable approximate configurations by leveraging libraries of proteins instead of using massive amounts of computation.

These heuristics do not find optimal structures, but they are much more able to explore a large space to find a promising point for later optimization. From a molecular dynamics standpoint, we do not need anything that is optimal, we only need things that are sensible. Molecular dynamics struggles more with global optimization, so this type of heuristic initialization makes the simulation process less difficult.

In FMO 3e0j crystal structure, there were two missing residue segments. The first missing segment is residue 214, a residue that lives on the surface of the structure. It is likely missing
because of the ubiquitous xray crystallography challenge of ascribing electron density to protein or to solvent. The second missing segment is the sequence of residues 1-6. This sequence is at the beginning of the structure, and is very likely intrinsically disordered when the protein is assembled \textit{in-vitro} rather than \textit{in-vivo}.

\textbf{Altloc}

While complete structural uncertainty is one possibility we have just explored, another common situation is that of few discrete alternate locations. An altloc will be represented in a crystal structure as an overlapping set of atoms labelled with different chain identifiers. These are used when the group modeling the crystal structure comes across two equally likely configurations that both score very well with respect to agreement with the electron density. In performing simulations, selecting between altlocs is often largely unimportant, as we simulate long enough that such minor differences are equilibrated away.

In the FMO 3eoj crystal structure, the altlocs were clustered in roughly two groups. Residues 164-170 were placed into ALTLOC B to allow for higher mobility of the BCL378
Figure 3.4: HIS and BCL form a biologically important contact in systems with Mg+ ions. In fact, four of the seven BCL molecules in FMO are held in place by this conjugation. A contact between one such of these pairs, BCL375 (magenta) and HIS (cyan, blue, and red), is shown with the conjugation represented by a yellow stick. The protein scaffold is represented in silver and alternate BCChl A molecules are in iceblue.

molecule. Residues 211-215 were placed in ALTLOC A because this configuration included a position for residue 215 (the ALTLOC B left residue 215 excluded). Residue 159 was placed in ALTLOC B to agree with previous work.

Protonation states

Even in the situation of complete conformational certainty in a crystal structure, we must always grapple with uncertainty of the crystal structure’s protonation state. Protonation varies in real compounds on a continuous basis. But unfortunately, averaging these protonation states is infeasible. We tend to prefer a rule-based standard protonation approach, where residues are chosen primarily by their pKa, but exceptions are cooked in for situations where binding is clear. These standard protonation rules lead to a reasonable configuration in most protein simulations, and are programmed into many molecular dynamics packages.
CHAPTER 3. MOLECULAR DYNAMICS OF THE FMO PIGMENT PROTEIN COMPLEX

For 3EOJ, we had to account for one additional biological consideration. The BCChl A molecules will conjugate their Mg+ atom to HIS groups, and in this binding, will require a particular protonation state of this group. The head of this group is asymmetric, so the crystal structure is able to identify for us which of the two possible HIS protonation states are occupied. After handling these residues manually, the remainder of the structure was treated with standard protonation rules.

3.1.4 Solvent and Ions

We have presumably at this point built a fine configuration, but if we want to avoid using an ad hoc implicit solvent forcefield, we need to manually insert solvent of our choosing. To insert a solvent, we always need to select our solvent/ion molecules and then select a desired ion concentration. Sometimes, though, a crystal structure already includes solvent molecules that may be unwanted. These solvents will be new molecules and will add complexity that is only present because these compounds were necessary to get the protein to crystallize. We can work around these additional molecule types by switching them out for the closest approximate substitution.

We must also be aware of the dangers inherent in the computational trick of periodic boundary conditions. Proteins will interact with their periodic neighbors if we do not construct a thick enough solvent layer. This is especially true when a protein possesses long strand-like residues, as in the 3EOJ residue 1-6 group. As a rule of thumb, 2 nm of solvent is sufficient, but it is important to check that the shape assumptions made for the solvent remain true over the course of the simulation. Position restraints can be included to prevent rotation of the molecule and thereby allow us to more tightly constrain the box size, saving a lot of computation from solvent molecules. Then with solvent added to the free space in the box, we conclude by converting some fraction of the solvent into ions. Choosing to include a physiological concentration of ions is often prudent as ions may be an important part of the surface of the protein.

For svnrhcdeni, the solvent configuration was straightforward. We used TIP3P-EW for the water model and NaCl as the ion pair. Ions were added to the solvent up to a physiological concentration. For the xmshbsri structure, the NH4 present in the system were exchanged for Na ions. This led to some ion trapping in the interior, the importance of which was never determined.

3.2 System stabilization

With the forcefield built and the configuration initialized, we enter the final stage before sample production whereby the forcefield and the configuration are gently relaxed into a state of mutual compatibility. Often, crystal structures have basic steric clashes. Furthermore, the crystal is built for generic purposes, and not to the specifications of any particular forcefield, so we should not expect any degree of compatibility to be guaranteed.
CHAPTER 3. MOLECULAR DYNAMICS OF THE FMO PIGMENT PROTEIN COMPLEX

Figure 3.5: Ions play an important role in the equilibrium structure of 3eoj. Left: cyan colored Cl$^-$ molecules accumulate in the top pore to interact with positively charged residues (blue). Right: blue colored Na$^+$ molecules aggregate in the bottom pore to interact with negatively charged residues (red) as well as in one cluster of residues on the perimeter consisting of negatively charged residues and polar residues (green). This binding is an example of the solvent structurally screening long-ranged electrostatics.

This stage may seem unimportant if we are to believe the axiom of ergodic sampling, but the realities of our computational capacities necessitate that we interpret ergodicity in only a limited capacity. The way in which ergodicity is usually interpreted in statistical mechanics can be explained with respect to “sampling bias.” A trajectory is sufficiently ergodic when the different slow behaviors can be thought of as being uncorrelated with the initial condition, and when these slow behaviors are sampled frequently enough that the trajectory could be modeled with some sort of reasonable stationary Markov process. Said differently, we do not need to sample all of the states to feel like we have met the assumption of ergodicity, we simply need to be unbiased with respect to which states we sample. When we stay within one major basin of attraction, we sample many, though not all of these states. Once we reach a stable equilibrium, we will hop around the network of state hubs in an unbiased fashion, and we can call the sampling sufficiently ergodic.

We therefore elaborate on the two primary tools, whether used in conjunction or separately, for system stabilization: energy minimization and position restraints.

3.2.1 Energy minimization

Energy minimization is an essential problem in computation, and the high-dimensional minimization of atomic positions with respect to the forcefield potential is a direct applica-
tion. While not the thermal (free energy) minimum – which is itself problematic to search for as it is not unique, and only defined with respect to some particular order parameter – the energy minimum is still very useful. Energy minimization techniques in fact do not often reach a particularly deep minimum. Actually, it is common for the potential energy to decrease further below the energy reached by the energy minimization thanks to global exploration. There are several methods that are common – steepest descent, which is the simplest and most robust; conjugate gradient, an accelerated method that is usually preferred to steepest descent, and L-BFGS, a fast semi-newton method that operates reliably with respect to minor perturbations, but which is attracted to saddle points in addition to minima [75].

If we energy minimize a symmetric structure, we expect the result to be symmetric as well. However, the solvent does not have the same symmetry as the protein; the solvent boxes added to the simulation are small boxes applied in a cubic tessellation, while the protein has a C3 symmetry. Especially in the case of steric clashes, the flow downhill can be very acute, and the differences in local solvent environments can make these minimizations very different between different monomers. For example, in energy minimization of FMO, the configuration of three-residue salt bridge 7y36y347 will fail to respect symmetry when a solvent is present. This salt bridge is a crucial and slow feature. Over 1.5 µs, this connection does not exchange conformational states.

### 3.2.2 Position restraints

Position restraints are a powerful tool with a variety of purposes. By holding a protein in place, we can preserve features during equilibration, model the environment, and control the approximations we are making during sampling. Let’s explore each of these purposes further.

- **Purpose 1**: During energy minimization and equilibration, position restraints are able to elevate the importance of particular structural elements while minimizing the importance of others. As discussed before, energy minimization can treat steric clashes quite aggressively, and if a solvent molecule takes the place that a residue is supposed to go, the two will be treated on equal footing without position restraints, potentially driving the protein into a configuration that is undesirable.

- **Purpose 2**: During production, position restraints can also be used to suggest and emulate physical features that are not explicitly modeled. For example, a constituent of a large complex can be modeled by position restraining its immediate environment, as in doing a single monomer simulation by restraining adjacent monomers of a large n-mer.

- **Purpose 3**: While they are not strictly rigorous, the reality is that when compared to an unrestrained simulation, a position restrained model can be much more sufficiently
Figure 3.6: A close up on the underside of the 3eoj protein showing violation of inter-monomer symmetry in the residue group 7-36-347. In the left monomer, circled residue 36 finds a relaxed position in which it is resting deeper in the bulk of the protein. In the middle and right monomers, residue 36’s position sits further outwards and with more contact to residue 347, which also causes interaction between residue 7’s charged oxygen groups and its charged nitrogen groups.

sampled. We can thereby be precise about the approximations made and validate our simulations by changing the position restraints.

Position restraints for xmslsbriP7y were placed on the 7y36y347 contact.

3.3 Quantitative interpretations of FMO configurations

Once we generate configurations, we must take the time to make structural interpretations. These interpretations attempt to give us perspective into why a configuration is favorable/unfavorable and physical/unphysical. Interpretations also offer a baseline of belief about the structure and fluctuations of the protein from which we can create a vocabulary of terms for a particular system. These terms function as a crude model for the actual behavior of the system; they are a form of uncontrolled dimensionality reduction.

Interpretations come in two main flavors: quantitative and qualitative. For pretty much any qualitative interpretation, a quantitative measure to probe it can be devised. However, qualitative interpretations allow a much broader range of data to be inspected simultaneously,
and such an overview will provide insights into the most essential aspects of the system to monitor quantitatively. These quantitative measures will help us quantitatively assess convergence and equilibration.

Beyond the importance of selecting good measures that probe features of interest, we must approach the interpretation of quantitative measures with an eye to the features of sampled signals. The field of signal processing is large, but a rudimentary understanding is necessary to get the most out of simulation data. Thus, a brief overview of sampling theory for molecular simulation is presented before discussing quantitative interpretations of protein structure.

### 3.3.1 On sampling rates for molecular simulations

It would be ideal to be able to generate perfectly independent samples of a protein’s configurations with an ideal sampling machine; unfortunately, time always plays a role in any sampling procedure that is either markovian or dynamical. With these time-dependent configurations, correlations in time are an essential feature. These correlations span a variety of time scales, and this time dependence of sample correlation is exactly why we must leverage sampling theory understanding to correctly parse a signal’s meaning.

Let’s assume for a moment that you take the plunge to sample a configuration from molecular dynamics at every time step of 1 fs — we will refer to this as complete sampling. For adjacent samples in the completely sampled signal, the configurations will be nearly identical except for the positions of hydrogens. If we average the time series in blocks of 10 configurations, the hydrogens will be frozen in their mean position, while the adjacent samples in this “smoothed” signal will show the motions of slightly heavier bonded atoms like carbons and nitrogens. We can again average these in blocks of 10, and arrive at a signal that only shows one slower level of motion. This sequence of operations is a naive filter, and shows how the motions in a system can be separated into high and low frequency constituents, even when those motions are not harmonic.

However, if we were to take the completely sampled signal and sparsify it (that is, take every 10th sample instead of averaging in blocks of 10), adjacent samples will still have the activity of hydrogens present. This process is called subsampling. Because we have subsampled the signal such that the hydrogen motion is not well resolved, we can no longer separate the slightly slower motions of the carbons from the motions of the hydrogens. When we sparsify our complete signal, we introduce aliasing.

Aliasing is a feature of all real sampled signals. Sampled signals have a maximum frequency that they can represent called the nyquist frequency, and all frequencies above the nyquist are “folded over” such that they cause frequency behavior other than their own. For very sharply resolved coherent frequencies, this can be to a disastrous effect, however, in stochastic signals this is less of a problem. High frequency noise in a stochastic signal will often fold over into what looks more like white noise. This type of noise is called alias pollution.
CHAPTER 3. MOLECULAR DYNAMICS OF THE FMO PIGMENT PROTEIN COMPLEX

Figure 3.7: Demonstration of the effect of direct sampling vs filtered samples for a particular desired sample frequency. The red line is a mixed high and low frequency signal, while the blue thick line is the low frequency isolated. The red dots are the red signal subsampled by a factor of 10,000, and the blue dots are the red signal subsampled by a factor of 100 and then filtered with a window size of 100. By filtering from a higher frequency subsample, we are able to capture the trend in the blue signal that was obfuscated by the high-frequency noise.

Because bond interactions in proteins are such high frequency compared to their dihedral motions, proteins can be analyzed with reference to least one separation of time scales. When observing equilibration over slow time scales, it is beneficial to reduce the noise from the high frequency part of the signal. Over long enough time scales, it becomes necessary to always use some degree of subsampling, though. Can we still recover the slow signal in the presence of alias pollution from the fast signal?

If we sample at an intermediate frequency, we can recapture major features of the low frequency signal through lowpass filtering. The demonstration from figure 3.7 shows that we can isolate the low frequency signal quite well even with just intermediate sampling. This demonstration took for the true signal a sum of 1000 high frequency cosines of random phase and angle and 400 low frequency cosines of random phase and angle. The low frequency part was isolated for methods comparison. After generating 100,000 data points, the signal is subsampled with two different approaches. The first approach is direct and subsamples down by a factor of 5000. The second approach uses filtering and subsamples by a factor of 50 before averaging in blocks of 100. When we compare the filtered signal to the subsampled signal, the filtered one shows the low frequency trend clearly, while the subsampled one fails to show the slow trends. The validity of this demonstration holds no matter how high the high frequency signal.

Occasionally, we get a break from alias pollution with a clever choice of analysis. Time autocorrelation function (ACF) is an excellent example of an analysis that avoids alias pollution. For time autocorrelation, averages are taken with respect to sample-pairs separated in time. Thanks to the fixed time difference and the averaging, we do not care about the
Figure 3.8: Radius of gyration plot for svnrhcdei simulation. The radius of gyration relaxes to a fairly consistent value after 600ns, indicating some degree of relaxation. Radius of gyration is sampled at 10ps and smoothed with a 20ns window. M1, M2, and M3 refer to monomers 1, 2, and 3, respectively.

sample rate except to the extent that we can only compute samples of the continuous ACF at integer multiples of our sample rate. Aliasing does not pollute the ACF in any way. Marginal distributions (with respect to time) are also unperturbed by aliasing.

With the details of sampling out of the way, let’s explore our dynamical features of interest.

### 3.3.2 Radius of gyration

The radius of gyration measures the second cumulant of the mass distribution, and is therefore statistically relevant and sensitive to relative movement. (explain relative motion). The radius of gyration is defined by the equation

\[ R_g = \sqrt{\langle (\mathbf{r} - \bar{\mathbf{r}})^2 \rangle_{\text{mass}}} \]  

where the averaging bracket \( \langle \rangle_{\text{mass}} \) is a mass-weighted average.

From figure 3.8, we can see that the distribution of masses broadens slightly but meaningfully. This type of relaxation of radius of gyration is typical of equilibration.

### 3.3.3 Potential energy

The potential is arguably the most fundamental measurement in a molecular dynamics simulation. It is the core of the forcefield. As molecular dynamics simulations equilibrate, the potential energy will often go downhill, particularly in protein simulations where the structure is hard to explore. As these potential energies are extensive, these energetic drops will
be substantive and will stick around due to their much higher Boltzmann weights. Overall, potential is an excellent feature to identify equilibration.

In figure 3.9, we can see that the equilibration occurs until 600 ns, after which the potential remains fairly stable after that.

3.3.4 Pair distances

Pair distances are a flexible but tremendously large category of quantitative measures. A good choice of pair distance is revealing, but a poor choice provides little to no information about a system. Figure 3.10 shows an important choice of pair distances between residues 37 and 347. While the large transitions of nearly 1 nm were visible to the eye (and in fact informed the decision to look at this residue pair), the subtler differences between purple and green trajectories were nearly impossible to detect by inspection. Selecting a pair of groups for distances is difficult and should be informed by some physical motivation. For example, in the xmshb-sri simulation, a significant opening formed between a pair of monomers. From this interface, several candidate group-pairs were identified. From these, this 36-347 pair was selected as the most causally relevant because of its intensity and temporal significance.

3.3.5 Radial distribution function

Solvent and ion analysis are difficult to join because solvents are intrinsically unstructured and fluctuating, while proteins have a significant amount of immutable structure that must be respected. We can consider the proximity of ions to particular residues using a radial
Figure 3.10: Trajectory of distances at different inter-monomer surfaces for residues 37 and 347. The trajectories show the breaking of symmetry even more clearly than figure 3.6. In fact, the two residues that looked quite similar from visual inspection actually show appreciable differences. Note that this is actually residue 37 and not the rendered residue 36 from 3.6, though the conclusion is the same due to the proximity of the two residues.

distribution function $g(r)$, defined by

$$g(r) = \frac{\langle N(r, \Delta r) \rangle}{4\pi r^2 \Delta r \rho},$$

(3.2)

where $N(r, \Delta r)$ is the observed number of particles in a window $\Delta r$ around radius $r$ and $\rho$ is the bulk density.

From Figures 3.11 and 3.12, we can see the affinity of these ions for particular binding locations, as well as an equilibration towards inter-monomer symmetry as time passes.

3.4 Qualitative interpretations

The main tool for qualitative interpretation is the visual inspection of molecular trajectories. In looking at these trajectories in a viewing program like VMD [76], the default representation never suffices. To see important structure, a wealth of representations are needed. A representation within VMD consists of a choice of drawing style, however, the usage here refers to the full set of choices to go from a collection of coordinates and topology information to a rendered image.

Some particularly complex representations can be made by iteratively nesting selections within one another. A selection of the interior solvent molecules was made by identifying
CHAPTER 3. MOLECULAR DYNAMICS OF THE FMO PIGMENT PROTEIN COMPLEX

Figure 3.11: g(r) in time across monomers. The amplitude of g(r) communicates how stiffly bound an ion is, while the peak position communicates how closely bound the ion is. As time elapses, the g(r) across different monomers converge.

Let’s consider some of the major qualitative insights that were made about the FMO complex.

3.4.1 Ion binding

Ions, being transient, are difficult to track quantitatively. In fact, the method used in the previous quantitative section depended on a reasonable choice of a reference residue. How do we select this residue?

From a time sweep which activates residues in proximity to ions, we can get a qualitative sense for which residues have a high ion affinity. This is possible using VMD’s “Update selections every step” timestep option.

Once we have selected a few candidates, we can watch them more closely and identify structural properties of the binding.
3.4.2 Solvent affinity

Solvent affinity is a major driver for binding in proteins, so its is also a feature that is worth being attentive to.

Let’s examine figure 3.13 of the upper surface of the FMO complex. The upper solvent-exposed surface is nearly 50% hydrophobic, quite a large number for a protein. Because this face attaches to the baseplate, this hydrophobicity likely plays a role in the attachment. We also observe that BCL378 has a hydrophobic tail which sticks into the solvent on this face. Because of the extent of hydrophobicity on this face, the phytol tail of BCL378 is able to find several conformations that are suitable to its stability.

Let’s examine figure 3.14 of the upper surface of the FMO complex. The lower surface is quite hydrophilic as would be expected of a protein. The structure has a chain of residues 1-4 which are all hydrophobic, and have an incredibly difficult time reaching equilibration. Their erratic structure percolates through the rest of the protein, breaking salt bridges that might otherwise be respected.
CHAPTER 3. MOLECULAR DYNAMICS OF THE FMO PIGMENT PROTEIN COMPLEX

3.4.3 Solvent penetration

While the FMO trimer seems well packed together, there are identifiable, persistent, and symmetric solvent channels running through the structure that can be identified. These channels stem from three major pockets: the grooves in the center at the top and bottom and the divet in the pocket of loop 2. The channels connect the upper groove and side pocket. The lower groove is disconnected from the other two pockets.

This water network suggests that it will not only be BCL372, which is solvent exposed through its own solvent pore, that interacts with unscreened solvent. In fact, the network approaches BCL375 quite closely. This detailed network further underscores the importance of explicit solvent in protein simulation.
CHAPTER 3. MOLECULAR DYNAMICS OF THE FMO PIGMENT PROTEIN COMPLEX

3.4.4 Interpreting the in-vivo analogous environment

Returning to the hydrophobic tail on the protein, hydrophobic residues 1-4 (as visible in figure 3.14), we are left to ask the function of such a residue chain. Given that FMO is a protein that docks onto a membrane, we might expect these hydrophobic tails to interact favorably with that membrane and act as one of several forces anchoring the system to the membrane.

If this membrane docking is part of the protein function in-vivo, what then should be done to simulate this protein in the absence of such a membrane? The residues 7, 36, and 347 are quite close to this hydrophobic tail, and position restraining these residues makes it possible to emulate the protein being docked, and the tail residues being relaxed into the docked membrane.

The upper surface poses another question. The largely hydrophobic surface is likely more disordered than it would be in the presence of contact with the chlorosome baseplate.
Figure 3.15: The network of water molecules in the complex rendered with a cut in the protein. The frontmost residue is rendered as a glass cartoon for minor context. Water molecules are represented with dynamic bonds to emphasize the network structure present. The interior channels are a notable feature of this hydration network, which could allow for ion presence inside of the trimer and have important electrostatic implications on the structure.

Baseplate protein CsmA has large hydrophobic patches that would rationalize this structure. But this leaves the question open of whether FMO would function differently in the presence of this protein.

Through just the meager position restraints on 7y36y347, we are able to reach an equilibrated structure in a reasonable amount of simulation time, which suggests that this configuration is acceptable, functionally speaking.

The protein without ions equilibrates much more quickly, but Green Sulfur Bacteria are actually notably salty, so the exclusion of ions seems unphysical. The protein has a net positive charge, so only Cl\(^-\) ions would be added for a neutral configuration. The observation that Na\(^+\) ions dock into the structure as well, despite the fact that they are not doing so simply to neutralize the net charge of the structure, is a strong indication that these ions are a part of the structure in-vivo, and should be modeled.
3.5 Model iterations

As discussed at the top of the chapter, model construction is an iterative process. I go through the various stages of model construction and explain their functions, good and bad.

3.5.1 4BCL structure

The first structure I simulated was the 4BCL structure. This is a structure from *C. tepidum*, and is fairly old. It is therefore missing several residues in their totality. This early iteration was run with these missing residues excluded from the structure, which surprisingly did not lead to any glaring structural deficiencies. More importantly was that the system was run 1. as a monomer and 2. with only neutralizing solvent. The expected effect of a collapse of loop 2 and alpha 1 was observed, as these groups are both held in place by their adjacent monomer in the biological assembly. More unexpected was the observed unwrapping of residues 8-10, on what would be the lower surface of the trimer. This site binds several NA ions, and it is possible that their absence led to a destabilization of this region.

3.5.2 3EOJ xmshbsri

After 4BCL, we switched to using 3EOJ, a structure from *P. aestuarii* which is seen more frequently in literature. This structure is missing fewer residues, but includes many ALTLOC choices. It improves upon 4BCL through the inclusion of BCL378, which is only stable in the biological assembly. Thus, the transition to trimer assembly was essential.

In addition to using the base structure, we followed a more rigorous preparation protocol. Each preparation step, roughly, is specified by one letter in its code.

- x: xtal structure; selection of ALTLOC A for simplicity
- m: model; modeled residues 1-6 and 214 with MODELLER program
- s: symmetrized; xm-monomer was fit to original xtal locations
- h: hydrogen; hydrogens were added to the structure and protonation states were selected
- b: box; box volume was adjusted to accommodate enough solvent for non-interacting pbc.
- s: solvent; solvent was added
- r: remove; solvent that embedded into the structure was manually removed (it later returned, confirming that it was not just an artifact of system setup)
- i: ions; solvent molecules were turned into ions.

This structure had a very difficult time equilibrating. Most notably, the modeled tail region (residues 1-4) did not equilibrate, and displaced other equilibrated structure due to its high hydrophobicity.
CHAPTER 3. MOLECULAR DYNAMICS OF THE FMO PIGMENT PROTEIN COMPLEX

3.5.3 3EOJ xmshbsriPR37x

To facilitate equilibration, position restraints were added. Three-residue salt bridge 7y36y347 was position restrained throughout the simulation. Unfortunately, this did not solve the problem of poor equilibration of residues 1-4. Several other position restraints were tried, but none succeeded in relaxing this group.

Ultimately, the largest problem with this configuration was that the BChl Mg-coordinated histidine groups were not correctly protonated by the standard protonation. These HIS groups should have a protonation state such that the coordination to the Mg\(^+\) of the BChl a occurs via an unprotonated nitrogen. The conformational state of the protein dictates which of the two nitrogens is involved in this coordination, and thus, we can determine the correct protonation state from this information. These incorrectly protonated histidines generated a lot of spurious conformational activity that had a disastrous effect on the gap calculation.

Less disastrous but still important, the symmetry mismatch between solvent and protein caused one monomer’s residue 36 to adopt an alternate configuration (as shown in 3.6). This was shown to be avoidable with a more careful energy minimization procedure that applied constraints immediately.

3.5.4 3EOJ svnrhP7y

This structure is a renovated structure based directly on a structure previously used in the literature that was shared by the original author [16] [77]. The structure started with 3EOJ and made several structural decisions:

1. Use ALTLOC B for alpha 1 to allow for higher BCL8 mobility
2. Protonate HIE for Mg-coordinating histidine groups for BCL 1, 3, 4, 7
3. Protonate HID for Mg-coordinating histidine group for BCL 6 and all other HIS.
4. Use MODELLER to construct residue 214
5. Solvent was included with neutralizing Na ions.

Beyond this, the structure was prepared for simulation as follows:

- sv: S. Valleau; author initials, the initial structure
- n: correction of “Na+” to “NA” for GROMACS
- r: residue re-numbering
- h: hydrogens; added hydrogen and used standard protonation for all groups other than HIS

Position restraints were included on the residues 7 (the “N-terminus” for the structure), 36, and 347 after the first energy minimization. Energy minimization was performed carefully not to disturb the basic conformational state of this salt bridge (as it had been affected in 3eoj xmshbsriPR37x during energy minimization).
3.5.5 3EOJ svnrhcdeiP7y

A further extension of the previous structure was made for physiological ion concentration.

The letters in this structure are as follows:

- svnrh: previous structure
- c: centered protein
- d: density-adjusted box
- e: energy minimization; to reduce risk of ill-behaved early trajectory after adding ions
- i: ions included
- P7y: position restraints on the 7-36-347 residue group nicknamed “7y36y347”.

3.6 Conclusion

From a poor simulation, there is little to be said about our fundamental topic of study (i.e. the optical properties of the pigment-protein complex), but still some to say about the molecule itself. We can use the insights from poor simulations through the process of iterative refinement. With iterative refinement, we can be confident about the final state of our simulation, its strengths, and its inevitable weaknesses.
Chapter 4

FMO’s high frequency fluctuations are protected

With the dominant time scale for energy transfer in the FMO complex hovering between 100 fs - 1 ps, the bulk of research has studied dynamics of gap fluctuations that occur on these fast timescales through a spectral density lens. A spectral density is the minimal description for any continuous stochastic zero-mean time series, just as the vector of variances is the minimal description for any collection of zero-mean vector of random variables. While this may not be a perfect description of the time series, the spectral density is perfectly sufficient to describe the emergent quantum dynamics for certain levels of theoretical approximation, as described in section 2.4.3. For the frequency range spanning periods of 1 fs - 10 ps, the spectral density is quite rich, and acquiring a good estimate of it is not a particularly simple task.

We are interested in not just the effect of the gap fluctuations on the excitation trajectories, but of building a sufficient model of the gap time series itself. When it comes to selecting a data model for a time series, the existence of a well behaved autocorrelation function for that time series with which a spectral density can be computed does not imply that the spectral density performs well as a data model. We would much prefer a data model, or a “generative process,” that will emulate the trajectories from the molecular dynamics. The spectral density is a good start, as it captures the marginal variance for sample trajectories of finite duration \( T \). If for example, we crudely consider additional features like skew and kurtosis as a function of sample length \( T \), we can refine the data model further.

In this chapter, we seek to address the disconnect between data model and theory by validating the use of a spectral density as a data model for FMO. We consider significantly longer trajectories (2 \( \mu \)s) than any others considered in the literature (typically about 50 ns). We observe a wealth of dynamical heterogeneity in the gap trajectories that extends well beyond simple equilibration. Despite this dynamical heterogeneity, we perform analysis on the fast features independently from the slow features, and find that the high frequency spectral density is robust to the dynamical heterogeneity. We therefore conclude that the use of a spectral density model for coupling to high-frequency fluctuations is well justified,
but this spectral density is a poor description of fluctuations on nanosecond timescales and longer.

The chapter proceeds as follows. First, we discuss the unique challenges that simulating a protein poses for acquiring independent and unbiased samples. Then we explore the dynamical heterogeneity present in the gap trajectories, and identify how they stem from essential physical fluctuations. Next, we look at how well the spectral density describes even the high frequency signal as a fundamental data model. Finally, we analyze the structure of gap contributions to argue that fast modes are protected through localization.

4.1 Protein dynamics create unique challenges to sampling the gap

The first requirement for computing any statistical estimates is a robust and reliable set of samples. To sample the CDC estimate of the gap (henceforth, the “gap”), we first generate samples of molecular configurations through molecular dynamics. These samples of molecular configurations give us the raw data from which the gap can be directly computed with the CDC method, according to the equation

$$G_{\kappa} = \sum_{i \in A \cap B_{\kappa}, k \in B_{\kappa}} \frac{q_i q_{\kappa}^{\text{CDC}}}{|r_i - r_k|},$$  \hspace{1cm} (4.1)$$

where $A$ is the set of all atom indexes, $B_{\kappa}$ is the set of all atom indexes in chromophore $\kappa$, $q_i$ is the forcefield charge of atom $i$, $r_i$ is the position of atom $i$, $q_{\kappa}^{\text{CDC}}$ is the CDC charge of atom $i$, and $f$ is the empirical screening factor. With this equation, we can compute a gap value for each of the 24 chromophores for any molecular configuration.

Thanks to the linearity of the CDC approximation $4.1$, the contributions to the total gap can be disaggregated into arbitrary atomic groups. By disaggregation into meaningful segments, we can use the gap samples to understand the impact of specific FMO landmarks on particular gaps in addition to understanding the spatial extent to which entities in the environment contribute to individual gaps. Here, we focus on disaggregation by residue because atomic contributions within a residue will be highly self-correlated as the motion within a residue is strongly constrained by bonds. Expanding our notation to define $R_I$ to be the atom indices in residue $I$, an index which wraps together both residue id and monomer, we can write

$$g_{\kappa I} = \sum_{i \in R_I, k \in B_{\kappa}} \frac{q_i q_{\kappa}^{\text{CDC}}}{|r_i - r_k|},$$  \hspace{1cm} (4.2)$$

$$G_{\kappa} = \sum_{I \in \{\text{RESID} \neq \kappa\}} g_{\kappa I}. $$  \hspace{1cm} (4.3)$$

Samples of the CDC gap can be considered for both dynamic and static properties. Each of these are important to characterize the gap dynamics, with static features taking somewhat
of a precedent because of their simplicity and robustness. Biased and advanced sampling methods can aid the calculation of static distributions, but at the expense of eliminating dynamical information.

Independent of how well we sample, we must be wary of treating the CDC as anything more than a collective variable. Values for the CDC are parametrized for relaxed, average calculations — not for fluctuations. However, being physics based in the coulomb interaction, the CDC is able to capture the impact of a variety of ranges of interactions that are physically meaningful, though not always necessarily in the correct proportions. In this sense, the disaggregated gap samples are perhaps more meaningful than the aggregate ones, as any single location’s modulation in fluctuation intensity will not be concealed by the aggregated error which emerges from combining all sites into one estimate.

With a grip on the mechanics of performing sampling, let’s now consider the realities of gap sampling. Sampling in this situation is quite difficult, owing both to the protein and the gap. Proteins are notoriously slow to relax, and these simulations are no exception to the rule. One challenge in protein relaxing a protein is determining exactly how to identify that relaxation is complete. A range of quantitative methods have been discussed in chapter 3.3. In addition to the protein relaxation problems, the gap is also a uniquely sensitive collective variable. Being coulomb based, it is long ranged, so it is impossible to rule out any source of perturbation as being the culprit for a particularly striking effect. Moreover, the coulomb coupling is formidable at close ranges, making local effects impossible to rule out as well. Protein summary statistics which describe any aspects of convergence and thoroughness of gap sampling are difficult to come by.

Owing to the precedent in the literature [11] [13] [12], we pursued a modest amount of equilibration in early attempts to sample the gap. These modest amounts of sampling failed in robustness and temporal homogeneity. Therefore, the collection of gap samples was performed in depth.

4.1.1 Biased simulations do not sufficiently accelerate protein sampling

In order to get very thorough sampling of the gap, biasing seemed viable. The fundamentals of biased sampling are covered briefly in chapter 2.2.3. Biasing accelerates sampling because it allows us to capture a fraction of a sample. Because samples are fundamentally discrete, to estimate the probability of a 1 in 1,000,000 event well, we need at least 1,000,000 samples, but more realistically 10,000,000 or 100,000,000. With a biasing potential, we can make the chance of observing that rare event quite large, say 1 in 10, but thanks to our known biasing procedure, we know that we have enhanced the probability by a factor of 100,000, and we can weight each observation in this biased ensemble by a factor of 1/100,000. Therefore, we can estimate the probability of rare events like these with 100,000 times the precision of the unbiased ensemble. With Gaussian distributed variables, such extreme probabilities lie just five standard deviations from the mean.
Figure 4.1: Despite significant variety in response mean, slope for response across all samples is consistent with the predicted slope of equation 4.10 as determined by the variance MLE.

The biasing was implemented in a linear fashion, as

$$H'_J = H_0 + \lambda G_J. \quad (4.4)$$

Following chapter 2.2.4, we can quickly estimate the response by approximating the process with the quadratic free energy function which is roughly observed empirically according to

$$F(G) \approx \beta \frac{G - \bar{G}}{2\sigma_G^2} \quad (4.5)$$

$$F'(G, \lambda) = F(G) + \lambda G. \quad (4.6)$$

$$ (4.7)$$

We then solve for the new minimum in the biased ensemble

$$\frac{dF'(G^*_\lambda)}{dG} = 0 \quad (4.8)$$

$$\left( G^*_\lambda - \bar{G} \right) = -\beta \sigma_G^2 \lambda \quad (4.9)$$

$$\text{slope} = -\frac{\sigma_G^2}{k_B T}. \quad (4.10)$$

The result is that the variance of the energy gap and the temperature are sufficient to predict the response slope of a biasing procedure. We validate that this hypothesis holds in figure 4.1.

Unfortunately, biasing was unable to accelerate sampling sufficiently, and itself failed to generate equilibrated results. One simple test for correctness is that two repetitions of the
same simulation sample the same distribution. To generate different trajectories in the biased ensemble, short 500 ps biased simulations were shot off of samples from a longer trajectory. We compared the results of selecting initial configurations for trajectories uniformly from a 50 ns trajectory to the results of using an identical initial configuration for all trajectories in the biased ensemble.

Figure 4.2 shows the challenging results of this biased sampling procedure. While both procedures produce similar slopes of shift, the procedure that randomly samples initial configurations produces samples that are strongly mutually inconsistent, even in terms of their mean. Simulation errors were taken by estimating the standard error of mean by splitting the trajectory into 5 smaller trajectories, then estimating the distribution of the mean for the full trajectory from these. For the simulations started from the same initial configuration, results were much more consistent, but were still not consistent enough to conclude that the trajectories were truly sampling the same distribution. These inconsistent biased simulations were the first indication that the system was seriously dynamically restrained, with very slow relaxation behavior that included time scales even as fast as 100 ps.

Oddly, the slope generated by biasing in the two scenarios was consistent. The broad distribution of gap means across initial configurations suggests that these configurations might also have different responses. We instead observed a consistent average response from all initial configurations.

Even more interestingly, the consistency of the response slope with the signal variance in figures 4.1 and 4.2 across a broad range of means implies that there is a consistent high-frequency variance across many different average environments. Such an observation

Figure 4.2: Left: a collection of 500ps biased trajectories where protein initial configurations were sampled from a 50ns trajectory uniformly. Right: a collection of 500ps biased trajectories where protein initial configurations were identical. The greater spread in the left figure proves that gap values depend strongly on subtle details of the protein conformation. Despite the differences in variance, though, the responses are not significantly different.
begins to lend evidence that there might be an effective spectral-density description of high-frequency modes that is consistent even across changing protein environments.

4.1.2 The multi-well model of protein landscapes explains slow dynamics

The failure to generate consistent biased data corroborates a common approach taken to understanding protein fluctuations, the multi-well model. This model describes a simplified two dimensional free energy landscape that is a function of a collective variable with two dimensions (though in general, this type of model can be posed for higher dimensional landscapes). In the multi-well model, this landscape possesses a wealth of local minima separated by free energy barriers. Over a modestly long dynamical trajectory, only a few observations of this two-dimensional collective variable hopping between wells are observed, and therefore, the average over all such wells is not properly converged.

Consider again figure 4.2. The slope generated by biasing in the two scenarios was consistent. The broad distribution of gap means in the scenario with multiple initial configurations suggests that these states are members of different wells, and therefore, one might expect that the well curvatures vary significantly. We can probe this as such: since we do not observe barrier crossings on the 500 ps time scale considered in this analysis trajectory, the response to bias will be inversely proportional to the curvature of the current well. From the data observed, there is no indication that it would be important to reject the hypothesis that well curvatures are comparable across the board.

Biased sampling typically is presented as a way to address well and barrier landscapes, but in this situation, we see a complete failure of biasing in this regard. Why is this? The order parameter being biased upon, the gap, is not one of the slow coordinates of the system. While transitions in the slow collective variables will correlate with extreme gap values, there are a wealth of configurations with extreme gap values that do not incur transitions. Thus, biasing the gap to extreme values will simply have no consistent effect on the transitions on the slow variables which bottleneck sampling in the system.

Because of this failure of the gap collective variable to induce global transitions, and because the biasing procedure is computationally expensive, we pursue the direct approach. For situations where the number of wells is sufficiently many, or if the important collective variables are unknown to the practitioner, direct sampling is a sensible way to manage their impact. Contrary to sampling far into the tails, where biasing is tantamount to being able to compute probabilities, we can allow direct sampling to do the work for us by sampling transitions randomly, and using the approximation that the directly sampled states are a fair and unbiased random sample of all of the wells.

The multi-well theory also presents a reasonable lens within which we can interpret what it means to have the fluctuations protected. If the wells lie along a coordinate orthogonal to the main fluctuating coordinates, then we should not expect well transitions to affect fluctuations. The orthogonality of the slow collective variables to the bias on the energy
CHAPTER 4. FMO’S HIGH FREQUENCY FLUCTUATIONS ARE PROTECTED

Figure 4.3: The trajectory for BCL373 shows fluctuations on both slow and fast time scales. The separation of scales is essential because the fast fluctuations from any point in time span the range explored by the low-frequency variability. When considering the original signal, monomers are somewhat difficult to distinguish; when considering only means, the differences between monomers becomes more stark. Left: full-data trajectory, of high-frequency quantiles, demonstrating a convolution of fast and slow data. Right: low-frequency trajectory, making clear the case for distinguishability of monomer due to small sample size.

gap provides indication that the bias fluctuations will be preserved as the system transitions wells.

4.1.3 Direct simulations sample conformational states slowly but reliably

With direct sampling, relaxation can be observed over time scales of 1 µs, and while this is expensive, it provides assumption-free confidence that the molecular simulation is sampling the space correctly.

Direct sampling was performed for 500 ns or more under several different protein models, each of which improved on the previous model’s deficits. These deficits are explained from a protein perspective in chapter 3.5, but the CDC samples are particularly able to emphasize the modeling failures.

In particular, for 4BCL, the atomistic trajectory showed unravelling of the N-terminus, but due to anomolous solvent exposure, the CDC trajectories were already showing exaggerated fluctuations. The CDC trajectory for 4BCL also showed a tremendous amount of dynamical heterogeneity from irreversible conformational changes stemming from the instability of the complex. For xmsihbsriP37x, the incorrect protonation states of HIS residues conjugated to BChl A Mg\(^{+}\) groups led to unphysical gap bimodalities stemming from a clash between the incorrect hydrogen on HIS sterically interfering with the Mg\(^{+}\) of the BChl A, leading to a bimodal configuration for HIS, and therefore a bimodal gap value.
4.1.4 Summary: Protein dynamics create unique challenges to sampling the gap

The protein environment contributes a complexity and richness to the FMO system that is worth careful investigation. In this spirit, we explored both biased and unbiased sampling procedures, and found that the kinetics of the landscape are dominated by features orthogonal to the gap. These sluggish dynamics suggest an unfilled need for a biasing methodology that is able to enhance dynamics for a system like this. Given that it is the space orthogonal to the order parameter that must be sampled, the ideal biasing would ignore the gap and instead identify slow dynamical features of the system and bias on those to achieve the enhancement. One such enhancement that could be developed could focus on enforcing symmetry among the monomers with a replica exchange set up; while this symmetry is artificial and likely not the optimal state of the system, a symmetry bias will allow for fast exchange of conformational trapping. Another such enhancement would be in the form of ion Monte Carlo moves. Ions play a crucial role in the protein structure, yet they are slow to penetrate into the interior of the complex. A Monte Carlo move that enhanced the exploration of the solvent space, particularly with respect to accessing the interior solvent, would speed equilibration many fold.

Fortunately, the direct samples are largely sufficient to perform an analysis of the complex, and they further carry the benefit of containing the proper dynamical information. We proceed with direct gap samples from the position-restrained svnrhcedeiP37y simulation.

4.2 Gap trajectories have multifaceted heterogeneity of physical origin

We begin our gap analysis with a descriptive overview of the data in order to validate the data integrity and sensibility of our simulations.

To assess the quality of direct gap samples, we can use summary statistics of the trajectory including statistics such as empirical distributions, time-disaggregated empirical distributions (i.e. empirical distributions for small segments of the full signal as a function of time), and trends in the signal. We know beforehand that our time series should be stationary. It is therefore prudent to apply some rudimentary tests for the quality of that assertion before we perform analyses that rely on such assumptions.

Assessing summary features like these is the first stage of identifying the feasibility of various models – including the spectral density. For example, it is a fool’s errand to compute the spectral density using a signal that has nonzero trend; this trend will prevent the full decay of the autocorrelation function, and depending on the nature of the trend, may grossly overestimate the contribution of low frequency signals to the correlation. In the common situation of equilibration, there will be trend between the beginning and the end of the
signal, and the autocorrelation function will treat this as a spontaneous fluctuation rather than a relaxation procedure.

### 4.2.1 Warning about generalization of results

We must remain vigilant of differences between chromophores, as they each have unique local environments (except, in the complete sampling limit, those with the same chromophore number). These local environments vary in charge composition and have been demonstrated to have significant effects on the mean gap [62] – there is therefore no reason to expect that they do not also incur different fluctuations. Ligation to HIS and exposure to solvent are two particularly notable features in addition to unavoidable differences in neighboring residues.

To make it easy to identify the particular BChl molecule being analyzed in any particular image, I introduce a color scheme in figure 4.4 based on ColorBrewer’s 8-class Dark2 [78]. Each residue is assigned a main color, and different monomers are distinguished by shades of that color. As with any 8-class color scheme, some pairs of colors are harder to distinguish than others. This color scheme is indicated for general presentation as it is colorblind-safe.

We must also remain vigilant that different molecular models will give different results (e.g. 4BCL against xmshbsriP37x). All data should be assumed to be from svnrhdeiP7y unless otherwise specified. For the meaning of particular simulation codes, consult chapter 3.5.

Despite all of the differences present in structures and residue locations, the findings about protection of fast fluctuations are generalizable. In any particular BCL’s environment, there may be additional insights to be made, but due to the challenges to protein sampling, such insights are at high risk of spurious feature identification. While we focus on one BCL molecule, 373, we focus on conclusions about this BCL molecule that generalize to the others,

<table>
<thead>
<tr>
<th>BCL</th>
<th>DoS 25%/75% (cm⁻¹)</th>
<th>solvent?</th>
<th>HIS?</th>
<th>top hf intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>371</td>
<td>-91, -44</td>
<td>No</td>
<td>Yes</td>
<td>HIS110, LYS150</td>
</tr>
<tr>
<td>372</td>
<td>-171, -121</td>
<td>Yes</td>
<td>No</td>
<td>SER72, LYS78</td>
</tr>
<tr>
<td>373</td>
<td>-211, -160</td>
<td>No</td>
<td>Yes</td>
<td>SER13, HIS289</td>
</tr>
<tr>
<td>374</td>
<td>-204, -164</td>
<td>No</td>
<td>Yes</td>
<td>HIS290, VAL352</td>
</tr>
<tr>
<td>375</td>
<td>-95, -46</td>
<td>Some</td>
<td>No</td>
<td>PRO244, LEU242</td>
</tr>
<tr>
<td>376</td>
<td>-198, -155</td>
<td>No</td>
<td>Yes</td>
<td>HIS145, TRP184</td>
</tr>
<tr>
<td>377</td>
<td>-90, -51</td>
<td>No</td>
<td>Yes</td>
<td>HIS297, GLN198</td>
</tr>
<tr>
<td>378</td>
<td>-96, -29</td>
<td>Yes</td>
<td>No</td>
<td>ARG125, ASP160*</td>
</tr>
</tbody>
</table>

Table 4.1: Table of features of individual BCL molecules in the complex. Each environment has a different local environment that leads to a different spread in gap values. Each also has a few key adjacent residues that most strongly impact its high-frequency fluctuations. Residues with an asterisk (*) are located on the positive adjacent monomer (i.e. for a chromophore on monomer 2 ASP160* is located on monomer 3).
CHAPTER 4. FMO’S HIGH FREQUENCY FLUCTATIONS ARE PROTECTED

Figure 4.4: A representation of the BCL pigments in FMO colored with their own identifying color that will be used throughout the chapter.

except when a result is stated to be specific to BCL 373.

4.2.2 Basic gap descriptions show substantial heterogeneity across most major factors

The guidance for what to validate first in a dataset comes from what one already believes to be true about the data – for example, you should validate that a time series is homogeneous in time if you believe that to be true. To show that a time series is homogeneous in time, it is simple to plot the data directly. In our case, as seen in figure 4.3, we struggle to reconcile a separation of time scales; that is, due to slow variations in the mean, it is difficult to assess whether the fast fluctuations are homogeneous, and conversely, because of the breadth of the fast fluctuations, it is difficult to identify trends in the mean precisely.

This is one case in which truly direct visualizations of all of the data are unable to communicate hardly anything; instead, we can understand the structure of the data with a variety of non-parametric summary statistics. Seeing these separate time scales at play, we can use a statistic that captures the features of the fast distribution (such as a few quantiles or a standard deviation) while simultaneously tracking a different statistic that captures features of the slow distribution (such as a mean). These methods are quite crude and
Involving nearly no assumptions about the data, but they are able to reveal structure that would have been difficult to detect by direct examination.

There are handful of factors to understand in this data set, most notably: (a) time, (b) chromophore, (c) monomer, and (d) position. We treat time and position marginally, but condition each analysis on chromophore and monomer due to large amounts of heterogeneity. We also marginalize both time and position to look at chromophore and monomer as individual factors conditioned on one another.

**Time as a factor**  To understand time as a factor, we must consider substantially better representations than the direct time series. The CDC data has a very large high-frequency aspect, yet a comparable low-frequency one as well. This separation in time scales makes representation a challenge.

One strategy of representation that is amenable to this separation of time scales is to treat the fast time scales as a distribution, and to observe changes to that distribution in time. Ideally, this is done in a non-parametric way, as the large amount of data makes non-parametric estimates quite feasible, while parametric estimates will require a handful of assumptions that we are seeking to validate and understand in the first place.

One easy non-parametric description is through the use of percentiles. The \( p \)-th percentile \( G_p(t, \Delta t) \) can be calculated for sampled function \( G(T) \) by considering a finite window of data points of width \( \Delta t \) spanning backwards and forwards in time about time \( t \), \( \{G(t - \Delta t/2), ..., G(t + \Delta t/2)\} \) via

\[
G_p(t, \Delta t; G(t)) = \text{percentile}(p, \{G(t - \Delta t/2), ..., G(t + \Delta t/2)\}). \tag{4.11}
\]

This percentile estimate is robust to any amount of subsampling, sufficing that the number of data points is large enough, so it is an effective to describe the time series evolution in the presence of this separation of scales.

Every percentile time series displays some degree of temporal heterogeneity. If we expect to describe these gap dynamics properly, we must be careful to both separate time scales, and if we hope to understand the source of this heterogeneous behavior, to sample sufficiently to capture the behavior.

**Chromophore as a factor**  For EET, the interchromophore relationships are essential. To understand just how essential they are, consider the overlap integral between two Gaussian distributions with identical variances.
Figure 4.5: A percentile timeseries for every chromophore in the trimer. The interquartile range is rendered in the interior, and the remaining range between the 5th and 95th percentiles is rendered in a lighter shade.
Figure 4.6: The chromophores have densities of states that are tightly packed. Separations between means are on the scale of 1-3 standard deviations.

$$I(\Delta) = N \int dx \exp \left( -\frac{x^2}{2\sigma^2} \right) \exp \left( -\frac{(x-\Delta)^2}{2\sigma^2} \right)$$  \hspace{1cm} (4.12)$$

$$= N \int dx \exp \left( -\frac{(x - \frac{1}{2}\Delta)^2}{\sigma^2} + \frac{1}{4}\Delta^2 \right)$$  \hspace{1cm} (4.13)$$

$$I(\Delta) \propto \exp \left( \left( \frac{\Delta}{2\sigma} \right)^2 \right)$$  \hspace{1cm} (4.14)$$

Between a pair of Gaussian distributions, the overlap integral between them decays in a Gaussian character in accordance with the square of their distance in units of $2\sigma$. Overlap between emission and absorption drives the energy transfer process, so understanding the energetic relationships between chromophores is paramount to understanding energy transfer.
CHAPTER 4. FMO’s HIGH FREQUENCY FLUCTUATIONS ARE PROTECTED

However, it is important to recognize that each chromophore is fairly unique. It behooves us to consider each chromophore as its own entity with unique inferences to be made about it, as seen in Table 4.1. We therefore focus on BCL373 for two reasons: (a) as the exit site of the complex and it is quite important (b) it is well protected from solvent and HIS-coordinated, suggesting it should be relatively well-behaved compared to other residues.

We can also consider the correlations between chromophores, both on different monomers and on the same monomer. This forms a 24x24 matrix of correlation values. This correlation can be written as

\[ C_{\kappa \lambda} = \frac{\left\langle G_\kappa G_\lambda \right\rangle}{\sqrt{\left\langle G_\kappa^2 \right\rangle \left\langle G_\lambda^2 \right\rangle}} \]  

(4.15)

Low frequency data in figure 4.7 contains such long (relative to the molecular dynamics sampling times) correlation times that any findings about correlation are dubious and should be subject to skepticism. However, inter-monomer correlations on BCL 376 seem compelling. The empirical joint distribution also in figure 4.7 gives a sense of why this result is dubious: the correlations cause the vast majority of the joint space to be poorly sampled.

In the high frequency data, contrary to what was found by Schulten et al. [79], there seem to be non-negligible (about .15) correlations present in the high-frequency signals of sites 3 and 4 on the same monomer as seen in figure 4.7. The inconsistency to their findings might have a simple and direct source in the differences between the analyses; for example, this analysis only considers the environmental contribution and removes the low frequency noise, so it has mitigated the impact of two major noise sources from their analysis that may have concealed the correlation in the analysis of Schulten et al.

Monomer as a factor To be confident about data integrity, we require plausible intermonomer symmetry. Each monomer in the simulation is perfectly identical up to minor differences in boundary conditions, which could themselves be made to share the symmetry of the protein in many cases. It is not theoretically forbidden that the preferred configurations are systematically monomer-asymmetric due to interactions, and that spontaneous exchanges of monomers roles occur on time scales longer than observed in the simulation. However, the most likely explanation for asymmetry in a trimer complex is that the monomers have each minimized independently to different regions of the acceptable phase space, and that they do not explore across these states within observed times. We did not enforce instantaneous symmetry across monomers with biasing, but the ability to monitor its presence or absence helps inform the extent to which we trust the absolute equilibration of our samples.

Two tools utilized often in statistics to compare distributions are the q-q plot and the empirical cumulative distribution function (ECDF). The q-q plot is defined as a transformation of two equally-sized sets of data \( G_{J_1} \) and \( G_{J_2} \) into data pairs \((qqx_i, qyy_i)\) as

\[ (qqx_i, qyy_i) = (\text{sort}(G_{J_1})_i, \text{sort}(G_{J_2})_i) \],  

(4.16)
Figure 4.7: Inter-chromophore and inter-monomer correlations for low frequency signal (top) and for high frequency signal (bottom). Left panes show covariance matrix while the right panes show hexbin 2d histogram for the largest absolute correlation elements in the matrix; that is, for BCL376 monomer 1 against BCL376 monomer 3 and for BCL373 monomer 1 against BCL374 monomer 2 for the low and high frequency, respectively. The erratic and incomplete shape of the low frequency histogram shows clearly the poor sampling, while the clean Gaussian shape of the high frequency histogram shows thorough and high-quality sampling.
Figure 4.8: Inter-monomer q-q plots and empirical CDF comparisons for BCL373. From the ECDF, we can test for exact identicality of distributions. Because these monomers are identical, their p-scores should be uniformly drawn from the $U(0, 1)$ distribution. Distributions are sparsified by a factor of 10 to eliminate as much correlation as possible to improve p-scores. Left: Before low-frequency filtering, p-scores are significant to a massive degree. q-q plots show the qualitative difference between these distributions. Right: After mean removal, only the tails vary slightly, as seen in the q-q plot. The values of .4 and .2 for p-scores are quite acceptable scores that do not cause us to reject the hypothesis that these distributions are identical.

Figure 4.9: Histograms of the gap for BCL373 before low frequency filtering (left) and after mean removal (right). In both situations, the distributions are quite Gaussian.
where sort() is the operation that sorts G in numerical order, thereby aligning all quantiles of the two data sets. For identically distributed data, the q-q plot forms a straight line of slope 1, which is to say, each quantile \( q \) from \( G_{J_1} \) is identical to the quantile \( q \) from \( G_{J_2} \). The ECDF is computed again with a sorted data set, but this time, assigning index \( i \) to a scaled index between 0 and 1 which is then plotted on the y axis. The ECDF is preferred to histograms for comparing distributions as it does not require any bin parameter and has a well understood null distribution between two datasets of IID values from identical distributions.

The sample trajectories show significant monomers asymmetry. Later in the chapter, we will demonstrate that simply removing the mean from these trajectories will restore symmetry between monomers. The simplicity of this fix suggests that the factors that influence fast fluctuations do not vary between monomers despite the factors that influence slow fluctuations varying significantly. This collapse to identical distributions can be seen in figures 4.8 and 4.9.

### 4.2.3 All major heterogeneity is removed by eliminating low-frequency components

The data set shows a lot of heterogeneity in time, and this feature is particularly damaging to the rest of the analyses that are possible. The simplest thing we can do is to remove the rolling mean from each segment in time. This way, we can remove the effect of large drifts in the mean. This does a good job at collapsing histograms that were previously quite separated. However, is it a principled approach?

Filtering was discussed in section 3.3.1, but we will address it again here in the context of this particular signal. In the context of low-frequency removal for the CDC, it is sensible to consider the low-frequency part of the signal to be the major source of noise. Just as we can remove the high-frequency noise and recover the low frequency signal, we can subtract the low frequency signal from the composite signal and retrieve just the high-frequency component. This filtering can be done in any number of ways, but for this work, an eighth order type 2 Chebyshev filter was used with a frequency cutoff of 1/100th the Nyquist frequency on a signal with a sampling rate of 10 ps. This filter has high-quality band-limiting behavior, so it is preferred to an otherwise theoretically simpler “filter” like a square window or triangle window convolution. The resulting signal for this filter had about half the correlation time as a square-window filter with the same “cutoff” frequency. (Note that “filter” and “cutoff” are quoted because the square window is not a proper lowpass filter [80].) This process is a linear filter.

One major upside of linear filter theory is, unsurprisingly, the linearity. Just as the CDC method’s linearity allowed us to split the contributions to the gap shift by protein residue, so can we apply filter to the constituent environmental entities rigorously and arrive at the correct compound signal. Thus, our linear filter allows us to not only filter the summed signal, but each of the individual constituent signals as well, meaning we can greatly increase
Figure 4.10: Quantile scatterplot matrix for BCL373. Left: before low-frequency removal we see lots of correlation between the 5th, 50th, and 95th percentile. Right: after low-frequency removal, nearly all of the correlation between these quantiles falls out, suggesting that the bulk of that correlation was simply from a consistent distribution with a continuously varying mean.

our ability to detect the origin of low frequency changes in the gap.

For analysis of the raw CDC data, we can consider filters through two different lenses. Filters are a way to lowpass and identify the equilibration status of the CDC trajectory. The filter also can clean up the substantial aliasing noise present in the signal. Filters also allow us to treat the low-frequency signal as the noise, and to isolate the high frequency fluctuations to assess the validity of the spectral density on those time scales.

It is important to note that filtering and thermodynamics do not usually pair well, as filtered samples will not respect the Boltzmann distribution. Despite this, filtering allows us to parse out dynamical features of an otherwise complex trajectory. We need slightly more than this if we want to justify the use of the filtered high-frequency CDC signal as a basis for using a spectral density. To this end, we sought to identify correlations in the high-frequency distribution as a function of time by computing the correlation present in the quantiles. If a distribution is truly just shifting in its mean, arbitrary quantiles will be perfectly correlated to each other, up to the statistical noise introduced by finite sample size. We compute the
correlation between quantiles as

$$C(q_L, q_H) = \langle \delta q_L \delta q_H \rangle,$$

and we see the approximate value of this computation by inspecting a scatterplot of $q_L$ against $q_H$, as in figure 4.10.

We see a high correlation between quantiles in the original signal, and a rather low correlation in the signal after filtering out the low-frequencies. In the original signal, the correlation between the 5th and 95th percentiles are around .9. After removing the low frequency signal, the correlation almost entirely disappears, leaving only the statistical noise which is mostly uncorrelated between the quantiles. This finding is robust and independent of time, which can be verified by inspection in figure 4.11.

While we have just shown that removing the low-frequency component of the CDC signal is statistically justified, we can also argue that this separation of frequencies has a well founded physical basis.
4.2.4 Consistency of high frequency covariance across monomers lends evidence to a physical basis for separating time scales

We saw in figures 4.8 and 4.9 that when we filter the low frequency components of the signal out of the CDC, our chromophore environments with quite different means collapse to statistically indistinguishable distributions. These identical high-frequency environments leads us to ask, how are these environments with significantly different means able to have the same variances? When we examine the spatial distribution of contributions, the picture begins to make more sense.

First, we can examine the distribution of variances being contributed from individual residues — if a residue contributes little variance, no amount of correlation can cause it to contribute a lot of variance in the collective picture. Said differently, correlation can only enhance or decrease the collective effect of residues that already have significant variance on their own.

The residues in the complex contribute to the variance in a way such that the aggregation of the small contributions is comparable to the aggregation of the large contributions. In this sense, we understand the process to be long ranged, and expect the long-ranged contributions to look like independent Gaussian noise in the high frequency band and like screening (in that we expect to see significant anti-correlations between the solvent and the rest of the tracked residues) in the low frequency band. Local structure, which is comparable to global contributions, has potential to contribute complicated fluctuations.

We can expand our analysis of covariance to include correlations as a way of understanding the relationships in the protein structure. It is important to identify first and foremost that there are substantive anti-correlations to solvent in the residue modes which act to screen contributions of the solvent (or visa versa), but these anti-correlations do not tell us anything about the specific protein structure. We instead want to investigate first and foremost large and specific correlations.

We can simultaneously observe these correlation features and observe the heterogeneity across monomers by selecting a set of residues which is the union of the top five variance contributions for each monomer, as in Figure 4.12. Correlations are much more heterogeneous for the low-frequency modes, and the top contributors to variance are different across the monomers. On the other hand, correlations are homogeneous and consistent across monomers for the high frequency modes. These correlations are a deeper identification of physically-based structure that leads to the consistency of high-frequency modes across different monomers.

4.2.5 Summary: Gap trajectories have multifaceted heterogeneity of physical origin

Getting a clean analysis of a data source requires clean data. Unfortunately, the energy gaps sampled from the FMO complex have a variety of heterogeneities across their varied factors.
These heterogeneities can be almost entirely removed by filtering out low-frequency components of the signal. While there is nothing guaranteeing that this analysis will be informative, it turns out that the high-frequency parts of the signal are consistent across time and monomer. This is excellent news for spectral density theories of open quantum dynamics, which are only able to couple to these high frequency components.

The low frequency components of the signal are themselves quite interesting, though. Their separability informs the use of static disorder/dynamic disorder descriptions of variability in spectroscopic line shape theory. Quite interestingly, the static disorder is anything but static — the fluctuations in the mean gap are just as often slow drifts as they are abrupt jumps. This analysis corroborates the idea that spectroscopic analyses are pragmatic, but it calls into question the cause and structure of static disorder.
4.3 High-frequency gap distributions suggest some complexity beyond the direct spectral density

With the low frequency fluctuations removed, we can directly analyze the usage of a spectral density model. We find a high degree of absolute normality in the short-time marginal density of states for window durations from 1 ns to 50 ns, which corroborates the validity of the spectral density for the FMO energy transfer process whose ejection time constant is around 4 ns. We compute the spectral density for the CDC process and compare it to previous results from the literature. We also examine the limitations of the spectral density in light of previously published results and propose a model which can begin to parametrize some of the unusual results found in full quantum mechanical simulations.

4.3.1 Spectral density in PPC bath modeling

It is difficult to take a reduced system, for which the effects of temperature are included in only an average way, and to introduce parameters that incrementally couple to the original dynamics. Fortunately, the spectral density is exactly this minimal parameter set. We saw in section 2.4.3 that for a redfield level of theory, the spectral density is the only parameter from the gap dynamics that couples to the energy transfer dynamics. A similar theory, the Heirarchy Equations of Motion [3], is also constructed at a level of theory that couples exclusively to the spectral density. The justification for this is somewhat circular; if we assume that the fluctuations are quite Gaussian, the only parametrization for the bath is the spectral density. It is necessary to examine the fluctuating signal itself to see if the justification for a spectral density holds.

To use a minimal model, it is necessary to scrutinize the assumptions of that model through validation. For a spectral density, there are two major assumptions: (1) the temporal fourier coefficients are Gaussian distributed with zero mean, and (2) the temporal fourier coefficients are independent of one another. The spectral density at frequency $\omega$ is then proportional to the variance of the fourier coefficient distribution for fourier coefficient $a_\omega$. In real space, the approximation central to the spectral density is that the joint distribution between any pairs of random variables is jointly Gaussian distributed. Using the fourier coefficients directly is challenging because of aliasing pollution; correct calculation of the fourier coefficients for any particular signal requires extremely high resolution sampling.

In this section, we explore the theory and estimation of the spectral density, then determine the extent to which FMO fits the bill for spectral density estimation.

4.3.2 Spectral density statistical overview

The physical application of the spectral density can be found in chapter 2.4.5. What we focus on here is the statistical definition of a spectral density, and its interpretation in this
CHAPTER 4. FMO’S HIGH FREQUENCY FLUCTUATIONS ARE PROTECTED

The spectral density is defined as

$$S(\omega) \equiv \langle |x(\omega)|^2 \rangle$$

(4.18)

for the extensivity-stabilized choice of Fourier transformation convention

$$x(\omega) \equiv \frac{1}{\sqrt{T}} \int_0^T x(t) e^{-i\omega t} dt.$$  

(4.19)

We can sketch out how this function approaches a single value for some stochastic process. Assume our process $x(t)$ has a finite correlation function. We can select a number $N(\omega)$ for a number of periods of oscillation $\omega$ over which two adjacent samples of $N$ have very little residual correlation. Define any one of these periods as

$$\hat{x}_{N,\omega}(\omega; t_0) = \int_{t_0}^{t_0+N2\pi/\omega} x(t) e^{-i\omega t} dt.$$  

(4.20)

We can write a probability distribution for the $P(\hat{x}_{N,\omega}(\omega))$, and because we can make the correlation between samples arbitrarily small by increasing $N$, we can eventually write the distribution of $x(\omega)$ as

$$x(\omega) = \frac{1}{\sqrt{T}} \sum_{n=0}^{T/(N2\pi/\omega)} \hat{x}_{N,\omega}(\omega; nN2\pi/\omega)$$  

(4.21)

$$\approx \frac{1}{\sqrt{T}} \sum_{n=0}^{T/(N2\pi/\omega)} \hat{x}_{N,\omega}(\omega),$$  

(4.22)

where equation 4.22 refers to the fully independent distribution. Since now we have expressed $x(\omega)$ as proportional to a sum of independent random variables, we know by the central limit theorem that the variance of such a quantity will grow extensively with the number of samples. Thus, the variance of $x(\omega)$ asymptotes to a single value, $S(\omega)$, thanks to the $\frac{1}{\sqrt{T}}$ convention.

4.3.3 FMO Spectral Density

We can apply this spectral density analysis to a fluctuating energy gap time series somewhat directly, at least for the purpose of understanding the amount of fluctuation that comes from each time scale. The burden then lies on the modeler to validate the use of a spectral density data model. From equilibrium statistics, we can compute the fluctuation-dissipation level of parametrization for a bath, aka a spectral density bath [11], as follows:

$$C(t) = \langle \delta G(t)\delta G(0) \rangle$$  

(4.23)

$$J(\omega) = \tanh(\beta\hbar\omega/2) \int_0^\infty dt J(\omega) \cos(\omega t)$$  

(4.24)
This spectral density encapsulates the dynamics of a Gaussian process completely. A Gaussian stochastic process is one in which the joint distribution of any pair of variables – measured at the same or at different times – is Gaussian distributed. One canonical example of a Gaussian process is White noise: it has correlation which decays instantaneously, and thus, the distribution of each point in time is an independent and identically distributed (IID) Gaussian distribution.

While this is incredibly strict, in practice, many of these joint distributions will not be particularly significant, particularly if each constituent distribution has low variance. It is easy to assume that the dynamics constitute a Gaussian process and then parametrize the correlation function and the spectral density, but it is a different challenge to attempt to verify this. We cannot just search the high-dimensional distribution for a pair of variables that are particularly non-Gaussian for the sake of throwing the baby out with the bath water. We must first ask “In what ways would a deviation from a Gaussian process affect our gap dynamics?”

The energy transfer dynamics are a highly nonlinear function of the gap. We can also take from our understanding of order parameters that when quantities are not Gaussian distributed into their tails, linear response dynamics may be quite broken. We thus want to explore a variety of perspectives through which we can rigorously assess both the practical efficacy and the core validity of this Gaussian stochastic process approximation.

4.3.4 Spectral density estimation

The most direct way to estimate the spectral density of a generic time series is through the method of correlograms. This method is robust and clean for computing spectral densities largely because the correlogram is easily represented as a time average with good properties of convergence, and can be fit with analytic functions to generate a smooth and continuous spectral density.

We follow the correlogram method and fitting procedure previously established [11]. The core premise of the correlogram method is that the spectral density is proportional to the Fourier transform of the autocorrelation function, as proven by the Wiener-Khinchin theorem. One extremely powerful aspect of this method is its robustness against aliasing. Though it will not eliminate aliasing numerically, the auto correlation function estimates are unbiased with respect to undersampling, and therefore if we fit the autocorrelation function with a sufficiently good fitting function to describe the underlying data process, we can get high-quality stable estimates of the spectral density.

To compute the spectral density, samples were taken with a period of 2fs for a duration of 1ns. This 1ns trajectory was split into 10 100ps trajectories, whose autocovariance functions were zeroed over the range of .25ps-1ps, then averaged to provide a more robust estimate of the high-frequency autocovariance. This function was then fit on a pair of exponentials, and the residual of that fit was fit to a set of 6 damped oscillators. The resulting estimated parameters were plugged into their functional representation in Fourier space and summed to provide an estimate of the spectral density [11].
Figure 4.13: Estimate of the spectral density for a short, high-resolution time series of the CDC method. Due to the trajectory’s heterogeneity, an average of sub-trajectories was necessary to reach a good estimate of the autocovariance function. Top left: quantile trajectory for the 1st, 25th, 75th, and 99th quantiles. Top right: Autocorrelation function for the full trajectory (thick red) shows greatly arrested decay after a brief period of rapid decay, while the ensemble of shorter trajectories (transluscent red) shows much faster decay in the majority of samples. Bottom left: autocorrelations (transluscent red) were "zeroed" at 1ps, and the residual was fit to a single exponential function and a collection of damped oscillators. Bottom right: parametrized values of exponential functions and damped oscillators allow for analytic evaluation of the spectral density.

The results for the spectral density analysis are as follows. When compared to the estimates made by Olbrich et al [11], the spectral densities here are are about 6 orders of magnitude smaller (This huge discrepancy in variances of CDC and ZINDO will be addressed in chapter 4.3.6). However, when compared to the experimental spectral densities of [7], the general shape of both the parametric and nonparametric functions show much more agreement. In particular, the parametric function shows a strong peak at about 50 cm$^{-1}$ with a fwhm of about 100 cm$^{-1}$. The non-parametric function shows a broad peakedness with similar characteristics as the experimental curve, e.g. peaks at 200 and 300 cm$^{-1}$, and an approximate decay to zero at 400 cm$^{-1}$. Each of these methods of spectral density estimation has different strengths – the parametric method requires difficult optimization routines, so it is not certain to find a good global minimum. It is also unconstrained, so it can permit unphysical negative values for the spectral density, as observed. However, it does capture the dominant spectral features quite well. On the other hand, the nonparametric
method captures the texture of the spectral density, but is quite numerically unstable. In fact, the spectral density shown has been modified to subtract out a linear trend from the high frequency tail because that tail persists up to the observed nyquist frequency, suggesting it is a symptom of some numerical instability.

4.3.5 Density of States, spectra, and response

The main tool of statistical mechanics is the probability distribution, as it is isomorphic to the thermodynamically critical free energy. The energy gap for each chromophore has a probability distribution, and therefore a free energy. In the photosynthetic literature, probability distributions of the gap as represented in figure 4.6 are referred to as the density of states. This terminology is used to distinguish the gap probability distribution from the absorption and emission curves, as discussed in section 2.4.2. With the density of states, we can perform a tremendous range of optical calculations.

Our primary interest in these density of states is to understand the behavior of linear response in these systems. For a perfectly quadratic gap-shift free energy, we can compute the reorganization energy from the free energy using linear response theory. This brings us back to a question we asked earlier: how linear must something be to treat it with linear response? In general, thanks to the binary nature of the excitation variable, we need about two or three sigmas of good fit to the Gaussian distribution on the red side of the distribution. Given that the shift always pushes towards the red, the blue tail is much less important.

We can test the distribution for Gaussian behavior at short time scales by splitting the existing trajectory to collect multiple sample trajectories. On these trajectories, we compute a standardized empirical cumulative density function (ECDF), scatterplot the sample means and variances, and test for skew and kurtosis, as is standard for tests for normality. The ECDF should look like the normal ECDF, and any deviations from that indicate skew or kurtosis; furthermore, the amount of spread around the ideal curve should decrease as we lengthen our trajectory and decrease our number of samples. The mean and variance should change randomly, but independently. The skew and kurtosis test p-values should be uniformly distributed.

The results of this analysis are presented in figure 4.14. The availability of many instances of trajectories of a particular window size makes it much easier to assess the quality of the model for the trajectory at large, as we can collect an ensemble of p-scores and can compare that p-score distribution to the null. The analysis results show that normality is a very strong assumption at the fastest time scales, and it weakens for longer trajectories. For 20 ns trajectories, there is a huge probability mass for skew trajectories with very small p-scores, indicating that there is indeed some skew to be expected at these time scales. On the other hand, over each of the time scales time scales, the ECDFs seem to be reasonablly gaussian, and the mean and standard deviation seem to carry very little correlation.

Assessing the full density of states is difficult given our trajectories’ prevalence of dynamically heterogeneous features. Without averaging over a sufficient number of heterogeneous states, the averages will suffer from correlation of samples, and the uncertainty in the es-
Figure 4.14: Normality tests for various window sizes of T=1 ns, T=5 ns and T=20 ns on BCL373 Monomer 2. Skewness and kurtosis p-values are expected to be uniform for a Gaussian distributed process. This seems to largely be the case, though the skewness test for the 20 ns window indicates the emergence of detectable skew in a number of windows.

The estimate will be astronomical despite the appearance of high certainty through many data points. It is also important to appreciate the difference between a density of states which is sufficiently Gaussian and a low frequency signal which is sufficiently Gaussian. While it is common in spectroscopic theory to broaden a high-frequency distribution with a Gaussian low-frequency noise, this underlying noise need not and does not seem to be Gaussian distributed. For the purposes of absorption and emission spectra, this detail can be sufficiently captured with an empirical inhomogeneous broadening function, but for correctly identifying the distribution of initial states for a calculation like the HEOM or the Redfield model, it would be useful to have the true joint distribution on the low-frequency gap shifts for the chromophores.

We have shown that we can compute a spectral density for the high frequency signal, and these high frequency signals appear nicely Gaussian. How does this CDC spectral density compare to previous analyses?
variable & observed std. dev. (σ) & min req’d & uncorr. est. & max allowed \\
\hline
\text{G}^{\text{CHARMM}}_{ZINDO, PC} & 258 \pm 10 \text{cm}^{-1} & \text{N/A} & \text{N/A} & \text{N/A} \\
\text{G}^{\text{CHARMM}}_{\text{TDDFT, PC}} & 445 \pm 10 \text{cm}^{-1} & \text{N/A} & \text{N/A} & \text{N/A} \\
\text{G}^{\text{CHARMM}}_{\text{TDDFT, vac}} & 291 \pm 10 \text{cm}^{-1} & \text{N/A} & \text{N/A} & \text{N/A} \\
\text{g}^{\text{CDC}} & 28.5 \pm 4 \text{cm}^{-1} & 151 \pm 15 \text{cm}^{-1} & 336 \pm 15 \text{cm}^{-1} & 736 \pm 15 \text{cm}^{-1} \\
g^{\text{CDC}, f=1} & 85.5 \pm 12 \text{cm}^{-1} & 151 \pm 15 \text{cm}^{-1} & 336 \pm 15 \text{cm}^{-1} & 736 \pm 15 \text{cm}^{-1} \\
\hline

Table 4.2: Table of values from quantum mechanical calculations [12] [16] benchmarked against CDC contributions from our simulations. CDC with screening falls about an order of magnitude short in scale to the contribution needed to explain the difference between vacuum and point-charge embedded DFT calculations, and CDC without screening still falls about a factor of two short of the smallest value possible to explain the difference in TDDFT variance estimates.

4.3.6 Overview of published ZINDO calculations

In the majority of the literature for FMO energy gap dynamics, the quantum chemical method of choice is ZINDO. ZINDO allows for cheap excited state calculations that were developed specifically for use in molecules with pyrrole motifs. The method is time-tested and used by many authors, so it serves as a good baseline.

Of course, the CDC method is only computing the gap’s environmental contribution, so it should not be compared to ZINDO directly. Instead, we can compute ZINDO with and without point charges in the environment. If the linear perturbative approximations for CDC are sufficient, then this pair of calculations will yield a difference equal to the CDC value. This is unreasonable to expect, as the ZINDO method itself has uncertainty in its calculation, a uncertainty only compounded by differencing two configurations, one with and one without point charges.

By inspection of table 4.2, we can identify that the high-frequency spread for CDC values is far too low to account for the effect of the embedded point charges in the ZINDO calculation. One obvious reason for why this is too low is to blame the ZINDO calculations. If each of the ZINDO measurements has an error associated with it, these errors convolve into the distribution and introduce broadening. Alternatively, the CDC method could be underestimating the gap. This is entirely feasible for fluctuating trajectories, as CDC was parametrized and optimized for use in energy minimized configurations. Fluctuations that affect the Mg$^{+}$ coordination, for example, will almost certainly be underestimated with a coulomb procedure instead of a quantum procedure.

If this is the case, then what do the CDC values actually mean? The CDC is certainly a valid collective variable, with a weighting function proportional to CDC-charge weighted inverse distance from the chromophore. In this sense, what the CDC tells us is somewhat generic about coulomb coupling to a chromophore-sized molecule with a permanent dipole moment embedded in the protein; the method just takes care of making a physically reason-
able set of parameters to potentially describe a real chromophore better than something ad hoc like a point dipole. The CDC gives us a sense of just how far the environment is pushing the gap, only without properly nonlinear effects from higher order perturbative terms.

4.3.7 Simple model for Density of States skew

After analyzing the CDC results for the time being, we can return to the densities of states computed from fully quantum calculations. In particular, if we will consider the facts that the cdc (1) substantially underestimates the variance contributed from the environment and (2) is much more Gaussian than the fully quantum mechanical gap densities of state, we can infer that the intramolecular contribution to the density of states itself has significant skew. We are thus brought to propose a theory that accounts for the density of states skew from an intramolecular source.

Our understanding of electronic energy transfer in photosynthesis is strongly informed by the use of spectral density models. These models have shown themselves to be predictive in estimating transition times for FRET, where the primary trait being monitored is the frequency at which a rare fluctuation between donor and acceptor occurs. The correctness of spectral density models for highly non-linear problems, however, is less easily tested. What is referred to here as a spectral density model has a particular form as such:

\[ H = H_e + \sum_{i,k} g_{ik} x_k |i \rangle \langle i | + \sum_k \frac{1}{2} \omega_k x_k^2. \]

This linear form comes directly from two assumptions:

1. The ground and excited states are well described by phonons (quadratic PES)
2. The ground and excited states have the same curvature for each phonon coordinate
3. The minimum value for some phonon coordinates in the excited state are shifted

If we compute \( G(x) \), the gap as a function of a phonon mode position \( x \):

\[ G(x) = U_e(x) - U_g(x) \]
\[ G(x) = \frac{1}{2} \omega (x - \Delta)^2 - \frac{1}{2} \omega x^2 \]
\[ G(x) = -\Delta \omega x + \text{const} \]

What we can get from this conclusion is that if \( p(x) \) is symmetric, as it will be in the Gaussian case, the absorption spectrum must also be Gaussian. Unfortunately for this level of theory, the vast majority of absorption spectra contain significant skew, with the absorption having a fat tail in the high-energy direction, and the emission having a fat tail in the low-energy direction.

For an arbitrary dimensional system of phonons, the result is different, which allows for a wider variety of spectral shapes. However, all of these shapes must still be symmetric about
the mean. This is easily observed by looking at the form of $G$.

$$G(x) = -\sum_i \omega_i \Delta_i x_i$$

Seeking a simple theory that describes skew, we can introduce a small quartic term into our one-dimensional potential, with the new sum quadratic + quartic potential denoted by potential energy $U^*$

$$G(x) = U^*_e(x) + U^*_g(x)$$
$$G(x) = \gamma (x - \Delta)^4 + \omega (x - \Delta)^2 - \gamma x^4 - \gamma x^2$$
$$G(x) = -(4\gamma \Delta^3 + \omega \Delta)x + (6\gamma \Delta^2)x^2 - (4\gamma \Delta)x^3$$

This function fails to retain the oddness that forced an even probability distribution to generate an even distribution for $G$ thanks to the $x^2$ term. Note that this function has no stationary points of this function due to the form of the coefficients. To interpret this function, note that the $x^2$ term carries a value greater than or equal to zero at all points, meaning that the negative range of $G$ is flattened and the positive range of $G$ is sharpened. This will push skew in the distribution of $G$ towards a positive value during absorption, which is exactly what we observe. Furthermore, it will push the value of $G$ symmetrically negative during emission due to the symmetric form of the potential difference. This shows that we do not need to implement a complicated PES in order to pick up skew in the absorption spectrum.

The two questions that remain are then these:

1. Can we write an expression that will allow us to quickly compute the energy gap as a function of phonon coordinates through parametrization instead of ab initio/empirical quantum calculations?

2. Can we get anywhere in the electronic energy transfer problem if we include this nonlinearity? Can we understand why this nonlinearity does/does not impact EET?

These remain interesting research directions.

### 4.3.8 Summary of gap overview

The raw CDC data shows complex behavior on a number of time scales. Its most notable feature, though, is a reduction into fast and slow degrees of freedom. This separation is well justified by correlations in quantiles; despite movement of the distribution’s center, the spread remains largely consistent.

With the bare features of our gap analyzed, we can proceed to computing assessing the sensibility of the spectral density.
4.4 Fast modes are independent and proximal while slow modes are collective and nonlocal

Where the structural analysis of chapter 4.2.4 lends evidence to a physical origin for the protection of high-frequency modes, it falls short in explaining the mechanism of protection in the presence of collective dynamical protein activity. In this section, we use a Principal Component Analysis on the residue gap contributions to capture collectivity by breaking fluctuations into modes that are characterized by their statistical independence. From this, we conclude that the slow modes are formed by collective activity in the protein, whereas the fast modes are much smaller systems of residues.

For systems of jointly Gaussian random variables, there is an indispensable concept of statistical normal modes. Statistical normal modes are the collective variables associated with the projections of the original modes onto the eigenvectors of the covariance matrix; they are the collective random variables of a jointly Gaussian system that are statistically independent of one another.

Statistical normal modes can be thought of in terms of their relationship to a dynamical system’s eigenmodes. A harmonic dynamical system will have an energetic coupling matrix, and the diagonalization of that energetic coupling matrix will define the collective variables (eigenmodes) that oscillate coherently. When a linear perturbation is applied to a harmonic dynamical system, the simplest way to look at the result is to project the perturbation onto the eigenmodes, as each of these independent modes will respond entirely on their own to the part of the perturbation which projects onto them. Similarly, for statistical normal modes, a linear perturbation will lead to a response in an exactly analogous way, in proportion to the eigenvalues of the different statistical normal modes.

Even without normality, the belief that contributions to a collective phenomena can be broken into independent constituents justifies a mode analysis. There are many mode analyses that attempt to separate independent factors, even in incredibly non-linear systems. The math of modes may be best posed in a linear framework, but the concept of identifying independent components is not limited to linear systems.

4.4.1 Principal component analysis defined

Computing a set of statistical modes is done through Principal Component Analysis (PCA). We will drop the chromophore index $\kappa$ for simplicity of notation, noting that we will not consider correlations in gap contributions across both chromophore and residue indices simultaneously, though it is perfectly possible to compute such a covariance matrix. In PCA, linear combinations of the input variables are formed to create a new set of variables that are all mutually independent. The algorithm to compute these linear combinations uses the eigenvectors from the covariance matrix as weights, thereby creating a basis rotation in which the covariance matrix is diagonal.
We need to thus formulate our PCA in terms of the variables of interest in our simulation. We have our collection of energy gaps, defined by our CDC method with inter-atomic coupling between atoms $i, j$ of $K(i, j)$ for atom groups $\{r_j\}$ of residue $j$ and atom groups $\{c_N\}$ of chromophore $N$ (with the $N$ index on $g_t$ suppressed for conciseness) as

$$\delta g_{tI} = g_{tI} - \langle g_{tI} \rangle$$
$$C_{IJ} = \langle \delta g_{tI} - \delta g_{tJ} \rangle$$

The PCA is then written as the eigenvalue problem:

$$w_i e_{iI} = \sum_j C_{IJ} e_{iJ},$$

that is, we solve the eigenvector problem for the covariance matrix $C$, giving us eigenvectors $u_{iI}$ and eigenvalues $w_i$ on which to project our original degrees of freedom. We can represent these eigenvectors as a change of basis matrix as $U_{iI} = u_{iI}$, where the columns are eigenvectors. We can also associate a value $e_i$ with each eigenvector as a function of time (called an eigenmode) by rotating the original values according to

$$g_J(t) = [g_1(t), g_2(t), \ldots]$$
$$e_i(t) = \sum_j U_{ij}^T g_J(t) g_{tI} = \sum_j U_{ij} e_{j}(t).$$

Such a transformation will preserve the sum of squares between $e_i$ and $g_J$, also called the $l^2$ norm. These eigenmodes are uncorrelated, though not necessarily completely independent.

If we wish instead to preserve the $l^1$ norm of our data set, as we might if we want to retain the meaning of a principal component’s value as the instantaneous contribution of that mode to the fluctuating gap, we must modify the modes. Any linear scaling of the eigenmodes will linearly scale pairwise correlations, and thus, our scaled eigenmodes will remain a collection of uncorrelated modes. As it turns out, we can compute the contribution of an eigenmode to the gap independent of all of the other eigenmodes because of the linear form and this correction takes the form of a linear scaling. These $l^1$ scaled eigenmodes will be referred to as the eigengaps $\bar{e}_i$. We prove the independence of eigengaps in what follows.

Let us write the original data vectors $g_I$ and their projections $e_i$ onto the eigenvectors through transformation $U$, whose columns are the eigenvectors.

We can then compute the total gap,

$$G(t) = \sum_I g_I(t)$$
$$G(t) = \sum_{i,J} U_{iJ} e_{iJ}(t)$$
CHAPTER 4. FMO’S HIGH FREQUENCY FLUCTUATIONS ARE PROTECTED

From this equation, it is clear that we can evaluate the sum over \( i \) before we acquire any data vectors \( e_j \). We can name this coefficient,

\[
M_j = \sum_i U_{ij}
\]

\[
G(t) = \sum_j M_j e_j(t).
\]

This \( M_j \) coefficient describes the impact of the eigenvector-represented data vector \( e \). It is easy to ask what properties these new scaled data vectors, \( \bar{e}_i = M_i e_i(t) \), have. They have in some sense reconstituted the \( l^1 \) norm into the data by scaling each eigenvector by the sum of its components. We can compute the correlation between scaled data vectors \( \bar{e} \) as

\[
\langle \bar{e}_i \bar{e}_j \rangle = \langle M_i e_i M_j e_j \rangle = M_i M_j \langle e_i e_j \rangle.
\]

Thus, if \( e_i \) are uncorrelated, \( \bar{e}_i \) must also be uncorrelated to one another. These new eigengaps have eigenvectors \( \bar{w}_i = M^2_i \bar{w}_i \). These eigengaps are a unique construction of the eigenvectors, and therefore, are a unique construction of the original data.

4.4.2 Practical challenges with PCA

Great care is needed when analyzing PCA results on data that is not obviously ergodically sampled.

Like all analysis procedures, PCA also presumes that samples are independent of one another. In time series analysis, this is particularly problematic, as the major feature defining time series is the continuity and correlation between adjacent samples. If we inspect figure 4.15, we can see this going terribly awry for a principal component analysis in dihedral angle space on a fairly short trajectory from 4BCL (i.e. a trajectory that was not considered for our main analyses). Eigenmodes are formed by a rotation of the original basis,

\[
e_i(t) = \sum_j U^T_{ij} f_j(t),
\]

and so we can consider the disaggregation of any particular \( e_i(t) \), into scaled eigenvector components \( F_{ij}(t) \) as

\[
e_1(t) = \sum_j F_{ij}(t)
\]

\[
F_{ij}(t) = U^T_{ij} f_j(t).
\]
Figure 4.15: Example of a PCA gone wrong from an unused analysis of dihedral angles. A single eigenmode is represented here, and is formed from the sum of all of these features. The shown features are the individual trajectories for which the covariance matrix has been constructed. Alignment of trajectories that contain a single transition will create a large principal mode with no statistical meaning, and elimination of all modes with drift will lead to a severely underpowered analysis.

Values of $j$ between 1 and 27 are shown in figure 4.15. We can see clearly why these modes add together to create a mode with very high variance; these transitions need not even be correlated as long as they happen at similar times. The drift in the modes causes the resulting modes to be very poor approximations of the eigenmodes at long times.

### 4.4.3 Results

The results of the PCA analysis are presented for chromophore 373. In just this chromophore, there are already many exceptional complexities justify independent consideration. To analyze each of the eight chromophores in FMO is an endeavor that is more informative about FMO than about pigment protein complexes as a class of entities. We thus continue to focus exclusively on this site.

The the inverse participation ratio (IPR) gives us a sense of how collective a particular mode is. For a normalized eigenvector, the sum of squares is equal to 1, and therefore, the sum of fourth powers is always less than or equal to 1. This sum of fourth powers
Figure 4.16: Basic summary of the largest modes of the principal component analysis for BCL373 for the high-frequency signal. Top row: pie charts for variance of each principal component, with the mode’s inverse participation ratio labelled on top. Middle row: residue against presence of residue within the top six principal components; As the sum of the squares of the eigenvector matrix sum to 1 in both the rowwise and the columnwise directions, this "presence" term is easy to compute directly from the eigenvectors. Bottom row: trajectories of the largest PC1 mode for each of the monomers.
is identically 1 when only a single element of the vector is nonzero; it is 1/2 when two elements are nonzero; it is 1/3 when three elements are nonzero. It is therefore a reasonable quantitative identifier for whether a mode is localized or delocalized, and is thus defined as the participation ratio. Its inverse, the IPR, describes in a slightly more intuitive way the number of elements that are similar in magnitude to the largest element. Thus, the IPR of eigenvector \( j \) is defined as

\[
IPR_j = \frac{\left(\sum_i v_j[i]^2\right)^2}{\sum_i v_j[i]^4},
\]

(4.39)

where \( v_j[i] \) is element \( i \) of eigenvector \( j \). For a set of normalized eigenvectors, the equation reduces to the inverse of the sum of the fourth power.

We can look at the IPR of the largest eigengaps modes of the system to assess whether fast or slow modes are more localized. To reduce the element to element variance, we computed the cumulative mean inverse participation ratio from the largest element to the sixth largest element in figure 4.17. With the exception of monomer 3, the trend is very clear: the most significant high frequency modes involve two or three residues acting collectively, while low frequency modes involve around 6-10. We can see some of the cause for the discrepancy by investigating figure 4.16. The different monomers appear to have almost exactly the same modes, but in differing orders of importance. These modes, identified by their largest components are:

1. Solvent, IPR 1, ranked: 3, 4, and 3.
2. HIS298, same monomer, IPR 1-2, ranked: 4, 2, and NA.
3. SER013, same monomer, IPR 1-3, ranked: 2, 1, and 1.
4. VAL301/GLY302, same monomer, IPR 2-7, ranked: 1, 5, and 4.
Most of these modes are well identified by a single major residue contributor, although not all monomers agree on exactly how this fluctuation proceeds. The monomer which is the best behaved, monomer 1, is also the one with the most localized fast modes and the most delocalized slow modes. This choice is in line with the practitioner’s bias of trusting that “boring” trajectories are more likely to be correct.

We can also consider the histograms as a source of evidence about the reliability of particular monomers’ trajectories. Consider the data in figure 4.18. Almost all of the modes look quite Gaussian with one very noteworthy exception. If we cross-reference these modes with the list we just created, it is clear that the SER013 mode is susceptible to deviations from Gaussianity on the blue end of the spectrum. However, as discussed previously, this end of the spectrum is not particularly important for nonlinearity upon absorption. What
is interesting, though, is that this blue tail of the spectrum will be of the most importance for energy transfer into this chromophore; BCL 373 is the most red shifted chromophore in the complex, so having enhanced fluctuations towards the blue would be able to increase the transfer rate into this chromophore. This SER013 mode’s behavior almost completely accounts for all of the differences between the different residues, but under this lens, it is hard to say which is the most physically likely. If a large nonlinear mode like this one is corroborated by more first-principles ab-initio calculations, the implications for quantum dynamics could be significant.

We can also ask why the fast fluctuations in residues act mostly independently while the slow fluctuations are more collective. The long rangedness of the coulomb interaction have the potential to provide a strong justification for this behavior. On long time scales, the motions that change and stay changed are delocalized around the pigment protein complex, and their impact as a group are comparable to the neighboring interactions due to the coulomb range. On short time scales, the motions of local residues cause large fluctuations, but their dynamic range changes quite little throughout the simulation. This separation into local and non-local suggests the inclusion of key local residues in quantum mechanical calculations to improve the theoretical correctness of results.

4.5 Conclusion

The energy gap dynamics in pigment protein complexes is a rich realm for physics that has long been taken for granted. The dynamical continuum of protein fluctuations leads to a picture that expands the perspective of static and dynamic disorder to one in which the low frequency mean changes both continuously and discontinuously. Meanwhile, the protection of high frequency modes allows for a consistent spectral density approach to quantum dynamics despite the labyrinthian voyage of the protein through conformational state space. These low frequency fluctuations are of potentially high significance, as their delocalized nature means that their fluctuations may have significant interchromophore and potentially intermonomer correlations, though it is just as likely that it is unrelated to the function of the protein and simply an unavoidable aspect of any feature that fluctuates through long ranged coupling. It may in fact lend evidence instead to the importance of fixing these chromophores in place so that the long ranged fluctuations do not have too deleterious an effect on the energy transfer dynamics.

These results also show that short protein simulations simply do not suffice to either equilibrate or explain features of protein complexes. Only after more than 1 µs were the basic protein energetics equilibrated, and even then, the individual environments of the separate monomers varied substantially at the precision of the gap calculation. It makes it difficult to consider the results of 50 ns as anything more than incidentally or qualitatively correct.

It is a tremendous boon for spectral density style theories that the high frequency fluctuations are protected. This may be why large-scale simulations of combined semiclassical
dynamics/molecular dynamics [14] did not unearth any fundamentally new physics. The main feature added by semiclassical models is the ability for the bath to respond nonlinearly. Due to the speed of energy transfer dynamics, only fast and local modes are strongly coupled to in these semiclassical dynamics, and so no nonlinearities are observed. It seems that the phenomenological improvements to FMO modeling do not lie in the high frequency dynamics of individual excitation trajectories but in the low frequency ensemble of excitation trajectory Hamiltonians.

The fact that the CDC failed to capture a sufficient degree of fluctuation to account for the environmental part of the ZINDO calculations was a bit troubling. Quantum chemical calculations have error that is difficult to characterize, and the amount of that error that translates into high frequency noise is likely quite high. In fact, Shim et al found that the high frequency part of their quantum calculations was extremely difficult to control[13], and argued in favor of a quantum mechanical origin of this behavior. It is certainly plausible that this high frequency noise is introduced by the method itself and has no physical origin. In this way, the quantum chemical methods may have a reduction of their variance that can meet the CDC’s underestimate somewhere in the middle.

While the spectral density seems overall quite effective, there are two features of the gap statistics that could be meaningful in certain parts of parameter space. The first is the skew in the density of states. While the presence of skew in absorption spectra is usually explained by a cartoon explanation of the Franck Condon principle, it should also be true that if both the ground and excited states are identical but shifted parabola, and the gap is really a linear function of the nuclear coordinates, then the absorption spectrum should be at the very least symmetric about a mean. Deviations from this indicate a non-trivial landscape for the energy gap potentially involving more than just the Q_y transition state, and therefore, more opportunity for complexity in the dynamics. In addition to this density of states and absorption skew, the presence of some anticorrelation between chromophores 3 and 4 is of some interest. If these chromophores were degenerate, this might be quite important, but as the model of FMO stands, these correlations are unlikely to be particularly important. It demonstrates, though, that some correlations are a realistic possibility in real pigment protein complexes, which is contrary to what has largely been argued in the field thus far.

As one final note, it is likely that spectral density behavior of this type is quite generic. If the source of this separation into local and non-local, fast and slow modes is the range of the interaction as hypothesized, it should generalize to other pigment protein complexes. Thus, the use of a spectral density for other complexes is likely also well justified.
Bibliography


[38] Dale E Tronrud et al. “The structural basis for the difference in absorbance spectra for the FMO antenna protein from various green sulfur bacteria”. In: Photosynthesis research 100.2 (2009), pp. 79–87.


[59] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. “Everything you wanted to know about Markov State Models but were afraid to ask”. In: Methods 52.1 (2010), pp. 99–105.


