2019

Exploring Novel Biophysical Mechanisms In Evolved Optical Systems

Dillion Fox

University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/edissertations

Part of the Biophysics Commons

Recommended Citation
https://repository.upenn.edu/edissertations/3402

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/3402
For more information, please contact repository@pobox.upenn.edu.
Exploring Novel Biophysical Mechanisms In Evolved Optical Systems

Abstract
Many sophisticated biological processes can be understood using basic concepts from physics. With this approach in mind, I will present two novel biophysical mechanisms derived from optical systems in marine animals. I will begin by describing the physical mechanism used by giant clams to increase diffusive flux of photosynthetic product from symbiotic algae. Over 50 years ago, it was reported that the addition of clam homogenate to algae stimulated photosynthate release, but the molecular effectors and mechanism of this "host release factor" have remained unresolved. Here we show that zwitterionic betaines, long known to exist in millimolar concentrations in coral and clam hosts, generally accelerate photosynthate diffusion. The electric field emanating from these molecules serves to organize water and generate a chemical potential for diffusion of small, polar molecules such as glucose and glycerol. The rest of my dissertation is dedicated to characterizing the sequence, structure, and function of the reflectin protein. The reflectin protein is responsible for giving cephalopods the remarkable ability to manipulate incident light to camouflage with their environment. This unique capability is made possible by iridescent cells composed of over 90% reflectin protein. The highly evolutionarily conserved region of reflectin that is unique to cephalopods has proven intractable to most standard biophysical characterizations including crystallization. Using a combination of sequence analysis and molecular modeling, I will explore the structure of motifs and the regions of sequence that connect them in bulk water and at membrane interfaces. I will then discuss the development of an experimental system designed to probe the interactions between reflectin and lipids. Here we show that reflectin, known to self-assemble into dense, membrane bound platelets, strongly interacts with lipid bilayers and increases the propensity for the membrane to adopt the flat structures observed in cells.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Physics & Astronomy

First Advisor
Alison Sweeney

Keywords
Biophysics, Computational, Intrinsically disordered protein, Molecular Simulation

Subject Categories
Biophysics

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/3402
EXPLORING NOVEL BIOPHYSICAL MECHANISMS IN EVOLVED OPTICAL SYSTEMS

Dillion Michael Fox

A DISSERTATION

in

Physics and Astronomy

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

________________________

Alison M. Sweeney, Associate Professor of Physics and Astronomy

Graduate Group Chairperson

________________________

Joshua Klein, Professor of Physics and Astronomy

Dissertation Committee

Philip Nelson, Professor of Physics and Astronomy

Eleni Katifori, Assistant Professor of Physics and Astronomy

Ravi K. Sheth, Professor of Physics and Astronomy

Loukas Petridis, Staff Scientist, Oak Ridge National Lab
EXPLORING NOVEL PHYSICAL MECHANISMS IN EVOLVED MARINE SYSTEMS

COPYRIGHT

2019

Dillion M Fox
None of this work would have been possible without the support, guidance, and generosity of my adviser and mentor, Alison Sweeney. Alison granted me the freedom to steer the project in whichever direction I wanted to take it, even if it meant she had to read up on a bunch of niche molecular simulation algorithms. Next, I would like to thank Jeremy Smith and Loukas Petridis for hosting me, mentoring me, and generously sharing their time and resources with me. I felt welcomed from the moment I arrived and I immediately benefited from their knowledge, advice, and mentorship. I would also like to thank Micholas Smith for helping me establish a connection with ORNL. Micholas was my first mentor. He taught me how to code, use Linux, and think about proteins, and those skills have carried me to where I am today.

I am very grateful to Seth Herzon and Xiaoshen Ma for being outstanding collaborators and for answering all of the random chemistry questions I’ve asked over the past few years. Thank you to Paul Heiney for the substantial guidance, support, and training you’ve provided me with. I would also like to thank Sophie Ettinger for training me on the Polscope and for always being there to help me find equipment in LRSM.

I owe a lot to Ryan Bradley, who helped me take my coding skills to the next level. I initially started working with Ryan so I could use a software package that he was working on, but he ended up teaching me a lot of new things about Python, linux, and
general coding practices. All of these skills have fundamentally changed the way I approach problems and I am very grateful for the time he spent teaching me.

None of this would have been possible without the support of my friends and labmates. First, I would like to thank everyone in the lab who helped me: James, Tom, Jill, Stephen, Emile, Amanda, Lincoln, and Asja. I would also like to thank Mouctar Diallo for working with me on the HRF project and for being patient as I learned how to be a mentor.

I am especially grateful to my MARC mentors, Chris MacDermaid, Dr. T, and Dr. Liz, for helping me get to where I am today. The MARC program totally rewired my perspective on issues relating to gender, race, class, and privilege. I am also grateful to Chris for continuing to support me, mentor me, and be a friend as I transitioned from Temple to Penn, and now from Penn to GSK. Chris has been one of my biggest role models since the day I met him.

Thank you to all of my friends, for listening to me blabber on and on about science, coding, and how Linux and Vim are superior to whatever system they’re using. Krystal Haislop, David Rivera, and Niloufar Khavari, you all helped me get through course work and you have all become very good friends. Lincoln and Asja – you guys suck. You have become two of my best friends. Chelsea, thanks for letting me use your desk. Dan, Brendan, Emily, Elgin – you guys rock.

Finally, I’d like to thank my family. My mom and Dad kept me on the right path while many of my friends lost their way. None of this would have been possible without your infinite support and patience. I’d like to thank my sister for always believing in me.
and wanting to know more about “whatever it is that I do.” Thank you to Chuck, Rose, Elyse, and Justin for your endless support and understanding.

Thank you to my partner, Leanna. You supported me through every difficult period of this journey and there is no way I could have done any of this without you.
ABSTRACT

EXPLORING NOVEL PHYSICAL MECHANISMS IN EVOLVED MARINE SYSTEMS

Dillion Fox
Alison Sweeney

Many sophisticated biological processes can be understood using basic concepts from physics. With this approach in mind, I will present two novel biophysical mechanisms derived from optical systems in marine animals. I will begin by describing the physical mechanism used by giant clams to increase diffusive flux of photosynthetic product from symbiotic algae. Over 50 years ago, it was reported that the addition of clam homogenate to algae stimulated photosynthate release, but the molecular effectors and mechanism of this “host release factor” have remained unresolved. Here we show that zwitterionic betaines, long known to exist in millimolar concentrations in coral and clam hosts, generally accelerate photosynthate diffusion. The electric field emanating from these molecules serves to organize water and generate a chemical potential for diffusion of small, polar molecules such as glucose and glycerol. The rest of my dissertation is dedicated to characterizing the sequence, structure, and function of the reflectin protein. The reflectin protein is responsible for giving cephalopods the remarkable ability to manipulate incident light to camouflage with their environment. This unique capability is made possible by iridescent cells composed of over 90% reflectin protein. The highly evolutionarily conserved region of reflectin that is unique to cephalopods has proven intractable to most standard biophysical characterizations including crystallization. Using a combination of sequence analysis and molecular modeling, I will explore the structure of motifs and the regions of sequence that connect them in bulk water and at membrane...
interfaces. I will then discuss the development of an experimental system designed to probe the interactions between reflectin and lipids. Here we show that reflectin, known to self-assemble into dense, membrane bound platelets, strongly interacts with lipid bilayers and increases the propensity for the membrane to adopt the flat structures observed in cells.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ III

ABSTRACT ........................................................................................................................ VI

LIST OF TABLES .................................................................................................................. X

LIST OF ILLUSTRATIONS .................................................................................................. XI

CHAPTER 1 INTRODUCTION ............................................................................................... 1

Giant Clams are Efficient Solar Energy Harvesters ........................................................... 2

Cephalopod Camouflage .................................................................................................... 4

The Cryptic Reflectin Protein ............................................................................................ 6

Iridocyte Morphology ......................................................................................................... 8

Intrinsically Disordered Proteins ....................................................................................... 11

Lipid Phases ........................................................................................................................ 13

Methodologies .................................................................................................................... 18

Polarization Microscopy ..................................................................................................... 18

Brief Overview of Solution X-ray Scattering .................................................................... 19

Introduction to Molecular Simulation ................................................................................ 22

Enhanced Sampling/Free Energy Methods ....................................................................... 25

Quantifying Water Organization ......................................................................................... 28

Overview of Chapters ......................................................................................................... 34

CHAPTER 2 BIOPHYSICAL MECHANISM AND MOLECULAR BASIS OF THE ENIGMATIC “HOST RELEASE FACTOR” IN MARINE SYMBIOSES .... 35

Results ................................................................................................................................. 36

Discussion ........................................................................................................................... 54

Methods ............................................................................................................................... 60

CHAPTER 3 STRUCTURAL INSIGHTS FROM REFLECTIN SEQUENCES .. 79

Methods and Materials ...................................................................................................... 83

Results ................................................................................................................................. 88

Conclusion .......................................................................................................................... 95
LIST OF TABLES

Table 2.1. Experimental design for the algal HRF assays.
LIST OF ILLUSTRATIONS

Figure 1.1. Spatial relationships between animal and algal cells within a photosymbiosis.  
a) Tridacnid giant clam in situ on a coral reef. Dark coloration of the tissue is due to high densities of unicellular Symbiodinium.  b) Low-magnification light micrograph showing brown Symbiodinium algae (indicated by white arrows) embedded within clam mantle muscle tissue (white tissue regions).  c) High-magnification light micrograph showing relationship of algal cells (shaded brown) to muscle cells (shaded green), with occasional clam iridocyte visible (shaded blue). There are no vascular structures for distribution of nutrients in this tissue.  d) Electron micrograph showing algal cell (shaded brown) next to clam cells with no specific distributive structures present.

Figure 1.2: Side view of iridocyte. All gray areas are intracellular, and all white areas are extracellular. The red arrows represent light reflecting off the reflectin platelets. The two length scales and indices of refraction required for Equations 1.1 and 1.2 are labeled.

Figure 1.3: Examples of reflectin containing structures.  a: Darkfield micrograph of dynamic iridophores at 40x magnification. Cells range in size from 20-40 μm.  b: Side view of iridophores. Dark bands are reflectin platelets. Scale bar represents 5 μm.  c: Photograph of eye silver. The dark circle in the center is the pupil.  d: TEM micrograph of
silver cells. e: Galiteuthis squid with ‘leaky’ light guide structures. f: Cross section of light guide fiber.

Figure 1.4 Lipid phases relevant to this study

Figure 1.5: A few examples of lamellar phases. a. The fluid phase is characterized by disordered tails and some irregularity in the spacing and arrangements of the head groups (green circles). b. The ordered gel phase has stiff, elongated tails and tight, hexagonal packing of the head groups. c. The interdigitated gel phase also has elongated tails but they interlock and drastically decrease the thickness of the bilayer.

Fig. 1.6. Diagram of 2 rings of a multilamellar vesicle. MLVs can have many layers. The space in between layers is filled with water.

Figure 1.7: Diagram of generic x-ray beam setup. The x-rays are generated by the source on the left side of the diagram and are carefully aimed at the sample in the center. A beam stop prevents the direct beam from reaching the detector, and only the elastic scattering from the sample is recorded by the detector (along with background noise, including cosmic rays).

Figure 1.8. Hydration shells surrounding trigonelline. Each consecutive shell is less organized. The red shell is almost as disorganized as liquid water.

Figure 1.9. Long-range electrostatic potential computed at the Willard-Chandler instantaneous liquid-liquid interface between a bilayer and solvent. The spheres laying on the rectangular grid map out the interface between the membrane and the solvent. The color is determined by the value of the long-range electrostatic potential, which ranges from red (negative) to blue (positive). The zwitterionic headgroups of the lipids have a
slight organizing effect on the water molecules in the system, as seen by the blue spheres on the surface. The molecule sitting on the surface is a protein with a net positive charge. This molecule has a stronger organizing effect than the membrane and causes the water molecules to flip orientation and point towards it.

Figure 2.1. HRF activity in cell culture assays. Colored diamonds show individual measurements of glycerol concentration after eight hours of incubation of cultured *Symbiodinium* cells in a given treatment, circles show normalized initial measurement. Treatment conditions are indicated across the top row. Relevant chemical structures are shown adjacent to their assay results. a) Comparison of LMWF and HMWF shows that the LMWF contains the HRF activity. b) The HRF-active molecule from the LMWF was isolated from the 16–30 min HPLC fraction. c) Trigonelline, identified from the 16–30 min fraction, showed maximum HRF activity at 100 mM concentration. d) The freebase form of trigonelline also shows maximum activity at 100 mM. e) Other betaines elicit HRF activity. The activity of sucrose was negligible, while sodium chloride was comparable to alanine betaine.

Figure 2.2. HRF activity in dialysis assay. a) Glycerol concentration as a function of time in a dialysis experiment. Glycerol was placed inside sealed dialysis tubing and HRF candidate molecules were placed in the solution external to the tubing. Average flux (units of mM m⁻² s⁻¹) was calculated for the first 50 min. of each experiment and is indicated by dashed lines on each dataset. Red circles indicate the condition of 100 mM trigonelline in the external solution. Dark blue circles indicate the condition of water only in the external solution. inset of a) Trigonelline diffusion out of the dialysis tubing
as a function of time. The diffusion of trigonelline from the tubing was much slower than that of the other molecules considered here. Accordingly, we considered trigonelline movement to be negligible during the first hour of equilibration. b) Control experiments for glycerol migration. The experiments shown here are similar to those in panel a, measuring 100 mM trigonelline in addition to glycerol inside the dialysis tubing (yellow circles), or 250 mM sodium chloride in the external solution (light blue circles). The resulting fluxes (units of mM m$^{-2}$ s$^{-1}$) are indicated by dashed lines.

Figure 2.3. Electrostatics and water polarization of glycerol and betaines, and their relationship to the HRF effect. a) Calculated LREPs at the interface of experimental molecules and bulk water. Two projections of each molecule are shown for clarity. Negative potentials are color-coded red while positive potentials are color-coded blue, with the magnitude of the potentials indicated by the color bar at right. The vector extending from the center of each molecule shows the relative magnitude of the calculated molecular dipole for each molecule in arbitrary length units. The magnitude of the dipole in Debye is indicated beneath each structure. b) Relationship between calculated molecular dipoles in Debye, and the HRF activity of each molecule as indicated by the HRF activity assay results shown in figure 2. The exponential fit and corresponding $R^2$ between the two quantities are shown on the figure.

Figure 2.4. Glycerol-glycerol interactions in the presence and absence of trigonelline. a) Glycerol-glycerol radial distribution functions with different concentrations of glycerol and trigonelline. Y-axis displays $g(r)$ for glycerol–glycerol spatial distributions, x-axis displays intermolecular distance in Å. b) Representative frames from MD simulation of
glycerol in solution, and glycerol in solution with trigonelline, with electrostatic potentials calculated at the glycerol-bulk water interface. Left panel shows points at which LREPs were calculated at the interface between glycerol and bulk water as individual spheres color-coded by the resulting potential. Negative potentials are coded red, while positive potentials are coded blue, with intensity indicating the magnitude of potential according to the color bar at right. Center panel shows glycerol in the presence of trigonelline. Spheres show points and magnitudes at which electrostatic potentials were calculated in the vicinity of glycerol, as in the left panel. Locations and orientations of trigonelline are shown by vectors. Overall magnitudes of electrostatic potentials in both the positive and negative directions are greater than in the glycerol-only case in the left panel. Right panel shows detail of an interaction between three h-bonded glycerols and three trigonellines in solution, represented in the same manner as the previous two panels. The bulk polarization of water around each glycerol molecule orients individual glycerols in a manner that promotes hydrogen bonding between them, either directly or through single water molecules.

Figure 3.1. Standard definitions for the N-terminus and internal reflectin motifs.
Figure 3.2: Structure of reflectin sequences with 3 examples of reflectin motifs.
Figure 3.3: a. Example periodic functions. Blue: sine wave. Green: square wave. Red: triangular wave. Orange: Sawtooth wave. b. The corresponding autocorrelation functions. The sine wave, triangular wave, and square wave all have similar autocorrelation functions. The sawtooth wave is never able to be totally out of phase with itself and therefore has less extreme wells in the autocorrelation curve. c. Random points. d.
Autocorrelation of random points. The curve is not periodic and therefore cannot be out of phase, so the function is never negative.

Figure 3.4. Graphical representation of the sequence dataset construction. The sequences derived from Illumina were used to create statistical profiles of internal motifs using HMMBuild. These statistical profiles were then used to identify sequences containing both N-terminal and internal motifs in the Illumina dataset using HMMSearch. The resulting sequences were clustering into self-similar groups and a representative was chosen for each cluster. These sequences were combined with the GenBank sequences to form the final dataset.

Figure 3.5. Skyline^2 graphics illustrating the relative probabilities of each residue in each sequence position for a: the N-terminal motif, and b: the internal motif. The height of the residue indicates the probability of it occurring there. The letters in each position are made thinner if there are insertions in that position in the sequence alignment.

Figure 3.6. Distribution of the lengths of a. N-terminal motifs, b: internal motifs. c: linkers connecting internal motifs. d: the distribution of the number of internal motifs per sequence.

Figure 3.7. a: Raw output of dG for all sequences in the sequence dataset. b: Example of a sine function fit to dG curve. c: dG curves aligned by phase shift. d: dG curves from charge scale. e: autocorrelation of dG for all sequences using interface scale. f: Correlation between interface affinity and charge.

Figure 3.8: Red points: Full-length sequences, Blue points: Reflectin motifs used in this study, Green points: Reflectin linkers used in this study. a. IDP plot shows almost all
sequences on disordered side. (Right) IDP Morphospace plot. All sequences are predicted to be weak-moderate polyampholytes/polyelectrolytes

Figure 3.9. Structural model of the reflectin protein. Reflectin linkers have high membrane binding affinity and reflectin motifs have low membrane binding affinity. Therefore we hypothesize that the linkers anchor the protein to the membrane.

Figure 3.10. Weblogo³ graphic showing statistics of multiple sequence alignment of representative sequences from spectral clustering

Figure 4.1: Sequence used in this study. Top: Full-length sequence. Middle: Sequence divided into motifs and linkers. Bottom: Interface affinity for chosen sequence.

Figure 4.2. Simplified diagram illustrating 5 systems (replicas) undergoing replica exchange. Each solid color line makes up a time-continuous trajectory, but the temperature changes over time due to swapping. Note: the figure is not drawn to scale.

The temperature swapping is instantaneous when performed in the simulation and therefore the sloped line representing the transition should be nearly vertical. The “room temperature ensemble” analysis is done on whichever frames happen to be in the gray dashed box at 300 K.

Figure 4.3. Monte-Carlo-generated structures for Motif 1 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log(R_g) vs. log(N). (c): R_g distributions. (d): secondary structure per residue determined by DSSP

Figure 4.4. Monte-Carlo-generated structures for Linker 1 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log(R_g) vs. log(N). (c): R_g distributions. (d): secondary structure per residue determined by DSSP
Figure 4.5. Molecular Dynamics derived structures for Motif 1 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of $\log(R_g)$ vs. $\log(N)$. (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.6. Molecular Dynamics derived structures for Linker 1 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of $\log(R_g)$ vs. $\log(N)$. (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.7: From left to right, structure for Motifs sequences 1, 2, and 3. Top row: Cumulative $R_g$ distributions. Bottom row: Average secondary structure determined by DSSP. Y-axis is probability, x-axis is residue number.

Figure 4.8: From left to right, structures for Linker sequences 1, 2, and 3. Top row: Cumulative $R_g$ distributions from 6 independent simulations. Bottom row: Average secondary structure determined by DSSP from 6 independent simulations. Y-axis is probability, x-axis is residue number.

Figure 4.9: From left to right, Motifs 1, 2, and 3. Cumulative contacts per residue from 6 independent simulations each. Normalized by number of frames in all simulations combined.

Figure 4.10: From left to right, Linkers 1, 2, and 3. Cumulative contacts per residue from 6 independent simulations each. Normalized by number of frames in all simulations combined.

Figure 4.11. a: Order parameter from 0-100 ns. b: Order parameter from 400-500 ns.
Figure 4.12. a: Average Thickness from 250-500 ns. Thickness is computed from the P atom in the headgroup. b: Snapshot of simulation. The blue lines outlines the periodic bounds of the system. The periodic images are displayed in the +/-X and +/-Y directions to help visualize the continuity of the curves, and therefore the system is displayed with 8 copies of itself total.

Figure 4.13: a: Average Interdigitation from 250-500 ns. b: Number of residues in contact with lipids vs. average interdigitation (nm). Color represents time in ns. c: Snapshots from simulation. First panel is at approximately 10 ns, second is approximately 100 ns, and third is approximately 300 ns. Dark gray lipids indicate upper leaflet and light gray indicates bottom leaflet.

Figure 4.14. Monte Carlo derived structures for Motif 2 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.15. Monte Carlo derived structures for Motif 3 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.16. Monte Carlo derived structures for Linker 2 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.17. Monte Carlo derived structures for Linker 3 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP
Figure 4.18. Representative structures from REST2 simulations.

Figure 4.19. Molecular Dynamics derived structures for Motif 2 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log($N$). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.20. Molecular Dynamics derived structures for Motif 3 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log($N$). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.21. Molecular Dynamics derived structures for Linker 2 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log($N$). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.22. Molecular Dynamics derived structures for Linker 3 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log($N$). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Figure 4.23. Distributions of radius of gyration from each motif simulated in the presence of a membrane.

Figure 4.24. Distributions of radius of gyration from each linker simulated in the presence of a membrane.
Figure 4.25. Distributions of end-to-end from each motif simulated in the presence of a membrane. Each distribution was fit to a Gaussian distribution. Many attempts at fitting were unsuccessful.

Figure 4.26. Distributions of end-to-end from each linker simulated in the presence of a membrane. Each distribution was fit to a Gaussian distribution. Many attempts at fitting were unsuccessful.

Figure 4.27. PCA results from each motif simulated in the presence of a membrane.

Figure 4.28. PCA results from each linker simulated in the presence of a membrane.

Figure 4.29. Lipid order parameter results from each motif simulated in the presence of a membrane.

Figure 4.30. Lipid order parameter results from each linker simulated in the presence of a membrane.

Figure 4.31. Lipid density as a function of the axis orthogonal to the membrane surface each simulation containing a reflectin motif simulated in the presence of a membrane.

Figure 4.32. Lipid density as a function of the axis orthogonal to the membrane surface each simulation containing a reflectin linker simulated in the presence of a membrane.

Figure 4.33. Number of motif residues in contact with lipids vs. average interdigitation (nm). Color represents time in ns.

Figure 4.34. Number of linker residues in contact with lipids vs. average interdigitation (nm). Color represents time in ns.

Figure 4.35. Local lipid interdigitation (nm) from bilayers simulated in the presence of motifs.
Figure 4.36. Local lipid interdigitation (nm) from bilayers simulated in the presence of motifs.

Figure 5.1: Phase separation in reflectin/MLV mixture. Phase a is clear and fluid. Phase b is viscous and cloudy.

Figure 5.3: Polscope images of silver cells from *Loligo pealeii*. Lighter pixels indicate higher retardance. The size and direction of the red lines indicate the degree and direction of organization.

Figure 5.4: Scattering intensity profile from reflectin in HFIP.

Figure 5.4. Left: DMPC MLVs above (blue) and below (green) the melting temperature. Right: MLV diagram showing the two correlation lengths in the scattering intensity profile.

Figure 5.5: Intermediate range x-ray scattering intensities for solutions containing a mixture of HFIP and MLV solution. The measurements were made for the same sample at 20 °C, 24 °C, and 28 °C.

Figure 5.6. Scattering intensity at 20 °C. The green curve comes from a mixture of HFIP and MLV solution in a ratio of 1:10. The blue curve comes from a mixture of reflectin solution and MLV solution in a ratio of 1:10. a: scattering intensity profile in the intermediate scattering range. b: scattering intensity profile from the low-q scattering range.

Figure 5.7. Temperature ramp for solutions containing reflectin stock and MLV stock in a ratio of a: 1:10, b: 1:15 (4/24), c: 1:20 (4/24), d: 1:50, and e. shows the intensity of the
secondary peak (\(q=0.136\)) for the three temperatures tested in a-d using the same color scheme.

Figure 6.1: Cartoon diagram showing how reflectin lowers the effective temperature of the membrane which causes an increase in the bending rigidity. We hypothesize that this mechanism counters the spontaneous curvature intrinsic to cells with spherical topologies.
Chapter 1

Introduction

The most sophisticated forms of “soft matter” (e.g., colloids, polymers, liquid crystals, etc.) on Earth likely occur in evolved, living matter. The field of soft matter is related to theories developed in hard condensed matter physics, which leverages symmetry to compute aggregate properties of particles arranged and interacting on a lattice. In soft matter physics, we use a similar approach, but without the formalism of a repeated lattice. These theories have allowed investigators to develop a first-principles approach to studying chemistry and biology. Our group seeks to link this first-principles approach to evolutionary insight in order to elucidate the physical driving forces behind the evolution of novel materials and biophysical mechanisms. We do this by constructing physical models of living systems and their evolutionary precursors and comparing the outputs of the models to experimental observables.

There are many ways we can quantify diversity in nature, but it is often difficult to discern whether certain properties were adapted in response to a particular demand or stimulus, or merely close correlates of another important character. Structural coloration
arises from precisely layered high- and low-refractive-index materials within a single cell. These materials are chitin and air in butterflies, and densely packed protein and water in cephalopods. Given clear roles for specific optical phenomena in these structured materials, biological systems that manipulate light are excellent candidates for understanding how evolving molecular properties contribute to evolving material function.

There are two systems that will be discussed. The first system I studied involves the mechanism used by giant clams to transfer photosynthetic sugar from symbiotic algae embedded in its mantle tissue out of the algae and into the clams’ metabolism. I also studied the structure and function of reflectin proteins, the protein family underlying the ability of cephalopods (squids, octopus, and cuttlefish) to self-assemble sub-visible-wavelength optical devices, allowing them to perform a variety of impressive optical tricks. In this chapter I will provide an overview of each system, general background information, methodology, and an overview of the structure of the dissertation.

**Giant Clams are Efficient Solar Energy Harvesters**

*Symbiodinium* dinoflagellates are photosymbionts in a vast range of marine species including all reef-building corals, sponges, giant clams, and sea slugs. This photosymbiosis is intrinsic to the coral reefs that cover 255,000 km² of the Earth’s surface and support 7.5% of the global human population. In the extremely nutrient-poor waters of the world’s tropical oceans, these photosymbionts use solar energy to fix
carbon dioxide into small organic building blocks. Gross carbon fixation of coral reef ecosystems is around 15 g m\(^{-2}\) day\(^{-1}\) (similar to that of temperate forests),\(^9\) but the net primary productivity, indicating carbon remaining in the system, is only 0–1 g m\(^{-2}\) day\(^{-1}\), such that the bulk of fixed carbon remains in easily respired forms in this system, not stored in long-lived structures, such as roots.\(^{10–12}\)

Attempts to understand the mechanism by which photosynthate is translocated from algal cells to their host date to 1967, when Muscatine reported that the addition of tissue homogenate from a reef coral or tridacnid giant clam induces glycerol release from *Symbiodinium*.\(^{13}\) This discovery raised the intriguing possibility that host-symbiont signaling is responsible for this nutrient transfer. Since Muscatine’s discovery, host-homogenate-stimulated photosynthate release from *Symbiodinium* has been documented *in vitro* for many symbioses.\(^{14–20}\) The list of proposed effectors, referred to as “host release factors” (HRFs) in the literature, is long and chemically disparate. Heat-sensitive proteins,\(^{15,17}\) amino acids,\(^{21,22}\) taurine\(^{23,24}\) and other low molecular weight but unidentified substances\(^{25}\) are among the HRFs advanced. One study also suggested a role for osmotic stress and/or nitrogen depletion in photosynthate release.\(^{26}\) There are no specific vascular or other distributive structures in these systems (in corals, the algae are located in a vacuole called the symbiosome, while in giant clams, algae are densely packed within the mantle tissue), such that all nutrient transport must be local between close-packed algal and host animal cells (Fig. 1.1).
The goal of Chapter 2 is to chemically identify the host release factor molecule and to elucidate the mechanism by which it increases diffusive flux of photosynthate across membranes.

Figure 1.1. **Spatial relationships between animal and algal cells within a photosymbiosis.**

- **a)** Tridacnid giant clam in situ on a coral reef. Dark coloration of the tissue is due to high densities of unicellular *Symbiodinium*.
- **b)** Low-magnification light micrograph showing brown *Symbiodinium* algae (indicated by white arrows) embedded within clam mantle muscle tissue (white tissue regions).
- **c)** High-magnification light micrograph showing relationship of algal cells (shaded brown) to muscle cells (shaded green), with occasional clam iridocyte visible (shaded blue). There are no vascular structures for distribution of nutrients in this tissue.
- **d)** Electron micrograph showing algal cell (shaded brown) next to clam cells with no specific distributive structures present.

**Cephalopod Camouflage**

Cephalopods are rich with evolved optical devices of sub-wavelength structure that self-assemble into bioluminescent light guides that provide camouflage, focusing mirrors that attract mates, and isotropic diffusers that lure prey. They have evolved several distinct light manipulation techniques: some cephalopods change the color of their skin, some are effectively transparent and utilize light guides reminiscent of fiber
optic cables to redirect light around their absorbing parts,\textsuperscript{28} and some use broadband Bragg reflectors to hide their eyes.\textsuperscript{29} These abilities to manipulate incident light derive from structurally colored skin cells, called iridocytes, that form elaborate dielectric structures.\textsuperscript{29}

These various sub-wavelength structures evolved in specific environmental optical and radiative contexts. Perhaps the most common evolutionary context in which these structures evolved is in manipulating light to camouflage prey from predators and vice versa.

Deep-sea squids live in a vast open void, never encountering a hard surface in their lives. There is no place to hide, so their only viable survival strategy is to camouflage with the radiance in the water that surrounds them by being transparent or guiding light around their bodies.

Eyes fundamentally cannot be made transparent, since they need to absorb light, and the squids cannot make their food transparent upon swallowing them, so they are left with a few parts that they physically cannot make invisible. However, they have evolved a very clever strategy for dealing with this. A paper by Holt et al\textsuperscript{30} shows how the squids have evolved proteinaceous tubes that act as light guides, redirecting light around the opaque organs and projecting it beneath them. These light guides “leak” light out the sides to match the horizontal radiance of light at the depths they occupy, replacing the light that would be emanating from the volume of the ocean that they instead occupy. The net effect is they are effectively indistinguishable from their environment, but instead of
relying on tuning reflected light, like the octopus, they rely almost entirely on emitted bioluminescent light.

The Cryptic Reflectin Protein

The reflectin protein is found in perhaps all cephalopod families and is the main building block giving rise to the structural optical effects described above.5,27,31 The high-index material in iridocytes of insects or vertebrates is typically chitin or crystals of purine.32 However, cephalopods seem to be alone in using dense proteinaceous platelets, whose curvature and complex geometry may allow the system to find new evolved functions.33–35 This system is so versatile that it can reflect the entire visible spectrum of light.36,37 Some cells in L. Pealeii reflect polarized light36,38, and they have been shown to do this during certain behavioral interactions.38

The dense protein platelets are wrapped in cell membrane39 and are interspersed with extracellular space. The protein platelets have a high index of refraction of about 1.4440, and the extracellular space has a low index of refraction of about 1.34, or just above water. Therefore, to a first approximation, cells reflect light using the rules of thin film interference41. The wavelength of light reflected off a Bragg stack can be calculated from Equation 1:

\[
d_H = \frac{\lambda}{4n_H} \quad (1.1)
\]

\[
d_L = constant \quad (1.2)
\]
The quantities in Equations 1.1 and 1.2 are defined in Figure 1.2.

The dynamic reflectors are controlled by tuning the thickness of the protein platelets. It has been shown that in the species *Loligo opalescens* and *L. pealei*, the neurotransmitter acetylcholine controls the thickness of the platelets by inhibiting the activity of a kinase.\textsuperscript{27,35,42,43} The hierarchical assembly of the protein can elicit large volume changes upon condensation,\textsuperscript{31,37} and the addition of phosphates can draw water into the platelets. The expansion and contraction of each layer is propelled by the intake and expulsion of water mediated by the Gibbs-Donnan effect.\textsuperscript{44} This is the first and only system to our knowledge that uses reversible post-translational modifications to control the state of a dynamic system.

![Figure 1.2: Side view of iridocyte. All gray areas are intracellular, and all white areas are extracellular. The red arrows represent light reflecting off the reflectin platelets. The two length scales and indices of refraction required for Equations 1.1 and 1.2 are labeled.](image)
Little is known about the history of the reflectin protein. There are no known related sequences outside of cephalopods. The ancestor of the reflectin gene was recently proposed to be a symbiotic bioluminescent bacterium *Vibrio fischeri*. There is also little known about how the cells form, but it has been shown that iridosomes (clusters of iridocytes) form before hatching.

There is significant interest in potential engineering applications of the reflectin protein. The most tantalizing application of the protein is to use it to inspire highly tunable, low latency camouflage coatings. Since reflectin has a natural tendency to form thin films, this is already being developed in laboratory settings. Reflectin has also been shown to be an excellent protonic transistor with a conductivity that rivals the best artificial proton conductors. The protein was also recently shown to be an excellent medium for neural stem cell growth. Finally, Iwasaka and colleagues showed that the reflectivity of reflectin thin films increases in the presence of magnetic fields, adding another avenue for potential engineering applications.

**Iridocyte Morphology**

Iridocytes come in a variety of shapes and sizes, with various names and functions. Figure 1.3 provides three common examples borrowed from literature. Each row describes a type of iridocyte. The column on the left shows the tissue to provide context for what it looks like in the animal, and the column on the right illustrates the structure of individual cells.
The first row, Fig. 1.3a, shows dynamically tunable camouflage in squid skin at 40x magnification from dark field microscopy, and Fig. 1.3b shows the cross-section of cells cut perpendicularly to the surfaces of the platelets. Each cell ranges in size from 20-40 microns wide, and each cell reflects a unique color. The alternating dark and light layers in Fig. 1.3b correspond to protein platelets and extracellular space. A close inspection of the micrograph shows the cell membrane wraps around each sheet in an accordion-like fashion.

Figure 1.3c shows the reflective tissue surrounding the eyes of a *Loligo pealeii* squid, called eye silver tissue. Squid eyes are fundamentally absorbing and are therefore very dark compared to their well camouflaged skin. They camouflage their eyes by minimizing the uncamouflaged areas by surrounding them with highly reflective tissue that reflects all of the light that hits it with an angular distribution consistent with the side-welling radiance of the region of the ocean in which they live. This tissue have “spindle” shapes (see Fig. 1.3d), and have been shown to be a Bragg reflector with a broad distribution of stack optical thicknesses.
Figure 1.3: Examples of reflectin containing structures. a: Darkfield micrograph of dynamic iridophores at 40x magnification. Cells range in size from 20-40 µm. b: Side view of iridophores. Dark bands are reflectin platelets. Scale bar represents 5 µm. c: Photograph of eye silver. The dark circle in the center is the pupil. d: TEM micrograph of silver cells. e: Galiteuthis squid with 'leaky' light guide structures. f: Cross section of light guide fiber.

Finally, Figures 1.3e and 1.3f describe the mid-water squid described above. The cross section of the light guide-like devices shows patterns of reflectin sheets that are somewhat elliptical but never totally close at the ends. These “parentheses” shapes, which
at first glance appear to be an inefficient design for wave guides, have been shown to closely match the radiance in the midwater depths where the squid are found.

**Intrinsically Disordered Proteins**

Reflectin sequences are unusual compared to canonically folding enzyme-like proteins. Nearly 20% of the residues in these sequences are tyrosine, and the sequences lack many of the hydrophobic residues commonly attributed to protein folding, such as leucine and isoleucine, such that there is almost no predicted secondary structure and no predicted tertiary structure. Therefore it is not surprising there has been speculation in the literature that reflectins might be intrinsically disordered.\textsuperscript{27,37} There are multiple ways to define intrinsic disorder in protein structures, but I use the following definition: an intrinsically disordered protein does not fold to one single conformation.

There are many methods to predict intrinsic disorder in protein sequences.\textsuperscript{59–64} One of the simplest methods, proposed by Uversky in 2000\textsuperscript{65}, is also one of the most intuitive. The method compares a scaled net hydrophobicity score of the sequence using the Kyte-Doolittle scale\textsuperscript{66} to a scaled net charge per residue score. The net charge is determined by summing the charge of each residue divided by the total number of residues. The mean hydrophobicity is calculated by taking the sum of the normalized hydrophobicity (Kyte-Doolittle scale\textsuperscript{66} with a window size of 5 and normalized to scale from 0 to 1) of all residues divided by the total number of residues minus 4. The authors showed that all known intrinsically disordered protein sequences occupy a well-defined
region of this plot and can be separated from all known canonically folded proteins by
drawing an empirically determined straight line that separates the two populations: \[<\text{Charge}> = 2.785 \times <\text{Hydrophobicity}> - 1.151\] (1.3)

If the protein falls on the left side of this line, then the protein is predicted to be
intrinsically disordered.

The logic supporting this method derives from the comparison of how
hydrophobic the protein is to how hydrophilic it is. If the protein does not contain a
minimum number of hydrophobic residues (or, more specifically, does not contain a
certain hydrophobicity) per total residue as determined by Equation 3.1, then it cannot be
expected to fold into a single conformation. Similarly, if there are enough charged
residues in the sequence, then no additional fraction of hydrophobic residues can
overcome the strong electrostatic interactions between the sidechains and the water
molecules.

If a protein is determined to be intrinsically disordered then the overall
compactness can be estimated by comparing the fraction of positively charged residues to
the fraction of negatively charged residues in the sequence. This method, proposed by
Pappu, can be understood using similar intuition as described above. For a sequence
containing a small fraction of both positively charged residues and negatively charged
residues, it will likely behave like a weak polyampholyte/polyelectrolyte. These
molecules tend to have a small radius of gyration, much like folded proteins, but they do
not fold to single conformations. If a protein has either an abundance of positively
charged residues or negatively charged residues, then the sidechains will strongly repel
each other and participate in strong electrostatic interactions with water molecules, which pulls the protein into a highly elongated shape, like a piece of spaghetti. This class of proteins is referred to as polyelectrolytes. If the protein has a lot of positive and negative charges then the sidechains repel each other less and there are more opportunities to form salt bridges. These interactions compete with water molecules and tend to make the protein a little more tightly packed than strong polyelectrolytes. Finally, there is a region that separates the weak polyelectrolyte/polyampholytes and the strong polyelectrolytes/ampholytes. The molecules in this region can adopt collapsed or extended conformations depending on the context of their environments. For example, a protein could be in an extended conformation in pure water but will collapse into a compact state if introduced to a chaperone protein.

**Lipid Phases**

Reflectin proteins have been speculated to associate with membrane interfaces based on their profile in an interface affinity scale. One of the main focuses of this dissertation is exploring the possibility that these interactions occur in nature, and characterizing any related membrane remodeling response. I use 1,2-dimyristoyl-sn-glycero-3-phosphocholine (DMPC) lipids in the simulations and experiments presented in Chapters 4 and 5. DMPC lipids have an unusually high melting temperature. This section serves as a primer on lipid phase behavior.
Lipids have highly complex phase diagrams. Shifts between lipid phases are determined by many factors, including acyl chain length, chemical properties of the headgroup, the overall lipid concentration, any cholesterol added to the system, solvents, solutes, and temperature, to name a few. Some commonly observed phases include bilayers, vesicles, and micelles. In this section I will discuss two phases that are less well known: the cubic bicontinuous phase and the sponge phase. Figure 1.4 shows simplified diagrams from Kulkarni et. al. of what these phases look like.

![Cubic Bicontinuous and Sponge Phase](image)

**Cubic Bicontinuous**  **Sponge Phase**

Figure 1.4 Lipid phases relevant to this study

The cubic bicontinuous phase is constructed from periodic repeats of minimal surfaces such that each point on the surface has zero mean curvature and negative Gaussian curvature (see Fig. 1.4). The sponge phase can be thought of as a melted cubic bicontinuous phase with less Gaussian curvature. The sponge phase can be accessed by introducing a surfactant to a microemulsion. The surfactant causes a first order phase transition by creating a tensionless liquid/liquid interface. The amount of surfactant added
to the mixture determines the tensionless area in the mixture and therefore the fraction of the lipids in the sponge phase.\textsuperscript{76}

Each lipid species is associated with a phase transition temperature, $T_m$, which is defined as the temperature required to transition from the ordered gel phase (Figure 1.6b) to the disordered fluid (or crystalline) phase (Figure 1.5a).\textsuperscript{78} Most biological membranes under physiological conditions occupy the fluid phase, which is less dense and more elastic than the gel phase.\textsuperscript{79,80} Lipids in the gel phase are tightly packed in a hexagonal lattice and have straight, rigid tails.\textsuperscript{81} At colder temperatures the tails from the top and bottom bilayer leaflets interdigitate and drastically decrease the thickness of the bilayer while increasing the rigidity, which we refer to as the interdigitated gel phase (Figure 1.5c).\textsuperscript{82} DMPC lipids form multilamellar vesicles (MLVs, Fig. 1.6) when mixed with water at room temperature and have a phase transition temperature of 24 °C.\textsuperscript{83}

![Figure 1.5: A few examples of lamellar phases. a. The fluid phase is characterized by disordered tails and some irregularity in the spacing and arrangements of the head groups (green circles). b. The ordered gel phase has stiff, elongated tails and tight, hexagonal packing of the head groups. c. The interdigitated gel phase also has elongated tails but they interlock and drastically decrease the thickness of the bilayer.](image)

Most lipid species have phase transition temperatures that are significantly lower than room temperature, and therefore most homogeneous bilayers prefer the fluid phase at room temperature. However, it has been shown that heterogeneous bilayers can prefer
the gel phase at a given temperature even if all of the constituent lipid species have lower phase transition temperatures. Membranes containing multiple lipid species can have mechanical and thermodynamic properties that differ greatly from all of its individual constituents.

Fig. 1.6. Diagram of 2 rings of a multilamellar vesicle. MLVs can have many layers. The space in between layers is filled with water.

Lipid composition is not the only factor that can affect phase transition temperatures. There are many molecules that can be incorporated into lipid bilayers, such as sterols and ceramides, that lower the phase transition temperatures. There is also evidence that proteins such as alpha synuclein can lower the phase transition temperature of membranes. However, apparently the only well-described way to increase the phase transition temperature of a membrane is to dehydrate or decrease the effective water concentration of the membrane. Water plays a fundamental role in membrane
structure and function, and there is a minimum molar ratio between lipids and water that is required to maintain equilibrium behavior for a given position in a lipid phase diagram. For example, DMPC bilayers require at least 27 water molecules for each DMPC lipid. Removing water molecules to a concentration below this critical threshold introduces a new, poorly described force commonly referred to as the hydration force, and the mechanical and thermodynamic properties of the membrane drastically change.

There are a few examples of natural systems that contain gel phases in the cell membranes. In most cases the gel phase is just one component of a multiphasic membrane. For example, there are gel phases that behave as lipid rafts in multiphasic membranes. It has been shown that some membrane proteins preferentially bind non-interdigitated gel phases because they can better accommodate the geometry of the protein. However, there is one example where the gel phase is more than just a subdomain of a fluid phase membrane. It has been proposed that the cell membrane of the mammalian skin barrier is a single gel phase, and the rigidity and decreased permeability of the membrane improves the efficacy of the barrier.

Membrane phase transitions not only affect the mechanical and thermodynamic properties of the membranes, they also affect the surface chemistry of the membrane. It has been shown that protein-lipid interactions are significantly weaker when the lipids are in the gel phase. For example, α-synuclein, which has been shown to decrease the phase transition temperature of membranes, preferentially binds fluid phase bilayer domains over gel phase domains.
Methodologies

Polarization Microscopy

The internal structure of materials can be probed by measuring birefringence. Birefringence is a material property that derives from orientation dependent refractive indices. The refractive index of a material is determined by the number of polarizable bonds along the axis of propagation of the incident light. Therefore if a material has a high degree of order, such as a crystal, there can be axes that have especially high densities of polarizable bonds. The axis with the highest index of refraction is called the extraordinary axis, and the orthogonal axis with the lowest index of refraction is called the ordinary axis.

Polarization microscopy is a technique that measures the retardance of a material. Retardance ($\delta$) is defined as the product of thickness of the material ($t$) and the birefringence, which is mathematically defined as the difference of the indices of refraction in the material along the extraordinary and ordinary axes ($n_e - n_o$): 

$$
\delta = t^* (n_e - n_o)\quad (1.4)
$$

Therefore if we know the exact thickness of our sample then we can directly compute the birefringence and estimate the degree of organization in the sample.
Brief Overview of Solution X-ray Scattering

One of the most effective methods for probing the structures of proteins and lipids is through x-ray scattering experiments. In these experiments, collimated x-rays are elastically scattered off of a sample and recorded on a detector (see Fig. 1.7). A beamstop prevents the direct beam from hitting the detector, so the detector only measures elastic scattering from the sample. This non-destructive technique can be used to determine the structure of crystals, chemical properties of soft matter systems, or even physical properties of thin films.

The uniform x-ray beam with wave vector $\mathbf{k}$ is generated by the source and scattered off the sample. The intensity of the scattered photons is detected as a function of the scattering angle, $2\theta$, or the momentum transfer vector, $\mathbf{q} = \mathbf{k} - \mathbf{k}_f$, also referred to as the...
scattering vector. The intensity pattern seen by the detector, \( I(q) \), is then the Fourier transform of the radial distribution function of the electron density.\(^{101}\) Therefore, elastic scattering experiments directly measure spatial correlations in materials. The intensity of the scattering pattern is determined by the contrast in the electron density and the total volume of the sample.\(^{101}\) For example, nanoparticles made from heavy metals dissolved in water will have much higher scattering intensities than methane dissolved in water because methane and water have very similar electron densities.

The x-ray scattering intensity as a function of the scattering vector can be used to extract spatial information about the arrangement, size, distribution, and shape of the internal structure of the sample, as well as interactions between the scattering elements in the sample.\(^{101}\) In order to determine physical quantities related to the scattering elements in the sample, such as molecular weight, particle volume, or specific surface area, the intensity curve must be normalized to an absolute scale.\(^{102}\)

The intensity curve can also be directly analyzed by identifying pronounced features at given scattering vectors, \( q_0 \), and converting the scattering vectors to correlation lengths, \( d \), with the following relationship:

\[
    d = \frac{2\pi}{q_0} \quad (1.5)
\]

We therefore think about features in terms of q-ranges. In the “high q-range”, sometimes called Porod’s region, we observe very small features (< 1 nm in real space) that occur at the interface between two scattering elements.\(^{102}\) This is the realm where it is possible to resolve the positions of atoms with precision on the Angstrom scale. On the opposite end of the spectrum we have the “low q-range”, which provides information about the
structural order of the system\textsuperscript{102} (typically > 10 nm in real space). Finally, the region bounding high- and low-q is the intermediate range (approximately 1-10 nm in real space). This range can detect medium-sized features in the system such as the size and shape of individual particles.

The slope of the intensity curve in log-log space correlates to a property of the system called the \textit{fractal dimension}.\textsuperscript{103} Mathematically, the intensity curve is proportional to the magnitude of the scattering vector raised to the power $D$, where $D$ is the fractal dimension:

$$ I(q) \propto q^{-D} \quad (1.6) $$

The concept of a fractal dimension can be conceptualized by envisioning a sphere of radius $r$ enclosing part of an object with mass $m$.\textsuperscript{104} If the object is a straight line, then the mass enclosed by the sphere is proportional to the radius. If the object is a plane, then the mass is proportional to the radius squared.\textsuperscript{104} Therefore, if the object is self-similar over a range of length scales then the mass is related to the radius of the sphere according to the following proportionality:

$$ m \propto r^D \quad (1.7) $$

The mass of the enclosed object is proportional to the scattering intensity, and the radius of the enclosing sphere is inversely proportional to the correlation length ($d$) being probed. The fractal dimension does not need to be an integer. For example, the fractal dimension of a polymeric system is the inverse of the Flory exponent.\textsuperscript{105}
Introduction to Molecular Simulation

The first realistic molecular simulation was run in 1974. Molecular simulations have since become some of the most valuable techniques for studying the structure of matter. The concept is simple: if we can represent a material by a discrete number of objects, and we can describe the interactions between the objects, then we can find energetic states accessible to the system by allowing the system to evolve according to the laws of thermodynamics and equations of motions. There are two standard methods for evolving a system through energetic states: molecular dynamics simulations and Monte Carlo simulations. In the former, we compute the forces on each object and predict where each object will be over a short intervals of time based on those forces. In the latter, we make a tree containing all of the possible ways an object in the system could move. We randomly choose a move and re-calculate the energetic state of the system, and we can either accept or reject the move.

If we try to interpret nature in the most literal sense, then we can describe a molecule by representing each atom as a singular point, and we can use quantum chemical models to describe the electrons with varying degrees of accuracy depending on the requirements of the study. It is often impractical to describe interactions using the full complexity of quantum mechanics, so simplified models, such as Density Functional Theory (DFT), have been developed that achieve sufficiently accurate results for a wide range of cases.
Biological systems are often large enough that computational constraints do not allow for even simplified quantum mechanical models. In these cases, we reduce interactions between electron clouds down to a limited number of simple cases: electrostatic interactions, Van Der Waals (VDW) interactions, and so-called “bonded” interactions. This representation allows us to represent the sharing of electrons between bonded atoms using harmonic oscillator restraints, and interactions between non-bonded atoms using partial charge assignments. For example, the electrons in a neutral molecule like water can be represented by assigning a charge of approximately $0.417*e$ to the hydrogens and $-0.834*e$ to the oxygen atom, as is the case for the TIP3P water model. Each atom is therefore assigned a partial charge and two VDW parameters, one for the effective radius of the atom, and another for the strength of the interaction. Each bond is modeled as a harmonic oscillator, and each bond type is assigned a spring-like stiffness and an equilibrium length. Therefore, in simulations of water, it is possible to reduce the full complexity of the quantum mechanical model describing the electron clouds down to just a few numbers, and this simple representation is able to recapitulate much of the experimentally verified phase behavior of water.

Atomistic modeling has many limitations. First of all, the bonds are modeled as harmonic oscillators, which does not allow them to break. Therefore, the most primitive (and the most common) implementation of this force field does not allow for chemical reactions. There are models that allow chemical reactions to take place (i.e. ReaxFF), but these models require significant calculations to be done before the simulation is run, and these pre-calculations are only valid for the specific systems for which they are
computed. Such models also do not recapitulate the interactions between intricate electron clouds with great accuracy, like aromatic rings, for example. Finally, the derivation of the values representing each atom type and each bond type is critical, and there are several opposing schools of thought for how this should be approached. The terms can be derived from quantum mechanical models, NMR data, or through machine learning to recapitulate experimental observables, but each method yields systematic error and the results of each method cannot be mixed.

Monte Carlo Simulations

Monte Carlo is a general simulation technique based on repeated random sampling to model deterministic processes. There are many varieties of Monte Carlo algorithms, but most are based on the following template: model a system based on probability density functions (PDF), sample states from the PDF, and perform some kind of statistical analyses on the results. Monte Carlo simulations have been used in a wide variety of contexts, including modeling galaxy formation, finance, climate change, and computational biology.

This principle is applied to sampling energetic states in molecular systems by constructing a tree that contains every possible next move in the system; for example, if the system contains CO molecules, then each molecule can be translated, rotated, or stretched. The user defines the probability density function that underlies all of the possible moves, and then moves are chosen at random. Each time a move is selected, the
energy of the new configuration is evaluated. Typically, if the move lowers the energy of
the system then it is accepted. However, if the move raises the energy of the system then
another random variable is invoked to determine whether or not to accept or reject the
move. This acceptance-rejection criteria allows the system transition out of local
energetic minima, which would otherwise limit conformational sampling.

Monte Carlo methods can rapidly cycle through physically realistic
conformational states using similar force fields as molecular dynamics simulations,\textsuperscript{119} but
they typically require implicit solvent models, meaning the simulation is run without
water molecules. Instead of explicitly representing the water, the effects of electrostatics
and collisions with water molecules are modeled by applying approximations such as the
Poisson Boltzmann equation for electrostatics and Langevin dynamics to modify the
equations of motion.\textsuperscript{120}

Enhanced Sampling/Free Energy Methods

Standard molecular dynamics is the most common and straightforward method to
evolve a protein through energetic states, but many molecular processes (such as protein
folding) occur over unattainably long time scales to directly simulate them. Several
research fronts aim to solve this problem, including: designing specialized hardware to
run the calculations, designing compilers and code to optimize calculations, and
designing better algorithms that cycle through physically realistic energetic states at
faster rates. Among these methods are umbrella sampling and free energy sampling, which will be the focus of this section.

Adaptive Biasing Force

Adaptive Biasing Force (ABF) simulations are designed to maximize sampling along a reaction coordinate, \( \xi \). The algorithm is one example of a larger class of algorithms that compute the Potential of Mean Force (PMF) to optimize sampling in molecular simulations. A PMF is a description of energy in the form of reversible work put into the system to bring two (or more) particles together from infinite separation. This PMF method estimates the Helmholtz free energy, \( A(\xi) \), as a function of the reaction coordinate by applying small forces to the molecules and measuring the response. In equation form, we are trying to approximate:

\[
\frac{dA(\xi)}{d\xi} = \langle F(\xi) \rangle \quad (1.8)
\]

where \( \langle F \rangle \) is the average force applied. The instantaneous force is calculated based on the change in the reaction coordinate:

\[
F_{ABF} = \nabla_\xi A = -\langle F(\xi) \rangle \nabla_\xi \xi \quad (1.9)
\]

The simulation begins with an equilibration phase by measuring \( A(\xi) \) in discrete bins \( (\Delta \xi) \). Once the bins have recorded a threshold of measurements, the calculation is switched on. The value of the reaction coordinate is measured at each step in the simulation \( {\xi}_i \) and is used to lookup the current estimate of \( A({\xi}_i) \) at that value, which is
used by Eq. 1.9 to compute the applied force. The applied force is then subtracted from the molecules of interest, which effectively removes the interactions between them and allows them to diffuse randomly according to random walk statistics. In my experience, this increases the sampling along the reaction coordinate by an order of magnitude because the molecules do not get stuck in energetic minima. Since the free energy is computed at each step, the method is self-correcting and is guaranteed to improve over simulated time. The values of $A(\xi)$ can then be used to calculate the radial distribution function, $g(r)$, with the following relation:

$$g(\xi) = e^{-\frac{A(\xi)}{kT}} (1.10)$$

Replica Exchange

Replica exchange simulations are designed to push systems out of energetic wells by constantly cycling through temperatures. The setup is as follows: N identical copies (replicas) of the simulation are created, and each replica is assigned a unique starting temperature, usually spaced out on an exponential scale. Each replica simulation is run at once, and every so often (typically 2000 time steps), the total energy of the system is computed, and the energies are compared between each system and the systems with the closest temperatures. If two neighboring systems have similar energies, then they swap temperatures. It has been shown that temperature swapping is a random variable, and therefore the temperature of each system is a random walk. Replica exchange simulations are typically run until the system that started with the highest and lowest temperature have both swapped. Since the temperature swaps are random walks
on a discrete lattice, the number of steps in the simulation required to achieve this can be estimated by computing the first passage time of the coldest replica.\textsuperscript{132}

The REST2 method improves on this framework by only scaling the effective temperature of the solute.\textsuperscript{133} Rather than literally heating and cooling the solute and the solvent, which takes time to equilibrate, the method scales solute-solute and solute-solvent interactions according to the following formula for replica \( m \):\textsuperscript{133}

\[
E^{\text{REST2}}_m (X) = \frac{\beta m}{\beta_0} E_{pp}(X) + \sqrt{\frac{\beta m}{\beta_0}} E_{pw}(X) + E_{ww}(X) \quad (1.11).
\]

The subscripts \( p \) and \( w \) stand for ‘protein’ and ‘water’, respectively, \( X \) is all of the coordinates in the system, and \( \beta \) is the inverse of the product of the Boltzmann constant and temperature. The factor in front of the protein-protein term decreases the energy barrier for conformational changes. The additional factor in front of the protein-water term follows from the combination rules of Lennard Jones interactions. These modifications to the standard replica exchange method leads to an order of magnitude increase in performance.\textsuperscript{133}

Quantifying Water Organization

One of the most important concepts in understanding the HRF effect, discussed in Chapter 2, is the structure of liquid water. Liquid water molecules tend to arrange themselves in a tetrahedron; the hydrogens (positively charged) point toward the oxygen (negatively charged) of another water molecule.\textsuperscript{134–136} The distances and angles
describing these hydrogen bonds form a distribution centered around roughly 2 Å and 109.5°. Therefore, one of the most intuitive ways to quantify the structure of water is to count the number of water molecules adhering to this tetrahedral order. This is most commonly done by computing the revised orientational tetrahedral order parameter:

\[
q = 1 - \frac{3}{8} \sum_{i=1}^{3} \sum_{k=i+1}^{4} \left( \cos \psi_{ik} + \frac{1}{3} \right)^2
\]

(1.12)

where the angle \(\psi_{ik}\) is defined as the angle formed between an oxygen atom and two neighboring oxygen atoms. If the molecules are arranged in a regular tetrahedron then \(q=1\). However, if the molecules have a totally random arrangement (such as an ideal gas), then \(q=0\). Liquid water at room temperature has an average \(q=0.67\).\(^{139}\)

The various orientational order parameters provide an estimate of the tetrahedrality of a snapshot of water.\(^{139,140}\) These estimates have been shown to be sensitive to subtle changes in temperature and pressure,\(^{139}\) but in my experience they are not sensitive to perturbations from dissolved solutes, such as salts. Molecules with strong electric fields have a polarizing effect on water. The organized water molecules surrounding a dissolved solid (such as an ion) are referred to as the hydration shell. Hydration shells typically only extend a few water molecule-lengths from the solute (~1 nm).\(^{141}\) These water molecules form a strong hydrogen bonding network that causes them to lose some of their translational and rotational kinetic energy.\(^{142}\) The water immediately outside of the canonical hydration shell also becomes more organized,\(^{143}\) but the increased order is difficult to detect from individual snapshots from a simulation.
The size of the hydration shell (called the Stokes radius, or hydrodynamic radius) can be determined by computing the diffusion rate of water molecules in shells, i.e. discrete radii, around the center of the solute (see Fig 1.8 for an example).\textsuperscript{144} The diffusion rate can be easily estimated from the Einstein relation:\textsuperscript{145}

\[ \langle (x(t) - x_0)^2 \rangle = 2Dt \] (1.13)

where \( x(t) \) is the position of a particle at time \( t \), \( x_0 \) is the initial position, and \( D \) is the diffusion constant. The diffusion constant decreases as the radius increases, and therefore the Stokes radius can be determined by setting an arbitrary cutoff diffusion rate and finding the radius that matches.\textsuperscript{146}
Figure 1.8. Hydration shells surrounding trigonelline. Each consecutive shell is less organized. The red shell is almost as disorganized as liquid water.

Despite the major perturbations to the liquid water structure that result from adding solutes, in my experience, the tetrahedral order parameters are unable to detect them. This subtlety will be made apparent in Chapter 2. I hypothesize that this is largely due to statistics; there are very few water molecules impacted by the addition of solutes compared to the total number of water molecules, and therefore the perturbation is washed out in the sum.
A method was proposed in 2014 that is more sensitive to long range ordering of polar molecules. Rather than compute all of the angles in the system, the authors propose computing a weighted electrostatic potential at various points in the system to represent the net polarization of water. While this method, referred to as electrostatic mapping, essentially reduces to summing over dot products of vectors, it can be computed at various points around the solute and detect differences in water structure as seen by those points instead of globally. The potential is weighted by an error function that is scaled to exclude contributions from the hydration shell so that the sum is not dominated by vitrified water. This modified potential, called the long-range electrostatic potential, can be computed at the interface between the solute and the water and account for long range ordering effects caused by the solute. Figure 1.9 shows an example of the long-range electrostatic potential computed at the Willard-Chandler instantaneous liquid-liquid interface between a POPC bilayer and the solvent.
Figure 1.9. Long-range electrostatic potential computed at the Willard-Chandler instantaneous liquid-liquid interface between a bilayer and solvent. The spheres laying on the rectangular grid map out the interface between the membrane and the solvent. The color is determined by the value of the long-range electrostatic potential, which ranges from red (negative) to blue (positive). The zwitterionic headgroups of the lipids have a slight organizing effect on the water molecules in the system, as seen by the blue spheres on the surface. The molecule sitting on the surface is a protein with a net positive charge. This molecule has a stronger organizing effect than the membrane and causes the water molecules to flip orientation and point towards it.
Overview of Chapters

I will begin by discussing the biophysical mechanism used by giant clams to extract nutrients from symbiotic algae in Chapter 2. Then I will transition to discussing the reflectin protein. I will begin Chapter 3 with a statistical analysis of reflectin sequences based on a transcriptome dataset from the Sweeney group. I will show that reflectin sequences as a whole are predicted to be intrinsically disordered, and I will present a new framework for interpreting reflectin sequences that is based on periodically spaced regions of high membrane interaction potential. Chapter 4 will provide an in-depth account of my molecular modeling results. I will present the first predicted structure of the reflectin protein along with the results of extensive molecular simulations in bulk water and at membrane interfaces. Finally, in Chapter 5 I will present experimental results that provide deep insight into the function of the reflectin protein in the cell and corroborate the modeling results.
Chapter 2

Biophysical mechanism and molecular basis of the enigmatic “host release factor” in marine symbioses

We have investigated the biophysical mechanism of HRF-stimulated photosynthate release and identified a passive transport mechanism driven by zwitterionic organic molecules known as betaines as responsible for this effect. Betaines are produced in millimolar concentrations in corals and giant clams. In this work, we show that in physiologically-relevant concentrations, betaines increase the flux of glycerol and glucose out of the alga via a physical mechanism that requires no energy from either cell other than the synthesis of the betaine itself. In the presence of an animal cell “sink”, this increased efflux could then be maintained indefinitely. Our identification of betaines as underlying the HRF effect is consistent with the vast preponderance of the studies in this field.
Results

Our experiments employed the giant clam *Tridacna derasa*, a *Symbiodinium*-bearing photosymbiont that has a large living tissue mass and can be purchased from aquaculturists. In contrast, reef-building corals are threatened, and have less than a millimeter of living tissue thickness covering their calcareous skeleton, rendering bulk biochemical extraction challenging.

In accord with the preponderance of earlier work in this area,\textsuperscript{14–20} we focused on HRF-stimulated glycerol release from alga. We utilized a commercial, enzyme-based colorimetric assay to measure glycerol concentration. A recent study raised the possibility that glucose is the primary photosynthate of *Symbiodinium*.\textsuperscript{150} We verified that betaines also accelerate glucose flux; simulations show that this mechanism is also physical in nature, though the details of the origin of the chemical potential are slightly different from the case of glycerol (see Supporting Information).

Chemical identification of HRF

We performed activity-guided fractionation to identify the compound(s) responsible for HRF activity. To assay HRF activity, we washed algae in fresh, glycerol-free medium, incubated them at a fixed cell density with a fraction or compound of
interest, and measured the concentration of glycerol in the medium initially and after eight hours. Unavoidably, there are significant quantities of glycerol present in the clam tissue, and some of the compounds we assayed also caused background cross-reaction in the assay. For these reasons, we report both the raw and background-subtracted glycerol assay data (Figs. 2, S1). Since glycerol production from *Symbiodinium* was variable between cultures, all data reported here were obtained from a single, freshly split and washed flask of our stock algal cell culture, unless otherwise noted. Raw data from colorimetric glycerol assays are included in the Supplementary Information.

On incubation with *Symbiodinium* culture, crude clam homogenate caused a mean increase of 7.18 µM glycerol in the medium after eight hours, compared to the initial background concentration. We then used spin-column fractionation to separate this crude homogenate into a high molecular weight fraction (HMWF, >3 kD) and a low molecular weight fraction (LMWF, <3 kD) (Fig. 2.1a). The LMWF resulted in a mean increase in glycerol concentration of 35.4 µM. The HMWF fraction unavoidably contained particulate animal material, and yielded an apparent decrease in glycerol concentration of 36.7 µM. We speculate that this apparent decrease in absorbance was due to the tissue background, glycerol entrainment by protein in these samples, and/or light-scattering from particles. Additionally, there was no obvious color shift during the assay, as was the case for other treatments with positive results. These observations, and the experiments below, suggest the HMWF does not contain HRF activity (Fig. 2.1a).

Accordingly, the LMWF was further analyzed by high-performance liquid chromatography (HPLC). HRF activity was more difficult to detect in HPLC fractions of
this LMWF due to the unavoidable dilution introduced by this separation technique. However, we observed reproducible HRF activity (an average increase in glycerol concentration of 7.18 µM) in the fraction collected between 16 and 30 min, and no average increase from any other fraction (Fig. 2.1b). We further fractionated the mixture obtained in the 16–30 min fraction via HPLC separation using the same conditions. In this separation, we isolated individual components absorbing at 290 nm or 320 nm (these eluted at 20.5, 21.1, 27.6, 28.6, 31.4, and 38.5 min) and analyzed their structures by nuclear magnetic resonance (NMR) spectroscopy (see Supplementary Information).

Figure 2.1. **HRF activity in cell culture assays.** Colored diamonds show individual measurements of glycerol concentration after eight hours of incubation of cultured *Symbiodinium* cells in a given treatment, circles show normalized initial measurement. Treatment conditions are indicated across the top row. Relevant chemical structures are shown adjacent to their assay results. a) Comparison of LMWF and HMWF shows that the LMWF contains the HRF activity. b) The HRF-active molecule from the LMWF was isolated from the 16–30 min HPLC fraction. c) Trigonelline, identified from the 16–30 min fraction, showed maximum HRF activity at 100 mM concentration. d) The freebase form of trigonelline also shows maximum activity at 100 mM. e) Other betaines elicit HRF activity. The activity of sucrose was negligible, while sodium chloride was comparable to alanine betaine.

NMR analysis revealed the peak eluting at 20.5 minutes to be an unresolvable mixture of two species, both of which contained aromatic and alkyl residues, and we
were not successful in further characterizing these molecules. The material eluting at 21.1 min was determined to be the cyclic betaine molecule trigonelline (Fig. 2.1c) by proton and carbon-13 NMR spectroscopy, and comparison to an authentic standard (see Supplementary Information). The materials eluting at 27.6, 28.6, and 31.4 min were too sparse for exact identification, and the material at 38.5 min was identified as the ubiquitous, constitutive biomolecule adenosine.

To determine if trigonelline exhibited HRF activity, we used an authentic sample of trigonelline hydrogen chloride and evaluated its HRF activity by performing the HRF assay described above.

When algae were incubated in fresh medium only, the final concentration of glycerol in the medium was indistinguishable from the initial concentration, as expected. When trigonelline HCl was added to the medium, the glycerol in the medium after eight hours increased as a function of increasing trigonelline HCl concentration, up to a concentration of 100 mM trigonelline (Fig. 2.1c). A concentration of 25 mM trigonelline HCl resulted in 29.0 µM glycerol after eight hours, while 100 mM trigonelline HCl resulted in 145 µM glycerol. However, when the concentration of trigonelline HCl was increased to 250 mM, the final glycerol concentration was 34.2 µM, less than that when 100 mM trigonelline HCl was present. The inner salt of trigonelline behaved similarly: 100 mM of trigonelline inner salt induced a maximum glycerol concentration increase of 119 µM (Fig. 2.1d). Given that the pKₐ of trigonelline HCl is approximately 2.8, we expect the inner salt form to be the predominant species in vivo in the clam. Glucose also
experienced increased flux from algal cells in the presence of 100 mM trigonelline in this assay: glucose in the medium increased from 2.1 to 30.5 µM after eight hours.

To distinguish photosynthesis-driven glycerol release from other mechanisms, such as membrane damage or permeabilization by trigonelline, we repeated the experiment above using 100 mM trigonelline in the dark. In darkness, the increase in glycerol in the medium (8.17 µM) was 18-fold less than when the assay was exposed to light (145 µM) (Fig. 2.1). Therefore, trigonelline-induced HRF activity requires active photosynthesis and is not a result of cell wall damage or otherwise increasing cell wall permeability.

Biophysical characterization of the HRF effect

The result that HRF activity increases linearly with trigonelline concentration up to 100 mM, and then decreases with further addition of trigonelline to the system led us to hypothesize that a purely physical mechanism, rather than a classical receptor-mediated biochemical mechanism or other form of active cellular transport, was responsible for the HRF effect. Under this hypothesis, discrete ligand-receptor interactions would not be required, and the physicochemical characteristics of trigonelline and similar molecules would then predict HRF activity. To test this hypothesis, we evaluated the HRF activity of glycine betaine, alanine betaine, proline betaine, and taurine (taurine has also been advanced as a candidate HRF molecule). At 100 mM concentration, these betaines all exhibited more HRF activity than the...
medium alone, with glycine betaine inducing a minimum activity among betaines (mean glycerol concentration increase of 8.92 µM) and taurine inducing the maximum activity we observed (mean glycerol concentration increase of 231 µM). Alanine betaine and proline betaine showed intermediate HRF activity, with average glycerol concentration increases of 52.3 µM and 107 µM, respectively (Fig. 2e). The activity of trigonelline was intermediate between that of proline betaine and taurine (Fig. 2c–e).

If movement of glycerol out of algal cells in the presence of HRF molecules is a purely physical phenomenon, then Le Chatelier's principle predicts that excess glycerol in the medium, even in the presence of HRF, should drive the system in reverse, sending glycerol in back into alga. To test this, we conducted an HRF assay with both 100 mM trigonelline and 250 mM glycerol in the algal medium. When glycerol was present in the medium in addition to trigonelline, there was a net decrease (by 34.7 µM) in glycerol concentration after eight hours, suggesting that glycerol diffused back into the algal cells. This result indicates that the efflux of glycerol from *Symbiodinium* in the presence of trigonelline is a physical process that does not directly leverage additional biochemical energy from either the algae or the clam such that glycerol movement may only occur in one direction, from the alga to the clam. Presumably the lack of glycerol efflux observed in medium alone is due to the presence of a metabolic sink for glycerol internal to the algae, resulting in a chemical potential for glycerol to remain within algal cells.

Since HRF activity is apparently shared by most or all betaines, but these molecules share the physical characteristics of being both zwitterionic (two oppositely-charged centers in each molecule) and polar, we assayed other compounds to identify the
physical characteristics of betaines that induce HRF activity. We conducted HRF assays using sucrose (polar but uncharged) and sodium chloride (charged, but forming two separated counterions in solution) to decouple these parameters. An HRF assay using 100 mM sucrose resulted in a mean increase in glycerol concentration of 3.06 µM, the smallest measurable increase of any compound tested (Fig. 2e). When we used 250 mM sodium chloride in an HRF assay, glycerol concentration increased by 4.08 µM (Fig. 2e). Consistent with osmotic stress, the cells darkened and shrank in this treatment; however, the observation of some glycerol release suggests that the cells survived the treatment. These experiments indicate that inorganic salts and polar carbohydrates possess some HRF activity, but it is an order of magnitude less than that observed for betaines, where counterions are combined in a single molecule. We conclude that it is the combination of paired charges in a single molecule that drives glycerol release.

To test whether living cells were required for the HRF effect we observed, we conducted a series of experiments using only glycerol and our compounds of interest on either side of a dialysis membrane, mimicking diffusion through an algal cell membrane and cell wall. We sealed a solution of 25 mM glycerol in water inside dialysis tubing, resulting in a reservoir of glycerol inside a semipermeable membrane similar to glycerol production within living algal cells. We placed this sealed dialysis tubing in a solution containing 100 mM of one of the molecules assayed above. Given that glycerol has been characterized in high concentrations (1–2 M) as an osmoprotectant in algae,\textsuperscript{151} and in moderate concentrations (~2 mM) as part of the photosynthate reservoir in land plants,\textsuperscript{152} we chose a starting concentration of 25 mM glycerol inside the tubing as an order-of-
magnitude estimate of the cellular concentration of glycerol in *Symbiodinium*. The permeability coefficient of glycerol with respect to cell membranes is similar to that of near-freely diffusing small molecules such as urea. Therefore, the initial conditions of this abiotic experiment approximate those that occur in the biological symbiosis of *Symbiodinium* and clams or corals.

We observed very different initial diffusive flux of glycerol from inside the dialysis tubing depending on the composition of the fluid external to the tubing. When the dialysis bag containing glycerol was immersed in water alone, the initial flux of glycerol into the external fluid (defined here as the average flux measured during the first 50 min of the experiment) was 2.18 mmol m\(^{-2}\) s\(^{-1}\) (Fig. 2.2). However, when the external fluid contained 100 mM trigonelline, the initial flux increased five-fold, to 11.5 mmol m\(^{-2}\) s\(^{-1}\). In contrast, when 100 mM trigonelline was initially present both inside and outside the bag, the rate of glycerol flux into the external fluid was near the water-only case (2.83 mmol m\(^{-2}\) s\(^{-1}\)) (Fig. 2.2b). Complete data are presented in the Supplementary Information.
Figure 2.2. **HRF activity in dialysis assay.**

**a)** Glycerol concentration as a function of time in a dialysis experiment. Glycerol was placed inside sealed dialysis tubing and HRF candidate molecules were placed in the solution external to the tubing. Average flux (units of mM m\(^{-2}\) s\(^{-1}\)) was calculated for the first 50 min. of each experiment and is indicated by dashed lines on each dataset. Red circles indicate the condition of 100 mM trigonelline in the external solution. Dark blue circles indicate the condition of water only in the external solution. **Inset of a)** Trigonelline diffusion out of the dialysis tubing as a function of time. The diffusion of trigonelline from the tubing was much slower than that of the other molecules considered here. Accordingly, we considered trigonelline movement to be negligible during the first hour of equilibration.

**b)** Control experiments for glycerol migration. The experiments shown here are similar to those in panel a, measuring 100 mM trigonelline in addition to glycerol inside the dialysis tubing (yellow circles), or 250 mM sodium chloride in the external solution (light blue circles). The resulting fluxes (units of mM m\(^{-2}\) s\(^{-1}\)) are indicated by dashed lines.

We also tested whether a fluid of high ionic strength alone would increase glycerol flux. When 100 mM sodium chloride was present in the external fluid, the initial rate of glycerol efflux was 2.48 mmol m\(^{-2}\) s\(^{-1}\), slightly higher than water alone, but 4.6-fold less than in the trigonelline treatment (Fig. 2.2b). These data are consistent with our observation of low HRF activity for sodium chloride in the algae assay described above, and suggests that that result is not due to osmotic stress or ionic strength.
We monitored glycerol concentration in the external fluid of these cell-free systems for the first 2 h of their equilibration, and then again after 48 h. Although the initial glycerol fluxes were very different between treatments, the glycerol concentration in the external fluid after 48 h was the same in each treatment (~1 mM), consistent with the eventual equilibrated concentrations of both glycerol and trigonelline on either side of the dialysis membrane. Since trigonelline can also diffuse across the membrane, albeit more slowly than glycerol (Fig. 2.2a, inset), the final equilibrium point of each of these systems was therefore not altered by the presence of trigonelline.

Some studies have reported that HRF is thermally unstable.\textsuperscript{13–15,17} To determine if the HRF activity of trigonelline was diminished on heat treatment, the LMWF obtained above was heated to reflux for 3.5 h, and we repeated the HRF assay using the refluxed sample. This assay resulted in a net decrease in glycerol concentration of 57 mM, consistent with elimination of HRF activity in the LMWF and diffusion of glycerol from the LMWF back into the algae (Fig. 2.2a). By comparison, heating trigonelline in water alone and then conducting an HRF assay did not alter its HRF activity (see Supplementary Information for details).

Probing the physical origins of glycerol flux

Diffusive flux is defined as the number of molecules passing through a plane per time, and is empirically described by Fick’s law. In general, the constant describing the behavior of a given molecule (the “diffusion constant”) in Fick’s law is an empirical
statement of the chemical potential for diffusion. Diffusion is therefore one example of a chemical process in which a chemical potential and the rate of a process are proportional. Therefore, for a given concentration gradient, if the chemical potential for diffusion changes, then the flux will also change. In the experiments outlined above, we observed that trigonelline and other betaines increase the diffusive flux across the algae or dialysis membranes to different degrees. We simulated the diffusion process using a simple 1-dimensional Monte Carlo model. To gain insight into the origin of the chemical potential for glycerol diffusion toward betaines, and the different potentials induced by different betaines, we performed molecular dynamics (MD) modeling of solutions of betaines and glycerol (see Methods for details).

Figure 3 shows the results of the Monte Carlo simulations compared to the experimental data. The random walk diffusion simulation nicely recapitulates the behavior of glycerol leaking though the dialysis bag into pure water. However, the model is unable to match the results of the dialysis experiment with trigonelline present on the outside of the bag. To account for this behavior, we incorporated an effective increase in chemical potential by biasing the random walk. The motivation for introducing the bias to the model stemmed from my hypothesis for how HRF molecules increase the rate of diffusion of glycerol molecules, and this hypothesis is analogous to osmotic shock. The driving force behind osmotic shock is the hydrogen bonding network in the water. The water molecules on the salty side of a membrane organize around ions and phase transition to a more ordered state. Meanwhile, water molecules pass freely across the membrane in both directions. However, there are less water molecules passing from the
salty side to the pure water side because a considerable fraction of the molecules on the salty side are entrenched in a more robust hydrogen bonding network. This is why we observe a decrease in the chemical potential on the salty side: it is easier to add a water molecule from the pure water side to the salty side because the salty side has a more robust hydrogen bonding network.

Our hypothesis for the HRF Effect was inspired by this concept. The idea was that HRF molecules organize glycerol molecules which effectively lowers the chemical potential. To account for this potential, I included a bias in the random walk to simulate the effect of net flux in one direction across the membrane. If we incorporate a 5% bias (so that 45% of particles move in the negative direction and 55% of particles move in the positive direction) then we achieve a significantly closer match to the experimental data. Notice it only takes a 5% bias to account for a nearly 4-fold increase in rate of diffusion.

To simulate the molecules in atomic detail, we calculated electronic structures to find the dipole moments of the molecules.\textsuperscript{155} We calculated water and glycerol to have similar dipole moments of 2.21 and 2.38 D, respectively, in good agreement with literature values for these molecules (Fig. 2.3a).\textsuperscript{156} In comparison, the betaines glycine betaine, trigonelline and taurine had dipole moments of 8.69, 16.9, and 23.5 D, respectively (Fig. 2.3a). We used the literature value of 3.4 D for sucrose.\textsuperscript{157} Therefore, the least polar betaine we studied had a dipole moment about four-fold that of water or glycerol\textsuperscript{156} while trigonelline and taurine were significantly more polar, with dipole moments eight-fold and ten-fold greater than water or glycerol.
Figure 2.3. Electrostatics and water polarization of glycerol and betaines, and their relationship to the HRF effect. a) Calculated LREPs at the interface of experimental molecules and bulk water. Two projections of each molecule are shown for clarity. Negative potentials are color-coded red while positive potentials are color-coded blue, with the magnitude of the potentials indicated by the color bar at right. The vector extending from the center of each molecule shows the relative magnitude of the calculated molecular dipole for each molecule in arbitrary length units. The magnitude of the dipole in Debye is indicated beneath each structure. b) Relationship between calculated molecular dipoles in Debye, and the HRF activity of each molecule as indicated by the HRF activity assay results shown in figure 2. The exponential fit and corresponding $R^2$ between the two quantities are shown on the figure.
We also evaluated how water equilibrated and organized around these dipolar molecules at the interface between the molecule and bulk water in the simulation. A proxy for the degree of water organization at a point is the long-range electrostatic potential (LREP) at that point, so to probe water organization in the presence of these molecular dipoles, we calculated LREPs on a grid defined on the instantaneous interface between the molecules and bulk water (Fig. 4a). The dipole present in glycerol had a modest organizing effect on water, with magnitudes of the electric potential around the glycerol molecules due to water organization of around ±0.06 pV. Water was more organized around glycine betaine, with potential magnitudes of ±0.11 pV, while trigonelline and taurine exhibited pronounced water organization, with potential magnitudes of ±0.14 pV and ±0.27 pV, respectively (Fig. 4a).

There was a tight relationship between the square of the calculated dipole moment of a given betaine and its activity in the algae-based HRF assays described above and reported in Figure 2. A regression relating molecular dipole of the compounds we tested to the final glycerol concentration observed in our algae-based assay showed a relationship of $A(d)=0.40d^2+1.1d-8.1n$ ($R^2 = 0.95$), where $A$ is the maximum observed glycerol concentration in µM, and $d$ is the molecular dipole moment of a given betaine in Debye (Fig. 4b).

We then simulated a system with two glycerol molecules and different concentrations of trigonelline in water and utilized an adaptive biasing force (ABF) approach to characterize the interactions between two glycerol molecules. The size of the simulated box was such that two trigonellines represent a concentration of about 20 mM.
This arrangement allowed us to determine the glycerol-glycerol radial distribution function (RDF) (see Methods for details). The RDF finds the frequency that glycerol is located at a given radial distance from a focal glycerol located at the origin, and then normalizes this frequency to the expectation for an ideal gas, converging to one at the separation at which glycerol pairs no longer influence each other’s position. The RDF will also therefore be related to the chemical potential for glycerol diffusion, as it captures information about both the enthalpy and the entropy of glycerol-glycerol interactions on the side of the system external to algal cells.

The pairwise behavior of glycerol molecules, as indicated by the RDF, changed markedly with trigonelline concentration. With no trigonelline in the system, the first peak of the glycerol pair-distribution function due to direct hydrogen bonding between glycerols occurred at 6 Å with a value of 1.4, and a second broader peak was present at ~9.5 Å, due to glycerols hydrogen bonded through a single water molecule, had a value of 0.8 (Fig. 2.4a). When trigonelline was present in the system, the RDF changed markedly. In the presence of trigonelline, peaks in the glycerol-glycerol RDF systematically increased in magnitude in a trigonelline concentration-dependent manner. The maximum values of the glycerol-glycerol RDF when glycerol was present at 20 mM concentration were observed with 100 mM trigonelline in the simulation; in this condition, the first peak of the RDF had a value of 1.75 and the second peak had a value of 1.2, reflecting a 25% and 50% increase in the number density of glycerols in the directly hydrogen bonded and single-water bonded positions, respectively, compared to no trigonelline (Fig. 2.4a). However, when trigonelline concentration in the simulation
was greater than 100 mM (250 mM), glycerol interactions became more disordered, with an RDF first peak value of 1.2 and no distinct second peak, such that in this condition, hydrogen bonding both directly and through water were less likely. The glycerol organization in these simulations was also therefore congruent with our experimental observations in the HRF activity assay of maximum HRF activity at 100 mM of trigonelline, and decreased HRF activity with 250 mM trigonelline present (Fig. 2.1). We also simulated different concentrations of glycerol with no trigonelline present, in order to determine the concentration at which glycerol-glycerol interactions in the presence of trigonelline approximated those without trigonelline present. We found that we had to raise the concentration of glycerol in the trigonelline-free simulation to 145 mM in order for glycerol to interact as frequently as 20 mM glycerol in the presence of 100 mM trigonelline.
Figure 2.4. Glycerol-glycerol interactions in the presence and absence of trigonelline. a) Glycerol-glycerol radial distribution functions with different concentrations of glycerol and trigonelline. Y-axis displays $g(r)$ for glycerol–glycerol spatial distributions, x-axis displays intermolecular distance in Å. b) Representative frames from MD simulation of glycerol in solution, and glycerol in solution with trigonelline, with electrostatic potentials calculated at the glycerol-bulk water interface. **Left panel** shows points at which LREPs were calculated at the interface between glycerol and bulk water as individual spheres color-coded by the resulting potential. Negative potentials are coded red, while positive potentials are coded blue, with intensity indicating the magnitude of potential according to the color bar at right. **Center panel** shows glycerol in the presence of trigonelline. Spheres show points and magnitudes at which electrostatic potentials were calculated in the vicinity of glycerol, as in the left panel. Locations and orientations of trigonelline are shown by vectors. Overall magnitudes of electrostatic potentials in both the positive and negative directions are greater than in the glycerol-only case in the left panel. **Right panel** shows detail of an interaction between three h-bonded glycerols and three trigonellines in solution, represented in the same manner as the previous two panels. The bulk polarization of water around each
glycerol molecule orients individual glycerols in a manner that promotes hydrogen bonding between them, either directly or through single water molecules.

Given that glucose also experiences an increase in flux out of cells in our assays, we also simulated glucose-glucose interactions in the presence and absence of 100 mM trigonelline. The frequency of direct glucose-glucose interactions decreases by about 20% when 100 mM trigonelline in the system (Fig. S2). However, there is a direct interaction between glucose and trigonelline, and our results show that trigonelline can both directly hydrogen-bond glucose, and the two molecules can also form a stable interaction via a single water molecule. This is in contrast to glycerol, where we observed little or no direct interaction between glycerol and trigonelline. Therefore, glucose-glucose interactions are likely reduced in the presence of trigonelline due to trigonelline outcompeting glucose for hydrogen-bonding interactions. This is likely a source of the observed chemical potential for glucose migration out of cells in our experimental results.

To gain insight into the physical basis for the increased frequency of glycerol-glycerol interactions in the presence of trigonelline, we simulated 100 mM glycerol in water with and without 100 mM trigonelline, and calculated LREPs in the water shells surrounding individual glycerol molecules (Fig. 2.4b). As described above, in the strongly polar and screening medium of water, LREPs probe both the local and the global degree of water organization in a system. With many glycerols present in the system, the maximum LREP local to a given glycerol increased relative to that of an isolated glycerol, with maximum potentials in the 100 mM system of ±~0.4 pV (Fig. 5b). However, when 100 mM trigonelline was present in the system with 100 mM glycerol,
the LREPs adjacent to glycerol molecules increased even further, tripling relative to the trigonelline-free experiment, with magnitudes of $\pm 1.29$ pV (Fig. 2.4b). The molecular orientations in these simulations suggests two ways in which the presence of trigonelline orders water and thereby increases the local bonding and organization of glycerol pairs in the system (Fig. 2.4b). First, we observed that in local clusters of glycerols near a few trigonellines, the increased organization of water in a shell around glycerol favors the polarized orientations of both water and glycerol that are required for hydrogen bonding two glycerols together via water. This increased propensity for hydrogen bonding between glycerols through one water molecule is reflected by the increase in magnitude of the peak at 9 Å in the RDF. Second, this increased polar organization of glycerol with respect to itself increases the likelihood that a pair of glycerols will become directly hydrogen bonded, since as described, they spend more time in the polarized, ordered orientation that favors this interaction.

**Discussion**

We isolated the betaine molecule trigonelline from giant clam tissue, and demonstrated that several betaines elicit the HRF effect. Further, we showed that the magnitude of the HRF effect increases as the square of the magnitude of the electric dipole of a given betaine.
Our experiments also indicate that the betaine-induced HRF effect is a purely physicochemical phenomenon, in which strongly dipolar molecules in the host generate a strong chemical potential for glycerol and glucose diffusion out of algal cells. The cell permeability of glycerol is high, comparable to urea. As long as animal cells maintain a “sink” for glycerol in the form of respiration, this elevated rate of diffusion out of algal cells and into animal cells could be maintained indefinitely. The chemical potential that causes this high rate of diffusion is related to long-range organization of water around the extended electric fields produced by the large betaine charge dipoles. When water is polarized throughout the system, which occurs at densities of ~100 mM betaine, intermolecular orientations are favored that increase the likelihood both of hydrogen bonding directly between glycerols, and of glycerol-glycerol hydrogen bonding that is mediated by single water molecules. An HRF mechanism mediated by water organization is also consistent with glycerol’s identity as the small metabolite whose release is stimulated by HRF molecules: any metabolites migrating in response to an organized aqueous field must resemble water in both polarity and geometry, which is true for glycerol but few or no other small sugars in cells. Correspondingly, water is essential to mediating this HRF ordering effect, and the interaction that contributes most to the chemical potential for glycerol diffusion from algal cells is that of two glycerols paired through a single water molecule (the second peak at ~9Å in the RDF).

The frequency of interactions of two glycerols in simulation (density of ~20 mM) in the presence of 100 mM trigonelline was about the same as those in a 150 mM solution of glycerol only (Fig. 2.4a). In the context of the symbiosis, this organization of the
glycerol molecules outside of algal cells will create small, isolated pockets of high glycerol concentration, “hiding” the already-migrated molecules, and maintaining a potential for further diffusion of glycerol out of algal cells. Similarly, interactions between glycerols will effectively double their molecular weight, further discouraging retrograde diffusion back through the algal cell membrane. Another way of stating this result is that the presence of trigonelline creates both increased enthalpy and decreased entropy for glycerol-glycerol interaction in the host, maintaining an energetic potential for diffusion. Given that HRF activity varies as a function of betaine structure and the associated dipole moment, and many betaines have been identified in bulk clam and coral tissue, our results suggest the hypothesis that different betaines, if spatially distributed within the host’s body, could create a mechanism for photosynthate distribution throughout the host, analogous to oxygen distribution via spatial gradients of hemoglobin and myoglobin within an organism.

An order-of-magnitude estimate of cell metabolic rate (a cell with volume of \( \sim 500 \mu m^3 \) requires \( 2 \times 10^8 \) ATP s\(^{-1} \) and a chemical energy conversion rate of 10 ATP glycerol\(^{-1} \), combined with our previous geometric characterization of giant clam tissue (\( 2.7 \times 10^7 \) algal cells cm\(^{-3} \) with cell radius 10 \( \mu m \)), shows that the rate of glycerol efflux from algae caused by the presence of 100 mM trigonelline in animal tissue (11.5–2.18 mmol m\(^2\) s\(^{-1}\)) is sufficient to supply the basic energy needs of animal cells over a 24-hour period with 8 hours of active photosynthesis.\(^{158–160}\) This estimate assumes that algal cells maintain an internal glycerol concentration of 25 mM via active photosynthesis, though in general, any rate of translocation via this mechanism will vary with concentrations internal to the
algal cell, and the rate at which glycerol is consumed by animal cells. In any case, the glycerol fluxes we measured in the presence of trigonelline are sufficient to supply the energetic needs of the animal side of the symbiosis over a wide range of plausible algal glycerol concentrations.

The result of a tight relationship between the square of the molecular dipole and the glycerol concentration in algal culture after eight hours is intriguing ($R^2 = 0.95$) (Fig. 2.3b). We speculate that this relationship reflects a systematic expansion of the electric field lines around individual molecular dipoles with dipole strength, and a corresponding volume around which waters and therefore glycerols can be organized around a single HRF molecule. This subject is complex, however, and further work will be required to understand the exact physical relationship between dipole strength, water organization and chemical potential for glycerol diffusion. Similarly, we speculate that this phenomenon is related to our result that the HRF effect decreases at concentrations greater than 100 mM trigonelline. In this picture, increasing concentration to the point where field lines between individual HRF molecules begin to interfere with each other may result in increasing water and glycerol dipolar disorder to the system relative to the case where field lines do not interfere.

Our data indicate that glycerol flux in the presence of dipolar betaines is primarily due to organization of water and subsequent organization and increased bonding interactions of glycerol. The effects of the degree of water organization in a system can be large and unpredictable; for example, organization of water due to confinement in carbon nanotubes can induce a voltage of 17 mV and thereby generate flow.\textsuperscript{103}
As others have noted, the existing literature aimed at identifying HRF molecules is inconsistent and often contradictory. Our model can account for most of the experimental observations recorded in earlier publications. For example, the association of HRF activity with low molecular weight betaines is consistent with the determination that the HRF activity of the coral *P. versipora* is contained in a low molecular weight fraction (<1 kDa) of homogenate. Additionally, several researchers have observed that boiling host homogenate abolishes HRF activity. This is in agreement with our determination that heating trigonelline in the chemical context of the LMWF resulted in decomposition of trigonelline and a reduction in HRF activity. We speculate that thiols or other nucleophilic species present in the LMWF may affect demethylation of trigonelline to generate a neutral species (nicotinic acid) that would not be expected to exhibit HRF activity. In addition, it has also previously been noted that HRF activity is dependent upon active photosynthesis, an observation that we recapitulated here and which is consistent with our model.

Significantly, Muscatine noted that HRF activity is not host-specific to a given algal subtype: clam homogenate can induce photosynthate release from coral zooxanthellae, and vice versa. The widespread prevalence of betaines in both clams and coral is in agreement with this observation. Finally, it has been suggested that HRF is a signaling peptide with an acidic isoelectric point. We would expect the betaines discussed here to have zero net charge at physiological pH (and therefore an acidic isoelectric point), also consistent with this prior result.
Glucose has also been reported as a directly exchanged photosynthate between the algal and animal parts of the photosymbiosis.\textsuperscript{150} We observed a 16-fold increase in the glucose concentration in the medium when trigonelline was present, suggesting that betaines create a potential for diffusion for this molecule as well. Our molecular dynamics simulations suggest that for glucose, the chemical potential for diffusion may arise from more typical direct bonding interactions between glucose and trigonelline, especially given that glucose–trigonelline interactions seem to be more favorable than glucose–glucose interactions.

Seemingly at odds with our model is an earlier publication that suggested free amino acids possess HRF activity, with glutamic acid and aspartic acid displaying the largest effect in an ex vivo model.\textsuperscript{21} However, a subsequent study showed that a mixture of free amino acids equivalent to those in the low molecular weight fraction of \textit{P. versipora} did not stimulate photosynthate release,\textsuperscript{22,163} and consistent with this, a separate study finding free amino acids stimulated photosynthesis but not photosynthate release.\textsuperscript{24} Additionally, Muscatine reported that up to 43.4 mM of external glycerol does not impede photosynthate release from zooxanthellae. However, in our model, the external concentration of glycerol at which efflux from algae ceases and then reverses direction will depend both on the concentration of algae and the concentration of HRF in the medium, and neither of these quantities were specified in the earlier study.\textsuperscript{15} Finally, it was reported that the active substance of \textit{P. versipora} did not pass through membrane filter with nominal 10 kDa cutoff, and that the release-inducing activity of \textit{P. versipora} extract did not decrease when the extract was dialyzed against seawater for 6 h at 4 °C;\textsuperscript{16}
to our knowledge, these are the only related observations in the existing literature for which we cannot account.

**Methods**

Nicotinic acid, trigonelline hydrochloride, glycine betaine hydrochloride, taurine, and glycerol were obtained from commercial sources. Alanine betaine hydrochloride and proline betaine hydrochloride were prepared according to literature procedures.

Proton nuclear magnetic resonance spectra ($^1$H NMR) were recorded at 500 MHz at 24 °C, unless otherwise noted. Chemical shifts are expressed in parts per million (ppm, δ scale) downfield from tetramethylsilane and are referenced to residual protium in the NMR solvent (HDO, δ 4.79). Data are represented as follows: chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet and/or multiple resonances, br = broad), integration, coupling constant in Hertz, and assignment. Proton–decoupled carbon nuclear magnetic resonance spectra ($^{13}$C NMR) were recorded at 125 MHz at 24 °C, unless otherwise noted. Chemical shifts are expressed in parts per million (ppm, δ scale) downfield from tetramethylsilane. Analytical ultra high–performance liquid chromatography/mass spectrometry (UPLC/MS) was performed on a Waters UPLC/MS instrument equipped with a reverse–phase C$_{18}$ column (1.7 µm particle size, 2.1 × 50 mm), dual atmosphere pressure chemical ionization (API)/electrospray (ESI) mass spectrometry detector, and a photodiode array detector. Samples were eluted with a
linear gradient of 20% acetonitrile–water containing 0.1% formic acid → 100% acetonitrile containing 0.1% formic acid over 3 min, followed by 100% acetonitrile containing 0.1% formic acid for 1 min, at a flow rate of 800 µL/min. Analytical high-performance liquid chromatography (HPLC) was performed on an Agilent HPLC instrument equipped with a hypercarb column (3 µm particle size, 2.1 × 50 mm) and a photodiode array detector. Samples were eluted with a gradient from 100% water containing 0.1% formic acid → 20% acetonitrile–water containing 0.1% formic acid over 15 min → 100% acetonitrile containing 0.1% formic acid over 5 min, followed by 100% acetonitrile containing 0.1% formic acid for 5 min, followed by grading to 100% water containing 0.1% formic acid over 2 min. Semi-preparative HPLC was performed on an Agilent HPLC instrument equipped with a hypercarb column (3 µm particle size, 4.6 × 50 mm) and a photodiode array detector. Samples were eluted with a gradient from 100% water containing 0.1% trifluoroacetic acid → 20% acetonitrile–water containing 0.1% trifluoroacetic acid over 30 min → 100% acetonitrile containing 0.1% trifluoroacetic acid over 10 min, followed by 100% acetonitrile containing 0.1% trifluoroacetic acid for 10 min, followed by grading to 100% water containing 0.1% trifluoroacetic acid over 4 min.

**Algal culturing technique.** *Symbiodinium* were obtained from Prof. Mary Alice Cornforth, Department of Geology, State University of New York at Buffalo. Medium for algal cultures was prepared by adding Guillard’s (F/2) marine water enrichment solution (20 mL, Sigma Aldrich) to an autoclaved solution of Marineland® instant ocean salt (36.5 g) in deionized water (1 L). All subsequent dilutions of clam mantle tissue and algal cultures were performed using this same culture medium. All algae cultures were
kept at 24 °C. Algae were initially grown in 30-mL cultures in autoclaved 125-mL Erlenmeyer flasks equipped with cotton stoppers under growth light on a 14 light:10 dark light cycle. After 28 d of growth, the supernatants of four of these culture flasks were decanted and the cells in each culture were resuspended in 5.0 mL of medium. These cell suspensions were combined in one 50-mL conical tube and centrifuged (1000 rpm, 5 min). This supernatant was also decanted and 12 mL of medium was added to the conical tube. The resulting suspension was divided into 1-ml aliquots into twelve autoclaved Erlenmeyer flasks equipped with cotton stoppers. Each Erlenmeyer flask was diluted with 29 mL medium and cells were kept as described above.

*Determination of glycerol concentration (general procedure):* The concentration of glycerol was determined using a commercial glycerol concentration assay kit based on a glycerol kinase–glycerol phosphate oxidase enzymatic assay (Sigma–Aldrich catalog # MAK117) according to the manufacturer’s instructions.

*Determination of glucose concentration (general procedure):* The concentration of glucose was determined using a commercial glucose concentration assay kit based on a glucose oxidase enzymatic assay (Sigma–Aldrich catalog # GAGO20) according to the manufacturer’s instructions.

*Preparation of clam homogenate.* *Tridacna derasa* were purchased from Flying Clams International (Koror, Palau). The clams were approximately 7 years old with shell lengths 30–45 cm. Mantle tissue was dissected from the clam and frozen in in ~35 g portions in 50 ml tubes at −78 °C. In our labs at Yale University, clam tissue was thawed and homogenized with 60 mL of medium (60 mL) using an immersion blender. The
resulting crude homogenate was evenly divided into two 50-mL conical tubes and centrifuged (3500 rpm, 15 min), resulting in a pellet of structurally intact *Symbiodinium* cells and collagen-like clam connective tissue, and a cloudy, opaque supernatant. This supernatant was decanted and then subjected to an additional round of centrifugation (3500 rpm, 15 min). This preparation constituted the clam homogenate used in subsequent fractionation steps.

**Spin column size filtration.** To separate the clam homogenate into a low-molecular weight fraction (<3 kD) and a high-molecular weight fraction (>3 kD), the supernatant of the clam homogenate was evenly divided into eight Macrosep® Advanced Centrifugal Devices (3 kDa MWCO, Pall Corporation, MAP003C37) (these were washed sequentially with 0.05 N aqueous sodium hydroxide solution (4 × 10 mL) and medium (2 × 10 mL)) according to manufacturer instructions). The low- and high-molecular-weight fractions were collected separately, and concentrated to dryness via lyophilization. The resulting dry powder from each was dissolved in 10 mL medium to provide the high-molecular weight fraction (HMWF) and the low-molecular weight fraction (LMWF).

**High-performance liquid chromatography (HPLC) analysis.** A ~25 µL aliquot of the LMWF was subjected to HPLC analysis using a Waters HPLC instrument equipped with a photodiode array detector and a semi-preparative hypercarb column (3 µm particle size, 4.6 × 50 mm). The samples were eluted with 100% water containing 0.1% trifluoroacetic acid initially, grading to 20% acetonitrile–water containing 0.1% trifluoroacetic acid over 30 min (linear gradient), grading to 100% acetonitrile containing 0.1% trifluoroacetic acid over 10 min (linear gradient), followed by 100% acetonitrile.
containing 0.1% trifluoroacetic acid for 10 min, at a flow rate of 4 mL/min. The eluent fractions from 0–15 min, 16–30 min, 31–45 min, and 46–60 min were collected separately, combined, frozen, and concentrated to dryness via lyophilization. The residue from each fraction was redissolved in 1.0 mL of medium and the resulting solutions were employed in the glycerol assays outlined below.

A second aliquot of the LMWF was subjected to the same HPLC protocol described above, and in this experiment, the eluent between 16–30 min was further fractionated by UV-VIS analysis of the signals detected at 290 and 330 nm wavelengths. Eluent from time points of 20.5 min (290 nm), 21.1 min (290 nm), 27.6 min (330 nm), 28.6 min (330 nm), 31.4 min (330 nm), and 38.5 min (360 nm) was collected separately. These fractions were also frozen and concentrated to dryness by lyophilization. The residues obtained were analyzed by nuclear magnetic resonance (NMR) and liquid chromatography-mass spectrometry (LC-MS) studies to determine their molecular structures.

After determining the molecular structure, a standard of trigonelline hydrochloride was obtained from Sigma-Aldrich. The concentration of trigonelline in the LMWF was determined by addition of a known amount of nicotinic acid and comparison of the peak areas against solutions of trigonelline and nicotinic acid of known concentrations.

**Heat stability studies of the LMWF from Symbiodinium.** A 500 µL aliquot of the LMWF obtained as above was concentrated to dryness by azeotropic distillation with anhydrous acetonitrile (10 mL). 1H NMR analysis of the residue obtained confirmed
the presence of trigonelline in the mixture. The NMR sample was diluted with deionized water (1 mL) and the diluted sample was concentrated to dryness by azeotropic distillation with anhydrous acetonitrile (10 mL). The residue obtained was dissolved in 100 µL of deionized water and heated to reflux for 3.5 h. The mixture was concentrated to dryness by azeotropic distillation with anhydrous acetonitrile (10 mL). The residue obtained was studied by 1H NMR analysis. The NMR sample was diluted with deionized water (1 mL) and the diluted sample was concentrated to dryness by azeotropic distillation with anhydrous acetonitrile (10 mL). The residue obtained was dissolved in 500 µL of deionized water for glycerol assays.

A second 100 µL aliquot of the LMWF obtained as above was concentrated to dryness by azeotropic distillation with anhydrous acetonitrile (10 mL). The concentration of trigonelline was determined by 1H NMR analysis of the residue obtained using potassium hydrogen phthalate as the internal standard. A third 100 µL aliquot of the LMWF obtained as above was kept at ambient environment for 7 days. The sample was concentrated to dryness by azeotropic distillation with anhydrous acetonitrile (10 mL). The concentration of trigonelline was determined by 1H NMR analysis of the residue obtained using potassium hydrogen phthalate as the internal standard.

**Assay of glycerol flux from Symbiodinium.** Prior to testing a given set of conditions for HRF activity in *Symbiodinium*, algal cultures were standardized as follows. Cells were washed by pelleting with gentle centrifugation and decanting of the existing medium and resuspension in 5.0 mL of fresh medium, followed by a second round of centrifugation (1000 rpm, 5 min). This pellet of cells was resuspended in 7.5 mL of
medium and the resulting algae suspensions were combined. The cell density of this preparation was determined to be $2.5 \times 10^9$ cells/L via haemocytometer. This suspension was evenly divided into seventy-five 1.5-mL centrifuge tubes. The tubes were subjected to centrifugation (4000 rpm, 30 s) and the supernatants were decanted. Each tube was diluted with 100 µL of the appropriate solution (see Table 1 at the end of the chapter), and the resulting suspensions were transferred to a 96-well plate (wells A2–H11). For algae-free wells, 100 µL of medium was added directly to the 96-well plate (wells A1–D1).

**Glycerol flux assays (conducted in the dark, negative control).** In a separate 96-well plate, 100 µL of 100 mM trigonelline hydrochloride in medium was added to well A1 and three samples of algae suspended in 100 µL of 100 mM trigonelline hydrochloride in medium were added to wells B1, C1, and D1. This plate was protected from light with aluminum foil. The two 96-well plates were placed under growth light for 8 hours. An aliquot (10 µL) was removed from each well and the concentration of glycerol was determined.

**Glycerol flux assays (sucrose control).** The supernatants of two algae cultures were decanted and each culture was resuspended in 5.0 mL of medium. The suspensions were combined, transferred to a 50-mL conical tube, and subjected to centrifugation (1000 rpm, 5 min). The supernatant was decanted and the algae were resuspended in 3.0 mL of medium. The cell concentration was determined to be $3.1 \times 10^9$ cells/L by analysis of an aliquot (10 µL) of this suspension with a haemocytometer. The suspension was evenly divided into three 1.5-mL centrifuge tubes. The tubes were subjected to
centrifugation (4000 rpm, 30 s) and the supernatant was decanted. 100 µL of 100 mM sucrose in medium was added to each centrifuge tube, and the resulting suspensions were transferred to a 96-well plate. 100 µL of 100 mM sucrose in medium was added to an algae-free well. The 96-well plate was placed under growth light for 8 hours. An aliquot (10 µL) was removed from each well and the concentration of glycerol was determined.

**Glycerol flux assays (salt stress control).** The supernatants of two algae cultures were decanted and each culture was resuspended in 5.0 mL of medium. The suspensions were combined, transferred to a 50-mL conical tube, and subjected to centrifugation (1000 rpm, 5 min). The supernatant was decanted and the algae were resuspended in 3.0 mL of medium. The cell concentration was determined to be $2.9 \times 10^9$ cells/L by analysis of an aliquot (10 µL) of this suspension with a haemocytometer. The suspension was evenly divided into three 1.5-mL centrifuge tubes. The tubes were subjected to centrifugation (4000 rpm, 30 s) and the supernatant was decanted. 100 µL of 250 mM sodium chloride in medium was added to each centrifuge tube, and the resulting suspensions were transferred to a 96-well plate. 100 µL of 100 mM sodium chloride in medium was added to an algae-free well. The 96-well plate was placed under growth light for 8 hours. An aliquot (10 µL) was removed from each well and the concentration of glycerol was determined.

**Glycerol flux assays (boiled LMWF control).** The supernatants of four algae cultures were decanted and each culture was resuspended in 5.0 mL of medium. The suspensions were combined, transferred to a 50-mL conical tube, and subjected to centrifugation (1000 rpm, 5 min). The supernatant was decanted and the algae were
resuspended in 6.0 mL of medium. The cell concentration was determined to be 2.3 \( \times 10^9 \) cells/L by analysis of an aliquot (10 \( \mu \)L) of this suspension with a haemocytometer. The suspension was evenly divided into six 1.5-mL centrifuge tubes. The tubes were subjected to centrifugation (4000 rpm, 30 s) and the supernatant was decanted. 100 \( \mu \)L untreated LMWF was added to three centrifuge tubes, and the resulting suspensions were transferred to a 96-well plate. 100 \( \mu \)L boiled LMWF was added to the other three centrifuge tubes, and the resulting suspensions were transferred to a 96-well plate. 100 \( \mu \)L of untreated LMWF and boiled LMWF were added to the algae-free wells. The 96-well plate was placed under growth light for 8 hours. An aliquot (10 \( \mu \)L) was removed from each well and the concentration of glycerol was determined.

**Glycerol assays (preaddition of glycerol to medium).** The supernatants of four algae cultures were decanted and each culture was resuspended in 5.0 mL of medium. The suspensions were combined, transferred to a 50-mL conical tube, and subjected to centrifugation (1000 rpm, 5 min). The supernatant was decanted and the algae were resuspended in 6.0 mL of medium. The cell concentration was determined to be 2.8 \( \times 10^9 \) cells/L by analysis of an aliquot (10 \( \mu \)L) of this suspension with a haemocytometer. The suspension was evenly divided into six 1.5-mL centrifuge tubes. The tubes were subjected to centrifugation (4000 rpm, 30 s) and the supernatant was decanted. 100 \( \mu \)L of 100 mM trigonelline hydrochloride in medium (“trigonelline medium”) was added to three centrifuge tubes, and the resulting suspensions were transferred to a 96-well plate. 100 \( \mu \)L of trigonelline medium was added to an algae-free well. 100 \( \mu \)L of 100 mM trigonelline hydrochloride and 250 mM glycerol in medium (“trigonelline–glycerol
medium”) were added to the remaining three centrifuge tubes, and the resulting suspensions were transferred to a 96-well plate. 100 µL of trigonelline–glycerol medium was added to an algae-free well. The 96-well plate was placed under growth light for 8 hours. An aliquot (10 µL) was removed from each well containing the trigonelline medium and the concentration of glycerol was determined using a glycerol kinase–glycerol phosphate oxidase assay (MAK117, Sigma–Aldrich). An aliquot (10 µL) was removed from each well containing the trigonelline–glycerol medium, diluted with deionized water (9990 µL), and the concentration of glycerol was determined.

**Glycerol flux assays (DMSP).** The supernatants of 6 algae cultures were decanted and each culture was resuspended in 5.0 mL of medium. The suspensions were combined, transferred to a 50-mL conical tube, and subjected to centrifugation (1000 rpm, 5 min). The supernatant was decanted and the algae were resuspended in 3.0 mL of medium. The cell concentration was determined to be $3.2 \times 10^9$ cells/L by analysis of an aliquot (10 µL) of this suspension with a haemocytometer. The suspension was evenly divided into three 1.5-mL centrifuge tubes. The tubes were subjected to centrifugation (4000 rpm, 30 s) and the supernatant was decanted. 100 µL of 100 mM trigonelline hydrochloride in medium was added to three centrifuge tube, and the resulting suspensions were transferred to a 96-well plate in wells B1, C1, and D1. 100 µL of 100 mM trigonelline hydrochloride in medium was added to an algae-free well (well A1). 100 µL of 100 mM DMSP in medium was added to three centrifuge tube, and the resulting suspensions were transferred to a 96-well plate in wells B2, C2, and D2. 100 µL of 100 mM DMSP in medium was added to an algae-free well (well A2). The 96-
well plate was placed under growth light for 8 hours. An aliquot (10 µL) was removed from each well and the concentration of glycerol was determined.

**Glucose flux assays (trigonelline hydrochloride).** The supernatants of six algae cultures were decanted and each culture was resuspended in 5.0 mL of medium. The suspensions were combined, transferred to a 50-mL conical tube, and subjected to centrifugation (1000 rpm, 5 min). The supernatant was decanted and the algae were resuspended in 3.0 mL of medium. The cell concentration was determined to be 3.2 ± 10⁹ cells/L by analysis of an aliquot (10 µL) of this suspension with a haemocytometer. The suspension was evenly divided into three 1.5-mL centrifuge tubes. The tubes were subjected to centrifugation (4000 rpm, 30 s) and the supernatant was decanted. 100 µL of 100 mM trigonelline hydrochloride in medium was added to three centrifuge tubes, and the resulting suspensions were transferred to a 96-well plate in wells B1, C1, and D1. 100 µL medium was added to an algae-free well (well A1). 100 µL of 100 mM trigonelline hydrochloride in medium was added to three centrifuge tubes, and the resulting suspensions were transferred to a 96-well plate in wells B2, C2, and D2. 100 µL of 100 mM trigonelline hydrochloride in medium was added to an algae-free well (well A2). The 96-well plate was placed under growth light for 8 hours. An aliquot (10 µL) was removed from each well and the concentration of glucose was determined.

**Dialysis assay 1: glycerol diffusion in water.** A solution of glycerol in water (2.0 mL, 25 mM) was sealed in dialysis tubing (Spectrum Labs Biotech cellulose ester membrane tubing, 0.1–0.5 kD MWCO, nominal flat width 10 mm, submerged in water (40 mL) in a 50-mL conical tube. The glycerol concentration outside of the membrane
tubing was determined by taking an aliquot (10 µL) of the solution every 10 min for 2 h and subjecting it to the glycerol assay described above. A final aliquot of the exterior solution was also taken after 48 h. This measurement was repeated in triplicate.

**Dialysis assay 2: trigonelline diffusion in water.** Water (2.0 mL) was sealed in a dialysis tube (Spectrum Labs Biotech cellulose ester membrane tubing, 0.1–0.5 kD MWCO). The sealed tube was submerged in a solution of trigonelline hydrochloride in water (40 mL, 100 mM) in a 50-mL conical tube. The concentration of trigonelline inside of the membrane tubing was determined by taking an aliquot (10 µL) of the solution every 10 min for 2 h and subjecting it to UV/VIS analysis. A final aliquot of the solution inside the membrane tubing was taken after 48 h and analyzed by UV/VIS. This experiment was repeated in triplicate.

**Dialysis assay 3: glycerol diffusion into trigonelline solution.** A solution of glycerol in water (2.0 mL, 25 mM) was sealed in a dialysis tube (Spectrum Labs Biotech cellulose ester membrane tubing, 0.1–0.5 kD MWCO). The sealed tube was submerged in a solution of trigonelline hydrochloride in water (40 mL, 100 mM) in a 50-mL conical tube. The glycerol concentration outside of the membrane tubing was determined by taking an aliquot (10 µL) of the solution every 10 min for 2 h and subjecting it to analysis. A final aliquot of the exterior solution was taken after 48 h. This experiment was repeated in triplicate.

**Dialysis assay 4: glycerol + trigonelline mixture diffusion in trigonelline solution.** A solution of glycerol and trigonelline hydrochloride in water (2.0 mL, 25 mM of glycerol and 100 mM of trigonelline) was sealed in a dialysis tube (Spectrum Labs
Biotech cellulose ester membrane tubing, 0.1–0.5 kD MWCO). The sealed tube was submerged in a solution of trigonelline hydrochloride in water (40 mL, 100 mM) in a 50-mL conical tube. The glycerol concentration outside of the membrane tubing was determined by taking an aliquot (10 µL) of the solution every 10 min for 2 h and subjecting it to analysis. A final aliquot of the exterior solution was taken after 48 h. This experiment was repeated in triplicate.

**Dialysis assay 5: glycerol + trigonelline mixture diffusion in glycerol solution.** A solution of glycerol and trigonelline hydrochloride in water (2.0 mL, 1.1 mM of glycerol and 100 mM of trigonelline) was sealed in a dialysis tube (Spectrum Labs Biotech cellulose ester membrane tubing, 0.1–0.5 kD MWCO). The sealed tube was submerged in a solution of glycerol in water (40 mL, 1.1 mM) in a 50-mL conical tube. The glycerol concentration inside of the membrane tubing was determined by taking an aliquot (10 µL) of the solution every 10 min for 2 h and subjecting it to analysis. A final aliquot of the solution inside the membrane tubing was taken after 48 h. This experiment was repeated in triplicate.

**Dialysis assay 6: glycerol diffusion in sodium chloride solution.** A solution of glycerol in water (2.0 mL, 25 mM) was sealed in a dialysis tube (Spectrum Labs Biotech cellulose ester membrane tubing, 0.1–0.5 kD MWCO). The sealed tube was submerged in a solution of sodium chloride in water (40 mL, 100 mM) in a 50-mL conical tube. The glycerol concentration outside of the membrane tubing was determined by taking an aliquot (10 µL) of the solution every 10 min for 2 h and subjecting it to analysis. A final
aliquot of the exterior solution was taken after 48 h. This experiment was repeated in triplicate.

**Monte Carlo Diffusion Simulations:** We created a 1-dimensional Monte Carlo model of the dialysis experiment by numerically integrating a random walk. The random walking particles represent glycerol, and a semi-permeable barrier representing the dialysis membrane is simulated by randomly deciding if glycerol particles can cross the boundary each time they try. Particles were assigned a 25% chance of bouncing off the membrane, and each attempt was calculated independently. The positions of the particles were updated by randomly choosing a number between (0, dx) for each particle, then randomly decide whether to move each particle in the positive or negative direction. This feature of the model was designed to approximate higher dimensional systems which would have particles move in fixed size steps spread out over multiple directions. This model captures the same dynamics but only keeps track of the motion perpendicular to the membrane. The effect of the host release factor is incorporated into the model by biasing the random walk. The biasing mechanism ensures that a fixed percentage of particles are chosen to go in the positive and negative directions pertaining to all particles on the left side of the membrane (inside the dialysis bag).

**Force field calculation:** Force field parameters for glycerol, trigonelline, glycine betaine, and taurine were generated using the CHARMM General Force Field (CGenFF) and the correlated ParaChem server$^{164–167}$. These parameters were then refined using the method outlined by Gumbart et. al.$^{168}$ using the Force Field Toolkit (fFTK) plugin in VMD 1.9.2$^{169}$ in conjunction with Gaussian 09$^{155}$. 
**Standard molecular dynamics:** All atom molecular dynamics simulations were carried out using NAMD 2.11\textsuperscript{170} software and CHARMM36 force field\textsuperscript{171} in TIP3P water.\textsuperscript{111} Langevin dynamics were applied with a temperature of 310 K and pressure of 1.0 atm. Short-range and long-range forces were calculated every 1 and 2 time steps respectively, with a time step of 1.0 fs. Water geometries were constrained to their equilibrium values using the SETTLE algorithm.\textsuperscript{172} Long-range electrostatic forces were taken into account using the particle mesh Ewald approach.\textsuperscript{173} All simulations were performed on in-lab resources.

**Electrostatic Maps:** MD simulations with candidate HRF molecules fixed in place using rigid constraints were run in an identical manner as described above. The simulations were run long enough for a water structure network to form around the molecules. Electrostatic maps were then computed using the procedure described by Remsing et. al.\textsuperscript{143} to visualize the collective polarization of water at the interface between the molecule and the water.

**Adaptive Biasing Force Molecular Dynamics:** An Adaptive Biasing Force (ABF) was applied to simulations using the Colvars module in NAMD.\textsuperscript{174} The reaction coordinate was defined as the distance between the centers of mass of the molecules in the simulation. The reaction coordinate was divided into bins of width 0.05 Å and ranged from 3.0 Å to 20.0 Å, and the force constant was defined as 20 kcal/mol. The method was used to measure the Helmholtz free energy ($A$) as a function of the distance between the molecules. This information was used to compute the radial distribution function, $g(r)$:\textsuperscript{125}

\[
A(r) = -kT \ln g(r)
\]
Table 2.1. Experimental design for the algal HRF assays.

<table>
<thead>
<tr>
<th>well #</th>
<th>algae (+/-)</th>
<th>solution</th>
<th>well #</th>
<th>algae (+/-)</th>
<th>solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>–</td>
<td>medium</td>
<td>E1</td>
<td>–</td>
<td>trigonelline HCl (50 mM)</td>
</tr>
<tr>
<td>B1</td>
<td>+</td>
<td>medium</td>
<td>F1</td>
<td>+</td>
<td>trigonelline HCl (50 mM)</td>
</tr>
<tr>
<td>C1</td>
<td>+</td>
<td>medium</td>
<td>G1</td>
<td>+</td>
<td>trigonelline HCl (50 mM)</td>
</tr>
<tr>
<td>D1</td>
<td>+</td>
<td>medium</td>
<td>H1</td>
<td>+</td>
<td>trigonelline HCl (50 mM)</td>
</tr>
<tr>
<td>A2</td>
<td>–</td>
<td>clam homogenate</td>
<td>E2</td>
<td>–</td>
<td>trigonelline HCl (25 mM)</td>
</tr>
<tr>
<td>B2</td>
<td>+</td>
<td>clam homogenate</td>
<td>F2</td>
<td>+</td>
<td>trigonelline HCl (25 mM)</td>
</tr>
<tr>
<td>C2</td>
<td>+</td>
<td>clam homogenate</td>
<td>G2</td>
<td>+</td>
<td>trigonelline HCl (25 mM)</td>
</tr>
<tr>
<td>D2</td>
<td>+</td>
<td>clam homogenate</td>
<td>H2</td>
<td>+</td>
<td>trigonelline HCl (25 mM)</td>
</tr>
<tr>
<td>A3</td>
<td>–</td>
<td>LMWF</td>
<td>E3</td>
<td>–</td>
<td>trigonelline HCl (10 mM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>B3</td>
<td>+</td>
<td>LMWF</td>
<td>F3</td>
<td>+</td>
<td>trigonelline HCl (10 mM)</td>
</tr>
<tr>
<td>C3</td>
<td>+</td>
<td>LMWF</td>
<td>G3</td>
<td>+</td>
<td>trigonelline HCl (10 mM)</td>
</tr>
<tr>
<td>D3</td>
<td>+</td>
<td>LMWF</td>
<td>H3</td>
<td>+</td>
<td>trigonelline HCl (10 mM)</td>
</tr>
<tr>
<td>A4</td>
<td>–</td>
<td>HMWF</td>
<td>E4</td>
<td>–</td>
<td>trigonelline (250 mM)</td>
</tr>
<tr>
<td>B4</td>
<td>+</td>
<td>HMWF</td>
<td>F4</td>
<td>+</td>
<td>trigonelline (250 mM)</td>
</tr>
<tr>
<td>C4</td>
<td>+</td>
<td>HMWF</td>
<td>G4</td>
<td>+</td>
<td>trigonelline (250 mM)</td>
</tr>
<tr>
<td>D4</td>
<td>+</td>
<td>HMWF</td>
<td>H4</td>
<td>+</td>
<td>trigonelline (250 mM)</td>
</tr>
<tr>
<td>A5</td>
<td>–</td>
<td>HPLC (0 – 15 min)</td>
<td>E5</td>
<td>–</td>
<td>trigonelline (100 mM)</td>
</tr>
<tr>
<td>B5</td>
<td>+</td>
<td>HPLC (0 – 15 min)</td>
<td>F5</td>
<td>+</td>
<td>trigonelline (100 mM)</td>
</tr>
<tr>
<td>C5</td>
<td>+</td>
<td>HPLC (0 – 15 min)</td>
<td>G5</td>
<td>+</td>
<td>trigonelline (100 mM)</td>
</tr>
<tr>
<td>D5</td>
<td>+</td>
<td>HPLC (0 – 15 min)</td>
<td>H5</td>
<td>+</td>
<td>trigonelline (100 mM)</td>
</tr>
<tr>
<td>A6</td>
<td>–</td>
<td>HPLC (16 – 30 min)</td>
<td>E6</td>
<td>–</td>
<td>trigonelline (50 mM)</td>
</tr>
<tr>
<td>B6</td>
<td>+</td>
<td>HPLC (16 – 30 min)</td>
<td>F6</td>
<td>+</td>
<td>trigonelline (50 mM)</td>
</tr>
<tr>
<td>C6</td>
<td>+</td>
<td>HPLC (16 – 30 min)</td>
<td>G6</td>
<td>+</td>
<td>trigonelline (50 mM)</td>
</tr>
<tr>
<td>D6</td>
<td>+</td>
<td>HPLC (16 – 30 min)</td>
<td>H6</td>
<td>+</td>
<td>trigonelline (50 mM)</td>
</tr>
<tr>
<td>A7</td>
<td>–</td>
<td>HPLC (31 – 45 min)</td>
<td>E7</td>
<td>–</td>
<td>trigonelline (25 mM)</td>
</tr>
<tr>
<td>B7</td>
<td>+</td>
<td>HPLC (31 – 45 min)</td>
<td>F7</td>
<td>+</td>
<td>trigonelline (25 mM)</td>
</tr>
<tr>
<td>C7</td>
<td>+</td>
<td>HPLC (31 – 45 min)</td>
<td>G7</td>
<td>+</td>
<td>trigonelline (25 mM)</td>
</tr>
<tr>
<td>D7</td>
<td>+</td>
<td>HPLC (31 – 45 min)</td>
<td>H7</td>
<td>+</td>
<td>trigonelline (25 mM)</td>
</tr>
<tr>
<td>A8</td>
<td>–</td>
<td>HPLC (46 – 60 min)</td>
<td>E8</td>
<td>–</td>
<td>glycine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>B8</td>
<td>+</td>
<td>HPLC (46 – 60 min)</td>
<td>F8</td>
<td>+</td>
<td>glycine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>C8</td>
<td>+</td>
<td>HPLC (46 – 60 min)</td>
<td>G8</td>
<td>+</td>
<td>glycine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>D8</td>
<td>+</td>
<td>HPLC (46 – 60 min)</td>
<td>H8</td>
<td>+</td>
<td>glycine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>A9</td>
<td>–</td>
<td>trigonelline HCl (24.7 mM)</td>
<td>E9</td>
<td>–</td>
<td>alanine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>B9</td>
<td>+</td>
<td>trigonelline HCl (24.7 mM)</td>
<td>F9</td>
<td>+</td>
<td>alanine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>C9</td>
<td>+</td>
<td>trigonelline HCl (24.7 mM)</td>
<td>G9</td>
<td>+</td>
<td>alanine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>D9</td>
<td>+</td>
<td>trigonelline HCl (24.7 mM)</td>
<td>H9</td>
<td>+</td>
<td>alanine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>A10</td>
<td>–</td>
<td>trigonelline HCl (500 mM)</td>
<td>E10</td>
<td>–</td>
<td>proline betaine HCl (100 mM)</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>--------------------------</td>
<td>------</td>
<td>-----</td>
<td>------------------------------</td>
</tr>
<tr>
<td>B10</td>
<td>+</td>
<td>trigonelline HCl (500 mM)</td>
<td>F10</td>
<td>+</td>
<td>proline betaine HCl (100 mM)</td>
</tr>
<tr>
<td>C10</td>
<td>+</td>
<td>trigonelline HCl (500 mM)</td>
<td>G10</td>
<td>+</td>
<td>proline betaine HCl (100 mM)</td>
</tr>
<tr>
<td>D10</td>
<td>+</td>
<td>trigonelline HCl (500 mM)</td>
<td>H10</td>
<td>+</td>
<td>proline betaine HCl (100 mM)</td>
</tr>
<tr>
<td>A11</td>
<td>–</td>
<td>trigonelline HCl (250 mM)</td>
<td>E11</td>
<td>–</td>
<td>taurine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>B11</td>
<td>+</td>
<td>trigonelline HCl (250 mM)</td>
<td>F11</td>
<td>+</td>
<td>taurine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>C11</td>
<td>+</td>
<td>trigonelline HCl (250 mM)</td>
<td>G11</td>
<td>+</td>
<td>taurine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>D11</td>
<td>+</td>
<td>trigonelline HCl (250 mM)</td>
<td>H11</td>
<td>+</td>
<td>taurine betaine HCl (100 mM)</td>
</tr>
<tr>
<td>A12</td>
<td>–</td>
<td>trigonelline HCl (100 mM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B12</td>
<td>+</td>
<td>trigonelline HCl (100 mM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C12</td>
<td>+</td>
<td>trigonelline HCl (100 mM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D12</td>
<td>+</td>
<td>trigonelline HCl (100 mM)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Acknowledgements**
We thank Professor Jason M. Crawford and Professor Alanna Schepartz (Yale University) for use of HPLC and centrifuge instrumentation. We thank Mr. Corey Perez for his assistance in semi-preparative HPLC analysis. Financial support from Yale University (S.B.H.), NSF-INSPIRE, and the Packard Foundation (A.M.S.) is gratefully acknowledged. NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign.

**Author contributions**
X.M. designed and performed algal and dialysis-based glycerol assays, activity isolation experiments, and analytical techniques required for activity isolation. D.F. designed and performed all molecular calculations and simulations. M.T. developed algal culture techniques and assays of clam lysates. S.H. and A.S. jointly conceived the project, performed fieldwork collecting and harvesting clams, supervised the work, and wrote the manuscript.
Chapter 3

Structural Insights from Reflectin Sequences

Reflectin proteins are found only in cephalopod skin cells, and there are no homologous proteins in any sequence database. They are very basic, methionine-rich, and aromatic, and they often lack the hydrophobic residues associated with canonical protein folding.\textsuperscript{1,2} Reflectin proteins are defined by the presence of at least one of two highly conserved motifs. One of these motifs always occurs near the N-terminus of the sequence, which I will refer to as the “N-terminal” motif, and the other motif follows and is typically repeated along the length of the sequence several times (see Fig. 3.1), which I will refer to as the “internal motif”. These motifs are generally defined as:\textsuperscript{1,3}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
N-Terminal Motif: & [MEPMSRTMDFQGRYMDSQGRMVDPRYY] \\
Internal Motif: & [M/FD(X)\_5MD(X)\_5MD(X)\_3/4] \\
\hline
\end{tabular}
\caption{Standard definitions for the N-terminal and internal reflectin motifs.}
\end{table}

79
Examples of the internal motif are shown in Figure 3.2. The function of the motif is not known, but recombinantly expressed versions of the encoded peptide have been shown to readily form thin films on glass. Some secondary structure prediction methods predict that the motifs have a propensity for forming various secondary structure patterns, but the predictions are inconsistent between methods and have low confidence scores associated with them.

The motifs are separated by regions of variable length, which I will refer to as “linkers”. There are no readily obvious patterns or predicted secondary structure patterns that characterize these linkers. However, a paper published in 2010 shows an archetypal reflectin sequence is periodic in a whole-residue interface-affinity scale with respect to sequence residue. We asked the question whether this periodicity is a common feature of reflectin proteins more generally, across transcriptomes and across species.

A simple mathematical procedure for identifying periodicity in data is the autocorrelation function, which is defined as the correlation of data with internal copies of itself. In simpler terms, the autocorrelation function checks for correlation between a
function and a copy of the function shifted by $dx$, $2*dx$, $\ldots$, $N*dx$, where $N$ is the number of points in the dataset and $dx$ is the spacing between them. If the data are perfectly periodic then the correlation will have a large value whenever the shift is equal to the period, because the function will correlate perfectly with the delayed copy of itself. When the delayed copy is shifted totally out of phase with the function, then the autocorrelation function yields a large negative value. Some example periodic functions and their corresponding autocorrelation functions are shown in Figure 3.3.

Figure 3.3: a. Example periodic functions. Blue: sine wave. Green: square wave. Red: triangular wave. Orange: Sawtooth wave. b. The corresponding autocorrelation functions. The sine wave, triangular wave, and square wave all have similar autocorrelation functions. The sawtooth wave is never able to be totally out of phase with itself and therefore has less extreme wells in the autocorrelation curve. c. Random points. d. Autocorrelation of random points. The curve is not periodic and therefore cannot be out of phase, so the function is never negative.
I will begin this chapter by discussing strategies for quantitatively characterizing reflectin sequences. I will provide an overview of how Hidden Markov Models (HMMs)\(^7-9\) can be used to both define and locate motifs in full-length sequences and to quantify the diversity of the sequences.

The second section of this chapter will utilize statistical tools to characterize transcriptome datasets from reflectin-expressing reflective squid tissues. Due to suboptimal assembly, these transcriptome data are highly redundant and full of partial sequences. There are also likely to be misassembled sequences alongside correctly assembled spliced sequence variants in the data. While every sequence of 17,444 total from the transcriptome datasets is unique, that is an implausibly large number to be encoded at individual loci in the DNA of the five species represented, since that number is likely only an order of magnitude less than the total genes encoded by the full genome. The goal of this section is to identify the sequences in our transcriptome dataset most likely to represent distinct DNA loci.

The next section will focus on the development of a more generalized approach to interpreting reflectin sequences by studying the periodic character of the sequences. Finally, I will discuss evidence that reflectin may be intrinsically disordered.
Methods and Materials

**Reflectin Sequence Dataset:** The Sweeney lab has an extensive transcriptome dataset derived from several squid species, including *Loligo pealeii, Loligo opalescens, Dosidicus gigas, Doryteuthis opalescens,* and *Galiteuthi sp.* Tom Dodson, a former graduate student in the group, collected reflectin-bearing tissue samples and sequenced them using Illumina technology (Illumina Inc, CA). For reasons that remain obscure, the raw data are of high quality and assemble well, with the exception of the highly prevalent reflectin sequences, and therefore there are many reflectin sequence fragments in the assembled dataset. I compiled all of the sequences containing a reflectin motif from each species into a single dataset totaling 17,444 sequences and sequence fragments.

I also queried the GenBank\textsuperscript{10} database with the term “reflectin” and manually filtered the results for experimentally validated reflectin sequences by looking at the publications associated with each sequence and checking for the presence of the conserved motifs. The search yielded 25 sequences total, 18 of which came from *Sepia officianalis,* a common cuttlefish, and 7 came from *Euprumna scolopes,* or bobtail squid.

The *reflectin sequence dataset* used for analysis in this chapter was made by combining the sequences from GenBank\textsuperscript{10} with sequences from the transcriptome dataset that represent DNA loci. The latter is described in the flow chart in Figure 3.4, and each step is described in the subsequent sections of the chapter.
Figure 3.4. Graphical representation of the sequence dataset construction. The sequences derived from Illumina were used to create statistical profiles of internal motifs using HMMBuild. These statistical profiles were then used to identify sequences containing both N-terminal and internal motifs in the Illumina dataset using HMMSearch. The resulting sequences were clustering into self-similar groups and a representative was chosen for each cluster. These sequences were combined with the GenBank sequences to form the final dataset.

**Hidden Markov Model Analysis:** Profile Hidden Markov Models (HMM) were originally developed to solve problems in speech recognition. Because they excel at aligning distantly related sequences and identifying commonalities between sequences, they are now used in many areas of science, including bioinformatics. A profile HMM is a probabilistic description of a multiple sequence alignment derived from a position-specific scoring algorithm that is able to capture information about conservation across
entire alignments. Therefore profile HMM analysis is often more sensitive than the more common pairwise methods like BLAST.

I constructed a Profile HMM for the N-terminal motif by compiling the N-terminal motifs listed by Izumi and colleagues, but I added four residues preceding and following their definition to allow for more residues to be included in the statistical profile. I aligned the resulting sequences using Clustal Omega and I used HMMBuild to construct the profile HMM from the alignment.

I constructed a Profile HMM for the internal motif using a similar strategy. However, I used a regular expression to identify all cases where three ‘MD’ repeats are separated by less than 5 residues (in Python, `re.compile('MD.{0,5}MD.{0,5}.MD')`), to extract all sequences containing a motif according to the definition above, and I added eight additional residues preceding and four additional residues following the matched residues. I aligned the potential internal motif candidate sequences using Clustal Omega and I constructed a HMM profile of the internal motifs from the alignment.

The profile HMM was then used to search the set of full reflectin sequences to identify motifs which significantly align with the HMM profiles using HMMSearch. The expectation value (computed by HMMSearch) with a cutoff of 10 was used to filter out subsequences that were not likely to be conserved motifs. The HMMSearch default calibration settings for determining expectation values were used. In the cases where two or more motifs overlapped in a given sequence, the motif with the highest raw HMMSearch score was used for analysis. In cases where two or more internal motifs were identified in a given sequence and there were no overlaps, the residues in between
the motifs were counted as linkers. All sequences containing both an N-terminal motif and at least one internal motif were saved to a master list.

I also built Profile HMMs for three transglutaminase motifs identified by a previous member of the Sweeney lab, Tom Dodson. A region of RNA encoding the transglutaminase domain was found at the beginning, middle, and end of sequences or sequence fragments that also contained reflectin motifs. I used these Profile HMMs to remove all sequences containing transglutaminase motifs from the master list.

The final reduced transcriptome dataset was constructed by searching for sequence fragments that contain an N-terminal motif and at least one internal motif but lack a transglutaminase motif. The HMM profiles were visualized using the SkyLign logo generator.

**Sequence Clustering:** We sorted the sequences in the reduced transcriptome dataset using a clustering algorithm called spectral clustering. This algorithm takes a collection of data points and computes a similarity matrix (defined by the user) and reduces the number of input dimensions by performing a spectral decomposition and representing the data in a new basis with fewer dimensions. This algorithm is commonly applied to sequence alignments using the BLOSUM matrix as the similarity matrix. We used a simpler approach and populated a similarity matrix using the `pairwise2` function from BioPython with the following settings: +1 match, 0 mismatch, -15 opening gap, 0 extending gap, and ignore end gaps. All pairwise distances were evaluated and then sorted into clusters using Scipy’s `SpectralClustering` module with default settings.
Periodicity Analysis: I computed the interface affinity (dG) for the sequences in the reflectin sequence dataset described above. I used the same whole residue interface scale\(^6\) as Izumi and colleagues\(^3\) as a lookup table for each residue in a given sequence, and the resulting curve was smoothed with a sliding window of 19 residues. I also constructed a net charge scale for each residue containing the values 0, +/-1 assigned to residues with formal assigned charges of 0, +1 and -1 respectively.

I used the SciPy \textit{curve_fit} function\(^18\) to fit each \(dG\) curve to a sine wave. I extracted the phase shift from each plot and I used it to align all of curves. Not all sequences were able to be fit to sine waves using the SciPy function, and there were many cases where the fits were not ideal. Therefore I sought out a simpler method of identifying sequences with measurable periodicity in residue character. I computed the autocorrelation of each \(dG\) curve to determine if the sequence exhibits periodic behavior in interface affinity, and considered curves to be periodic if the minimum value of the first well of the autocorrelation curve was below -0.1 (note: this value was chosen somewhat arbitrarily). The period of the curves was determined by the location of the first peak on the autocorrelation function.

IDP Analysis: Intrinsic disorder of compiled reflectin sequences from our transcriptomes was predicted using the Uversky method\(^19\) as outlined in the Introduction chapter. The net charge was determined by summing the formal charge of each residue divided by the total number of residues. The mean hydrophobicity was calculated by taking the sum of the normalized hydrophobicity using the Kyte-Doolittle scale\(^20\) with a window size of 5 and normalized to scale from zero to one of all residues divided by the
total number of residues minus four. If the charge is greater than the right hand side of the following equation then the protein is predicted to be intrinsically disordered.

\[ \text{<Charge>} = 2.785 \times \text{<Hydrophobicity>} - 1.151 \]

**IDP Morphospace:** Sequences predicted to be intrinsically disordered were categorized into five distinct groups by plotting the number of positively charged residues against the number of negatively charged residues,\(^{21}\) as described in the Introduction chapter. The sequences were sorted into the following categories: weak polyampholytes/polyelectrolytes, collapsed or extended (context dependent), strong polyampholyte, or strong polyelectrolyte with +/- charge.\(^{21}\)

**Results**

**HMM Analysis**

The HMMSearch algorithm identified 960, 1778, and 115 sequences that contain at least one N-terminal motif, internal motif, or transglutaminase motif, respectively. The reduced transcriptome dataset contains 404 sequences that contain both an N-terminal motif and at least one internal motif but do not contain a transglutaminase motif.

The HMM profile for the N-terminal motifs encoded in our transcriptome data (Figure 3.5a) shows that the residues at positions 12, 16, 22, and 26 are the most variable, with no single residue having a probability higher than 20%. The rest of the positions show conservation/identity rates above 50% for each residue position (Figure 3.5a). The internal motif is overall more variable than the N-terminal motif, with 9/28 sequence
positions showing low rates of conservation (<20% for any residue type). The HMM profile detected residues that are conserved between 60 and 100% of the time at position 2 (methionine), 5 (proline), 6 (glutamic acid), and 7 (arginine), which, interestingly, are not typically included as part of the definition of the motif\(^1\) (Figure 3.5b). In fact, the proline in the fifth position of the sequence has the highest rate of conservation at 100%.

![Figure 3.4](image)

**Figure 3.4.** Skylign\(^1\) graphics illustrating the relative probabilities of each residue in each sequence position for a: the N-terminal motif, and b: the internal motif. The height of the residue indicates the probability of it occurring there. The letters in each position are made thinner if there are insertions in that position in the sequence alignment.

**Sequence Clustering**

The spectral clustering of the reduced transcriptome dataset yielded 30 distinct clusters of sequences. Each member of an individual cluster is at least 90% similar to the other members of the cluster. Of the 30 clusters, 15 (50%) were periodic in both interface affinity and charge, 2 (7%) were periodic in interface affinity and aperiodic in charge, 13
(43%) were aperiodic in interface affinity and periodic in charge, and there were no clusters found to be aperiodic in both. Overall, 56% of the sequences were periodic in interface affinity and 93% were periodic in charge. A by-eye sequence alignment can be found in the Supplemental Information section at the end of the chapter.22

Reflectin Sequence Statistics

The average N-terminal motif identified in our data contains 26.5 amino acids (Figure 3.6a), which is 1.5 residues shorter than the definition commonly used in literature.1,3 The HMM profile often included two tyrosine residues at the end of the sequence that are not typically included in the formal definition of the motif. The average internal motif we identified contains 26.8 amino acids (Figure 3.6b), which is 6.8 residues longer than the definition used in the literature.3 The additional residues occur at both the beginning and the end of the sequence. The most common motifs begin with methionine at position 2 in Figure 3.6b and end with a charged residue such as arginine (R) or lysine (K) at position 28 in Figure 3.6b.

Most sequence fragments (i.e., sequences within a cluster shorter than the longest sequence in the cluster) contained only one or two internal motifs, but as many as five were found. The average linker contains 35.3 amino acids (Figure 3.6c). The linkers vary in length from 20 to ~100 amino acids. The average full-length sequence is 281 amino acids long.
Figure 3.6. Distribution of the lengths of a. N-terminal motifs, b: internal motifs. c: linkers connecting internal motifs. d: the distribution of the number of internal motifs per sequence.

**Periodicity Analysis**

The raw output of the interface affinity analysis of the reflectin sequence dataset is shown in Figure 3.7a. Each sequence’s interface affinity as a function of residue number that could be fit to a sine wave (for example, see Figure 3.7b) was plotted and aligned by setting the phase shift to 0 (Figure 3.7c). Aligning the curves by the phase shift aligns the underlying sequences so that the conserved motifs align within a two-residue shift of each other. The peaks of each curve roughly correspond to the part of the sequence encoding reflectin motifs (highlighted in pink), and the negative regions, which
have high interface affinities, roughly correspond to reflectin linkers (highlighted in violet). This procedure was repeated with the charge scale, and these relationships are shown in Figure 3.6d.

The periodicity of dG (for both the interface affinity and charge scales) as a function of residue position were evaluated by computing the autocorrelation functions. The autocorrelations are plotted in Fig. 3.7e; the curves that do not have minima less than -0.1 were considered non-periodic. 71% and 79% of these sequences were found to be periodic on the interface scale and charge scale, respectively. The extrema in the dG curves on the interface scale shift in the negative direction toward the C-terminus of the sequence, but the extrema on the charge scale are constant across the sequence. Despite the differences in amino acid composition, length, and sequence patterns in each linker sequence, they each share approximately the same shape on the dG curve in both scales.

I computed the correlation between the interface and charge scales and found they are anti-correlated, meaning the motifs tend to have a negative net charge and a low interface affinity (positive dG), and the linkers tend to have a positive net charge and a high interface affinity (negative dG) (Figure 3.7f).
Figure 3.6. a: Raw output of dG for all sequences in the sequence dataset. b: Example of a sine function fit to dG curve. c: dG curves from interface scale aligned by phase shift. Pink boxes denote locations of motifs in alignment, and violet represents the regions in between them. d: dG curves from charge scale with same color patterns. e: Autocorrelation of dG for all sequences using interface scale. f: Correlation between interface affinity and charge.
IDP Analysis

The relationship of average charge with scaled hydrophobicity\textsuperscript{19} suggests that all full-length reflectin sequences are intrinsically disordered. When each sequence is divided into motifs and linkers we find that these isolated regions also score as intrinsically disordered. The linkers tend to fall farther to the left of line dividing ordered and disordered structures, perhaps suggesting a greater tendency to disorder or random-coil-like behavior. An example sequence that will be used throughout this dissertation was broken into motifs and linkers and plotted in Figure 3.8a, represented by blue and green points, respectively.

The IDP morphospace plot shows all reflectin sequences are predicted to be either weak polyampholytes/polyelectrolytes or special cases where the structures could be in collapsed or extended conformations (Figure 3.8b). Note: two of the motifs occupy the same point on the morphospace plot (on the right), giving the appearance that there are only two blue points. Motifs generally favor the boundary region (light gray) and the linkers occupy the weak polyampholyte/polyelectrolyte region.
Conclusion

The HMM profiles identified motifs consistent with definitions previously published in literature but extended the internal motif definition to include ‘PER’ at the N-terminus. The linker sequences are too variable to be analyzed with HMMs. However, the interface affinity analysis shows a global periodic trend in the full-length sequences; the conserved regions of the sequence have low interface affinity, and the regions linking them have high interface affinity. Moreover, the period of oscillation describing the overall shape of the dG curve is conserved across reflectin sequences, and the shape and depth of the energetic wells in the dG curves are similar, with a variance of 3 in self-similar minima. Therefore the linker regions all share the property of having approximately equal interface affinities of about -6 kcal/mol, which is about two average
hydrogen bonds.\textsuperscript{23} Given there are several regions with this binding affinity per sequence the interaction between linkers and cell membrane, just based on sequence analysis, is likely strong enough to prevent diffusion of the sequence away from the membrane. This observation in turn suggests that the conserved regions of the reflectin motif are cytosolic and the regions linking them are membrane-adjacent (Figure 3.9).

![Figure 3.9. Structural model of the reflectin protein. Reflectin linkers have high membrane binding affinity and reflectin motifs have low membrane binding affinity. Therefore we hypothesize that the linkers anchor the protein to the membrane.](image)

The overall periodic trend in the dG curves makes it possible to perform sequence alignments by fitting the curves to sine waves and aligning the curves by their phase shifts. All of the other sequence alignment tools I tested failed to produce consistent alignments due to the extreme polydispersity in the system. This observation reinforces the idea that reflectins evolve with conserved biophysical properties, even if the primary sequence is not stringently conserved. The sequences that were not able to be fit to sine curves had regions of consistently high dG values on the interface affinity at the C-terminus (see Figure 3.4a for these cases). Two of these sequences strongly resemble a reflectin variant called Reflectin B, which is a special type of reflectin that lacks internal motifs.\textsuperscript{3}
Our analyses predict that, based on their amino acid composition, reflectins are intrinsically disordered, an observation that is also consistent with experimental measurements of secondary structure\textsuperscript{4,5}. Both the conserved motifs and the linker regions are predicted to be intrinsically disordered, but the linkers are predicted to have more exaggerated disordered characteristics due to the lack of hydrophobic residues. The reflectin linkers are predicted to form a highly collapsed state. The motifs fall in the specific region of the IDP morphospace plot where sequences are predicted to have the ability to either adopt collapsed or extended conformations. This could be consistent with a structure that undergoes a conformational change upon associating with a membrane.
Supplementary Information: Reflectin Sequence Dataset.

Sequences labeled ‘Lop’ comes from the Loligo Opalescens transcriptome, sequences label Gali come from the Galiteuthis transcriptome, and sequences labeled Dos come from the Dosidicus Opalescens transcriptome.
Chapter 4

Structural Modeling of the Reflectin Protein

There are no crystal structures of the reflectin protein, and its unusual amino acid composition and material properties make it unlikely to ever crystallize. It has been shown that isolated reflectin in vitro has a high propensity for self-assembling into films and nanoparticles of various sizes, however, the physical origins of the complex, layered, hierarchical dielectric structures that emerge in reflectin-expressing cells have remained elusive. Many studies have attempted to pinpoint the function of the reflectin motifs by constructing recombinant sequences made from several repeated motifs, but this approach assumes the motifs can carry out the self-assembly process without the linkers. There is equivocal evidence from circular dichroism for some stable helix- or sheet-like structure in the conserved sequences, but there is no evidence one way or the other about ordered structure in the linkers.

Intrinsically disordered regions linking structured regions of proteins may be responsible for driving the self-assembly of proteinaceous systems. Dennis et. al. found that a particular linker sequence is responsible for recapitulating the light scattering
behavior of reflectin subjected to vapor pulses. Furthermore, Izumi et. al. showed that a simple whole-residue interface affinity scale predicts the reflectin linkers have a high propensity for interfacial interactions, while the motifs do not. I therefore hypothesize that the disordered sequences between the conserved regions that define reflectin proteins play an important structural role in their self-assembly.

To date, no structural models of the reflectin protein have been published. Protein structure prediction is a powerful tool that can provide insight into the function of proteins when nothing is known other than the sequence, much like the reflectin protein. Despite the recent developments in protein structure prediction technology, the reflectin protein adds several unique complications to an already challenging problem. Full-length reflectin sequences are intractably long for structure prediction. Moreover, reflectin sequences even within a single squid genome are highly variable. Computational constraints require that only one particular sequence can be studied at a time, and it is not clear whether or to what extent the structural model of one sequence is generalizable for the entire family of proteins. Finally, reflectin protein sequences are so unusual compared to any other entries in either DNA or protein sequence or structural databases that they are not ideally suited for comparative modeling or fragment based prediction techniques. These problems are exacerbated by the lack of predicted secondary structure, which for canonically folding proteins can provide a set of structural constraints that are commonly used to improve protein structure prediction.

I used a set of computational models to estimate the structure of reflectin protein. I divided a representative full-length reflectin sequence into subsequences representing
the conserved regions and the unstructured regions linking them, which were conveniently appropriately sized for computational structure prediction techniques. Given the evidence that reflectin is an intrinsically disordered protein (IDP), presented earlier in Chapter 3, I utilized a structure prediction technique that is optimized for proteins that do not undergo canonical folding into compact enzyme-like structures. Rather than search for a single, neatly folded structure at in a narrow well of minimum energy, this IDP-specific method utilizes a Monte Carlo algorithm to fold a set of randomly generated structures to characterize the overall conformational landscape accessible to a protein. The structures can then be analyzed by computing a set of geometric quantities for each structure such as radius of gyration, asphericity, end-to-end distance, and solvent accessible surface area. Each of these scalar geometric quantities then serves as an input dimension in principal component analysis (PCA), reducing the complex structural data to fewer dimensions. The structures are then clustered in principal component (PC) space using K-means clustering, and the centroid of each cluster serves as a representative of the conformational state within a given cluster.

The structures derived from Monte Carlo folding are a rough estimate of physically realistic conformational states, but they need to be carefully refined. Monte Carlo simulations of macromolecules dissolved in water must be simulated with an implicit solvent model, or else the molecules would overlap every time a move is made. The measures required to achieve both a physically realistic and computationally feasible approximation of liquid water necessarily decrease the accuracy of the energetic landscape compared to traditional molecular dynamics simulations with explicit water.
However, traditional molecular dynamics simulations are too time consuming to sample the representative set of conformational states accessible to an IDP. Furthermore, when applied to IDPs, most validated molecular mechanics force fields are known to over-compact structures, over-predict secondary structure, and form spurious stabilizing interactions. However, there are force fields designed specifically to correct these issues and I will utilize one of them in this chapter.

I will utilize the updated Replica Exchange with Solute Tempering (REST2) method to greatly increase conformational sampling efficiency. Finally, I will present the results of molecular dynamics simulations of reflectin interacting with a lipid bilayer to test the hypothesis that linkers associate with cell membranes. Replica exchange methods are often not implemented in systems containing lipids because of the strict temperature dependence of various mechanical properties of membranes. While it is theoretically possible to implement such a method, I used a more direct approach by simulating six independent copies of each system.

**Methods and Materials**

**Sequences Used in Simulations**

I chose one full-length reflectin sequence to represent the complex set described in Chapter 3. The sequence was chosen by systematically iterating through the transcriptome dataset to search for a sequence with an average length and a high degree of periodicity. The sequence I ultimately chose is found in the genome of *Loligo*.
Due to computational resources, only the three motifs and the three linkers closest to the N-terminus of the sequence were considered in this analysis. The following figure shows the full-length sequence with the motifs highlighted in pink and the linkers highlighted in violet:

![Sequence used in this study. Top: Full-length sequence. Middle: Sequence divided into motifs and linkers. Bottom: Interface affinity for chosen sequence.](image)

**Structure Analysis Suite (SAS)**

All of the analyses described in this section are contained in a purpose-written software package that can be found in Appendix X, and on GitHub: https://github.com/dillionfox/IDP_analysis. Python 2.7 and MDTraj were utilized for scripting, atom selection, and coordinate extraction. NumPy was used to optimize calculations and store data, and Matplotlib was used to visualize results.

**SAS.1: Gyration Tensor**
The gyration tensor (S) is a special quantity that describes the second moments of the positions making up an object. $S$ is defined as the sum of the dyadic products of the position vectors in a coordinate system where the center of mass is shifted to the origin (Eq. 4.1).

$$S_{mn} = \sum_{i}^{N} \mathbf{r}_{m}^{(i)} \mathbf{r}_{n}^{(i)} \quad (4.1)$$

The eigenvalues of $S$, $(L_1, L_2, L_3)$, describe the distribution of the objects mass in space, and we can use them to compute the radius of gyration and the asphericity.

The radius of gyration ($R_g$) describes the spatial distribution of an object, and it is mathematically equivalent to the root mean squared distance of all of the objects parts to its center of mass. This quantity is often used to approximate the size of irregular objects, such as polymers. It is defined as the trace of $S$, or, equivalently:

$$R_g = \sqrt{L_1^2 + L_2^2 + L_3^2} \quad (4.2)$$

The asphericity ($\Delta$) is a normalized parameter that describes how similar an objects distribution is to a sphere. For a perfect sphere, $\Delta = 0$. For a perfectly aspherical object (i.e. a straight line), $\Delta = 1$. We compute the asphericity with the following relationship.

$$\Delta = \left< \frac{(L_1 - L_2)^2 + (L_1 - L_3)^2 + (L_2 - L_3)^2}{2(L_1 + L_2 + L_3)^2} \right> \quad (4.3)$$

**SAS.2: N-Terminal to C-Terminal Distance**

The distribution of the end-to-end distances (N-terminal to C-terminal for proteins) is commonly used to classify the overall behavior of polymers. Polymers with regularly spaced backbones and randomly rotated angles have Gaussian distributed end-
to-end distances, and are therefore called Gaussian chains (or random walk polymers). For this to be possible, the interactions between the polymer and the solvent have to be equally favorable as the interactions between the polymer and itself.

The end-to-end distribution of each protein structure was calculated by computing the length of the vector connecting the position of the N-terminal atom to the C-terminal atom.

**SAS.3: Persistence Length**

The stiffness of a polymer (P) can also be estimated from the end-to-end distance distribution. This quantity, called the persistence length, defines the length scale at which the polymer is decorrelated. At length scales smaller than this number, the polymer behaves like a flexible rod; at length scales larger than this number, the polymer behaves like a random walking polymer. Since the quantity is proportional to the Young’s modulus, it is often used to describe the bending rigidity of a polymer. The persistence length was computed using the Polymer Analysis module in MDAnalysis.²¹⁰

**SAS.4: Flory Exponent**

The Flory exponent, $\chi$, is used to establish the scaling relationship between size of a polymer ($R_g$) to the number of bond segments (N):

$$R_g \sim bN^\chi$$

This is a very useful quantity that succinctly classifies polymers. For Gaussian chains, $\chi = 1/2$. For a polymer in a poor solvent (which causes it to collapse into a small conformation), $\chi = 1/3$; for a polymer in a good solvent (which corresponds to an extended conformation) $\chi = 3/5$. 

107
The Flory exponent was estimated by computing the radius of gyration, $R_g$, for all consecutive $n$ residues for $n \in \{3, 4, \ldots, N-3\}$, where $N$ is the number of residues in the chain. The resulting distribution of $R_g$ vs. $N$ can be fit to Eq. 4.4, where $b$ is proportional to molecular weight and $v$ is the Flory exponent. The curve fitting was performed using SKLearn’s $\text{curve.fit}$ code.\textsuperscript{211}

**SAS.5: SASA**

$\text{MDTraj}^{207}$ was used to compute the solvent accessible surface area (SASA) with the $\text{MDTraj}^{207} \ \text{shshake\_rupley}^{212}$ function.

**SAS.6: Secondary Structure**

The secondary structure per residue was computed using the $\text{MDTraj}^{207} \ \text{compute\_dssp}^{213}$ function.

**SAS.7: Protein Structure Fluctuations**

Fluctuations in the protein structures were estimated by computing the root mean squared deviation (RMSD) of the backbone atoms. The structures were first aligned using the $\text{MDTraj superpose}$ function, and then compute RMSD relative to the first frame with the $\text{rm\_sd}$ function.

**SAS.8: Principal Component Analysis**

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of a dataset. The method works by decomposing the covariance matrix of an input matrix where each column represents an axis in real space. The eigenvalues of the covariance matrix are correlated to the amount of variation contained in the corresponding eigenvector. If the input axes are linearly dependent then some of the
eigenvalues will be close to zero which indicates they contain a very small percentage of the total variance in the data. In cases like this, the eigenvalues with the largest value can be chosen to represent the system without losing much variance. The corresponding eigenvectors, which are linear combinations of the input axes, are then used as the basis for the data.

I will use four input dimensions in this section: $R_g$, asphericity, SASA, and end-to-end distances. These four quantities are computed for each structure and therefore represented by a point in 4D space. PCA is then applied using the SKLearn implementation.\textsuperscript{211} Since the input dimensions are highly correlated, the number of input axes can be reduced to 2 or 3 without losing more than 15\% of the variance. The points are then viewed in a 2D space made from the 2 eigenvectors containing the most variance, which I will refer to as PC-1 and PC-2.

If the protein structures find unique, stable energetic minima, then the points will form clusters in PC-space. If no distinct clusters form then the protein structure might be intrinsically disordered. I used SKLearn’s KMeans clustering implementation\textsuperscript{211} to cluster the points in PC-space.\textsuperscript{211} K-means is a commonly used algorithm for sorting an arbitrary number of points into $k$ clusters. The clusters are formed so that the mean distance between each point and the center of the cluster are minimized. The optimal number of k-means clusters was chosen by running k-means on the data set for 2-20 clusters, and the associated score from SKLearn was used to determine the minimum number of clusters to be used. The structure nearest to the geometric centroid of each cluster was extracted to represent the cluster. The structure chosen from the centermost
cluster (along PC-1) was used for further refinement with molecular dynamics simulations.

**SAS.9: Protein-Lipid Contacts**

Protein-lipid contacts were computed using a simple distance cutoff of 3 Å.

**SAS.10: Lipid Density**

The average density of the lipids was calculated along the axis normal to the bilayer using the GROMACS gmx density function. This method is capable of detecting asymmetry in the upper and lower leaflets, the average thickness of the bilayer, and fluctuations in the density of the bilayer midplane caused by interdigitated residues.

**SAS.11: Lipid Order Parameter**

The length and rigidity of lipid tails is dependent on phase behavior. Ordered gel phases have elongated lipids while fluid phases have disordered lipids. The lipid order parameter is computed by summing the cosines of the angles made between the carbons in the lipid tails and it is sensitive to minor perturbations to the lipid tail structure. The lipid order parameter was computed using the GROMACS gmx order function.

**SAS.12: Lipid Mesh Construction**

A mesh object was constructed for each leaflet in the lipid bilayer by mapping the centroid of each lipid molecule onto a rectangular grid in the X-Y plane. The grid spacing was chosen so there would be approximately a one-to-one ratio between lipids in each leaflet to grid points. If multiple lipids were mapped to the same grid point then the values were averaged.
SAS.12a: Lipid Interdigitation

Phase transitions can be detected by measuring the interdigitation of the lipid tails. Interdigitation is measured by calculating the distance between the average height of the distal carbons in the lipid tails (C214 and C314 in the CHARMM36 force field) for each leaflet. The average interdigitation was recorded for each frame, and the local interdigitation of the lipids was calculated by imposing the bilayer onto the lipid mesh object.

SAS.12b: Lipid Bilayer Thickness

The local thickness of the bilayer was computed by mapping the coordinates of the P atoms in the DMPC head groups to the lipid mesh object and subtracting the average value at each mesh point.

SAS.13: Replica Exchange Ensemble Reconstruction

Replica exchange simulations were analyzed by constructing constant temperature ensembles (see Fig 4.2 for description). Rather than analyze time-continuous replicas, the snapshots from the simulations (both Monte Carlo and REST2) are reorganized to form groups of structures of the same effective temperature. In the example shown in Figure 4.2, the 500 K ensemble contains 2 structures from Replica 5, then 4 structures from Replica 4, then 1 structure from Replica 1, and finally 3 more structures from Replica 4. The 300 K ensembles, referred to as the room temperature ensembles, were analyzed using the methods described in the Protein Structure Characterization section below.
Figure 4.2. Simplified diagram illustrating 5 systems (replicas) undergoing replica exchange. Each solid color line makes up a time-continuous trajectory, but the temperature changes over time due to swapping. Note: the figure is not drawn to scale. The temperature swapping is instantaneous when performed in the simulation and therefore the sloped line representing the transition should be nearly vertical. The “room temperature ensemble” analysis is done on whichever frames happen to be in the gray dashed box at 300 K.

**De Novo Structure Prediction**

I generated initial structures for each sequence fragment using CAMPARI, a software package that implements flexible Monte Carlo sampling methods for the study of biopolymers. CAMPARI was used to generate random extended conformations of each sequence. Structures were then evolved using Monte Carlo (MC) moves as determined by the ABSINTH force field. These MC simulations were performed using replica exchange with 16 temperatures (250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 400, 420 K). The simulations were run for 100,000,000 steps, with the first 100,000 steps used for an initial equilibration. The replica simulations were
evaluated for exchange every 2,000 steps. The remaining parameters were chosen consistent with the CAMPARI tutorial. The script I wrote to auto-generate and run simulations with CAMPARI is provided in Appendix X, and all input values can be found in the code.

After the equilibration period, every 10,000th frame was included in further analysis. Secondary structure was computed in parallel using the DSSP algorithm. The room temperature ensemble (300 K) was analyzed using the Structure Analysis Suite described above.

**Molecular Dynamics Simulations Configuration**

All REST2 MD simulations were performed using GROMACS 5.0.2 on the Titan Supercomputer, and all standard MD simulations were performed using GROMACS 2016.3.knl on the CORI supercomputer. The following settings were used in both sets of simulations: Langevin dynamics was applied for temperature 300 K with the velocity rescaling temperature coupling algorithm, which rescales all velocities to match the defined temperature. The equations of motion were integrated using the ‘md’ integration scheme in GROMACS with a 2.0 fs timestep. This method applies a leapfrog algorithm for integrating the equations of motion. Long-range electrostatic forces were taken into account using the particle mesh Ewald approach. Straightforward electrostatic interactions are not guaranteed to converge in periodic systems. However, the Ewald summation approach replaces the direct summation of long-range interactions with a calculation in Fourier space which avoids a singularity in real space. The algorithm rapidly converges and executes very quickly. The recently
released CHARMM36m\textsuperscript{113} force field was also utilized in conjunction with the tip3p water model,\textsuperscript{111} which was shown to work well for both folded and intrinsically disordered proteins. The CHARMM36m force field builds off of the CHARMM36 force field but has improved peptide phi/psi angle conformational energies and corrects a previously overestimated salt bridge interaction.

The room temperature ensemble (300 K) was analyzed using the Structure Analysis Suite described above.

**Standard Molecular Dynamics Simulations: Membrane Interfaces**

The lipid bilayer systems were created using the CHARMM-GUI Input Generator tool. Each bilayer contained 77 DMPC lipids in each leaflet to give a total membrane area of 49 nm\textsuperscript{2} or \textasciitilde7.0 nm on each edge. Each system was solvated with approximately 15,000 water molecules and 150 mM NaCl. The systems were then equilibrated and simulated with GROMACS for 500 ns each. I ran six independent simulations of each reflectin subsequence in contact with these lipid systems. The first simulation was constructed so that the middle residue was approximately 3 Å from the lipid at the center of the system. The rest of the simulations were constructed by iterating the first system, rotating the protein five times in 90 degree increments in two planes, effectively allowing each face of the protein to start out in contact with the lipids. The lipid bilayers were constructed by CHARMM-GUI independently for each system.

The simulations were analyzed using the Structure Analysis Suite described above. All calculations were performed on 500 frames separated in time by 1 ns.
REST2 Simulations

Each starting structure was derived from a snapshot of one of the Monte Carlo simulations described above. The structures were solvated with the GROMACS command `gmx solvate` and ionized with 150 mM NaCl using the GROMACS command `gmx genion`. These structures were then equilibrated for 1 ns to prepare for production scale REST2 simulations. Each simulation was equilibrated for 500 ns, and the subsequent 500 ns were used for analysis. These simulations were analyzed in the same way as the Monte Carlo simulations.

Hamiltonian Replica Exchange simulations\(^{126}\) were prepared using the PLUgin for MolEcular Dynamics package (PLUMED), version 2.3.\(^{131}\) The PLUMED `partial_tempering` code was used to make 10 copies of the parameter file, and then set the effective temperature of each system to ranging from 300–450 K, separated exponentially (See Appendix X for the scripts automating this process). The REST2 simulations were then performed using GROMACS 5.0.2 in conjunction with PLUMED. The CHARMM36m force field\(^ {113}\) was used with TIP3P water.\(^ {111}\) The rest of the parameters were chosen to match the standard MD simulations described above.

The room temperature ensemble (300 K) was analyzed using the Structure Analysis Suite described above.
Results

De Novo Structure Prediction

The Monte Carlo folding simulations of the motif sequences yielded a wide variety of conformations. PCA identified SASA and $R_g$ to contain the most variability of the input dimensions (Figure 4.3a). The Flory exponent for the motifs ranged from 0.38-0.41 (Figure 4.3b), which is consistent with structures that are slightly compact. The majority of the structures had a $R_g$ of ranging from 0.8-0.9 nm (Fig 4.3c). Approximately 20% of the structures generated for Motif 1 (Figure 4.3d) exhibit tertiary contact patterns consistent with alpha helices in residues 3-8, 11-16, and 18-24. Motifs 2 and 3 demonstrated a low propensity (<15%) for forming alpha helices in residues 11-23 (See Figures 4.14 and 4.15 in the SI at the end of the chapter). No obvious clusters formed in PC-space in any of the simulations, although there was a higher density of structures occupying PC-1 in the range of -1 to 0.
The linker sequences also accessed a wide range of conformations without forming distinct clusters in PC-space (Figure 4.4a). The input axes with the highest variability were once again SASA and $R_g$. Most structures were measured to have a $R_g$ between 0.8-1.0 nm (Figure 4.4c), which is slightly larger than the range observed by the motif sequences. However, this quantity does not account for the length of the polymer. The Flory exponent for the linkers, which depends on the number of residues, was estimated to be 0.33-0.35 (Figure 4.3b), which is consistent with highly compact
structures. Despite the qualitative differences in the sequences, each linker was predicted to have a probability of around 20% to form alpha helical structure between residues 3-8. Linkers 2 and 3 had a roughly equal probability of forming an alpha helix between residues 17-24 (See Figures 4.16d and 4.17d at the end of the chapter). There were a few predictions of beta-sheet formation, particularly involving residues 8, 10, 11, 17, and 28 in Linker 2 (See Figure 4.16d at the end of the chapter).

Figure 4.4. Monte-Carlo-generated structures for Linker 1 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP.
**Structure of Reflectin in Solution**

The REST2 simulations of the motif sequences generated less compact structures compared to the Monte Carlo simulations. The average radius of gyration in the REST2 simulations was approximately 0.4 nm larger than the Monte Carlo simulations, which is a 50% increase. PCA found the radius of gyration is one of the most variable parameters of the simulations (Figure 4.5a), ranging from 0.8-2.0 nm (Figure 4.5c). The motif structures generated by this method have size distributions and Flory exponents consistent with structures that are in between collapsed and extended conformational states. The Flory exponents for Motifs 1, 2, and 3 were simulated to be 0.46, 0.39, and 0.48 (Figure 4.5b), respectively, which are in the vicinity of Gaussian chains. I also computed the simulated persistence lengths of the motifs and found they are 0.45 nm, 0.61 nm, and 0.58 nm for Motifs 1, 2, and 3, respectively.

Given that the motif is highly evolutionarily conserved, there was a surprising amount of structural variability predicted by these simulations. Simulations of Motif 1 showed a probability of about 20% of forming beta-sheet contacts between residues 6-7 and residues 11-15 (Figure 4.5d). Motif 2 is more compact than a Gaussian chain but less compact than a polymer in poor solvent. This sequence demonstrates the highest propensity for forming secondary structure (Figure 4.19d) with a 50% probability of forming beta-sheet structure between residues 9-10 with residues 15-16 (see Figure 4.19d). The higher propensity for forming secondary structure may contribute to its smaller size compared to the other motifs (Figure 4.19c). In contrast, Motif 3 demonstrated a low probability for forming beta-sheet contacts between residues 2 and 9,
and in contrast showed a high probability (>50%) for forming helical structure for residues 17-21 (see Figure 4.20d).

Figure 4.5. Molecular Dynamics derived structures for Motif 1 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP

Linker sequences demonstrated very different structural behavior in the REST2 simulations compared to the Monte Carlo simulations. No distinct clusters formed in PC-space, and the input dimensions with the most variability were $R_g$ and SASA (Fig 4.6a). Linker 1 formed almost no secondary structure in the REST2 simulation (Figure 4.6d) and had a Flory exponent of 0.42 (Figure 4.6b), suggesting that the structure is in
between a collapsed state and a Gaussian chain. Linker 2 was a little more compact than Linker 1 and had a Flory exponent of 0.39 (Figure 4.21b), but it also had a high propensity for forming secondary structure (Figure 4.21d). Approximately 70% of structures formed an alpha helix in residues 20-25, consistent with the result of Monte Carlo folding. Nearly half of the structures from the Linker 2 room temperature ensemble had beta-sheet contacts focused around residue 16, making contacts with either residues 10-11 or residues 28-29 (Figure 4.21d). Linker 3 was measured to have a Flory exponent of 0.49 which is consistent with a Gaussian chain (Figure 4.22b).

![Molecular Dynamics derived structures for Linker 1 sequence.](image)

Figure 4.6. Molecular Dynamics derived structures for Linker 1 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log(Rg) vs. log(N). (c): Rg distributions. (d): secondary structure per residue determined by DSSP.

Structure of Reflectin at Membrane Interfaces
The reflectin motif structures have similar distributions of $R_g$ in the membrane simulations and the REST2 simulations. Motif 3 sampled fewer conformational states and favored more compact structures. More than 75% of the structures had a radius of gyration measuring between 0.8-1.0 nm (Figure 4.7). There was a small amount of beta-sheet that formed in all 3 motif sequences, and there was very little alpha-helix formation.

![Figure 4.7](image.png)

Figure 4.7: From left to right, structure for Motifs sequences 1, 2, and 3. Top row: Cumulative $R_g$ distributions. Bottom row: Average secondary structure determined by DSSP. Y-axis is probability, x-axis is residue number.

The linkers also adopted similarly compact size distributions in the simulations containing membranes, as demonstrated by the distributions of $R_g$ (Figure 4.8). Linker 1 formed some short-lived secondary structure patterns (Figure 4.8). Linker 2 rapidly formed stable beta-sheet interactions between residues 10-11 and 16-17 and alpha helical structure in residues 20-25. These secondary structure patterns are also formed in the
REST2 simulation but are not as strong. Linker 3 did not form any secondary structure patterns and had the largest variation in $R_g$.

![Graphs showing cumulative $R_g$ distributions and average secondary structure for Linker sequences 1, 2, and 3.](image)

Figure 4.8: From left to right, structures for **Linker** sequences 1, 2, and 3. Top row: Cumulative $R_g$ distributions from 6 independent simulations. Bottom row: Average secondary structure determined by DSSP from 6 independent simulations. Y-axis is probability, x-axis is residue number.

**Interactions Between Reflectin Sub-Sequences and Membranes**

The motif structures demonstrate an overall moderate propensity for interacting with DMPC lipids (Figure 4.9). Most residues are in contact with lipids in less than 20% of the frames, but some arginine and tryptophan residues are in contact in as many as 40% of the frames. Each sequence had at least one simulation where the protein never made strong contacts with the interface, and at least one simulation where the protein made strong contacts with the interface right away but later dissociated. Motif 3 exhibited the highest rate of protein-lipid interactions (Figure 4.9). Three of the six simulations
with this sequence formed strong contacts with the membrane within the first nanosecond of the simulation.

The residues that make the most frequent contacts in descending order, are arginine, tryptophan, and tyrosine. The simulations with Motif 1 had the fewest contacts overall. The characteristic ‘PER’ and ‘MDM’ regions of the sequence do not show an especially high affinity for interface interactions. The residues associated with high interface affinities (tyrosine (Y) and tryptophan (W)) do not account for the majority of contacts with the bilayer. However, these residues account for a very low proportion of the residues in the sequences. The most common contacts are purely electrostatic in origin, occurring between arginine, which is positively charged, and the negative charge on the zwitterionic headgroups of the lipids. Given not all lipid headgroups have this charge separation, the high frequency of charge-charge interactions may be an artifact of our choice of lipid.

Figure 4.9: From left to right, Motifs 1, 2, and 3. Cumulative contacts per residue from 6 independent simulations each. Normalized by number of frames in all simulations combined.
The initial conditions also strongly influenced the outcome of the simulations with linker sequences. Each sequence was simulated six times, and all six simulations yielded a different frequency of contacts and different residue combinations binding the membrane. Linker 1 had one simulation that never made strong contacts, yet there were two simulations that had contacts in every frame. Linker 2 had contacts in every frame in 4/6 simulations, but the contacts were not always with the same residues (hence the bimodal distribution). Linker 3 had three simulations where half of the frames contained contacts and three simulations where nearly every frame had contacts.

The linker sequences have a very high proportion of tyrosine residues which account for the majority of the contacts in those simulations. Tyrosine has one of the highest scores for interface affinity (second to tryptophan), and therefore the high density of these aromatic residues gives the linkers a significantly higher propensity for associating with the interface. This is predicted both by the interface affinity scale discussed in Chapter 3 and the results of the simulations.

Figure 4.10: From left to right, Linkers 1, 2, and 3. Cumulative contacts per residue from 6 independent simulations each. Normalized by number of frames in all simulations combined.
Membrane Remodeling

The figures in the following sections are taken from one simulation that best represents the group, and the rest of the figures can be found in the end of the chapter (Figures 4.23-4.36). I observed a significant disruption in the density of the lipids in the upper leaflet (which is what the protein touches) in the last 100 ns of the simulations with linker sequences (See Figure 4.11a for an example and Figures 4.31-4.32 for the full results). The displaced lipids from the upper leaflet shifted to the bilayer midplane, which became more dense over time and the density of the layer becomes more homogeneous. There was an appreciable increase in the order parameter when comparing the first 100 ns of the simulation to the last 100 ns (See Figure 4.11b for an example and Figures 4.29-4.30 for the full results).

Figure 4.11. a: Order parameter from 0-100 ns. b: Order parameter from 400-500 ns.

The locally averaged thickness of the bilayer ranged from 2.8 to 4.4 nm (Figure 4.12a). The membrane immediately adjacent to the protein contacts was always thinner than the membrane far from the protein contacts (Figure 4.12b). Note: these values are calculated from the coordinates of the P atoms in the DMPC head groups, so the fluid bilayer thickness should be roughly 3.7 nm.
The lipid interdigitation proximal to protein-lipid contacts showed similar patterns in the average membrane thickness (See Figure 4.13a for an example and Figures 4.35-4.36 for the full results). The average interdigitation decreases as the number of contacts increases (See Figure 4.13b for an example and Figures 4.33-4.34 for the full results). The points are colored by time so the evolution of the system can be tracked. The general trend shows a spike in the number of contacts is typically followed by a decrease in the interdigitation (i.e. the lipid tails become more interdigitated). A time series of this process is shown in Figure 4.13c.
Figure 4.13: a: Average Interdigitation from 250-500 ns. b: Number of residues in contact with lipids vs. average interdigitation (nm). Color represents time in ns. c: Snapshots from simulation. First panel is at approximately 10 ns, second is approximately 100 ns, and third is approximately 300 ns. Dark gray lipids indicate upper leaflet and light gray indicates bottom leaflet.
Conclusion

The conserved regions of the reflected protein are predicted by Monte Carlo simulation to form overall compact structures with a small propensity for forming alpha helices. The structures did not form distinct clusters in PC-space, suggesting that the system has no specific energy minima. Instead, the structures may fluctuate near an equilibrium, without accessing local minima strong enough to overcome thermal fluctuations. These observations are consistent with the prediction that the structures are intrinsically disordered and it is exhibited by all sequences and simulation methods.

The conserved motifs are predicted to be less compact in the REST2 simulations compared to the linkers and have Flory exponents consistent with Gaussian chains. The REST2 simulations of the linkers between the conserved motifs form compact structures consistent with the IDP-morphospace prediction (Figure 3.X) based on the ratio and number of charged residues in the sequences. The Flory exponents suggest the linkers resemble a polymer in a poor solvent. The interactions with itself are more energetically favorable than with water, such that the structures apparently collapse rather than folding to a single state. The linker structures demonstrate a small probability of forming beta-sheets or alpha helices, neither of which are strongly predicted by secondary structure prediction algorithms.

The initial orientation of the protein near the bilayer strongly influenced the outcome in simulations of both conserved motifs and the linker regions connecting them.
In cases where one or more residues were able to form contacts from the starting configuration, the overall number of contacts was greatly increased. The motifs exhibit a low propensity for interacting with lipids and therefore had few opportunities to adopt new conformations in response the interacting with the membrane. However, three of the simulations with Motif 3 formed strong contacts with the membrane within the first nanosecond of the simulation and the contacts persisted for several hundred nanoseconds. The distribution of $R_g$ for these structures show that they compact upon forming contacts with the membrane. The simulations that did not form strong interactions with the membrane showed similar structural properties as the REST2 simulations in bulk water.

The linkers formed persistent, stable contacts with the bilayer and formed compact structures on average. The simulation of the Linker 2 sequence showed a very high propensity for forming beta-sheet structure upon binding the membrane. The same secondary structure pattern was observed in the REST2 simulation but at a probability of 20% compared to the 60% observed at the membrane interface. I hypothesize that the interactions with the membrane stabilized the structure of the protein and allowed it to stabilize in an energetic minimum.

The DMPC bilayers showed strong responses to the interactions with reflectin. The simulations were conducted at 3 °C above the melting temperature of DMPC lipids. Below the melting temperature, the lipids elongate and the headgroups pack into a dense hexagonal structure. This effect is observed in the lipid order parameter, which is significantly increased when reflectin forms stable contacts with the lipids. The lipid density as a function of the direction normal to the bilayer shows the density of the
The bilayer midplane is increased, which can indicate interdigitation, curvature, or the ripple phase.\textsuperscript{218}

In some cases, the bilayers appear to exhibit the “ripple phase” (see Fig 4.12b).\textsuperscript{218} However, I think the ripple-like appearance comes from the fact that the lipids interdigitate right where the protein contacts the lipids, but the rest of the bilayer is in an ordered gel phase. This is reflected both in the local interdigitation and the local average thickness plots. This suggests the reflectin protein initiates a phase transition in the membrane upon binding. To test this, I compared the number of protein-membrane contacts to the average interdigitation, and found that they correlate. The protein and membrane form contacts and the lipids begin to interdigitate within the first 100 ns.

Before the lipids can fully transition to the interdigitated gel state, the lipid headgroups need to increase their packing density and rearrange into a hexagonal lattice. The timing of this process correlates with a decrease in protein-membrane contacts, but once the membrane reaches the steady-state mixed interdigitated phase, the number of protein-membrane contacts increases again.

There are two cohesive forces that favor the ordered gel phase: tail attraction, driven by Van der Waals, and the minimization of the hydrophobic/polar (tail/head) surface area. However, tail entropy prefers the liquid phase and disordered tails have a larger excluded volume and therefore are less confined.\textsuperscript{219} Water decreases the effective tail-head surface tension because it interacts favorably with the heads and therefore decreases the phase transition temperature. I hypothesize that reflectins increase the
lipid/solvent surface tension by inserting polar residues near the heads and hydrophobic residues near the tails.

**Acknowledgements**

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This research also used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.
Supplemental Information

Monte Carlo Structure Analysis

Motif 2

Figure 4.14. Monte Carlo derived structures for Motif 2 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP
Motif 3

Figure 4.15. Monte Carlo derived structures for Motif 3 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of \( \log(R_g) \) vs. \( \log(N) \). (c): \( R_g \) distributions. (d): secondary structure per residue determined by DSSP.
Figure 4.16. Monte Carlo derived structures for **Linker 2** sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of $\log(R_g)$ vs. $\log(N)$. (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP.
Figure 4.17. Monte Carlo derived structures for **Linker 3** sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of $\log(R_g)$ vs. $\log(N)$ (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP
Figure 4.18. Representative structures from REST2 simulations.
Figure 4.19. Molecular Dynamics derived structures for **Motif 2** sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP.

**Motif 2**
Motif 3

Figure 4.20. Molecular Dynamics derived structures for Motif 3 sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP
Figure 4.21. Molecular Dynamics derived structures for **Linker 2** sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log($R_g$) vs. log(N). (c): $R_g$ distributions. (d): secondary structure per residue determined by DSSP.
Figure 4.22. Molecular Dynamics derived structures for **Linker 3** sequence. (a): PCA results colored by k-means clusters. (b) Estimate of Flory exponent from distribution of log(R_g) vs. log(N). (c): R_g distributions. (d): secondary structure per residue determined by DSSP.
Figure 4.23. Distributions of radius of gyration from each motif simulated in the presence of a membrane.
Figure 4.24. Distributions of radius of gyration from each linker simulated in the presence of a membrane.
Figure 4.25. Distributions of end-to-end from each motif simulated in the presence of a membrane. Each distribution was fit to a Gaussian distribution. Many attempts at fitting were unsuccessful.
Figure 4.26. Distributions of end-to-end from each linker simulated in the presence of a membrane. Each distribution was fit to a Gaussian distribution. Many attempts at fitting were unsuccessful.
Figure 4.27. PCA results from each motif simulated in the presence of a membrane.
Figure 4.28. PCA results from each linker simulated in the presence of a membrane.
Figure 4.29. Lipid order parameter results from each motif simulated in the presence of a membrane.
Figure 4.30. Lipid order parameter results from each linker simulated in the presence of a membrane.
Figure 4.31. Lipid density as a function of the axis orthogonal to the membrane surface each simulation containing a reflectin motif simulated in the presence of a membrane.
Figure 4.32. Lipid density as a function of the axis orthogonal to the membrane surface each simulation containing a reflectin linker simulated in the presence of a membrane.
Figure 4.33. Number of motif residues in contact with lipids vs. average interdigitation (nm). Color represents time in ns
Figure 4.34. Number of linker residues in contact with lipids vs. average interdigitation (nm). Color represents time in ns.
Figure 4.35. Local lipid interdigitation (nm) from bilayers simulated in the presence of motifs.
Figure 4.36. Local lipid interdigitation (nm) from bilayers simulated in the presence of **motifs**.
Chapter 5

Experimental Insights into Reflectin-Membrane Interactions

The sequence analyses and modeling results from Chapters 3 and 4 suggest the reflectin protein has a high affinity for associating with interfaces and interacting with lipids. These predictions can be translated into experimentally testable hypotheses. However, native and recombinant reflectin proteins have proven to be difficult to work with in experiments. The proteins have low solubility in water, and the only known good solvents are solutions of sodium dodecyl sulfate (SDS)\(^{45,192}\) and hexafluoroisopropanol (HFIP)\(^{31,47,48,51}\). Given that these solvents have very different properties than the aqueous cytoplasm, it is unclear if the results obtained from experiments utilizing these solvents accurately reflect the behavior of reflectin in its natural environment. Furthermore, reflectin protein as expressed in squid cells is both heterogeneous and polydisperse, so it is unclear whether or how the typically monodisperse, recombinant preparations used in experiments\(^{31,37,47,48,55,176}\) reflect the properties of the natural system. The most common approach to performing biochemical assays on reflectin is to prepare a recombinant peptide designed from a repeated reflectin motif\(^{31,37,47,48,55,176}\). In contrast, I am interested
in exploring how the complexity of the natural system may inform its biophysical function.

The goal of this chapter is to develop an experimental understanding of the structure and function of the reflectin protein in a lipidic environment. I will begin by discussing the organization of reflectin in native cells using polarization microscopy. Then I will discuss biophysical insights gained from solution x-ray scattering experiments on a system of DMPC lipids and native reflectin protein solvated with HFIP. I will compare the experimental findings of shifts in lipid assembly in the presence of reflectin protein to the modeling results reported in Chapter 4.

**Materials and Methods**

**Reflectin Protein Extraction:** Reflectin protein was extracted from eye silver tissue of *Loligo pealeii* squid obtained from Marine Biological Laboratory of Woods Hole MA. Nearly 90% of the transcripts in eye silver tissue encode reflectin proteins (Chapter 3), so we prepared reflectin samples by adding the silver tissue directly to HFIP and mechanically disrupting the cells with a Dounce homogenizer to release the intracellular protein. The tissue was further dissociated by shearing it with a metal 0.7 mm x 25.4 mm needle attached to a 2 ml glass syringe. The tissue quickly broke apart and created a uniform, slightly light-scattering solution. 0.6 ml of HFIP was added for each 10 mg of protein, which appeared to be the amount of protein that would disperse in the solution without forming a pellet. Despite the fact that this preparation was not
filtered and contains all parts of the cell, I will refer to this solution as the reflectin solution.

**Polarization Microscopy:** The Zeiss optical microscope was fitted with a Polscope to measure the optical retardance (see Chapter 1) of materials (Laboratory of Rudolf Oldenbourg at the Marine Biological Laboratory in Woods Hole, MA). The samples were prepared by adding 50 µl of water with of Instant Ocean (1% w/vol) to a glass slide, and then a small piece of silver tissue was gently agitated in the water droplet. This gentle motion separates individual silver cells from the tissue without rupturing them and they subsequently settle onto the slide. The solution was covered with a coverslip and then placed under the microscope for imaging. The images were rendered with Zeiss Zen Microscope Software and the Polscope was operated using the OpenPolScope plugin in the Micro-Manager Software package.

**MLV Preparation:** 1,2-Dimyristoyl-sn-glycero-3-phosphocholine (DMPC) lipids form MLVs in water at room temperature. Multilamellar vesicles (MLV solution) were prepared using DMPC: a 20% wt/vol solution was created by mixing 0.1 g of DMPC and 0.5 ml of water. This solution was prepared fresh at the start of every day of SAXS experimentation. The resulting mixture was viscous and white. This mixture was cycled several times between 60 °C for 15 s and room temperature vortexing for 15 s. We observed experimentally that this thermal cycling favored fluid phase in the MLVs throughout the sample and reduced the amount of gel phase present. The MLV solution was subsequently kept at 60 °C until ready for use.
**Reflectin and DMPC Mixture Preparation:** There were two types of samples prepared for x-ray experiments: solutions containing DMPC, water, HFIP, and reflectin; and solutions containing DMPC, water, and HFIP with no reflectin added, referred to as the “control” solution. The control sample consisted of 10% HFIP and 90% MLV solution by volume. I used four concentrations of reflectin in this study, referred to by the ratio of the reflectin solution to the total solution volume: 1:10 (20 µl of reflectin solution and 180 µl of MLV solution), 1:15 (14 µl of reflectin solution, 7 µl HFIP, and 190 µl of MLV solution), 1:20 (10 µl of reflectin solution, 10 µl HFIP, and 180 µl of MLV solution), and 1:50 (4 µl of reflectin solution, 16 µl HFIP, and 180 µl of MLV solution). Therefore, the reflectin-containing samples consisted of a variable amount of reflectin solution with a constant volume fraction of HFIP. The molarity of DMPC was 0.265 M for all solutions, and the molarity of reflectin was 0.0345 mM, 0.023 mM, 0.0173 mM, and 0.0069 mM for the 1:10, 1:15, 1:20, and 1:50 concentrations, respectively.

The reflectin solution was added to the MLV solution while it was being gently vortexed. The mixture was then vortexed for at least one minute and then left to sit at room temperature until bulk separation between the aqueous and HFIP-containing fractions was completed (approximately 5 minutes, Fig. 5.1). The top phase was transparent and thin, and the bottom phase was a viscous, translucent gel-like material with a gray cobweb-like structure running through it. Mixtures containing reflectin protein also contained small opaque particles (order of magnitude 100 µm) in the bottom phase of the mixture.
Solution X-Ray Scattering: Small and wide angle x-ray scattering (SAXS and WAXS, respectively) were performed using a Xeuss 2.0 Dual Environment X-Ray Source (DEXS) instrument (Xenox, France). The x-ray beam used in this study was generated using a copper anode with energy of 30 W and collimated using two motorized scatterless slits at a separation of 0.7 mm for “high resolution” measurements, and of 1.2 mm for “high flux” measurements. The instrument has separate detectors for small and wide angle scattering; a 1M Pilatus solid state detector was used for small angle scattering (angles of 0.2° to 36°), and a 100k Dectris detector was used for wide angle scattering at angles up to 45 °C. Both detectors were calibrated using silver behenate powder as a standard prior to each experiment, and the sample-to-detector distance was computed using the Foxtrot software package (EnableSoft, USA). A mask was created to isolate the elastic scattering signal from the sample by collecting a scattering pattern for 60 s using a glassy carbon sample.
The solutions were injected into a reusable glass capillary in a metal housing using a 20 µl pipette tip filled with 20 µl of solution. The pipette tip was carefully positioned in the bottom phase of the solution (Phase b in Fig. 5.1), and the solution was drawn into the pipette tip very slowly. The pipette tip was then inserted into the capillary tube and the solution was expelled very slowly to avoid introducing air bubbles.

The capillaries were held in a Julabo temperature controller rack (JULABO, USA) that was positioned approximately 150 mm from the wide angle detector and approximately 2500 mm from the small angle detector for all measurements unless otherwise specified. Therefore data was measured in q-ranges of approximately $q \in (1.2, 3.1) \text{ Å}^{-1}$ and $q \in (0.025, 1.1) \text{ Å}^{-1}$, for the wide and small angle detectors, respectively. A Julabo temperature controller, controlled by Specfe (open source software by Guy Jennings), was used to set the temperatures of the solution during x-ray scattering. The sample chamber and the beam path were kept under a vacuum with pressure less than 1 Torr to minimize air scattering.

**X-Ray Scattering with Reflectin Solution**

A solution containing reflectin protein and HFIP (0.01 g in 0.6 ml HFIP) was measured in the x-ray beam at high resolution (collimation slits were set to 0.7 mm, setting the beam diameter to approximately 0.7 mm). The sample to detector distance was set to approximately 2500 mm, and the exact distance was determined using Silver Behenate as a calibration standard. The solution was kept at room temperature and scattering data were collected for 20 minutes.
X-Ray Scattering with a Mixture of MLV Solution and HFIP

Solutions containing a mixture of MLV solution and HFIP (1 volume HFIP in 9 volumes MLV solution) were measured in the x-ray beam at high resolution (collimation slits were set to 0.7 mm, setting the beam diameter to approximately 0.7 mm). The temperature of the solution was initially set to 20 °C and scattering data were collected for 20 minutes. The temperature was then increased first to 24 °C and then to 28 °C, and the scattering pattern was measured for 20 minutes at each temperature step. The data were processed using the Foxtrot Software package (EnableSoft, USA) and plotted in Python (Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org) with the Matplotlib208 plotting utility. These measurements were then repeated but with the temperature started at 28 °C and then lowered to 24 °C, and then 20 °C, respectively.

X-Ray Scattering with a Mixture of Reflectin Solution and MLV Solution

Solutions containing a mixture of MLV solution and varying concentrations of reflectin protein with constant HFIP volume fraction in the final sample (detailed above) were measured in the x-ray beam at high resolution. The sample to detector distance was set to 1300 mm, 2500 mm, or 3800 mm for the measurement of each sample, resulting in q-ranges of \( q \in (0.01, 0.3) \, \text{Å}^{-1} \), \( q \in (0.008, 0.2) \, \text{Å}^{-1} \), and \( q \in (0.004, 0.14) \, \text{Å}^{-1} \). With the sample-to-detector distance set to 2500 mm, the 1:10 reflectin solution was measured at 20 °C, 24 °C, 28 °C, 32 °C, 36 °C, 40 °C, and 44 °C. Data was collected for 20 minute intervals, and the measurements were again repeated in reverse order, starting with 44 °C.
and decreasing to 36 °C in 4 °C increments. The 1:15, 1:20, and 1:50 concentrations were measured at 36 °C, 40 °C, and 44 °C, and then again in reverse order.

The 1:10 solution was also measured with sample-to-detector distances of 1300 mm and 3800 mm. The solution was measured at 36 °C, 40 °C, and 44 °C at high resolution for 20 minute intervals.

**Results**

**Whole-Cell Imaging**

In polarization microscopy, all of the reflectin-containing cells imaged showed the same general shape and features (Figure 1.3). All cells were shaped like elongated ellipses, and the edges showed birefringence (see Chapter 1 for background). The retardance at the birefringent edge of the reflectin cells was measured to be approximately 1.6 nm, and the retardance in the middle was measured to be approximately 0 nm. The red vectors near the cell membrane, indicating the extraordinary axis of birefringence, tends to point in the direction perpendicular to the surface of the membrane (Figure 1.3b). The red vectors in the center of the cell did not appear to be organized. However, they had approximately the same magnitude as the background, which was effectively zero.
Figure 5.3: Polscope images of silver cells from *Loligo pealeii*. Lighter pixels indicate higher retardance. The size and direction of the red lines indicate the degree and direction of organization.

**X-Ray Scattering with Reflectin Solution**

The scattering intensity profile for the reflectin solution is shown in Figure 5.4. There are no prominent features in the profile. The slope is approximately -2 from $0.015 \text{ Å}^{-1} < q < 0.04 \text{ Å}^{-1}$, and the slope is approximately zero from $0.04 \text{ Å}^{-1} < q < 0.2 \text{ Å}^{-1}$.
X-Ray Scattering with MLV Solution

Fig. 1.8 shows the scattering intensity profile for DMPC lipids in a MLV phase. Each curve shows two prominent features, each one corresponding to a major repeated length scale in the MLV structure. The length scales are labeled in the diagram on the right side of Figure 5.4. The blue curve shows the structure of the lipid assemblies at 60 °C, so the MLVs should be in the fluid state. There are two sharp features located at q=0.100 Å\(^{-1}\) (real space distance of 62.5 Å) and q=0.199 Å\(^{-1}\) (real space distance of 31.6 Å). The green curve shows the scattering intensity profile of the MLV solution at 15 °C, well below the phase transition temperature. The two major features found at 60 °C are shifted slightly, and small secondary feature form. The feature at q=0.100 Å\(^{-1}\) shifts to q=0.095 Å\(^{-1}\), and a secondary feature forms at q=0.104 Å\(^{-1}\). The intensity of the secondary feature (q=0.095 Å\(^{-1}\)) is 20% of the intensity of the more prominent feature at 165
q=0.100 Å⁻¹. The other peak is located at q=0.199 Å⁻¹ and has a secondary feature at 0.192 Å⁻¹. The secondary feature is 90% of the intensity of the dominant feature.

![Figure 5.4. Left: DMPC MLVs above (blue) and below (green) the melting temperature. Right: MLV diagram showing the two correlation lengths in the scattering intensity profile.](image)

**X-Ray Scattering with a Mixture of MLV Solution and HFIP**

The scattering intensity profile of the solution containing HFIP and DMPC in water (10% HFIP by volume) is shown in Figure 5.5. The two sharp peaks characteristic to the MLV phase (see Figure 5.4) are absent, and instead there is one wide peak at q=0.18 Å⁻¹, corresponding to a D-spacing of 35 Å. The full width-half-max of the peak (q=0.18 Å⁻¹) is approximately 11 Å in real space. The full width-half max and the intensity are the same for all temperatures tested in this study.

At 24 °C, a small feature emerged at q=0.136 Å⁻¹(real space distance of 46 Å), and this feature increased in intensity as temperature decreased, forming an intense, narrow secondary peak with a full width-half max of approximately 1 Å in real space. At temperatures greater than 24 °C, the secondary peak disappears and there is an overall
decrease in scattering intensity in the q-range of 0.01-0.08 Å⁻¹. This q-range corresponds to D-spacings of 80-600 Å.

Figure 5.5: Intermediate range x-ray scattering intensities for solutions containing a mixture of HFIP and MLV solution. The measurements were made for the same sample at 20 °C, 24 °C, and 28 °C.

Reflectin Interaction with Lipid Interfaces

Fig. 5.6 shows the scattering intensity profile of the control solution (10% HFIP by volume in 20% aqueous DMPC wt/vol.) compared to the solution with 1:10 reflectin stock in HFIP in 20% DMPC wt/vol., at two different q-ranges: q∈(0.01, 0.3) Å⁻¹ (Figure 5.6a) and q∈(0.004, 0.14) Å⁻¹ (Figure 5.6b) at 20 °C. The scattering intensity profile from the reflectin-containing solution has the same general features as the control solution. Both solutions have a wide peak at q=0.18 Å⁻¹, a similar slope at low-q, and the relative intensities of the features are within an order of magnitude of each other. However, there are several differences between the two solutions (Fig. 5.6a). The height of the wide peak (q=0.18 Å⁻¹) is decreased in the solution containing reflectin protein by 42%. An additional feature forms at q = 0.055 Å⁻¹ in the solution containing reflectin, which
corresponds to 11 nm in real-space. The intensity of this feature is 40% larger than the control solution. The intensity of the secondary peak formed at q=0.136 Å⁻¹ is increased by 13% in the reflectin-containing solution. Another small feature forms around q=0.3 Å⁻¹ in the solution contain reflectin, which corresponds to 21 Å in real-space. The intensity of this feature is 12% greater than the control solution.

Fig. 5.6b shows both plots have the same overall low-q behavior, but the solution containing the reflectin protein had an overall higher scattering intensity in the low q-range. The additional feature in the reflectin solution at q=0.055 Å⁻¹ is also present in Fig. 5.5b but it is not as prominent.

Figure 5.6. Scattering intensity at 20 °C. The green curve comes from a mixture of HFIP and MLV solution in a ratio of 1:10. The blue curve comes from a mixture of reflectin solution and MLV solution in a ratio of 1:10. a: scattering intensity profile in the intermediate scattering range. b: scattering intensity profile from the low-q scattering range.

The 1:10 reflectin stock in HFIP in 20% DMPC wt/vol. sample was measured again at 36 °C, 40 °C, and 44 °C in a q-range of q∈(0.008, 0.2) Å⁻¹ (Fig. 5.7a). The secondary peak at q=0.136 Å⁻¹ was the only feature that changed appreciably as the temperature increased. The secondary peak had an intensity of 105% relative to the peak at q=0.18 Å⁻¹ and a full-width half-max of 1 Å at 36 °C. The intensity of this peak
dropped to 76% of the intensity of the peak at $q=0.18\ \text{Å}^{-1}$ after increasing the temperature to $40^\circ\text{C}$, and the full-width half max was 1 Å. The peak was barely detectable at $44^\circ\text{C}$. The measurement made at $36^\circ\text{C}$ has a slightly more intense feature at $q=0.055\ \text{Å}^{-1}$ compared to the measurements made at $40^\circ\text{C}$ and $44^\circ\text{C}$, and a slightly less intense feature at $q=0.18\ \text{Å}^{-1}$ compared to the measurements made at $40^\circ\text{C}$ and $44^\circ\text{C}$.

Solutions with lower reflectin concentrations share the same general features as the 1:10 concentration (Fig. 5.6a-d). The intensities of the peak at $q=0.18\ \text{Å}^{-1}$ are approximately constant between all of the concentrations. The secondary peaks at $q=0.136\ \text{Å}^{-1}$ followed the same trend as the 1:10 concentration: the measurement taken at $36^\circ\text{C}$ has the highest intensity secondary peak, and the intensity of the peak drops as the temperature is increased. At $44^\circ\text{C}$, the intensity of the peak is diminished. However, the intensity of the features at $q=0.10\ \text{Å}^{-1}$ and $q=0.055\ \text{Å}^{-1}$ vary for each concentration. We observed that the intensity of these features depended on the exact position of the beam relative to the sample – a $<1\ \text{mm}$ movement of the beam on the capillary often resulted in an arbitrary change in intensity of this feature.
Figure 5.7. Temperature ramp for solutions containing reflectin stock and MLV stock in a ratio of a: 1:10, b: 1:15 (4/24), c: 1:20 (4/24), d: 1:50, and e. shows the intensity of the secondary peak (q=0.136) for the three temperatures tested in a-d using the same color scheme.
Discussion

The whole-cell imaging showed us that the material near the edge of the cell membrane is organized and the material in the rest of the cell is not. The exact thickness of the material was not measured and therefore the birefringence cannot be computed directly. However, I estimate thickness has an order of magnitude of 1 µm. Therefore, since the birefringence is the ratio of the retardance to the thickness, the birefringence has an order of magnitude of 0.001, which is indeed very small. It is possible that this signal is derived from the lipids in the membrane and not the protein, but this does not seem likely since the retardance at the corners of the cells is close to zero. Given the likelihood that the reflectin protein is intrinsically disordered, it is surprising that it has a measurable birefringence value, even if it is very small. However, since we observe non-zero birefringence, it suggests that the reflectin proteins near the cell membrane may have a semi-regular arrangement, but the reflectins far from the membrane lack a bulk orientation. These observations suggest that the cell membrane has an orientational ordering effect on the reflectin proteins, and without it the reflectin proteins lack orientation.

The solution made from adding reflectin to HFIP did not show any major features in the x-ray scattering intensity profile, and therefore we conclude the protein was adequately solvated by HFIP. There is not enough low-q data in Figure 5.4 to confidently measure the fractal dimension, which is equal to the slope at low-q, but if the curve in
Figure 5.4 is extrapolated then the fractal dimension is estimated to be 2. The modeling results in Chapter 4 predict the Flory exponent of the motifs in solution (water) to be approximately 0.5, which is consistent with a Gaussian chain. The linkers in solution (water) were predicted to be slightly more compact, with Flory exponents in the range of 0.4-0.5. The fractal dimension of such polymers (which is the inverse of the Flory exponent) is then predicted to be close to 2, which is consistent with the experimental results. Since HFIP is a better solvent for reflectin than water, we expect it to adopt a slightly larger structure (i.e. larger Flory exponent) in HFIP compared to water, and indeed the Flory exponent estimated from extrapolating the curve in Figure 5.4 is consistent with that.

The MLVs demonstrated a temperature dependence that can be attributed to the lipid structure in a mixed interdigitated gel state, similar to what was observed in the molecular simulations in Chapter 4. The leftward shifts in both of the peaks in the green curve (15 °C) relative to the blue curve (60 °C) in Figure 5.5 indicates the bilayer has increased in thickness, which is consistent with the ordered gel phase. The shoulder at q=0.104 Å⁻¹ (real space distance of 60.4 Å) on the green curve (15 °C) shows a fraction of the bilayer got smaller, which is consistent with an interdigitated gel phase. The distance between the two peaks (q=0.104 Å and q=0.095 Å) in real space is approximately 5 Å, which matches the average interdigitation distance in the modeling results reported in Chapter 4.

When HFIP is added to the DMPC system in the absence of reflectin, the overall shape of the scattering intensity as a function of q changed relative to the aqueous system.
A major shift in the scattering intensity profile is consistent with a phase transition because the location and widths of the features changed, which means the size and distribution of molecular assemblies has changed. The shape of the scattering intensity profile above 24 °C has one broad feature at 35 Å, which does not match the correlation lengths from the MLV phase. However, this correlation length is similar to the size of water channels reported in other lipidic sponge phases. Therefore I hypothesize that HFIP causes the lipids to phase transition to a sponge phase. At temperatures below 24 °C, the scattering intensity profile shows the emergence of a secondary peak with a real-space distance of 45 Å, which corresponds to the length scale of a DMPC bilayer. The presence of both features at q=0.18 Å⁻¹ and q=0.136⁻¹ suggests there is a coexistence of both sponge phase and some sort of lamellar phase, which is interpreted as a cubic bicontinuous phase.

The transition from a cubic bicontinuous phase to a sponge phase is typically induced by reducing surface tension. Since the melting temperature of DMPC lipids is 24 °C, which is also the temperature that shifts the system between the cubic phase and the sponge phase, the phase transition is likely caused by the lipids melting from the ordered gel phase to the liquid crystalline phase. The surface tension of fluid bilayers is significantly smaller than that of gel phase bilayers. A study using DPPC lipids (similar to DMPC but each lipid tail has 2 more carbons, and the phase transition temperature is 41 °C) found the relationship between surface tension and temperature is approximately linear with a slope of -1.27 µN/cm/K.
When reflectin is added to the system, we again observe a wide peak at $q=0.18 \text{ Å}^{-1}$ and a secondary peak around $q=0.136 \text{ Å}^{-1}$. Therefore, the data suggest that the reflectin solution, which is predominantly HFIP by weight, also causes the DMPC lipids to phase transition to a cubic phase or a sponge phase. However, reflectin affects the bicontinuous phases in several ways. First, the increased height of the secondary peak ($q=0.136 \text{ Å}^{-1}$) indicates the system includes a higher proportion of lipids in the gel phase, which is apparently a direct consequence of the presence of reflectin protein. Furthermore, the secondary peak was present up to temperatures around 40 °C, which is 16 °C above the temperature at which this peak disappeared in the DMPC/HFIP system without reflectin. These data suggest that the cubic phase was stabilized by the reflectin protein. Our data suggests that the phase transition we observe is driven by melting ordered gel phase bilayers, so it seems likely that the presence of reflectin raises the apparent DMPC melting temperature.

Each reflectin concentration (Figure 5.7) exhibited the same relationship between the height of the feature at $q=0.136 \text{ Å}^{-1}$ and the temperature. The heights of the peaks varied from sample-to-sample, but the features were all totally diminished at temperatures above 40 °C, which suggests there is not a dose dependence between the concentration of reflectin and the temperature that will cause a phase transition.

The scattering pattern at low-$q$ ($0.004 \text{ Å}^{-1} < q < 0.01 \text{ Å}^{-1}$) was not altered by the presence or absence of reflectin. This either indicates that the fractal dimension is only affected on larger length scales (> 60 nm), or reflectin does not remodel the bicontinuous phase on a larger order and therefore the fractal dimension is unchanged.
An additional feature that forms only in the reflectin condition at q=0.055 Å⁻¹, at a d-spacing of 115 Å corresponds to the size of the “protopeptide” reported by Guan and colleagues. This study reported that reflectins form stable assemblies of eight reflectin proteins. Given the only major feature that changes when reflectin is added to the control sample is this additional peak at a temperature consistent with the fluid phase, I hypothesize that this peak is a direct measurement of an assembly of reflectin proteins in a similar arrangement as the protopeptide.

These experimental results inform and help interpret the modeling results reported in Chapter 4. In experiment, the phase transition temperature of the DMPC lipids was significantly increased when reflectin was introduced to the MLV solution. Similarly, the modeling results also suggest that the membrane begins a phase transition to the interdigitated gel phase when a critical number of contacts (~10) with reflectin are made. The lipids directly in contact with the protein interdigitate with the lipids in the opposite leaflet (by 1.5 nm), but the lipids not in direct contact transition to an ordered gel phase. The net effect in our simulations is then a mixed interdigitated gel phase. Given a critical density of reflectin proteins, it is possible that the entire simulated membrane could adopt the interdigitated gel phase.
Chapter 6

Conclusion

The two systems outlined in this dissertation are rich with physical insight. The biophysical mechanism driving the HRF effect comes down to organizing and orienting water molecules to facilitate hydrogen bonding between glycerol molecules. The mechanisms behind reflectin self-assembly into membrane bound platelets are simple rules of geometric packing and phase transitions in lipids to increase order. Both of these mechanisms are driven by local and global molecular ordering, and are therefore governed by entropy.

The mechanism of the mysterious HRF molecule in clam homogenate was long thought to be driven by a complex biological process. However, we elucidated the mechanism using basic concepts from physics like diffusion and free energy. Science research is becoming more interdisciplinary and this paradigm is leading to creative new solutions. Chapter 2 is an exciting example of how collaboration across fields can produce deep insight into nature that otherwise would not be possible.
Betaines are found in millimolar concentrations in marine photosymbiotic organisms (for example, a recent study found the mean concentration of each of seven betaines in coral reefs was 75 mM, range of 12–204 mM), making this a plausibly general mechanism explaining the HRF effect throughout coral reef symbioses. Previous work has inconclusively suggested that these betaines have photoprotective effects; our results suggest the possibility that a primary role for these metabolites is the transfer of photosynthate from algal cells to animal cells within the symbiosis.

Reflectin sequences are both polydisperse, meaning they can be different lengths, and heterogeneous, meaning reflectins of a given size have different sequences (Chapter 3). I hypothesize that this extreme variability has evolved to maximize the packing fraction of the protein. The proteinaceous platelets need to pack at the highest density physically possible to achieve a maximum refractive index. The heterogeneity in the reflectin sequences precludes the structures from forming a repeated periodic lattice, and therefore prevents crystallization at extremely high densities. Crystal lattices have suboptimal packing densities and are highly inflexible. On the other hand, random close packing is the most efficient packing arrangement, and it is only made more efficient by introducing polydispersity. The likely intrinsically disordered, heterogeneous, and polydisperse reflectin sequences are well suited to take advantage of these properties and form a highly flexible, maximally electron dense platelet.

My approach to analyzing transcriptome data for membrane binding potential (Chapter 3) shows the polydispersity has little effect on the interfacial properties of the protein. There is often a push towards producing highly pure samples in protein
experiments, but in this case the structure and function of the protein likely depends on the natural variability of the sequences. I believe future studies aimed at characterizing the role of polydispersity in reflectin self-assembly will lead to a deeper understanding of how nature takes advantage of stochasticity.

IDPs make up roughly 10% of all proteins, and it is estimated that at least 40% of eukaryotic proteins contain at least one intrinsically disordered region. Despite this, most molecular modeling tools are not well suited for IDPs, and there are not many computational studies of IDPs. Furthermore, the common vocabulary used to describe protein structures is not well suited for describing IDPs. In this dissertation I utilized a hybrid approach drawing from both polymer physics and protein biochemistry to analyze protein conformations. There is growing interest in IDPs because of their prevalence in neurodegenerative diseases, but there is no standardized procedure for characterizing these structures. The software package I wrote to perform analyses of reflectin structures combines tools appropriate for both ordered and disordered proteins and is freely available (https://github.com/dillionfox/IDP_analysis).

There are few simulations published in literature detailing interactions between intrinsically disordered proteins and membranes. One of the main reasons for this is because the best force field for lipids (CHARMM36) is notorious for over-compacting intrinsically disordered protein structures. Many of the best force fields for intrinsically disordered proteins utilize artificially modified water models that render them incompatible with membrane force fields. Furthermore, deriving physically realistic structures of intrinsically disordered proteins is very challenging and is still a growing
field. This study takes advantage of an improved version of the CHARMM36 force field (CHARMM36m\textsuperscript{10}) that has been corrected to work with intrinsically disordered proteins without modifying the water model. There are so few examples of IDP-membrane interactions that we as a community cannot make confident predictions about how other IDPs should interact with membranes, and therefore this is an important case study. We find that the linkers, which interact strongly with the lipids, compact and have an increased propensity for forming beta-sheet structure upon binding the membrane, but the structures appear disordered in solution. Therefore reflectin might adopt a folded structure in the presence of an interface.

Our results are the first report of a protein that increases the phase transition temperature of lipids. It remains unknown whether living iridocytes actually contain gel-phase bilayers, and we would need to determine the lipid composition of iridocyte cell membranes to find out. However, I hypothesize that this cooling effect could contribute to iridocyte function whether the membrane undergoes a phase transition \textit{in vivo} or not. Careful imaging of iridocytes shows that low-curvature reflectin platelets are blanketed by the cell membrane, and my initial hypothesis was that the reflectin protein bound the membrane and induced curvature to flatten the membranes. However, gel phase membranes and “cool” fluid membranes have larger bending rigidities compared to fluid phase membranes.\textsuperscript{21} Therefore I postulate one of the primary functions of reflectin is to associate with the cell membrane and increase its rigidity.
Figure 6.1: Cartoon diagram showing how reflectin lowers the effective temperature of the membrane which causes an increase in the bending rigidity. We hypothesize that this mechanism counters the spontaneous curvature intrinsic to cells with spherical topologies.

It is possible that the mechanism by which reflectin lowers the phase transition temperature is similar to dehydration. Reflectin has been shown to have a very high propensity for forming thin films\textsuperscript{4,22–24}, and the sequence studies and modeling results demonstrate a strong propensity for associating with interfaces. Therefore, I propose that reflectin forms a thin film-like assembly at membrane interfaces, effectively coating and dehydrating the membrane surface. The mechanical and thermodynamic properties of membranes are highly sensitive to the structure of the water in and around the lipids. By replacing the robust water-lipid and water-water hydrogen bonding network with electrostatic and hydrophobic interactions between reflectin and lipids, the lipids diffuse at a slower rate and the effective temperature of the membrane is decreased.
16. Muscatine, L., Falkowski, P. G., Porter, J. W. & Dubinsky, Z. Fate of
32. Denton, E. J. & Land, M. F. Mechanism of Reflexion in Silvery Layers of Fish and


70. Simons, K. & Van Meer, G. Lipid Sorting in Epithelial Cells. Biochemistry 27,
87. Mayer, J. E. & Streeter, S. F. Biomembrane Phase Transitions STUDIES OF LIPID-WATER SYSTEMS USING DIFFERENTIAL SCANNING
104. Johansen, D., Trewhella, J. & Goldenberg, D. P. Fractal dimension of an


560 (1908).


164. Vanommeslaeghe, K. *et al.* Charmm General Force Field: A Force Field for Drug-


