# COMPUTATIONAL MODELING AND DESIGN OF PROTEIN AND POLYMERIC NANO-ASSEMBLIES

Christopher D. Von Bargen

A DISSERTATION

in

Chemistry

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

---

Jeffery G. Saven, Professor of Chemistry

Graduate Group Chairperson

---

Gary Molander, Hirschmann-Makineni Professor of Chemistry

Dissertation Committee

Joseph Subotnik, Associate Professor of Chemistry
E. James Petersson, Assistant Professor of Chemistry
A. T. Charlie Johnson, Professor of Physics, Nano/Bio Interface Center Director

COMPUTATIONAL MODELING AND DESIGN OF
PROTEIN AND POLYMERIC NANO-ASSEMBLIES

© COPYRIGHT

2015

Christopher Daniel Von Bargen

*For my father.*

*I thought I could get my degree faster than you,*
*but I must have misheard 'weeks' for 'years'.*

*Tomato, tomato.*

## ACKNOWLEDGEMENT

Thanks to my thesis advisor Professor Jeff Saven. I have learned an immense amount from you. As a mentor, you gave me the space to explore and build a diverse scientific knowledge and intuition. Thank you for your guidance and trust, including allowing me to dig into the details and rebuild the lab's protein design software. You were at once accommodating and challenging, and your confidence in me in turn gave me the confidence to wear many hats. I am a much better scientist for that. I am happy to have been part of your lab.

To the many members of my graduate committee, Professors Joe Subotnik, James Petersson, Charlie Johnson, Dawn Bonnell, and Kent Blasie. Thank you for your insight and honesty and advice, setting clear milestones and bearing by my best interest. Thanks to my chair, Joe Subotnik, for your sincere interest in my future, and for taking me to lunch way back at the beginning of things.

To all the members of the Saven lab with whom my path crossed, Dr. Seung-gu Kang, Dr. Jose-Manuel Perez-Aguilar, Dr. Chris Lanci, Dr. Chris MacDermaid, Matthew "Will" Eibling, Dr. Lu Gao, Huixi "Violet" Zhang, Jose Villegas, Wenhao Liu, and Krishna Vijayendran. We have the best offices in the whole building – hidden and quiet and with reliably outrageous lunch hours. I count you as collaborators and friends, and can't thank you enough for sharing your knowledge, and more importantly, your opinions.

To Chris Lanci and Chris MacDermaid. Lanci, for 3D glasses, the LEGO Friends doppelgänger, and always holding your convictions far past the point of reason. Thanks for all the experimental insight, always pushing to get things done, and having my back. MacDermaid, for Salt Lake City and rare books, witnessing the "convince me", and tempering Lanci. Thanks for kickstarting hacking away at COMPASS and sharing the most enjoyable days of grad school. More importantly, thank you for introducing me to Axel and Mike and a pantheon of inspiring computational scientists. Thank you both for including me from the first day I joined the lab. We'd have a lot of wonderfully silly times together.

To Matthew Eibling, or by his stage name, Will. I don't know how grad school could have been better without visiting the Arboretum on the hottest day of the year. (N.B., sarcasm.) Thank you for constant and detailed updates, filling in my knowledge gaps and always taking the time to explain things to me. The ups and downs of scientific progress were always most extreme with you, and getting our projects to work – which they will! – will be all the sweeter for it.

To Lali Pazos and Ethan Alguire. We've come a long way from kicking Ed awake in class and trolling Kate. Thank you for always comparing notes, lending an empathetic ear, and the countless pep talks. Lali, you are responsible for me knowing more than a dozen people at Penn. Thank you for the latest department buzz, and more so, the sincere encouragement. Ethan, we will always have Count Colin Mochrie. Thank you for always including me no matter how much of a fuddy-duddy I am, and for all your candid advice. You have both been good friends from the beginning, and despite what I said, I always enjoyed you showing up in my cube unannounced ~~for lunch~~ to eat your own lunch.

To Grant Scribner. For making fun of science and bringing levity to six years of school. As serious as I could be, it only took a *blah, blah* to put things in perspective. Littering these last years with two man quizzo and historian talks and salmon pants made for the best of times. Thank you for coercing Kristen and I to live in Mole Street, and the incredible subsequent year of mashed potatoes and garlic bread. For taping every invitation and chopping peppers with gloves and moving all our stuff over and over. I hope I have the privilege of getting pulled over for speeding taking you to Henry's.

To my family, all of you, for your support of the choice to be in college for a decade. I probably would have lost track of time if no one asked me if I was almost done. In all seriousness, thank you for the encouragement, and your willingness to listen to me ramble about nanotubes and proteins. To my parents, for pointing me at Cooper and letting me go. You have always spurred me forward with confidence. You make me intensely prideful to do this work and be here today.

And to Kristen, who came into view less than a month before all this began. For sharing Jack and the Beanstalk and Duck and that time the toilet overflowed. For becoming intimate with 95 traffic, getting towed in an ice storm, living in the mouse house, and banning yourself from the state of Maryland, for me. I can't imagine facing Irshy or Ron BELLamy or Jim Stroh without you. You are incredible and inspiring, and I hope that this degree gets me anywhere near the success and fortune of your blossoming career. You have this incredible power to still my mind and give me the courage to take a step forward. Today is a clear demarcation of the next part of our huge, incredible life. I am on the edge of my seat to see where our master speed takes us next.

Also thanks, Moxie. Thoxie.

# ABSTRACT

## COMPUTATIONAL MODELING AND DESIGN OF PROTEIN AND POLYMERIC NANO-ASSEMBLIES

Christopher D. Von Bargen

Jeffery G. Saven

Advances in nanotechnology have the potential to utilize biological and polymeric systems to address fundamental scientific and societal issues, including molecular electronics and sensors, energy-relevant light harvesting, "green" catalysis, and environmental cleanup. In many cases, synthesis and fabrication are well within grasp, but designing such systems requires simultaneous consideration of large numbers of degrees of freedom including structure, sequence, and functional properties. In the case of protein design, even simply considering amino acid identity scales exponentially with the protein length. This work utilizes computational techniques to develop a fundamental, molecularly detailed chemical and physical understanding to investigate and design such nano-assemblies. Throughout, we leverage a probabilistic computational design approach to guide the identification of protein sequences that fold to predetermined structures with targeted function. The statistical methodology is encapsulated in a computational design platform, recently reconstructed with improvements in speed and versatility, to estimate site-specific probabilities of residues through the optimization of an effective sequence free energy. This provides an information-rich perspective on the space of possible sequences which is able to harness the incorporation of new constraints that fit design objectives. The approach is applied to the design and modeling of protein systems incorporating non-biological cofactors, namely (i) an aggregation prone peptide assembly to bind uranyl and (ii) a protein construct to encapsulate a zinc porphyrin derivative with unique photo-physical properties. Additionally, molecular dynamics simulations are used to investigate purely synthetic assemblies of (iii) highly charged semiconducting polymers that wrap and disperse carbon nanotubes. Free energy

calculations are used to explore the factors that lead to observed polymer-SWNT superstructures, elucidating well-defined helical structures; for chiral derivatives, the simulations corroborate a preference for helical handedness observed in TEM and AFM data. The techniques detailed herein, demonstrate how advances in computational chemistry allot greater control and specificity in the engineering of novel nano-materials and offer the potential to greatly advance applications of these systems.

TABLE OF CONTENTS

*Adapted from Christopher D. Von Bargen, Christopher M. MacDermaid, One-Sun Lee, Pravas Deria, Michael J. Therien, and Jeffery G. Saven. "The Role of Ionic Side Chains in the Helical Wrapping of Phenylene Ethynylene Polymers about Single-Walled Carbon Nanotubes." *J. Phys. Chem. B.*, 2013, 117 (42).

†Adapted from Pravas Deria, Christopher D. Von Bargen, , Jean-Hubert Olivier, Amar S. Kumbhar, Jeffery G. Saven, and Michael J. Therien. "Single Handed Helical Wrapping of Single-Walled Carbon Nanotubes by Chiral, Ionic, Semiconducting Polymers." *J. Am. Chem. Soc.*, 2013, 135 (43).

# LIST OF TABLES

# LIST OF ILLUSTRATIONS

# 1 Introduction

Current technology in nano-material engineering offers a diversity of potential applications; for example, carbon nanotubes for nanoscale electronics, polymer bio-nano matrices for water purification, small organic molecules for photovoltaic devices, and proteins for biosensors. Such materials promise the ability to address fundamental scientific questions while tackling concerns germane to society. Yet understanding the details of these materials, and in turn the design of new systems, is often met with enormous complexity. The task requires simultaneous consideration of large numbers of degrees of freedom which can include structure, composition, and functional properties. Protein design is the quintessential example, wherein even simple sequence variability scales exponentially with protein length. The following work encompasses theoretical and computational methods to model and design nano-scale polymeric systems, with an emphasis on understanding their well-defined structures and spontaneous self-assembly.

## 1.1. Protein Engineering and Computational Protein Design

Novel protein assemblies can drive the advent of new nanotechnologies, either through the redesign of naturally occurring proteins or the design of completely *de novo* structures. Redesign of the common TIM-barrel structure to stabilize high energy reaction intermediates in reactions[1–4] underscores the ability to design catalysts for which no known natural enzyme exists. Designing fluorophores into transfer proteins to mark binding of fatty acids showcases the creation of novel biosensors[5]. A variety of *de novo* proteins have been designed to encapsulate synthetic nonbiological cofactors with interesting and unique nonlinear optical responses or light induced electron transfer properties[6–13], which provide promising frameworks with which to develop materials. These designed proteins can provide insulation between neighboring cofactors and serve as the means by which to direct organization into the macro-scale structures required for optical communications applications or efficient har-

vesting of photo-generated charges required for organic photovoltaics. The many successes of protein engineering and design stand at the forefront of nanotechnology.

Protein design seeks to identify a sequence, or set of sequences, that folds to a particular structure which confers a targeted function[14]. It both promotes our knowledge of biological processes and facilitates the engineering of novel proteins[15]. The determination of the key physical and chemical features of a sequence that dictate native structures is often difficult to identify and isolate in natural proteins, despite the expanding set of protein structures available[16]. Early design efforts utilized empirically derived knowledge of protein secondary structural motifs and amino acid propensities for those motifs to select new protein sequences with limited success – often less structurally defined than their natural analogues[17–22]. The complexities associated with protein structure formation are well-known: proteins have tens to thousands of amino acids, those residues can take on an exponentially large number of conformations, and stabilizing forces are relatively weak noncovalent interactions. Identification of a sequence alone requires traversal of the huge sequence space; e.g., a modest 100-residue protein using the 20 natural amino acids can have more than $10^{130}$ possible sequences. Furthermore, the pliancy of proteins often allows similar structures and function between sequences with little similarity. The structure and function of some sequences may be highly sensitive to mutation, thereby impeding an evolutionary approach to potentially discover new proteins with targeted functionality. While combinatorial protein experiments provide a means to exhaustively explore sequence space, they are burdened by the vast complexity of the exponentially large number of possible sequences[23,24].

Tools developed from theoretical methods can directly address the inherent complexity of designing and redesigning proteins. In order to address the large number of concerted interactions within a targeted fold, these methods introduce simplifications that make the sequence/structure search space more tractable. As such, the complexity reduction is generally determined by (1) the target protein structure, (2) admissible flexibility, (3) the

sequence composition, and (4) energetic details. Targeted folds, chosen as to confer desired functional properties, are usually defined by protein backbone atoms and constrained during calculations. These structures can be extracted from existing x-ray crystal structures or NMR ensembles[25], assemblies of protein fragments[26–28], simple elements from secondary structures[29–31], or through homology modeling[32]. To mitigate overly constraining permissible sequences within a target structure, methods have introduced minute backbone flexibility by sampling neighboring configurations[33–39]. Conversely, the flexibility of side chains is directly addressed in discretizing conformational states. Often these states are classified on a rotameric basis (only varying side chain dihedrals). Rotamer libraries derive states from structural databases to cohere with energetically favorable structures[40], and can be categorized in a variety of ways: e.g. backbone independent, secondary structure dependent, and backbone dependent libraries[40–47]. Full conformational libraries (including variation in bond lengths and bond angles) can be used, but such practice is reserved for placement of small molecules[48,49]. While the sequence composition in design generally consists of all 20 natural amino acids, restrictions can be applied where appropriate. Regions of the target fold can be patterned, e.g. restricting buried positions to hydrophobic residues[50] to drive hydrophobic collapse[51], excluding known helix disrupting residues from helical regions, and fixing the identity of a residue required for metal binding. Conversely, the sequence can be broadened to included additional monomer units including nonnatural amino acids, or rotamers positioned with well defined water molecules[52]. Lastly, the choice of energy function quantifies the sequence-structure consonance and modeling accuracy[53,54]. Much design work borrows molecular mechanics force fields developed for molecular dynamics simulations[55–58]. While such atomistic potentials are obtained from experimental data and electronic structure calculations to account for the physical chemistry of interactions, alternate potentials can be employed, including knowledge-based methods or coarse grain energy functions[59]. Additionally, a hybridized approach can be taken to develop an optimized force field[60]. Solvation in protein design is modeled implicitly, as an explicit representation is generally impractical. Implicit solvation models can be treated as solvent exposure

propensities based on local protein density[36] or with detailed terms adopted from the generalized Born model[61,62]. A biasing against misfolded structures (negative design) can also be introduced into physico-chemical potentials through an unfolded state, or ensemble of states, represented as a set of reference energies. These can be obtained empirically as to reconstruct established amino acid frequencies in nature, or calculated from model peptide systems[63–65].

Protein design methods can harness these elements in a variety of ways to arrive at a sequence, though the techniques comprise two distinct camps of approach. Directed protein design is chiefly concerned with optimization procedures that select sequence identity based upon estimating global energetic minima. Stochastic methods, including Monte Carlo (MC) methods, sample sequences from a selected probability distribution as to escape local barriers in the sequence landscape and improve sampling. MC employs an effective temperature to determine acceptance rates which in practice can be altered during the course of sampling. MC techniques include simulated annealing[63,66], quenching[67], biasing methods[68], and replica exchange algorithms[69]. While such stochastic techniques attempt to converge upon a global minimum, they often require multiple independent calculations to fully explore the sequence landscape. Alternatively, deterministic optimization techniques are widely used, including genetic algorithms[70], graph search (including A*[71] and other heuristic optimizations[72]), linear programming[73], and pruning or elimination methods[74]. Elimination methods (namely, dead-end elimination) iteratively prune unfavorable rotamers not part of the optimal sequence until no further eliminations are possible. From the limited set of states remaining, an exhaustive search is performed to identify a global minimum. While elimination techniques are useful for small proteins and proteins with a limited number of amino acids/rotamers, computational time increases exponentially with the number of residue variables and they perform poorly for large systems[67]. Consequently, improvements have been suggested, including relaxing pruning criteria[75], comparing clusters of rotamers (generalized DEE)[76], revised elimination with flagging criteria[77], and placement of constraining boxes to bound rotameric interactions[78]. Nonetheless, directed methods in

general are limited by the size and number of sequences considered and require extensive calculations to inform the space of possible or nearby sequences.

Conversely, probabilistic methods provide a direct estimate of sequence variability rather than specific sequences. Such methods are able to access the immense sequence space and able to characterize ensembles of sequences that may fold to a selected structure. With a probabilistic approach, site-specific probabilities are estimated among the ensemble of sequences as placed within a targeted folded structure. The method is a means of addressing the inherent uncertainties in computational protein design, namely approximations in energy functions, discretization of rotameric states in the amino acids, and simplified solvation models that are often employed. Instead of simply identifying global minima, the site-specific probabilities act as a guide in design to identify the sensitivity of residue mutation to the overall structure or functional properties. Such methods provide broad perspective on the sequence-rotamer landscape. The formalism for such an approach is usually derived from mean field theory[79,80], which foregoes sequence enumeration to estimate probabilities informed by local energetic interactions. Where mean field methods traditionally lower an associated effective temperature to identify low energy sequences, a probabilistic approach seeks to maximize an entropy of sequence variability and utilize probability profiles to identify a consensus sequence[36,65,81–83]. It is this maximization of an effective entropy that is cardinal to the methodology. Probabilities are determined by optimizing an effective sequence-rotamer entropy, subject to constraints which can specify a variety of properties (e.g. mean energy of all possible sequences or local site-specific residue composition). The constraints are defined as average values, and are assumed to have minimal fluctuations about their mean. The method permits large protein structures (>100 residues) to be considered, with a time and space complexity bounded only by constraint dimensionality. Moreover, the method can be reduced to a directed mean field approach in the limit of lowering the effective temperature significantly to arrive at a most probable sequence. Likewise, the calculated probabilities can be used to efficiently bias MC trial sequences to arrive at a low energy candidate[84].

This entropically based formalism has been applied to a variety of protein design work, showcasing the advantages of identification of the diversity of neighboring sequences able to satisfy targeted structure and function[85]. Due to the difficulty in isolating and purifying membrane proteins, the probabilistic approach was used to target the redefinition of exterior residues in the transmembrane domain to ease processing and solubility while retaining the membrane protein structure and function. The designs of a water-soluble variant of a bacterial potassium ion channel (KcsA) showed close agreement of the solution phase (NMR) structure with that of the membrane-soluble wild type structure[86], and the design of a water soluble variant of a G-protein-coupled receptor exhibited comparable antagonist affinity as the native receptor[87]. Alternatively, these methods have been applied to *de novo* protein design, namely in the context of encapsulation of the aforementioned nonbioloical guest molecules (cofactors). Where natural proteins may not be sufficient for increasingly complex macromolecular cofactors, probabilistic protein design is able to adeptly apply constraints to the sequence space of both oligomeric proteins as well as *de novo* single-chain structures. Examples of such designs include a tetrahelical protein that selectively binds multiple copies of a nonbiological diphenyl iron porphyrin (DPP-Fe)[30], a A2B2 heterotetramer that selectively binds a nonbiological photoactivatable zinc porphyrin with correct stoichiometry[88], and single-chain helical proteins able to bind and encapsulate a variety of different cofactors and chromophores[13,89]. Recently, this design approach was extended to lattice systems targeting pre-specified crystal structures – a three-helix protein was designed to form a polar crystal in the P6 space group, exhibiting sub-Å agreement between the X-ray crystal structure and computation model[90].

## 1.2. Polymer-Nanotube Assemblies

A variety of synthetic macromolecules composed of organic and organometallic polymer assemblies have been key in developing new technologies involving liquid crystalline[91,92] optoelectronic,[93,94] and spintronic[95] materials. Such polymeric systems feature polarizable, hyperpolarizable, or low band gap building blocks[94,96–99]. For this reason, there is interest in

integrating such polymers with other nanoscale semiconducting materials to evolve entirely new classes of optoelectronic and spintronic materials. Semiconducting single-walled carbon nanotubes posses unique electro-optic properties and tunable valence and conduction bands[100], which make them prime for integration into hybrid nano-materials. Moreover, much work has been done in the way of utilizing semiconducting polymers as nanotube solubilization agents[101–106], including highly charged aryleneethynylene polymers which form distinct, well-formed helical superstructures[101,102,106].

Electron microscopy, atomic force microscopy, and spectroscopic methods can elucidate the structures of polymer-nanotube complexes with limited resolution. Molecular dynamics simulations have afforded a means to gain a molecular-level understanding of superstructural formation which will greatly advance the application and further design of these systems. The use of simulations to provide insight into the structure, fluctuations, and energetics of polymer-nanotube is well documented[102,107–114,114–125], building an atomistic picture of monomer interactions with the nanotube surface, conformation energetics preferences of the polymer backbone, and exploring the different superstructures these assemblies adopt. Additionally, relative free energies of polymer-nanotube superstructures have been calculated from these molecular simulations[126–128].

Obtaining free energy estimates from atomistic simulations can relate the precise details of atomic interactions to structural biases at the macromolecular scale. Unfortunately, such estimates are often difficult to calculate from conformational fluctuations in molecular simulations due to rugged and complex energy landscapes[129]. Advances in the techniques employing configurational averages over classical simulations[130–133], including free energy perturbation, non-equilibrium approaches applying exponential averaging of work estimates[134], adaptive biasing force sampling[135], thermodynamic integration[136], and replica exchange methods[137], have become indispensable tools to providing accurate thermodynamic estimates for macromolecular assemblies[138]. Joined with the time scales molecular dynamics simulations can access with current advances in hardware, free energy techniques

build molecularly detailed chemical and physical understanding of nanoscale assemblies, and offer the potential to greatly advance applications of next-generation soft matter materials.

## 1.3. Overview of Thesis

This thesis details computational techniques to model and design nano-assemblies. The majority of this work details and utilizes a probabilistic design approach to identify novel protein sequences. The statistical methodology is detailed extensively, with an emphasis on building a versatile and parallelized software suite (Chapter 2). Advances in the theory are presented, including identifying entropically weighted residue probabilities, inclusion of new constraints, and leveraging robust third party nonlinear optimization techniques to minimize the sequence free energy. Choices for energy functions, conformational libraries, and an elementary solvation model are validated in a series of benchmarking studies. Discussion of the program architecture makes a point that the methodology is easily extendable to any set of polymeric species as to include biological, nonnatural, and hybrid systems.

The probabilistic methodology is then applied to the *de novo* design of proteins which incorporate non-biological cofactors. Standing on the previous success designing a protein trimer in a predetermined crystal structure, we target the design of an environmentally relevant peptide trimer which aggregates in the presence of uranyl (Chapter 3). The work summarizes identification of multiple binding sites in the trimer core and a sequence consistent with packing in the P6 space group. Experimental results for selected sequences show the peptides readily aggregate upon binding uranyl. The design strategy is then extended to identify a single-chain protein which binds and orients a donor-bridge-acceptor electron transfer chromophore (Chapter 4). Emphasizing *de novo* design principles, the work traces out the identification of a coiled-coil motif with a hydrophobic core tailored to provide shape complementarity to the cofactor. A residue offering an explicitly positioned hydrogen bond to the cofactor is engineered into the core, *de novo* loops threaded onto the scaffold, and an exterior sequence identified with an emphasis on a crystalline assembly.

The remaining work employs molecular dynamics simulations to examine the superstructures associated with synthetic polymers wrapped about single-walled carbon nanotubes. The simple phenylene derivative of an amphiphilic, semiconducting poly-arylene ethylene is investigated with a series of free energy calculations to explore factors that drive the formation of well-defined helical motifs when non-covalently wrapping nanotubes (Chapter 5). The effects charged side chains of the polymer play in these structures is explored, including water-mediated conformations that promote specific conformations. For more complex derivatives of these semiconducting polymers, including those that utilize a binapthol component in their repeat unit, dynamics simulations are employed to verify helical handedness preferences dictated by polymer chirality (Chapter 6). All results are compared to the modest sampling of TEM and AFM images to provide an experimental reference for the simulations.

# 2 Probabilistic Computational Protein Design

## 2.1. Introduction

We focus on developing a tool for computational protein design, with the goal of providing a probabilistic approach to guide the identification of protein sequences and the properties of sequences that fold to predetermined structures. As designing such sequences requires simultaneous consideration of large numbers of degrees of freedom (including the structure, sequence, and functional properties of the protein system), it harnesses a theoretical framework for estimating site-specific probabilities of the amino acids among the ensemble of sequences likely to be compatible with the target structure. Such a probabilistic approach is partly motivated by the uncertainties associated with identifying sequences consistent with targeted functions and structures, e.g., energy functions are approximate; side chain conformations are treated discretely; backbone atoms are highly constrained; solvation is treated using simplified models. The site-specific probabilities identify residues that tolerate variation without affecting structure or other properties. As a result, such a method provides a broad perspective on the space of possible sequences and yields information that is useful in designing functional protein systems. To do so, an entropy based formalism is applied which estimates residue probabilities directly; this is done through an effective free energy minimization subject to a set of sequence constraints on the ensemble.

This methodology has been expanded and our implementation reorganized. The following both reviews and expands upon new contributions to the formalism. The overall theory and implementation of this protein design methodology is presented, namely, the use of statistical mechanics principles to create an effective sequence-structure free energy object, the application of an effective temperature to identify low energy sequences, the utilization of molecular force fields and rotamer libraries, and the choice of an optimization method and toolkit for obtaining solutions of residue probabilities. Furthermore, the methodology

is easily extendable to any molecular species with variable substituents. While we only present analysis of the program for protein systems, the methodology holds for any any general molecular ensemble, allowing the computational package to assess the variability and design of any biological, nonnatural, or hybrid systems.

## 2.2. Probabilistic Theory

While the statistical theory of sequence ensembles has been detailed previously[36,81–83], the following section aims to summarize and review all theoretical contributions to the formalism used herein. The theory aims to describe a probabilistic description of sequences that are likely to fold to a given structure through the determination of ensemble probabilities; in previous work this has targeted using the natural set of amino acids to determine protein sequences for given folded states, but we note that the formalism is extendable to any set of monomer moieties that assemble into a targeted polymer structure. Furthermore, the generality of the method allows for the application of constraints upon the ensemble of sequences, which is accounted for in the implementation of a probabilistic computational design program.

The following equations establish a formalism for counting the number of states associated with a particular polymer scaffold through the maximization of an entropy function. The ensemble of allowable polymer states is comprised of both monomer types and monomer conformations (e.g., amino acids and associated side chain conformations). We define the total sequence-state entropy, $S$, which quantifies the number of polymer states likely to adopt a particular polymer scaffold and is exactly analogous to the eponymous Boltzmann equation

**Eq. 2-1**

$$\frac{S}{k_b} = \ln(\Omega_{seq})$$

where $k_b$ is the Boltzmann constant, and $\Omega_{seq}$ is the total number of sequence-structure microstates. For example, a protein having N residue positions with the possibility of all 20 amino acid identities at each position and only a single conformation per amino acid states still possesses $\Omega_{seq} = 20^N$ possible sequences. Similarly, we can describe the ensemble of sequences as a discrete set of accessible polymer microstates, such that the ensemble sequence entropy is given by the formula

**Eq. 2-2**

$$\frac{S}{k_b} = -\sum_{seq}^{\Omega_{seq}} W_{seq} \ln W_{seq}$$

where $W_{seq}$ is the probability of a particular sequence-state. Monomer sites are assumed to take on discrete monomer identities and conformational states at each position on that polymer chain. As is done with mean field theories, we approximate the multivariable probability $W_{seq}$ by factorizing it into the joint probability of individual probabilities for all subunits of that sequence.

**Eq. 2-3**

$$W_{seq}(c_1(t_1, n_1); ...; c_N(t_N, n_N)) = \prod_{n}^{N} w_{(n,t,c)}$$

where we introduce a compactified indexing scheme to denote the probability of conformational state $c$ for monomer type $t$ at the polymer site $n$, denoted $c(t, n)$. The following definitions reflect the computational implementation of the theory, where the indices can be rolled into a single index $i$ over all allowed combinations of $c$, $t$, and $n$.

**Eq. 2-4**

$$w_{c(t,n)} \equiv w_{(n,t,c)} \equiv w_i \qquad \text{and} \qquad \mathscr{D} = \sum_{n}^{N} \sum_{t}^{T_n} \sum_{c}^{C_{nt}}$$

Here, $N$ is the total number of variable site locations in the polymer chain, $T_n$ is the total number of monomer types allowed at site $n$, and $C_{nt}$ is the total number of conformational states of monomer type $t$ at site $n$. This provides the total number of individual probabilities present in the ensemble of sequences as $\mathscr{D}$. It will become clear that this is the dimensionality of the formalism, which is in general a significantly smaller number than $\Omega_{seq}$; for the N residue protein with $20^N$ possible sequences, this reduces to $20 \cdot N$ individual probabilities.

Within this factorization approximation, we can then write the total sequence entropy as

**Eq. 2-5**
$$\frac{S}{k_b} = -\sum_n \sum_t \sum_c w_{(n,t,c)} \ln w_{(n,t,c)} \equiv -\sum_i^{\mathscr{D}} w_i \ln w_i$$

As each site must be occupied, this constraint is imposed on all $N$ sites so that all probabilities across each $n$ are appropriately normalized.

**Eq. 2-6**
$$\sum_t^{T_n} \sum_c^{C_{nt}} w_{(n,t,c)} \equiv \sum_t^{T_n} \sum_c^{C_{nt}} w_i = 1 \qquad \text{for all } n$$

If we now maximize the sequence entropy subject to these normalization constraints, it is clear that we then recover our original formulation of the Boltzmann entropy **Eq. 2-1**. For such a solution, the probabilities of the individual conformers are independent of each other; however, introducing a pairwise energy function (or any such quadratic or higher order constraint) to the optimization couples the probabilities.

The energy function is defined using the mean field approach, which naturally arises from the factorization approximation. The energy for a particular sequence microstate can be written as a sum over each intra-monomer potential term and the sum over all pairwise inter-monomer interactions in that sequence. Note that $\gamma_{(n,t,c;n',t',c')}$ for $n = n'$ is zero –

that is, there are no interactions between states at the same site.

**Eq. 2-7**

$$E_{seq} = \sum_{n}^{N} \gamma_{(n,t,c)} + \frac{1}{2} \sum_{n \neq n'} \gamma_{(n,t,c;n',t',c')}$$

where $\gamma_i$ is the potential term for conformer $i$, $\gamma_{ij}$ is the potential term between conformers $i$ and $j$ in that sequence, and $\frac{1}{2}$ avoids double counting ($\gamma_{ij} = \gamma_{ji}$). Furthermore, we define $\gamma_{ii} = 0$, removing self pairwise energies. For the local energy at index $(n, t, c) = i$, we can define

**Eq. 2-8**

$$\epsilon_(ntc) \equiv \gamma_{(n,t,c)} + \sum_{n',t',c'} \gamma_{(n,t,c;n',t',c')}$$

or using the compactified notation,

**Eq. 2-9**

$$\epsilon_i \equiv \gamma_i + \sum_{j} \gamma_{ij}$$

In a mean field treatment, we would assume that the the value of $\epsilon_i$ can be replaced with its mean value in the limit that fluctuations in energy about the mean due to variation in the sequence are small.

**Eq. 2-10**

$$\epsilon_(ntc) \approx \langle \epsilon_(ntc) \rangle = \gamma_{(n,t,c)} + \sum_{n',t',c'} \gamma_{(n,t,c;n',t',c')} w_{n',t',c'}$$

or using the compactified notation,

**Eq. 2-11**

$$\epsilon_i \approx \langle \epsilon_i \rangle = \gamma_i + \sum_j \gamma_{ij} w_j$$

Similarly, we make the assumption that the fluctuations of the internal energy are small about the average energy,

**Eq. 2-12**

$$\langle E \rangle \equiv U = \sum_i \gamma_i w_i + \frac{1}{2} \sum_{ij} \gamma_{ij} w_i w_j$$

Previous descriptions of this probabilistic approach to sequence design target an entropy maximization subject to a series of constraints $f_k(\mathbf{w})$ held at $f_k^o$, including an energetic constraint at $E^o$,

**Eq. 2-13**

$$\begin{aligned} \max \frac{S}{k_b}(\mathbf{w}) \ \text{ subject to } \ U(\mathbf{w}) - E^o = 0 \\ 0 \leq w_i \leq 1 \\ \text{for all } n, \ \sum_{tc} w_i - 1 = 0 \\ f_k(\mathbf{w}) - f_k^o = 0 \end{aligned}$$

However, it is often useful to specify a desired effective temperature (conjugate Lagrange multiplier of the energy) at which to optimize the probability profiles. This has the advantage of providing a means by which to compare energies across many related structures, as one does when probing a sequence-structure energy landscape. Alternatively, it can be viewed as a means of tuning the relative strength of the energetic contribution to the ef-

fective free energy landscape. Simply, this is accomplished by either a maximization of sequence entropy with an energetic function at a fixed Lagrange multiplier $\frac{S}{k_b}(\mathbf{w}) - \beta U(\mathbf{w})$, or equivalently, a minimization of mean sequence free energy,

**Eq. 2-14**

$$\begin{aligned} \min F(\mathbf{w}, \beta) = U(\mathbf{w}) - \frac{1}{\beta} \cdot \frac{S}{k_b}(\mathbf{w}) \\ \text{subject to} \qquad 0 \leq w_i \leq 1 \\ \text{for all } n, \sum_{tc} w_i - 1 = 0 \\ f_k(\mathbf{w}) - f_k^o = 0 \end{aligned}$$

In practice, we solve the effective free energy minimization problem with current nonlinear optimization techniques. As in statistical thermodynamics, $\beta$ corresponds to the conjugate variable of energy and offers the interpretation that its inverse is an effective temperature, $\beta = \frac{1}{k_b T}$, at the optimized free energy. In the case of high effective temperature and correspondingly high mean energy, many degenerate sequences exist causing increased variance associated with sequence energy. In the limit that $\frac{1}{k_b \beta}$ goes to infinity, all sequences are equally likely. Conversely, at low effective temperatures, the degeneracy is reduced and low energy sequences are favored. In the limit that the effective temperature goes to 0, the probabilities go to either 0 or 1 choosing a single low energy sequence. In such a case with only normalization constraints are imposed, this is exactly the optimization of the mean field approach in **Eq. 2-9** and **Eq. 2-11**.

A statistical thermodynamic quantity of particular interest to the formulation of the mean field sequence energy is the rate of change for the mean sequence energy with respect to change in the conjugate Lagrange multiplier, $\beta$; that is, the sequence heat capacity of the

system (see Appendix C for derivation). $C_v$ is estimated as

**Eq. 2-15**
$$C_v = -k_b\beta^2\frac{\partial U}{\partial \beta} \approx k_b\beta^2 \sum_{n}^{N}\left[\sum_{tc}{\epsilon_i}^2 w_i - \left(\sum_{tc}\epsilon_i w_i\right)^2\right]$$

### 2.2.1. Probability Profiles

From the individual probabilities, $w_i$, obtained by solving **Eq. 2-14** we also obtain reduced quantities that do not depend explicitly upon the conformational state of a given monomer type. The simplest is to define a simple aggregation of probabilities, $w_{(n,t)}$, of the same type $t$, capturing the potential entropy associated with allowed states for that monomer. The summation of probabilities over conformers of a given monomer type $t$ at site $n$ is given as

**Eq. 2-16**
$$w_{(n,t)} = \sum_{c}^{C_{nt}} w_i = \sum_{c}^{C_{nt}} w_{(n,t,c)}$$

However, this relies on an accurate representation of the entropy associated with the various monomer species allowed in a given ensemble. Often, it is useful to estimate the probability of a given monomer type at a particular position independent of the number of allowed conformational states. This both provides equality among type probabilities at high temperatures, as well as emphasizes the most probable state at a site among all possible states. We define an entropically-normalized type probability that is a function of the aggregated type probability presented in **Eq. 2-16**.

**Eq. 2-17**
$$\tilde{w}_{(n,t)} = \frac{w_{(n,t)} \cdot exp(\frac{S_{nt}}{k_b})}{\sum_{t}^{T_n} w_{(n,t)} \cdot exp(\frac{S_{nt}}{k_b})} \qquad \text{for} \qquad \frac{S_{nt}}{k_b} = -\sum_{c}^{C_{nt}} \frac{w_i}{w_{(n,t)}} \ln \frac{w_i}{w_{(n,t)}}$$

### 2.2.2. Approximating the Unfolded State

In the particular instance of protein design, an estimate of nontarget conformations is included. Explicitly sampling unfolded protein states is computationally prohibitive, given the flexibility of protein structures and large inter-residue interaction fluctuations[64]. As is commonly done, we approximate unfolded interactions by sampling discrete conformations for amino acids threaded onto a "unfolded" structure to estimate energetic reference values. As a means of crudely simulating variation in beta carbon interactions in neighboring backbone residues, the model peptide N-acetyl-**X**-N'-methylamide is chosen as this "unfolded" structure, where **X** signifies the placement of a single amino acid.

Values are dependent only upon amino acid type $t$, and calculated as free energies of the unfolded ensemble, $\gamma_{\mathrm{u},t}$, at a chosen unfolded temperature factor ($\beta_{\mathrm{u}}$).

**Eq. 2-18**

$$\gamma_{\mathrm{u},t} = -\frac{1}{\beta_{\mathrm{u}}} \ln z_{\mathrm{u},t}(\beta_{\mathrm{u}})$$

where the unfolded ensemble partition function is defined as

**Eq. 2-19**

$$z_{\mathrm{u},t} = \sum_{\phi,\psi} \sum_{c} \exp(-\beta_{\mathrm{u}} \cdot \epsilon(\phi, \psi, t, c))$$

The partition function is estimated by miming unfolded states for each amino acid $t$ by evaluating energies $\epsilon$ across possible backbone configurations for each available residue conformation state. This translates to a sum over (i) a set of discretized model peptide states $(\phi, \psi)$ and (ii) the conformational states $c$ available to the that residue $t$[43,47]. Tripeptide states are selected as a systematic variation over backbone dihedrals $\phi$ and $\psi$ by 10° increments ($-180° < \phi < 180°; -180° < \psi < 180°$; 1,296 configurations) with all bond lengths,

all bond angles, and the N-acetyl and N'-methylamide groups held fixed. Energies, $\epsilon$, are calculated using identical components to the energetic coefficients identified in Eq. 2-12 used to quantify interaction energies in the folded ensemble. The unfolded temperature factor, $\beta_{\mathrm{u}}$, is chosen as a constant. $\beta_{\mathrm{u}}$ can be chosen as $\beta$ (**Eq. 2-14**), or an alternate value such as $\beta_{\mathrm{u}} \approx 1.69 mol/kcal$ consistent with room temperature.

The coefficients are used to formulate a mean unfolded free energy estimate in the context of the target structure. To provide a contextualization for these free energies, we offset the unfolded energies such that the unfolded free energy of glycine is zero. Note that while the summation is across the entire folded ensemble, the unfolded free energies are dependent upon only the amino acid type $t$ and not the local position or conformation ($n$ and $c$ respectively).

**Eq. 2-20**
$$\langle F_{\mathrm{u}} \rangle = \sum_{i} (\gamma_{\mathrm{u},t} - \gamma_{\mathrm{u,GLY}}) w_i$$

This then provides a means to rewrite the free energy objective function as a difference in free energies of the folded and unfolded states. In practice, this is the objective function which is minimized during the calculations.

**Eq. 2-21**
$$\Delta F = \langle F_{\mathrm{f}} \rangle - \langle F_{\mathrm{u}} \rangle = \langle E \rangle - \frac{1}{\beta} \frac{S}{k_b} - \langle F_{\mathrm{u}} \rangle$$

### 2.2.3. Lattice Approximation

We introduce the ability to perform a sequence design within the context of an arbitrary infinite lattice. Previous work[83] has detailed the calculation of a symmetric energy in the context of $M$ symmetrically related chains; we begin there and reduce those equations to

the energy of a single asymmetric unit within an infinite lattice.

The symmetry approximation imposes the following restrictions on all symmetrically related asymmetric units

1. The allowed monomer type probabilities on each unit (chain) are the identical, i.e. $w_t^m = w_t$ for all $m$ units

2. Conformer probabilities at equivalent sites on different units are equivalent, i.e. $w_c^m = w_c$ for all $m$ units

The approximation thus supposes that in lieu of designing all possible positions in the lattice, we instead symmetrically link equivalent positions to reduce the computational cost associated with a sequence free energy optimization. The total lattice energy, $\langle E_{lattice} \rangle$, can be written as the sum of all intramolecular energies associated with each asymmetric unit, $\langle E_f \rangle$, with the sum of all intermolecular energies between all asymmetric units, $\langle E_a \rangle$.

**Eq. 2-22**
$$\langle E_{lattice} \rangle = \langle E_f \rangle + \langle E_a \rangle$$

For $M$ asymmetric units in a lattice, the total intramolecular energy of all individual units is

**Eq. 2-23**
$$\langle E_f \rangle = \sum_m^M \left( \sum_i \gamma_i^m w_i + \frac{1}{2} \sum_{i \neq j} \gamma_{ij}^m w_i w_j \right)$$

where $\gamma_{ij}^m$ denotes the interactions between conformers $i$ and $j$ on asymmetric unit $m$. We assign the $\frac{1}{2}$ to avoid the double counting caused by the equivalency of $\gamma_{ij} = \gamma_{ji}$. The form of this energetic term is equivalent to the previous discussion of a mean field approximation

in the energy of a system, namely **Eq. 2-12**. We denote the individual energy of each asymmetric unit as $\langle E_{au}^{m} \rangle$. Furthermore, because the units are equivalent, that is $\langle E_{au}^{m} \rangle = \langle E_{au}^{n} \rangle$, we remove the superscript and factor out the number of asymmetric units, $M$.

**Eq. 2-24**

$$\langle E_f \rangle = \sum_{m}^{M} \langle E_{au}^{m} \rangle = \sum_{m}^{M} \langle E_{au} \rangle = M \cdot \langle E_{au} \rangle$$

The energy of interactions between each of the $M$ asymmetric units is the defined by all allowable interactions within the context of the symmetric approximation, and thus quantifies the intermolecular interaction between an asymmetric unit and all neighboring copies.

**Eq. 2-25**

$$\langle E_a \rangle = \frac{1}{2} \sum_{m}^{M} \sum_{n}^{M} \sum_{i} \gamma_{ii}^{mn} w_i + \frac{1}{2} \sum_{m}^{M} \sum_{n}^{M} \sum_{i \neq j} \gamma_{ij}^{mn} w_i w_j$$

where $\gamma_{ii}^{mn}$ denotes the interactions between copies of conformer $i$ between asymmetric units $m$ and $n$, and $\gamma_{ij}^{mn}$ denotes the interactions between conformers $i$ and $j$ on the asymmetric units $m$ and $n$, respectively. We assign the $\frac{1}{2}$ to both sums to avoid the double counting caused by the equivalencies $\gamma_{ii}^{mn} = \gamma_{ii}^{nm}$ and $\gamma_{ij}^{mn} = \gamma_{ji}^{nm}$ in the respective terms.

As before, we aim to factor out the number of asymmetric units, $M$, to create an intensive lattice energy. To do so, we assert that in the context of an infinite lattice, any two asymmetric units, $m$ and $m'$, have identical interactions amongst all $n$ neighbors.

**Eq. 2-26**

$$\sum_{n}^{N} \gamma_{ij}^{mn} = \sum_{n}^{N} \gamma_{ij}^{m'n}$$

such that

**Eq. 2-27**

$$\langle E_a \rangle = \frac{M}{2} \sum_{m}^{M} \sum_{i} \gamma_{ii}^{0m} w_i + \frac{M}{2} \sum_{m}^{M} \sum_{i \neq j} \gamma_{ij}^{0m} w_i w_j$$

We note that while the inversion of both pairs of indices provides equivalent interactions, that is $\gamma_{ij}^{mn} = \gamma_{ji}^{nm}$, the inversion of only one pair of indices does not map to identical pairs of conformers, $\gamma_{ij}^{mn} \neq \gamma_{ji}^{mn}$. As such, the summation is *not* further reduced by by $\frac{1}{2}$. By design, the singly inverted pair should have a symmetric interaction as placed in an ideal lattice, but do not qualify as double-counting. Returning to the expression for the lattice energy, we obtain

**Eq. 2-28**

$$\langle E_{lattice} \rangle = M \cdot \left[ \sum_{i} \gamma_i^m w_i + \frac{1}{2} \sum_{i \neq j} \gamma_{ij}^m w_i w_j + \frac{1}{2} \sum_{m}^{M} \sum_{i} \gamma_{ii}^{0m} w_i + \frac{1}{2} \sum_{m}^{M} \sum_{i \neq j} \gamma_{ij}^{0m} w_i w_j \right]$$

which can be factored to take a form analogous to **Eq. 2-12**

**Eq. 2-29**

$$\langle E_{lattice} \rangle \equiv U_{lattice} = M \cdot \left[ \sum_{i} \left( \gamma_i + \frac{1}{2} \sum_{m}^{M} \gamma_{ii}^{0m} \right) w_i + \frac{1}{2} \sum_{ij} \left( \gamma_{ij} + \sum_{m}^{M} \gamma_{ij}^{0m} \right) w_i w_j \right]$$

## 2.3. Constraints

While the theory is able to address sequence variability, it is often useful to impose constraints on the calculations. Constraints can address design requirements, experimental necessities, or aid in the selection of a sequence from diverse profiles. For example, where highly interacting and buried sites are able to resolve well defined sequence choices, solvent

exposed positions often boast diffuse distributions due to minimal interactions and the omission of an implicit solvent model. Constraints can take a variety of forms: physico-chemical energy functions, most commonly adapted from potentials developed for molecular simulations[55–58]; knowledge-based statistical functions, such as those used to quantify secondary structure propensities; simple compositional restrictions based on residue counts. The following is a summary of several of the constraints that have been implemented, several of which are used in later chapters.

### 2.3.1. Sequence Composition

The simplest way to impose a composition constraint is to specify the total occurrence of a specified amino acid in the sequence. Naively, this means constraining the total number of residues in the final sequence to a fixed value. More precisely, we formulate the total composition of a residue type $t^*$ across all sequences considered by writing the mean number of that residue in the ensemble.

**Eq. 2-30** $$f(\mathbf{w}, t^*)_{\text{composition}} = \sum_i \delta_{t^*} \cdot w_i \qquad \text{where} \qquad \delta_{t^*} = \begin{cases} 1 & \text{if } i = (n, t^*, c) \\ 0 & \text{if } i \neq (n, t^*, c) \end{cases}$$

where the delta function picks out probabilities with indices that belong to the specified type $t^*$.

A more complex variant of this constraint is to consider the composition across the set of amino acids. The sensitivity of the probabilistic calculations in the context of a molecular mechanics force field has a tendency to over select for a single type of residue (in part due to large side chain, more rotameric states, etc.). Imposing a required amount of diversity among the more probable sequences can alleviate this, and is often necessary when considering experimental requirements on NMR chemical shift assignment or protein expression[139]. An expression for sequence diversity can be obtained from an inverse participation ratio of

the above defined residue composition.

**Eq. 2-31**

$$f(\mathbf{w})_{\text{diversity}} = \sum_{t}^{T} \left( \frac{f(\mathbf{w}, t)_{\text{composition}}}{N} \right)^2$$

where we sum over all amino acid types in the ensemble, $T$. At one extreme, the diversity can be constrained to its lower limit such that only one residue is permitted, i.e. a homopolymer with $f(\mathbf{w})_{\text{diversity}} = 1$. At the other, the largest diversity can specify equal likelihood among the residues. For the natural 20 amino acids, the type composition produces a value of $\frac{N}{20}$, and in turn $f(\mathbf{w})_{\text{diversity}} = 0.05$. Often, the diversity constraint is applied such that it is lowered (the sequence diversified) from the unconstrained diversity value within the function limits described above.

### 2.3.2. Linear Sequence Properties

Constraints can be used to impose any variety of sequence properties dependent only upon the residue type (that is, independent of the residue conformation). These are often useful when targeting particular characteristics that must be met across a broad range of calculations. In the chapters to follow, such constraints are applied across a structural landscape such that (i) all designs are valid in the context of these characteristics and (ii) comparison across the landscape can be based purely upon the objective function or parts of it – namely, the sequence free energy or the mean field energy.

These constraints can be generalized:

**Eq. 2-32**

$$f(\mathbf{w}, t) = \sum_{i} \epsilon_t \cdot w_i$$

recalling that the index $i$ is a contraction of $(n, t, c)$. The coefficients $\epsilon_t$ depend only upon

the monomer/residue type.

**Mean Net Charge**

Obtaining a mean net charge across the ensemble of sequences requires establishing coefficients $q_t$ that reflect the charge of each amino acid. These can be obtained by specifying the value, or can leverage the partial charges inherent to atomistic force fields by simply summing across all atoms in a given conformer. This constraint can be imposed to enforce charge neutrality, or place bounds on the overall charge of the sequence.

**Eq. 2-33**

$$f(\mathbf{w})_{\text{Net Charge}} = \sum_i q_t \cdot w_i = 0$$

**Mean Molar Absorptivity**

Similarly, the mean molar absorptivity across sequences can be constrained to guarantee most probable sequences bear residues which can be tracked via spectroscopic methods. The molar extinction coefficient of a protein is estimated in an additive way.

**Eq. 2-34**

$$\epsilon_{\text{Protein}} = \sum_t \epsilon_t \cdot N_t$$

where for a protein in water measured at 280 nm, the coefficients (in $\text{M}^{-1}.\text{cm}^{-1}$) are $\epsilon_{\text{TRP}}$=5500, $\epsilon_{\text{TYR}}$=1490, $\epsilon_{\text{CYS}}$=125, and $\epsilon$=0 for all other residues. We express the con-

straint as an ensemble average

**Eq. 2-35**

$$f(\mathbf{w})_{\text{Extinction}} = \langle \epsilon_{\text{Protein}} \rangle = \sum_i \epsilon_{\text{Ext},t} \cdot w_i$$

**Mean Helix Propensity**

Such constraints can be formulated in a variety of ways. Linear functions based on the residue type along can leverage estimations of the protein pI[140], an aggregation scale[141], or secondary structure propensities[142–145]. For example, the incorporation of the O'Neil-DeGrado helix propensity scale[146,147] can be formulated as a mean over sequences in the following way

**Eq. 2-36**

$$f(\mathbf{w})_{\text{Helix Propensity}} = \sum_i \delta_n \cdot \epsilon_{\text{hp},t} \cdot w_i \qquad \text{where} \qquad \delta_n = \begin{cases} 1 \text{ if } n \text{ is a helical position} \\ 0 \text{ otherwise} \end{cases}$$

such that $\epsilon_{\text{hp},t}$ is obtained from the O'Neil-DeGrado helix propensity scale for type $t$ given that site $n$ is at a helical position as picked out by the delta function $\delta_n$. The constraint can be used to promote the formation of secondary structures given in the target fold (scaffold). Values can be estimated from databased derived averages over similar structures (helical regions); as the function form suggests, the value should scale linearly with the number of helical positions.

### 2.3.3. Environmental Exposure*

Adapting solvation models, either as an explicit or implicit addition[148–151] to the pairwise energy model employed in the probabilistic approach is difficult. While such models are commonplace in molecular dynamics simulations[152–155], we instead derive a linear knowledge-based constraint based on the overall solvent exposure of each residue. Coined as the "environmental energy" in previous work[36], the constraint estimates a protein density score, interpreted as a potential, based upon statistical sampling of amino acid propensities in regions of various $C_\beta$ density. The simple potential leverages the rationale that in native folds, certain amino acids are common to dense regions of the protein (hydrophobic residues in the core), while others are preferential to sparser areas (hydrophilic residues at surface exposed positions).

The formulation of the environmental energy is identical to the linear constraints detailed above. It is the mean value of the one-body coefficients $\epsilon_{\text{env}}$ associated with the potential.

**Eq. 2-37**
$$f(\mathbf{w}, t^*)_{\text{Environmental}} = \langle E_{\text{Environmental}} \rangle = \sum_i \epsilon_{\text{Env}}(t, \rho) w_i$$

Values for $\epsilon_{\text{env}}(t, \rho)$ are dependent only upon the residue type $t$ and a local $C_\beta$ density $\rho$ within the protein scaffold (backbone). They are statistically derived from a training set of 423 protein crystal structures (PDBs), consisting of hydrolases, transferases, isomerases, ligases, lyases, and oxidoreductases. Each of the structures is a single chain at least 40 residues in length with resolution $\leq$ 2Å. The maximum sequence identity between any sequences within the set is kept below 30 % to void sampling bias. These coefficients are

*Construction of updated potential and corresponding chain length correlation plot provided courtesy of Krishna Vijayendran.

constructed by sampling across all residues within these sequences as defined by

**Eq. 2-38**

$$\epsilon_{env}(t,\rho) = -\frac{1}{\beta_{env}} \ln \frac{p(t,\rho)}{p(t)p(\rho)}$$

where $p(t)$ is the probability of observing a given residue $t$ in the structure set, $p(\rho)$ is the probability of observing a given local $C_\beta$ density $\rho$ for side chain geometric center across the structure set, $p(t,\rho)$ is the joint probability of observing a given residue type $t$ with a local $C_\beta$ density $\rho$ across the structure set, and the inverse temperature is chosen as $\beta_{env}$ = 0.5. The local $C_\beta$ density is estimated as density of $C_\beta$ atoms within the free volume of a shell centered on a given rotamer's side chain geometric center, given as

**Eq. 2-39**

$$\rho(t) = \frac{n(C_\beta)}{\frac{4}{3}\pi R_c^3 - \langle V_{access}(t)\rangle}$$

where $n(C_\beta)$ is the $C_\beta$ count within a the specified sphere of radius $R_c = 8$Å, and $\langle V_{access}(t)\rangle$ is the mean accessible side chain volume for residue type $t$ across the structure set. Estimates of $\epsilon_{\text{env}}(t,\rho)$ were compiled into type dependent histograms, where density values accrued in bins of 0.003475044 $C_\beta/\text{Å}^3$ size; this corresponded to four midpoint centered bins in each histogram. Each histogram was fit to a fourth order polynomial to smooth estimates. Calculation of coefficients $\epsilon_{\text{env}}(t,\rho)$ in **Eq. 2-37** are then obtained from this polynomial for a given local $C_\beta$ density. As with the previous parameterization, these values are in agreement with other hydrophobicity scales[36]. As the environmental coefficients are independent of neighboring residue identities, this extrapolates to a simple linear model in the protein's total "environmental energy". As before[156], values for a total $E_{\text{Environmental}}$ were calculated for each of the 423 structures in the training set to build a linear mapping to a protein's total length. This linear relationship is shown in Figure 2-1.

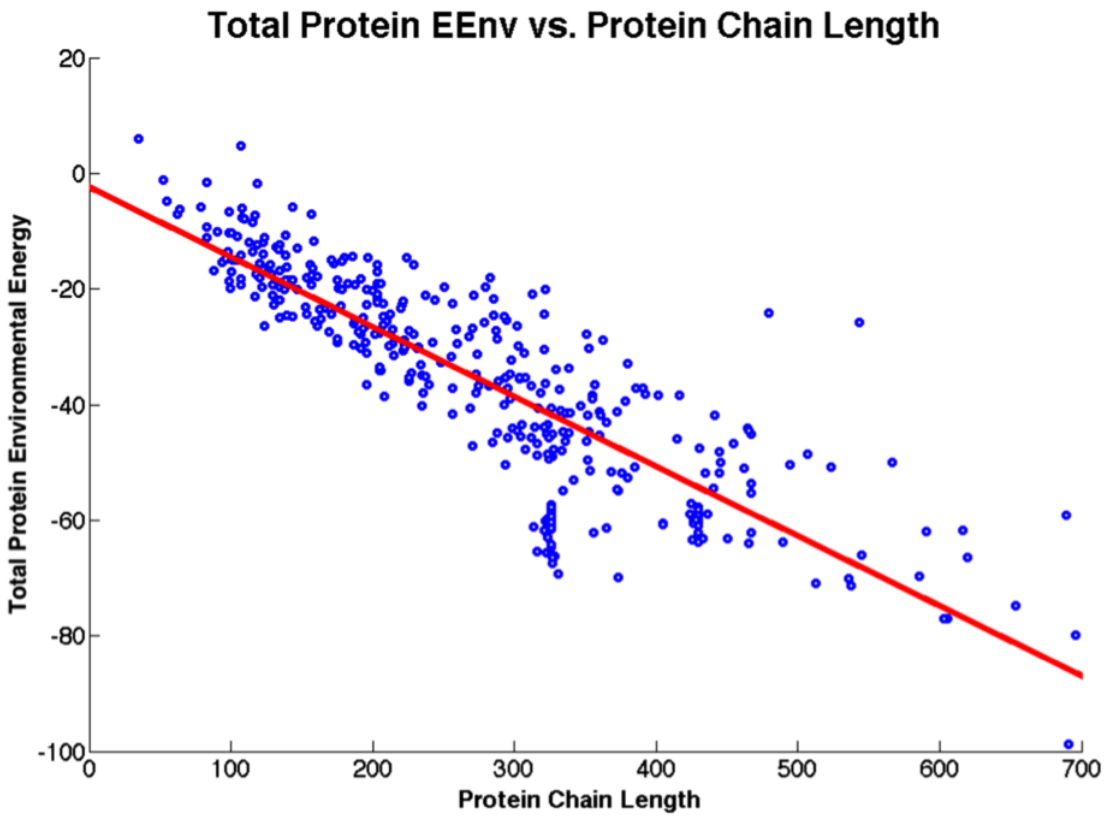



**Figure 2-1.** Linear correlation between Environmental Potential and the chain length of globular single-chain proteins. A linear fit (red line) to the data produces a model corresponding to the equation $E_{\text{Environmental}} = -0.1207n - 2.514$ where $n$ is the protein chain length.

## 2.4. Computational Elements

### 2.4.1. Minimization Routine

The VERGIL package solves the nonlinear programming problem in **Eq. 2-14** through a direct optimization of the effective free energy function subject to the set of constraints. We note that while the theory mimes a heterogeneous mean field theory, it is significantly easier to find minimizers utilizing numerical optimization routines in lieu of the traditional self-consistent approach. Furthermore, the theory allots for the addition of arbitrary constraints on the set of probabilities which can be easily implemented when solving the nonlinear programming problem.

To obtain a solution for the set of amino acid-conformation probabilities, we utilize the method of Lagrange multipliers. Namely, this requires defining a Lagrangian for **Eq. 2-14**

**Eq. 2-40**
$$V(\mathbf{w}) = U(\mathbf{w}) - \frac{1}{\beta} \cdot \frac{S}{k_b}(\mathbf{w}) - \sum_k \lambda_k (f_k(\mathbf{w}) - f_k^o)$$

and obtaining a stationary point, that is where the gradient of the Lagrangian is zero

**Eq. 2-41**
$$\vec{\nabla} V(\mathbf{w}, \vec{\lambda}) = \mathbf{0}$$

The current implementation relies upon the Interior Point OPTimizer (IPOPT) open-source software package for solving **Eq. 2-14**. IPOPT is a robust optimization package used to address general, large-scale nonlinear programming problems utilizing a primal-dual interior-point algorithm with a filter line-search. The interior-point method (or barrier method) introduces a logarithmic barrier term to the objective function to be minimized. For some

minimization,

**Eq. 2-42**

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t. } c(x) = 0 \\ &\qquad x \geq 0 \end{aligned}$$

the variable boundary constraints are reformulated into a term with some scalar $\mu$.

**Eq. 2-43**

$$\begin{aligned} &\min_x \varphi_\mu(x) = f(x) - \mu \sum_{i=1}^{n} \ln(x_i) \\ &\text{s.t. } c(x) = 0 \end{aligned}$$

For some positive non-zero value of $\mu$, this form as the objective function goes to infinity at the zero boundaries and guarantees than an optimal solution of **Eq. 2-43** must be within the bounds. Moreover, as $\mu$ is decreased to 0, an optimal solution of **Eq. 2-43** converges to the optimal solution of the original formulation **Eq. 2-42**. Thus the overall strategy of the algorithm is to solve a series of barrier problems each by decreasing $\mu$ until the optimality conditions are met.

The package has shown it is particularly suited for large problems (with up to millions of variables and constraints) and is able to handle both convex and non-convex programming problems. Both equality and inequality constraints are handled, where the package internally introduces necessary slack variables for inequality constraints. We emphasize that such programming problems are NP-hard, and as such IPOPT is only able to guarantee a local minimizer.

In practice, obtaining an optimal solution requires that a nonlinear programming problem

satisfies the Karush-Kuhn-Tucker (KKT) conditions. When the mean field energy is linear (that is, only considers one body terms), the free energy function is convex; provided that any additional inequality constraints are convex and equality constraints are affine, this assures that any local minimizer is a global minimizer. However, the introduction of two body terms into the mean field energy creates possible non-convexities in the energy function. This means that while we are able to obtain local minimizers of the Lagrangian, it is possible that other minima exist. Here, we utilize the theory to estimate probability profiles that are energetically consistent with a particular target structure instead of identifying the single-lowest energy sequence-structure. The degree to which solutions are affected by the mean field energy's non-convexity are controlled by the relative strength of the effective temperature term, $\beta$; we take care further on to detail appropriate choices for this parameter when identifying optimized probability profiles.

When multiple stationary points exist, it is important to recall that the success of the IPOPT solver depends upon choices in a starting point for the probabilities, as well as choices in the algorithm's method and convergence criteria. Based on performance testing, we have made the choice to use the monotone (Fiacco-McCormick) strategy to update the barrier parameter, $\mu$, instead of the adaptive update strategy. The relative convergence tolerance was set to $10^{-5}$, which terminates the optimization if the IPOPT optimality error function is smaller than this tolerance. All other algorithm settings, including the maximum number of iterations, were held at the default values.

Finally, it is of note that the implemented interface for IPOPT is able to accomodate the addition of any additional constraints, so long as they are twice differentiable. All objective function elements and constraint functions herein are implemented using analytic first and second order mixed partial derivatives. All such currently implemented constraints are linear and provide a sparse Jacobian matrix with respect to the optimization space which maintains the overall efficiency of the IPOPT solver.

### 2.4.2. Target Structure

We represent the target structure as a fixed scaffold upon which the sequence ensemble is constructed. That is, for any design considered, we constrain the common coordinates among monomer types located at a given site such that their molecular connectivity to neighboring monomers is consistent throughout the ensemble. This allows the identification of sequences that are consistent with the targeted folded structure.

In the case of protein design, this is a set of atomic coordinates that comprise the protein backbone that dictates the secondary, tertiary, and even quaternary structure of the fold. Despite limiting conformational readjustments that sequence mutations may impart upon the backbone, the specification of a target structure greatly reduces the computational dimensionality of the design problem. Furthermore, modeling an ensemble of sequences on an existing structural motif from a natural protein allows for the ability to leverage known structures to potentially achieve new functionalities. While these scaffolds can be obtained through *de novo* modeling, naturally occurring structures can be utilized, including those determined from x-ray crystallography and NMR spectroscopy in the Protein Data Bank (PDB) file format.

### 2.4.3. Sequence-Structure Degrees of Freedom

The dimensionality of the ensemble of sequences is explicitly the permissible degrees of freedom associated with each position in the target structure. In proteins, this constitutes the amino acid identities allowed at each of the backbone positions, as well as their respective side chain conformational states. These can include all type/conformation possibilities, a narrowed range of types (e.g., hydrophobic patterning of a core), or only specified conformations (a subset of the most probable states). Furthermore, the theory can incorporate non-standard types including ligands and cofactors, non natural amino acids, or even structurally well-defined water molecules such they they can be discretized to a site specific location in the ensemble of states.

For studies herein, we consider only the set of all 20 natural amino acids as possible monomer types in design calculations. Bonding information is obtained from standard force field topology information (Amber, Charmm, etc.) as to remain consistent with conventional atomic naming schemes. However, instead of relying up the topological construction routines found in such files, each amino acid is modeled from a residue template. Template structures were generated for each of the 20 natural amino acids by accruing mean values of all inclusive bond length, bond angles, and dihedral angles across a set of structures. The set was comprised of residues with full electron density and no partial occupancy from the high resolution HiQ54+ dataset[157].

The discretization of conformation states permits atomic resolution of a given monomer's variable structure to include the level of detail associated with directed function properties (binding, catalysis, etc.) or atomic interactions which stabilize the targeted folded structure. While it is possible to utilize any arbitrary set of discretized amino acid conformational states, it is often more useful to choose a subset of more probable conformation space that reduces the dimensionality of the design ensemble. For protein design calculations, this is achieved two-fold: conformational states are reduced to rotameric states (only variance of side chain dihedral angles) and selected from statistically significant values in representative protein structures. Furthermore, by inferring side chain states from structural databases, rotamer states are typically consistent with energy minima of molecular potential functions. The libraries of states used in subsequent calculations were taken from amino acid backbone-dependent rotamer libraries inferred from Dunbrack et. al.[43,47] These libraries model the smallest amino acids (glycine and alanine) with a single conformational state and most variable amino acids (e.g. glutamine) with several dozens of rotamer states specific to local backbone conformations.

In general, finer grain rotamer libraries and even full conformer libraries may be more able to capture atomic placement, as is necessary with small molecules in enzymes or other highly specific binding motifs. Likewise, when structure specific information is sparse (e.g.,

limited x-ray structures) or unavailable (e.g., non natural monomers), rough discretization may be necessary.

### 2.4.4. Energy Functions

The population of the coefficients in **Eq. 2-12** relies upon conformer specific energies. These energies can be estimated in a variety of ways, from simplified, coarse-grain energy functions[158,159] to detailed, all-atom potentials[55,56]. For the general energy function applied to the minimization in **Eq. 2-14**, an atomistic potential is more prone to capture sequence-structure compatibility at sequence free energy minima. Atomistic potentials can identify highly complementary noncovalent interactions that highlight energy stabilizing packing associated with a particular structure.

Including such an explicit atomistic potential can be most easily achieved by harnessing established classical molecular force fields as developed for molecular simulations. The form of **Eq. 2-12** dictates so called one-body ($\gamma_i$) and two-body ($\gamma_{ij}$) terms. For all calculations performed, we consider atomistic one-body terms to include all interaction terms of the potential for atoms contained within a particular conformer $i$; this is a means of assaying the energetic stability of the isolated monomer unit in its present conformation. We consider atomistic two-body terms to be the sum of all pairwise potential terms between conformer $i$ and $j$, including all nonbonding terms as well as bonding terms that may span the two conformers, i.e. dihedral potential of neighboring residues on a protein chain. In many of the classical molecular force fields, multi-residue dihedrals ($\phi, \psi$) depend upon the identity of the residue on not simply the position of the backbone atoms. The two-body terms are able to capture interactions with other conformations of side chains in the ensemble as well as interactions with the fixed scaffold and any other fixed moieties in the design.

Calculations performed herein utilized a modified version of the united-atom Amber force field[55]. The standard Amber84 force field is stripped of all polar hydrogens, and partial charges absorbed into the corresponding heavy atoms. Van der Waals radii for heavy polar

atoms are kept unchanged and united carbons use the smaller van der Waals radii specified by Dunfield et al.[160] (not the larger set as specified by the scaled Jorgensen et al.[161] values). To compensate for the removal of polar hydrogens, the energy function uses a simplified version of the hydrogen bonding potential developed by Kono[44] using parameters outlined by Stickle[162] with scaled minimum hydrogen bonding energy of -1 kcal/mol. Electrostatic interactions utilize a distance dependent dielectric of $4r$. All nonbonded interactions introduce a hard cutoff at 8 Å.

While it is possible to utilize other standard molecular mechanics force fields (Charmm, OPLS, etc.), we focus on the the described Amber-derived fully united atom force field to remove the complexities of hydrogen placement while still being able to capture energetics associated with hydrogen bonding. Future considerations include the usage of alternate force fields, as well as alternate means of assaying potential hydrogen bonding.

# 3 De Novo Design of a Uranyl Binding NanoBio Matrix

Extraction of radioactive heavy-atom derivatives like the uranyl cation ($UO_2^{2+}$) through a highly selective ligand binding is of environmental relevance; immediate concerns include using adsorbent materials for sequestration to serve in environmental cleanup[163–165], as well as extraction from sea water for sustaining the nuclear fuel cycle[166]. Well-known uranyl-binding motifs offer distinctive handholds for rational design of proteins that selectively bind uranyl[167,168]. Here we present a methodology for the *de novo* protein design of a peptide sequence capable of forming a trimeric bundle with multiple uranyl binding sites at it core. The sequence is tailored within the P6 space group to promote crystallization of the protein-uranyl complexes, building on the recent success in designing a targeted protein crystal[90].

## 3.1. Introduction

Uranium and its derivatives are key components in the nuclear fuel cycle. Developing efficient binding materials to harvest uranium is one way to address the necessary remediation near rare earth processing sites. Rare earth metals are critical components to a number of modern technologies, including wind turbines, hybrid electric vehicles, and defense applications. It is their unique chemical properties than make them prime for use in solid oxide fuel cells, superconductors and laser technology[169]. However, rare earth extraction, separation, and refining operations generate radioactive thorium and uranium bearing waste susceptible to leaching, mobilization, and distribution into the environment in more water/air labile states[165]. The increasing demand for heavy metals has led to radioactive waste concentrations above natural levels, raising concerns about ecotoxicity around rare earth processing operations[164]. As such, it is of environmental significance to identify means by which to efficiently capture and remediate uranium[163].

The uranyl cation is the most stable, common aerobic form of uranium for which there are well known uranyl-binding motifs. A variety of approaches have targeted sequestering the uranyl form in aqueous media, including functionalized polymers with amidoxime ligands[170,171], small-molecules[172], sophisticated chelating ligands[173,174], metal-organic frameworks[175], and proteins[168,176,177]. Proteins offer the distinct advantage of leveraging examples of efficiently and selectively bound protein-metal assemblies from nature. Furthermore, proteins can be incorporated into biological systems (e.g. bacteria, plants) capable of regenerating the protein at minimal cost. Zhou and coworkers recently reported the computational identification and rational development of a thermally stable uranyl-binding protein which offers a high $K_d$ of 7.4 femtomolar (fM); the protein has a significant selectivity over other metal ions, including those commonly found in seawater[168].

Protein design has been utilized to engineer a wide range of unique metalloproteins with novel functions[13,178–184], including significant work regarding the incorporation of metals and nonbiological cofactors[30,31,65,88,89]. We aim to couple these strategies to protein design methods that select protein sequences which crystallize in a chosen lattice arrangement. Lanci et. al. employed computational design to select a peptide sequence that crystallizes in a predetermined three dimensional lattice[90]. A three helix coiled-coil was designed for the polar, layered P6 space group, and was confirmed to agree with the model structure at the subangstrom level ($C_\alpha$ RMSD $< 0.7$ Å). The crystal design approach is able to capture the commonly weak intermolecular forces that stabilize crystalline ordering, and is easily extendable to the positioning of symmetric cofactors or guest molecules within a lattice framework. By applying such a strategy to a uranyl binding peptide, we are able to target (a) the design of a crystalline nano-bio matrix which is activated simply upon the addition of uranyl, as well as (b) a means to assess our design capability at atomistic resolution via x-ray structure determination.

This work employs probabilistic protein design to engineer an assembly for efficiently binding the uranyl cation. The construction of such an assembly presents an opportunity to

investigate uranyl binding, coordination, structure, and sequestration, potentially offering insight into improving the removal and cleanup of waste sites and harvesting for the nuclear fuel cycle.

## 3.2. Overview of Design Strategy

Here we seek to build upon the protein crystal design work of Lanci et. al.[90] to design peptide trimer assemblies able to bind multiple uranyl cations throughout the trimer core. Computational methods identify amino acid positions that can accommodate binding uranyl within a coil-coil structure, and the subsequent designed sequences target a specific crystalline lattice. Designed sequences are currently being evaluated experimentally.

The overall design procedure employed herein is as follows: From (1) a uranyl-glutamate super-rotamer modeled after a uranyl-acetate crystal structure, (2) identify peptide trimer structures which accommodate a uranyl-glutamate binding site to be replicated at multiple core locations in the bundle core. From trimer candidates with suitable binding geometries, (3) identify low energy structures with hydrophobic residues comprising the remainder of the core. (4) Upon placing the lowest energy trimer structure in the P6 space group, perform full sequence design on the remaining exterior positions. If necessary, (5) introduce sequence constraints to address biological rationales and experimental concerns. A visualization of these steps can be found in Figure 3-1.

The author thanks and acknowledges the significant contribution made by Christopher M. MacDermaid to this particular project. He devised much of the design strategy described hereafter, provided a first generation of sequences submitted for experimental verification, and much guidance in redesigning this system[185].

## 3.3. Uranyl Binding Geometry and Parameterization

The uranyl cation is known to accommodate up to six equatorial ligands in a overall hexagonal bipyramidal geometry; this can easily be obtained with binding to either three aspartate

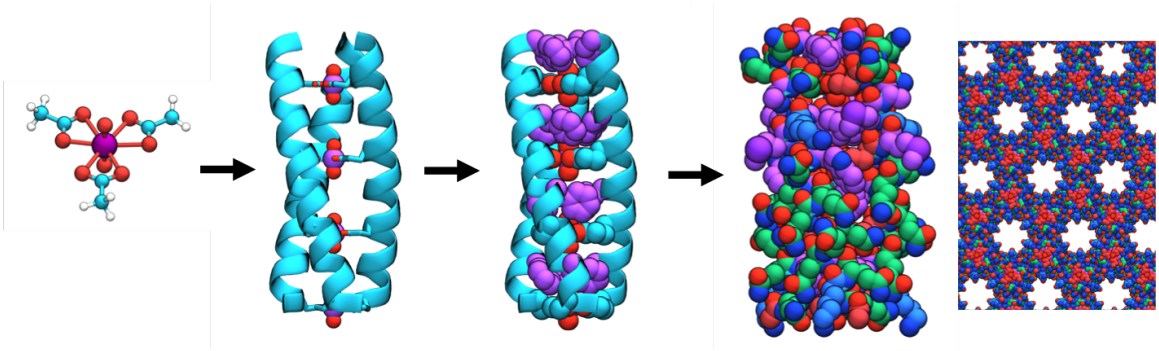


**Figure 3-1.** Overview of the design for a uranyl binding protein crystal. Each step is fully elaborated in subsequent sections, which identify: (1) a model uranyl binding motif, (2) a peptide coiled-coil trimer accommodating multiple interior binding sites, (3) an optimized trimer core, and (4) a full sequence in the context of a predetermined space group.

or glutamate amino acids in a 3-fold planar arrangement. Given glutamate has two torsional degrees of freedom, which should provide a larger subset of potential symmetric binding motifs inside a coiled-coil structure than accessible with aspartate, a GLU-uranyl super-rotamer was created. This was accomplished by overlaying the glutamate carboxylate plane onto acetate in the crystal structure obtained for uranyl acetate. The $UO_2(Ac)_3$ structure was obtained from the Cambridge Crystallographic Database[186]. Distances between the uranium and glutamate carboxylate oxygens were set to 2.49 Å, and the carboxylate plane was enforced to be perpendicular to the linear uranyl (Figure 3-2).

Non-bonding parameters for the uranyl were obtained from Lins et al.[187] and converted to units used by the AMBER force field, as listed in Table 3-1. Topologies and parameter for the GLU were taken from the AMBER84 force field, as modified by Kono et al. to exclude all explicit hydrogen atoms[36]. These were then combined into a single residue with a topology specifying the coordination between the glutamate carboxylate oxygens and uranium as bonds (Figure 3-2).

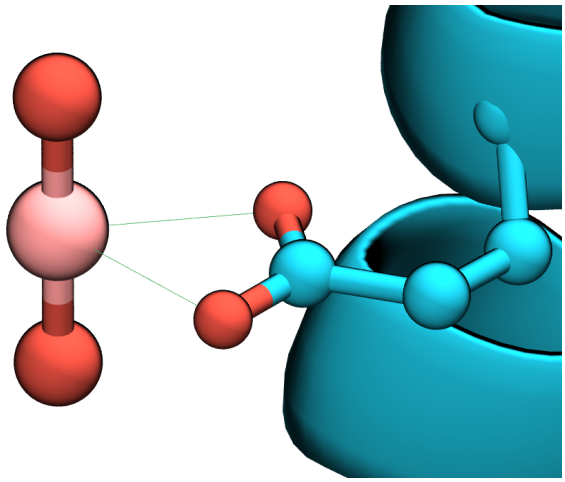


**Figure 3-2.** Glutamate-Uranyl super rotamer rendering, as attached to an alpha helix. The coordination between the glutamate carboxylate oxygen and the uranium atom considered a bond by the force field (green lines). Additionally, the plane spanned by the uranium atom and glutamate carboxylate oxygens is strictly perpendicular to the vector spanned by the uranyl molecule.

| Atom | $q_i$ ($e$) | $\epsilon_i$ (kcal/mol) | $R_{min}/2$ (Å) |
|---|---|---|---|
| U | 2.50 | 0.131902 | 1.768 |
| OU1 | −0.25 | 0.151898 | 2.06 |
| OU2 | −0.25 | 0.151898 | 2.06 |

**Table 3-1.** Uranyl parameters as used by the AMBER force field. Listed are the partial charge on the atom ($q_i$), the depth of the van der Waals potential well ($epsilon_i$), and half the radius of van der Waals potential depth ($R_{min}/2$). These parameters have been converted from the values reported in Lins et al.[187]

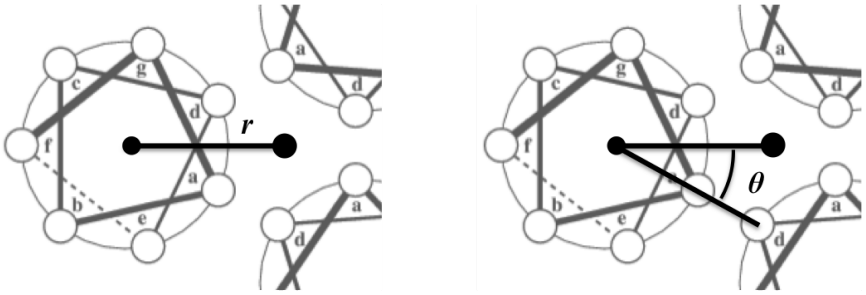



**Figure 3-3.** Diagram of varied trimer parameters. (Left) The superhelical radius, $r$, defined as the distance from the coiled-coil axis to the alpha-helical axis. (Right) The minor helical rotation, $\theta$, defined as the rotation of the first alpha carbon about the alpha helical axis. For this definition, $\theta = 0\,^{\circ}$ corresponds to the first alpha carbon on the alpha helix directed at the coiled-coil axis.

## 3.4. Modeling Uranyl Binding at the Core of a Coiled-Coil

### 3.4.1. Generating Coiled-Coil Scaffolds

The design method targets placement of multiple potential binding sites for the uranyl on the interior of a trimeric coiled-coil structure – each binding site should comprise three symmetric GLU at interior positions for the uranyl to satisfy an analogous complex to the uranyl acetate structure. The previous success in designing a trimeric coiled-coil crystal structure[90] was used a starting point for specifying criteria for coiled-coil scaffold geometries. Each alpha helix in the P6 coiled-coil protein crystal structure was comprised of 26 residues (positions 2-27) with an acetyl N-terminal group (positions 1) and an amidated C-terminal capping group (positions 28). If uranyl coordinating motifs are positioned every seven residues, such a helical bundle structure should then afford 4 potential binding locations (more than 3 heptads). The identification of potential scaffolds (backbone coordinates) was achieved *de novo*, and structures were generated using the well know Crick parameterization of coiled-coil structures[188]. Fitting the previous designed crystal structure trimer unit (PDB Code: 4DAC)[90] to a parallel coiled-coil provided a Crick parameterization with a rise per residue of 1.518 Å and superhelical pitch of 128.0 Å.[189].

Each trimeric coiled-coil was generated with a computational builder. The builder places

alpha carbons along the alpha helical pathway as described by the Crick parameterization, and then adds the remaining protein backbone heavy atoms for poly-L-glycine. For a specific superhelical radius ($r$) and helical rotation ($\theta$), the rise per residue and superhelical pitch ($P_o$) were fixed to values of the aforementioned crystal structure fit, and the minor alpha helical radius was fixed to 2.26 Å(i.e. a standard coiled-coil). The Fraser MacRae constraint[190] was imposed upon the residues per turn (**Eq. 4-1**), $\rho$, (based on the current superhelical radius, superhelical pitch, and rise per residue) to ensure an exact heptad repeat throughout the trimeric coiled-coil. No axial offset ($\Delta Z$) was applied. All helices were varied symmetrically, such that all minor helices positioned symmetrically related starting heptad positions with respect to the coiled-coil axis. For a full description of coiled-coil generation, please see Section 4.3.1.

### 3.4.2. Glutamate Rotameric States That Satisfy Trimeric Uranyl Binding Symmetry

The GLU-uranyl super rotamer was placed at the third residue position (counting the acetyl group as the first position) and all other amino acid positions were fixed as glycine. Additionally, all minor helical rotations were adjusted so that the third alpha carbon position was rotated to the equivalent rotation of the first alpha carbon position (Figure 3-3). With the heptad repeat is enforced, this is accomplished by an additional rotation of $\theta$ through 2 positions of a heptad ($4\pi$) of $-\frac{2}{7} \cdot 2 \cdot 360\,^\circ$.

For each coiled-coil considered, acceptable binding geometries were obtained by altering the three dihedral angles associated with the glutamate $[\chi_1, \chi_2, \chi_3]$ in the super rotamer (Figure 3-4). Acceptable symmetric binding orientations of the super rotamer consist of structures where the uranyl cation lies on the axis of the coiled-coil (here, the z-axis). By applying the C3 symmetry to a generated alpha helix in the coiled-coil context, three glutamates can be placed to fully satisfy the six equatorial ligation points on the uranyl molecule. The uranyl maintains a locked orientation with respect to the glutamate, and as such, the simplest way to achieve such orientations of the glutamate is to systematically vary the dihedrals until

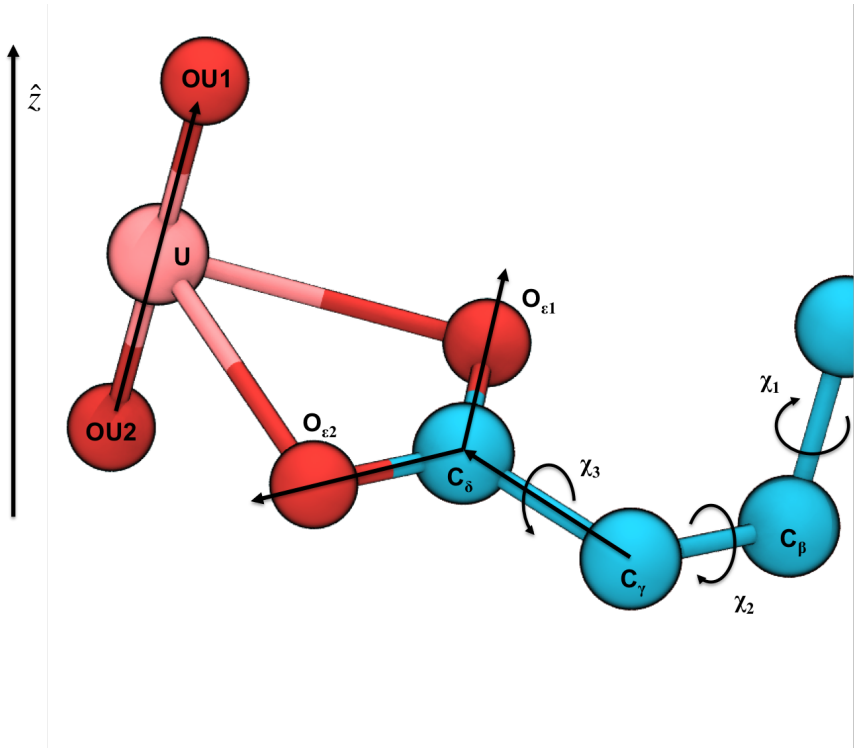



**Figure 3-4.** Atomic diagram of the GLU-uranyl super rotamer with vectors utilized in solving for $\chi_3$. The atoms in the GLU-uranyl super rotamer are labeled according to the AMBER molecular topology, along with the uranyl naming. Each of the dihedral angles of the glutamate are illustrated, $(\chi_1, \chi_2, \chi_3)$. For a particular set of dihedral values for $(\chi_1, \chi_2)$, **Eq. 3-1** is used to solve for the value of $\chi_3$ that most aligns with the $z$ axis.

the uranyl molecule aligns to the z-axis.

For each state of $\chi_1$ and $\chi_2$, there exist exactly two (symmetric) values of $\chi_3$ that maximize the orientational alignment of uranyl with the z-axis. By solving for this optimal value of $\chi_3$, the complexity of the *GLU* rotamer search is reduced from $N^3$ rotamer states to $N^2$ rotamer states where $N$ is the number of allowed values for each dihedral angle. The solution to $\chi_3$ is obtained by simply enforcing that for a positioning of $C_\delta$ (i.e. for some $\chi_1, \chi_2$), the dihedral between the uranyl ($\vec{b_1}$ vector between OU1 and OU2) and the z-axis is minimized. As detailed in Figure 3-4, the axis of rotation for the dihedral is simply the bond between $C_\gamma$ and $C_\delta$. The value of rotation ($\chi_3$) about the $C_\gamma$-$C_\delta$ bond is defined such

that the uranyl vector, $\vec{b_1}$, is brought parallel to the z axis.

**Eq. 3-1**

$$\vec{b_1} = \overrightarrow{A_{OU1} - A_{OU2}} = (\overrightarrow{A_{C\delta} - A_{O\epsilon 1}} \times \overrightarrow{A_{C\delta} - A_{O\epsilon 2}})$$
$$\vec{b_2} = \overrightarrow{A_{C\gamma} - A_{C\delta}}$$
$$\chi_3 = atan2(\vec{b_1} \cdot (\hat{b_2} \times \hat{z}), (\vec{b_1} \times \hat{b_2}) \cdot (\hat{b_2} \times \hat{z}))$$

where $\overrightarrow{A_i}$ corresponds to the 3-dimensional coordinates of the given atom, and $\chi_3$ uses the well known computational function for the dihedral value between three non-collinear vectors. Note that $\vec{b_2}$ is normalized in **Eq. 3-1**, denoted as $\hat{b_2}$. Furthermore, the definition of $\vec{b_1}$ explicitly acknowledges the planar orientation of the super rotamer, which equates the uranyl molecular vector with the normal vector to the carboxylate plane.

For each coiled-coil structure, $\chi_1$ and $\chi_2$ were scanned over the full angular range $[-180\,^\circ, 180\,^\circ]$ with $\Delta\chi_i = 1\,^\circ$, solving for the optimal value of $\chi_3$ using **Eq. 3-1** (32,400 orientations of GLU-uranyl per coiled-coil structure). The root-mean-squared deviation (RMSD) between the uranyl and the z-axis was then evaluated with

**Eq. 3-2**

$$RMSD(\hat{z}) = \sqrt{\frac{1}{N}\sum_{i}^{N}(A_i \cdot \hat{x})^2 + (A_i \cdot \hat{y})^2}$$

where $\hat{x}$ and $\hat{y}$ are the x- and y-axes respectively, and for $i \in \{\text{U, OU1, OU2}\}$ with $N = 3$. For an RMSD$(\hat{z}) = 0.0$, the uranyl sits exactly on the z-axis, providing perfect superposition when the C3 symmetry is applied to the entire monomer unit. If the RMSD$(\hat{z}) \leq 0.1$ Å, the rotameric state $[\chi_1, \chi_2, \chi_3]$ was added to a particular coiled-coil's ensemble of GLU-uranyl rotameric states. If the current coiled-coil allowed a non-zero set of GLU-uranyl states, then the remaining binding site residue positions of the alpha helix were typed with the GLU-uranyl super rotamers (sites 3, 10, 17, 24). This ensemble was then subjected to

| Atom | Scaled $q_i$ ($e$) | Scaled $\epsilon_i$ (kcal/mol) | $R_{min}/2$ (Å) |
|---|---|---|---|
| U | 0.83333 | 0.002931 | 1.768 |
| OU1 | −0.08333 | 0.016878 | 2.06 |
| OU2 | −0.08333 | 0.016878 | 2.06 |

**Table 3-2.** Scaled uranyl parameters as used by the AMBER force field, as to accomodate overlapping uranyl molecules. Listed are the partial charge on the atom ($q_i$), the depth of the van der Waals potential well ($epsilon_i$), and half the radius of van der Waals potential depth ($R_{min}/2$).

energetic calculations both for the single chain and the symmetric trimeric coiled-coil unit, as described below.

### 3.4.3. Symmetric Energy Function Modifications

As each GLU-uranyl super rotamer was to be placed on a three-fold axis (the center of the coiled-coil trimer), the non-bonding terms were overcompensated for by a factor of 3 given three uranyl molecules are overlaid in the crystal structure. For partial charges, this simply means dividing each partial charge by 3. For the van der Waals 12-6 potential, $\epsilon_i$ is divided by 9 – this is due to the AMBER geometric combining rules for $\epsilon_{ij} = \sqrt{\epsilon_{UO_2} * \epsilon_j}$ which we want scaled by a third.

To prevent collision energies between symmetric GLU-uranyl super rotamers, which were targeted to overlap in the final crystal structure, the symmetric self interaction for GLU-uranyl was ignored in all calculations. As described by **Eq. 2-29**, we denote $\gamma_{ii}^{0m}$ as the interaction between conformer $i$ in the asymmetric unit and that same conformer in $m$th symmetric copy; this energy coefficient was set to 0 if conformer $i$ is a GLU-uranyl super rotamer. Justification for the omission of this term arises from the distinction that we use a known binding geometry for uranyl, viz. the uranyl acetate crystal structure. Additionally, the initial domain trimming routine was modified in a similar way; for clashing pairs of symmetric rotamers, we excluded the removal of GLU-uranyl super rotamer states to ensure they were part of the calculation.

### 3.4.4. Coiled-Coils Which Accommodate Uranyl Binding

The coiled-coil ensemble considered accesses a wide variety of beta carbon placements for the glutamate interior positions by scanning through the superhelical radius, $r$, and the minor (alpha) helical rotation, $\theta$. These parameters are defined in Figure 3-3. Values for the superhelical radii were drawn from observed parallel trimeric radii for natural structures[189], 6.5 Å $\leq r \leq$ 8.5 Å, incremented at $\Delta r = 0.02$ Å. The minor helical rotation was scanned from $-90.0\,^\circ \leq \theta \leq +90.0\,^\circ$, incremented at $\Delta\theta = 0.5\,^\circ$, which roughly corresponds to sweeping from the $g$ to $e$ helical wheel positions (Figure 3-5). That is, $\theta$ is varied from $-90.0\,^\circ$ near the *ge'*-interface at the outer-edge of the coiled-coil core, through the $a$ and $d$ positions within the core, and finally out near the *eg'*-interface on the other outer edge at $-90.0\,^\circ$.

For each identified coiled-coil that accommodated an acceptable binding position, an ensemble energy average was calculated in the context of C3 symmetry by placing the monomer in the P3 space group ($a = b \neq c$; $\alpha = \beta = 90\,^\circ$; $\gamma = 120\,^\circ$). To ensure the elements in the unit cell (a single trimer) had no interactions with neighboring cells, the unit cell dimensions were dramatically enlarged ($a = b = c = 100$ Å). Values of the internal energy were obtained by minimizing the sequence free energy at $\beta = 0.5$ (mol/kcal). Potential terms used in the energetic calculation include the dihedral potential from the AMBER84 forcefield, van der Waals (Lennard-Jones) and electrostatic potentials from a modified AMBER84 forcefield[36] (absorption of polar hydrogens into associated heavy atoms), and a 12-10 hydrogen bonding potential[36].

Figure 3-6 details the full energetic landscape of the poly-L-glycine trimer with four glutamate-uranyl binding sites. The landscape indicates the region of coiled-coil structures that permit the uranyl binding site, which is symmetrically shaped and highly restrictive. Trimers with larger bundle radii bear a GLU-uranyl orientation which points more directly at the core; for structures with smaller radii, the GLU-uranyl instead adopts an offset position, where the C$\alpha$-C$\beta$ bond points tangential to the core allowing the GLU to twist back to the bind-

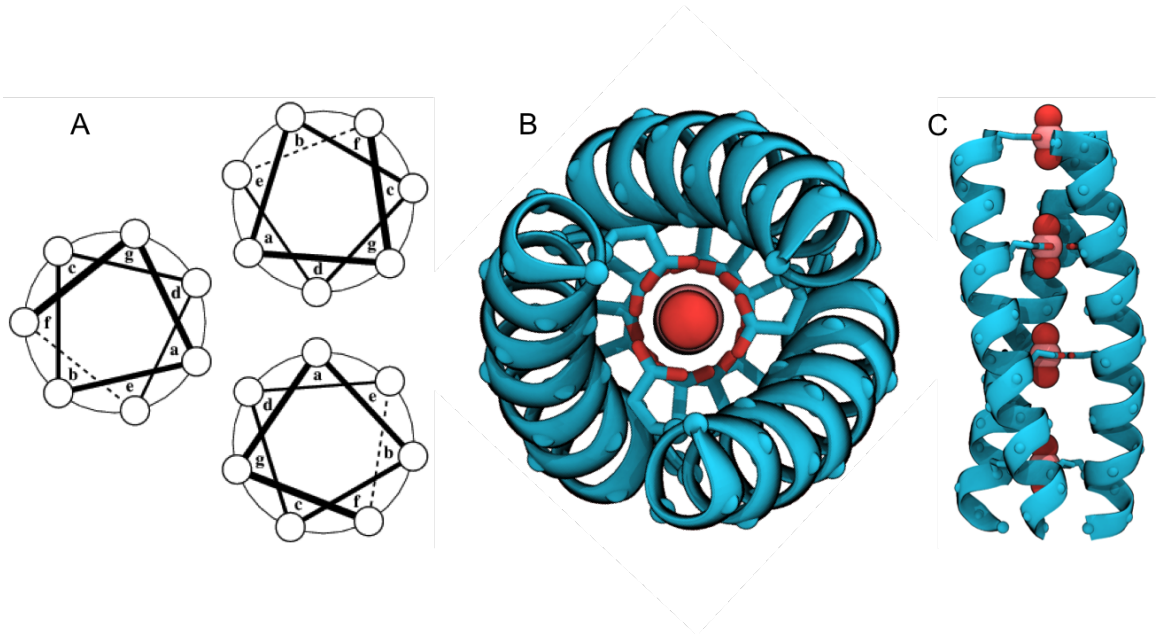



**Figure 3-5.** (A) Coiled-coil representation of the parallel trimer, indicating the heptad repeat. Each heptad position is denoted by the lower case letters *abcdefg*. The structural variation of the trimer changes $\theta$, by rotating the helices about their individual axes $\pm 90\,^{\circ}$ with respect to zero position where the alpha carbon points at the coiled-coil center; this effectively allows potential positions for the GLU-uranyl super rotamer at the *ge'*, *da'*, *ad'*, and *eg'* interfaces. (B) Top-down and (C) side views of one such trimer that accommodates the uranyl binding motif. Alpha carbons are rendered as spheres on the helical backbone to show the glutamate placement is close to the *a* position.

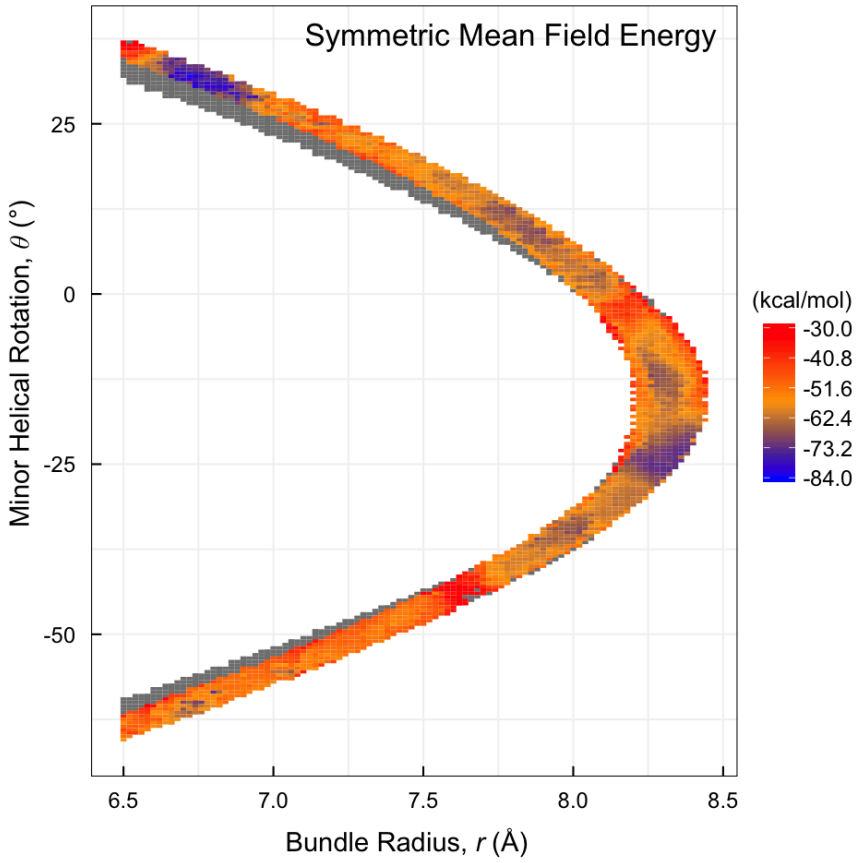



**Figure 3-6.** Mean Field Energy of the poly-L-glycine trimer with GLU-uranyl binding sites, using the dihedral, van der Waals, hydrogen bonding and electrostatic potentials. White space denotes coiled-coils that do not exhibit an orientation of the GLU-uranyl super rotamer that satisfies a uranyl-z axis alignment RMSD($\hat{z}$) $\leq$ 0.1Å. Gray tiles indicate energies above -30.0 kcal/mol. The global minimum is at $r = 6.72$ Å and $\theta = 32.0\,^{\circ}$.

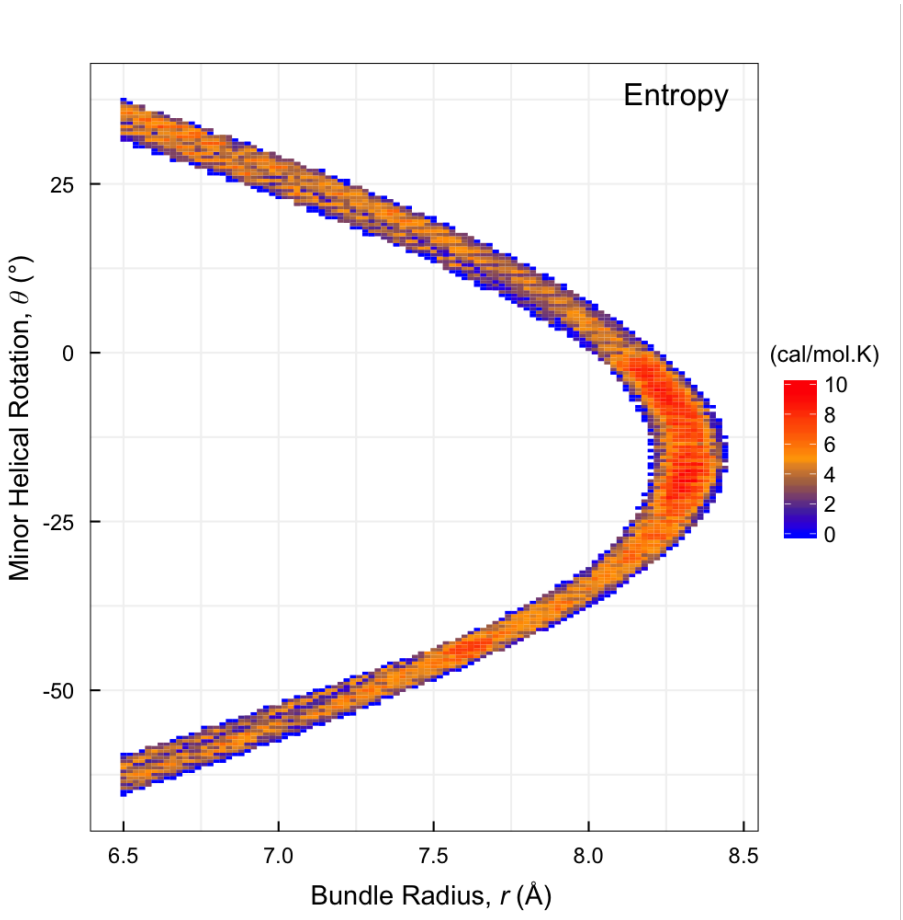



**Figure 3-7.** Entropy of the poly-L-glycine trimer with GLU-uranyl binding sites. White space denotes coiled-coils that cannot find an orientation of the GLU-uranyl super rotamer that satisfies a uranyl-z axis alignment RMSD($\hat{z}$) $\leq$ 0.1Å.

ing site. The global energy minimum lies in the positive rotation region at $r = 6.72$ Å and $\theta = 32.0\,^{\circ}$. The most probable GLU-uranyl super rotamer state in this coiled-coil consists of dihedral values $[\chi_1 = -136.0\,^{\circ}; \chi_2 = 66.0\,^{\circ}; \chi_3 = -8.7^{\circ}]$.

Figure 3-7 shows the entropy associated with each ensemble of satisfactory GLU-uranyl super rotamer states. The landscape possesses expected characteristics; namely, the edges of the landscape are low entropy regions (few super rotamer states in the ensemble) while the the highest entropy region lies at large radii which permit the most states. The coiled-coil structure at the energetic minimum ($r = 6.72$ Å, $\theta = 32.0\,^{\circ}$) has an value of 4.73 cal/mol.K.

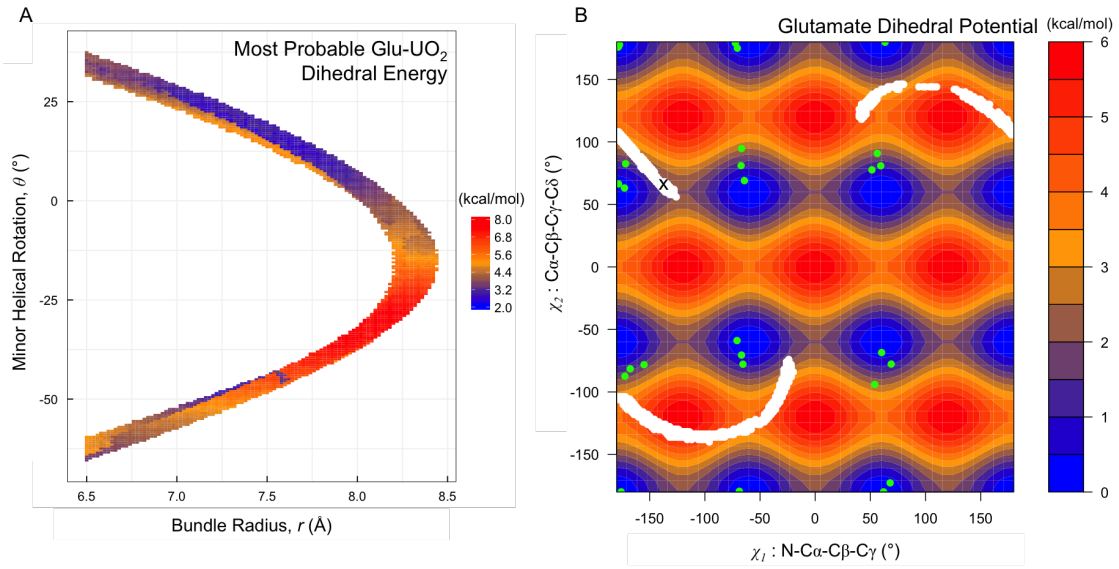



**Figure 3-8.** (A) Total dihedral energy for each of the most probable GLU-uranyl for each coiled-coil satisfying the positioning criteria. White space denotes coiled-coils that cannot find an orientation of the GLU-uranyl super rotamer that satisfies a uranyl RMSD($\hat{z}$) $\leq$ 0.1Å. (B) Dihedral potential map across the first two dihedral states in glutamic acid. The ensemble of rotational states found for the GLU-uranyl super rotamer in A are marked in white. The 2002 Dunbrack rotamer library[43] glutamic acid rotamers are marked in green. The rotamer corresponding to the lowest energy in 3-6 is marked at $[-136.0\,°; 66.0\,°]$

Characterization of the glutamate dihedral state underscores the global energetic minimum choice at $r = 6.72$ Åand $\theta = 32.0\,^{\circ}$. Figure 3-8A projects full dihedral potential of the most probable GLU-uranyl super rotamer onto the surface of the coiled-coil parameters ($r$ and $\theta$). The upper 'arm' (positive rotations of $\theta$) collects the lowest energy conformations; the super rotamer configuration at $[-136.0\,^{\circ}; 66.0\,^{\circ}; -8.7^{\circ}]$ has a dihedral energy of 3.4 kcal/mol. Figure 3-8B places all accepted super rotamer states collected in the coiled-coil search on the potential surface associated with the two primary dihedrals of the glutamate (white). As a means of reference, the dihedral states found in the Dunbrack 2002 rotamer library[43] are highlighted in green – it is clear that the low z-axis RMSD configurations identified do not overlap with these states. The GLU-uranyl configuration at $[-136.0\,^{\circ}; 66.0\,^{\circ}]$ sits near the saddle point between the $[-180.0\,^{\circ}; 60.0\,^{\circ}]$ and $[-60.0\,^{\circ}; 60.0\,^{\circ}]$ wells in the potential surface. As an additional measure, performing a Rotamer Analysis on this structure with MolProbity[191] did not flag the glutamate rotamer state as unacceptable $[-136.0\,^{\circ}; 66.0\,^{\circ}; -8.7^{\circ}]$, despite assigning a low rotamer percentile score of 1.0%.

It is worth noting that energetic landscapes for subcomponents of the total internal energy highlight other potential coiled-coil parameter candidates. Figure 3-9A depicts the symmetric mean field energy landscape in the absence of electrostatic interactions. The global energetic minimum corresponds to the coiled-coil at $r = 6.82$ Åand $\theta = -59.5\,^{\circ}$, placing the GLU-uranyl in the negative rotation region at the $g$ helical position. While this structure minimizes the van der Waals contact and provides the appropriate binding motif for the uranyl, the dihedral potential of the glutamate rotamer is less favorable than the full potential global minimum. At $r = 6.82$ Å and $\theta = -59.5\,^{\circ}$, the most probable GLU-uranyl super rotamer state is $[-172.0\,^{\circ}; -112.0\,^{\circ}; -130.1^{\circ}]$. This both exists in the higher energy region of the dihedral potential (Figure 3-8A) and is flagged by MolProbity during the Rotamer Analysis with a rotameric percentile score of less than 0.1%. Investigation of other rotameric states nearby have similar near-zero MolProbity rotamer percentile scores, suggesting the ensemble of glutamate binding rotamers at the $g$ position take on non-natural configurations. Conversely, the complete energetic landscape for only the asymmetric unit

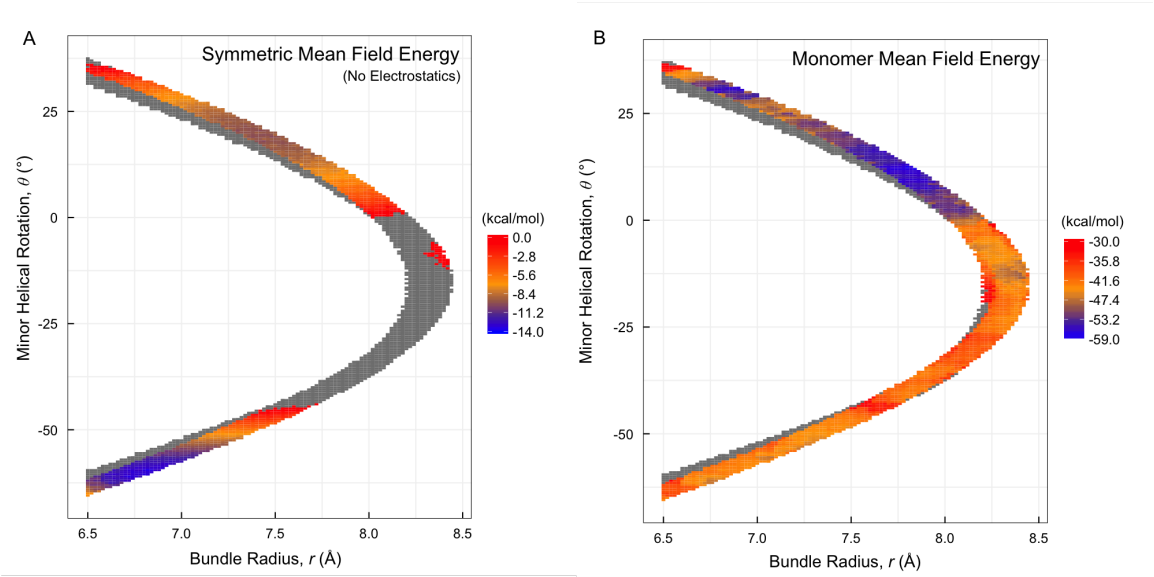



**Figure 3-9.** Alternate energetic landscapes. (A) Mean Field Energy of the poly-L-glycine trimer with GLU-uranyl binding sites, using only the dihedral, van der Waals, and hydrogen bonding potentials (omission of electrostatics). Gray tiles indicate energies above 0.0 kcal/mol. The global minimum is at $r = 6.82$ Åand $\theta = -59.5\,^{\circ}$. (B) Mean Field Energy of the poly-L-glycine monomer with GLU-uranyl binding sites, using dihedral, van der Waals, electrostatic, and hydrogen bonding potentials. Gray tiles indicate energies above -30.0 kcal/mol. The global minimum is at $r = 7.8$ Åand $\theta = 12.5\,^{\circ}$. For (A) and (B), white space denotes coiled-coils that cannot find an orientation of the GLU-uranyl super rotamer that satisfies a uranyl RMSD($\hat{z}$) $\leq 0.1$Å.

(a single monomer) is presented in Figure 3-9B. While this landscape is able to recover the well near $r = 6.72$ Å and $\theta = 32.0\,^\circ$, it highlights a global minimum at $r = 7.8$ Åand $\theta = 12.5\,^\circ$. This is obviously an incomplete picture of the system to be designed (omitting interhelix energetics in the desired trimer), but does provide an interesting minimum that corresponds to the lowest dihedral states in Figure 3-8A. The most probable GLU-uranyl state is $[-166.0\,^\circ, 92.0\,^\circ, -21.2^\circ]$, which passes the MolProbity Rotamer Analysis with a percentile score of 4.2%. It is worth noting this larger radius exists at an extreme of observed trimer radii[189]. In all, these additional landscapes provide a more detailed picture of what interactions and orientations (atomic overlaps, high torsional potentials, favorable symmetric electrostatics) bias energetic minima of the more complete energetic landscapes.

### 3.4.5. Patterned Design of Hydrophobic Core

While the candidate at $r = 6.72$ Åand $\theta = 32.0\,^\circ$ could simply have been chosen as the coiled-coil structure used for full sequence calculations, we instead revisit the structural landscape in the context of a hydrophobic core. We make the assumption that optimization of an interior sequence will be independent of choice of the remainder of the sequence (exterior positions interacting with copies in the crystal). Again, to achieve the symmetry, the monomer unit was placed in the P3 space group such that the coiled-coil axis coincided with the axis of three-fold symmetry, while making the unit cell dimensions large enough that no other unit cells were within interaction range of the trimer unit ($a = b = c = 100$ Å). The interior residues were allowed to be a subset of hydrophobic amino acids (A, V, L, I, F) while keeping the four GLU-uranyl binding sites and the remainder of the sequence as glycine. By targeting the 'upper arm' region (positive $\theta$) of the coiled-coil structures that allow uranyl binding, namely to avoid ill-favored glutamate rotamer states, the position of the GLU-uranyl lay at the *a* position rotation ($\theta = 25.7\,^\circ \pm 25.7\,^\circ$). This in turn identified the remaining interior positions as the *d* repeat; where the GLU-uranyl was placed at residues 3, 10, 17, and 24, the hydrophobic ensemble was placed 4 positions away at residues 6, 13, and 20. Residue position 27 was omitted and reserved for subsequent calculations of the

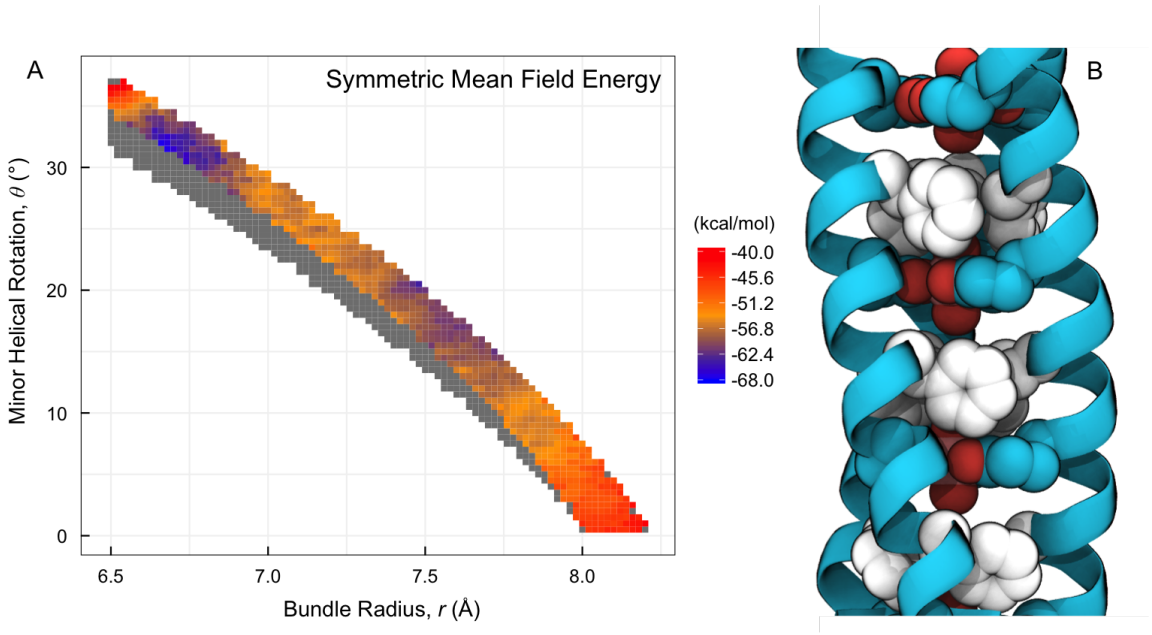



**Figure 3-10.** Design landscape incorporating hydrophobics at the $d$ position. (A) Mean Field Energy of the poly-L-glycine coiled-coil trimer with GLU-uranyl binding sites and hydrophobic interior at $\beta$=0.5. White space denotes coiled-coils that cannot find an orientation of the GLU-uranyl super rotamer that satisfies a uranyl RMSD($\hat{z}$) $\leq$ 0.1Å. Gray tiles indicate energies above -40.0 kcal/mol. The global minimum is at $r$ = 6.66 Åand $\theta = 32.0\,^{\circ}$. (B) Rendering of the most probable sequence for the global minimum, depicting the interstitially placed coiled phenylalanine (white) motif.

inter-trimer interface. As before, reside position 1 was typed with an acetyl N-terminal capping group and position 28 with an amidated C-terminal capping group. An energetic landscape was generated as before with the probabilistic design methodology by minimizing the sequence free energy at $\beta$ = 0.5 mol/kcal. No additional constraints were used, and the Dunbrack 2010 rotamer library[47] (all potential rotamers) were used for the hydrophobic amino acid rotamer ensemble. An unfolded reference was applied to the sequence free energy, wherein amino acid specific unfolded reference energies were estimated for an ensemble of dipeptide states at room temperature $\beta$ = 1.69 mol/kcal.

The corresponding energetic landscape for positive $\theta$ is depicted Figure 3-10A. The global minimum lies at $r$ = 6.66 Åand $\theta = 32.0\,^{\circ}$, quite close to the previously identified poly-glycine parameters. Furthermore, the GLU-uranyl super rotamer state is almost identical,

making a slight adjustment to the last dihedral at the smaller radius: $[-136.0\,^\circ; 66.0\,^\circ; -8.76^\circ]$. The most probable amino acids at the interior positions were identified as predominantly phenylalanine, with an equal probabilities distributed among its three most probable rotamers at $[-73.3\,^\circ; -17.30\,^\circ]$, $[-73.3\,^\circ; 132.60\,^\circ]$, and $[-73.3\,^\circ; 11.90\,^\circ]$. Each of these states places the phenylalanine in position where the C$\beta$-C$\gamma$ bond is perpendicular to the bundle axis, allowing a coiled F motif at the core. A rendering of the $[-73.3\,^\circ; 132.60\,^\circ]$ state creating a hydrophobic core around the GLU-uranyls is shown is Figure 3-10B.

A note regarding the interior probability profile: at $\beta$=0.5, this structure is predominantly F (probability $> 0.8$) with small (but non-zero) probabilities for A, I, L and V. As the radius of the structure increases, the probability distribution diffuses across the other hydrophobics, such that at the local minimum at $r = 7.48$ Åand $\theta = 20.5\,^\circ$, L, I, and F are about equally likely. Subsequent designs explicitly target the global minimum at $r = 6.66$ Åand $\theta = 32.0\,^\circ$.

### 3.5. Sequence Design in the Context of a Targeted Space Group

As was done with previous designs targeting a controlled crystallization in a specific space group[90], full sequence design was performed in the context of a lattice though a systematic grid search of unit cell parameters. Again, the P6 space group ($a = b \neq c$; $\alpha = \beta = 90\,^\circ$; $\gamma = 120\,^\circ$) was chosen as it possesses high symmetry, solvent channels spanning the lattice, a parallel orientation of the proteins (polarity), and rarity (0.1%) among known protein crystal structures.

In order to decouple the degrees of freedom associated with placement in the P6 unit cell, a separate search was carried out over the crystalline "layer" separation, $c$ (z-component of the unit cell). This design step focuses on optimizing the z-component interface to create an extended coiled-coil structure throughout the lattice; we focus on the interaction between a pair of trimers across this layer interface. For simplicity, this was performed in the context of the P3 space group since both $c$ dimensions operate in the same way between the P3 and P6 symmetry operations; that is, a $c$ value chose in P3 will provide the same separation

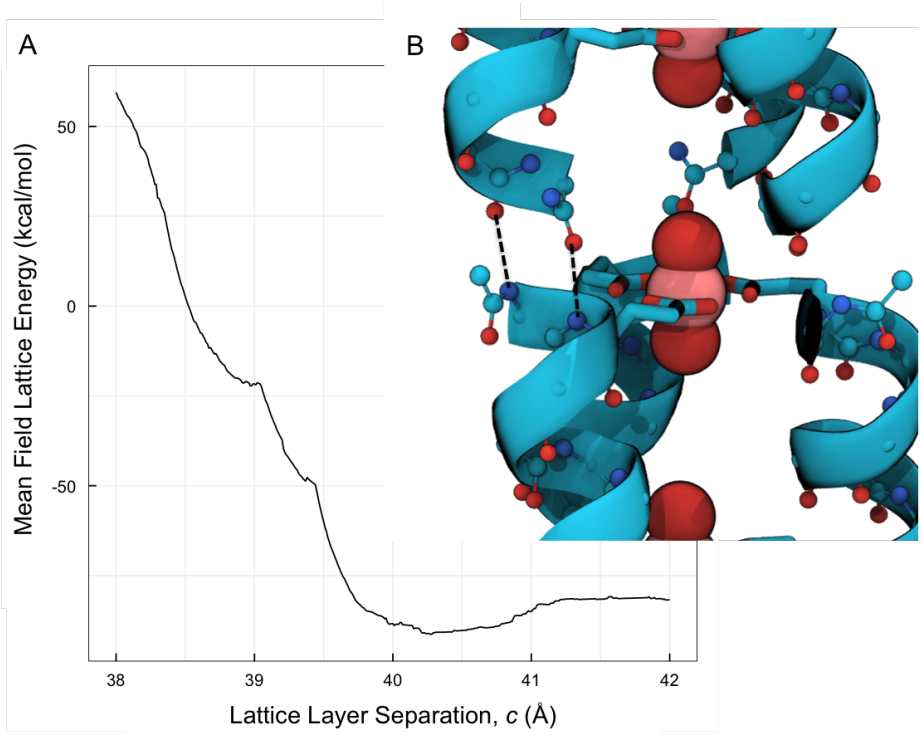



**Figure 3-11.** (A) Mean Field Lattice Energy (kcal/mol) in the context of infinitely large P3 dimensions ($a = b = 100$ Å) as a function of the crystal layer separation dimension ($c$). (B) Rendering of the energy minimum at $c = 40.28$ Å. Hydrogen bonding between helices is highlight with dashed lines. Phenylalanines (at the $d$ position) are omitted for clarity.

in P6. The remaining unit cell dimensions were kept large to avoid trimer interactions in the x-y directions of the unit cell ($a = b = 100$ Å). The GLU-uranyl states were kept at the $a$' positions in the monomer, only phenylalanine allowed at the $d$ positions, and glycine at all remaining positions. The search over $c$ (38 Å$\leq c \leq$ 42 Å, incremented at $\Delta c = 0.01$ Å), is shown in Figure 3-11A. The lowest energy separation lies at 40.28 Å and possess the characteristic interhelical backbone hydrogen bonding for alpha helices which creates an extendable coiled coil structure in this dimension of the unit cell (Figure 3-11B).

Subsequent calculations placed the monomer unit in the P6 space group by translocating each trimer's $C_3$ axis to a P6 $C_3$ axis. This is achieved by translating the single alpha helix

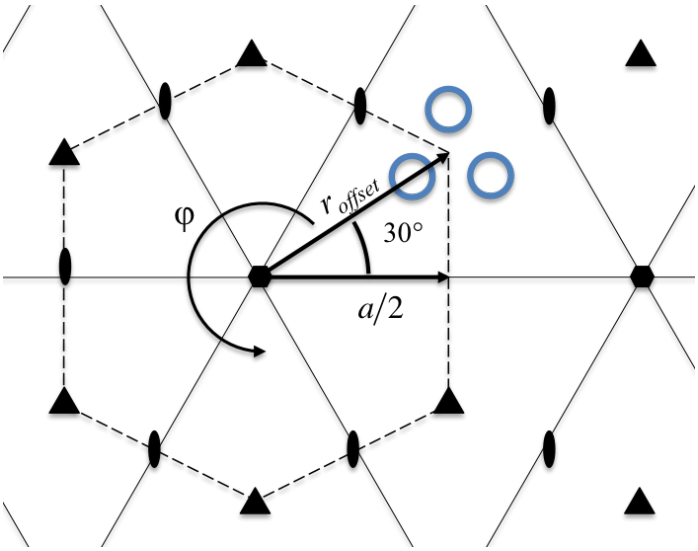



**Figure 3-12.** Illustration of translocation of trimer axis from P6 6-fold axis to a 3-fold axis, given as $r_{\text{offset}}$. The unit cell length, $a$, is highlighted as the distance between 6-fold axes. The angle $\phi$ denotes the rotation applied to the asymmetric unit about the 6-fold axis prior to translation along $r_{\text{offset}}$.

(which originally places the coiled-coil axis at the origin) by $r_{\text{offset}}$, given as

**Eq. 3-3**

$$r_{\text{offset}} = \frac{a}{\sqrt{3}}$$

Given the chosen coiled-coil parameters (radius, rotation) that satisfy the designed binding sites and the crystal layer dimension, the remaining degrees of freedom associated with placement in the P6 space group are the planar unit cell dimension, $a = b$, and the rotation of the entire trimer unit about its axis, $\varphi$. To satisfy these parameters, the initial placement of the monomer unit requires a rotation about the z-axis (coiled-coil axis) by $\varphi$, a translation in the $a$ direction by $r_{\text{offset}}$, and then finally a rotation by 30 ° about the z-axis. This is detailed in Figure 3-12, demonstrating the two free lattice parameters.

The replicated monomer unit remains as the single alpha helix with initial placement as specified by the designed coiled-coil parameters ($r$ = 6.66 Å, $\theta$ = 32.0 °). Generation

of the P6 lattice employs the appropriate space group symmetry operations to create each symmetric copy within the unit cell. Creation of the full lattice of symmetric copies requires specification of the number of unit cells to generate in each dimension. Unit cells were generated with respect to the placement of the monomer unit in the positive $a$ direction of the unit cell origin, creating two cells in $+a$ direction, one cell in both the $+b$ and $-b$ directions, and one cell in both $+c$ and $-c$ directions. The lattice contained a total of 27 unit cells, each with 6 monomers, for a total of 162 monomer units; these copies were then iterated through to discard any copies which did not interact with the asymmetric unit due to cut-offs in the intermolecular potential energy function. This routine guarantees that for any rotation of $\varphi$, only units with the energetic interaction cutoff distance of the asymmetric monomer unit were present to quicken the speed of the calculation. Furthermore, the symmetric images exist temporarily in memory – only for the evaluation of lattice energetic coefficients as detailed by **Eq. 2-29**.

### 3.5.1. Identification of Low Energy Sequences

An initial search through the lattice parameters space was performed, where the monomer alpha helix retains the GLU-uranyl binding site and phenylalanine interior with the remainder of the sequence as a glycine. This initial scan serves as a means to highlight low energy, closely-packed regions of the lattice landscape that accommodate glycine-glycine contacts between neighbors in the crystal. That is, identification of sequences that possess the GXXXG motif should leverage the success seen with the P6d sequence. The grid search is computationally inexpensive, here scanning over $a$ (31 Å$\leq a \leq$ 38 Å, incremented at $\Delta a = 0.1$ Å) and $\varphi$ ($90^\circ$ Å$\leq \varphi \leq 210^\circ$ Å, incremented at $\Delta\varphi = 1^\circ$ Å). Note that the variation in $\varphi$ sweeps the full $120^\circ$ associated with the three-fold axis, but is offset by $90^\circ$. This was done based on implementation details of the lattice generation, and ensures that the monomer unit was kept within the interior of the P6 unit cell.

The mean field lattice energy for this glycine trimer exterior is shown in Figure 3-13, which highlights the energetic minima associated with glycine contacts in the crystal. The surface

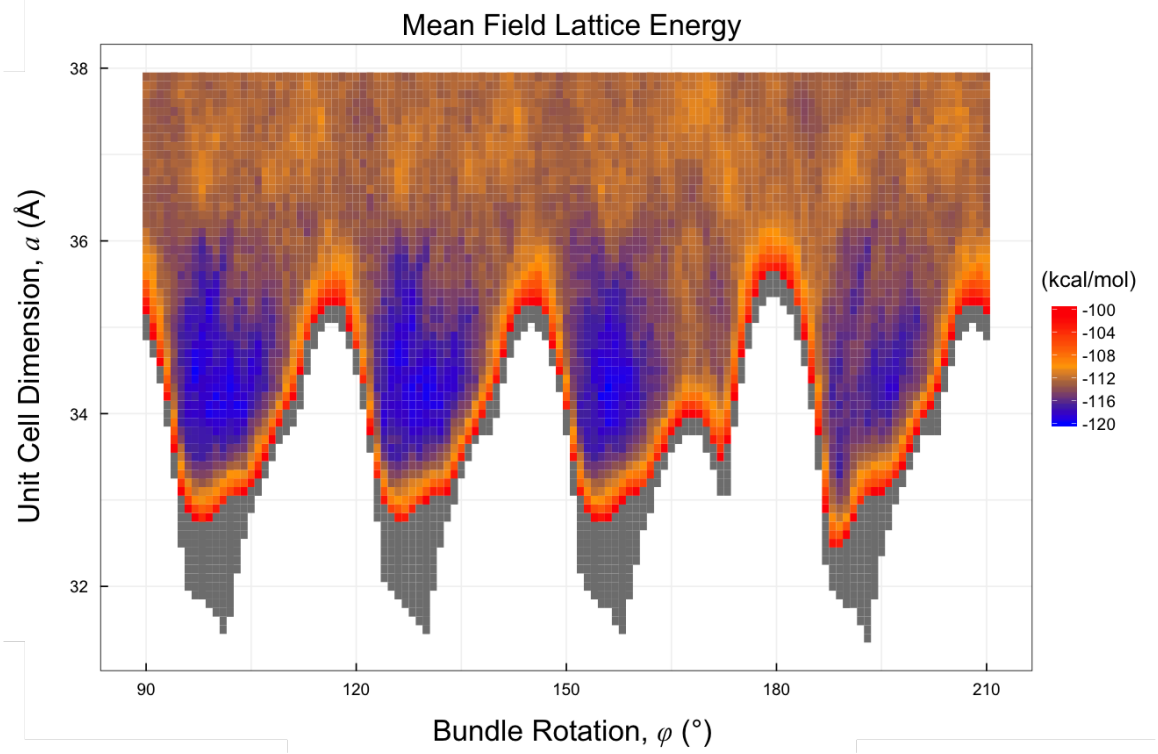



**Figure 3-13.** Glycine crystal contact surface. Mean Field Lattice Energy of the poly-L-glycine coiled-coil with GLU-uranyl binding sites and phenylalanine interior in the P6 space group, at $\beta$=0.5. White space denotes crystal configurations with backbone overlap. Gray tiles indicate energies above -100.0 kcal/mol.

possess four main wells, each corresponding to different rotated contact surfaces where the glycine trimer surfaces are within optimal contact. This surface was then used as a guide to narrow the lattice search space for full designs, limiting the $a$ parameter to the low energy region of interest between 33 Åand 37 Å.

Identification of the remainder of the sequence was carried out through full designs covering the aforementioned lattice sub-landscape. Designs were carried out utilizing the Dunbrack 2010 library[47] with that library's 10 most probable rotamers for each amino acid type. For each design, the four GLU-uranyl binding sites and phenylalanine core were fixed as identified in previous sections. While the conformation of the GLU-uranyl rotamer remained fixed, the phenylalanines are the core were allowed all conformations in the Dunbrack library. The remainder of the positions were typed with the remaining available 18 amino acids (all except proline and cysteine). Site-specific type probabilities were determined computationally, holding the effective inverse temperature, $\beta$, at 0.5 mol/kcal. No additional constraints were applied to the calculations. Lattice parameters were swept across $a$ (33 Å$\leq a \leq$ 37 Å, incremented at $\Delta a = 0.1$ Å) and $\varphi$ ($90^\circ$ Å$\leq \varphi \leq 210^\circ$ Å, incremented at $\Delta\varphi = 1^\circ$ Å).

Energetic wells within the mean field energy landscape at $\beta = 0.5$ guided the selection of sequences for expression. Previous design protocol identified a sequence by iteratively calculating profiles at this inverse effective temperature value, fixing some number of sites to their most probable amino acid at each round. Here a different approach is taken; instead, a final sequence is selected by repeating the calculation at a higher $\beta$ (lower effective temperature) at room temperature ($\beta$=1.69, $T$=298.15 K). Where the lower $\beta$ landscape guides the identification of structural candidates, the higher $\beta$ landscape yields less diffuse site probability profiles by emphasizing crystalline contact interactions. In general this tends to emphasize charged pair interactions, which provides low lattice energies. We note that the repeated calculation at the higher $\beta$ parameter generally preserves the highly resolved type probabilities at the $\beta$=0.5 calculation, but does not guarantee such.

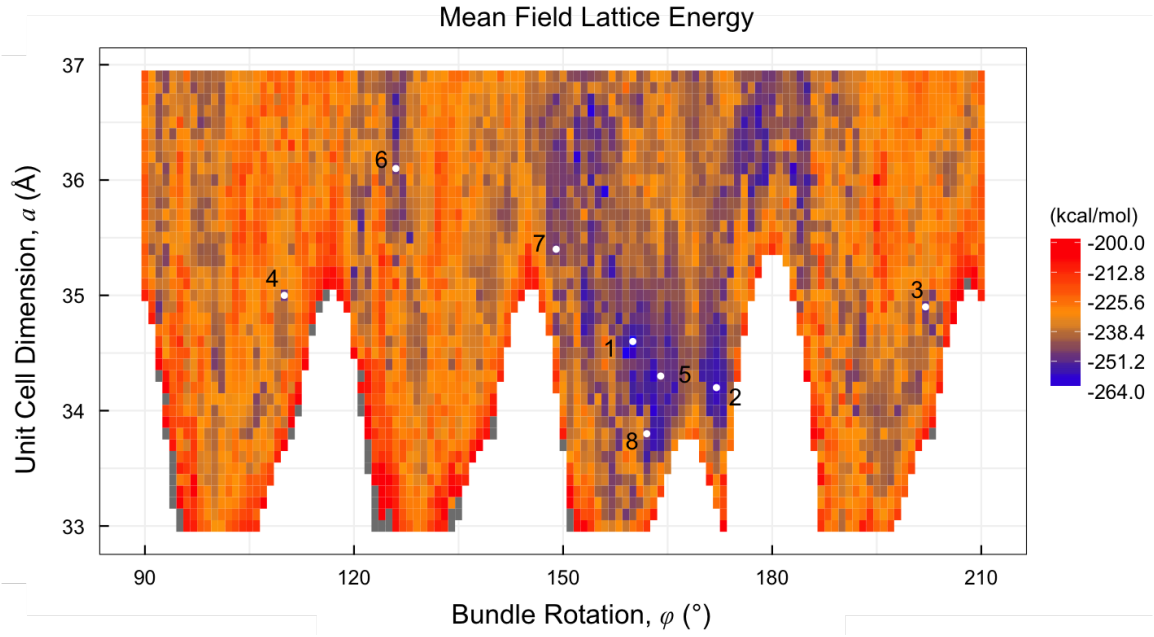



**Figure 3-14.** Mean Field Lattice Energy landscape for the complete design of the peptide in the P6 space group. Each sequence fixes all four GLU-uranyl binding sites and the phenylalanine core, while allowing all other residues (save C and P) at all other positions. Solutions are obtained at $\beta$=0.5. White space denotes crystal configurations with backbone overlap. Gray tiles indicate lattice energies above −200.0 kcal/mol. Markers (white) are placed to correspond to choices made for each of the sequences in (A), which include the global minimum (4) among others.

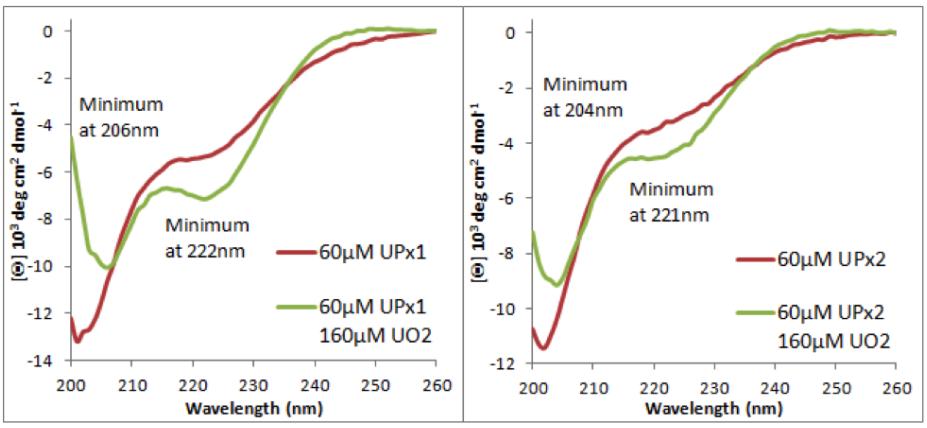



**Figure 3-15.** Circular Dichroism (CD) measurements for the UPx-1 (left) and UPx-2 (right) peptides. The apo peptide spectra (red) are characteristic of random-coil secondary structure. The peptides in the presence of two equivalents of uranyl (peptide:uranyl ratio of 3:8) have a spectra (green) consistent with a mixture of alpha-helical and random-coil secondary structure. Measurements were performed in Starna 0.1 cm path length quartz cuvettes using an AVIV Circular Dichroism Spectrometer Model 410. Isothermal wavelength scans were collected at 20 °C. Bandwidth and wavelength step were both set to 1 nm.

Two candidates, UPx-1 and UPx-2, were chosen for synthesis and purified. Both peptides were solubilized in a buffer (20mM MOPS 150mM NaCl pH7) and exhibited immediate visible aggregation upon the addition of uranyl. When solubilized in a 20mM ammonium acetate pH5.5 buffer, 60$\mu$M ( 0.2mg/mL) peptide with 2 equivalents of uranyl (160$\mu$M for a peptide:uranyl ratio of 3:8), neither peptide showed visible aggregation. CD measurements taken of these samples are shown in Figure 3-15. Both peptides, when in the presence of 2 equivalents of uranyl, possess a CD spectra consistent with a mixture of alpha-helical and random-coil secondary structure. Additionally, CD of peptide concentrations up to 5mg/mL in the presence of one equivalent of uranyl showed a significant random-coil population. These samples were thus non-suitable for crystallization. As such, new peptide designs were considered.

## 3.6. Constrained Redesign of Lattice Interactions

Redesigned sequences were identified to address the lack of alpha-helical character for the UPx-1 and UPx-2 sequences in the presence of multiple equivalents of uranyl. The sus-

pected non-specific aggregation of those peptides was the main driver for redesign of the system, which focuses on excluding potential binding sites for uranyl on the exterior of the trimer, and strengthening the helical stability of the monomer peptide. Here we detail the imposition of specific constraints on the sequence ensemble to target such features.

### 3.6.1. Constraint Choices

The previous unconstrained uranyl peptide designs, as well as the the P6d designs[90], picked out candidate sequences with desirable properties from a broad range of low energy structures. Instead, we take the approach of applying a variety of constraints that focus the design and search for viable sequences as well as address biological concerns. The following describes the number of ways in which we adjust the free energy optimization routines to estimate sequence probabilities subject to constraints which are experimentally motivated.

#### Sequence Composition

The simplest constraint to impose is in limiting the amino acid ensemble at designable sites. We exclude certain amino acids from the design so as to prevent any possible uranyl binding sites from existing outside the folded core. Previous uranyl binding peptide designs possessed multiple carboxylate bearing residues (ASP and GLU) at exterior positions. The possibility of these residues acting as uranyl binding sites is corroborated by experimental details, which suggest that the peptides with large numbers of GLU and ASP residues are not folding in the presence of uranyl. Furthermore, these peptides exhibit nonspecific aggregation. In an attempt to mitigate such concerns, we remove all GLU and ASP residues from the design, save the GLU placed at the four core binding positions. Furthermore, GLY was removed as an allowed amino acid from the designable sites in an effort to foster helical structure formation.

### Total Charge

By removing all ASP and GLU from the exterior positions, the only ionizable residues that remain are positively charged at neutral pH, which raises concerns about the overall charge of the protein. As we wish to design a sequence that is near neutral to promote crystallization, the imposition of a net charge constraint is applied. The constraint is linear in the conformer probabilities. For each conformer indexed $i$ (that is, the abbreviated indexing used in Chapter 2), its total charge $q_i$ is simply the sum of the force field partial charges of atoms in that residue. The expression for the mean net charge constraint is given as a sum over residue positions, types, and conformers

**Eq. 3-4**

$$\langle \text{Net Charge} \rangle = \sum_i q_i \cdot w_i = 0$$

which is applied to each design optimization at the value 0 (to impose neutrality). It should be noted that the sum of the partial charges on each GLU-uranyl super rotamer state is $-\frac{1}{3}$. The overall charge of each binding site consisting of 3 GLU (-3) and a uranyl (+2) is -1. There are 4 binding sites across the trimer for a total charge of -4, a charge per helix of $-\frac{4}{3}$, and an overall charge per GLU bound to uranyl of $-\frac{1}{3}$.

### Extinction Coefficient

To enforce a means of monitoring protein concentration throughout experimental trials, we include the addition of a mean extinction coefficient constraint. To estimate the molar extinction coefficient of a protein, we extend the following estimation over all amino acid

types, $t$ for the number of each type in the sequence, $N_t$

**Eq. 3-5**

$$\epsilon_{\text{Protein}} = \sum_t \epsilon_t \cdot N_t$$

where for a protein in water measured at 280 nm, the coefficients (in $\text{M}^{-1}.\text{cm}^{-1}$) are $\epsilon_{\text{TRP}}$=5500, $\epsilon_{\text{TYR}}$=1490, $\epsilon_{\text{CYS}}$=125, and $\epsilon$=0 for all other residues. Taking the average of the above equation over all states in the design ensemble provides the following mean extinction coefficient constraint

**Eq. 3-6**

$$\langle \epsilon_{\text{Protein}} \rangle = \sum_i \epsilon_{\text{Ext},i} \cdot w_i \geq 5690.0 \text{ M}^{-1}\text{cm}^{-1}$$

which is imposed to be greater than or equal to the molar extinction coefficient of at least one tryptophan. This guarantees that protein and peptide concentrations are able to be monitored easily.

### Helix Propensity

In order to further promote the helicity of the peptide sequences a helix propensity constraint was applied. As per the design of the P6d peptide[185], the constraint target was estimated from 304 non-redundant, trimetric, parallel coiled-coil structures from the CC+ database[192]. A linear regression from the calculated helix propensities of each coiled-coil region yielded the equation $E_h(n) = 0.317n$+0.1407 ($\text{R}^2 = 0.81$), and in turn the value $E_h$ = -8.09 kcal/mol for n=26. This was applied as an upper boundary for the constraint (more negative energies infer more helicity). The equation for the mean helix propensity across the design ensemble

takes a similar to the linear constraints above.

**Eq. 3-7** $$\langle E_{\text{Helix Propensity}} \rangle = \sum_i \epsilon_{h,i} w_i \leq \text{-8.09 kcal/mol}$$

**Environmental Energy**

Initial trials including only these three constraints in the uranyl-binding peptide system tended to cluster TRP in the solvent channel due to the absence of a solvent component in the force field. This can be circumvented by excluding hydrophobic residues from the lattice calculation altogether with the rationalization of avoiding "sticky" patches on the protein that may cause experimental difficulties. We attempt to harness the power of the environmental energy to engineer a hydrophobic handle to engineer specific aggregation points. This is achieved by modifying the local $C_\beta$ density calculation to include the $C_\beta$ of lattice neighbors as well. The environmental energy, as described in Chapter 2 and utilizing the updated parameterization, was applied as a constraint in the familiar linear form

**Eq. 3-8** $$\langle E_{\text{Environmental}} \rangle = \sum_i \epsilon_{\text{Env},i} w_i \leq \text{-3.0}$$

The contraint is applied as an inequality against the boundary of -3.0. This value arises from evaluating the upper boundary of the 95% confidence level for the linear fit for the proteins utilized in parameterizing the environmental energy. We estimate the environmental energy as one third of the upper boundary for a 78 residue protein (the number of residues in the trimer structure) which is -3.0. Without the constraint, sequences had environmental energies over +5.0 due to the placement of ionizable residues at the lattice interface.

### 3.6.2. Constrained Designs

#### Alanine Exterior Landscape

The lattice parameters were scanned to identify key regions of the landscape. Again, glycine was removed as an allowed amino acid in the calculation to foster helical structure formation. As before, the monomer alpha helix retained the GLU-uranyl binding site and phenylalanine interior with the remainder of the sequence now alanine. Ideally, such a scan will identify favorable AXXXA motifs (much like the GXXXG motif in the P6d design work). Calculations were performed over the parameter space of $a$ (31 Å$\leq a \leq$ 38 Å, incremented at $\Delta a = 0.1$ Å) and $\varphi$ ($90^\circ$ Å$\leq \varphi \leq 210^\circ$ Å, incremented at $\Delta\varphi = 1^\circ$ Å). Figure 3-16 highlights low energy regions of the lattice landscape as more restrictive than the poly-glycine search, here between values of 34 Å and 37 Å for the $a$ parameter.

#### Constrained Sequence Design Landscape

Full designs were then carried out over this subregion as per the constraints described above. In summary, the lattice energy landscape was recalculated at $\beta$=0.5 under the following conditions:

- Fixing the GLU-uranyl super rotamer at the 'a' position binding sites
- Typing the remaining core position 'd' to PHE
- Allowing the remaining sites to be all residues excluding PRO, CYS, GLU, ASP, and GLY
- Constraining the mean net charge to 0
- Constraining the mean extinction coefficient to be $\geq$ 5690.0 $M^{-1}cm^{-1}$
- Constraining the mean helix propensity to be $\leq$ -8.09 kcal/mol
- Constraining the mean environmental energy to be $\leq$ -3.0 kcal/mol

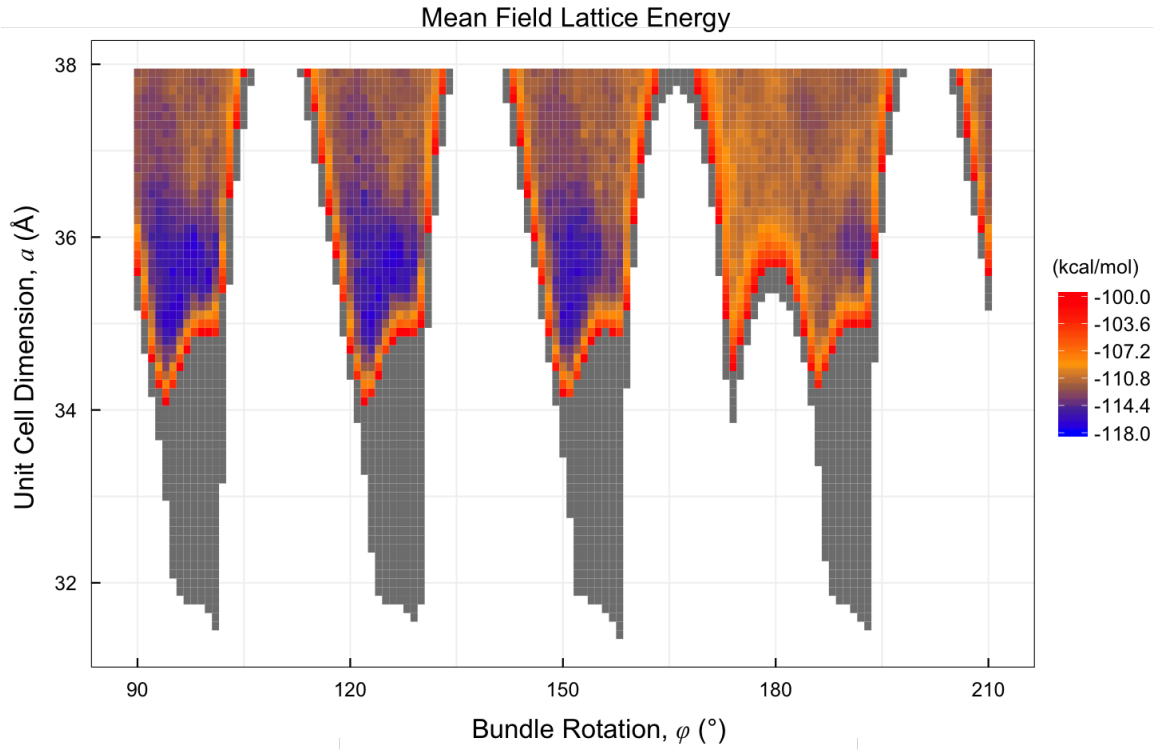



**Figure 3-16.** Alanine crystal contact surface. Mean Field Lattice Energy of the poly-L-alanine coiled-coil with GLU-uranyl binding sites and phenylalanine interior in the P6 space group, at $\beta$=0.5. White space denotes crystal configurations with backbone overlap. Gray tiles indicate energies above -100.0 kcal/mol.

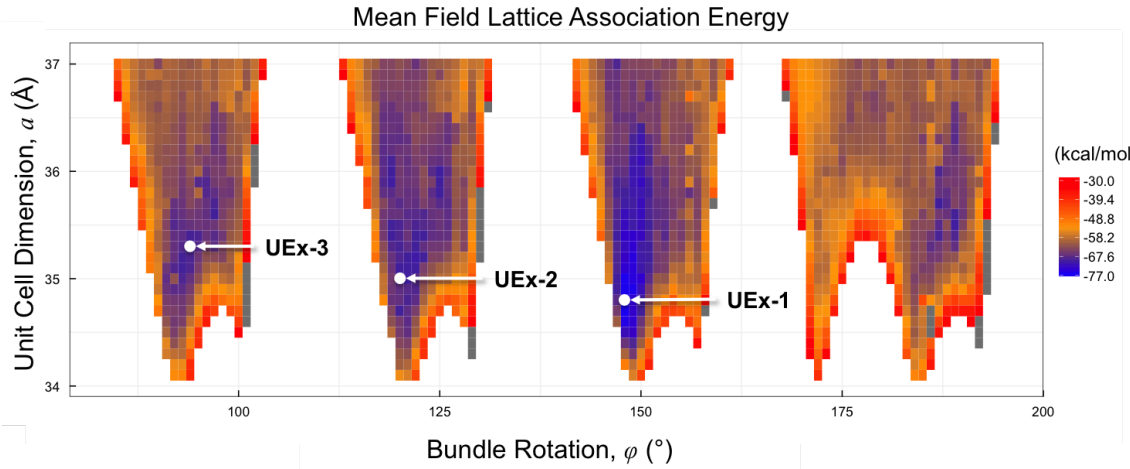



**Figure 3-17.** The Mean Field Lattice Association Energy landscape for the constrained re-design of the peptide in the P6 space group. Energies are expressed as the Mean Field Lattice Energy with the Mean Field Energy of the monomer subtracted off. Each sequence fixes all four GLU-uranyl binding sites and the phenylalanine core, while allowing only (RHKSTNQAVILMFYW) at all other positions. Constraints are applied to helix propensity, net charge, extinction coefficient, and lattice environmental energy. Solutions are obtains at $\beta$=0.5. Gray tiles indicate lattice energies above -30.0 kcal/mol; white tiles indicate structures with atomic overlap between symmetric backbones. Markers (white) are placed to correspond to choices made for each of the sequences, which comprise the three main energy minima.

The landscape rendered in Figure 3-17 is the resulting mean field lattice association energy for these calculations. The lattice association energy is given as the difference of the mean field lattice energy and the monomer mean field energy at the solved probabilities:

**Eq. 3-9**
$$\langle E_{assoc} \rangle^* = \langle E_{lattice} \rangle^* - \langle E \rangle^*$$

where the probabilities are taken from the solution of minimizing the mean field lattice energy, $\langle E_{lattice} \rangle$. The quantity highlights the energetics of interactions both within the trimer unit and the full P6 lattice. The surface possess four main regions, from which three local energy minima are selected. UEx-1 is the both the lattice association energy and lattice energy global minimum.

The three highlighted minima are rendered in Figure 3-18. Collectively, they show the migration of a hydrophobic interface along the trimer surface; where UEx-1 has an ALA-MET-MET contact towards one end, UEx-3 has a MET-ALA-TRP patch at the other, and UEx-2 a large hydrophobic strip spanning the interface. Additional properties of these sequences are given in Table 3-3.

### 3.6.3. Sequence Mutation Considerations

While a majority of the trimer exterior is hydrophilic, there is a clear hydrophobic MET 'handle' at the lattice interface providing an interlocking glove. This should provide an energetically favorable contact for the entire structure to come together in the P6 space group. The sequence/structure is rendered to highlight these features in Figures 3-19 and 3-20.

In addition to selecting UEx-1, we chose to synthesize two variants of this sequence. The first, UEx-1-W09, is the final sequence obtained from running the same UEx-1 calculation at $\beta$=1.69. The calculation estimates an identical sequence with the exception of an additional

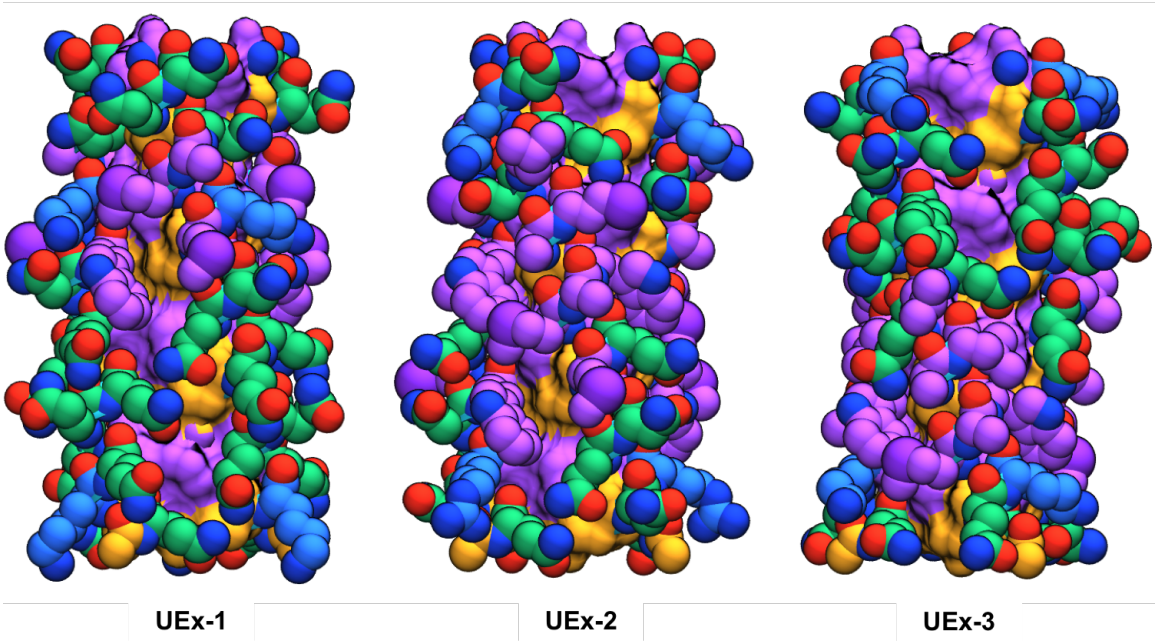



**Figure 3-18.** Renderings of the three sequences which correspond to the three energy minima in Figure 3-17. The renderings highlight the translation of the hydrophobic lattice contact along the trimer exterior for the three different lattice rotation values. Coloring scheme indicates carbon atoms in positively charged residues (light blue), hydrophobic residues (light purple), hydrophilic residues (light green), the specially placed glutamate and cap residues (orange). Heavier atoms are colored as red (oxygen), blue (nitrogen), purple (sulfur), and pink (uranium).

| **Sequence Property** | **UEx-1** | **UEx-2** | **UEx-3** |
|---|---|---|---|
| Number of Residues | 28 | 28 | 28 |
| Molecular Weight (Da) | 4528.28 | 4517.32 | 4431.13 |
| Percent Charged (%) | 21.43 | 21.43 | 21.43 |
| Percent Hydrophobic (%) | 28.57 | 42.86 | 39.29 |
| Net Charge | +0.67 | +0.67 | +0.67 |
| Ionizable Charge, pH=7.0 | -2.0 | -2.0 | -2.0 |
| Isoelectric Point, pI | 4.1 | 4.1 | 4.1 |
| Helix Propensity (kcal/mol) | -8.80 | -9.06 | -8.20 |
| Total Hydropathy Index | -47.9 | -27.9 | -30.8 |
| Average Hydropathy Index | -1.711 | -0.996 | -1.100 |
| Extinction Coefficient ($M^{-1}.cm^{-1}$) | 11380.0 | 11380.0 | 6970.0 |
| Unit Cell Volume ($Å^3$) | 42245.32 | 42732.29 | 43467.98 |
| Matthews coefficient ($Å^3$/Da) | 1.55 | 1.58 | 1.63 |
| Crystal Solvent Content (%) | 20.89 | 21.98 | 24.77 |

**Table 3-3.** Summary of sequence properties for **UEx-1**, **UEx-2**, and **UEx-3**.

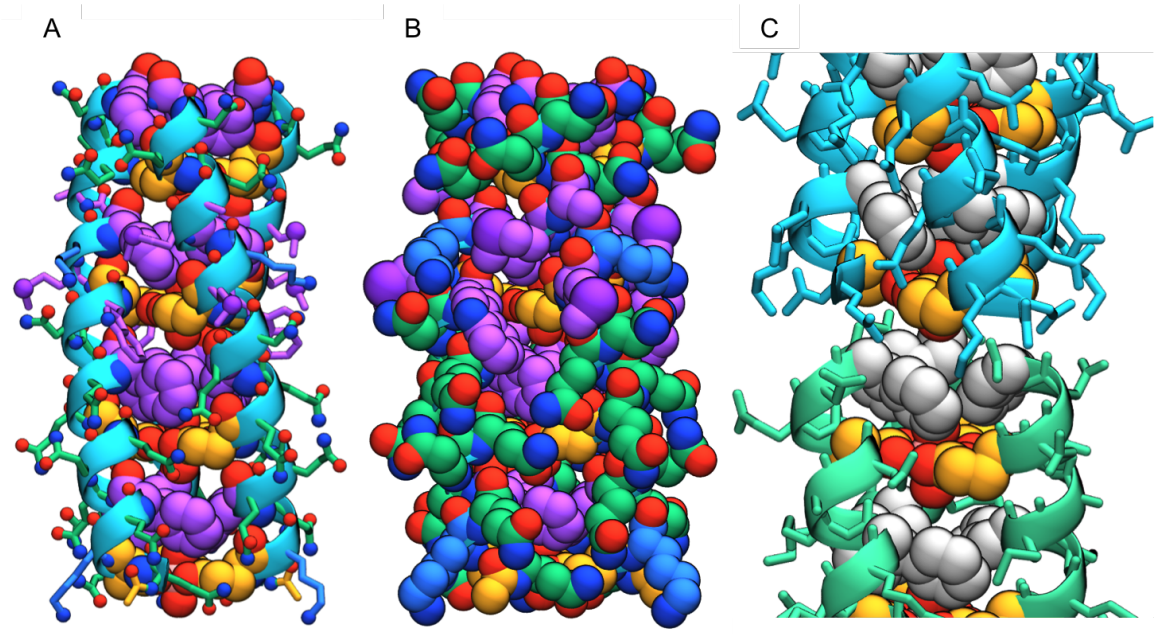



**Figure 3-19.** Multiple renderings of UEx-1. (A) Stick and (B) sphere renderings of the UEx-1 sequence which highlight both the composition of the core (A) and the trimer exterior (B). Coloring scheme indicates carbon atoms in positively charged residues (light blue), hydrophobic residues (light purple), hydrophilic residues (light green), the specially placed glutamate and cap residues (orange). Heavier atoms are colored as red (oxygen), blue (nitrogen), purple (sulfur), and pink (uranium). (C) Rendering of the lattice packing along the $c$ coordinate. Residues are colored as by different segments to highlight the interface, separately from the GLU-uranyl (orange/red) and F (white) core.

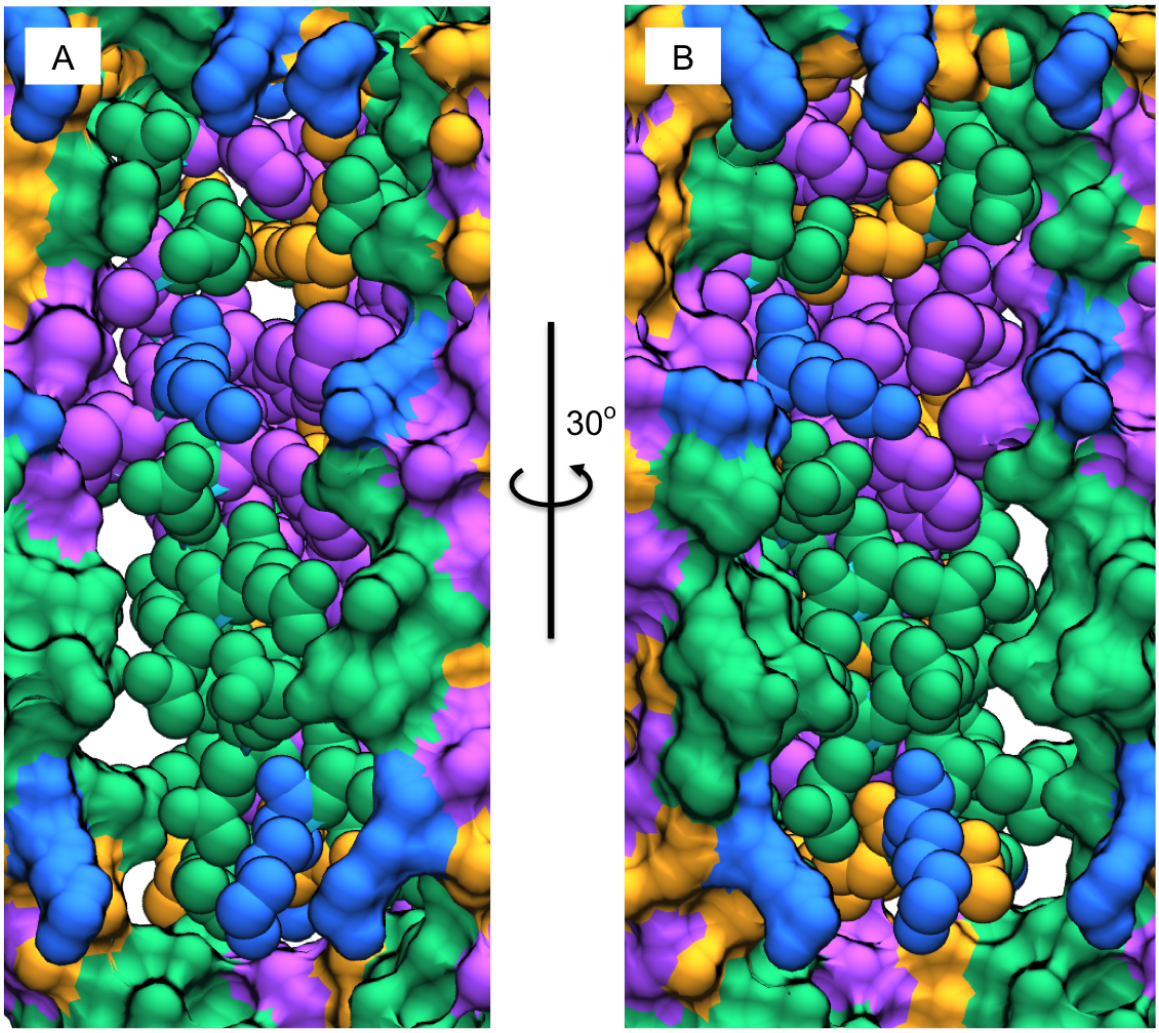



**Figure 3-20.** Rendering of UEx-1 (spheres) packed against lattice neighbors (surface) in the P6 space group. Renderings are slightly rotated from each other to show both sides of the contact surface, meant to highlight the two methionine residues on the exterior which interlock with their symmetry mates. Coloring scheme indicates carbon atoms in positively charged residues (light blue), hydrophobic residues (light purple), hydrophilic residues (light green), the specially placed glutamate and cap residues (orange).

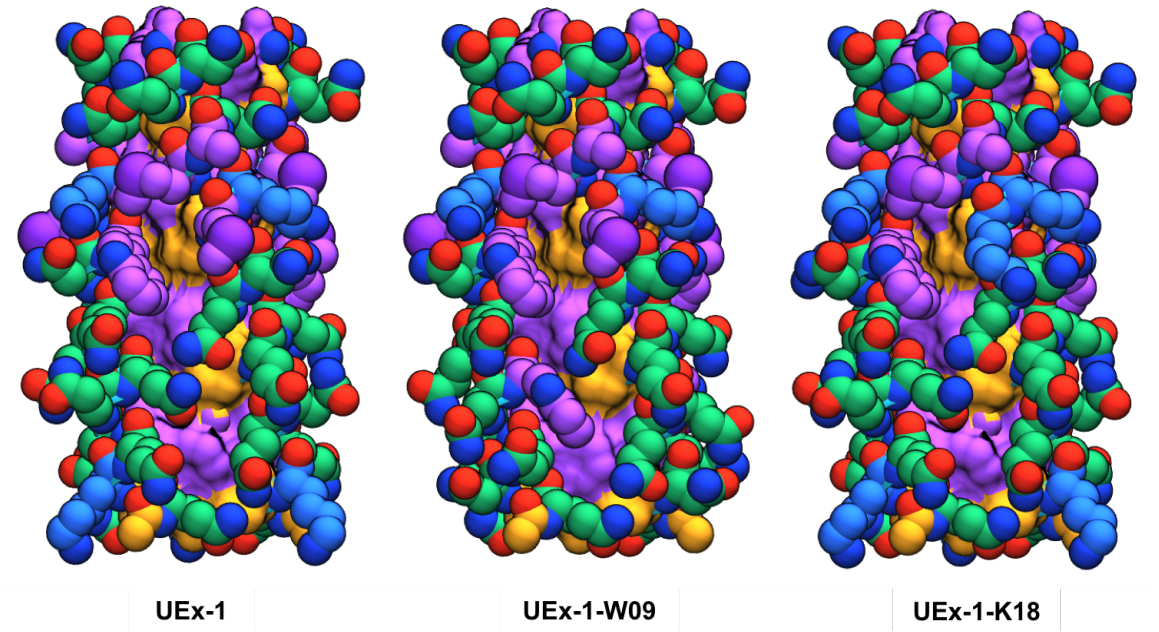



**Figure 3-21.** Renderings of the three sequences selected for synthesis, which are UEx-1 and two mutants. The renderings highlight the extension of the hydrophobic interface (UEx-1-W09) and the disruption of the hydrophobic lattice contact (UEx-1-K18). Coloring scheme indicates carbon atoms in positively charged residues (light blue), hydrophobic residues (light purple), hydrophilic residues (light green), the specially placed glutamate and cap residues (orange). Heavier atoms are colored as red (oxygen), blue (nitrogen), purple (sulfur), and pink (uranium).

tryptophan placed at position 9. This in effect extends the hydrophobic interface which is supported by the lower energy of this higher $\beta$ sequence. The second sequence, UEx-1-K18, is intended to disrupt the clear hydrophobic contact at the interface to make the peptide more soluble. Here, the MET at position 18 is substituted for a LYS by fixing the identified UEx-1 sequence and allowing only ARG or LYS at position 18. It should be noted that while the energy of this sequence/structure is higher than that of UEx-1, the most probable LYS exhibits a conformation consistent with this lattice structure. Rendered interfaces are shown in Figure 3-21.

The UEx-1 and UEx-1-K18 sequences were selected for synthesis and purified. For both peptides solubilized in a non-uranyl-coordinating buffers, aggregated material is visible immediately upon addition of uranyl. This is depicted in Figure 3-23 for the UEx-1-K1

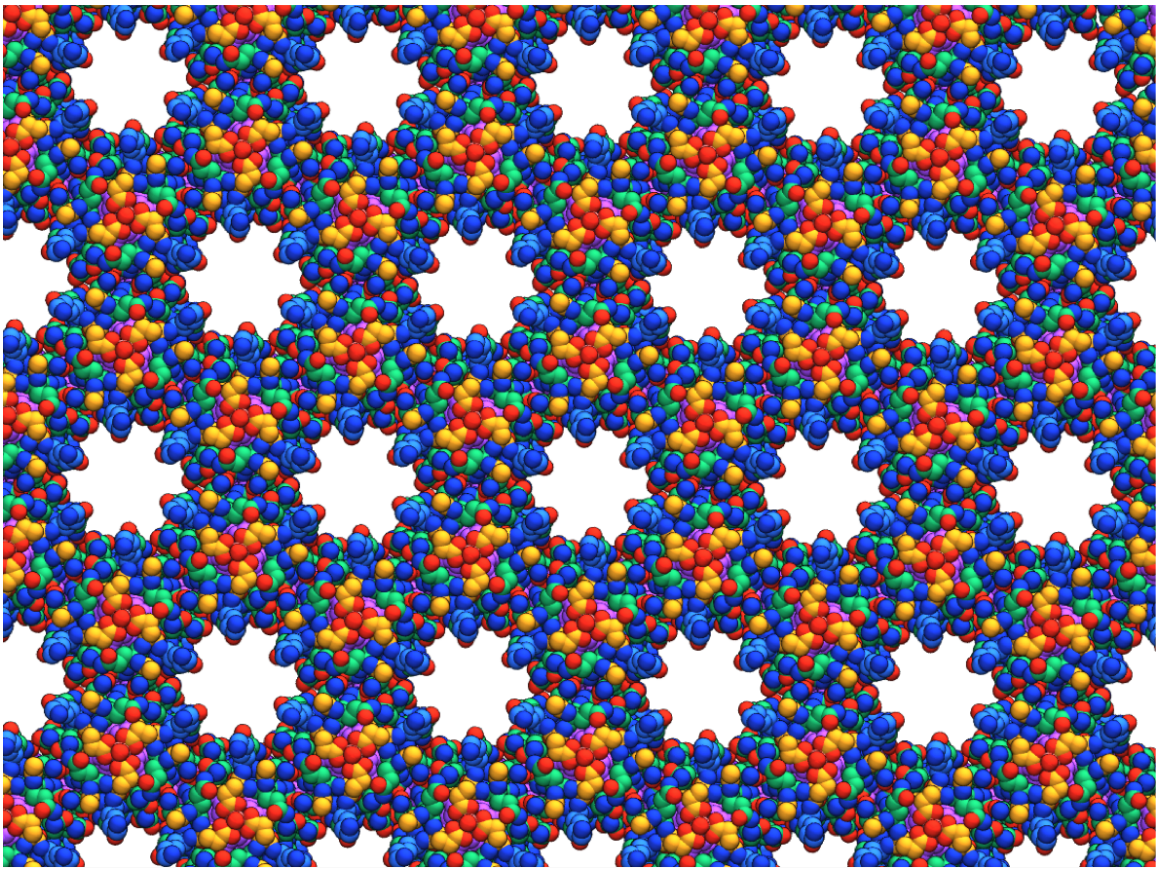


**Figure 3-22.** Full rendering of the UEx-1 lattice. Coloring scheme indicates carbon atoms in positively charged residues (light blue), hydrophobic residues (light purple), hydrophilic residues (light green), the specially placed glutamate and cap residues (orange). Heavier atoms are colored as red (oxygen), blue (nitrogen), purple (sulfur), and pink (uranium)

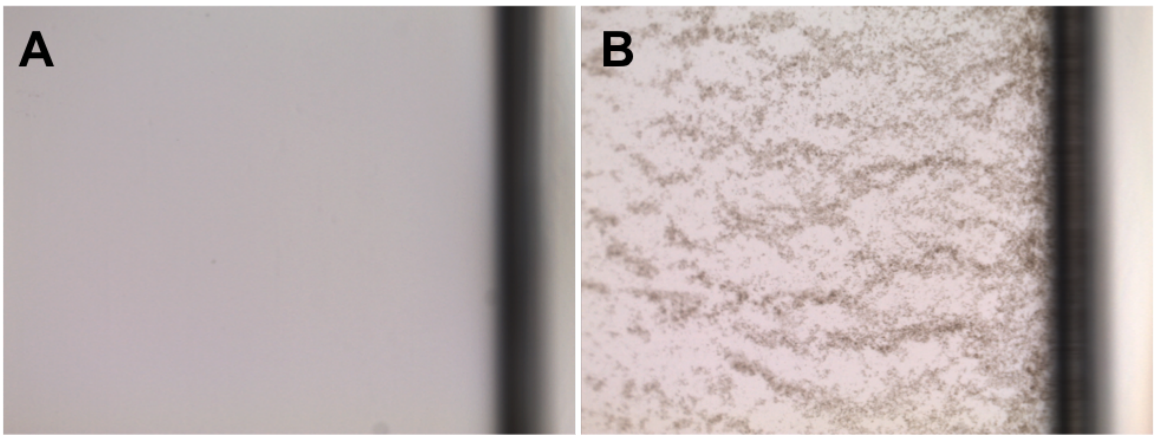



**Figure 3-23.** Images of of 73$\mu$M (0.25mg/mL) UEx-1-K18 in a 20mM MOPS 150mM NaCl pH7.5 buffer. (A) The apo peptide after 12 hours. (B) The peptide in the presence of one equivalent of uranyl (peptide:uranyl ratio of 3:4) after 12 hours.

peptide. The apo peptide remained solubilized and free of aggregation after 12 hours (3-23A), whereas aggregated material is visible in significant quantities for the peptide in the presence of one equivalent of uranyl (a peptide:uranyl ratio of 3:4) after 12 hours (3-23B).

When solubilized, low concentrations of the peptides remained soluble in the presence of one equivalent of uranyl. CD wavelength scans for a 10 $\mu$M sample of UEx1-K18 in 10 mM Glycine pH7 at varying levels of addition of uranyl are shown in Figure 3-24A. There titration shows a dramatic transition from random-coil to alpha-helical secondary structure. Figure 3-24B monitors the transition of the 222 nm CD data and suggests the system saturates at approximately 1 equivalent of uranyl.

## 3.7. Conclusion

The approach outlines the *de novo* design of a peptide system which binds four uranyl cations at a trimer core. The methods developed offer a means of aligning a known uranyl binding geometry to an amino acid motif, and identifying coiled-coil structures capable of accommodating the multiple binding sites. Having established a coiled-coil trimer structure and interior sequence, the remainder of the peptide sequence was designed to be compatible within the P6 space group; the method is an extension of the first computationally designed

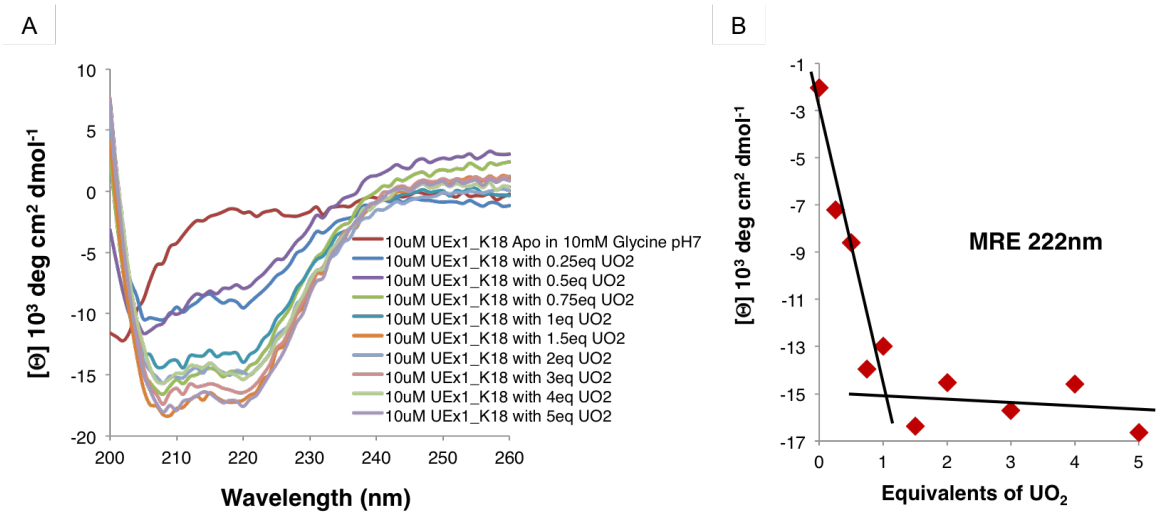



**Figure 3-24.** (A) Circular Dichroism (CD) measurements for UEx1-K18 peptide in the presence of varying concentrations of uranyl. Up to 5 equivalents of uranyl were titrated allowing 15 minutes equilibration time in-between each step. (B) The mean residue ellipticity of the signals presented in (A) at 222nm as a function of uranyl equivalents. All measurements were performed in Starna 0.1 cm path length quartz cuvettes using an AVIV Circular Dichroism Spectrometer Model 410. Isothermal wavelength scans were collected at 20 °C. Bandwidth and wavelength step were both set to 1 nm.

peptide crystal[90]. Sequences from both unconstrained calculations sampling from diverse low energy structures, and constrained calculations targeting the global energetic minima were identified. The latest sequence, UEx1-K18, readily aggregates upon the addition of uranyl and shows a strong transition from random coil in the apo form to a helical peptide in the presence of uranyl. CD measurements indicate minimal change in secondary structure at peptide:uranyl ratios higher than the idealized 3:4 ratio. Further experimental characterization is currently underway.

# 4 Computational Design of a Protein Bundle That Selectively Binds a Non-Biological Cofactor

Electron transfer reactions are pervasive in biochemical processes, crucial for respiration, photosynthesis, and water oxidation[193–195]. Probing biologically relevant, photo-induced charge separation reactions is challenging, requiring intricate knowledge of the surrounding protein structure and energetics. Computational protein design can afford methods to target the binding and orientation of specific nonbiological cofactors with interesting charge-transport pathways. The work here leverages design strategies for identifying proteins which bind such cofactors, and builds a complete *de novo* methodology for the encapsulation of a donor-bridge-acceptor electron transfer chromophore. Utilizing the previously described probabilistic technique, a single chain protein-cofactor construct was targeted with an emphasis on a crystalline assembly.

## 4.1. Introduction

Protein-cofactor assemblies are the building block of many critical biological functions. Oxygen binding in respiration, transduction of light energy in photosynthesis, and enzymatic catalysis[193–195] all arise from specific functions conferred to proteins by natural cofactors. Yet the means by which nature has identified efficiently bound and organized protein-cofactor assemblies is still poorly understood. The use of synthetic proteins and nonbiological cofactors offers one such potential route to exploring the processes underlying natural systems. Furthermore, designed proteins which incorporate nonbiological cofactors can provide novel properties that are not afforded by natural systems. In turn, the ability to mime biological design with precise bio-nano-structures would enable the realization of novel materials.

For natural cofactors, ranging is size from single metal atoms to metalloporphyrins, non-

convalent interactions predominantly govern positioning. *De novo* protein design seeks to leverage these interactions to bind and manipulate the protein-cofactor assembly to modulate and tune desired cofactor properties. For proteins encapsulating cofactors, this control comes from the specification of cofactor conformation and orientation, as well as solubility and local dielectric environment. Moreover, composition and arrangement of the protein-cofactor complex exterior dictates larger ordering and assembly required for placement at interfaces or surfaces. Computational methods have been successful in engineering proteins able to encapsulate synthetic nonbiological cofactors[30,31,65,88,89], where several *de novo* proteins have been designed to bind cofactors composed of extended $\pi$-electron systems, either of which exhibit specific nonlinear optical (NLO) responses or light induced electron transfer (ET) over large distances[6–13].

A subset of these cofactors includes covalently linked donor-acceptor (D-A) chromophores, which undergo ultrafast, photo-induced charge separation and thermal charge recombination ET reactions[196–203]. Such cofactors are of interest to organic photovoltaic (OPV) applications, where a high-density of 2-D vectorially oriented cofactors might act as a nanoscale active layer in bilayer heterojunction devices[204–209]. The electron transfer process in such D-A systems (Figure 4-1 A) can be described by (1) excitation by light to an excitonic state (D*-A), after which some population relaxes via (2) charge separation to form a free electron and hole charge ($^{+}$D-A$^{-}$). This is then followed by (3) charge recombination and relaxation back to the ground state (D-A). In addition to the bulk organization required to be useful in OPV applications, these chromophores present the challenge of stabilizing the charge separated state. Drawing from natural systems, we seek to do so by housing such cofactors in the low dielectric environment of a protein core.

This work considers the encapsulation of N-[5-(10,20-Diphenylporphinato)zinc(II)] N-(octyl)-pyromellitic diimide (PZnPI), which features a D-A system as a (porphinato)zinc(II) chromophore (PZn) covalently bound to a pyromellitimide (PI) acceptor[202,203] (Figure 4-1B). Previously, Koo et al. designed an amphiphilic peptide tetramer which was able to bind,

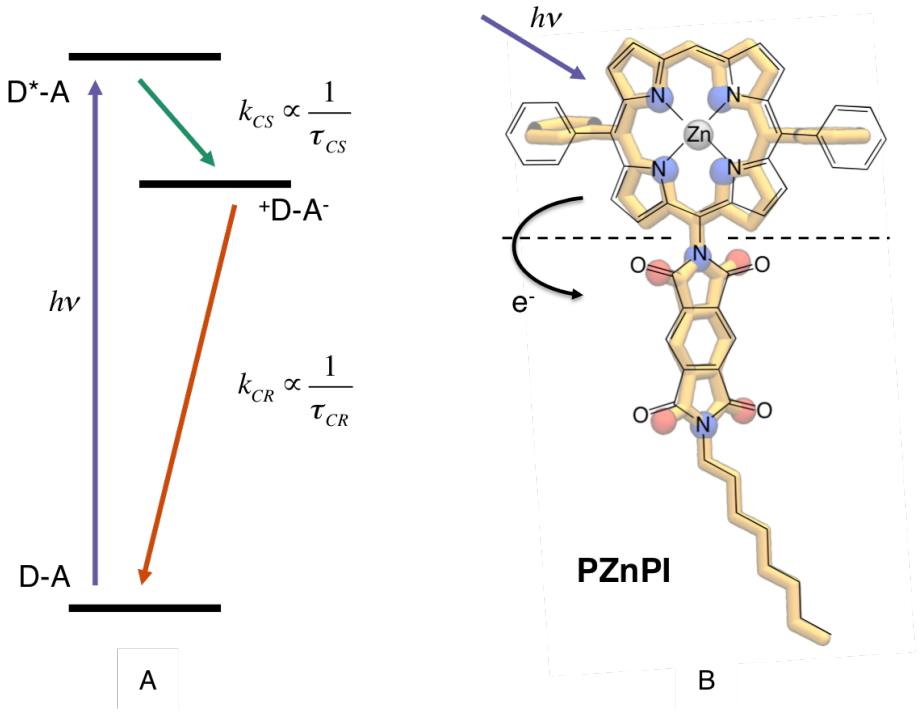



**Figure 4-1.** (A) Light induced electron transfer diagram for a donor-acceptor (D-A) system, indicating charge separation ($\tau_{CS}$) and charge recombination ($\tau_{CR}$) rates. (B) Chemical structure of the PZnPI cofactor (N-[5-(10,20-Diphenylporphinato)zinc(II)] N-(octyl)pyromellitic diimide). Overlay indicates the charge separation which occurs between the porphyrin and diimide moieties.

sequester, and orient the PZnPI cofactor at its core with high specificity[12]. The bundle was designed coarsely by inspection, relying upon amino acid patterning along the well-known heptad repeat to achieve hydrophilic and hydrophobic domains. Upon incorporation into the peptide bundle, the PZnPI cofactor exhibited enhancement of charge separation and charge recombination rates as compared to organic solvents. Specifically, time-resolved transient absorption spectroscopic experiments indicated an accelerated charge separation lifetime ($\tau_{CS}$) from 1.2 ps to 0.3 ps, and a protracted charge recombination lifetime ($\tau_{CR}$) from 4.6 ps to 78 ps, between the interior of the four helix bundle and DMSO solvent[12]. The results suggest a successful enabling of a hydrophobic sheath for the PZnPI cofactor, and further postulate that one or more hydrogen-bonding interactions with the diimide carbonyl oxygens are responsible for the promoted electron transfer dynamics[202,203,210,211]. The exploration of tuning the D-A's local environment poses the ability to gain control over charge migration dynamics, and provide a detailed insight to protein ET processes.

To test the conjecture that specific hydrogen bonding interactions are able to alter the ET dynamics of the PZnPI cofactor within a protein core, we utilize computational protein design techniques with atomic precision to target single chain constructs both with and without specified H-bonding interactions. In lieu of a simple peptide tetramer, we instead opt to leverage Fry et al.'s success designing a single chain protein to form a four helix bundle which binds the nonbiological chromophore in a 1:1 ratio[13]. The single chain construct allows the construction of a single binding site for coordination of the porphyrin zinc while sculpting a protein core specific to the shape and desired orientation of the cofactor. What follows is a presentation of our design technique to identify protein sequences able to fold, bind, and sequester the PZnPI cofactor. Throughout we maintain a focus on atomistic precision of interactions within the cofactor bearing core. Identifying pairs of sequences that differ by a single core position residue should test how a tailored hydrogen bonding interaction affects the ET dynamics of the encapsulated cofactor. Furthermore, given recent success with computationally designing a protein crystal[90], we apply similar constraints to the exterior sequence of each design to promote crystallization. A protein crystal of

the protein-cofactor assembly would offer resolution of the PZnPI microenvironment and potentially confirm the ability of computational protein design to target residue placement with atomistic control.

## 4.2. Overview of *De Novo* Design Strategy

The project aims to identify protein sequences compatible with a donor-acceptor chromophore, providing a hydrophobic core offering shape-complementarity. Moreover, the design should incorporate a tailored hydrogen-bonding interaction within the core; the placement of a proton donor within an appropriate configuration to the diimide's carbonyl oxygens should provide a hydrogen bond which will further stabilize an induced charge separated state. We expect the local dielectric anisotropy about the cofactor to dictate the nature of photoinduced charge transfer dynamics. As such, the simple mutation of the tailored hydrogen-bonding residue to an analogue lacking a proton donor (e.g. serine to alanine, tyrosine to phenylalanine, etc.) should support the conceit. The complete protein sequence should be robust as a supermolecular assembly, as to afford structural verification and identify a sequence in a crystallographic context.

The overall design strategy employed herein builds through increasing complexity by gradually adding variable degrees of freedom to the design target. Starting from (1) a set of mathematically describable coiled-coil structures, identify (2) suitable binding geometries and orientations which accommodate the PZnPI chromophore. This set of initial structures seeds a Monte Carlo optimization to relax the coiled-coil around the cofactor, followed by (3) additional Monte Carlo trajectories to identify structures that support both a complementary hydrophobic core and a specific hydrogen-bonding residue in ideal proximity to the cofactor diimide (acceptor). From these structures, (4) a series of loops are generated by satisfying loop closure to form a single chain, drawing upon natural backbone probabilities of flexible poly-glycine segments. Finally, (5) a full sequence is identified as contextualized in a well ordered lattice environment to promote potential crystallization trials. A visualization of these steps can be found in Figure 4-2.

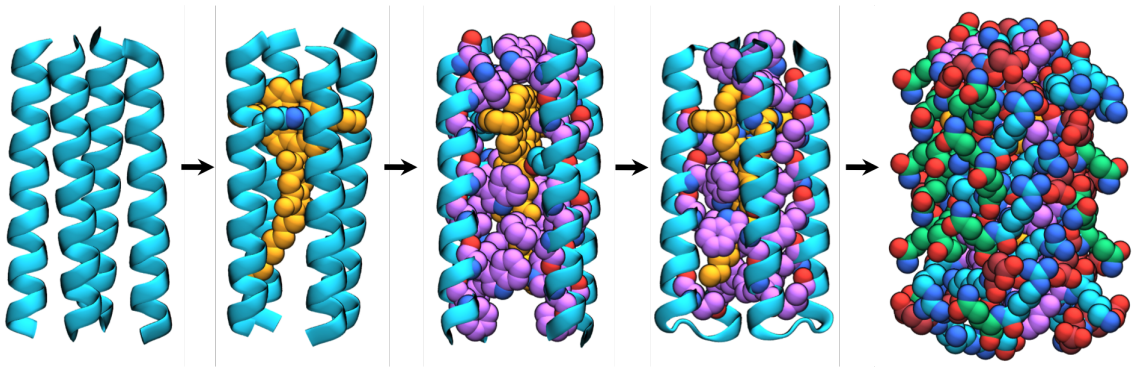


**Figure 4-2.** Overview of the design for a single chain protein construct to bind, orient, and order the PZnPI chromophore. Each step is fully elaborated in subsequent sections, which identify: (1) a set of coiled-coil structures, (2) a suitable binding geometry to encase the PZnPI cofactor, (3) an optimized hydrophobic core, (4) flexible loop segments, and (5) a full sequence in the context of a predetermined space group.

## 4.3. Modeling of Bundles of Helices Encapsulating the PZnPI Cofactor

As with previous work[13,65,88,89], the design targets the encapsulation of the PZnPI chromophore within a coiled-coil protein bundle. Placement within the core provides a variety of advantages including (1) a means to solubilize the cofactor, (2) a sheath to both inhibit cofactor-cofactor aggregation and mitigate cofactor-cofactor interactions, (3) control over the orientation of the cofactor within the assembly, and (4) the potential to leverage protein interactions to immobilize the assembly on a surface. As such, we address these advantages by tailoring a protein core which provides shape-complementarity to the cofactor.

### 4.3.1. The Coiled-Coil Scaffold

The coiled-coil scaffold is an excellent candidate for accommodating large guest molecules into a protein structure. Many naturally occurring coiled-coil proteins leverage the tubular structure to encapsulate hemes, anesthetics, and metal ions[212,213]. The coiled-coil motif is a widely studied protein fold, contained in an estimated 5-10% of protein structures[189,214,215]. Furthermore, the fold is well defined and predictable. Modeling and manipulating the coiled-coil backbone can be easily achieved through Crick's parameterization[188], which has been shown to closely capture most naturally occurring coiled-coils[189].

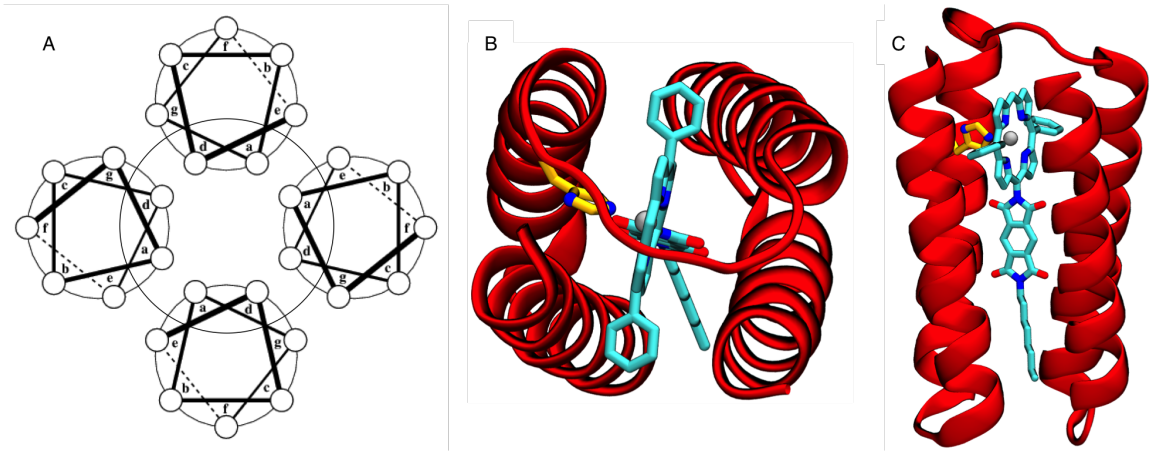



**Figure 4-3.** (A) Coiled-coil (helical wheel) representation of an antiparallel tetramer, indicating the heptad repeat. Each heptad position is denoted by the lower case letters *abcdefg*. The core is highlighted to indicated the interior *a* and *d* positions. (B) Top-down and (C) side views of one such antiparallel tetramer (red) that accommodates the PZnPI cofactor (cyan). Binding is achieved by placement of a single histidine residue (yellow) close to the *a* position.

A four helix bundle was selected as the protein scaffold. A helical wheel diagram is given in Figure 4-3A, indicating positions that correspond to the core of a coiled-coil tetramer, where Figures 4-3B and 4-3C offer renderings of a possible placement for the cofactor. The bundle was modeled as an idealized coiled-coil from the well known Crick parameterization. Helices were oriented in an antiparallel configuration to allow for a single chain construct and target an aspect ratio between protein and cofactor of 1:1.

*De novo* template structures were generated utilizing a coiled-coil builder. The builder manipulates the coiled-coil structure via a set of mathematically describable parameters (many derived from Crick[188]) by placing alpha carbons along the coiled-coil pathway and adding the remain poly-L-glycine heavy atoms upon completion[189]. The five tunable parameters of interest in this design are illustrated in Figure 4-4. Helices can be rotated about their (the minor) helical axes, $\theta$, either individually or in concert (4-4A). For this work, we consider the rotation of each alpha helix as a separate parameter, $\theta_i$. The value of $\theta = 0$ directs the first alpha carbon at the coiled-coil axis. The superhelical pitch, $P$, alters the coiled nature of the bundle (4-4B top), and is related to $\rho$, the projected alpha helical residues per

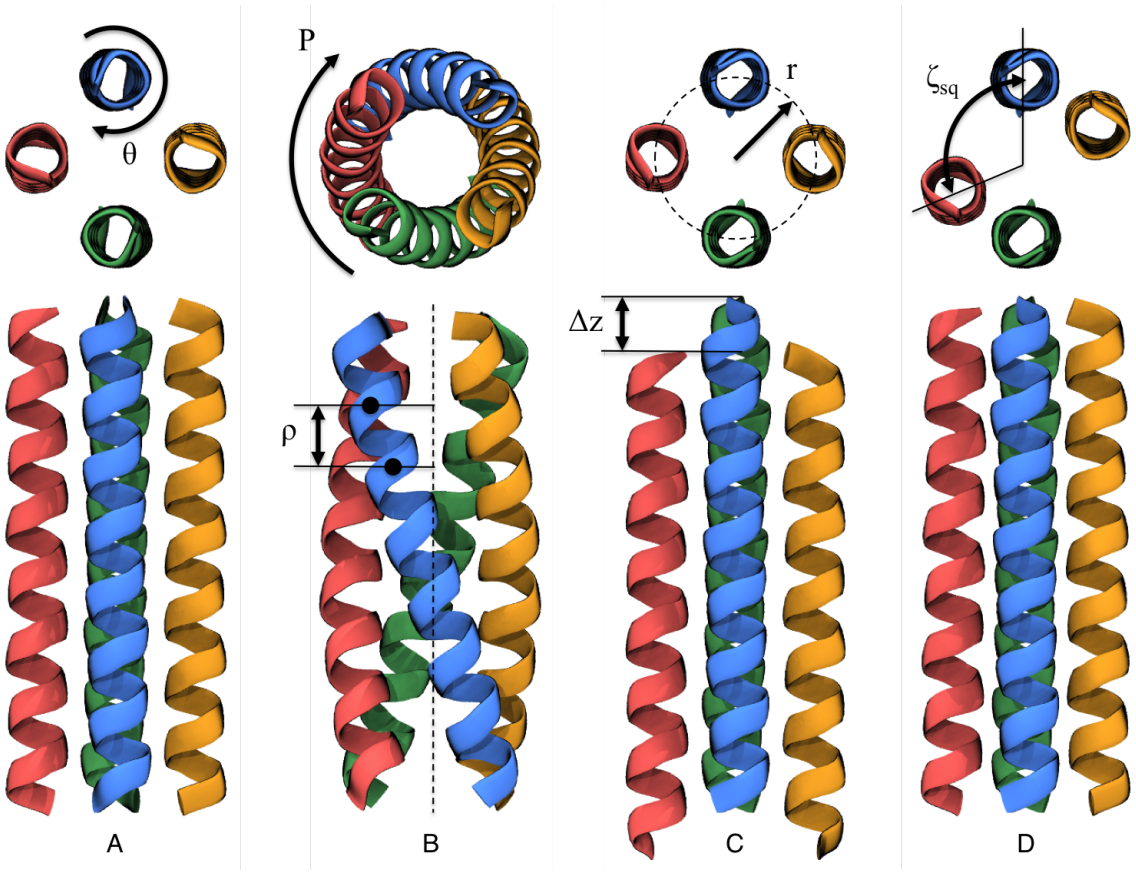



**Figure 4-4.** Structural parameters associated with variations in the coiled-coil structure. (A) Rotation of individual helices about their alpha helical axis, i.e. the minor helical axis. (B) Variation in the super helical pitch of the coiled-coil, as well as the projected residues per turn of the alpha helix onto the superhelical axis. (C, top) Variation in the super helical radius. (C, bottom) For anti-parallel helices, variation in inter-helical offset along the bundle axis. (D) Variation in the inter-helix offset in the x-y plane, hereafter termed 'bundle squareness'.

turn onto the superhelical axis (4-4B bottom). Manipulation of the super helical radius, $r$, allots for the expansion/contraction of space in the core (4-4C top). Antiparallel helices are permitted to transverse the coiled-coil axis with respect to each other by some axial offset, $\Delta Z$ (4-4C bottom). Lastly, variation in the inter-helical offset within the plane perpendicular to the coiled-coil axis results in modifying the "squareness", $\zeta_{sq}$ of the bundle. When $\zeta_{sq} = 0\,^{\circ}$, the bundle is in an idealized square orientation; for values that approach $\zeta_{sq} < 90\,^{\circ}$, the bundle becomes more and more rectangular (4-4D).

For all coiled-coil structures, the rise per residue, $d$, is fixed to 1.495 and the minor helical radius is fixed to 2.26 Å (i.e. a standard coiled-coil). While the superhelical pitch can be manipulated directly, but we instead opt to alter $\rho$, the projected residues per turn onto the superhelical axis. This requires the use of the Fraser MacRae constraint[190] to enforce the presence of a heptad repeat through the structure. The constraint is given as

**Eq. 4-1**

$$P = \sqrt{\left(\frac{d3.5\rho}{3.5-\rho}\right)^2 - (2\pi r)^2}$$

The presence of 3.5 in **Eq. 4-1** arises from guaranteeing the heptad repeat through a minor helical angular frequency of $\frac{7}{2}$. This twist differential thus bounds $\rho$ by idealized straight helices (no twist differential) at $\rho = 3.5$. Values of $\rho$ are enforced to create right-handed coiled-coil structures, where pitch is expressed as a positive quantity.

### 4.3.2. Zinc Porphyrin Coordination Geometry

For the zinc porphyrin to be positioned on the interior of the bundle, a binding site was engineered to satisfy a Zn(II) penta-coordination via the placement of a single histidine[216]. The binding geometry was drawn from x-ray crystal coordinates of a horseheart cytochrome/Zn porphyrin substituted cytochrome C peroxidase complex (PDB Code: 1U75)[217]. The crystal structure features the desired histidine axial ligation of a zinc porphyrin. These coordinates are used to model the coordination, and are defined by the bond length between the histidine epsilon nitrogen and zinc, fixed to 1.95 Å, and the bonding angle fixed between the zinc, histidine epsilon nitrogen, and histidine epsilon carbon, fixed to 131.5° (Figure 4-5).

Force field parameters were adapted from previous modeling of synthetic cofactors[13] as to be consist with parameterization of the modified AMBER84 force field excluding all explicit hydrogen atoms[36]. The porphyrin zinc atom nonbonding parameters were set to $R_{min} =$ 2.18 Å and $\epsilon = 0.25$ kcal/mol. The four carbonyl oxygens on the diimide were given hydrogen bonding acceptor parameters of a 1.6 Å acceptor radius and an $sp^2$ hybridization consistent

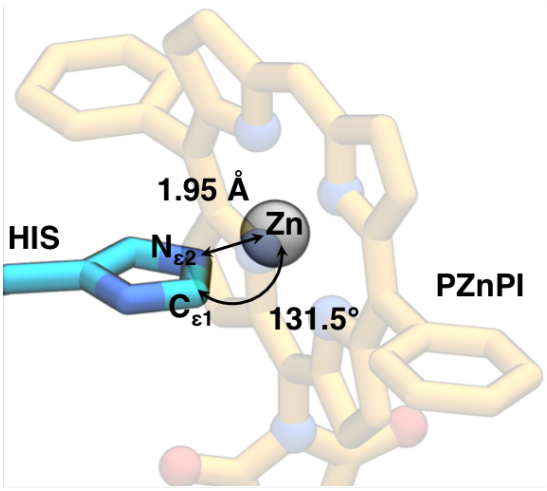



**Figure 4-5.** Binding geometry defining the pentacoordinated zinc metal at the center of the PZnPI porphyrin ring. The rendering highlights the positioning of a histidine residue (cyan) to satisfy the coordination geometry obtained from a cytochrome/Zn porphyrin substituted cytochrome C peroxidase crystal structure (1U75). While the defined bond lengths and angles are held fixed, the dihedral centered about the histidine-porphyrin ligation is permitted to rotation freely during modeling.

with other carbonyl acceptors specified by Stickle et. al.[162] A topology was created for the entire histidine-PZnPI complex (HIS-PZnPI) consistent with the AMBER84 histidine topology and the bonding structure of the PZnPI, such that the axial zinc coordination was treated as a bond. The initial geometry of the remainder of the cofactor was taken from a low energy structure obtained through a minimization, specifying the torsional angle between porphyrin and diimide at 90°, and the octyl tail and a full extended trans-trans configuration.

## 4.4. Optimization of a Hydrophobic Binding Pocket

Placement of the PZnPI cofactor within the bundle requires balancing constraints on satisfying the axial histidine ligand, degrees of freedom associated with the cofactor's structural variability, and alteration of the coiled-coil parameters. Instead of independently varying the cofactor's placement and optimizing a histidine coordination for each candidate bundle, we opt to treat the entire HIS-PZnPI complex as a super rotameric state of histidine. Where HIS bears two degrees of freedom (side chain torsional angles $\chi_1$ and $\chi_2$), the HIS-PZnPI

super rotamer can be permitted many more. Here, the original histidine side chain torsional angles are kept variable, and three additional dihedrals are introduced into the super rotamer. To allow variable rotation about the ligation bond, the torsional angle spanning the histidine epsilon nitrogen - porphyrin zinc is allowed to rotate. Additionally, the torsional angle spanning the bond between the porphyrin and PI diimide is allowed to rotate, as is the torsional angle defining the position of the first aliphatic carbon in the octyl tail. All five dihedrals are illustrated in Figure 4-6. In this way, each probabilistic design calculation can accommodate multiple possible cofactor positionings/conformations, and in turn obtain a probability distribution for these configurations within the context of the design sequence ensemble – and in turn a high probability/optimal placement among that set. We detail at each step what ensemble of HIS-PZnPI states are introduced into the sequence ensemble calculations.

As indicated in Figure 4-4, the HIS-PZnPI super rotamer is placed at an interior helical position: here the second *a* position on the first chain (chain A, resid 8). Because the design exists in the context of a *de novo* structure, the position at **A8** does not necessarily guarantee appropriate placement of the cofactor in the bundle's core. Choosing the histidine side chain dihedrals alone will place the extremities of cofactor is wildly different orientations. Different minor helical rotations will potentially sample both helical interior positions, *a* and *d*, as well as the helical interface, *g* and *e*. In short, the choice of A8 does not specify anything beyond a position about which the coiled-coil and super rotameric dihedral angles must be optimized.

### 4.4.1. Initial Core-Positioned PZnPI Geometries

As the cofactor-protein assembly possesses a large number of degrees of freedom (up to eight associated with the coiled-coil, and up to five for the HIS-PZnPI super rotamer), a stochastic sampling of structures provides an efficient means of identifying low energy structures. Instead of sampling these structures from random initial conditions, we opt to find a subset of reasonably positioned HIS-PZnPI conformations to act as structural seeds.

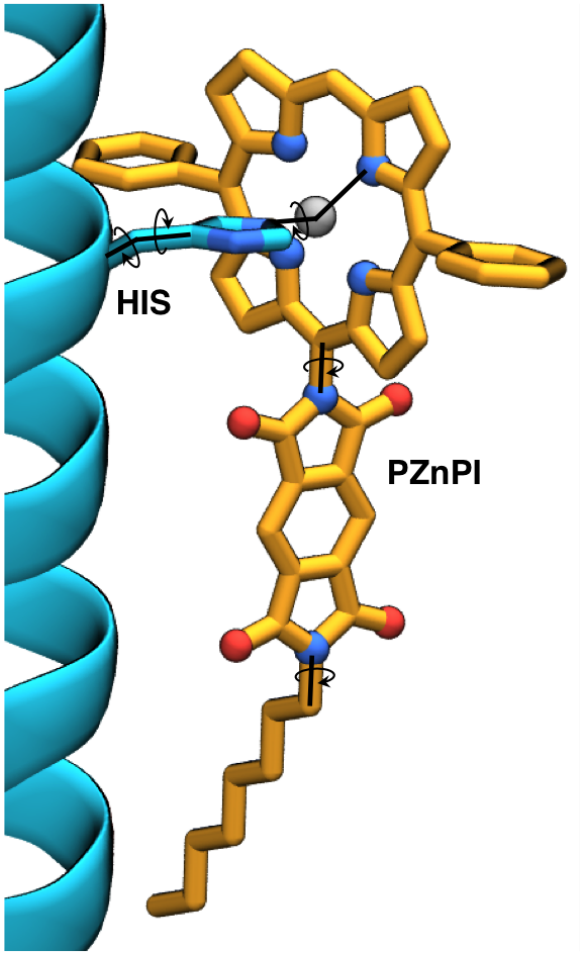



**Figure 4-6.** Diagram of the HIS-PZnPI super rotamer state. Variable dihedrals include the (i, ii) the side chain dihedral angles of the histidine residue, (iii) the rotation about the ligation between the histidine $N_{\epsilon 2}$ and the PZnPI zinc, (iv) the rotation about the bond between the zinc porphyrin and the diimide, and (v) the rotation about the bond connecting the octyl tail to the diimide.

Initial configurations are taken from a grid search over only the coiled-coil radius and minor helical rotation. To simplify the search space, all helices are rotated symmetrically and all other bundle parameters held at fixed values. Additionally, the protein sequence is kept as poly-alanine except the single HIS-PZnPI residue position. This is done to exclude glycine from the helical segments of the protein and prevent spurious packing of the cofactor against any interior glycine. Suitable configurations for the HIS-PZnPI cofactor are obtained from a discretized ensemble of possible conformations, as given by the rotameric variability depicted in Figure 4-6. For this round, the cofactor was modeled without the octyl tail and $\chi_4$ held at 90 °.

The most efficient means of obtaining HIS-PZnPI configurations is to leverage an alignment methodology developed elsewhere (see Chapter 3). Given the goal of restricting the cofactor orientation within the bundle's core, this will provide a fast, accurate way of selecting super rotamers which sufficiently align PZnPI to the coiled-coil core. From these structures, the usual elimination based upon high energetic overlap with the backbone scaffold should provide conformations commensurate with the superhelical parameters of the bundle.

For the HIS $\chi_1$, $\chi_2$ pair, we are trying to obtain a rotation about the zinc ligation bond which optimizes the cofactor's alignment with the coiled-coil axis. This maps directly on the previous alignment equations for specifying a super rotamer dihedral $\chi_3$. Figure 4-7 maps these angles onto the HIS-PZnPI cofactor, specifying that the vector spanning the porphyrin face vertically best aligns to the z-axis. This is given by the equation

**Eq. 4-2**

$$\vec{b_1} = \overrightarrow{A_{CHD} - A_{CHB}}$$
$$\vec{b_2} = \overrightarrow{A_{Zn} - A_{N_{\epsilon 2}}}$$
$$\chi_3 = atan2(\vec{b_1} \cdot (\hat{b_2} \times \hat{z}), (\vec{b_1} \times \hat{b_2}) \cdot (\hat{b_2} \times \hat{z}))$$

such that for any given $\chi_1$, $\chi_2$, there exist a value of $\chi_3$ that maximizes the orientational alignment to the coiled-coil axis. This solution both restricts the possible placement of the PZnPI cofactor, while reducing the search space for super rotamer states (from $N^3$ to $N^2$

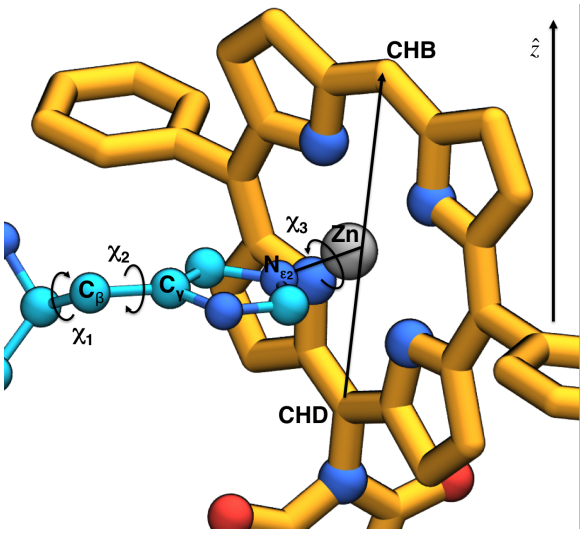



**Figure 4-7.** Atomic diagram of the HIS-PZnPI super rotamer with vectors utilized in solving for $\chi_3$. The figure directly cites the formulation used to solve for $\chi_3$ in the GLU-uranyl (Figure 3-4). Atoms in the HIS-PZnPI super rotamer are labeled according to the AMBER molecular topology, along with the designated atomic names for the PZnPI cofactor (see Appendix). Each of the dihedral angles of the the histidine are illustrated, $(\chi_1, \chi_2, \chi_3)$. For a particular set of dihedral values for $(\chi_1, \chi_2)$, **Eq. 4-2** is used to solve for the value of $\chi_3$ that most aligns the verticality of the cofactor with the $z$ axis.

states where $N$ is the total possible states for any of the dihedrals). $\overrightarrow{A_i}$ corresponds to the 3-dimensional coordinate of the named atom, as defined in Figure 4-7. Note that $\vec{b_2}$ is normalized in **Eq. 4-2**, denoted as $\hat{b_2}$. Furthermore, the definition of $\vec{b_1}$ explicitly assumes a planar orientation of the porphyrin ring as approximated by the positioning of carbons $CHB$ and $CHD$.

For each coiled-coil structure, $\chi_1$ and $\chi_2$ are scanned over the full angular range $[-180\,°, 180\,°]$ with $\Delta\chi_i = 1\,°$, solving for the optimal value of $\chi_3$ using **Eq. 4-2** (32,400 orientations of HIS-PZnPI super rotamer per coiled-coil structure). These orientations are then pruned to assure a desired PZnPI placement within the bundle.

(1) For each super rotamer structure, three evaluations determine retention in the super rotamer ensemble. The first is a simple evaluation of nonbonding potential with the protein scaffold (here, a poly-alanine coiled-coil). If the total nonbonding energy between the super rotamer and scaffold exceeds 30 kcal/mol, the the state is removed.

(2) The orientation of the porphyrin face is assessed to anticipate the reintroduction of the octyl tail. If the angular alignment of the porphyrin spanning vector and the coiled-coiled axis exceed some angular tolerance, $\delta_\theta$, then the super rotamer is removed. Here, we select a strict tolerance of $\delta_\theta = 2\,^\circ$. The vectors are defined between Figure 4-7 and **Eq. 4-2**, and the tolerance test given as

**Eq. 4-3** $$\text{acos}(\hat{b_1} \cdot \hat{z}) \leq \delta_\theta$$

(3) The positioning of the PZnPI within the tubular interior of the coiled-coil is evaluated. This is done to avoid orientations that are not within the coiled-coil core but still energetically favorable, such as placement to one side of the bundle between only two alpha helices. We remove any super rotamer whose zinc position is more than some distance in the plane perpendicular to coiled-coil axis (here, x-y plane to z-axis), $\delta_{xy}$. The modest choice of $\delta_{xy} = 3\mathring{A}$ affords a wide range of PZnPI orientations while still retaining the cofactor in the bundle core.

**Eq. 4-4** $$\sqrt{{A_{x,Zn}}^2 + {A_{y,Zn}}^2} > \delta_{xy}$$

where $A_{x,Zn}$ and $A_{y,Zn}$ are the Cartesian x- and y-coordinates of the zinc atom and the z-axis is coincident with the superhelical axis of the bundle. The remaining subset of super rotamer states comprise the ensemble of HIS-PZnPI states permissible in a given coiled-coil structure. We note that in practice, these eliminations are performed in reverse order to minimize computational cost; that is, the distance calculation is less expensive than the arccos operation, which is less expensive than the energetic scoring.

A grid search is performed to obtain a rough estimate of initial bundle/cofactor configurations. For this calculation only, each chain in the 4 helix bundle was extended to 28 residues in length, and the HIS-PZnPI (no octyl tail) was placed an additional heptad position down (chain A, resid 15). This is meant to model an extended bundle and remove

| Label | Radius ($r$) | Rotation ($\theta$) | Squareness ($\zeta_{sq}$) | $\chi_1$ | $\chi_2$ | $\chi_3$ |
|---|---|---|---|---|---|---|
| $CC_1$ | 8.06 Å | 18.0 ° | +15 ° | -104 ° | -152 ° | -41.8 ° |
| $CC_2$ | 8.44 Å | 4.0 ° | +15 ° | -132 ° | -108 ° | -77.4 ° |
| $CC_3$ | 9.20 Å | 10.0 ° | -15 ° | -100 ° | +42 ° | 121.4 ° |
| $CC_4$ | 9.38 Å | -11.5 ° | -15 ° | -166 ° | +18 ° | -165.0 ° |
| $CC_5$ | 8.78 Å | -37.5 ° | -15 ° | +166 ° | +46 ° | -164.3 ° |

**Table 4-1.** Structural parameters for selected initial configurations in Figure 4-8. The three $\chi$ angles describe the first three dihedral angles of the HIS-PZnPI super rotamer.

any possible cofactor interactions with termini. All remaining residues were fixed as alanine. The landscape spans variation in the bundle radius, $r$, and minor helical rotation, $\theta$ (where all helices are rotated by the same amount). The axial offset is fixed at 0 Å, and the residues per turn fixed at 3.5 to model "straight" helices. The superhelical pitch was obtained through the Fraser MacRae constraint to ensure the heptad repeat. To provide space for the rectangularity of the cofactor without requiring extremely large bundle radii, we generate two landscapes at nonzero values of the squareness parameter.

The generated landscapes are shown in Figure 4-8. In each, the bundle radius was varied from 7.5 Å to 9.5 Å at increments of 0.02 Å. Minor helical rotation was varied from $-60\,^\circ$ to $60\,^\circ$ at increments of $0.5\,^\circ$. Drawing from squareness values in previous design work[13], the value was fixed at either $+15\,^\circ$ or $-15\,^\circ$. Calculations determined super rotamer probabilities, and in turn mean field energies, for each ensemble holding the effective inverse temperature $\beta$ at a room-temperature value of $\sim$1.69 mol/kcal.

From the two generated landscapes, five distinct initial constructs were chosen (Table 4-1). Renderings of the variation in cofactor placement are shown in Figure 4-8B, orange indicating selected bundles with positive squareness and green indicating those with negative squareness. An overlay of the histidines is shown in Figure 4-8C, highlighting the dramatically different starting rotamer states. These five structures are then used as initial configurations (seeds) for subsequent Monte Carlo sampling; the following description of Monte Carlo search runs five calculations in parallel from these starting constructs.

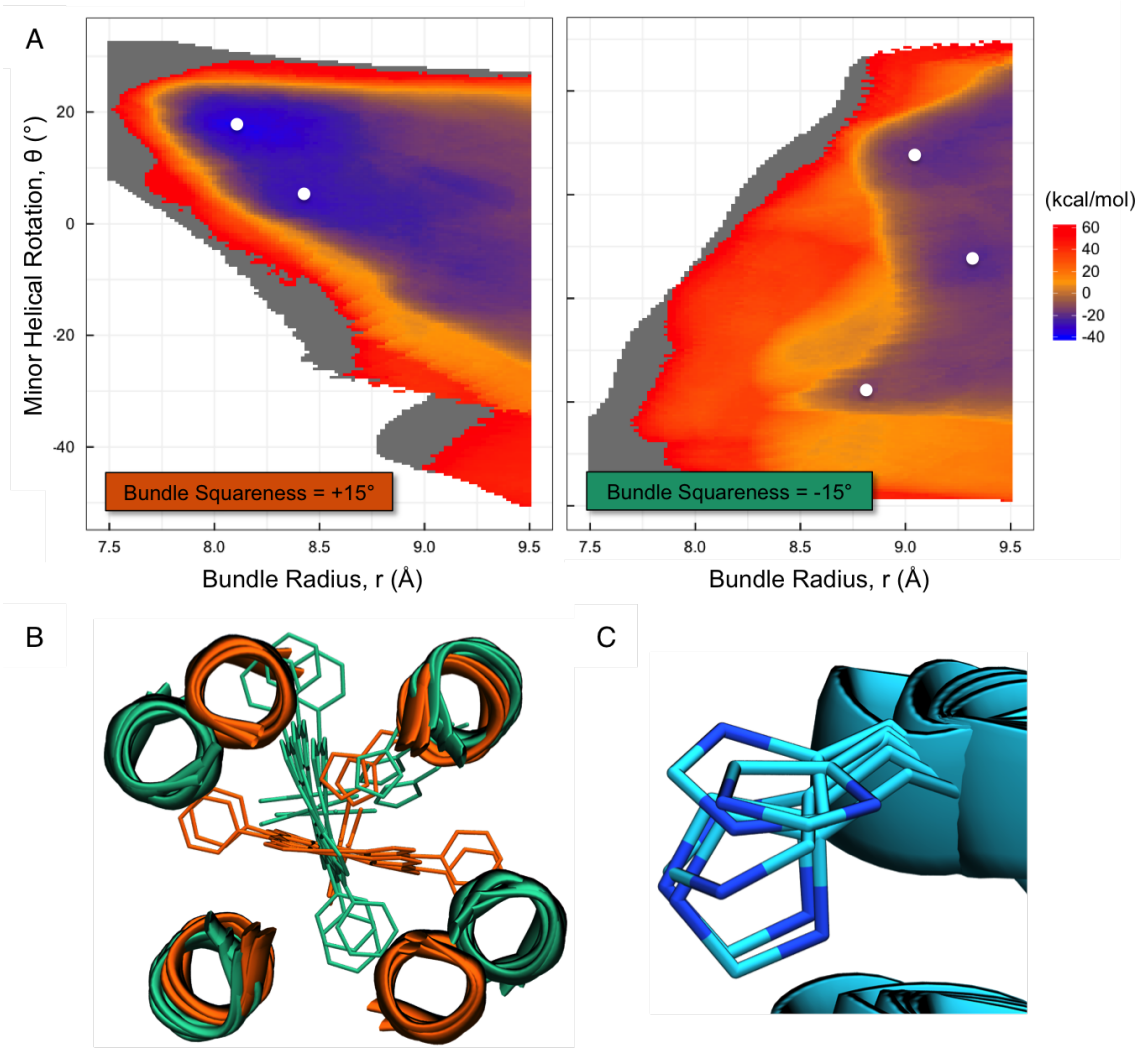



**Figure 4-8.** (A) Mean Field Energy of the poly-L-glycine antiparallel tetramer with a single HIS-PZnPI binding site (chain A, residue position 15). Each point draws from configurations specified by **Eq. 4-2** and the series of described eliminations. The energy encompasses the dihedral, van der Waals, hydrogen bonding and electrostatic potentials. White space denotes coiled-coils that, after cofactor elimination, cannot accommodate the PZnPI cofactor on the interior. Gray tiles indicate energies above 60 kcal/mol. Five minima are drawn from the surfaces, two from bundles with a fixed squareness of $+15\,^\circ$ (left), and three from bundles with a fixed squareness of $-15\,^\circ$ (right). (B) Alignment of the five selected structural minima drawn from (A). Renderings in orange denote structures with a fixed squareness of $+15\,^\circ$, renderings in green denote structures with a fixed squareness of $-15\,^\circ$. (C) Alignment of histidine states in each of the five selected structural minima drawn from (A). The overlay highlights the different starting configurations for valid PZnPI encapsulation.

### 4.4.2. Monte Carlo Sampling of Coiled-Coil Structures

As we wish to search over a large number of structural parameters associated with both the coiled-coil and the super rotamer, exhaustive enumeration over the conformational search space quickly becomes intractable. Instead, Monte Carlo (MC) sampling is employed to stochastically sample the structural parameters and hasten the identification of low energy structures. MC utilizes an effective inverse temperature, $\beta_{MC}$, which defines the likelihood of sampling high energy regions of the search space. The structural searches targeted are discrete and the ability to escape possible local minima makes MC advantageous. The search is set up in the usual way: each step is a generated trial on a MC Markov chain with an associated acceptance criterion according to the Metropolis acceptance probability[218] for some scoring function $f(\mathbf{w})$

**Eq. 4-5** $$a(\mathbf{w}, \mathbf{w}') = \min(1, \exp(-\beta_{MC}(f(\mathbf{w}') - f(\mathbf{w}))))$$

A slow cooling of $\beta_{MC}$, i.e. simulated annealing, provides a good means for estimating minima. An initial temperature is set to a large value to improve sampling and lowered by some cooling schedule. We choose an exponential decay schedule from an initial temperature ($T_{0,MC}$) to some temperature at MC step $n$ ($T_{MC}(n)$)

**Eq. 4-6** $$T_{MC}(n) = T_{0,MC} \exp(-n/\tau_c)$$

or expressed in terms of $\beta_{MC}$,

**Eq. 4-7** $$\beta_{MC}(n) = \frac{\beta_{0,MC}}{\exp(-n/\tau_c)}$$

The decay constant $\tau_c$ is chosen such that after some $N$ MC steps, the acceptance is a chosen final inverse temperature $\beta_{MC,f}$.

**Eq. 4-8**

$$\tau_c = \frac{N}{ln(\beta_{f,MC}/\beta_{0,MC})}$$

At each MC step, the structural update occurs by selecting one of the conformational variables at random, and adjusting it by some bounded change based an an allowed maximum change. Here, the conformational variables are defined as the set of eight coiled-coil parameters (bundle radius $r$, individual helical rotation of the four chains $\theta_i$ , superhelical residues per turn $\rho$, axial offset $\Delta Z$, and bundle squareness $\zeta_{sq}$) and four super rotamer dihedrals (histidine side chain dihedrals $\chi_1$, $\chi_2$, PZn-PI dihedral $\chi_4$, and PI-octyl main dihedral $\chi_5$) – twelve degrees of freedom in total. The super helical pitch is constrained by the Fraser MacRae formula, the rise per residue set to 1.495 Å, and the HIS-PZn axial ligation dihedral, $\chi_3$, is assigned for a given ($\chi_1$, $\chi_2$) based on **Eq. 4-2**. At each step, a parameter is randomly selected from this set and assigned a random update amount bounded by value and some maximal change ($\delta x$). That is, for a random value $p(i)$ from $[-1, 1]$, an update at $i$ to structural parameter $x$ is given as

**Eq. 4-9**

$$x_{i+1} = \begin{cases} x_{min} & : x_{i+1} \leq x_{min} \\ x_i + p(i) \cdot \delta x & : x_{min} < x_{i+1} < x_{max} \\ x_{max} & : x_{i+1} \geq x_{min} \end{cases}$$

A summary of values is given in Table 4-2. Bundle radius is bounded to encompass the natural range of antiparallel tetramers (6.3 Å to 7.8 Å)[189], as well as a reasonable expansion up to 9.5 Å should the cofactor require it. Helical rotations are bounded such that the starting heptad $a$ position only sweeps the $e$ position, through the core, to the $g$ interface position. The bundle squareness is bounded by $\pm 20^\circ$ to prevent interhelical contacts from

| **Parameter** | $x_{min}$ | $x_{max}$ | $\delta x$ |
|---|---|---|---|
| $r$ | 6.5 Å | 9.5 Å | 0.5 Å |
| $\theta_A$ | -60 ° | 60 ° | 10 ° |
| $\theta_B$ | -60 ° | 60 ° | 10 ° |
| $\theta_C$ | -60 ° | 60 ° | 10 ° |
| $\theta_D$ | -60 ° | 60 ° | 10 ° |
| $\rho$ | 3.5 | 3.65 | 0.03 |
| $\Delta Z$ | -2.0 Å | 2. Å | 0.5 Å |
| $\zeta_{sq}$ | -20 ° | 20 ° | 5 ° |
| $\chi_1$ | -180 ° | 180 ° | 10 ° |
| $\chi_2$ | -180 ° | 180 ° | 10 ° |
| $\chi_4$ | 80 ° | 100 ° | 10 ° |
| $\chi_5$ | -180 ° | 180 ° | 10 ° |

**Table 4-2.** Monte Carlo structural parameter update criteria, including value boundaries and maximal change per MC step. The parameters listed are the bundle radius $r$, individual helical rotation of the four chains $\theta_i$, superhelical residues per turn $\rho$, axial offset $\Delta Z$, bundle squareness $\zeta_{sq}$, the histidine side chain dihedrals $\chi_1$, $\chi_2$, PZn-PI dihedral $\chi_4$, and PI-octyl main dihedral $\chi_5$. Omitted structural parameters are either held at a fixed value, or calculated based on the update as described in the text.

dominating calculation energetics. Lastly, the PZn-PI dihedral, $\chi_4$, is bound as $90\,^\circ \pm 10\,^\circ$ to allow flexibility without severe twisting of the donor-acceptor bridge.

The first MC trajectory, starting from the structures identified in Table 4-1, identifies low energy poly-alanine bundle/cofactor configurations in the context of all 12 structural degrees of freedom. This trajectory was run for $N$ = 1,000,000 steps, at a constant MC temperature ($\beta_{0,MC} = \beta_{f,MC} = 0.5$). The four helix antiparallel bundle was shorted to 25 residues per helix. The cofactor's octyl tail was reintroduced and rotated in each structure to be positioned between the wider interhelix interface. The sequence was fixed with the HIS-PZnPI rotamer at position **A15** with ALA elsewhere – in effect a fixed structure with no sequence/conformational variability. As such, the choice for $f(\mathbf{w})$ was the local energy of the HIS-PZnPI super rotamer, $\epsilon_{HIS-PZnPI}$. Lowest energy structures from each of the five trajectories were taken as seeds for the subsequent round of MC sampling (Figure 4-9A).

A second round of MC sampling introduced hydrophobic residues to the bundle core. At each MC step, the interior heptad positions $a$ and $d$ were dynamically identified based on

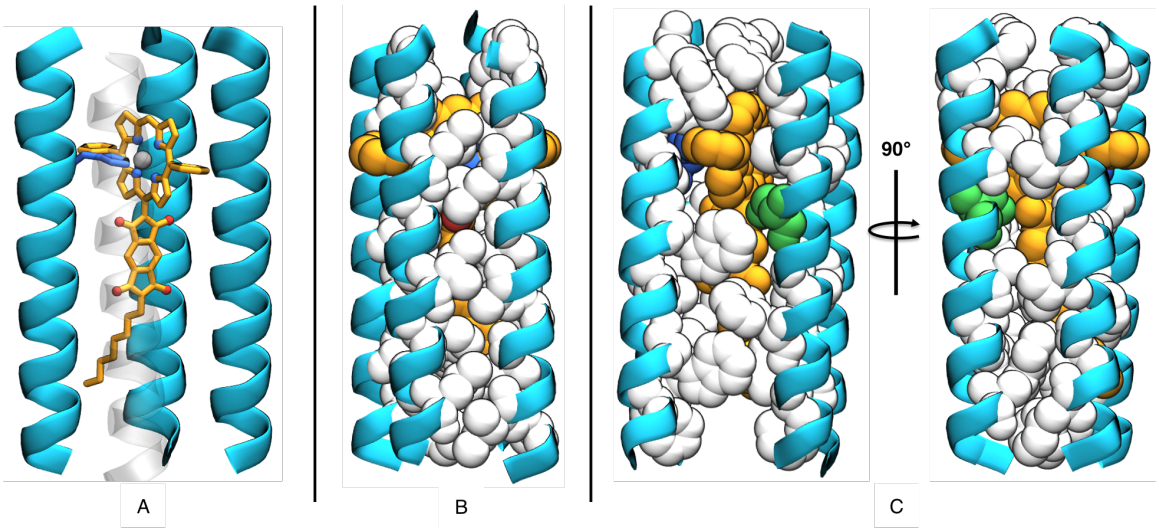



**Figure 4-9.** Representative low energy structures/sequences from each round of Monte Carlo sampling. (A) MC trajectory energy minimum for poly-alanine coiled-coil encapsulating the HIS-PZnPI super rotamer. (B) MC trajectory mean field energy minimum for hydrophobic core packing around the HIS-PZnPI super rotamer. (C) MC trajectory combinatorial energy minimum for hydrophobic core packing around the HIS-PZnPI super rotamer with an optimized hydrogen bonding interaction (rendered: SER, green).

the current coiled-coil parameters. These positions were typed with a restricted set of hydrophobic residues (A,V,I,L,F) which excludes the larger W and M residues and encourages a close packing core. This trajectory was run for $N$ = 200,000 steps, at a constant MC temperature ($\beta_{0,MC} = \beta_{f,MC} = 0.5$). In including the hydrophobic ensemble in the core, $f(\mathbf{w})$ is chosen as an ensemble average at the calculated probabilities, here the change in mean field local energy, $\langle \epsilon_{HIS-PZnPI} \rangle$. Again, lowest energy structures from each of the five trajectories were taken as seeds for the subsequent round of MC sampling (Figure 4-9B).

For each of the five low energy structures identified in the previous MC trajectory, potential cofactor hydrogen bonding residues were identified within the interior of the protein. Each residue position in these five coiled-coil structures was evaluated for possible hydrogen bonding interactions between the diimide carbonyl oxygen and any candidate. Candidate residues were chosen such that they would lend to simple mutation – recall the motivation to identify the role hydrogen bonding interactions play in charge transfer dynamics by engineering a core with and without a hydrogen bonding interaction. As such, we consider

| Label | Chain | Position | Residue |
|---|---|---|---|
| $CC_1$ | C | 19 | TYR |
| $CC_2$ | C | 11 | SER |
| $CC_3$ | B | 19 | SER |
| $CC_4$ | D | 11 | SER |
| $CC_5$ | B | 11 | SER |

**Table 4-3.** Cofactor hydrogen bonding candidates for each of the five bundle structures, listing position in the bundle and amino acid identity.

serine, threonine, and tyrosine with the intention of working with a second core where the nonpolar residues alanine, valine, or phenylalanine, respectively, are placed in their stead (S $\rightarrow$ A; T $\rightarrow$ V; Y $\rightarrow$ F).

Iterating through each residue location, each of the three candidate residues (S T Y), and respective Dunbrack rotamers[47], a hydrogen bonding interaction[36] is calculated between the candidate and the HIS-PZnPI super rotamer. The lowest scoring candidates for each structure are presented in Table 4-1. Moving forward, the positions identified in Table 4-3 were fixed to the candidate hydrogen bonding amino acid and corresponding conformation. An illustration of the identified positioning for $CC_2$ is presented in Figure 4-10.

These results were then incorporated into a final MC trajectory. As before, the interior $a$ and $d$ positions were dynamically typed for each sampled structure, save the hydrogen bonding candidate which was fixed in position and conformation. The 12 structural parameters were again allowed to vary, this time with an applied annealing schedule. The trajectory was run for $N$ = 30,000 steps, at the cooling decay constant specified by MC and **Eq. 4-8** temperatures $\beta_{0,MC} = 0.5$ to $\beta_{f,MC} = 2.0$. Where the previous trajectories sampled according to $f(\mathbf{w})$ as a simple energetic difference, here we wish to track energetic changes for both the entire bundle ensemble, as well as maintain optimality in the selected hydrogen bonding residue. The requires $f(\mathbf{w})$ be defined as a linear combination of the two

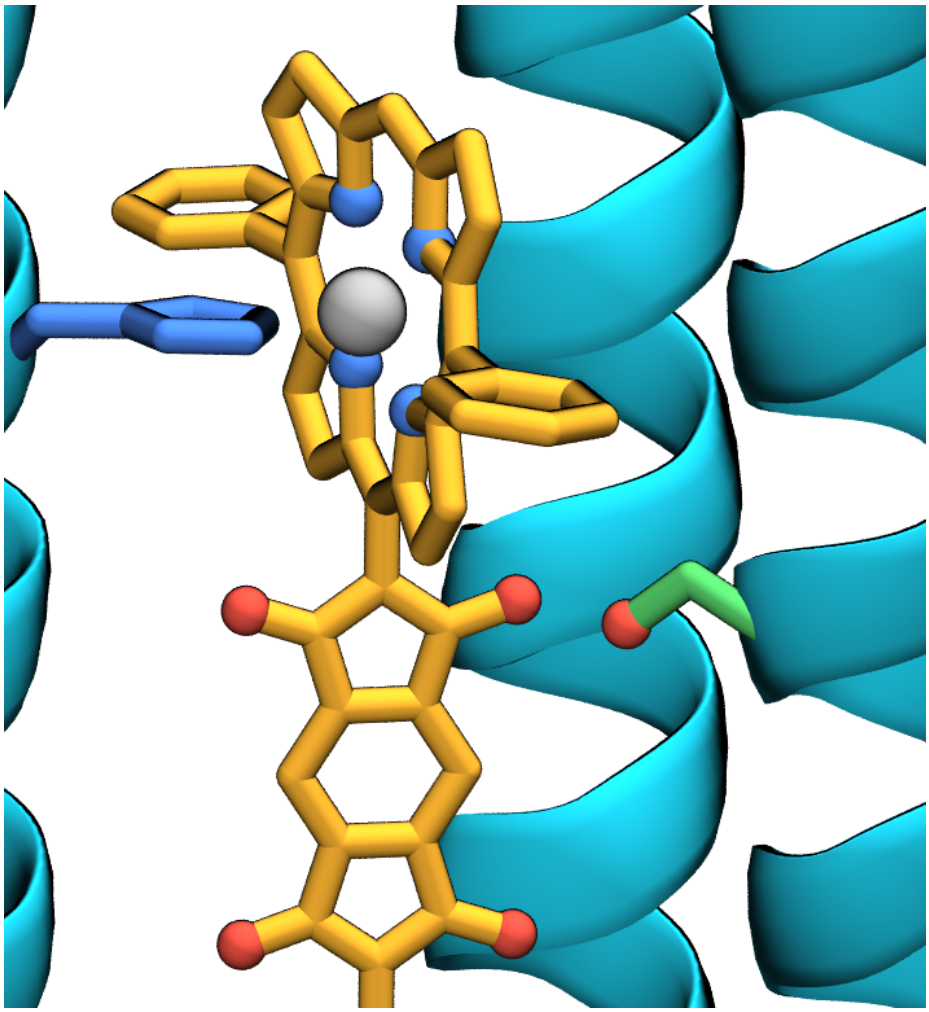


**Figure 4-10.** Rendering of specified hydrogen bonding residue to the diimide carbonyl oxygens of the PZnPI cofactor (yellow) for the $CC_2$ structure. The bundle here specifies a single histidine (blue) to axially ligate the porphyrin zinc (gray), and a serine (green) on the opposing chain to hydrogen bond to the diimide.

target energies.

**Eq. 4-10**

$$f(\mathbf{w}) = \langle E \rangle + \frac{\beta_{HBond}}{\beta_{MC}} \cdot \epsilon_{HBond}$$

where $\langle E \rangle$ is the mean field energy of the entire ensemble, $\epsilon_{HBond}$ is the hydrogen bonding contribution to the interaction between the HIS-PZnPI super rotamer and the selected hydrogen bonding candidate, and $\beta_{HBond}$ an associated acceptance temperature for the hydrogen bonding contribution. The estimation of $\beta_{HBond}/\beta_{MC}$ is derived from energetic statistics for each of the energy terms across the previous MC trajectory ($\beta_{MC} = 0.5$). Across those 200,000 structures, the accepted structures yielded mean field energies of approximately $\langle E \rangle = 460 \pm 6$ kcal/mol. The hydrogen bonding energy, due to the functional form, is optimal at a strength of -1 kcal/mol. As such, we estimate the ratio of fluctuations between the two energies as $\approx \frac{10}{1}$, and set $\beta_{HBond}/\beta_{MC} = 10$. This combinatorial acceptance ratio criterion, combined with the annealing schedule, produced a lowest energy structure from each of the five trajectories (Table 4-4). An example structure is illustrated in Figure 4-9C.

## 4.5. Inverse Kinematics and Loop Design

Transforming the coiled-coil tetramer into a single chain construct requires modeling loops between the helical segments. This may be done by stitching natural loops onto a structure through PDB Select[219] or a database of loop structures[89], with connections identified by visual inspection and overlap scored by backbone RMSD alone. Here, we instead choose to employ inverse kinematics to model *de novo* loops onto the established tetramer constructs.

Loops were modeled with the Cyclic Coordinate Descent (CCD) algorithm[220–222], known to be borrowed from the robotics task of inverse kinematics. The algorithm operates as follows: create an initial loop of specified length $l + 3$, aligned to the last three backbone

| **Label** | $r$ **(Å)** | $\theta_A$ (°) | $\theta_B$ (°) | $\theta_C$ (°) | $\theta_D$ (°) | $\rho$ |
|---|---|---|---|---|---|---|
| $CC_1$ | 8.127 | 17.68 | -32.59 | -19.51 | -54.56 | 3.501 |
| $CC_2$ | 8.321 | 18.79 | -39.13 | 18.18 | -41.57 | 3.520 |
| $CC_3$ | 7.881 | -35.85 | 4.49 | -32.21 | 16.16 | 3.501 |
| $CC_4$ | 7.900 | -32.28 | 10.98 | -32.10 | 11.22 | 3.506 |
| $CC_5$ | 7.764 | -38.59 | 12.07 | -7.71 | 13.55 | 3.505 |
| **Label** | $\Delta Z$ **(Å)** | $\zeta_{sq}$ (°) | $\chi_1$ (°) | $\chi_2$ (°) | $\chi_4$ (°) | $\chi_5$ (°) |
| $CC_1$ | -0.745 | 19.42 | -107.37 | -151.90 | 89.71 | -119.27 |
| $CC_2$ | 0.348 | 16.95 | -107.11 | -148.16 | 99.26 | 85.29 |
| $CC_3$ | 1.026 | -18.65 | -175.57 | 45.80 | 85.76 | 100.63 |
| $CC_4$ | 1.489 | -19.86 | -174.42 | 42.60 | 81.59 | -114.78 |
| $CC_5$ | 1.318 | -18.94 | 174.40 | 41.94 | 83.73 | 78.80 |

**Table 4-4.** Structural parameters for selected lowest energy bundles after the final round of MC annealing. The parameters the bundle radius $r$, individual helical rotation of the four chains $\theta_i$ , superhelical residues per turn $\rho$, axial offset $\Delta Z$, bundle squareness $\zeta_{sq}$, the histidine side chain dihedrals $\chi_1$, $\chi_2$, the PZn-PI dihedral $\chi_4$, and the PI-octyl main dihedral $\chi_5$.

positions on the C-terminus of the first chain. For some number of attempts at closing the loop or until the loop is closed, attempt a loop closure. Each loop closer consists of some number of CCD steps which 1) randomly select a degree of freedom to vary, 2) adjust that degree of freedom to minimize the closure metric, and 3) check for overall closure sufficiency. Loop closure is assessed as the RMSD between the backbone atoms in the last 3 residues of the loop (moving) and the backbone atoms in the first three residues on the N-terminus of the second chain (target). Here, the degrees of freedom selected to vary are the loop backbone dihedrals $(\phi, \psi)$ and are rotated by the geometrical solution outlined by Canutescu et. al.[220] for minimizing the moving/target backbone atoms. RMSD threshold was set to 0.1 Å, the maximum number of CCD step was set to 10,000, and the maximum number of closure attempts was set to 10.

For each of the five low energy structures, three loops were modeled: two short loops between closer adjacent helices, and one longer loop spanning the remaining opposite pair. An illustration of an ensemble of generated loops is shown in Figure 4-11A. Note that for loops to span with minimal length, the starting position on a helix is critical. The helices are trimmed such that the loop termini are oriented in the direction of the opposite helix.

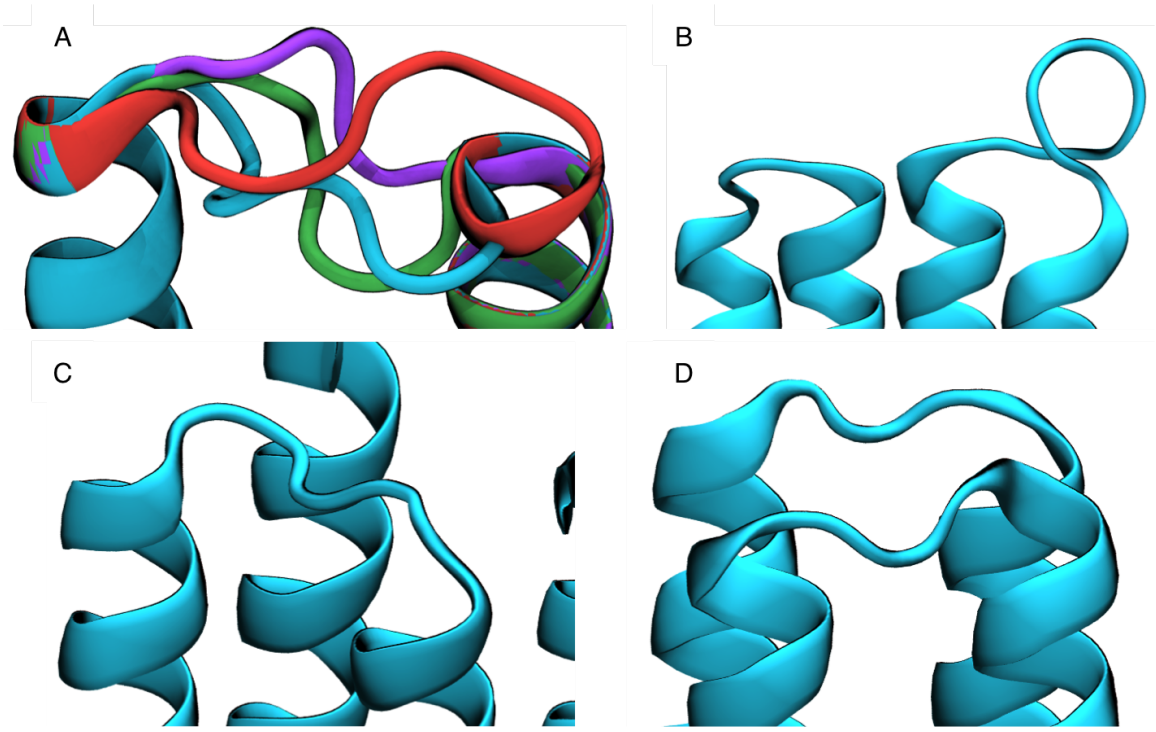



**Figure 4-11.** Renderings of loop modeling onto the antiparallel tetramer. (A) Ensemble of closed loops, showcasing variations in satisfactory conformation given loop length. (B) Example of comparable loop closure problems with wildly different lowest energy solutions. On the left, the loop is of minimal length; on the right, the loop coils back onto itself overcompensating its energetic value with self-interaction. (C) Selected low energy/high probability loop at the further interhelix interface. (D) Selected low energy/high probability loops at the closer interhelix interfaces.

Specifically, for $CC_1$ and $CC_2$ loops are modeled between C25-B2, B24-A2, and A25-D2, and for $CC_3$, $CC_4$, and $CC_5$, loops are modeled between A24-B1, B24-C2, and C24-D1. Henceforth, these loops will be referred to as loops e, f, and g respectively. Loops were modeled as entirely poly-glycine sequences for simplicity.

Based on distance between termini, the e and g (shorter) loops were generated between lengths 3-8 and the f (longer) loop between lengths 5-8. For each loop, for each candidate length, 1000 loops were generated using the CCD algorithm. Each loop was scored by energetic interaction both within the poly-GLY loop and with the poly-alanine coiled-coil containing the HIS-PZnPI (absence of hydrophobic core) by summing over local site energies at each position in the loop, $\epsilon_{loop,i}$. However, this metric alone is likely to bias

towards longer, self-interacting loops. An example of this is illustrated in Figure 4-11B, where extended length provides an a "doubling-back" of the loop. In addition to energetics, neighbor-independent $\phi$,$\psi$ probabilities were estimated from the Dunbrack neighbor-dependent Ramachandran probability distributions[223]. A negative log probability average, given as

**Eq. 4-11**
$$\langle -\ln p \rangle_{loop} = \frac{1}{l} \sum_{i}^{l} -\ln(p_{GLY}(\phi_i, \psi_i))$$

was estimated for each loop. An examination of natural structures suggested a negative log probability average cutoff of 10. From the ensemble of generated loops across all considered lengths, the lowest energy loop with a negative log average probability below the cutoff was chosen.

Examination of low energy structures and loops suggested moving forward with the first two structures, $CC_1$ and $CC_2$. For both of these structures, the optimal e and g loops were modeled with $l$=3 and the optimal f loop modeled with $l$=5. These final single chain structures were renumbered so sites were contiguous and all labeled by the same chain. The 107 residues were relabeled from A1 to A107 thusly: C1 → C25 → e1 → e3 → B2 → B24 → f1 → f5 → A2 → A25 → g1 → g3 → D2 → D25. Additionally, the first position (previously C1) was constrained in all subsequent calculations to be GLY to account for the experimental requirement that a terminal GLY residue is left upon cleavage of the TEV protease recognition sequence ENLYFQG.

## 4.6. Full Sequence Design in a Targeted Space Group

Having established a set of bundle parameters which are capable of positioning the co-factor within a tailored hydrophobic core and containing compact loops, calculations were performed to determine residue identities at exterior positions of the single chain protein.

While this could be accomplished with a wide variety of hydrophilic and polar residues, we instead aim to identify a sequence commensurate with protein crystallization. This provides a means to bias the probabilistic sequence calculations towards well-defined protein-protein interactions at lattice interfaces; should crystallization trials be successful, it would also serve to confirm the designed structure with atomic resolution.

There can be up to twelve parameters positioning the asymmetric unit in a lattice: those associated with the space group (three lattice dimension and three lattice angles) and those with the placement of the asymmetric unit within the unit cell (three translational and three rotational degrees of freedom). A high order space group is targeted to, among other reasons, restrict this search space. The layered space groups like P3, P4, and P6 all fix most lattice degrees of freedom, including the layer spacing, $c$ and the constrained in-plane lattice dimension $a$=$b$. Furthermore, by considering the coiled-coil structure as a cylindrical unit placed within the lattice, we constrain its orientation such that it can only translate within the $a$-$b$ plane and rotate about the coiled-coil axis.

To reduce the dimensionality further, the distance between asymmetric cylinder centers can be constrained to be equal throughout the lattice. This is done as a preliminary step to allow a larger sampling of bundle rotations and lattice spacings. Figure 4-12 illustrates this constraint, and Figure 4-13 provides an accompanying rendering highlighting potential packing of such coiled-coils in the P3, P4, and P6 space groups.

The P6 lattice is selected. It is a layered, porous space group, which offers high symmetry (to expedite x-ray data collection and structure determination), with large tubular solvent channels extending through the crystal (a solvent content in the appropriate protein crystal range of 26% - 65%[224]). The P6 space group ($a = b \neq c$; $\alpha = \beta = 90\,^{\circ}$; $\gamma = 120\,^{\circ}$) posseses a six-fold axis of symmetry at the origin; to compensate for this, the asymmetric unit must be translated away to prevent overlap between copies. Parameters associated with the space group and this translocation are defined in Figure 4-14. The P6 lattice possesses five parameters: the two free unit cell dimensions $a$=$b$ and $c$, the the radial offset from the

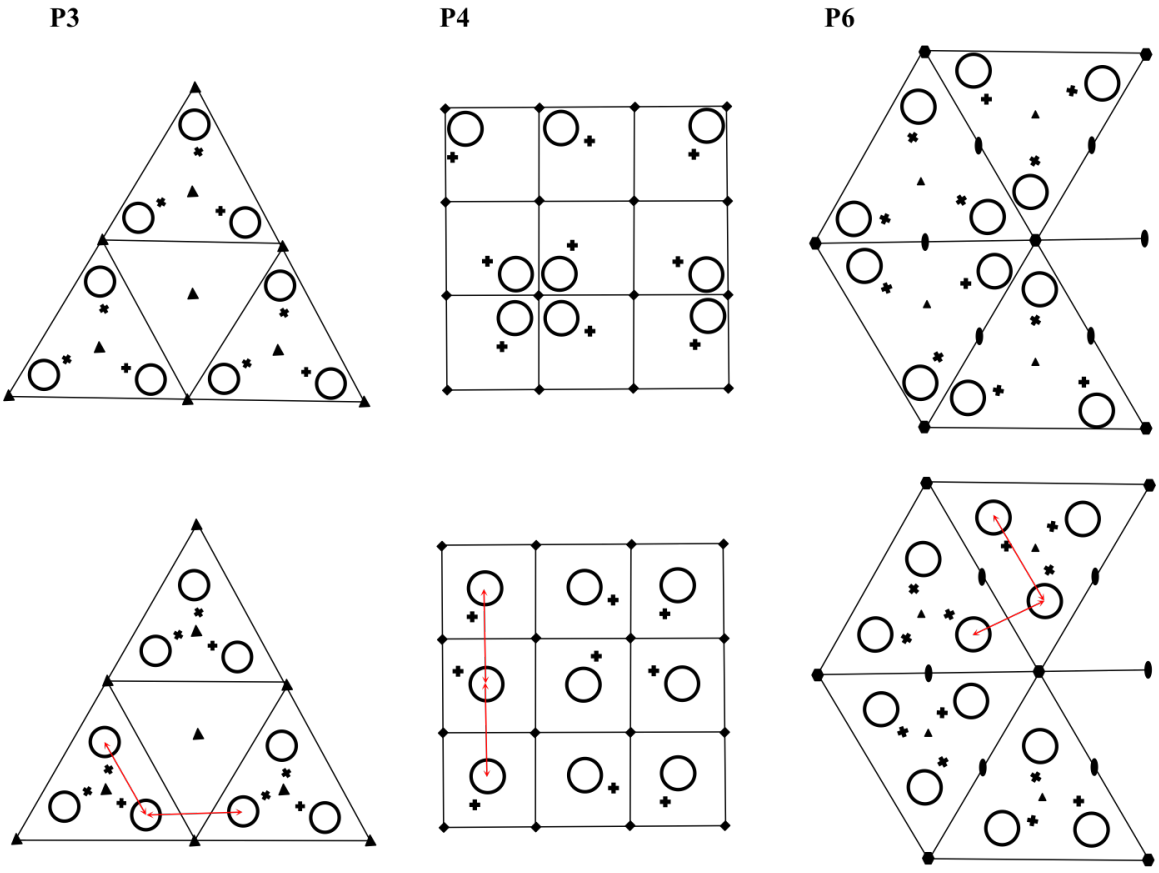



**Figure 4-12.** Depiction of spacing in various layered space groups (P3, P4, and P6). Potential orientations of a cylindrical asymmetric unit (circles) in each requires specification of rotation about the cylindrical axis (denoted by each plus), the unit cell spacing, and translation in the *a*-*b* plane (top). By constraining the distance between asymmetric unit centers (bottom, red lines), the search space can be reduced to exclude the two degrees of freedom associated with translation in the *a*-*b* plane.

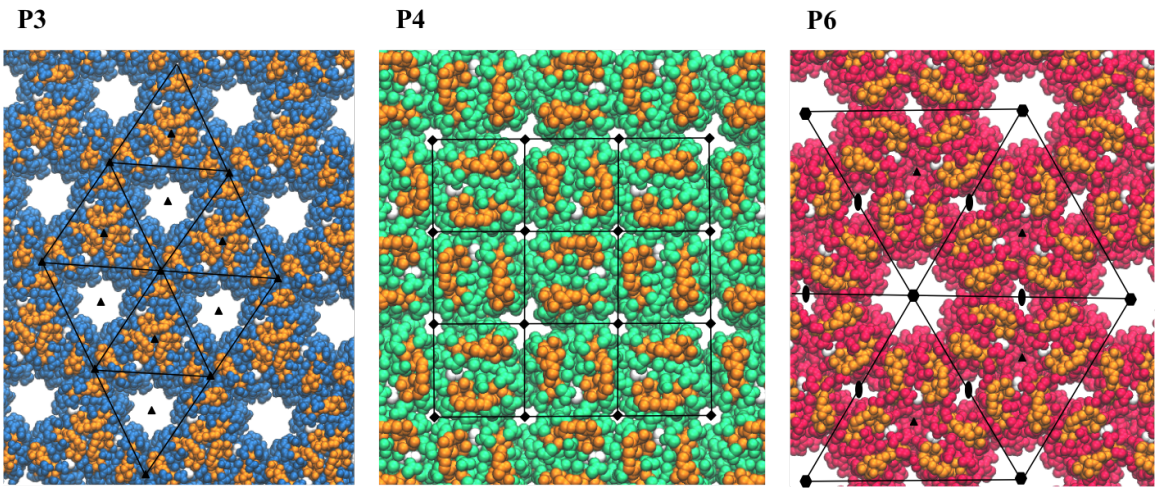



**Figure 4-13.** Overlay of constrained spacing in Figure 4-12 onto sample renderings of coiled-coil constructs encapsulating the PZnPI cofactor (orange). The P3 (blue), P4 (green), and P6 (red) space groups are represented. Ultimately the P6 space group was selected for full sequence design.

six-fold axis $r_{\text{offset}}$, the rotation away from the two-fold axis $\alpha$, and the rotation about the bundle axis $\varphi$.

As described above, we initially restrict some of the P6 search space to compensate for costly lattice calculations. Constraining the equidistant bundle spacing requires fixing $\alpha = 30\,^{\circ}$ and defining the translation $r_{\text{offset}}$ as

**Eq. 4-12**

$$r_{\text{offset}} = \frac{a}{\sqrt{3}+1}$$

such that variation of the lattice relies only upon the unit cell dimension $a$ and the superhelical rotation about the bundle's coiled-coil axis $\varphi$ (Figure 4-14). For all lattice calculations, mean lattice association energies are obtained, which are simply the difference between the mean lattice energy and the mean energy of the asymmetric unit (the full single chain protein construct) given as

**Eq. 4-13**

$$\langle E_{association} \rangle = \langle E_{lattice} \rangle - \langle E \rangle = \frac{1}{2} \sum_{i} \sum_{m}^{M} \gamma_{ii}^{0m} w_i + \frac{1}{2} \sum_{ij} \sum_{m}^{M} \gamma_{ij}^{0m} w_i w_j$$

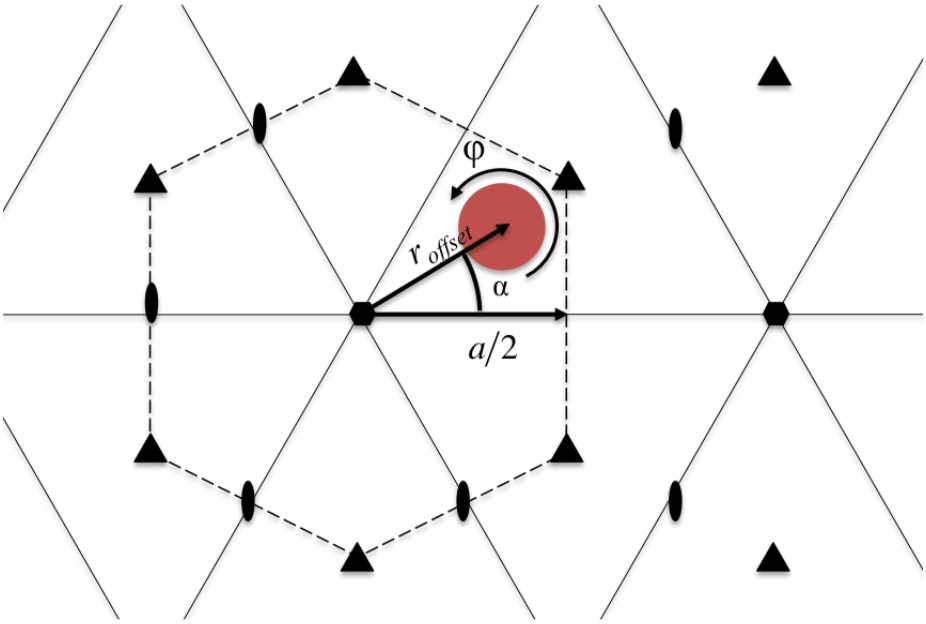



**Figure 4-14.** Illustration of translocation of tetramer bundle away from the P6 6-fold axis, given as $r_{\text{offset}}$. The unit cell length, $a$, is highlighted as the distance between 6-fold axes. The rotation away from the two-fold axis is given as $\alpha$. The angle $\varphi$ denotes the rotation about the asymmetric unit's axis$r_{\text{offset}}$.

To further reduce the computational cost, we assume that the unit cell parameter $c$ (z-component of the unit cell) can be optimized independent of the remaining in-plane parameters. Here, optimizing sequence design along the $c$ separation component can identify optimal inter bundle interactions between neighboring loop segments. This is achieved by enforcing a large unit cell size in the $a-b$ plane, $a$=$b$=100Å, such that symmetry interactions only occur with neighbors above or below the asymmetric unit. For both $CC_1$ and $CC_2$ structures, an optimal spacing was obtained across 36Å $< c <$ 46Å at 0.1Å increments (Figure 4-15). Each calculation allowed the ensemble specified for interior positions as described before (placement of the HIS-PZnPI super rotamer, specific hydrogen bonding residue, and all hydrophobics A,V,I,L,M,F,W at the remainder of the core positions), ALA at the remaining helical positions, and sequence variation at all positions in each of the three loops. Loops positions were allowed to be all amino acids except histidine and cysteine. The lowest lattice association energy corresponded to values of $c$ = 41.1Å and 42.24Å for $CC_1$ and $CC_2$ respectively.

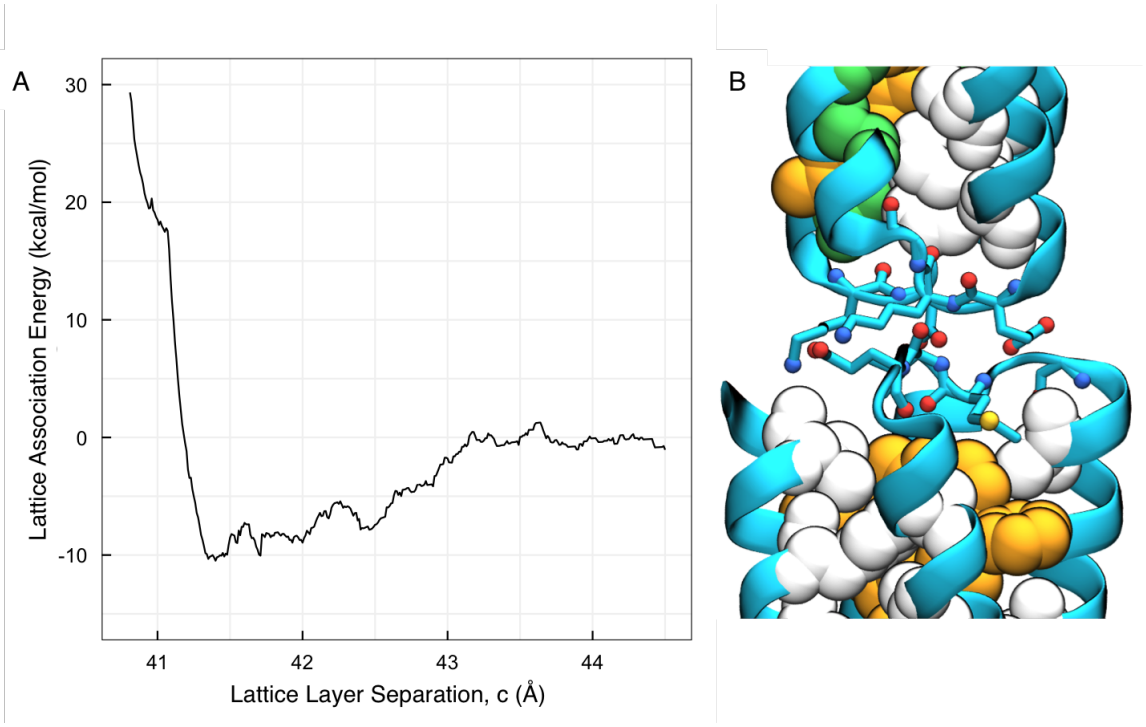



**Figure 4-15.** (A) Lattice Association Energy between protein bundles at the z-interface, as a function of the lattice layer separation parameter $c$ for the $CC_1$ structure. (B) Rendering of the lattice association energetic minimum along $c$ for $CC_1$. Core residues are depicted as spheres (PZnPI: orange, hydrophobics: white) and loop residues are rendered as ball and stick by atom type (N: blue, O: red). This particular interface highlights close packing glycines, as well as complimentary electrostatic interactions between charged pairs.

An initial scan over the simplified lattice space with a fully alanine exterior expedites identifying low energy ALA-ALA lattice contacts. For each calculation, interior positions were again typed as previously described (placement of the HIS-PZnPI super rotamer, specific hydrogen bonding residue, and all hydrophobics A,V,I,L,M,F,W at the remainder of the core positions), exterior positions on helices typed as alanine, and loop positions typed as glycine. While allowing glycine at the flexible loop regions, we exclude it from the helical domains to discourage potential helix breakers. Because the search is efficient, we are able to scan over an extensive landscape quickly, here over $a$ (57 Å$\leq a \leq$ 70 Å, incremented at $\Delta a = 0.1$ Å) and $\varphi$ ($0^\circ$ Å$\leq \varphi \leq 360^\circ$ Å, incremented at $\Delta\varphi = 2^\circ$ Å). Each calculation solves for the site-specific amino acid probabilities as those that minimize the sequence free energy at $\beta$=0.5. The mean field lattice association energy landscape (**Eq. 4-13**) for $CC_2$ is presented in Figure 4-16.

While the landscapes for $CC_1$ (not shown) and $CC_2$ (Figure 4-16) have multiple local minima, the global maximum is selected for the remaining parameter annealing and sequence design; it is entirely reasonable to repeat the following steps for any of the other identified local wells, we initially make this choice to winnow structural possibilities. To relax the severe constraints imposed on inter-bundle spacing posed in **Eq. 4-12**, the structures first undergo a MC annealing with the alanine exterior to further promote well packed bundles within the lattice. Lattice parameter update criteria are listed in Table 4-5. This trajectory was run for $N$ = 3,000 steps, at the cooling decay constant specified by **Eq. 4-8** and MC temperatures $\beta_{0,MC} = 0.5$ to $\beta_{f,MC} = 2.0$. The selection criteria uses the difference in mean field lattice energy values, $f(\mathbf{w}) = \langle E_{lattice} \rangle$.

Upon annealing, the lattices were significantly more compacted structures with clear ALA-ALA contacts. A preliminary full design calculation (allowing all amino acids except GLY, HIS, PRO, and CYS at the exterior) on the lowest energy structures possessed both AXXXA and Small-XXX-Small motifs to validate this. Trials of further annealing this lattice structure in the context of full design led to larger spacing between bundles with more favorable

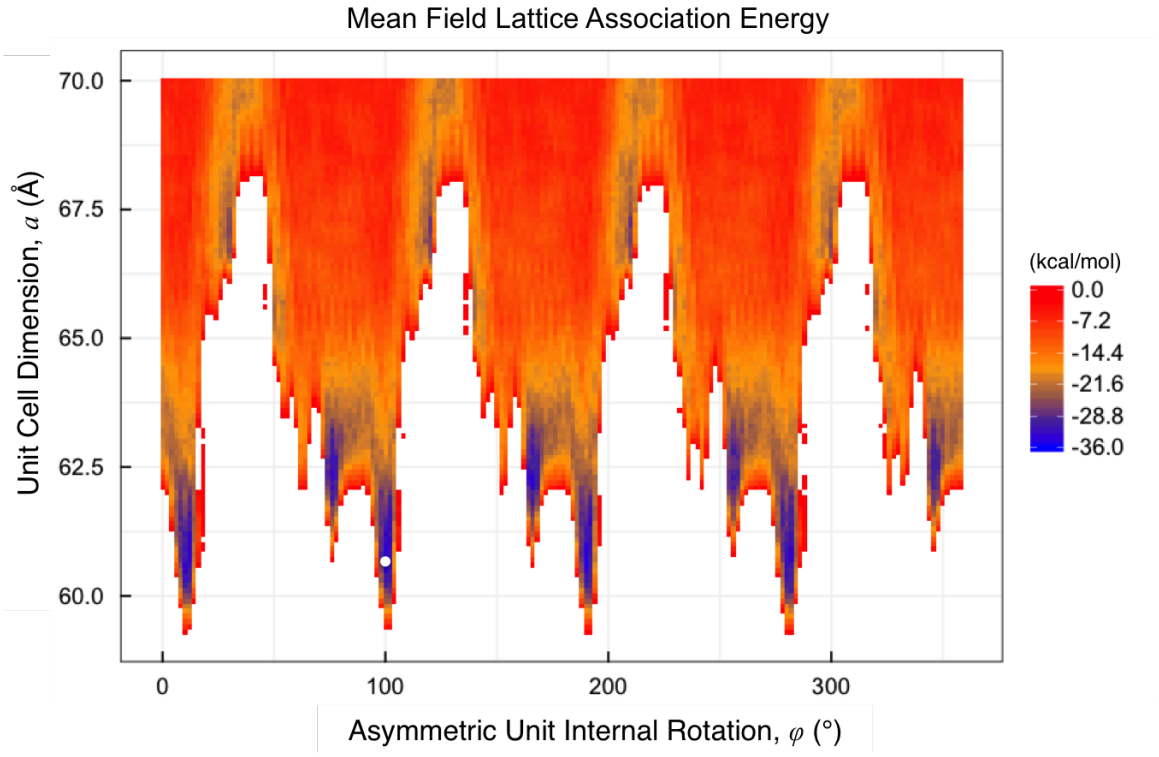



**Figure 4-16.** Mean Field Lattice Association Energy of the $CC_2$ single chain construct with all alanine exterior in the P6 space group, as given by the VERGIL package at $\beta$=0.5. The global minimum is denoted with a white circle at $a = 60.7$Å, $\varphi = 100.0\,^\circ$. White space denotes crystal configurations with association energies above 0.0 kcal/mol.

| **Parameter** | $x_{min}$ | $x_{max}$ | $\delta x$ |
|---|---|---|---|
| $a$ | 55 Å | 70 Å | 0.5 Å |
| $\varphi$ | -180 ° | 180 ° | 5 ° |
| $r_{\text{offset}}$ | 10 Å | 30 Å | 0.5 Å |
| $\alpha$ | -180 ° | 180 ° | 5 ° |

**Table 4-5.** Monte Carlo lattice parameter update criteria, including value boundaries and maximal change per MC step. The parameters listed are the in-plane lattice dimension $a$, the rotation about the bundle axis $\varphi$, the radial offset from the six-fold axis $r_{\text{offset}}$, and the rotation away from the two-fold axis $\alpha$. The axial lattice dimension $c$ was held fixed as per values specified in the text.

interactions between larger residues (e.g., ARG or GLN). While such structures may have favorable energetics, they largely remove all small residue contacts in previously identified structures. Instead of modifying the parameters further, we chose to retain the final low energy lattice structure obtained from the ALA-exterior MC trajectory. The $CC_2$ structure was selected to identify a full sequence, with final lattice parameters of $a = 59.34$ Å, $\varphi = 101.16\,^\circ$, $r_{\text{offset}} = 22.29$ Å, $\alpha = 29.90\,^\circ$, $c = 42.24$ Å.

A final full sequence calculation was performed using the VERGIL package at $\beta$=1.69 for the $CC_2$ structure. In addition to the HIS-PZnPI super rotamer, hydrogen bonding serine, and hydrophobics (A,V,I,L,M,F,W) at the core, exterior helical positions were typed with all amino acids except GLY, HIS, PRO, and CYS and loop positions typed with GLY in addition to the exterior set. Again, to promote crystallization we enforce a mean net neutrality constraint,

**Eq. 4-14**
$$\langle \text{Net Charge} \rangle = \sum_i q_i w_i = 0$$

a mean molar extinction coefficient constraint to assure the ability to monitor the protein concentration throughout experimental trials to a value of at least one TRP,

**Eq. 4-15**
$$\langle \epsilon_{\text{Protein}} \rangle = \sum_i \epsilon_{\text{Ext},i} w_i \geq 5690.0 \text{ M}^{-1}\text{cm}^{-1}$$

and a limit on the total number of TRPs in the sequence to be no more than 6.

**Eq. 4-16**
$$\langle N_{\text{TRP}} \rangle = \sum_i w_{n,TRP,c} \leq 6$$

Trials without the constraint on the tryptophan composition often left to a prevalence of TRP in the sequence at solvent channel positions. To address unusual solvent accessible placement of TRP, several trial designs were conducted with the environmental potential constraint; however, due to the irregular shape of the bundle, the constraint was unable to capture similar patterning to globular water soluble proteins such at those for which

the environmental potential was parameterized. Furthermore, TRP rich sequences offer clustered TRP around the PZn porphyrin (electron donor) which may act as potential electron donors during experiments. The simplest solution is to simply restrict the number of TRP residues in the sequence. The number 6 was arrived at by counting high probability TRP positions from the unconstrained calculation. The full calculation solved for site-specific amino acid probabilities as those that minimize the sequence free energy for the mean field lattice energy at $\beta$=1.69. The final sequence was selected from the most probable entropically weighted amino acid at each position, denoted as **SCPZnPI-2A**.

The final sequence for **SCPZnPI-2A** is rendered in Figure 4-17. The sequence features a neatly packed hydrophobic core with aromatic residues flanking the diimide of the PZnPI cofactor. While the larger porphyrin ring occupies a large fraction of the core in the upper bundle, the octyl tail is positioned to have a clear channel among the dense hydrophobic block in the lower bundle. A sense of packing within the P6 space group is given in Figure 4-18 which indicates the orientation of the bundle about the six-fold axis (solvent channel). This face of the bundle bears a stripe of ionizable residues while the other helices and corresponding interfaces possess small residues to encourage inter bundle packing. A TRP at the top of the bundle is positioned to pack into the hydrophoic core of what would be the exposed hydrophobic underside of neighboring bundles to favor a layered crystalline configuration. Full diagrams of the lattice arrangement are given in Figures 4-19 (helical backbone and PZnPI cofactor), 4-19 (colored by reside type), and 4-19 (colored by residue).

In addition to this sequence, the final calculation was revisited with further constraints to suggest a more hydrophobic variant. Inspection of the RuPZn single chain protein[13] suggests a hydrophobic interface at the closer inter-helix $e/g$ positions. As such, hydrophobic residues (A,V,I,L,M,F,W) were typed at the corresponding interface positions: A7, A12, A14, A21, B3, B10, B17, B24, C7, C12, C14, C21, D3, D10, D17, and D24. The design calculations were altered to optimize the sequence free energy using the asymmetric mean field energy (instead of the mean lattice energy) at $\beta$=0.5 to focus on stabilizing the indi-

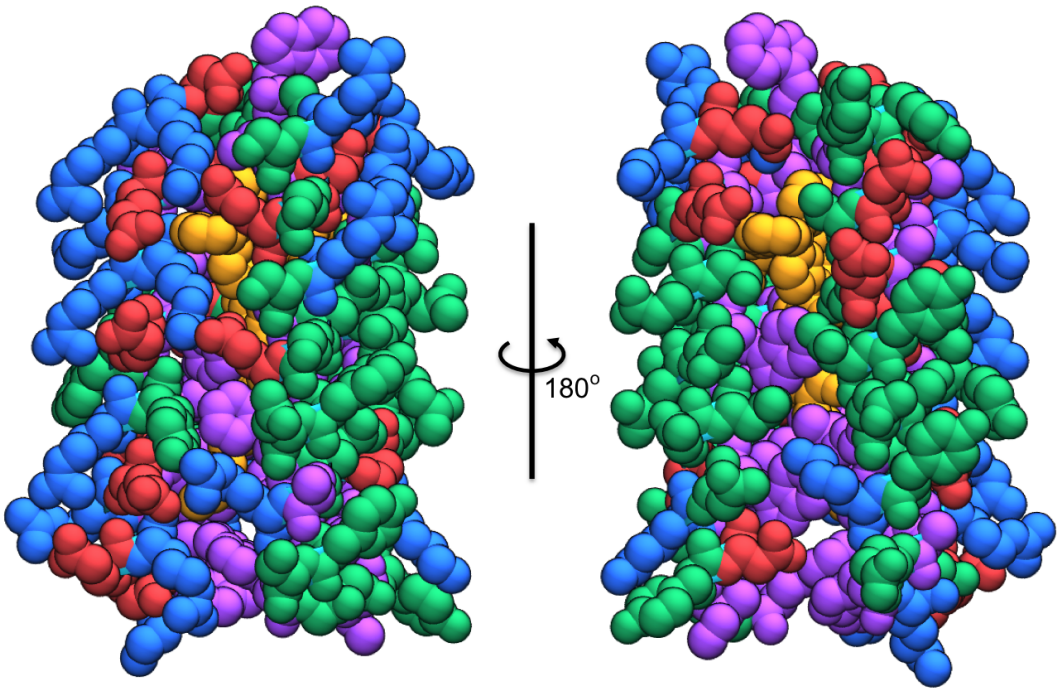



**Figure 4-17.** Rendering of **SCPZnPI-2A**. Coloring scheme indicates atoms in positively charged residues (blue), negatively charged residues (red), hydrophobic residues (purple), hydrophilic residues (green), the PZnPI cofactor (orange).

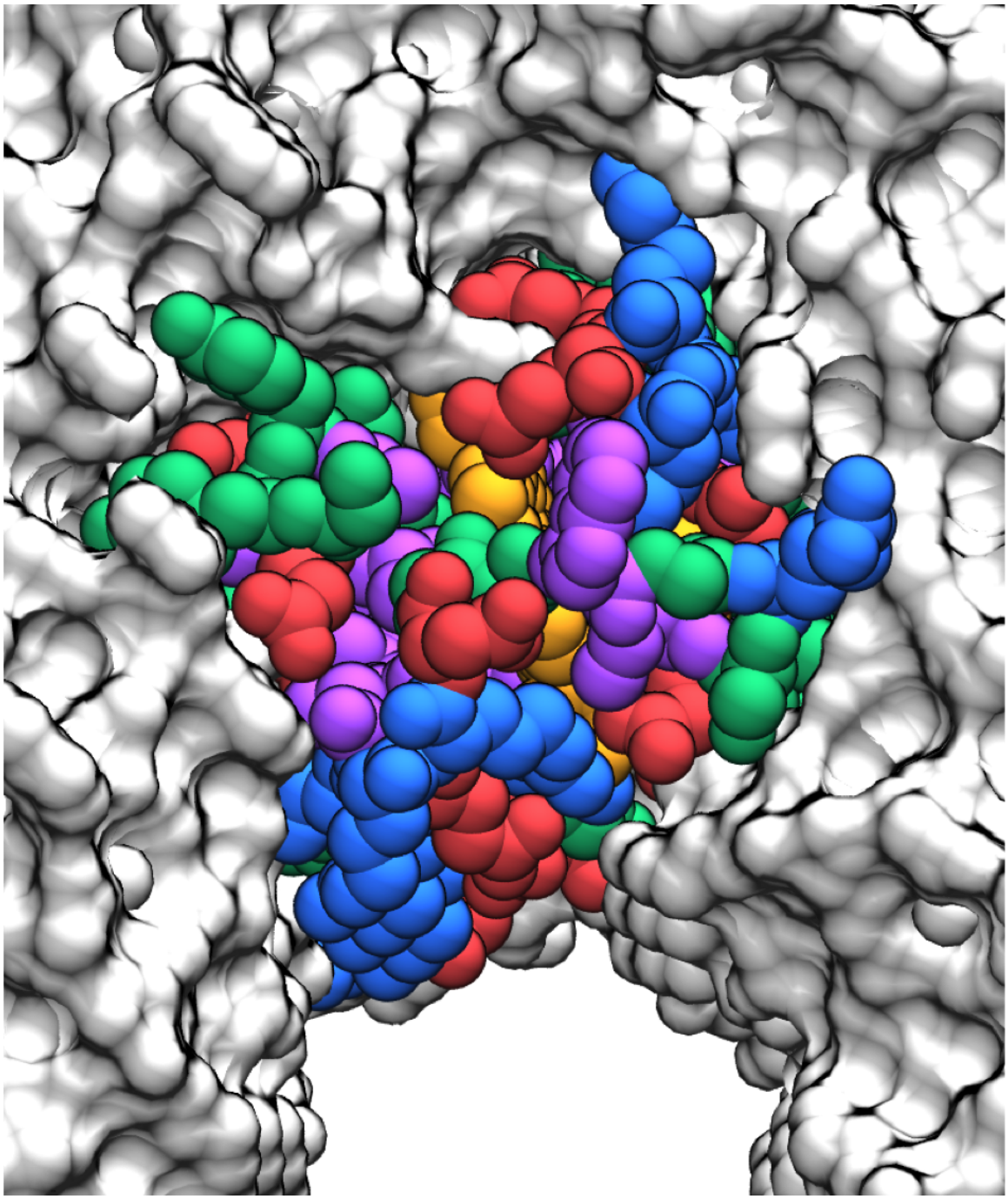


**Figure 4-18.** Rendering of **SCPZnPI-2A** as packed in the P6 space group. All neighboring units are rendered as a uniform surface, to highlight how the designed structure packs into the crystal contextually. Coloring scheme indicates atoms in positively charged residues (blue), negatively charged residues (red), hydrophobic residues (purple), hydrophilic residues (green), the PZnPI cofactor (orange).

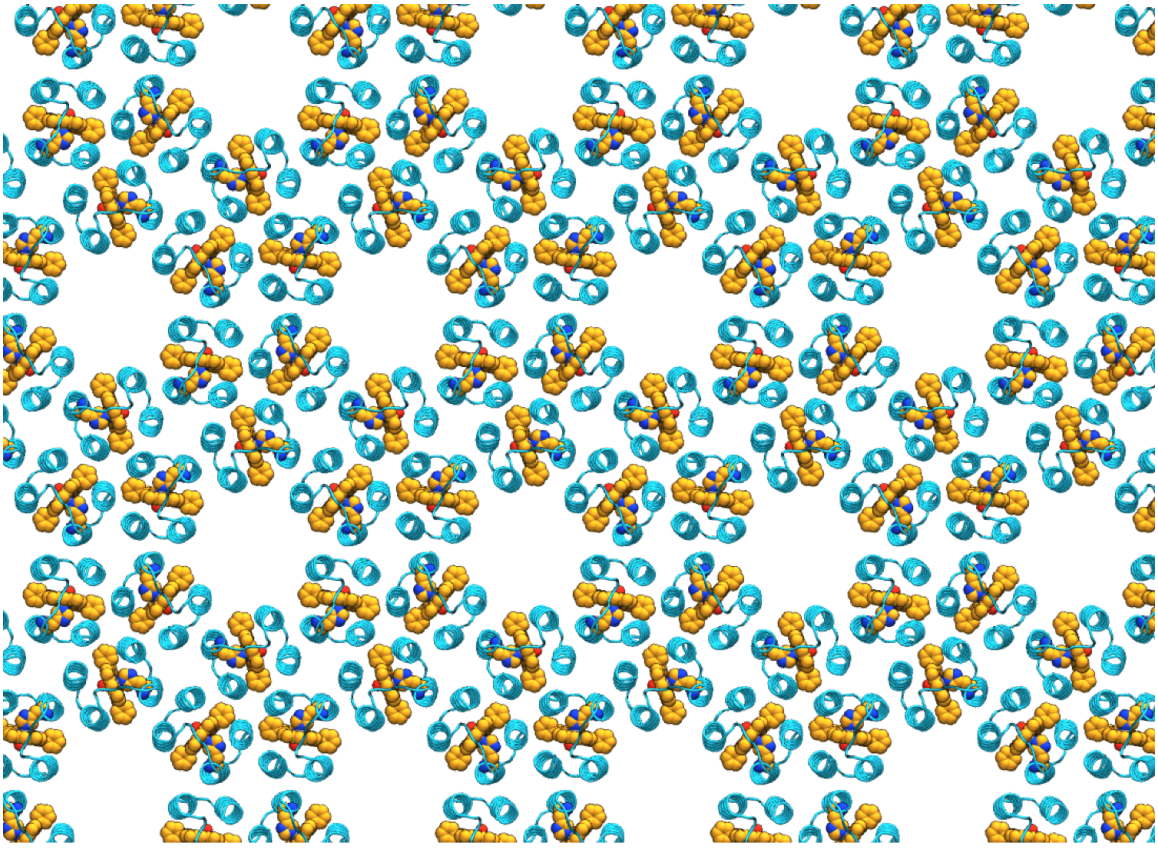


**Figure 4-19.** Rendering of the extended **SCPZnPI-2A** lattice in the P6 space group. Only the protein backbone (cyan), HIS, and PZnPI cofactor (orange) are rendered for clarity. Heavy atoms are colored as red (oxygen), blue (nitrogen), and gray (zinc).

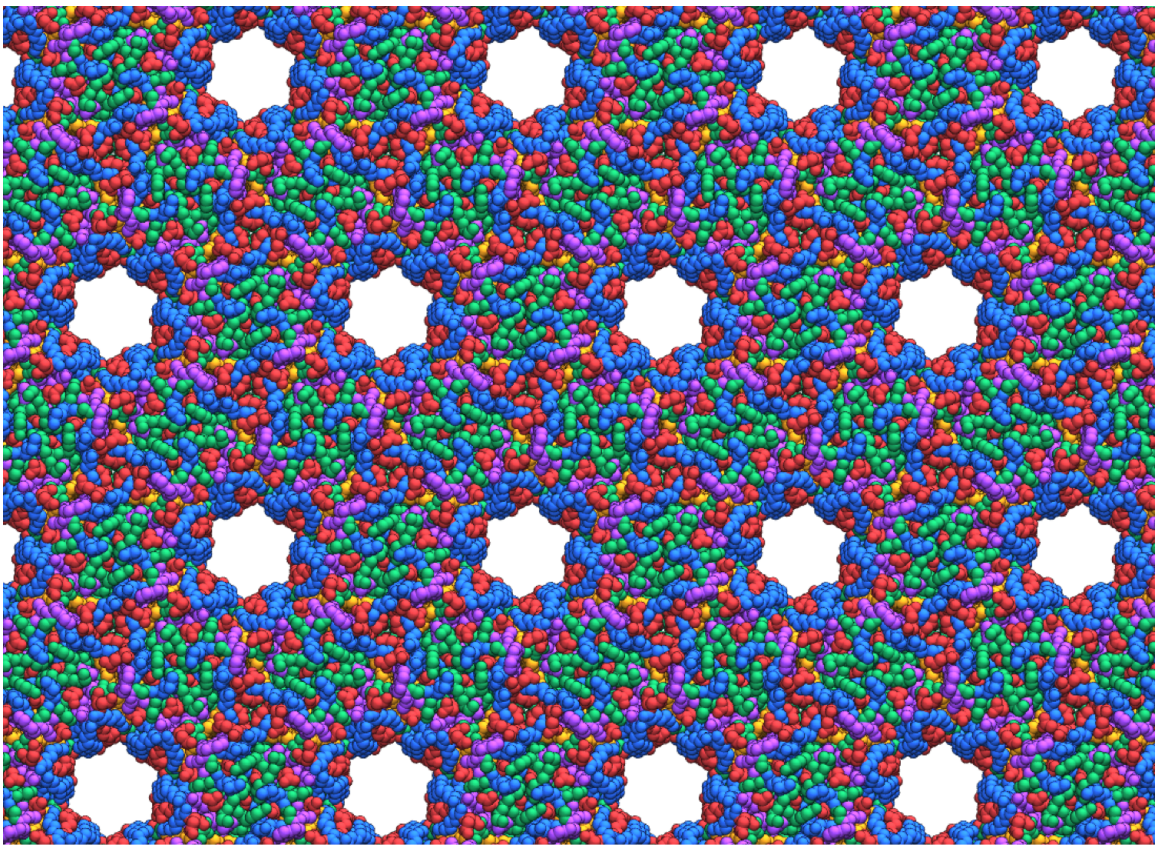


**Figure 4-20.** Rendering of the extended **SCPZnPI-2A** lattice in the P6 space group. Coloring scheme indicates carbon atoms in positively charged residues (light blue), hydrophobic residues (light purple), hydrophilic residues (light green), the PZnPI cofactor (orange). Heavier atoms are colored as red (oxygen), blue (nitrogen), purple (sulfur), and gray (zinc).

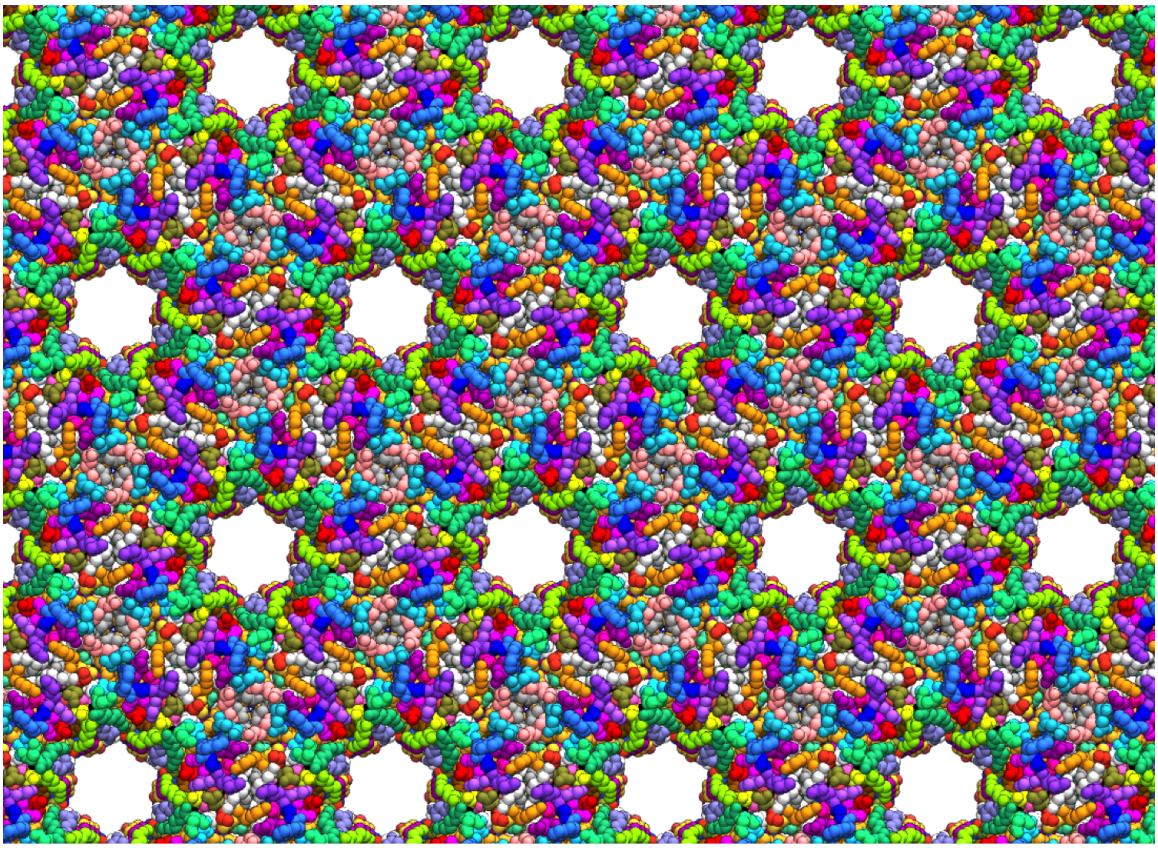


**Figure 4-21.** Rendering of the extended **SCPZnPI-2A** lattice in the P6 space group. Each residue is rendering a distinct color to highlight sequence diversity.

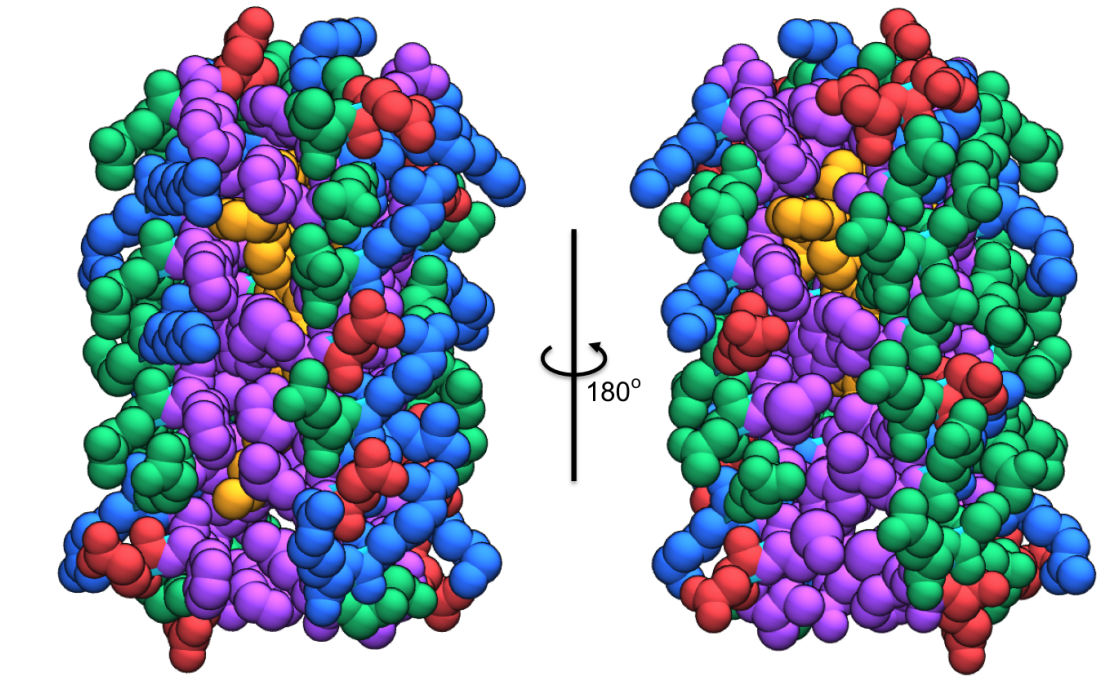



**Figure 4-22.** Rendering of **SCPZnPI-2B**. Coloring scheme indicates atoms in positively charged residues (blue), negatively charged residues (red), hydrophobic residues (purple), hydrophilic residues (green), the PZnPI cofactor (orange).

vidual protein over a full assembly. The final sequence was selected by simply choosing the most probable amino acid at each position from the aggregated type probabilities, denoted as **SCPZnPI-2B**. A rendering of **SCPZnPI-2B** is given in Figure 4-22. The sequence features a notably richer hydrophobic region at the wider interfaces, which should dehydrate the cofactor further than the first sequence.

Four sequences were submitted for expression - both the original **SCPZnPI-2A**, the redesigned variant **SCPZnPI-2B**, and then each of these sequences with the hydrogen bonding serine at residue position 11 mutated to an alanine.

## 4.7. Conclusion

The methods detailed here outline a strategy for designing fully *de novo* proteins, including the ability to precisely identify ligand binding motifs, construct energetically favorable he-

lical bundle structures connected by novel loops, and determine compatible sequences. The technique has been shown to be easily extendable to a lattice context, influencing sequence design to favor protein-protein interactions in a crystal. For the sequences identified to bind and encapsulate the PZnPI cofactor, experimental trials are underway to assess protein secondary structure and ability to bind the PZnPI cofactor. Additional studies are aimed at studying electron transfer properties of these constructs, as well as crystallization trials to determine the protein/cofactor assembly structure.

Moreover, this work adequately sets up computational protein design techniques to handle a variety of nuanced cofactor binding schemes. One can imagine designing a hydrophobic core purely based upon shape complementarity (foregoing histidine ligation), or the design of a protein construct to selectively bind two or more distinct cofactors with defined locations in the protein. Where this work seeks control over modulation of the microenvironment of the sequestered cofactor, the design of a single protein which incorporates cofactors at controlled distances would be able to probe how physical separation in the core affects charge transfer dynamics.[30]. Computational protein design affords the ability to probe such questions and explore novel bio-assemblies.

# 5 Understanding the Helical Wrapping of Semiconducting Polymers Wrapped about Carbon Nanotubes *

Shape-persistent conjugated polymers possess unique optical, electronic, and structural properties[225] that yield a wide range of applications, including light emitting diodes, organic molecular transistors, and biochemical sensors[226–228]. Poly(*p*-arylene-ethynylene) (PAE)-based polymers have been widely studied and consist of aromatic units bridged by acetylene (ethyne) units. The ostensibly linear backbone of these molecules is conducive to extended electronic conjugation and facilitates the control of their structures[229,230]. As a result, there has been much interest in elaborating derivatives of PAE as components in molecular electronics and sensors[118,229], as their optical and electronic properties have been characterized in a variety of solvents[230–232]. The conformational data obtained for these molecules is typically consistent with rigid-rod type molecules having long persistence lengths[233–235]. By leveraging their linear, aromatic-ring-based structures, appropriately functionalized PAE polymers provide vehicles for both improving dispersion of single wall carbon nanotubes (SWNTs) in a variety of solvents (including water) and functionalizing such nanotubes noncovalently, so as to not adversely perturb the tube's electronic properties[101,102,236]. Somewhat surprisingly, PAE polymers have been observed to wrap in a helical manner about SWNTs[102].

Previous studies of poly[*p*-{2,5-bis(3-propoxysulfonicacidsodiumsalt)}phenylene]ethynylene (PPES) have indicated the polymer's ability to solublize SWNTs with a constant helical morphology; transmission electron micrographs and atomic force microscopy suggested a

*Adapted from Christopher D. Von Bargen, Christopher M. MacDermaid, One-Sun Lee, Pravas Deria, Michael J. Therien, and Jeffery G. Saven. "The Role of Ionic Side Chains in the Helical Wrapping of Phenylene Ethynylene Polymers about Single-Walled Carbon Nanotubes." *J. Phys. Chem. B.*, 2013, 117 (42).

helical superstructure of pitch 13 ± 2 nm. All-atom molecular dynamics (MD) simulations of a solvated PPES 20-mer and (10,0) SWNT complex in aqueous media exhibited wrapping of the PPES polymer about the nanotube to form a helix of pitch 14 ± 1 nm[102]. The following chapter employs a quantitative, molecular-level insight into the helical wrapping of SWNTs by the PPES polymer. MD simulations are used to obtain potentials of mean force as functions of the polymer end-to-end displacement, which is a monotonic function of the helical pitch. The simulations also provide a molecular perspective on the helical PPES-SWNT assembly, including the roles played by aqueous solvent and the propoxysulfonate side chains in determining conformational properties of the polymer-SWNT helical superstructure.

## 5.1. Introduction

SWNTs possess optical, electrical, structural, and tensile properties[103–105,237,238] that make them promising candidates for a variety of applications, including energy storage[239], electronic devices[240], and sensors[228]. These applications are limited, however, by difficulties associated with processing and dispersing SWNTs in solvent systems. Carbon nanotubes are highly insoluble in both organic solvents and in aqueous solutions due to strong intertube van der Waals interactions[241]. While solubilization of SWNTs can also be achieved via covalent modification of the nanotube,[242,243] doing so affects SWNT electronic structure, introducing both mechanical defects and disrupting critical semiconducting and conducting properties[244–247]. As such, a wide range of surfactants, small molecules, and polymers have been identified as noncovalent dispersion agents[248–251]. Many of these noncovalent methods, which utilize ultrasonication to drive the solubilization in the presence of surfactants and polymers[252,253], yield systems that have no apparent regular structure when associated with the SWNT. For example, sodium dodecyl sulfate (SDS) non-selectively forms columnar micelles about SWNTs and is limited to water as the solvent[252].

A variety of flexible polymers are capable of wrapping SWNTs, facilitating nanotube dispersion, but these systems generally provide ill-defined assemblies[254–257]. A wide assortment

of polymers are known to wrap SWNTs in aqueous media: polystyrene sulfonate (PSS)[254], polyvinyl pyrrolidone (including poly acrylic and maleic acid-containing copolymers),[254] $\alpha$-helical amphiphilic peptides[258,259], proteins[260], DNA[261–265], gum arabic (polysaccharide)[266], sodium carboxymethylcellulose (Na-CMC)[267], alginic acid[268], and $\beta$-1,3-glucans[269]. In addition to the poly(aryleneethynylene)s, the $\beta$-1,3-glucans[269], $\alpha$-helical peptides[259] and specific sequences of DNA[264] are known to wrap individualized SWNTs with a regular helical periodicity. However, the biologically derived polymers are unsuitable for providing well-defined polymer-SWNT superstructures in organic solvents. Nonnatural polymers can also be similarly problematic; upon transfer from an aqueous to an organic phase, SWNTs wrapped with an amphiphilic polymer undergo polymer dewrapping, resulting in SWNT precipitation[254,269]. Hydrophobic conjugated polymers based on PmPV[257], polythiophene[270], polyfluorene[271], and poly(*p*-phenylene)ethynylene[248] frameworks can disperse SWNTs in a variety of nonaqueous solvents. Indeed, changing organic solvent has been used to disperse and release SWNTs with poly(*m*-phenylene)ethynylene (PPE) polymers[272]. These polymer-SWNT systems often display microscopy consistent with indistinct rodlike structures, and the polymer/SWNT molar ratios are consistent with multilayer polymer aggregates associated with the nanotubes. For example, flexible polyvinyl pyrrolidone can wrap SWNTs with very short pitch lengths but does not yield a regular structure for the polymer monolayer adsorbed on the SWNT[254]. Relative to the vast array of SWNT solubilization agents that have been utilized, highly charged aryleneethynylene polymers provide unique combinations of structure and utility[101–105].

PAE-based polymers comprising para-connected monomer units are ostensibly linear polymers but have been observed to acquire both linear and helical structures when adhered to SWNTs. A PPE polymer with neutral side chains appeared to solubilize SWNTs based on a parallel interaction mode with no evidence for helical wrapping[248]. For the solubilization of a boron nitride nanotube (BNNT) by a PPE polymer, the polymer backbone appeared to adhere to the BNNT surface in linear fashion[273]. Poly[*p*-{2,5-bis(3-propoxysulfonicacidsodiumsalt)}phenylene]ethynylene (PPES) and the related poly[2,6-{1,5-bis(3-

propoxysulfonicacidsodiumsalt)}naphthylene]ethynylene (PNES), each forms a distinct self-assembled polymer-SWNT helical superstructure[101,102], in which a polymer monolayer wraps the nanotube surface to form a well-defined helix. These highly charged aryleneethynylene polymers exfoliate individual SWNTs in aqueous media, and they also afford a means to solubilize SWNTs in a wide range of organic solvents, while preserving the polymer-SWNT helical morphology[101]. For PPES, transmission electron micrographs and atomic force microscopy (AFM) observations were consistent, suggesting a single-stranded, regular helical super-structure of pitch 13 $\pm$ 2 nm wrapped about the nanotubes. The distinctive features of SWNTs wrapped by aryleneethynylene polymers such as PPES and PNES underscore the need to understand at a molecular level the observed regular, helically wrapped polymer-SWNT assemblies in aqueous environments. Of particular interest are: (i) the critical molecular interactions present in the highly charged aryleneethynylene polymer-SWNT assembly, (ii) the nm-scale geometric features of the polymer-SWNT superstructure including the helical pitch, and (iii) the relative stabilities of various helical superstructures of a given polymer-SWNT system.

Though electron microscopy, AFM, and spectroscopic methods can be highly informative, atomistically detailed information on polymer-SWNT assemblies is generally difficult to achieve, due to structural heterogeneity and the limited resolution of these methods. Alternatively, molecular simulations have proved insightful and provide vehicles to investigate the structure, fluctuations, and energetics of these assemblies. Simulation-based studies have been extensively applied to a wide variety of nanotube/polymer systems, including SWNTs complexed with: DNA[107–112]; amylose[113]; polythiophene;[114,115] polymers that are rich in aromatic groups such as polystyrene (PS), poly(phenylacetylene) (PPA), poly(p-phenylenevinylene) (PPV), and poly(m-phenylenevinylene-co-2,5-dioctyloxy-p-phenylenevinylene) (PmPV)[114,116–118]; polyethylene[119,120]; poly-alkylsilanes[121]; poly-phenyleneethynylenes[102]; poly(N-decyl-2,7-carbazole) and poly(9,9-dialkyl-2,7-fluorene)[122]; alginic acid[126,127]; poly[9,9-dioctylfluorenyl-2,7-diyl] (PFO)[123]; and proteins[124,125]. Simulations and theoretical calculations have provided detailed information on monomer interactions

with the SWNT, including the preference of aromatic moieties to associate with the nanotube surface[114,116,124,125,258,274] and conformational backbone energetics such as the balance between torsional and electrostatic energies in the phosphate linkages of DNA[108,110]. In addition, such simulations can assist in understanding the wide range of structures possible when polymers adhere to SWNTs and the relative free energies of different superstructures[126,127]. For PPES, the superstructures obtained from molecular dynamics (MD) simulations of a PPES 20-mer and (10,0) SWNT complex in aqueous media were in excellent agreement with experimental findings[102]; the simulations exhibited wrapping of the PPES polymer about the nanotube to form a helix of pitch $14 \pm 1$ nm[102].

Simulation data can provide a detailed molecular-level rendering of the polymer-nanotube complex, as well as the origins and free energetics of the helical wrapping. A molecular understanding of the determinants of such helical structures in terms of the chemical properties of the polymer and its solvent environment would be useful both in understanding why different morphologies are observed upon variation of solvent and chemical structure of the monomeric units and in engineering poly(aryleneethynylene)/SWNT systems having specific properties.

## 5.2. Simulation Tools

### 5.2.1. Molecular Dynamics Simulations

All simulations were performed using the molecular dynamics program NAMD2.7[275]. Orthorhombic periodic boundary conditions were applied in all three Cartesian dimensions, and the average dimensions were 55 Å x 55 Å x 285 Å. The equations of motion were integrated with a time step of 2 fs. Covalent bonds involving hydrogen atoms were constrained to their equilibrium length by means of the SHAKE/RATTLE algorithms[276,277]. Long-range electrostatic forces were evaluated by means of the particle-mesh Ewald (PME) approach[278] with a 1 Å mesh, and van der Waals interactions were truncated smoothly with a spherical cutoff of 12 Å. For solvated simulations, the TIP3P[279] water model was

used. The simulations in the aqueous phase included ∼23,000 TIP3P waters placed using the VMD module SOLVATE[280], yielding a total of ∼75,000 atoms including the PPES polymer, counter ions and the (10,0) nanotube. The aqueous system was ensured to be electrostatically neutral via the addition of 50 sodium and 8 chloride ions using the VMD module AUTOIONIZE,[280] consistent with a salt concentration (ionic strength) of 0.2 mol/L. Aqueous simulations were carried out in the isothermal-isobaric ensemble; temperature was maintained at 300K by employing Langevin dynamics with damping coefficient of 5 $ps^{-1}$, and the pressure was maintained at 1 bar employing the Langevin piston pressure control with an oscillation period of 100 fs and decay time of 50 fs[281]. For simulations without solvent, only the polymer (PPE or PPES with absent atom-centered partial charges) and SWNT were present, and temperature (300K) was maintained using the aforementioned Langevin dynamics damping coefficient. All initial configurations were minimized for 1,000 steps using conjugate gradient energy minimization prior to each simulation. Preparation, visualization, and analysis of structures and trajectories utilized the VMD package[280].

### 5.2.2. Molecular Models of Phenylene Ethynylene Polymers

Two polymers were considered in the simulations (Figure 5-1): poly[*p*-phenylene]ethynylene (**PPE**) and poly[*p*-2,5-bis(3-propoxysulfonicacidsodiumsalt)phenylene]ethynylene (**PPES**). The molecular potential parameters were taken from previous work[102,282–284]. For the aqueous simulations, each sulfonate group was fully ionized with net charge of −1. For simulations in the absence of solvent ("vacuum" simulations), all atom-centered partial charges were set to zero. For all simulations, polymers composed of 20 monomer units were considered ($n = 20$ in Figure 5-1). For both linear and helical initial configurations of each polymer, all phenyl rings were positioned in the same manner on each monomer; there were no "ring flips," and the monomers were translationally invariant along the contour of the polymer.

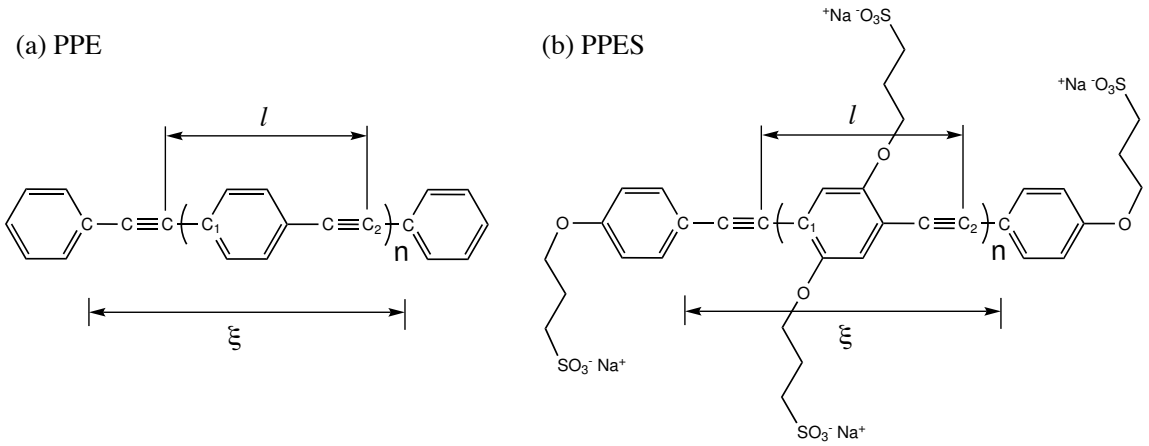



**Figure 5-1.** Chemical structures of the poly[*p*-phenylene]ethynylene polymers considered in the simulations. (i) Poly[*p*-phenylene]ethynylene featuring terminal phenyl units (PPE). (ii) Poly[*p*-2,5-bis(3-propoxysulfonicacidsodiumsalt)phenylene]ethynylene featuring terminal *p*-{4-(3-propoxysulfonicacidsodiumsalt)}phenyl units (PPES). For all simulations, $n = 20$, $l$ is the distance between equivalent carbons in adjacent monomer units, and $\xi$ is the difference in z-coordinates of the indicated carbon atoms.

### 5.2.3. Carbon Nanotube Model

An achiral semiconducting (10,0) SWNT consistent with previous solubilization studies of PPES was selected[102]. Coordinates for an ideal tube were generated using the VMD Nanotube Builder[285] with no additional chemical functionalization or geometric deformations at the end of the tube. No relaxation of nuclear coordinates of the tube were performed. Each atom of the SWNT was parametrized as an $sp^2$ carbon atom of the CHARMM force field[286] with zero partial atomic charge. The nanotube length in each simulation was 25.4 nm with an internuclear diameter of 0.793 nm. The nanotube length was chosen to be twice that of the fully extended, linear 20-mer of PPES, and at no point in any simulation did the polymer approach the nanotube ends. The coordinates of all atoms within each nanotube were fixed in all simulations, resulting in C-C bonds constrained to their equilibrium lengths of 1.42 Å. Nanotubes selected in the following simulations were chosen as achiral with comparable diameters to nanotube diameters in respective experiments, as given by

**Eq. 5-1**
$$d = \frac{a}{\pi}\sqrt{(n^2 + nm + m^2)}$$

for $a = 0.246$ nm.

### 5.2.4. Potential of Mean Force Calculations.

The potentials of mean force of the PPE and PPES polymers adsorbed onto a SWNT were calculated using the adaptive biasing force (ABF) method[135,287–289]. This technique estimates the average force $\langle F \rangle_\xi$ acting along a given reaction coordinate $\xi$ and applies a force that counteracts this average force, thus allowing the system to freely diffuse along the chosen coordinate.

**Eq. 5-2**
$$\frac{dA(\xi)}{d\xi} = -\langle F \rangle_\xi = \left\langle \frac{\partial U}{\partial \xi} \right\rangle_\xi$$

Here $dA(\xi)/d\xi$ is the average force that is applied along the coordinate $\xi$, and $U$ is the potential energy of the system. The average $\langle ... \rangle_\xi$ is over configurations having a particular value of the reaction coordinate $\xi$. This calculated average force is used to estimate the potential of mean force $\Delta A(\xi)$ relative to a fiducial reference value of the order parameter $\xi_o$.

**Eq. 5-3**
$$\Delta A(\xi) = -\int_{\xi_o}^{\xi} d\xi \ \langle F \rangle_\xi$$

The axis of cylindrical symmetry of the SWNT in each simulation was defined to be collinear with the Cartesian z-axis. The order parameter, $\xi$, associated with the linear-helical transition of the polymer was chosen as $z$, the difference in z-coordinates between the two $sp^2$ carbon atoms in the first and last monomers as denoted in Figure 5-1. With this choice, the expression for the average force is

**Eq. 5-4**
$$\frac{dA(\xi)}{d\xi} = \left\langle \frac{\partial U}{\partial z} \right\rangle_z$$

For the potential of mean force calculations, the order parameter was considered in the range $11.0 \leq \xi \leq 13.7$ nm, where $\xi$=13.7 nm corresponds to the fully extended, linear

20-mer. This range was divided into 15 windows each of 0.4 nm width, overlapping 0.2 nm on each side to improve sampling continuity. To confine sampling within each window, a harmonic potential was applied if the coordinate exceeded the window boundary; the potential had a force constant of 10 kcal/mol/Å and was centered at the window boundary. Each calculation sampled instantaneous force values for 20 ns, collected in bins 0.01 nm wide. To reduce possible non-equilibrium artifacts, 20,000 samples were accrued in each bin before introducing the biasing force (Equation Eq. 5-2) within each bin.

In simulations of PPES/SWNT in vacuum and in aqueous solution, the polymer was observed to remain adsorbed to the SWNT in a helical conformation, and thus for the potential of mean force calculations, a helical initial configuration was chosen to more rapidly sample "equilibrium" configurations for each value of $\xi$. For each of the 15 windows of $\xi$, the initial configuration of the polymer was selected as a helix having the value of $\xi$ at the center of the window. These initial structures were generated by aligning the polymer to an ideal helix having radius $r$ and pitch $p$. This helical contour $\vec{r}_h(t)$ can be defined parametrically in terms of the variable $t$; in Cartesian coordinates $\vec{r}_h(t) = (r\cos(2\pi t), r\sin(2\pi t), pt)$. For a given window, an ethynyl carbon in each monomer (see Figure 5-1) was positioned on a helical path, $\vec{r}_h(t)$, having radius $r = 7.35$ Å, which provides near van der Waals contact between polymer and nanotube carbon atoms (internuclear distance of 3.4 Å). The pitch $p$ was varied to specify a given value $\xi$. In constructing the helical conformation, rigid body motions were used to position adjacent monomers so that equivalent ethynyl carbons are at positions $\vec{r}_h(t)$ and $\vec{r}_h(t+\tau)$, respectively. The parametric one-monomer increment $\tau$ was determined by specifying that equivalent ethynyl carbons on adjacent monomers lie on the helical contour: $|\vec{r}_h(t+\tau) - \vec{r}_h(t)| = l$, where $l$ is the Euclidean distance between the equivalent ethynyl carbons (Figure 5-1). The phenyl rings of the monomers were then rotated about an axis collinear with the ethyne bridge such that the plane of the aromatic ring was perpendicular to the normal of the nearest point on the nanotube surface. In all cases, the helical axis is chosen as the z-axis (SWNT axis). These initial structures were minimized for 1,000 steps using the conjugate gradient energy minimization and then run

for 10,000 time steps (20 ps), prior to the 20 ns of conformational sampling.

### 5.3. Spontaneous Wrapping of a Single-Walled Carbon Nanotube

To explore the spontaneous wrapping of PPES about the carbon nanotube, unconstrained MD simulations of the PPES 20-mer (Figure 5-1) and a (10,0) SWNT were carried out in an aqueous environment. The polymer strand was initially set in a linear conformation parallel to the SWNT axis ($p = \infty$). The chain was positioned so that the center of mass of the phenyl subunit was 3.4 Å from the cylinder that contains the carbon nuclei of the SWNT, consistent with van der Waals contact between the phenyl backbone of the polymer and the nanotube surface. The simulation was carried out for 40 ns. This initial simulation employed a water solvated polymer-nanotube system with explicit counter ions. Within 15 ns, the solvated polymer had spontaneously wrapped about the SWNT forming a helical structure.

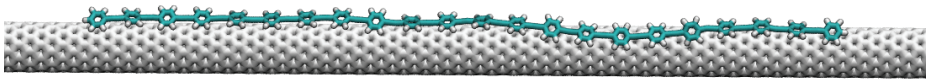
To quantify the polymeric superstructure, parameters specifying a helix centered about the longitudinal axis of the SWNT were calculated from sampled configurations. The pitch $p$ and radius $r$ of helical conformations were determined from the coordinates of the centroids of the phenyl rings of the PPES monomer units (centers of mass of the aromatic carbon atoms). These centroid coordinates were fit to a standard helical contour using a least squares method[282,290]. Configurations were sampled every 20 ps from the final 20 ns of simulation to estimate the pitch, and the uncertainties reported are plus/minus one standard deviation. For all sampled configurations, the root-mean-square deviations (RMSD) with respect to ideal helices were less than 2 Å. The average pitch of the helical polymeric structure was $p = 13 \pm 1$ nm, in good agreement with both that measured from transmission electron micrographs and previous simulation studies of the PPES polymer[102].

The molecular features of PPES that promote helical wrapping about the SWNT were further explored. The solvated PPES/SWNT system is complex, but via computer simulation, the structural and energetic properties of simplified systems, systems which may not

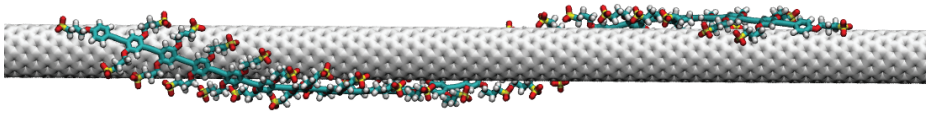
be experimentally realizable, can used to probe the roles of particular interactions and structural features in driving the helical superstructure formation. To this end, simulations of PPES/SWNT and PPE/SWNT systems were carried out in the absence of solvent, counter ions, and electrostatic interactions (here termed "vacuum" simulations) as a means to explore the intrinsic helical propensities of these polymer/SWNT systems. As in the solvated simulations, the initial conformation of the polymer in each case was a linear configuration. In vacuum, the PPES polymer wrapped the SWNT in approximately 3 ns, forming a persistent helical superstructure. A larger helical pitch and larger pitch fluctuations ($p = 16 \pm 4$ nm) were observed compared to that observed for the PPES/SWNT simulations carried out in solvent. The PPES helical structure obtained in this vacuum simulation is structurally distinct, however, from that observed in the solvated simulation with regard to the orientation of the side chains relative to the cylindrical axis of the carbon nanotube (Figure 5-4).

In simulating the PPE/SWNT system in vacuum, the impact of the monomer side chains was explored; compared to PPES, the polymer is simplified via the replacement of the propoxysulfonate side chains with hydrogen atoms (Figure 5-1). PPE maintains van der Waals contact with the nanotube surface but had no persistent global helical structure. The segments of the polymer intermittently acquire helical conformations locally (Figure 5-2). The polymer only transiently adopts overall helical structures and explores an ensemble of linear and helical conformations, including the initial collinear state. No persistent helical structure was observed.

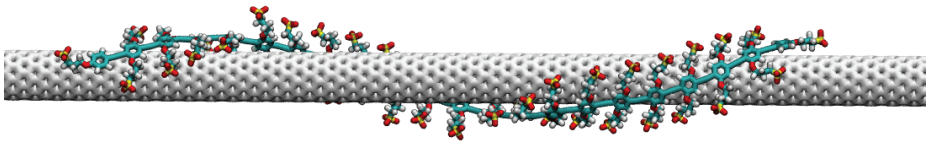
An evaluation of the direct polymer-SWNT interactions and adsorption of the polymer on the SWNT is presented in Figure 5-3. The interaction energy, $E_{inter}$ between the polymer and nanotube (nonbonded energy) per monomer unit for each system (Figure 5-3(a)) quantifies the energy of adsorption due to direct noncovalent interactions between the polymer and the SWNT. For each of the systems, $E_{inter}$ fluctuates about a value established within the first few nanoseconds of the trajectory. For the solvated PPES/SWNT system, $E_{inter}$



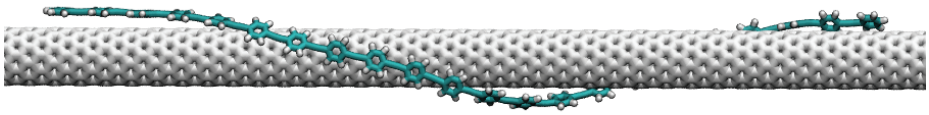
(i) PPE Backbone *in vacuo:* $\xi$ = 13.5, pitch > 90 nm



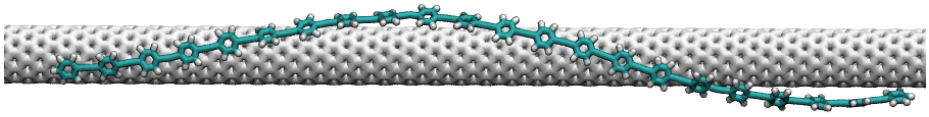
(ii) PPES *in vacuo*: $\xi$ = 13.1 nm, pitch = 15 nm, $\theta = +17^{o}$, $\phi = 43^{o}$



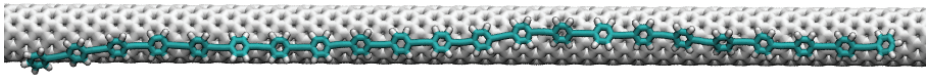
(iii) PPES solvated: $\xi$ = 12.8 nm, pitch = 13 nm, $\theta = -19^{o}$, $\phi = 79^{o}$



(iv) 27.8 ns: $\xi$ = 13.4 nm, $\theta = 10.0^{o}$



(v) 28.5 ns: $\xi$ = 13.3 nm, $\theta = 6.6^{o}$



(vi) 29.1 ns: $\xi$ = 13.6 nm, $\theta = -0.2^{o}$

**Figure 5-2.** Representative configurations of PPE/SWNT and PPES/SWNT from unbiased 40 ns simulations. (i) PPE/SWNT in vacuum, (ii) PPES/SWNT in vacuum, and (iii) PPES/SWNT in aqueous solvent. (iv-vi) Sampled configurations of PPE/SWNT in vacuum during the simulation at (iv) 27.8 ns, (v) 28.5 ns, and (vi) 29.1 ns.

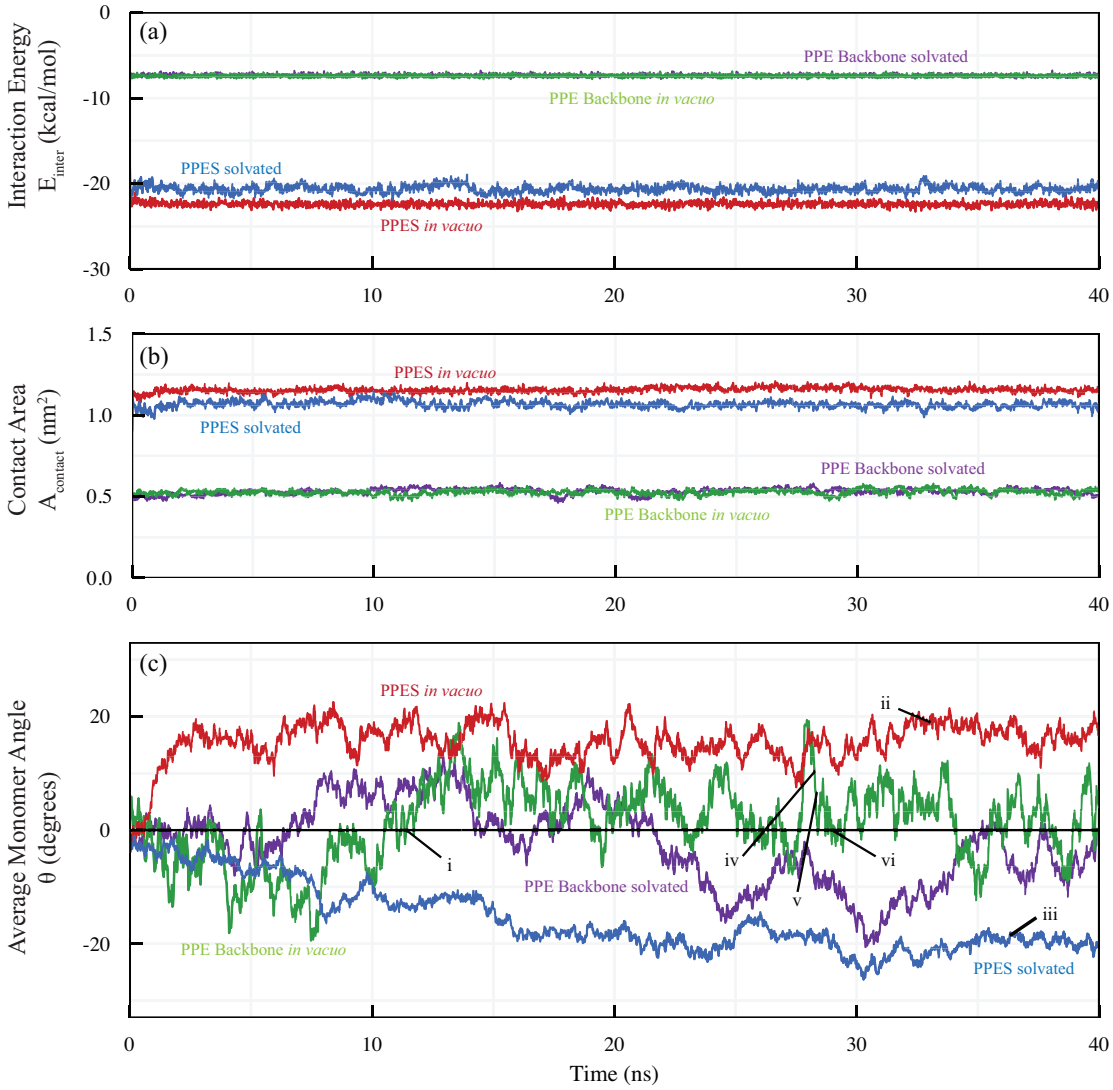



**Figure 5-3.** (a) Evolution of the interaction energy per monomer unit between the polymer and the SWNT, $E_{inter}$. $E_{inter}$ is calculated as the total nonbonding energy of the polymer (per monomer unit) with the (10,0) nanotube. (b) Evolution of the contact area between the polymer and the SWNT, $A_{contact}$, per monomer unit with time for the three 40 ns MD simulations. $A_{contact}$ is calculated as given by Equation Eq. 5-5. (c) Evolution of wrapping angle, $\theta$, with time for the three 40 ns MD simulations. Initial configuration in each case is the linear polymer, $\theta = 0$, adsorbed onto the SWNT surface. Sampled configurations i–vi in (Figure 5-2) correspond directly to the indicated positions in each trajectory. PPE/SWNT in vacuum (green), PPE/SWNT in aqueous solvent (purple), PPES/SWNT in vacuum (red) and PPES/SWNT in aqueous solvent (blue). For each sampled configuration, $\theta$ values are obtained from the 20 interior *p*-{2,5-bis(3-propoxysulfonate)}phenylene]ethynylene units (Figure 5-1).

= -20.6±0.4 kcal/mol. The vacuum simulations of PPES/SWNT have a relatively higher mean interaction energy, $E_{inter}$ = -22.4±0.3 kcal/mol due to the extensive van der Waals contact of the side chains onto the tube. The PPE/SWNT has a much lower mean interaction energy, $E_{inter}$ = -7.4±0.1 kcal/mol due to its lack of side chain interactions with the SWNT. As further evidence of the strong adsorption of the polymer, the contact area of the polymer with the surface of the SWNT per monomer unit is presented in (Figure 5-3(b)) and is defined as[291]

**Eq. 5-5**
$$A_{contact} = \frac{A_{polymer} + A_{SWNT} - A_{complex}}{2}$$

where for each sampled configuration, $A_{polymer}$ is the surface area of the isolated polymer chain, $A_{SWNT}$ is that of the isolated nanotube, and $A_{complex}$ is that of the complex comprising only the polymer and SWNT. The surface area in each case was calculated with the solvent-accessible surface area method available in VMD[280], using the Shrake-Rupley algorithm[292] and a probe radius of 1.4 Å. The contact area per monomer unit fluctuates about a value established in the first few nanoseconds of each trajectory. The PPES/SWNT solvated system has a mean contact area of $A_{contact} = 1.07 \pm 0.02\ nm^2$, while the PPES/SWNT in vacuum system has a slighter higher mean at $A_{contact} = 1.16 \pm 0.02\ nm^2$. The PPE/SWNT system in vacuum has the lowest mean contact area of $A_{contact} = 0.53 \pm 0.02\ nm^2$. These results are summarized in Table 5-1. These findings involving interaction energy and contact area support the choice of the initial conditions where the polymers are within van der Waals contact of the SWNT, as the polymers remain adhered to the nanotube surface and exhibit little variation in $E_{inter}$ and $A_{contact}$ throughout each simulation.

| Polymer System | After Initial Minimization | | | Average over last 20 ns | | | | |
|---|---|---|---|---|---|---|---|---|
| | $E_{inter}$ (kcal/mol) | $A_{contact}$ (nm$^2$) | $\theta$ ($^\circ$) | $E_{inter}$ (kcal/mol) | $A_{contact}$ (nm$^2$) | $\theta$ ($^\circ$) | $p$ (nm) | $\varphi$ ($^\circ$) |
| PPE *in vacuo* | $-8.1$ | 0.49 | 0 | $-7.4 \pm 0.1$ | $0.53 \pm 0.02$ | $+3 \pm 5$ | - | - |
| PPE solvated | $-7.9$ | 0.49 | 0 | $-7.3 \pm 0.2$ | $0.53 \pm 0.02$ | $-7 \pm 6$ | - | - |
| PPES *in vacuo* | $-19.0$ | 0.91 | $-3$ | $-22.4 \pm 0.3$ | $1.16 \pm 0.02$ | $+16 \pm 3$ | $16 \pm 4$ | $44 \pm 3$ |
| PPES solvated | $-18.1$ | 0.90 | $-4$ | $-20.6 \pm 0.4$ | $1.06 \pm 0.02$ | $-20 \pm 2$ | $13 \pm 1$ | $80 \pm 2$ |

**Table 5-1.** Summary of mean values for the equilibrated structures (last 20 ns) of the three unbiased systems; $E_{inter}$, the nonbonding energy per monomer (kcal/mol); $A_{contact}$, contact area between the complex per monomer (nm$^2$); $\theta$, average monomer contact angle($^\circ$); $p$, average helical pitch (nm); and $\varphi$, the average side chain orientation angle ($^\circ$). The last two values ($p$ and $\varphi$) are not reported for the PPE Backbone simulations due to the polymer's lack of persistent helical structure and the absence of propoxysulfonate side chains. All uncertainties are $\pm$ one standard deviation.

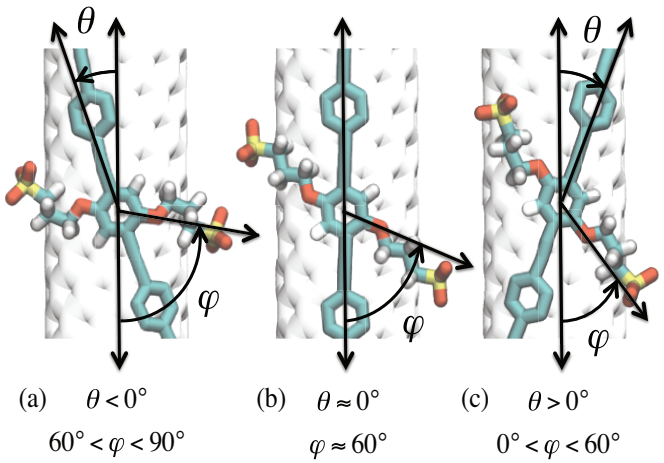



**Figure 5-4.** Illustration of wrapping angle, $\theta$, and side chain angle, $\varphi$. $\theta$ is the angle between a vector defined by atoms in the monomer unit ($C_1$ and $C_2$ in Figure 5-1) and the longitudinal axis of the nanotube. $\varphi$ is the angle subtending the vector defined by a C-O bond (between a phenylene carbon and the ether oxygen to which it is bonded) and the nanotube axis (Figure 5-4). (b) The linear configuration $\theta = 0\,^\circ$. (a) $\theta < 0\,^\circ$ is uniquely associated with $\varphi > 60\,^\circ$. (c) $\theta > 0\,^\circ$ helical direction is uniquely associated with $\varphi < 60\,^\circ$.

Additionally, the collinear PPE/SWNT system was solvated and allowed to equilibrate. Much like the PPE/SWNT in vacuum, the polymer maintained van der Waals contact with the nanotube surface, yielding similar values of $E_{inter}$ = -7.3±0.2 kcal/mol and $A_{contact}$ = 0.53±0.02 $nm^2$. The polymer had no persistent global helical structure. The resulting configurations from the 40 ns trajectory indicate similar behavior to the system in vacuum: the polymer explores an ensemble of linear and helical structures, never adopting a well-defined superstructure and often returning to the collinear state.

For the solvated and vacuum simulations, conformations of PPES and PPE were sampled and analyzed with regard to the local orientations of the monomer units. The local helical wrapping angle per monomer, $\theta$, is the angle between a vector defined by atoms in the monomer unit ($C_1$ and $C_2$ in Figure 5-1) and the longitudinal axis of the nanotube (Figure 5-4). $\theta$ characterizes the orientation of polymer locally on the nanotube (Figure 5-4). For

an ideal helix, the relation between $\theta$ and the corresponding pitch $p$ is

**Eq. 5-6**

$$p = \left| \frac{2\pi r}{tan\theta} \right|$$

A unique helical wrapping can be described by $\theta$ and $\varphi$, where $\varphi$ is the angle subtending the vector defined by a C-O bond (between a phenylene carbon and the ether oxygen to which it is bonded) and the nanotube axis (Figure 5-4). In the case of the solvated PPES/SWNT simulation, the helical conformation has $\varphi > 60\,^\circ$ and $\theta < 0\,^\circ$, where side chains are oriented roughly perpendicular to the nanotube cylindrical axis. Conversely, the vacuum PPES/SWNT simulation has $\varphi < 60\,^\circ$ and $\theta > 0\,^\circ$, where the side chains are closer to colinear with the tube and maintain contact with the nanotube surface. These observations of particular values of $\theta$ do not indicate a preferred chirality or helical handedness of PPES, in solvent or in vacuum. For any of the helical configurations, e.g., the three depicted in Figure 5-4, there is a mirror image possible that has the opposite helical handedness, for which $\theta$ and $\varphi$ are of opposite sign. Accessing such enantiomeric conformations would require desorption of the polymer and/or internal torsional rotation of the phenylene units to reorient the side chains. As mentioned, PPES remains adsorbed in van der Waals contact with the SWNT throughout the simulations, and such rare, high-energy events (desorption and internal rotation) were not observed in any of the simulations. The evolution of $\theta$ for the solvated PPE/SWNT system has similar fluctuations to the vacuum case. The presence of solvent has little impact on the range superstructures sampled by the PPE polymer and does not lead to persistent helical structures. For this reason, all further discussions of the PPE system concern the simpler study of PPE/SWNT in vacuum.

Representative configurations from each of three simulations are presented in Figure 5-2(i-iii). In a vacuum simulation, the PPE/SWNT system adopted no well-defined superstructure and sampled multiple helical conformations, often returning to a collinear state where $\theta = 0\,^\circ$. Sampled configurations of the PPE/SWNT system are shown in Figure 5-2(iv-vi), wherein large fluctuations in $\theta$ for the backbone (see Figure 5-3c) indicate the

lack of persistent, well-defined helicity. Persistent helical structures were observed, however, in the PPES/SWNT simulations, both in vacuum and in solution. The fully solvated PPES/SWNT system reached $\theta = -20 \pm 2\,^\circ$ with pitch of $p = 13\pm$ 1 nm within 15 ns. The vacuum simulation of PPES equilibrated within 5 ns to $\theta = +16 \pm 3\,^\circ$, with pitch $p = 16\pm$ 4 nm.

To verify the observed helical preferences, additional calculations were performed. For each of the systems in Figure 5-3, two repeat trajectories of 20 ns each were simulated (data not shown), and the observed equilibrated conformations possessed helical parameters within statistical uncertainty of the values in Table 5-1. In addition, the equilibrated structures of PPES in vacuum and solution were exchanged. That is, the final equilibrium conformation ($\theta = +17\,^\circ$) of the vacuum simulations of PPES-SWNT was fully solvated, and charges and counter ions were introduced. Within 20 ns, the newly solvated helix had reorganized, passed through a linear conformation, and changed its helical wrapping to arrive at a structure having $\theta = -20 \pm 2\,^\circ$, i.e., the polymer converted from structure *iii* to structure *ii* in Figure 5-2. Similarly, a vacuum simulation was performed using the final conformation ($\theta = -20\,^\circ$) from the previous solvated simulation of PPES-SWNT as the initial condition. The polymer also changed helical wrapping to form the superstructure ($\theta = +16 \pm 3\,^\circ$) that was previously observed in vacuum simulations (evolution from *ii* to *iii* of Figure 5-2). Each of these simulations was extended an additional 20 ns, and the helical parameters remained unchanged, suggesting that the observed helical structures are independent of the initial conformation.

Several inferences can be made from these simulation results. PPES and PPE adhere strongly to the SWNT. For each polymer, the phenyleneethynene backbone maintained contact with the SWNT for each simulation (in solvent and in vacuum), and the polymer did not dissociate from the nanotube in any of the simulations. Inspection of Figure 5-3 shows that the interaction energy $E_{inter}$ and contact area $A_{contact}$ are essentially invariant after the first few nanoseconds, whereas the local monomer contact angle (and thus the global

polymer conformation) continues to evolve on a longer multi-ns timescale, particularly for the solvated simulation where persistent helical structure is achieved only after 15-20 ns. The backbone accommodates deviations from linearity on nm length scales, as observed by persistent helical structures of the PPES/SWNT systems and the conformational fluctuations observed in the simulations of PPE/SWNT. Both vacuum and solvated simulations of PPES/SWNT yielded persistent helical structures, suggesting that the propoxysulfonate side chains play an essential role in specifying the observed helical wrapping. The vacuum and solvated helical superstructures of PPES/SWNT are different, however, as is evident in the respective values of $\theta$ and $\varphi$, i.e. the orientations of the monomers in each helix formed under these two different conditions. Thus solvent and electrostatic interactions play a role in determining the precise helical structure of the PPES/SWNT. For the solvated simulation, the helical pitch matches that observed in transmission electron micrographs.[102] Interestingly, neither the interaction energy nor the contact area track with the formation of helical superstructure. This observation suggests that a subtle set of interactions involving the polymer, nanotube, and solvent are responsible for the formation of persistent helical superstructures and motivates the use of free energy calculations to quantify the energetics of such helical structure formation.

## 5.4. Potential of Mean Force and Helical Pitch

To quantify the relative stabilities of different helical superstructures of PPES/SWNT in vacuum and in solution, potentials of mean force $\Delta A(\xi)$ were calculated using the adaptive biasing force method (Figure 5-5)[135,287–289]. The order parameter (or reaction coordinate) is $\xi$, the difference in z-coordinates of carbon atoms in the terminal monomers of each polymer as denoted in Figure 5-1, i.e., $\xi$ is the displacement between the indicated atoms projected onto the longitudinal symmetry axis of the SWNT. In simulations of PPES/SWNT in vacuum and in aqueous solution, the polymer was observed to remain adsorbed to the SWNT in a helical conformation, and thus for the potential of mean force calculations, a helical initial configuration was chosen to more rapidly sample "equilibrium" configurations

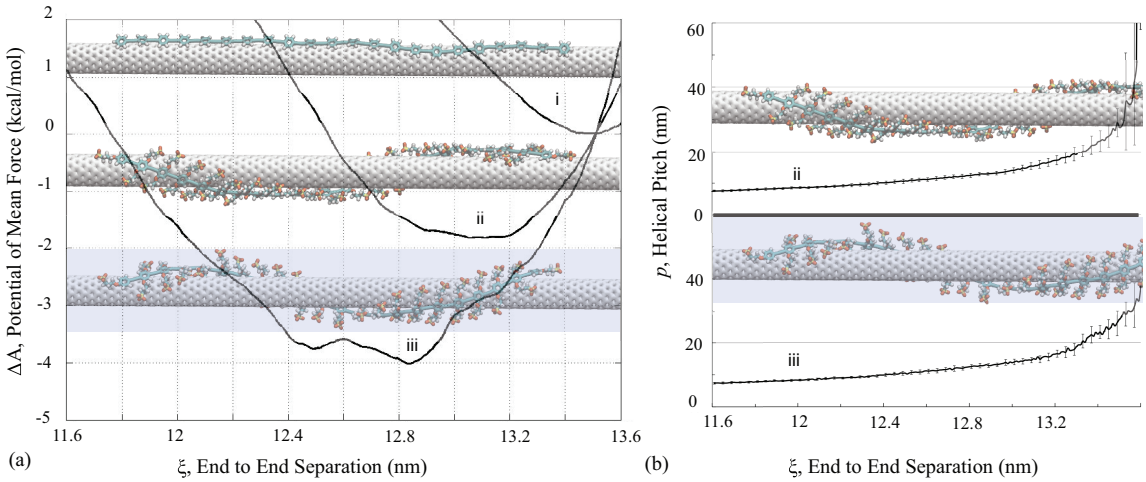



**Figure 5-5.** (a) Calculated potential of mean force, $\Delta A(\xi)$, as a function of $\xi$, the displacement between end monomers projected on the nanotube longitudinal axis (Figure 5-1). $\Delta A$ provides the relative free energies of helically wrapped polymer/SWNT structures. (i) PPE/SWNT in vacuum, (ii) PPES/SWNT in vacuum, and (iii) PPES/SWNT in aqueous solution. (b) The average helical pitch $p$ vs $\xi$ for (ii) PPES in vacuum, and (iii) PPES in aqueous solution; $p$ monotonically increases with $\xi$ in each case.

for each value of $\xi$. The order parameter was considered in the range $11.0 \leq \xi \leq 13.7$ nm, where $\xi$=13.7 nm corresponds to the fully extended, linear 20-mer. This range was divided into 15 windows each of 0.4 nm width, overlapping 0.2 nm on each side to improve sampling continuity. For each of the 15 windows of $\xi$, the initial configuration of the polymer was selected as a helix having the value of $\xi$ at the center of the window ($r$ = 7.35 Å). Each calculation sampled instantaneous force values for 20 ns, collected in bins 0.01 nm wide. To reduce possible non-equilibrium artifacts, 20,000 samples were accrued in each bin before introducing the biasing force (Equation Eq. 5-2) within each bin.

The simulation results indicate that for these systems with their well-defined helices, the pitch $p$ is a monotonic function of $\xi$ (see Figure 5-5 b). The potential of mean force $\Delta A(\xi)$ is the difference in free energy of the system at a particular value of $\xi$ (or equivalently, the helical pitch $p$), and that at a reference value $\xi_o$ (see Equation Eq. 5-3). $\Delta A(\xi)$ thus provides the relative free energies of different helical PPES/SWNT superstructures.

For the PPE/SWNT system in vacuum, $\Delta A(\xi)$ has a single minimum at $\xi$ = 13.51 nm,

where the polymer fluctuates about the linear conformation, consistent with no preference for helically wrapped configurations under these conditions. The position of this minimum is slightly less than that of the fully extended, linear polymer ($\xi = 13.7$ nm), as expected from the observed fluctuations of PPE polymer when adhered to the SWNT (Figure 5-2(i)). This value $\xi_o = 13.51$ nm is chosen as the reference value of $\xi_o$ with which to calculate $\Delta A(\xi)$ since $\xi_o$ is the value expected of a poly-[*p*-phenlyene]ethynylene polymer that adheres to the SWNT with no persistent helical structure and fluctuates about the linear configuration.

For the vacuum simulations of PPES/SWNT, $\Delta A(\xi)$ has a single minimum at $\xi = 13.05$ nm. As $\xi$ decreases below this value, $\Delta A(\xi)$ increases as the polymer takes on unfavorable, tightly wrapped configurations. At this minimum, $\Delta A = -1.8$ kcal/mol, suggesting a modest stabilization of the helical conformation at 300 K. The value $\xi = 13.05$ nm corresponds to a helical pitch of $p$ =16 nm, which concurs with that observed in the previously described, unconstrained molecular dynamics simulation of this system (Figure 5-3). Furthermore, the breadth of the minimum is consistent with the fluctuations in pitch observed in the unconstrained simulations ($p = 16 \pm 4$).

For the solvated PPES/SWNT system, $\Delta A(\xi)$ has a minimum at $\xi$ =12.84 nm where $\Delta A = -4.0$ kcal/mol. This minimum corresponds to a helical pitch of $p = 13$ nm, in agreement with previous simulation results (Figure 5-3) and with previous experimental and simulation studies[102]. There is a second local minimum at $\xi$ =12.49 nm, which has $\Delta A = -3.8$ kcal/mol and corresponds to a helical pitch of $p = 10$ nm. The metastable character associated with this conformation was revealed using a subsequent simulation (data not shown). A sampled configuration of the solvated PPES/SWNT system having $\xi$ =12.49 nm was used as the initial condition of a solvated simulation of 12 ns. The polymer retained the helically wrapped structure of smaller pitch for 2 ns and then evolved to the global minimum of $\Delta A(\xi)$ within 6 ns; the polymer remained in the lower free energy helically wrapped structure ($\xi = 12.9 \pm 0.1$ nm) for the remainder of the simulation.

## 5.5. Propoxysulfonate Side Chain Conformations

In addition to providing the global helical superstructure and relative free energies of different helical superstructures, simulations can provide molecular information that can further our understanding of the helical superstructures. Side chain conformations were followed as functions of $\xi$ in the potential of mean force simulations. Four dihedral (torsional) angles of the propoxysulfonate side chain were monitored (see Figure 5-6 inset). Each of these angles were binned into one of three intervals: $0\,^\circ < \chi < 120\,^\circ$, $gauche^+$ or $p$ ; $120\,^\circ < \chi < 240\,^\circ$, $trans$ or $t$; or $240\,^\circ < \chi < 360\,^\circ$, $gauche^-$ or $m$[293]. The abbreviations $p, t$, and $m$ label the $gauche^+, trans$, and $gauche^-$ states, and a particular side chain rotamer state is denoted by an ordered quartet of these labels $\chi_4\chi_3\chi_2\chi_1$, e.g., *tttt* indicates the all trans side chain conformation. This discretization results in 81 potential rotamer states for each side chain, and the probabilities of these side chain rotamer conformations were calculated using all side chains of [*p*-{2,5-bis(3-propoxysulfonate)}phenylene]ethynylene monomers (Figure 5-1) every 20 ps from the collection of biasing force trajectories so as to sample conformations of the PPES/SWNT system at a given value of $\xi$ (see Figure 5-6 inset).

The populations of the side chain rotamer states are presented in Figure 5-6. In the vacuum PPES/SWNT simulations, the side chains populate predominantly three conformations, *tptt, tttt, tmtt*, as seen in Figure 5-6 (a). The dihedral rotameric states of the side chains are almost exclusively populated by states that only differ with regard to $\chi_3$. For $\xi = 11$ nm and thus small pitch, the three conformers are roughly equally populated, but as $\xi$ increases a single state *tptt* becomes favored. For the solvated PPES/SWNT simulations, a single rotamer state, *tptt*, predominates and its population is independent of $\xi$ (see Figure 5-6 (b)).

The most populated side chain conformations are rendered in Figure 5-7. In the solvated simulations, the *tptt* rotamer state dominates throughout. Figure 5-7 illustrates this side chain state for $\theta = -20\,^\circ$ and $\varphi = 80\,^\circ$, the values associated with the preferred helically wrapped structure of the solvated PPES/SWNT system. For these values of $\theta$ and $\varphi$,

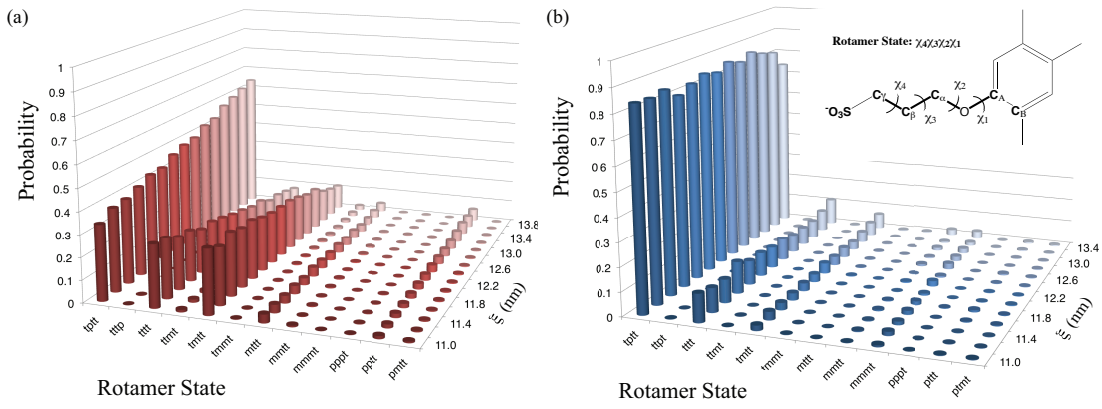



**Figure 5-6.** Populations of propoxysulfonate side chain conformations. Conformations are classified according to the side chain dihedral angles $\chi_4\chi_3\chi_2\chi_1$: $\chi_4$ ($SC_\gamma C_\beta C_\alpha$); $\chi_3$ ($C_\gamma C_\beta C_\alpha O$); $\chi_2$ ($C_\beta C_\alpha O C_A$); $\chi_1$ ($C_\alpha O C_A C_B$), where the ordered atoms in parentheses are those that specify the corresponding dihedral angle (inset). Each dihedral angle is grouped into one of three rotameric states: *p*: gauche$^+$, *t*: trans, *m*: gauche$^-$. A side chain rotamer state is denoted by an ordered quartet of these labels $\chi_4\chi_3\chi_2\chi_1$, e.g., *tptt* indicates that $\chi_4\chi_3\chi_2\chi_1$ take on the trans, gauche$^+$, trans, and trans dihedral states, respectively. Only side chain conformational states with probabilities greater than 0.001 are shown. Populations (probabilities) are calculated from final 20 ns of unbiased molecular dynamics simulation (Figure 5-2). (a) PPES/SWNT in vacuum. (b) PPES/SWNT in aqueous solution.

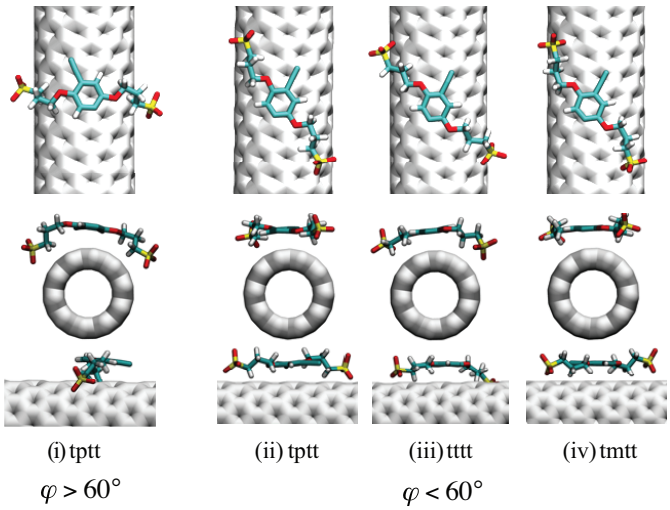



**Figure 5-7.** Rendering of a single *p*-{2,5-bis(3-propoxysulfonate)}phenylene]ethynylene monomer from the PPES/SWNT system for "equilibrated" structures to illustrate the positioning of the side chains (Figure 5-2). In each case, the rotamer side chain configuration is shown using 3 orthogonal views. (i) PPES/SWNT in aqueous solvent. $\theta = -20\,^\circ$. (ii-iv) PPES/SWNT in vacuum. $\theta = +16\,^\circ$.

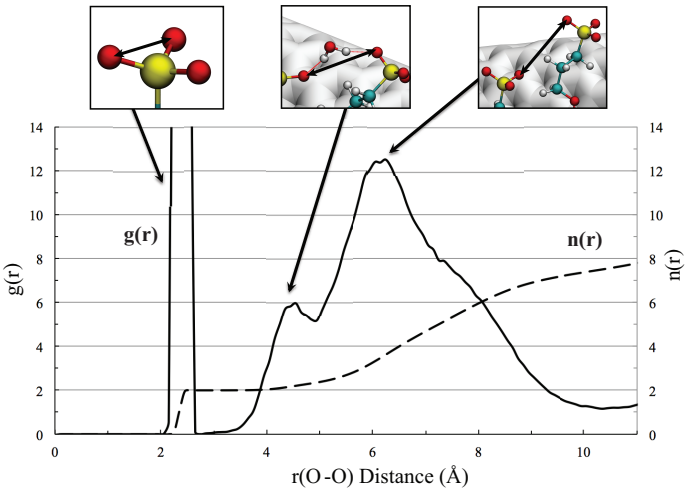



**Figure 5-8.** Radial distribution function $g(r)$ for pairs of sulfonate oxygen atoms, i.e., $g(r)$ is the relative density of sulfonate oxygen atoms at the distance $r$ given one such oxygen is at the origin. The insets illustrate structural elements corresponding to peaks in $g(r)$: O-O pair within a sulfonate group ($r$ = 2.2 Å), O-O pair on adjacent side chains bridged by a water molecule ($r$ =4.5 Å), and O-O pair for adjacent sulfonates not hydrogen bonded to the same water molecule ($r$ =6.2 Å). The average number of sulfonate oxygens within a distance $r$ of another, $n(r)$, is also shown (dashed) and obtained from integrating $g(r)$.

the *tptt* side chain conformation maintains van der Waals contact between the SWNT and the side chain methylenes, while leaving the sulfonate group exposed and accessible to aqueous solvation. The *tptt* rotamer configuration seen in Figure 5-7 illustrates the solvent accessibility of the sulfonate group, which allows for formation of on average 5 hydrogen bonds to water per side chain (Figure 5-9).

In the vacuum PPES/SWNT simulations ($\theta = +16\,^\circ$ and $\varphi = 44\,^\circ$), the three most probable rotamer states align the side chain along the curved surface of the nanotube so as to essentially maintain van der Waals contact of the entire side chain, including the sulfonate, with the SWNT. These particular orientations of the side chains relative to the surface of the SWNT are somewhat artificial since no electrostatic interactions are present due to the absence of partial charges and counter ions. Notable, however, is the fact that a helical structure is strongly preferred even in the absence of electrostatic interactions and solvent; van der Waals interactions of the side chains with the SWNT play a key role in specifying such structures.

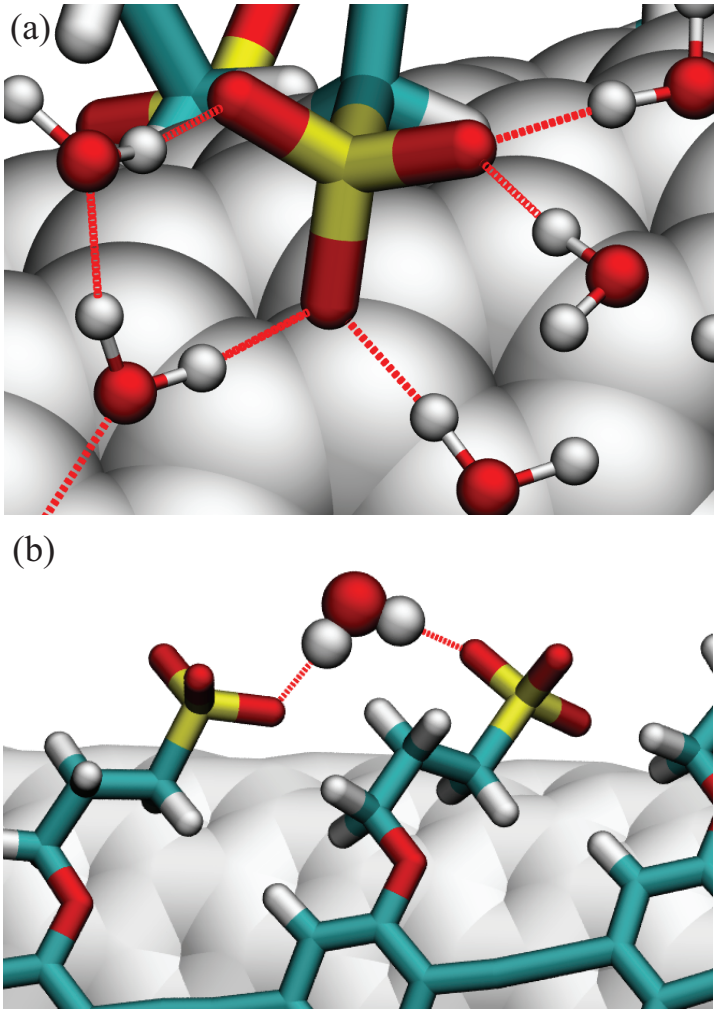



**Figure 5-9.** (a) Representative configuration of five hydrogen bonds in the water shell surrounding a single sulfonate side chain. (b) A bridging water molecule forming two hydrogen bonds with adjacent sulfonate side groups of the helically wrapped PPES.

In the unconstrained solvated simulations of PPES/SWNT (see the last 20 ns of Figure 5-3 (c)), the distribution of the distance $r$ between sulfonate oxygen atoms can be used to characterize the orientations of side chains relative to each other. The resulting distribution of distances, the radial distribution function $g(r)$, is shown in Figure 5-8. The first peak at 2.3Å corresponds to the average distance between oxygens that are part of the same $SO_3^-$ group. The peak at 6.2Å corresponds the average distance between oxygen atoms of side chains that are on adjacent monomers when each is in a conformation that falls into the *tptt* rotamer state. The number of $SO_3^-$ oxygens within a chosen distance can be obtained from integrating $g(r)$: the integrated curve between $0\,\text{Å} < r < 11.35\,\text{Å}$ yields 8 oxygens, which can be assigned to 2 other oxygens in the sulfonate and 3 in each of two nearest neighboring side chains. The peak located at 4.5Å occurs at a distance where neighboring side chains can hydrogen bond to a shared water molecule (Figure 5-9). The fraction of adjacent monomer pairs that are hydrogen bonded to the same water molecule via their sulfonate groups is 24% (Figure 5-9). An analysis of the rotamer states for the side chains participating in these hydrogen bond water bridges shows that on average 82% are in the *tptt* rotamer state. The relatively high frequency with which the sulfonate side chains are observed to be in in the *tptt* rotamer state suggests that these water bridges between side chains play an important role in specifying the overall helical superstructure.

## 5.6. Helical Superstructures

The simulations provide an atomistically detailed perspective on helical PPES/SWNT superstructures. For all simulations in water and in vacuum, the PPE and PPES polymers were observed to adhere via the phenylene backbone and not desorb from the tube. The backbone phenyleneethynylene units and nanotube surface were within van der Waals contact throughout the simulation. The noncovalent association of the aromatic units of the polymer backbone with the nanotube surface is responsible for the observed polyaryleneethynylene adhesion to the SWNT and, in the case of those polymers functionalized with propoxysulfonate side chains, the ability to disperse carbon nanotubes in water and a va-

riety of organic solvents.[101,102] The poly[*p*-phenylene]ethynylene backbone is flexible and adopts structures other than the linear conformation, as is evidenced by both the fluctuations in the global structure observed in the vacuum PPE/SWNT simulations and the helically wrapped superstructures observed in the simulations of PPES/SWNT in vacuum and in solution. These excursions from the linear conformation are accommodated by variation of bond angles and internal rotations in the polymer. Deviations of bond angles from their ideal values are well known and have been addressed in the parameterization of the potential energy used in the simulations[282]: (a) the exocyclic angle where the alkyne is attached to a phenyl ring can exhibit deviations of several degrees from 120 °[294]; and (b) the deviations from linearity of the bond angle that includes the two ethyne carbons of up to 8 ° are consistent with both experimental and theoretical studies.[282,294,295] In addition, the internal rotation of one phenyl ring relative to another due to torsional rotation about the ethyne linkage is known to have a low barrier (0.6 kcal/mol), that is comparable to $k_bT$ at $T = 300K$.[282,296–298] Thus the phenyl rings can readily rotate so as to maintain van der Waals contact with the SWNT to accommodate helical superstructures. The underlying surface of the (10,0) SWNT is smooth and essentially featureless. Multiple conformations of a polymer adhered to the tube are expected, which is consistent with the large fluctuations in superstructure and absence of persistent helical wrapping observed for the vacuum simulations of the PPE/SWNT system. Thus the presence of the propoxysulfonate side chains and their (2,5) substitution pattern on each monomer appear to specify the particular 13 nm pitch helical superstructure observed in the simulations and in experiment.[102]

In the aqueous environment, the side chains have a strong preference for the *tptt* side chain conformation, and in the preferred helically wrapped structure, this conformation maintains contact of the side chain methylene groups with the nanotube surface while making the sulfonate groups accessible for aqueous solvation. In protein simulations, similar coordination of electrostatic and hydrophobic interactions has been observed[299]. Here, we note that the side chain conformations are dictated by both the optimal van der Waals contact of the aliphatic segments with the nanotube surface, while maintaining solvation

of the sulfonate groups, creating a well-defined configuration for the side chains to adopt. Within this work, only aqueous systems have been considered. For other solvents and side chains, alternate structures could potentially be observed, which may suggest why no evidence of helical wrapping was reported in studies of related aryleneethynylene polymer systems having uncharged side chains.[248]

## 5.7. Conclusion

PPES/SWNT adopts a superstructure where the polymer is helically wrapped about the carbon nanotube, and herein atomistic molecular dynamics (MD) simulations were performed to better understand the origins of this helical wrapping for what is ostensibly a "linear" polymer. In unbiased, aqueous simulations of a 20-monomer PPES and a (10,0) SWNT, a helical structure was observed to form spontaneously on the nanosecond time scale. The helical pitch of the polymer matched that measured experimentally: $p = 13$ nm[102]. In simulations carried out in vacuum and in solution, no persistent helical structure was observed for a related system PPE/SWNT, in which the phenylene units of the poly[*p*-phenylene]ethynylene polymer lacked 2,5-bis(3-propoxysulfonicacidsodiumsalt) side chains. For PPES/SWNT simulated in vacuum, the system relaxed to a helical structure distinct from that of the solvated case with regard to the local monomer orientation and overall pitch ($p = 16$ nm). Potential of mean force calculations provided the relative stabilities of different helical configurations of the PPES/SWNT system, and the experimentally observed helical pitch in aqueous solution was found to be a global free energy minimum. Interactions of the propoxysulfonate side chains with the nanotube and with solvent were found to specify the helical superstructure observed in water, where one side chain rotamer state, the *tptt* rotamer, was almost exclusively populated independent of the range of helical pitch values accessed in the course of these simulations. In the preferred helical superstructure, the side chain maintained hydrophobic contact with the carbon nanotube while exposing the sulfonate group for aqueous solvation. Specific interactions between side chains were also observed: 24% of the propoxysulfonate side chain population was found to form a

hydrogen-bonded interaction with a shared water molecule, where 82% of the participating side chains took on the *tptt* rotamer state necessary for the solvated helical superstructure. In water, the calculated difference in free energy between the preferred helically wrapped PPES/SWNT structure and that of the polymer adsorbed to the tube in a state that fluctuates about the linear conformation is only 0.2 kcal/mol/monomer. The relatively small value of this stabilization suggests that the global superstructure of the PPES/SWNT system could potentially be controlled by engineering weak, noncovalent interactions, e.g., by modulating van der Waals interactions and solvation forces involving the monomer units through chemical modification of the PPES archetype. As with proteins[300,301] and other folding polymers[302,303], these polymer/SWNT systems possess potentially complex energy landscapes due to their many noncovalent and frustrated interactions, but this complexity does not preclude design. Such engineering can be facilitated by molecular simulations, which address the myriad molecular structures and interactions present in solvated polymer/SWNT systems while providing molecularly detailed insight into the relative stabilities of possible superstructures.

# 6 Simulations of Chiral Polymers Wrapped about Carbon Nanotubes *

For highly charged semiconducting polymers that utilize a (S)-1,1'-bi-2-naphthol component in their repeat unit, TEM and AFM images of SWNTs wrapped by these polymers suggest preferences for helical wrapping handedness commensurate with the chirality of the polymer. However, a modest statistical analysis of these images indicates that roughly 20% of the helical structures are formed with the opposite handedness. CD spectroscopic data and a set of basic TDDFT calculations that attempt to correlate the spectral signatures of the chiral 1,1'-binaphthalene unit offer two binding modes, what are denoted as *cis-facial* and *cis-side* (Figure 6-1). For similar polymers which possess a 2,2'-1,3 benzyloxy bridging the 1,1'-bi-2-naphthol, the restricted set of torsional angles available to the binaphthol unit elicit an even stronger preference for the 'expected' helical handedness given a polymer chirality. An analysis of TEM images reveals that these bridged-binaphthalene-based polymers form polymer-wrapped CNT constructs in which chiral polymer helical wrapping manifests an overwhelming preference (96%) for the expected left-handed helical superstructure (pitch-length = 8 ± 2 nm). To provide a comprehensive molecular perspective that spans length scales ranging from the local conformational restrictions of the bridged binaphthyl moiety to the global helical superstructures observed in the AFM and TEM data, molecular dynamics (MD) simulations were conducted for such chiral polymer-nanotube systems. The following chapter offers a series of equilibration simulations of SWNTs that are helically wrapped by *S*-**PBN(b)-Ph**$_3$ and *S*-**PBN(b)-Ph**$_5$ in the presence of water and counter ions so as to characterize the persistence of helical superstructures and the local orientation and conformation of the constituent monomers. In the simulations, *S*-**PBN(b)-Ph**$_5$ is the most robust and able to persistent in a helical superstructure.

*Adapted from Pravas Deria, Christopher D. Von Bargen, , Jean-Hubert Olivier, Amar S. Kumbhar, Jeffery G. Saven, and Michael J. Therien. "Single Handed Helical Wrapping of Single-Walled Carbon Nanotubes by Chiral, Ionic, Semiconducting Polymers." *J. Am. Chem. Soc.*, 2013, 135 (43).

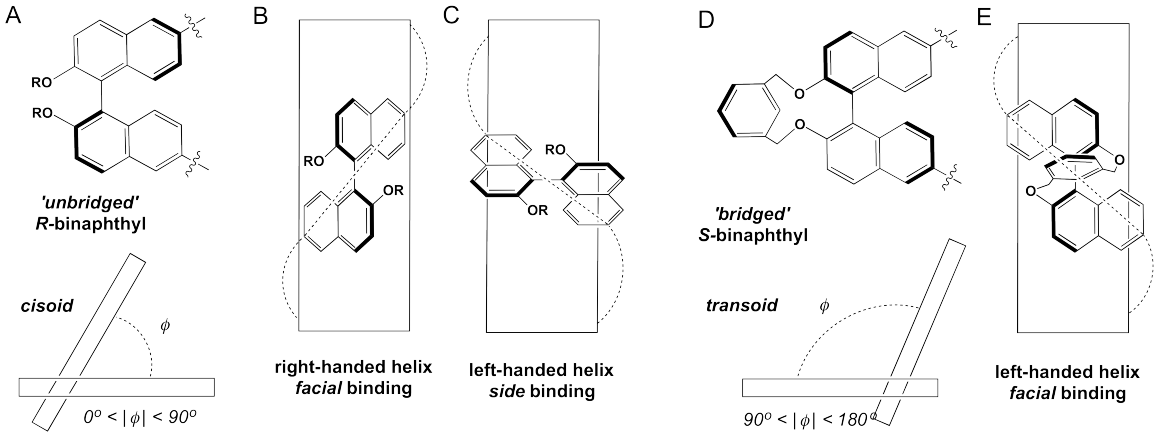



**Figure 6-1.** Conformations of 1,1'-bi-2-naphthol-derived polymer chain components, and their possible binding modes at SWNT surfaces: (a) the *cisoid* conformation adopted by the *unbridged* **R**-chirality binaphthalene unit; (b) cartoon depicting *cisoid-facial*' binding of an **R**-chirality binaphthalene to the SWNT surface in a right-handed helical superstructure; (c) cartoon depicting the *cisoid-side*' binding of an **R**-chirality binaphthalene to the SWNT surface in context of the "unexpected" left-handed helical superstructure; (d) the *transoid* conformation adopted by a 2,2'-(1,3-benzyloxy)-*bridged*-1,1'-bi-2-naphthol unit, and (e) *transoid-facial*' binding mode of the 2,2'-(1,3-benzyloxy)-*bridged*-1,1'-bi-2-naphthol moiety with the SWNT surface in the context of a left-handed helical superstructure.

## 6.1. Simulation Tools

### 6.1.1. Molecular Dynamics Simulations

All simulations were performed using the molecular dynamics program NAMD2.7[275]. Orthorhombic periodic boundary conditions were applied in all three Cartesian dimensions, and the average dimensions were 60 Å x 60 Å x 330 Å. The equations of motion were integrated with a time step of 2 fs. Covalent bonds involving hydrogen atoms were constrained to their equilibrium length by means of the SHAKE/RATTLE algorithms[276,277]. Long-range electrostatic forces were evaluated by means of the particle-mesh Ewald (PME) approach[278] with a 1 Å mesh, and van der Waals interactions were truncated smoothly with a spherical cutoff of 12 Å. For solvated simulations, the TIP3P[279] water model was used. The aqueous system was ensured to be electrostatically neutral via the addition of sodium and chloride ions using the VMD module AUTOIONIZE,[280] consistent with a salt concen-

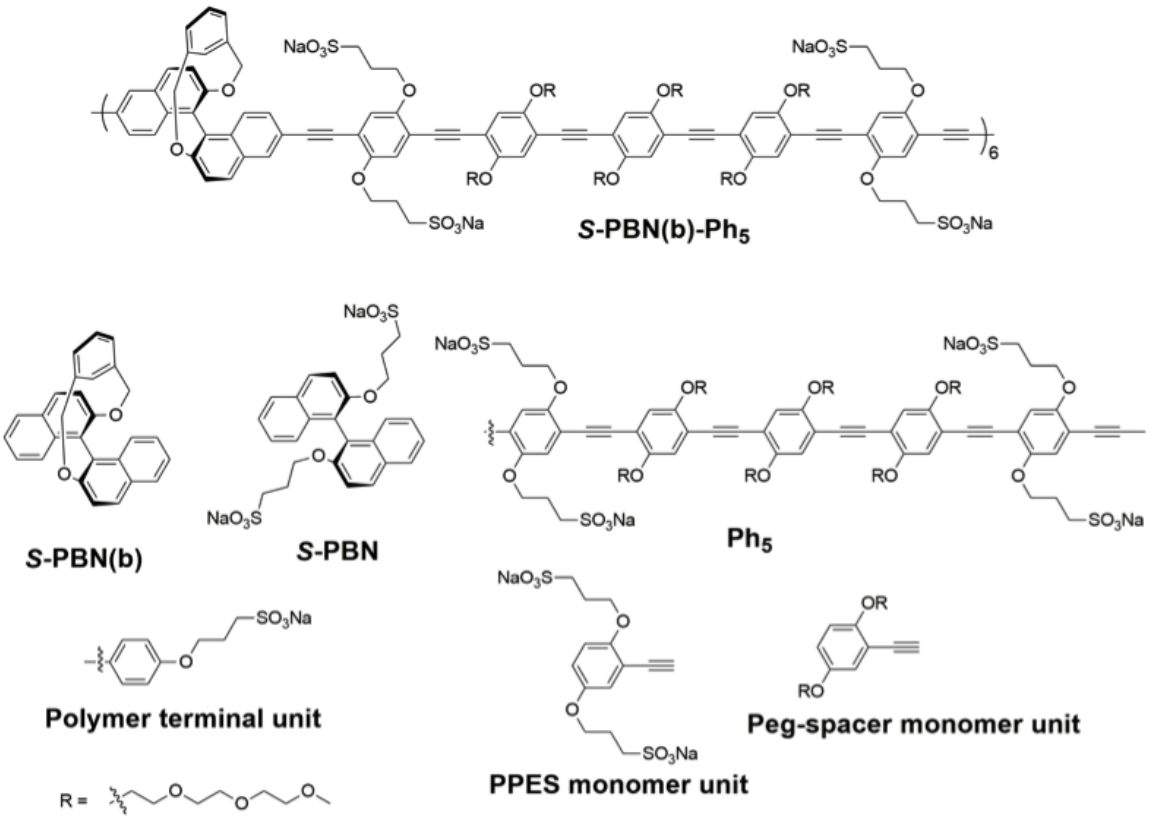



**Figure 6-2.** Ionic aryleneethynylene polymer ***S*-PBN(b)-Ph$_5$** based on 1,1'-bi-2-naphthol derivatives and various monomeric units used in molecular dynamics (MD) simulations.

tration (ionic strength) of 0.2 mol/L. The simulations in the aqueous phase included 35,960 TIP3P waters placed using the VMD module SOLVATE[280], yielding a total of 111,902 atoms including the PPES polymer, counter ions and the (10,0) nanotube. Aqueous simulations were carried out in the isothermal-isobaric ensemble; pressure and temperature were maintained at 1 bar and 300K by employing Langevin dynamics with damping coefficient of 5 ps$^{-1}$ and the Langevin piston pressure control with an oscillation period of 100 fs and decay time of 50 fs[281]. Preparation, visualization, and analysis of structures and trajectories utilized the VMD package[280].

## 6.2. Molecular Models of Chiral Binaphthylene Ethynylene Polymers

The molecular potential parameters for chiral polymer repeat units were developed using quantum mechanical calculations and parameters reported in previous work[102,282–284].

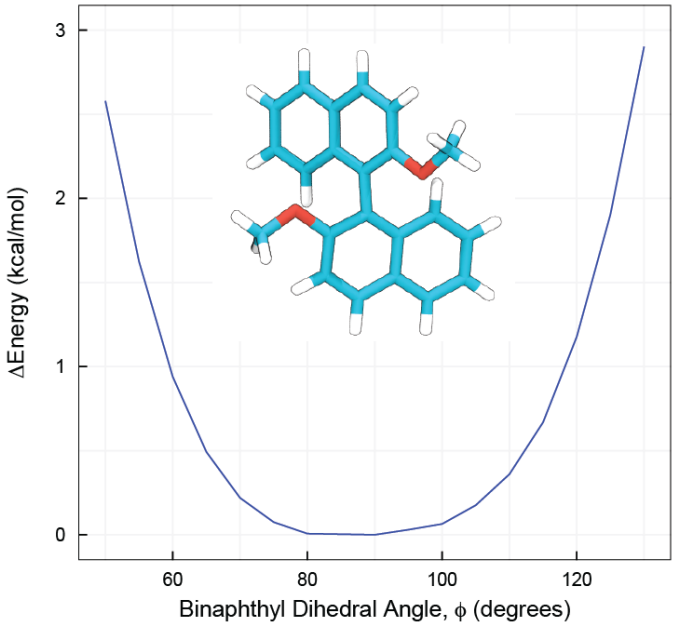



**Figure 6-3.** Dihedral potential energy of the binaphthyl bond between naphthyl rings for methoxy-binaphthyl. Values are obtained by rotation through the dihedral angle. The potential is commensurate with known potential of methoxy-binaphthyl units[305–307]

The restrained electrostatic potential method was used to obtain effective atomic charges subject to overall neutrality of each aromatic unit. The effective charges were fit using electrostatic energies computed using the HF/6- 311G** basis set in Gaussian98[304]. The naphthalene-naphthalene dihedral potential of the binaphthyl unit for both polymer derivatives was parameterized to be consistent with known potentials of methoxy-binaphthyl units[305–307] (Figure 6-3). Two representative polymers, both with *S* chirality, were considered in the simulations: ***S*-PBN(b)-Ph$_3$** and the "bridged" ***S*-PBN(b)-Ph$_5$**. Both polymers variants were prepared with the same number of subunits (35) and terminal p-4-(3-propoxysulfonicacidsodiumsalt)phenylene]ethylene units; ***S*-PBN(b)-Ph$_3$** contained 8 monomers with the addition of an additional **Ph$_3$** group, while ***S*-PBN(b)-Ph$_5$** contained 5 monomers with an additional **Ph$_5$** group. (Figure 6-2). For the helical initial conditions of the polymers, the binaphthyl interplanner torsional angles were chosen as $\theta = 90\,^\circ$ (Figure 6-4) and all side chains were positioned in the same manner (all trans) on each aromatic unit.

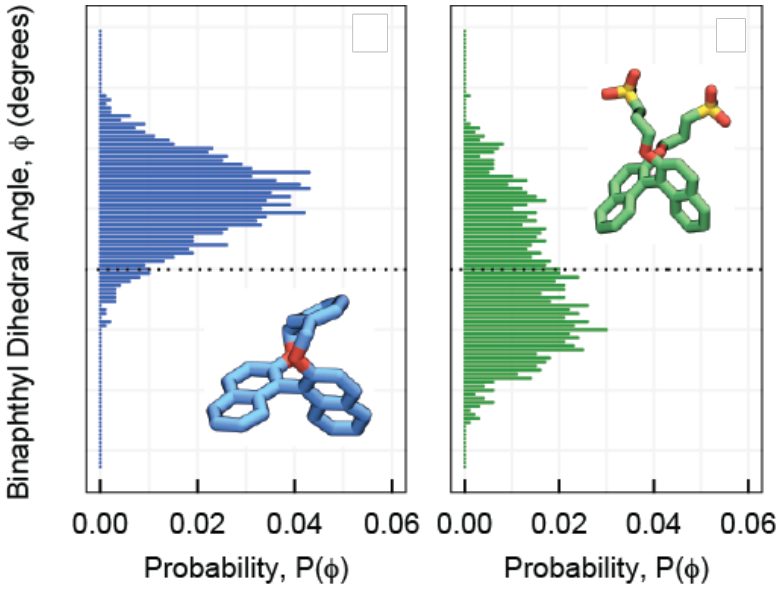



**Figure 6-4.** Aggregate histogram of the binaphthyl dihedral for 40 ns simulations of a single *S*-**PBN** and *S*-**PBN(b)** unit, respectively. For *S*-**PBN** (left, green), the dihedral angle is centered at 87 ± 17°; for *S*-**PBN(b)** (right, blue) the dihedral angle is centered at 107 ± 10°.

To improve the efficiency of these simulations and compare the properties of right- and left-handed structures, the initial conformations of the polymers were those resulting from alignment to right- and left-handed ideal helical contours. The helical parameters were chosen such that the helical pitch was p = 8.0 nm and the helical radius was r = 0.736 nm, yielding polymer configurations in van der Waals contact with the nanotube carbon atoms. For all helical initial configurations of each polymer, all phenyl rings were positioned in the same manner on each monomer; there were no "ring flips", and the aryl subunits were translationally invariant along the contour of the polymer. In building the helix, rigid body motions were used to position each monomer by equivalent ethynyl carbon positions that span each binaphthyl and phenyl subunit.

Simulations of the isolated binaphthyl units, unbridged *S*-**PBN** and bridged *S*-**PBN(b)**, in TIP3P aqueous solvent (300 K, 1 atm) were performed. The number of water molecules and atoms were: 6,871 atoms (2,267 water molecules) for *S*-**PBN**, and 5,390 atoms (1,768 water molecules) for *S*-**PBN(b)**. The length of each trajectory was 40 ns, and configurations were

sampled every 20 ps for a total of 2,000 configurations. The data were consistent with the expected properties of unbridged and bridged moieties[308]; for ***S*-PBN**, the average dihedral angle was $\varphi = 87 \pm 17\,^{\circ}$,, while for the bridged ***S*-PBN(b)**, $\varphi = 107 \pm 10\,^{\circ}$ (Figure 6-4).

### 6.2.1. Carbon Nanotube Model

An achiral (10,0) carbon nanotube, as with previous polymer/nanotube studies of PPES[128], was selected. Coordinates for the tube were generated using the VMD Nanotube Builder[309]. An ideal nanotube was used, and no relaxation of nuclear coordinates of the tube was performed. Each atom of the SWNT was parameterized as $sp^2$ carbon atoms of the CHARMm force field[286] with zero net atomic charge. The nanotube length in each simulation was 29.6 nm with an internuclear diameter of 0.793 nm. The nanotube length was more than twice that of the extended polymer, and at no point in any simulation did the polymer approach nanotube ends. The coordinates of all atoms within each nanotube were fixed in all simulations with C-C bonds constrained to their equilibrium lengths of 1.42 Å.

### 6.2.2. Helical Polymer Alignment

All helical initial structures were generated by aligning the polymer to an ideal helix having radius $r$ and pitch $p$. This helical contour $\vec{r}_h(t)$ can be defined parametrically in terms of the variable $t$, and in Cartesian coordinates $\vec{r}_h(t) = (r\cos(2\pi t), r\sin(2\pi t), pt)$. An ethynyl carbon in each monomer (see Figure 5-1) was positioned on a helical path, $\vec{r}_h(t)$, having radius which provides near van der Waals contact between polymer and nanotube carbon atoms (internuclear distance of 3.4 Å in addition to the nanotube radius). In constructing the helical conformation, rigid body motions were used to position adjacent monomers so that equivalent ethynyl carbons are at positions $\vec{r}_h(t)$ and $\vec{r}_h(t+\tau)$, respectively. The parametric one-monomer increment $\tau$ was determined by specifying that equivalent ethynyl carbons on adjacent monomers lie on the helical contour: $|\vec{r}_h(t+\tau) - \vec{r}_h(t)| = l$, where $l$ is the Euclidean distance between the equivalent ethynyl carbons (Figure 5-1). The aromatic units were rotated about an axis collinear with the ethyne bridge such that the plane of the

aromatic ring was perpendicular to the normal of the nearest point on the nanotube surface. In all cases, the helical axis is chosen as the z-axis (SWNT axis). These helical structures were minimized for 1,000 steps using the NAMD conjugate gradient energy minimization.

### 6.3. Helical Stability and Preference

Four simulations were considered: each polymer, (*S*-**PBN(b)-Ph**$_3$) and (*S*-**PBN(b)-Ph**$_5$), was initially wrapped about the SWNT in either a left-handed and or a right-handed helical conformation. Each simulation was extended to 80 ns to ensure sufficient sampling. Final configurations of each 80 ns trajectory are shown in Figure 6-5. A qualitative analysis of the trajectories indicates that the polymers remain adhered to the nanotube throughout but that neither initial conformation of the *S*-**PBN(b)-Ph**$_3$ maintained persistent helicity. The **Ph**$_3$ phenyl subunits of the polymer adhere to the tube and remain in contact with the tube's cylindrical surface. These segments can, however, locally change orientation relative to the nanotube axis. On the other hand, the *S*-**PBN(b)-Ph**$_5$ polymer initially placed in a left handed helix maintained persistent left-handed helicity for the duration of the simulation. Figure 6-6 depicts a time evolution of both polymer variants initially placed in a left handed helix; where *S*-**PBN(b)-Ph**$_3$ fluctuates to the point of losing its initial helical structure, *S*-**PBN(b)-Ph**$_5$ preserves the initial helical configuration.

For each of the four simulations, conformations of the polymers were sampled during the entirety of the 80 ns trajectories (every 20 ps for a total of 4,000 configurations) and analyzed with regard to the local orientations of the phenylene subunits from within the interior 27 subunits. The local helical wrapping angle per monomer, $\theta$, is the angle between the vector defined by the 1,4 carbon atoms in each phenyl-based subunit and the longitudinal axis of the nanotube (see Figure 6-7 for definition of vector). $\theta$ characterizes the orientation of the polymer locally on the nanotube. For an ideal helix, the relation between $\theta$ and the corresponding pitch $p$ is given by **Eq. 5-6**. Figure 6-7 depicts the evolution of $\theta$ average and standard deviation across all subunits at a given time step for each of the simulations.

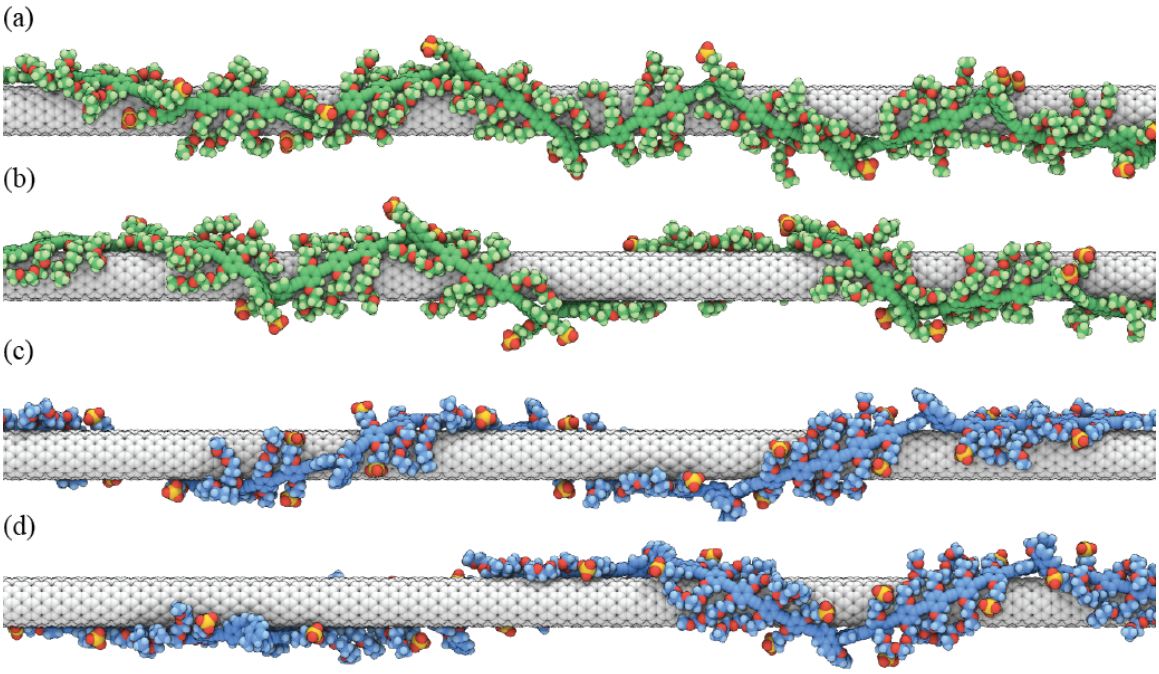



**Figure 6-5.** Final configurations of the four polymer systems. (a) *S*-**PBN(b)-Ph$_3$** initially placed in a left handed helical conformation, (b) *S*-**PBN(b)-Ph$_3$** initially placed in a right handed helical conformation, (c) *S*-**PBN(b)-Ph$_5$** initially placed in a left handed helical conformation, and (d) *S*-**PBN(b)-Ph$_5$** initially placed in a right handed helical conformation. After 80 ns, only the left handed *S*-**PBN(b)-Ph$_5$** configuration is able to maintain its helicity for the duration of the simulation.

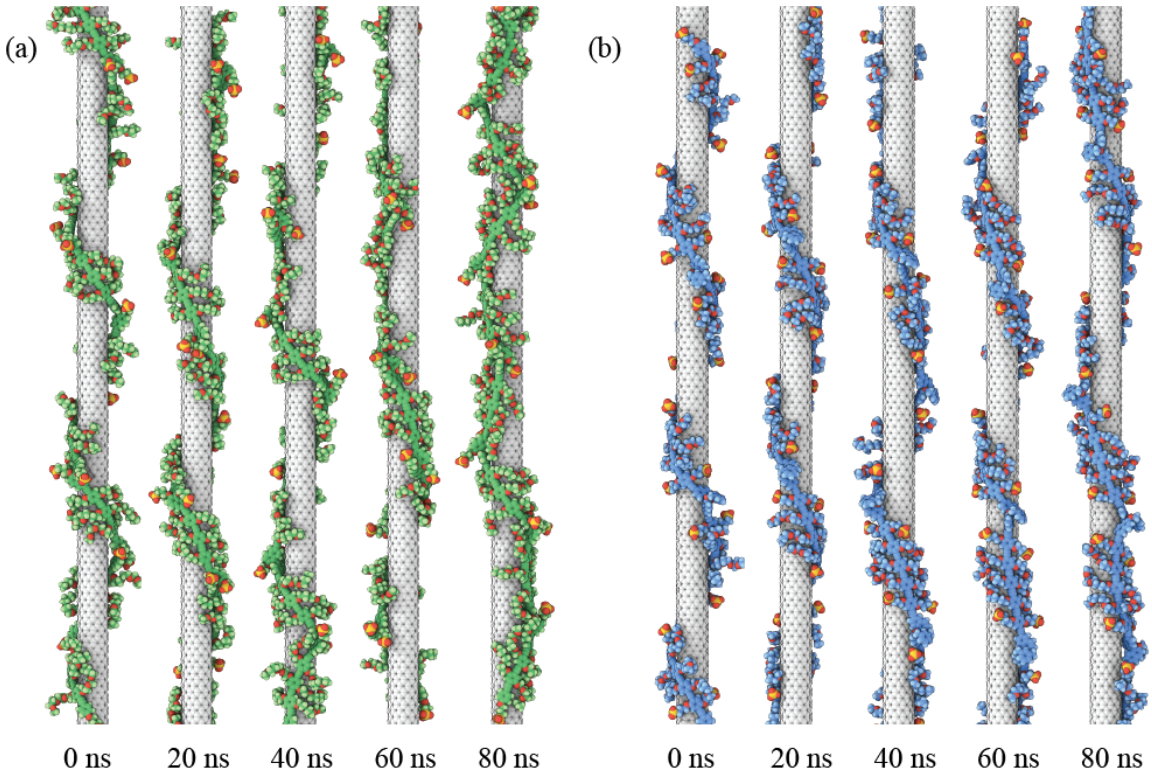



**Figure 6-6.** Time evolution at 20 ns intervals, depicting the different configurations adopted by the initially left handed helices of the two polymer variants. (a) ***S*-PBN(b)-Ph$_3$**, and (b) ***S*-PBN(b)-Ph$_5$**. Note ***S*-PBN(b)-Ph$_3$** configurations adopting a zigzag conformation, wherein the binaphthyl unit orients subsequent monomers to have values of $\theta$ that change sign.

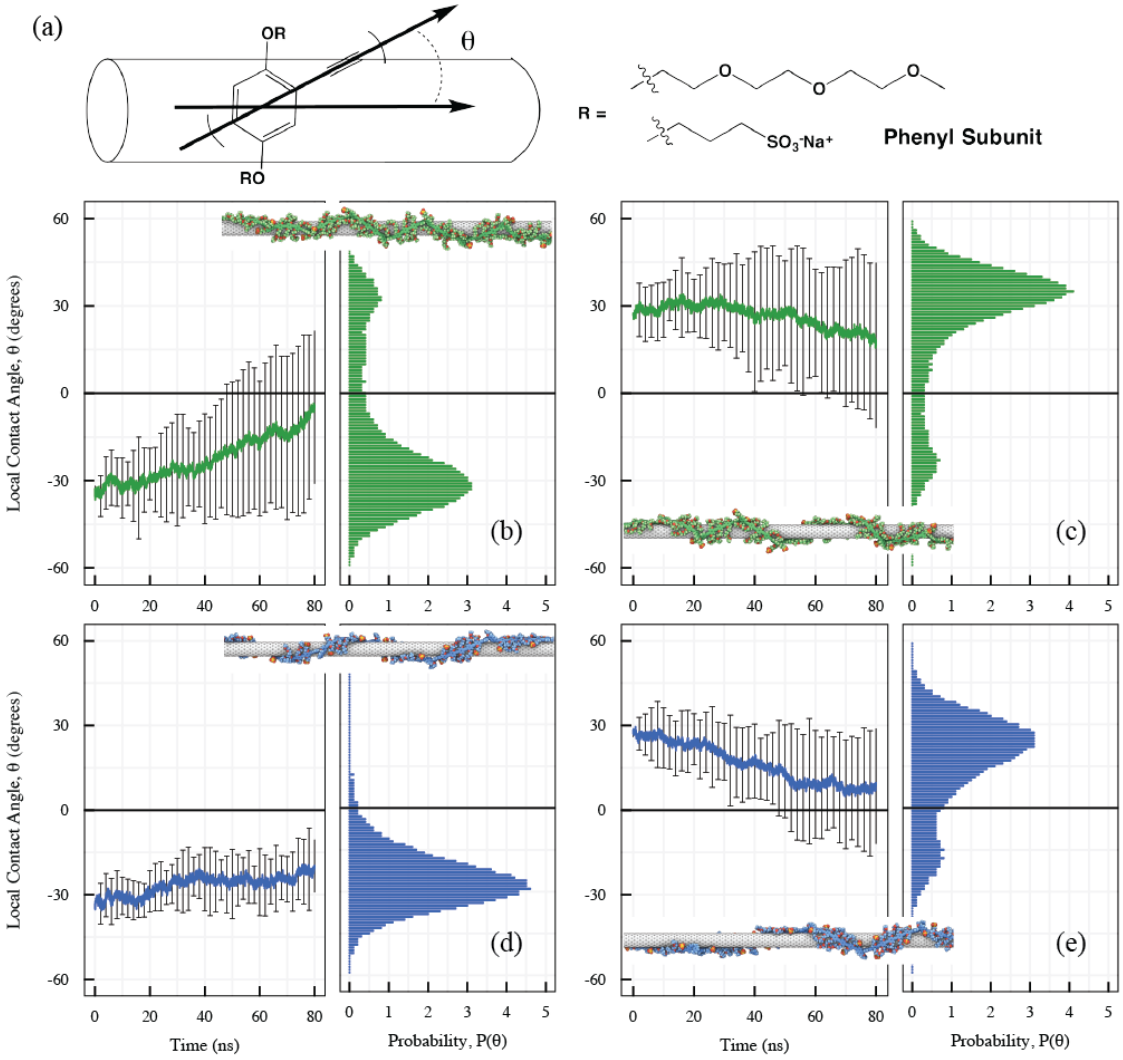



**Figure 6-7.** Evolution of average local contact angle, $\theta$, for each of the polymer-SWNT simulations, with corresponding distributions across the entire 80 ns simulation. (a) Vector description for Ph local contact orientation vector. $\theta$ is the angle between the projection of this vector on the nanotube and the nanotube axis, (b) ***S*-PBN(b)-Ph$_3$** initially placed in a left handed helical conformation, (c) ***S*-PBN(b)-Ph$_3$** initially placed in a right handed helical conformation, (d) ***S*-PBN(b)-Ph$_5$** initially placed in a left handed helical conformation, and (e) ***S*-PBN(b)-Ph$_5$** initially placed in a right handed helical conformation.

For both of the *S*-**PBN(b)-Ph$_3$** cases, $\theta$ exhibits large fluctuations and it appears the average has not yet reach a plateau after 80 ns. Where the left-handed configuration maintains a value of $\theta$ = -25 ° ± 3 ° ($p$ = 10 ± 2 nm) during the first 20 ns of the simulation, the magnitude of this mean angle continues to decrease as the simulation progresses. Likewise, the right-handed configuration maintains a value of $\theta$ = +27 ° ± 2 °($p$ = 9 ± 2 nm) for the first 20 ns of the simulation and then continues to decrease. What is apparent in each of these trajectories is the tendency for the binaphthyl units to reorient phenylene segments units into the opposite value of $\theta$, creating local zigzag-like conformations of the polymer though it remains adhered to the nanotube. This is apparent in the distribution of $\theta$ over the course of the simulation (Figure 6-7). For both of these *S*-**PBN(b)-Ph$_3$** simulations, this local contact angle $\theta$ is distributed between both the left (negative) and right (positive) orientations. A depiction of several of the zigzag features can be seen in the final configurations of Figure 6-5, where the binaphthyl moieties appear at vertices and the **Ph$_3$** segments take on alternating negative and positive values of $\theta$.

Conversely, the simulations of the *S*-**PBN(b)-Ph$_5$** polymer yield a robust helical structure and a preferred helical handedness. The initially left-handed structure maintains a persistent helical structure throughout the 80 ns simulation (Figure 6-7). This is apparent in the equilibrated average value of $\theta$ over the last 40 ns for the left-handed configuration at $\theta$ = -24 ° ± 3 °, which corresponds to a pitch of $p$ = 10 ± 2 nm. On the other hand, the initially right-handed configuration almost immediately begins to take on conformations that do not wrap the SWNT in a helical fashion, and the average value of $\theta$ drifts from the initial value of $\theta$ = +30 ° ($p$ = 8 nm) towards $\theta$ = 0 °. These observations are further corroborated by the distributions $P(\theta)$. $P(\theta)$ for the left-handed configuration contains a single maximum at $\theta$ = -29 °. $P(\theta)$ for the initially right-handed configuration contains two maxima at $\theta$ = +25 ° and $\theta$ = -17 °, consistent with the observed zigzag structure (Figure 6-5).

For the left-handed *S*-**PBN(b)-Ph$_5$** polymer, its ability to maintain a helical structure for the duration of the simulation matches the expectation that this derivative has a strong pref-

erence for this helical handedness. Since the polymer comprises multiple types of monomer unit, analyzing the trajectory via fits to ideal helical contours is nontrivial. We opt instead to consider a quantity that represents the density of polymer backbone atoms adhered to the surface of the tube. Such a quantity is also in harmony with how helical structural parameters are inferred experimentally from TEM measurements. An evaluation of a distribution pair separations (density-density correlations) parallel to the nanotube axis for all (35 subunits) backbone carbons in the polymer (aromatic phenyl, aromatic naphthyl, and ethyne carbons) shows characteristics expected for a helical configuration. The pairwise distribution function, $p(\Delta z)$, is evaluated using the following histogram function

**Eq. 6-1**

$$p(\Delta z) = \frac{\sum_{i<j} \delta(\Delta z, \Delta z_{ij}, \delta z)\delta(\alpha_i, \alpha_j, \delta\alpha)}{\sum_{i<j} \delta(\alpha_i, \alpha_j, \delta\alpha)}$$

**Eq. 6-2**

$$\text{where } \delta(x, y, \delta x) = \begin{cases} 1, & \text{for } |x - y| < \delta x \\ 0, & \text{otherwise} \end{cases}$$

where $\Delta z_{ij}$ denotes the z-coordinate difference involving atoms $i$ and $j$, $\alpha$ denotes the radial angle in the x-y plane perpendicular to the nanotube axis (see Figure 6-8), and $\delta z$ and $\delta\alpha$ are the corresponding bin sizes for $z$ and alpha respectively. Here $\delta z = 1$ Å and $\delta\alpha = 10\,^\circ$. Configurations with a helical structure will have a peak in this distribution corresponding to the value of $z$ that is the pitch of the helix, as well as values or $z$ for which are improbable if the helix is persistent and well maintained. Figure 6-8 shows this distribution for the *S*-**PBN(b)-Ph**$_5$ polymer initially placed in a left handed helix, is evaluated for the final 40 ns of the trajectory. The distribution has a maximum for $\delta z < 3$ nm, corresponding to carbon atoms within the same and adjacent monomers. The the distribution has near zero amplitude over 4-8 nm and large peak located at 10 nm, consistent with a persistent helical

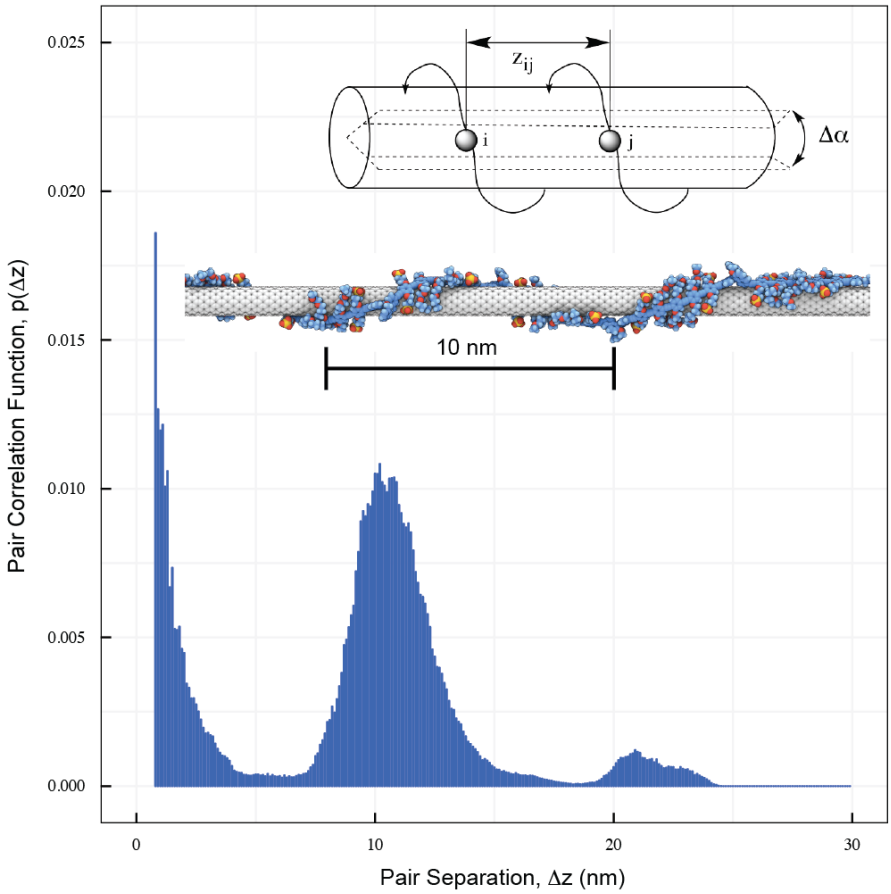



**Figure 6-8.** Pairwise linear distribution of monomer subunits within the same angular subsection of the nanotube for *S*-**PBN(b)-Ph**$_5$ initially placed in a left handed helical conformation. (subset) Depiction of the cylindrical coordinate system in which $p(z)$ is calculated. The evaluation of the delta function aligning points within the same angular section are grouped by some $\Delta\alpha$, and placed in the corresponding bin for $\Delta z$.

structure of pitch 10 nm. This is consistent with to the estimated pitch from **Eq. 5-6** ($p$ = 10 ± 2 nm) and the experimentally inferred helical pitch from the TEM micrographs.

## 6.4. Binaphthyl Dihedral Angle Distribution

The nature of the binaphthyl interior dihedral angle, $\phi$, plays a role in the helical stability of each of these chiral polymer derivatives when adsorbed to the SWNT. For the *S*-**PBN(b)-Ph**$_3$ polymer, Figure 6-9 shows both conformations have a broad distribution of allowed $\phi$ angles for all binaphthyl subunits throughout the course of the simulations. Where

the interior dihedral angle has been estimated from the CD spectra to be $\phi$ = 60-70 °, the mean of the simulations with the initially left- and right-handed configurations of the *S*-**PBN(b)-Ph$_3$** polymer are $\phi$ = 111 ° ± 14 ° and $\phi$ =106 ° ± 13 °, respectively. The distribution is broad, however, and values as low as 60 ° are sampled. Similar simulations were performed for the *S*-**PBN(b)-Ph$_3$** polymer derivative (data not shown), wherein the initial configuration of $\phi$ was set to values both greater than and less than 90 °, and all such simulations yielded distributions centered around ∼110 °. For the *S*-**PBN(b)-Ph$_5$** simulations, Figure 6-9 shows the distribution of $\phi$ is more narrow and centered upon slightly larger, more oblique angles. While the interior dihedral angle estimated from the CD spectra and TDDFT calculations suggests a narrow distribution between 95 °-100 °, the left- and right-handed *S*-**PBN(b)-Ph$_5$** simulation distributions of $\phi$ are 120 ° ± 8 ° and 114 ° ± 9 °, respectively. Here, more stringent range for the bridged derivative is apparent. Nonetheless, relative to the unbridged systems, the average value of $\phi$ is larger (more oblique) in bridged polymers for both the experimental and the simulation studies.

To address the orientation and contact of the binaphthyl units on the nanotube, an orientation angle, $\chi$, is defined between the binaphthyl bond (vector from $C_1$ to $C_1$' in the binaphthyl subunit) projected onto the surface of the nanotube and the nanotube axis. $\chi$ addresses the binaphthyl subunits' placement on the nanotube surface. For both polymers initially placed in a left-handed helix (the expected helical conformation for the *S* binapthyl enantiomer), $\chi$ fluctuates about a peak at $\chi$ = 25 °. Conversely, both systems where the polymers were initially placed in a right-handed helix exhibit an evolution of $\chi$ from close to $\chi$ = 90 ° to configurations where $\chi$ takes on smaller values and the binaphthyl bond vector is more aligned with the nanotube axis (Figures 6-10 and 6-11).

For both *S*-**PBN(b)-Ph$_3$** initial conditions, the large fluctuations and the inability to maintain a persistent helix result in part from the larger range of values accessible to $\phi$ in the unbridged systems. Where *S*-**PBN(b)-Ph$_5$** remains helical with a relatively narrow distribution of $\phi$ = 120 °, in the unbridged systems, the sampling of $\phi$ < 90 ° allows a

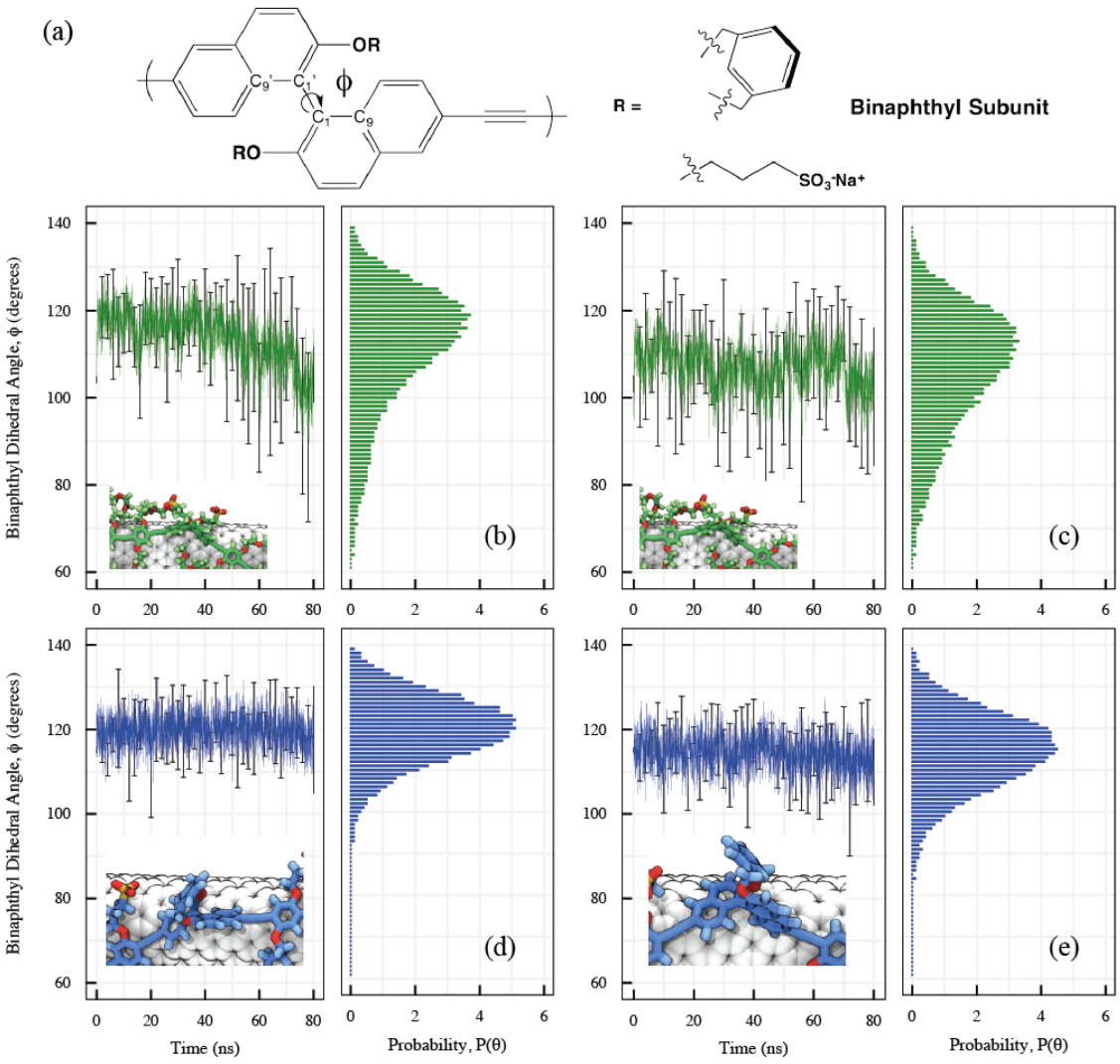



**Figure 6-9.** Evolution of the average interior binaphthyl dihedral angle, $\phi$, (Fig 2) for each of the polymer-SWNT simulations, with corresponding distributions across the entire 80 ns simulation. (a) Depiction of ***S*-PBN** dihedral angle about the binaphthyl bridge bond, $\phi$. (b) ***S*-PBN(b)-Ph$_3$** initially placed in a left handed helical conformation, (c) ***S*-PBN(b)-Ph$_3$** initially placed in a right handed helical conformation, (d) ***S*-PBN(b)-Ph$_5$** initially placed in a left handed helical conformation, and (e) ***S*-PBN(b)-Ph$_5$** initially placed in a right handed helical conformation.

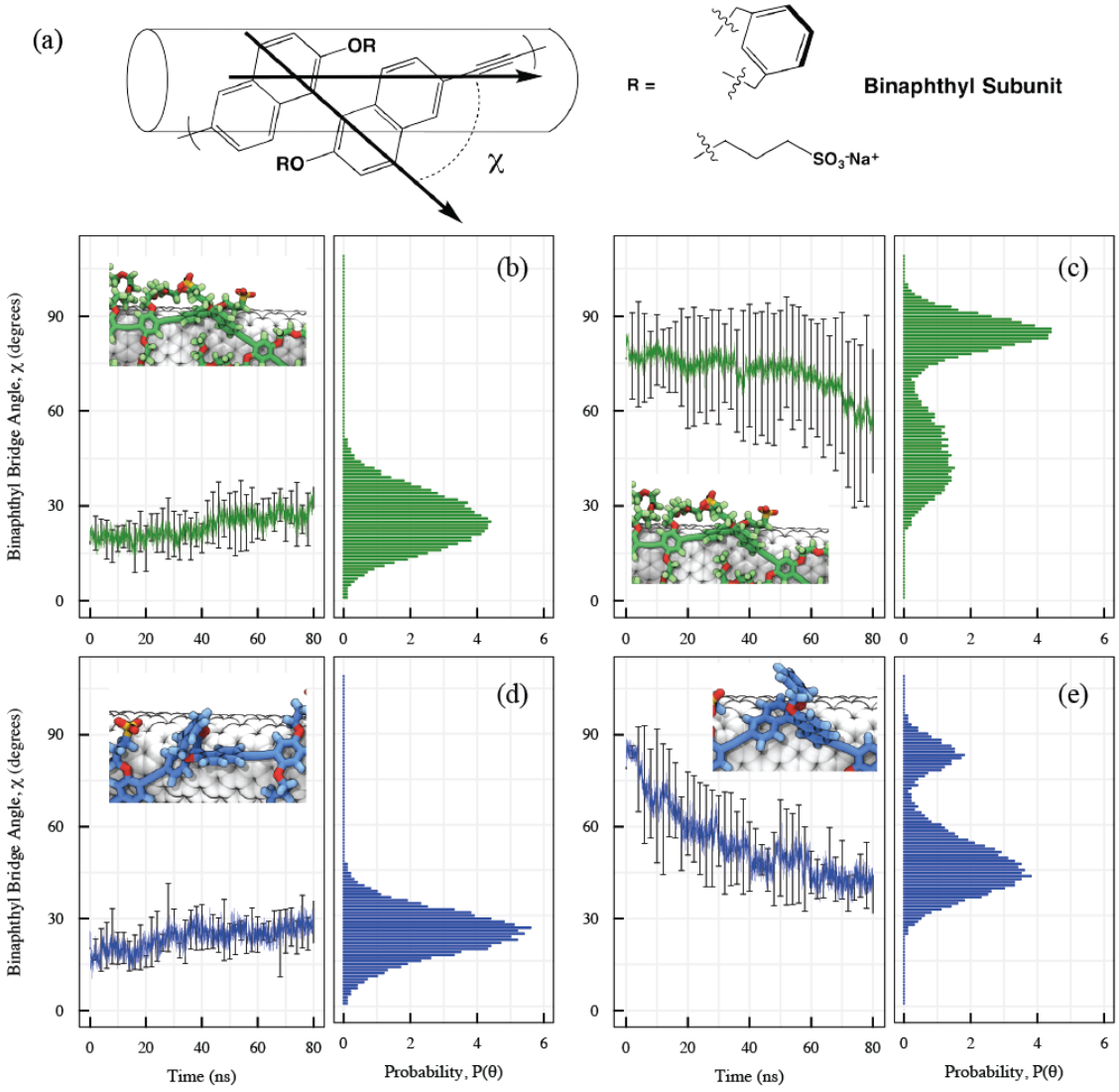



**Figure 6-10.** Evolution of the average binaphthyl bond angle with the nanotube axis, $\chi$, (Fig 2) for each of the polymer-SWNT simulations, with corresponding distributions across the entire 80 ns simulation. (a) Vector description for ***S*-PBN** local vector describing the binaphthyl bridge bond. $\chi$ is the angle between the projection of this vector on the nanotube and the nanotube axis. (b) ***S*-PBN(b)-Ph$_3$** initially placed in a left handed helical conformation, (c) ***S*-PBN(b)-Ph$_3$** initially placed in a right handed helical conformation, (d) ***S*-PBN(b)-Ph$_5$** initially placed in a left handed helical conformation, and (e) ***S*-PBN(b)-Ph$_5$** initially placed in a right handed helical conformation.

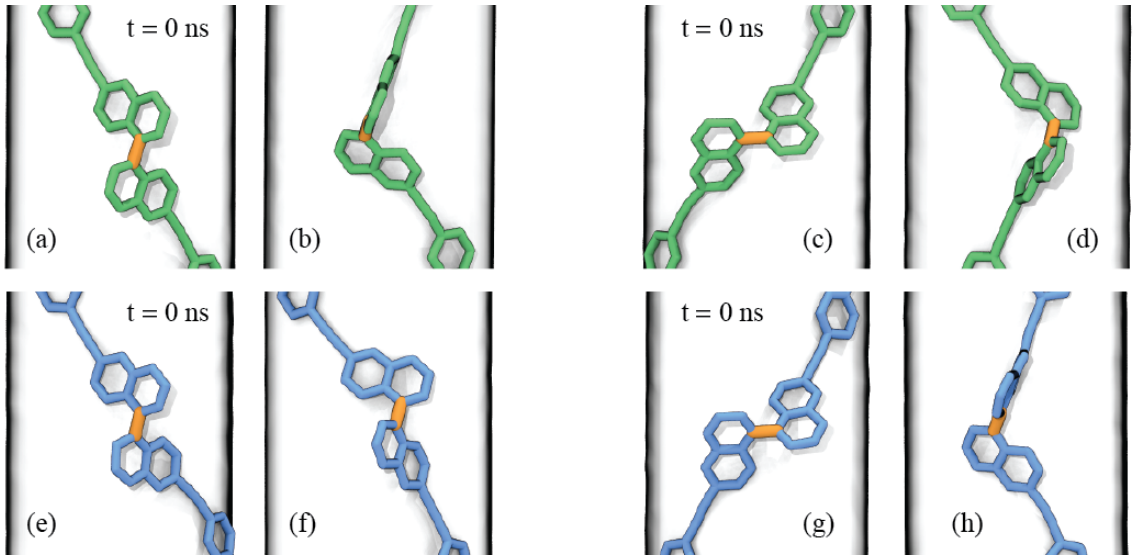



**Figure 6-11.** Configurations for the binaphthyl units in each of the polymer simulations. For all renderings, only the polymer carbon backbone is show for clarity, with the binaphthyl bridge highlighted in orange. (a) initial placement of ***S*-PBN(b)-Ph$_3$** in a left handed helix, and (b) a representative configuration from the final 40 ns. (c) initial placement of ***S*-PBN(b)-Ph$_3$** in a right handed helix, and (d) a representative configuration from the final 40 ns. (e) initial placement of ***S*-PBN(b)-Ph$_5$** in a left handed helix, and (f) a representative configuration from the final 40 ns. (g) initial placement of ***S*-PBN(b)-Ph$_5$** in a right handed helix, and (h) a representative configuration from the final 40 ns.

local contraction of the polymer at the binaphthyl unit such that the "zigzag" pattern is now accessible. Figure 6-10 illustrates this; where the left handed helix ***S*-PBN(b)-Ph$_5$** (Figure 6-10a) begins in a similar conformation to the left handed helix of ***S*-PBN(b)-Ph$_5$**, the wider distribution of $\phi$ allows for conformations like Figure 6-10b to exist. Here, the binaphthyl unit has contracted sufficiently such that the remaining phenyl portion of the polymer has slid across the nanotube surface adopting the "zigzag". As the initially right handed trajectories evolve (Figure 6-10c), $\chi$ reorients to values aligned with the nanotube (perfect alignment would be $\chi = 0\,^\circ$ or $180\,^\circ$) to foster aromatic contact of the binaphthyl with the SWNT. Such local structures can involve contact of one or both naphthyl rings with the nanotube surface. Figures 6-10d and 6-10h illustrate such configurations, where one of the naphthylene units makes contact with the nanotube surface while the other is oriented edge to face.

The simulation results support the interpretation that restraining the chiral binaphthyl subunits' conformation restricts the conformational availability of the polymer when adsorbed onto the SWNT surface. The simulations suggests that the broad range of allowable dihedral values in all four simulations is afforded by the broad internal dihedral potential, which is consistent with that studied in previous work[308]. In restraining the dihedral angle to a more oblique value ($\phi > 90\,^{\circ}$) and limiting the fluctuations to a narrower range by bridging the binaphthyl units, the polymer is able to maintain the "expected" helicity (here left-handed for the $S$ derivative) without creating any of the "zigzag" patterning both $S$-**PBN(b)-Ph$_3$** simulations obtained. Where the $S$-**PBN(b)-Ph$_3$** simulations exhibited a wide variety of conformations and general instability in the initial placement along a helical contour, the $S$-**PBN(b)-Ph$_5$** simulations are consistent with the superstructure inferred from the TEM and AFM images.

## 6.5. Conclusion

While chiral polymers, such as ssDNAs, have been established to helically wrap single-walled carbon nanotubes (SWNTs), the resulting ssDNA-SWNT hybrids manifest both right- and left-handed helical SWNT assemblies, indicating that intrinsic polymer helical chirality is in general insufficient to dictate the helical chirality of polymer-nanotube constructs. The experimental studies of the $S/R$-**PBN(b)-Ph$_3$** and $S$-**PBN(b)-Ph$_5$** polymers suggests the first general method to rigorously control the handedness of the helically wrapped polymer-SWNT superstructures. The above simulations point to the presence of interstitially placed binaphthyl units that permit these polymers to adopt directed structures when adhered to the SWNT surface. Preferred helical superstructures as adopted by these poly-arylene ethynylene polymers are influenced by the presence of charged side chains, alteration of backbone chemistry, and restriction of backbone conformation. Most notably, distinct modes associated with the polymer-SWNT complex environments have helped elucidate expectations for the increasing/decreasing of helical pitch, either through solvation or chemical modification of monomer flexibility. The design insights described herein enable enantiose-

lective control of the helical screw axis of semiconducting polymers that single-chain wrap SWNT surfaces. We posit that this work opens up new opportunities to: (i) regulate the strength of excitonic and electronic interactions between an aryleneethynylene polymer and the nanotube surface, (ii) engineer robust conjugated polymer-SWNT superstructures in which optoelectronic and chirooptic properties can be extensively modulated, and (iii) develop new approaches to organize SWNTs in the solid state.

# A | Mean Field Theory Derivations

## A.1. Enumerated Solution

If we consider an exact enumeration of the sequences in a given sequence ensemble of size $\Omega_{seq}$, it is possible to find an exact solution. Computationally, this is only limited by the tractability of the ensemble size. Starting from the sequence entropy, **Eq. 2-2**, individual sequence probabilities $W_{seq}$ can be obtained for each polymer microstate. The sequence probabilities are multivariate objects, given as

**Eq. A-1**
$$W_{seq} = W(tc_1, tc_2, ..., tc_N)$$

where each polymer microstate is defined as the set of monomer conformations for a given sequence across all positions in the polymer chain; that is, at each site $n$, a particular choice of monomer type $t$ and conformation $c$. The mean energy of the sequence ensemble, as expressed by the sequence energies in **Eq. 2-7** is given as

**Eq. A-2**
$$U_{seq} = \sum_{seq}^{\Omega_{seq}} E_{seq} W_{seq}$$

such that the effective free energy of the sequence is expressed as

**Eq. A-3**
$$F_{seq} = U_{seq} - \frac{1}{\beta}\frac{S}{k_B} = \sum_{seq}^{\Omega_{seq}} E_{seq} W_{seq} - \frac{1}{\beta}\sum_{seq}^{\Omega_{seq}} W_{seq} \ln W_{seq}$$

A minimization of this effective free energy in the context of a normalization constraint applied to the sequence probabilities,

**Eq. A-4**
$$\sum_{seq}^{\Omega_{seq}} W_{seq} = 1$$

produces an exact solution for predetermined $\beta$ in the form of

**Eq. A-5**
$$W_{seq} = \frac{\exp(-\beta E_{seq})}{\sum_{seq}^{\Omega_{seq}} \exp(-\beta E_{seq})}$$

where we can express the individual probabilities of the independent conformers as the sum over all sequence probabilities that contain the selected monomer, $t$, of conformation state $c$ at position $n$

**Eq. A-6**
$$w_{n't'c'} = \sum_{tc_1} \sum_{tc_2} ... \sum_{tc_N} W(tc_1, tc_2, ..., tc_N)\delta(ntc, n't'c')$$

where the delta function picks out sequences that contain the specified set of indices $n'$, $t'$, and $c'$. **Eq. A-6** can otherwise be written as

**Eq. A-7**
$$w_{n't'c'} = w_i = \sum_{seq \ni i}^{\Omega_{seq}} W_{seq}$$

Furthermore, obtaining an expression for the heat capacity of the mean sequence energy

is straight forward, utilizing expressions from **Eq. 2-15** and **Eq. A-21** to evaluate the derivative of the mean sequence energy as

**Eq. A-8** $$\frac{\partial U_{seq}}{\partial \beta} = \frac{\partial}{\partial \beta} \frac{\sum_{seq}^{\Omega_{seq}} E_{seq} \exp(-\beta E_{seq})}{\sum_{seq}^{\Omega_{seq}} \exp(-\beta E_{seq})} = -\sum_{seq}^{\Omega_{seq}} (E_{seq}^2 - E_{seq} \sum_{seq}^{\Omega_{seq}} E_{seq} W_{seq}) W_{seq}$$

obtaining the sequence heat capacity as proportional to the variance of that mean sequence energy

**Eq. A-9** $$C_{v,seq} = -k_b \beta^2 \frac{\partial U}{\partial \beta} = k_b \beta^2 ((\sum_{seq}^{\Omega_{seq}} E_{seq}^2 W_{seq}) - U^2) = k_b \beta^2 Var(U_{seq})$$

## A.2. Recasting the Optimization with the Method of Lagrange Multipliers

In practice, we simply solve **Eq. 2-14** utilizing standard nonlinear optimization techniques. However, it is sometimes useful to partially solve the set of equations cast in the traditional Gibbs form to obtain new constraints or metrics (i.e., the mean field energy derivative). To do so, we treat the minimization of the effective free energy using the method of Lagrange multipliers by defining the variation functional with objective function $-\beta F(\mathbf{w})$

**Eq. A-10** $$V(\vec{w}, \vec{\alpha}, \vec{\lambda}) = -\sum_i w_i \ln w_i - \beta(U(\vec{w})) - \sum_n \alpha_n (\sum_{tc} w_i - 1) - \sum_k \lambda_k (f_k(\vec{w}) - f_k^o)$$

for normalization constraints and arbitrary constraints, $f_k(\vec{w})$, and their corresponding Lagrange multipliers, $\alpha_n$ and $\lambda_k$, respectively. The optimum is obtained at the stationary

point of $V(\vec{w})$, that is, where the gradient is a zero vector.

**Eq. A-11**

$$\vec{\nabla} V(\vec{w}, \vec{\alpha}, \vec{\lambda}) = \vec{0}$$

Gradient elements of $V$ with respect to probability $w_i$ are expressed as

**Eq. A-12**

$$\frac{\partial V}{\partial w_i} = -\ln w_i - 1 - \alpha_n - Q_i = 0$$

and we define $Q_i$ as

**Eq. A-13**

$$Q_i \equiv \beta \frac{\partial U}{\partial w_i} + \sum_k \lambda_k \frac{\partial f_k}{\partial w_i}$$

noting that our definition of $Q_i$ provides a way to summarize any non entropic contributions to $V$. Rearranging **Eq. A-12** obtains,

**Eq. A-14**

$$w_i = \exp(-1 - \alpha_n) \cdot \exp(-Q_i)$$

Gradient elements of $V$ with respect to the normalization Lagrange multipliers are expressed

as

**Eq. A-15**

$$\frac{\partial V}{\partial \alpha_n} = \sum_{tc} w_i - 1 = 0$$

Applying this definition to these normalization constraints solves for the normalization Lagrange multipliers and removes them from the system of equations, providing a definition for a sequence partition function, $z_i$,

**Eq. A-16**

$$\sum_{tc} w_i = \sum_{tc} \exp(-1 - \alpha_n) \cdot \exp(-Q_i) = 1$$
$$\exp(-1 - \alpha_n) = \frac{1}{\sum_{tc} \exp(-Q_i)} \equiv \frac{1}{z_i}$$

and finally providing a series of equations for the individual probabilities.

**Eq. A-17**

$$w_i = \frac{\exp(-Q_i)}{z_i}$$

For the particular instance of no additional constraints to the normalization, the local energy definition arises naturally. Eq. A-13 reduces to

**Eq. A-18**

$$Q_i = \beta \frac{\partial U}{\partial w_i} \equiv \beta \epsilon_i \qquad \text{thus} \qquad w_i = \frac{\exp(-\beta \epsilon_i)}{z_i}$$

This equation includes $\epsilon_i$, the mean local energy of $i$, which we define in **Eq. 2-9** when establishing the mean field energy **Eq. 2-12**.

Again, we note that **Eq. A-17** provides insight into the theory's analog to statistical mechanics, and that in practice solving these equations with self-consistent techniques is inadequate when dealing with possible nonlinearities in the energy function and associated with constraints.

## A.3. Heat Capacity

As usual, the heat capacity is defined as

**Eq. A-19**
$$C_v = \frac{\partial U}{\partial T} = -k_b \beta^2 \frac{\partial U}{\partial \beta}$$

Obtaining an expression for the heat capacity of the mean field energy as defined in **Eq. 2-15** requires obtaining a partial derivative of **Eq. 2-12** with respect to $\beta$,

**Eq. A-20**
$$\frac{\partial U}{\partial \beta} = \sum_i \gamma_i \frac{\partial w_i}{\partial \beta} + \frac{1}{2} \sum_{ij} \gamma_{ij} (w_i \frac{\partial w_j}{\partial \beta} + \frac{\partial w_i}{\partial \beta} w_j)$$

Utilizing the self-consistent equation established in **Eq. A-17**, we evaluate the partial derivative of $w_i$ with respect to $\beta$

**Eq. A-21**
$$\frac{\partial w_i}{\partial \beta} = \frac{1}{z_i^2} \left[ \frac{\partial \exp(-Q_i)}{\partial \beta} z_i - \exp(-Q_i) \frac{\partial z_i}{\partial \beta} \right]$$

Evaluating both derivatives reduces to

**Eq. A-22**
$$\frac{\partial w_i}{\partial \beta} = -w_i \frac{\partial Q_i}{\partial \beta} - \frac{w_i}{z_i} \sum_{tc} -\frac{\partial Q_i}{\partial \beta} \exp(-Q_i) = -w_i \left( \frac{\partial Q_i}{\partial \beta} - \sum_{tc} \frac{\partial Q_i}{\partial \beta} w_i \right)$$

In assuming that the first of the $k$ constraints is the mean field energy function it is clear that we simply recover the form presented in **Eq. A-13**. We evaluate the derivative of $Q_i$ with respect to $\beta$, choosing to ignore any of the mixed partial derivatives.

**Eq. A-23**
$$\frac{\partial Q_i}{\partial \beta} = \frac{\partial U}{\partial w_i} + \beta \frac{\partial^2 U}{\partial w_i \partial \beta} + \sum_k \lambda_k \frac{\partial^2 f_k}{\partial w_i \partial \beta} \approx \frac{\partial U}{\partial w_i}$$

arriving at a derivative of the probability $w_i$ with respect to $\beta$ that evaluates to a weighted difference between the local mean energy at $i$ and the site mean energy across all monomer types and conformers at that site.

**Eq. A-24**
$$\frac{\partial w_i}{\partial \beta} = -w_i \left( \epsilon_i - \sum_{tc} \epsilon_i w_i \right) \qquad \text{for} \qquad \epsilon_i \equiv \frac{\partial U}{\partial w_i} = \gamma_i + \sum_j \gamma_{ij} w_j$$

Inserting this definition into **Eq. A-20**, we are able to recover the mean sequence energy

derivative with respect to $\beta$ corresponding to the sum of all polymer site variances.

**Eq. A-25**

$$\frac{\partial U}{\partial \beta} = -\left[\sum_i \gamma_i w_i \left(\epsilon_i - \sum_{tc} \epsilon_i w_i\right) + \frac{1}{2}\sum_{ij} \gamma_{ij} w_i w_j \left(\left(\epsilon_j - \sum_{tc} \epsilon_j w_j\right) + \left(\epsilon_i - \sum_{tc} \epsilon_i w_i\right)\right)\right]$$
$$= -\sum_i \left({\epsilon_i}^2 - \epsilon_i \sum_{tc} \epsilon_i w_i\right) w_i = -\sum_s^S \left[\sum_{tc} {\epsilon_i}^2 w_i - \left(\sum_{tc} \epsilon_i w_i\right)^2\right] = -\sum_s^S Var(\epsilon_s)$$

The final expression, limited by the choice made in **Eq. A-23**, looks in many ways similar to the traditional expression for the heat capacity in statistical mechanics.

**Eq. A-26**

$$C_v = k_b \beta^2 \sum_n^N \left[\sum_{tc} {\epsilon_i}^2 w_i - \left(\sum_{tc} \epsilon_i w_i\right)^2\right] = k_b \beta^2 \sum_n^N Var(\epsilon_n)$$

## A.4. Entropically-Normalized Monomer Type Probabilities

In the absence of any energetic constraints (or an infinite temperature ensemble), a discrepancy between monomer type probabilities by simply aggregating conformational probabilities (Eq. 2-16) is clear should $C_{nt}$ differ for indices $n$ and $t$.

**Eq. A-27**

$$w_{(n,t)}\Big|_{S_{max}} = \sum_c^{C_{nt}} \frac{1}{(\sum_t^{T_n} C_{nt})} = \frac{C_{nt}}{\sum_t^{T_n} C_{nt}} \qquad \text{given} \qquad w_i\Big|_{S_{max}} = \frac{1}{\sum_t^{T_n} C_{nt}}$$

In order to recover monomer type probabilities that are independent of the number of allowable conformational states, we weight each monomer type probability by a conditional monomer type entropy, $S_{nt}$, as to recover the property that at the infinite temperature

solution, all $w_{(n,t)}$ are equal.

**Eq. A-28** $$\tilde{w}_{(n,t)} = w_{(n,t)} \cdot A \cdot \exp(-\frac{S_{nt}}{k_b}) \qquad \text{for} \qquad \frac{S_{nt}}{k_b} = -\sum_{c}^{C_{nt}} \frac{w_i}{w_{(n,t)}} \ln \frac{w_i}{w_{(n,t)}}$$

These weighted probabilities are then normalized

**Eq. A-29** $$\sum_{t}^{T_n} \tilde{w}_{(n,t)} = 1 \qquad \text{thus} \qquad A = \left( \sum_{t}^{T_n} w_{(n,t)} \cdot \exp(-\frac{S_{nt}}{k_b}) \right)^{-1}$$

yielding

**Eq. A-30** $$\tilde{w}_{(n,t)} = \frac{w_{(n,t)} \cdot \exp(\frac{S_{nt}}{k_b})}{\sum_{t}^{T_n} w_{(n,t)} \cdot \exp(\frac{S_{nt}}{k_b})}$$

The definition affords the appropriate limits associated with this reweighting. For the infinite temperature sequence-state entropy maximum, we obtain conformational independence for each type probability; that is, the probabilities of each type at position $n$ on the polymer are equal.

**Eq. A-31** $$\tilde{w}_{(n,t)}\Big|_{S_{max}} = \frac{\frac{1}{\sum_{t}^{T_n} C_{nt}}}{\sum_{t}^{T_n} \left( \frac{1}{\sum_{t}^{T_n} C_{nt}} \right)} = \frac{1}{\sum_{t}^{T_n}} = \frac{1}{T_n}$$

In the limit of the zero temperature solution, (i.e., the reduction to a single conformer

identity and zero probabilities for all others), the reweighted type probability reduces exactly to the original summation over conformational probabilities, being either identity or zero.

**Eq. A-32** for $w_{(n,t)} = \begin{cases} 1 \\ 0 \end{cases}$ , $\tilde{w}_{(n,t)} = \dfrac{w_{(n,t)} \cdot \exp(0)}{\sum_{t}^{T_n} w_{(n,t)} \cdot \exp(0)} = \dfrac{w_{(n,t)}}{\sum_{t}^{T_n} w_{(n,t)}} = w_{(n,t)} = \begin{cases} 1 \\ 0 \end{cases}$

# B Vergil Implementation Details

The methodology detailed in Chapter 2 was implemented in the completely reorganized software package VERGIL (previously versions were referred to as SCADS – Statistical Computationally Assisted Design Strategy). Architecting the VERGIL suite placed emphasis on the utilization of parallel computing, the ability to run calculations on laptops and desktops to massively parallel machines, and user interfaces in the Tcl scripting languages. However, the most critical aspects of the software are the internal data structures and routines housed in the Domain and the Optimizer. The Domain is responsible for the atomic organization of the ensemble of sequences considered, containing positional and chemical information. The Optimizer is the work-horse, which handles the probability functions of interest and solves for the desired sequence probabilities. Below we discuss both, including associated data structures, interfaces, and general implementation details.

## B.1. Domain Levels and Transversal

The main organization of the molecular system subject to design lies within the implementation of the domain; it is here that information relevant to the system variability is stored, from molecular segregation of motifs down to the atomistic coordinates. The domain serves as the physical location for the molecular ensembles in question, organized in a tree where depth of traversal corresponds to the level of detail.

Each of the levels of the domain has been derived from the template class **Container**. This class acts as the base class for each of the domain levels. Their commonality lies in that each has an ElementType – of which that level contains a collection – and a ParentType – of which that level belongs to.

The container class is meant to provide a general interface for any level of the domain. For example, any container can add new children, add a copy of a child, delete children, etc.

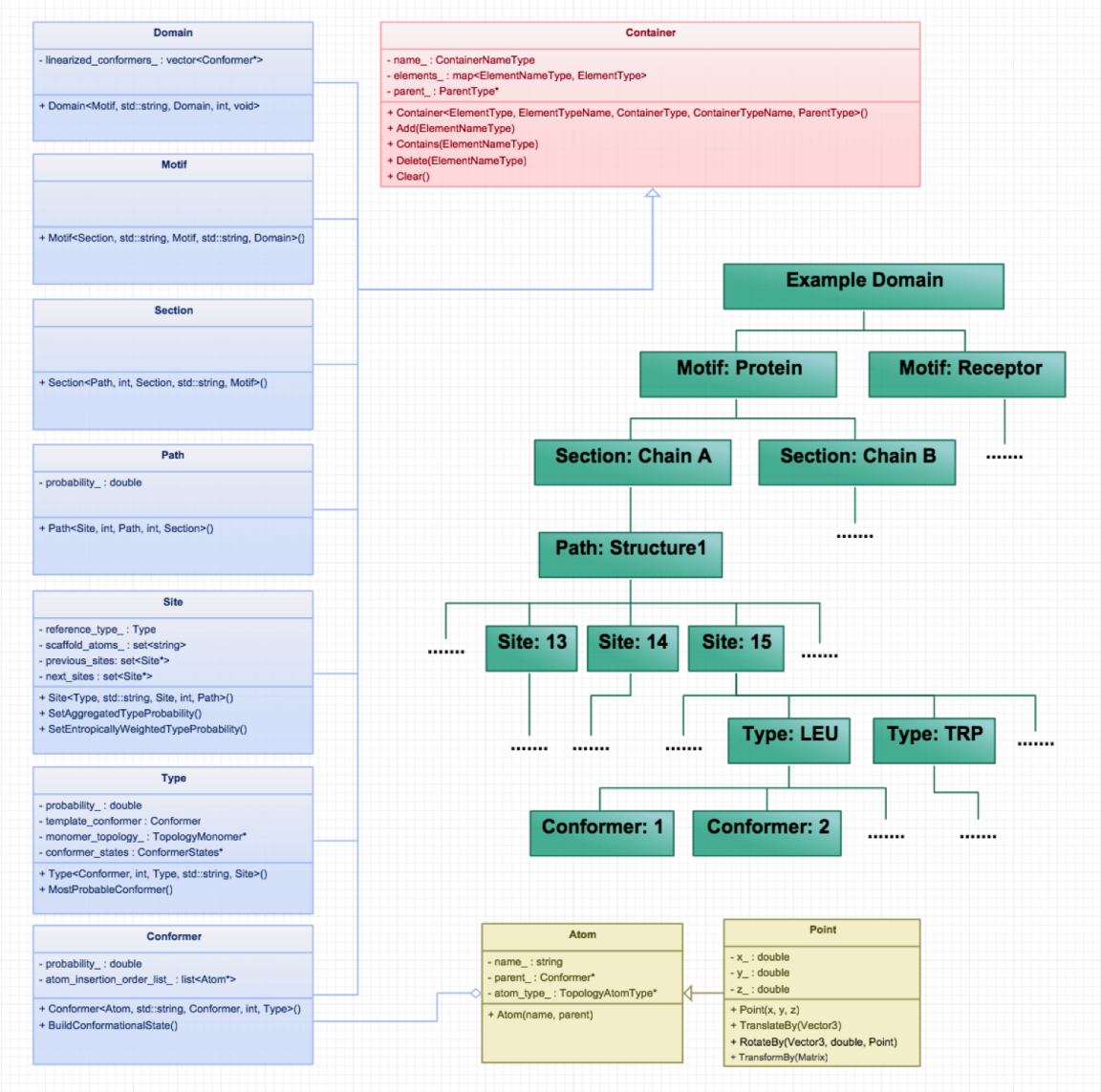



**Figure B-1.** Simple UML diagram indicating the Domain organization. All levels of the Domain (blue) inherit from the Container Interface (red), with the exception of the Atom class (yellow). An example domain is depicted to indicate how a particular protein structure might be organized into this tree system (green).

Perhaps the most useful of these is the ability to access a child element with the accessor [ ] operator, or evaluate if a container contains a particular child. Additionally, each container is equipped with an iterator (and a corresponding const iterator) to iterate through its children.

More importantly, all levels of the domain are equipped with iterators through all sublevels through the use of the ContainerIterator template class. That is, while a Container can iterate through all its children, it can use the ContainerIterator to also iterate through subchildren, subsubchildren, and so on. For example, the domain is equipped with iterators through all levels, down to an iteration through all atoms contained within the domain. This is equivalent to iterating in a depth first search manner, thus providing efficient shorthand that is useful through the library.

The domain is organized into eight levels as indicated in Figure B-1. A general domain is composed of a series of motifs (eg. different protein chains, a nonbiological cofactor, etc.). Each motif is composed of sections that divide the motif into discrete units, and in turn each section can have multiple paths (pathways) for the scaffolding of the design. This is done so that within a motif, the interface between connected sections is fixed as to provide continuity between the possibility of alternate pathways. Each path has an associated probability that provides a means of determining the most probable pathway/structure. For design work where only one backbone path is considered, this probability is always 1.

A given path is composed of a series of sites (residue locations). Sites are explicitly connected with pointers to indicate molecular connections, creating a directed graph throughout the domain. Additionally, a site has a reference child which guarantees molecular information when no ensemble variability is present. Each site is composed of a series of types (monomer types allowed to exist in the ensemble at that site). A type object is a container of conformers, and includes a default conformational state (conformer) when no explicit conformers are specified. Additionally, a type object has an associated probability to specify the most probable sequence through the domain. Conformer objects are specific conformations of a given

type at a given site location, each with an associated probability. Finally, all conformers are composed of a collection of atom objects.

The **Atom** class, the lowest level of the domain, is an extension of a Point object. An atom stores the cartesian coordinates it occupies, has an atom name, a pointer to its atom type in the topology libraries, and like all levels of the domain, a pointer to its parent – in this case its parent conformer. We have provided an additional method for Atoms that attempts to evaluate the smallest number of bonds between the atom and some other atom in the domain (used in nonbonding interactions, etc.). We note that this only works for atoms that belong to conformers that belong to a type that has been appropriately linked to the topology library. This is due to the routine used for evaluating the smallest number of bonds between atoms according to the molecular topology in the topology library.

## B.2. Generalized Probability-based Functions

As discussed in Chapter 2, the basis for the thermodynamic quantities lies in the individual probabilities of the ensemble of conformers. Recall the probability of the $i$th conformer being $w_i = w_{(n,t,c)}$, corresponding to site $n$, type $t$, and conformational state $c$. These become the free variables used to evaluate the various quantities defined throughout Chapter 2, and are described below.

The implementation of the function class is implemented as a base abstract Function class (defined in **function.h**), for which the function interface is defined. The three main class methods are **Value**, **FirstDerivative**, and **SecondDerivative** (Listing B.1) each of which take a vector **w** (and partial derivative indices) and return the value of the function at that point. **Value** must be defined for each of the functions, but the derivative methods have been provided with default forward-finite differences routines should an exact formula be difficult to arrive at.

**Listing B.1.** Function Interface

```
virtual double Value(const Vector &x) = 0;
virtual double FirstDerivative(const Vector &x, size_t i);
virtual double SecondDerivative(const Vector &x, size_t i, size_t j);
```

The abstract Function class also provides a means of managing which indices of the vector **w** are included in the evaluation of the member methods, as well as interfaces that must be defined for adding, removing, and updating those indices. This creates an efficient lookup for functions which only include a specific set of variables (eg., SiteProbability or Composition). The following sections discuss the implementation of commonly used functions and their corresponding equations.

### B.2.1. Conformational Entropy

The total entropy of the sequence conformational states is defined in **Eq. 2-5** as the total entropy over the Boltzmann factor. To keep the quantity unitless, we divide through by the gas constant, $R$, when the entropy is in per molar values.

**Eq. B-1**

$$\frac{S}{R} = -\sum_{i}^{N} w_i \ln w_i$$

This function utilizes all $i$ in the summation, and the implementation is straightforward. The evaluation of the first partial derivative with respect to the $i$th variable is given as

**Eq. B-2**

$$\frac{\partial \frac{S}{R}}{\partial w_i} = -\ln w_i - 1$$

while the second mixed partial derivative with respect to the $i$th and $j$th variables is given

as

**Eq. B-3**

$$\frac{\partial^2 \frac{S}{R}}{\partial w_i \partial w_j} = \begin{cases} -\frac{1}{w_i} & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

The implementation of these is found in **function_entropy.cpp**.

It is worth noting that the definitions above are real only for positive values of the probabilities. While solutions to the entropic maximization problem should not include negative probabilities, solvers may not necessarily be bounded by $[0, 1]$. For each of the above definitions (the real parts of the entropy), we also define the corresponding imaginary definitions. The imaginary part of the function is given as

**Eq. B-4**

$$Im(\frac{S}{R}) = \begin{cases} -\sum_i^N \pi * w_i & \text{for } w_i < 0 \\ 0 & \text{for } w_i \geq 0 \end{cases}$$

which is nonzero for negative probabilities. The corresponding first derivative is given as

**Eq. B-5**

$$Im(\frac{\partial \frac{S}{R}}{\partial w_i}) = \begin{cases} -\pi & \text{for } w_i < 0 \\ 0 & \text{for } w_i \geq 0 \end{cases}$$

while the imaginary part of the second derivative is always zero.

**Eq. B-6**

$$Im(\frac{\partial^2 \frac{S}{R}}{\partial w_i \partial w_j}) = 0 \qquad \text{for all } w_i$$

These metrics have been helpful for identifying solutions of the entropy maximization problem where the normalization constraints have been satisfied, but contain negative probabilities. The class as implemented will output the imaginary part of the entropy should it exist, as a way to alert the user that conformers have negative probability values.

### B.2.2. Site Probability

The critical constraint to the entropy maximization problem is the normalization of the conformer probabilities. As described in Section 2.2, the normalization occurs across all conformers at a particular site, invariant of their type. We define a function to perform such a summation (which we will later constrain to be 1 in the context of the optimization problem) as given by

**Eq. B-7**

$$f = \sum_{t}^{T_n} \sum_{c}^{C_{nt}} w_i$$

Because the function is linear in $w_i$, its first partial derivative is constant and is nonzero for variables that exist at site $n$,

**Eq. B-8**

$$\frac{\partial f}{\partial w_i} = 1 \qquad \text{for } i \in \text{site } n$$

and its second mixed partial derivative is always zero

**Eq. B-9**

$$\frac{\partial^2 f}{\partial w_i \partial w_j} = 0$$

All implementations can be found in **function_siteprobability.cpp**.

### B.2.3. Mean Field Energy

Section 2.2 describes the formulation for a mean field based approach to obtaining an energy for the sequence ensemble. The implementation of **Eq. 2-12** is defined as a function of energy coefficients in both one and two indices ($\gamma_i$, $\gamma_{ij}$). These coefficients are defined by a sum of potential compute objects passed on construction, such that for any conformer $i$, $\gamma_i$ is defined by the **InterConfomer** routine, and for any pair of conformers $i$ and $j$, $\gamma_{ij}$ is defined by the **IntraConfomer** routine, as summed across those potential functions. The mean field energy, a quadratic function in the probabilities, is thus given as

**Eq. B-10**

$$U = \sum_i \gamma_i w_i + \frac{1}{2} \sum_{ij} \gamma_{ij} w_i w_j$$

where the first partial derivative is defined as

**Eq. B-11**

$$\frac{\partial U}{\partial w_i} = \gamma_i + \frac{1}{2} \sum_j (\gamma_{ij} + \gamma_{ji}) w_j$$

and the second mixed partial derivative is defined as

**Eq. B-12**

$$\frac{\partial^2 U}{\partial w_i \partial w_j} = \frac{1}{2}(\gamma_{ij} + \gamma_{ji})$$

It is prudent to note that the above definitions explicitly take into account the difference in the quadratic coefficients $\gamma_{ij}$ and $\gamma_{ji}$ (as is necessary for the definition of the lattice energy, see Section 2.2.3). However, given our simple definition of the energy coefficients, we can reduce **Eq. B-11** and **Eq. B-12** by taking advantage of the inversion symmetry in $i$ and $j$ when considering the pairwise interactions. This means that the energy interaction of conformer $i$ with conformer $j$ is equivalent to the energy interaction of conformer $j$ with conformer $i$. Because of this, the evaluation of the energetic coefficients is reduced considerably as detailed in Listing B.2.

This then affords a reduced form of **Eq. B-11** and **Eq. B-12** which incorporates this inversion. The first partial derivative of the mean field energy then becomes

**Eq. B-13**

$$\frac{\partial U}{\partial w_i} = \gamma_i + \sum_j \gamma_{ij} w_j$$

and second mixed partial derivative

**Eq. B-14**

$$\frac{\partial^2 U}{\partial w_i \partial w_j} = \gamma_{ij}$$

The specific implementation of the mean field energy utilize the **Eq. B-11** and **Eq. B-12** in the class implementation in **function_meanfieldenergy.cpp**.

**Listing B.2.** One- and Two-Body Energies

```
//One-Body Energies
for (size_t i = 0; i < n; ++i) {
  for (list<Compute*>::const_iterator it = computes_.begin(); it !=
      computes_.end(); ++it) {
    one_body_energies_[i] += (*it)->Intra(*conformers[i]);
  }
}
//Two-Body Energies
for (size_t i = 0; i < n; ++i) {
  for (size_t j = i + 1; j < n; ++j) {
    for (list<Compute*>::const_iterator it = computes_.begin(); it !=
        computes_.end(); ++it) {
      two_body_energies_(i, j) += (*it)->Inter(*conformers[i], *conformers[j
          ]);
    }
    two_body_energies_(j, i) = two_body_energies_(i, j);
  }
}
```

### B.2.4. Sequence Conformational Free Energy

As defined in Section 2.2, it is often desirable to maximize the entropy function in the context of the mean field energy constraint at a particular value of the mean field energy Lagrange multiplier, $\beta$ (here denoted as the effective temperature). To do so, we create a new function, taking the form of a sequence conformational free energy fixed at a particular value of $\beta$.

**Eq. B-15**
$$F = U - TS = U - \frac{1}{\beta}\frac{S}{R} = U - RT\frac{S}{R}$$

As seen here, this is just a linear combination of the previously defined functions of the mean field energy (Eq. B-10) and the sequence conformational entropy (Eq. B-1). Because of that, the partial derivatives are trivial to define, in the sense that we have already done so. The implementation of this function simply sums function calls from each of those classes. The first partial derivative is given as

**Eq. B-16**

$$\frac{\partial F}{\partial w_i} = \frac{\partial U}{\partial w_i} - RT\frac{\partial \frac{S}{R}}{\partial w_i}$$

and second mixed partial derivative as

**Eq. B-17**

$$\frac{\partial^2 F}{\partial w_i \partial w_j} = \frac{\partial^2 U}{\partial w_i \partial w_j} - RT\frac{\partial^2 \frac{S}{R}}{\partial w_i \partial w_j}$$

Each implementation can be found in **function_freeenergy.cpp**.

## B.3. The Optimization Interface

The main purpose of Vergil is find a solution to the entropy maximization/free energy minimization problem detailed in Chapter 2; that is, to solve constrained optimization problems using the domain-based probability functions defined in Section B.2. Vergil is organized in a way that allows the definitions of these optimization problems to be formed in a modular way: adding and deleting constraints, changing the objective function, and finding the stationary point of a partially solved problem are just a few such examples.

The primary means by which to interface the probability-based functions with optimization routines is the **Problem** class. The class is composed of six private data members as detailed in Listing B.3. The n-dimensional point **x** stores some value which serves as a means of accessing values of the problem. Vectors detailing the upper and lower boundaries

are included for each variable in **x**, which default to the largest positive and lease negative double values respectively. A pointer to an objective function to optimize is stored, along with a map of Constraints indexed by a string key (id for the constraint).

**Listing B.3.** Problem Data Memebers

```
Vector x_;
Vector lower_boundary_;
Vector upper_boundary_;
Function* objective_function_;
std::map<std::string, Constraint> constraints_;
```

The problem class provides an interface for the general form of an optimization problem: minimize some objective function $f(x)$ subject to equality constraints $g_k$ and boundaries on the variables $x$.

**Eq. B-18**

$$\begin{gathered} \min f(\vec{x}) \\ \text{subject to } g_k = g_k^o \\ \text{bounded by } x_L \leq x \leq x_U \end{gathered}$$

Constraints are handled by the routines listed in Listing B.4. To impose a constraint on the problem, one simply calls the **AddConstraint** method, passing a string identifier for that constraint, a pointer to the constraint function ($g_k$), and the constrained value ($g_k^o$). As constraints are organized in a map indexed by string ids making search or deletion simple map operations. Additionally, the class provides a means to obtain a vector of all constraints (maintained as insertion order) for ease of use in the optimizer classes.

**Listing B.4.** Optimizer Interface for Constraints

```
void AddConstraint(std::string id, Function* function, double target);
Constraint& constraint(std::string id);
void DeleteConstraint(std::string id);
const std::vector<Constraint*>& constraints();
```

The **Constraint** object is organized as a combination of a function pointer and target value. The class also supports booleans indicating whether the constraint is an equality, greater than, or less than constraint. It is important to note is the various public methods of the constraint class implement manipulations of the **primal form** of a constraint, indicated in Listing B.5.

**Listing B.5.** Constraint Methods

```
double Value(const Vector &x);
double PrimalValue(const Vector &x);
double PrimalFirstDerivative(const Vector &x, size_t i);
double PrimalSecondDerivative(const Vector &x, size_t i, size_t j);
bool IsPrimalFirstDerivativeZero(size_t i);
bool IsPrimalSecondDerivativeZero(size_t i, size_t j);
```

Calculating the value of the constraint at some point $x$ is simply evaluating the constraint function's **Value** method, $g_k(x)$.

This is contrasted by evaluating the primal value of the constraint, through **PrimalValue**, which is the value of the constraint function minus the target value. This is the same as evaluating how close the constraint is to being satisfied.

**Eq. B-19** $$g_k(x) - g_k^o(x)$$

Additionally, both the first and second partial derivatives of the primal value of the constraint are supplied by **PrimalFirstDerivative**

**Eq. B-20**

$$\frac{\partial}{\partial w_i} g_k(x) - g_k^o(x) = \frac{\partial g_k(x)}{\partial w_i}$$

and **PrimalSecondDerivative**.

**Eq. B-21**

$$\frac{\partial^2}{\partial w_i \partial w_j} g_k(x) - g_k^o(x) = \frac{\partial^2 g_k(x)}{\partial w_i \partial w_j}$$

Lastly, as with the definitions for all **Function** classes, methods are provided to assess if the first or second partial derivatives will always be zero for certain indices with **IsPrimalFirstDerivativeZero** and **IsPrimalSecondDerivativeZero**.

The interface for obtaining information about the objective function in the problem is provided in Listing B.6, which gives the user access to the value of the objective function, the gradient of that function, and its Hessian.

**Listing B.6.** Problem Interface

```
double ObjectiveFunctionValue();
void Gradient(Vector* gradient);
void Hessian(Matrix* hessian);
```

To handle the specific case of optimization problems working on functions given in the theory chapter, there is the derived class **ProblemDomainProbability**, which considers all functions provided to work on the set of conformer probabilities associated with the domain. This class is unique in its ability to get and set probabilities from corresponding

conformers in the domain. Upon instantiation, this class sets the lower and upper bounds of all variables (probabilities) to 0 and 1, respectively. Additionally, normalization constraints are automatically added for each site in the domain.

### B.3.1. Optimizers

The default optimization interface for the **Optimizer** class is detailed in Listing B.7. It provides a class method which takes a problem and performs the optimization. If the problem is able to be solved, it will store the solution $x^*$ in the problem. There is also a means of obtaining the multipliers from the optimizer directly.

**Listing B.7.** Problem Interface

```
bool optimize(Problem& problem);
Vector& multipliers();
```

The default (base class) optimizer casts the problem as a Lagrangian system of nonlinear equations, and creates a nonlinear solver based on the modified Newton-Raphson solver with a line search as implemented in *Numerical Recipes in C, Second Edition*. In the future, this optimizer will be adjusted to a more robust method.

Vergil also provides an interface to the IPOPT (Interior Point Optimizer) software package. **CMake** will automatically detect if the IPOPT package is installed on your system, and compile the corresponding derived class if necessary. The class provides the same interface as in Listing B.7, with the addition of a handful of routines for adjusting IPOPT settings.

### B.3.2. Lagrangian Optimization

The default optimizer uses the method of Lagrange multipliers to cast the problem as a system of nonlinear equations. For a problem as given by **Eq. B-18**, we can define a

Lagrangian as

**Eq. B-22**
$$V = f(x) + \sum_k \lambda_k (g_k(\vec{w}) - g_k^o)$$

where the solution $x^*$ is a local optimizer of $V$ as given by

**Eq. B-23**
$$\vec{\nabla} V = \vec{0}$$

To find this solution, Vergil contains a series of classes that abstract solving this system of nonlinear equations, through both the **NonlinearSystem** and **NonlinearSolver** classes.

The **NonlinearSystem** class contains an interface for setting the value of the system's Functional, as well as the Jacobian of the Functional at some point $x$ (Listing B.8).

**Listing B.8.** Nonlinear System Interface

```
virtual void Functional(const Vector &x, Vector* functional);
virtual void Jacobian(const Vector &x, Matrix* jacobian);
```

The **NonlinearSystemLagrangian** is derived from the **NonlinearSystem** class. The Functional and Jacobian methods are then defined by the following equations, forming the Gradient and Hessian of the Lagrangian variational functional, respectively. The Gradient consists of the first partial derivative of V with respect to the variables (using the first partial derivative of the objective function and the first partial derivative of the constraints), and the first partial derivative with respect to the constraint multipliers (using the primal value

of the constraints).

**Eq. B-24**

$$\mathbf{F}(\vec{x}) = \begin{pmatrix} \frac{\partial V}{\partial x_i} \\ \frac{\partial V}{\partial \lambda_k} \end{pmatrix} = \begin{pmatrix} \frac{\partial f(x)}{\partial x_i} + \sum_k \lambda_k (\frac{\partial g_k}{\partial x_i}) \\ g_k - g_k^o \end{pmatrix} = 0$$

where

**Eq. B-25**

$$\vec{x} = \begin{pmatrix} x_i \\ \lambda_k \end{pmatrix}$$

The Jacobian of the gradient of the variational function is exactly the Hessian of the variational functional.

**Eq. B-26**

$$\begin{pmatrix} \frac{\partial^2 V}{\partial x_i \partial x_j} & \frac{\partial^2 V}{\partial x_i \partial \lambda_k} \\ \frac{\partial^2 V}{\partial \lambda_k \partial x_j} & \frac{\partial^2 V}{\partial \lambda_k \partial \lambda_l} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} + \sum_k \lambda_k (\frac{\partial^2 g_k}{\partial x_i \partial x_j}) & \frac{\partial g_k}{\partial x_i} \\ \frac{\partial g_k}{\partial x_j} & 0 \end{pmatrix}$$

To solve a given system of nonlinear equations, a **NonlinearSolver** class is implemented. The interface, given in Listing B.9, takes a NonlinearSystem on construction, contains a means of finding the root of that system for an initial estimate of the root, a boolean indicating if the root finder was successful, an accessor to the root, and a measure of the residual for that system of nonlinear equations.

**Listing B.9.** Nonlinear System Interface

```
NonlinearSolver(NonlinearSystem& nonlinear_system, int dimension);
virtual void FindRoot(const Vector& initial_guess) = 0;
bool root_found();
Vector root();
virtual double Residual() = 0;
```

This interface is implemented specifically in the **NonlinearSolverNumericalRecipes** class, which finds the root of the system of equations using the Newton-Raphson inspired implementation detailed in *Numerical Recipes in C, Second Edition*.

### B.3.3. Gibbs Form of Nonlinear System of Equations

Historically, the probabilistic approach has taken advantage of the partial solution afforded by **Eq. A-17**. Taking this equation across all probabilities along with any additional equality constraints, one can then form a system of nonlinear equations whose root is a solution. This looks like

**Eq. B-27**
$$\mathbf{F}(\vec{x}) = \begin{pmatrix} \frac{\exp(-Q_i)}{z_i} - w_i \\ g_k - g_k^o \end{pmatrix} = 0$$

where

**Eq. B-28**
$$\vec{x} = \begin{pmatrix} w_i \\ \lambda_k \end{pmatrix}$$

and where the definition for $Q_i$ and $z_i$ are given in Chapter 2. As a reference, we will begin

with **Eq. A-13** as the current definition, where $U$ is the mean field energy and $\beta$ is the "effective temperature".

**Eq. B-29**

$$Q_i = \beta \frac{\partial U}{\partial w_i} + \sum_k \lambda_k \frac{\partial g_k}{\partial w_i}$$
$$z_i = \sum_{tc} \exp(-Q_i)$$

The **NonlinearSystemGibbs** class is a derived class from **NonlinearSystem**, which defines the Functional as **Eq. B-27**, and the Jacobian of that Functional as

**Eq. B-30**

$$\begin{pmatrix} \frac{\partial}{\partial x_i}(\frac{\exp(-Q_i)}{z_i} - w_i) & \frac{\partial}{\partial \lambda_k}(\frac{\exp(-Q_i)}{z_i} - w_i) \\ \frac{\partial}{\partial x_i}(g_k - g_k^o) & \frac{\partial}{\partial \lambda_k}(g_k - g_k^o) \end{pmatrix}$$

or

**Eq. B-31**

$$\begin{pmatrix} \frac{1}{z_i}\frac{\partial \exp(-Q_i)}{\partial x_i} - \frac{\exp(-Q_i)}{z_i^2}\frac{\partial z_i}{\partial x_i} - \frac{\partial w_i}{\partial x_i} & \frac{1}{z_i}\frac{\partial \exp(-Q_i)}{\partial \lambda_k} - \frac{\exp(-Q_i)}{z_i^2}\frac{\partial z_i}{\partial \lambda_k} \\ \frac{\partial g_k}{\partial x_i} & 0 \end{pmatrix}$$

It is clear that to populate the Functional vector and Jacobian matrix, we need to be able to evaluate the first partial derivatives of the Boltzmann factors and partition functions with respect to both the probabilities and Lagrange multipliers. The **NonlinearSystemGibbs** class keeps computational time down by storing the current estimate of both $Q_i$ and $z_i$, each stored in a vector, and has a private interface for evaluating each (using **Eq. B-29**) as well as their first partial derivatives, as given in Listing B.10.

**Listing B.10.** NonlinearSystemGibbs Private Interface for Boltzmann Factors and Partition Functions

```
double BoltzmannFactor(size_t i);
double BoltzmannFactorFirstDerivative(size_t i, size_t j);
double PartitionFunction(size_t isite);
double PartitionFunctionFirstDerivative(size_t isite, size_t j);
      void AddFunctionTerm(std::string id, Function* function, double
          constant_scalar = 1.0);

ProblemDomainProbability probability_problem_;
std::vector<double> boltzmann_factor_;
std::vector<double> partition_function_;
std::vector<size_t> partition_function_index_;
std::vector<std::vector<size_t> > partition_function_list_;
std::map<std::string, std::pair<Function*, double> > function_terms_;
Vector multipliers_;
```

It is worth calling attention to the vector of vector indices, **partition_function_list_**, which stores the list of indices associated with each partition function index; as well as the vector of integers **partition_function_index_** which stores the corresponding index to **partition_function_list_** for each of the conformer indices. These two objects are created on instantiation when the domain is passed to the constructor.

The equations defining the first partial derivatives in Listing B.10 are as follows. The first partial derivative of the Boltzmann factor with respect to the probabilities $w_i$ is given as

**Eq. B-32** $$\frac{\partial}{\partial x_j}\exp(-Q_i) = -\exp(-Q_i)\frac{\partial Q_i}{\partial x_j} = -\exp(-Q_i)(\beta\frac{\partial^2 U}{\partial w_i \partial w_j} + \sum_k \lambda_k \frac{\partial^2 g_k}{\partial w_i \partial w_j})$$

while the first partial derivative with respect to the multipliers is given as

**Eq. B-33**
$$\frac{\partial}{\partial \lambda_k} \exp(-Q_i) = -\exp(-Q_i)\frac{\partial Q_i}{\partial \lambda_k} = -\exp(-Q_i)(\frac{\partial g_k}{\partial w_i})$$

The first partial derivative of the partition function simply uses the above definitions for the first partial derivatives of the Boltzmann factors with respect to both the probabilities and Lagrange multipliers as shown in the following equations.

**Eq. B-34**
$$\frac{\partial}{\partial w_j} z_i = \frac{\partial}{\partial w_j} \sum_{tc} \exp(-Q_i) = -\sum_{tc} \exp(-Q_i)\frac{\partial Q_i}{\partial w_j}$$

**Eq. B-35**
$$\frac{\partial}{\partial \lambda_k} z_i = \frac{\partial}{\partial \lambda_k} \sum_{tc} \exp(-Q_i) = -\sum_{tc} \exp(-Q_i)\frac{\partial Q_i}{\partial \lambda_k}$$

To minimize having to call the evaluation of the Boltzmann factors and partition functions, the **set_x** functions have been overridden to make sure that any changes to the variables also updates the appropriate terms. Additionally, we call special attention to the routine which fills in the Jacobian matrix. Because the same derivatives will be called in each column, the fill is done in a column-wise fashion as detailed in Listing B.11.

**Listing B.11.** Gibbs Fill Jacobian Routine

```
void NonlinearSystemGibbs::Jacobian(const Vector &x, Matrix* jacobian) {
  //Update the variables, boltzmann factors, and partition functions
  set_x(x);
  //For each column j in the Jacobian
  for (size_t j = 0; j < jacobian->number_cols(); j++) {
    ...
    //Grab all derivates of Qi and Zi for this fixed column of j
    for (size_t i = 0; i < boltzmann_factor_.size(); i++) {
      boltzmann_factor_1D[i] = BoltzmannFactorFirstDerivative(i, j);
    }
    for (size_t isite = 0; isite < partition_function_.size(); isite++) {
      partition_function_1D[isite] = PartitionFunctionFirstDerivative(isite,
          j);
    }
    //Fill probability rows of the jth column
    for (size_t i = 0; i < probability_problem_.NumberVariables(); i++) {
      size_t isite = partition_function_index_[i];
      //Fill each element in that column
      ...
    }
  }
}
```

# BIBLIOGRAPHY


1. L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard, and D. Baker, "De novo computational design of retro-aldol enzymes.," *Science*, vol. 319, pp. 1387–1391, 2008.

2. D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, and D. Baker, "Kemp elimination catalysts by computational enzyme design.," *Nature*, vol. 453, pp. 190–195, 2008.

3. J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael, and D. Baker, "Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction.," *Science*, vol. 329, pp. 309–313, 2010.

4. H. K. Privett, G. Kiss, T. M. Lee, R. Blomberg, R. A. Chica, L. M. Thomas, D. Hilvert, K. N. Houk, and S. L. Mayo, "Iterative approach to computational enzyme design," 2012.

5. E. J. Choi, J. Mao, and S. L. Mayo, "Computational design and biochemical characterization of maize nonspecific lipid transfer protein variants for biosensor applications.," *Protein Sci.*, vol. 16, pp. 582–588, 2007.

6. S. Ye, B. M. Discher, J. Strzalka, T. Xu, S. P. Wu, D. Noy, I. Kuzmenko, T. Gog, M. J. Therien, P. L. Dutton, and J. K. Blasie, "Amphiphilic four-helix bundle peptides designed for light-induced electron transfer across a soft interface," *Nano Lett.*, vol. 5, pp. 1658–1667, 2005.

7. T. Xu, S. P. Wu, I. Miloradovic, M. J. Therien, and J. K. Blasie, "Incorporation of Designed Extended Chromophores into Amphiphilic 4-Helix Bundle Peptides for Nonlinear Optical Biomolecular Materials," *Nano Lett.*, vol. 6, pp. 2387–2394, 2006.

8. J. Strzalka, T. Xu, A. Tronin, S. P. Wu, I. Miloradovic, I. Kuzmenko, T. Gog, M. J. Therien, and J. K. Blasiert, "Structural studies of amphiphilic 4-helix bundle peptides incorporating designed extended chromophores for nonlinear optical biomolecular materials," *Nano Lett.*, vol. 6, pp. 2395–2405, 2006.

9. H. Zou, M. J. Therien, and J. K. Blasie, "Structure and dynamics of an extended conjugated NLO chromophore within an amphiphilic 4-helix bundle peptide by molecular dynamics simulation.," *J. Phys. Chem. B*, vol. 112, pp. 1350–7, Mar. 2008.

10. G. Gonella, H.-l. Dai, H. C. Fry, M. J. Therien, V. Krishnan, A. Tronin, J. K. Blasie, D. U. V, and N. Carolina, "Control of the Orientational Order and Nonlinear Optical Response of the Push-Pull Chromophore RuPZn via Specific Incorporation into

Densely Packed Monolayer Ensembles of an Amphiphilic 4-Helix Bundle Peptide: Second Harmonic Generation at High Chromoph," *J. Am. Chem. Soc.*, no. 21, pp. 9693–9700, 2010.

11. V. Krishnan, A. Tronin, J. Strzalka, H. C. Fry, M. J. Therien, and J. K. Blasie, "Control of the orientational order and nonlinear optical response of the "push-Pull" chromophore RuPZn via specific incorporation into densely packed monolayer ensembles of an amphiphilic four-helix bundle peptide: Characterization of the peptide-chromoph," *J. Am. Chem. Soc.*, vol. 132, no. 14, pp. 11083–11092, 2010.

12. J. Koo, J. Park, A. Tronin, R. Zhang, V. Krishnan, J. Strzalka, I. Kuzmenko, H. C. Fry, M. J. Therien, and J. K. Blasie, "Acentric 2-D ensembles of D-br-A electron-transfer chromophores via vectorial orientation within amphiphilic n -helix bundle peptides for photovoltaic device applications," *Langmuir*, vol. 28, pp. 3227–3238, Mar. 2012.

13. H. C. Fry, A. Lehmann, L. E. Sinks, I. Asselberghs, A. Tronin, V. Krishnan, J. K. Blasie, K. Clays, W. F. DeGrado, J. G. Saven, and M. J. Therien, "Computational de novo design and characterization of a protein that selectively binds a highly hyperpolarizable abiological chromophore.," *J. Am. Chem. Soc.*, vol. 135, pp. 13914–26, Sept. 2013.

14. C. Pabo, "Molecular technology. Designing proteins and peptides.," 1983.

15. S. M. Lippow and B. Tidor, "Progress in computational protein design," 2007.

16. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank.," *Nucleic Acids Res.*, vol. 28, pp. 235–242, 2000.

17. L. Regan and W. F. DeGrado, "Characterization of a helical protein designed from first principles.," *Science*, vol. 241, pp. 976–978, 1988.

18. C. P. Hill, D. H. Anderson, L. Wesson, W. F. DeGrado, and D. Eisenberg, "Crystal structure of alpha 1: implications for protein design.," *Science*, vol. 249, pp. 543–546, 1990.

19. S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht, "Protein design by binary patterning of polar and nonpolar amino acids.," *Science*, vol. 262, pp. 1680–1685, 1993.

20. T. P. Quinn, N. B. Tweedy, R. W. Williams, J. S. Richardson, and D. C. Richardson, "Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, pp. 8747–8751, 1994.

21. J. W. Bryson, S. F. Betz, H. S. Lu, D. J. Suich, H. X. Zhou, K. T. O'Neil, and W. F. DeGrado, "Protein design: a hierarchic approach.," *Science*, vol. 270, pp. 935–941, 1995.

22. J. W. Bryson, J. R. Desjarlais, T. M. Handel, and W. F. DeGrado, "From coiled coils to small globular proteins: design of a native-like three-helix bundle.," *Protein Sci.*, vol. 7, pp. 1404–1414, 1998.

23. S. Roy, K. J. Helmer, and M. H. Hecht, "Detecting native-like properties in combinatorial libraries of de novo proteins.," *Fold. Des.*, vol. 2, pp. 89–92, 1997.

24. Y. Wei, S. Kim, D. Fela, J. Baum, and M. H. Hecht, "Solution structure of a de novo protein from a designed combinatorial library.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, pp. 13270–13273, 2003.

25. M. Schneider, X. Fu, and A. E. Keating, "X-ray vs. NMR structures as templates for computational protein design," *Proteins Struct. Funct. Bioinforma.*, vol. 77, pp. 97–110, 2009.

26. M. Levitt, "Accurate modeling of protein conformation by automatic segment matching.," *J. Mol. Biol.*, vol. 226, pp. 507–533, 1992.

27. K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.," *J. Mol. Biol.*, vol. 268, pp. 209–225, 1997.

28. C. J. Tsai, J. Zheng, C. Alemán, and R. Nussinov, "Structure by design: from single proteins and their building blocks to nanostructures," *Trends Biotechnol.*, vol. 24, pp. 449–454, 2006.

29. B. North, C. M. Summa, G. Ghirlanda, and W. F. DeGrado, "D(n)-symmetrical tertiary templates for the design of tubular proteins.," *J. Mol. Biol.*, vol. 311, pp. 1081–1090, 2001.

30. F. V. Cochran, S. P. Wu, W. Wang, V. Nanda, J. G. Saven, M. J. Therien, and W. F. DeGrado, "Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor.," *J. Am. Chem. Soc.*, vol. 127, pp. 1346–7, Mar. 2005.

31. K. a. McAllister, H. Zou, F. V. Cochran, G. M. Bender, A. Senes, H. C. Fry, V. Nanda, P. a. Keenan, J. D. Lear, J. G. Saven, M. J. Therien, J. K. Blasie, and W. F. DeGrado, "Using $\alpha$-helical coiled-coils to design nanostructured metalloporphyrin arrays," *J. Am. Chem. Soc.*, vol. 130, no. 36, pp. 11921–11927, 2008.

32. T. Cui, D. Mowrey, V. Bondarenko, T. Tillman, D. Ma, E. Landrum, J. M. Perez-Aguilar, J. He, W. Wang, J. G. Saven, R. G. Eckenhoff, P. Tang, and Y. Xu, "NMR structure and dynamics of a designed water-soluble transmembrane domain of nicotinic acetylcholine receptor," 2012.

33. J. R. Desjarlais and T. M. Handel, "New strategies in protein design.," *Curr. Opin. Biotechnol.*, vol. 6, pp. 460–466, 1995.

34. P. B. Harbury, B. Tidor, and P. S. Kim, "Repacking protein cores with backbone freedom: structure prediction for coiled coils.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, pp. 8408–8412, 1995.

35. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim, "High-resolution protein design with backbone freedom.," *Science*, vol. 282, pp. 1462–1467, 1998.

36. H. Kono and J. G. Saven, "Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure.," *J. Mol. Biol.*, vol. 306, pp. 607–28, Mar. 2001.

37. E. L. Humphris and T. Kortemme, "Prediction of Protein-Protein Interface Sequence Diversity Using Flexible Backbone Computational Protein Design," *Structure*, vol. 16, pp. 1777–1788, 2008.

38. I. Georgiev, D. Keedy, J. S. Richardson, D. C. Richardson, and B. R. Donald, "Algorithm for backrub motions in protein design," *Bioinformatics*, vol. 24, 2008.

39. D. J. Mandell and T. Kortemme, "Backbone flexibility in computational protein design," 2009.

40. R. L. Dunbrack, "Rotamer libraries in the 21st century," 2002.

41. J. W. Ponder and F. M. Richards, "Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.," *J. Mol. Biol.*, vol. 193, pp. 775–791, 1987.

42. P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery, "A new approach to the rapid determination of protein side chain conformations.," *J. Biomol. Struct. Dyn.*, vol. 8, pp. 1267–1289, 1991.

43. R. L. Dunbrack and M. Karplus, "Backbone-dependent rotamer library for proteins. Application to side-chain prediction.," *J. Mol. Biol.*, vol. 230, pp. 543–74, Mar. 1993.

44. H. Kono and J. Doi, "A new method for side-chain conformation prediction using a Hopfield network and reproduced rotamers," *J. Comput. Chem.*, vol. 17, pp. 1667–1683, Nov. 1996.

45. M. De Maeyer, J. Desmet, and I. Lasters, "All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination.," *Fold. Des.*, vol. 2, pp. 53–66, 1997.

46. S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. De Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, "Structure validation by C$\alpha$ geometry: $\phi$,$\psi$ and C$\beta$ deviation," *Proteins Struct. Funct. Genet.*, vol. 50, pp. 437–450, 2003.

47. M. V. Shapovalov and R. L. Dunbrack, "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.," *Structure*, vol. 19, pp. 844–58, June 2011.

48. Z. Xiang and B. Honig, "Extending the accuracy limits of prediction for side-chain conformations.," *J. Mol. Biol.*, vol. 311, pp. 421–430, 2001.

49. R. P. Shetty, P. I. W. De Bakker, M. A. DePristo, and T. L. Blundell, "Advantages of fine-grained side chain conformer libraries.," *Protein Eng.*, vol. 16, pp. 963–969, 2003.

50. S. A. Marshall and S. L. Mayo, "Achieving stability and conformational specificity in designed proteins via binary patterning.," *J. Mol. Biol.*, vol. 305, pp. 619–631, 2001.

51. K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem.," *Annu. Rev. Biophys.*, vol. 37, pp. 289–316, 2008.

52. L. Jiang, B. Kuhlman, T. Kortemme, and D. Baker, "A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces," *Proteins Struct. Funct. Genet.*, vol. 58, pp. 893–904, 2005.

53. G. I. Makhatadze and P. L. Privalov, "Energetics of protein structure.," *Adv. Protein Chem.*, vol. 47, pp. 307–425, 1995.

54. F. E. Boas and P. B. Harbury, "Potential energy functions for protein design," 2007.

55. S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner, "A new force field for molecular mechanical simulation of nucleic acids and proteins," *J. Am. Chem. Soc.*, vol. 106, pp. 765–784, Feb. 1984.

56. A. MacKerell and D. Bashford, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *J. ...*, vol. 5647, no. 97, pp. 3586–3616, 1998.

57. A. D. Mackerell, M. Feig, and C. L. Brooks, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulation," *J. Comput. Chem.*, vol. 25, pp. 1400–1415, 2004.

58. K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *J. Comput. Chem.*, vol. 31, pp. 671–690, 2010.

59. W. Jin, O. Kambara, H. Sasakawa, A. Tamura, and S. Takada, "De novo design of foldable proteins with smooth folding funnel: Automated negative design and experimental verification," *Structure*, vol. 11, pp. 581–590, 2003.

60. Y. Song, M. Tyka, A. Leaver-Fay, J. Thompson, and D. Baker, "Structure-guided

forcefield optimization," *Proteins Struct. Funct. Bioinforma.*, vol. 79, pp. 1898–1909, 2011.

61. W. Im, M. S. Lee, and C. L. Brooks, "Generalized Born Model with a Simple Smoothing Function," *J. Comput. Chem.*, vol. 24, pp. 1691–1702, 2003.

62. C. L. Vizcarra and S. L. Mayo, "Electrostatics in computational protein design," 2005.

63. B. Kuhlman and D. Baker, "Native protein sequences are close to optimal for their structures.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, pp. 10383–10388, 2000.

64. J. G. Saven, "Designing Protein Energy Landscapes," *Chem. Rev.*, vol. 101, pp. 3113–3130, Oct. 2001.

65. J. R. Calhoun, H. Kono, S. Lahr, W. Wang, W. F. DeGrado, and J. G. Saven, "Computational Design and Characterization of a Monomeric Helical Dinuclear Metalloprotein," *J. Mol. Biol.*, vol. 334, pp. 1101–1115, Dec. 2003.

66. H. W. Hellinga and F. M. Richards, "Optimal sequence selection in proteins of known structure by simulated evolution.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, pp. 5803–5807, 1994.

67. C. A. Voigt, D. B. Gordon, and S. L. Mayo, "Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design.," *J. Mol. Biol.*, vol. 299, pp. 789–803, 2000.

68. A. P. Cootes, P. M. G. Curmi, and A. E. Torda, "Biased Monte Carlo optimization of protein sequences," *J. Chem. Phys.*, vol. 113, pp. 2489–2496, 2000.

69. X. Yang and J. G. Saven, "Computational methods for protein design and protein sequence variability: Biased Monte Carlo and replica exchange," *Chem. Phys. Lett.*, vol. 401, pp. 205–210, 2005.

70. J. R. Desjarlais and T. M. Handel, "Side-chain and backbone flexibility in protein core design.," *J. Mol. Biol.*, vol. 290, pp. 305–318, 1999.

71. A. R. Leach and A. P. Lemon, "Exploring the conformational space of protein side chains using dead- end elimination and the A algorithm," *Proteins Struct. Funct. Genet.*, vol. 33, pp. 227–239, 1998.

72. L. Wernisch, S. Hery, and S. J. Wodak, "Automatic protein design with all atom force-fields by exact and heuristic optimization.," *J. Mol. Biol.*, vol. 301, pp. 713–736, 2000.

73. G. Grigoryan, A. W. Reinke, and A. E. Keating, "Design of protein-interaction specificity gives selective bZIP-binding peptides.," *Nature*, vol. 458, pp. 859–864, 2009.

74. J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters, "The dead-end elimination theorem and its use in protein side-chain positioning.," *Nature*, vol. 356, pp. 539–542, 1992.

75. R. F. Goldstein, "Efficient rotamer elimination applied to protein side-chains and related spin glasses.," *Biophys. J.*, vol. 66, pp. 1335–1340, 1994.

76. L. L. Looger and H. W. Hellinga, "Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics.," *J. Mol. Biol.*, vol. 307, pp. 429–445, 2001.

77. D. B. Gordon, G. K. Hom, S. L. Mayo, and N. A. Pierce, "Exact rotamer optimization for protein design," *J. Comput. Chem.*, vol. 24, pp. 232–243, 2003.

78. I. Georgiev and B. R. Donald, "Dead-End Elimination with backbone flexibility," in *Bioinformatics*, vol. 23, 2007.

79. P. Koehl and M. Delarue, "Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy.," *J. Mol. Biol.*, vol. 239, pp. 249–275, 1994.

80. J. Mendes, C. M. Soares, and M. A. Carrondo, "Improvement of side-chain modeling in proteins with the self-consistent mean field theory method based on an analysis of the factors influencing prediction," *Biopolymers*, vol. 50, pp. 111–131, 1999.

81. J. G. Saven and P. G. Wolynes, "Statistical Mechanics of the Combinatorial Synthesis and Analysis of Folding Macromolecules," *J. Phys. Chem. B*, vol. 101, pp. 8375–8389, Oct. 1997.

82. J. Zou and J. G. Saven, "Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure.," *J. Mol. Biol.*, vol. 296, pp. 281–94, Mar. 2000.

83. X. Fu, H. Kono, and J. G. Saven, "Probabilistic approach to the design of symmetric protein quaternary structures.," *Protein Eng.*, vol. 16, pp. 971–7, Dec. 2003.

84. J. Zou and J. G. Saven, "Using self-consistent fields to bias Monte Carlo methods with applications to designing and sampling protein sequences," *J. Chem. Phys.*, vol. 118, pp. 3843–3854, 2003.

85. S.-g. Kang and J. G. Saven, "Computational protein design: structure, function and combinatorial diversity.," *Curr. Opin. Chem. Biol.*, vol. 11, pp. 329–34, June 2007.

86. A. M. Slovic, H. Kono, J. D. Lear, J. G. Saven, and W. F. DeGrado, "Computational design of water-soluble analogues of the potassium channel KcsA.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, pp. 1828–33, Mar. 2004.

87. J. M. Perez-Aguilar, J. Xi, F. Matsunaga, X. Cui, B. Selling, J. G. Saven, and R. Liu,

"A computationally designed water-soluble variant of a G-protein-coupled receptor: the human mu opioid receptor.," *PLoS One*, vol. 8, p. e66009, Jan. 2013.

88. H. C. Fry, A. Lehmann, J. G. Saven, W. F. DeGrado, and M. J. Therien, "Computational design and elaboration of a de novo heterotetrameric alpha-helical protein that selectively binds an emissive abiological (porphinato)zinc chromophore.," *J. Am. Chem. Soc.*, vol. 132, pp. 3997–4005, Mar. 2010.

89. G. M. Bender, A. Lehmann, H. Zou, H. Cheng, H. C. Fry, D. Engel, M. J. Therien, J. K. Blasie, H. Roder, J. G. Saven, and W. F. DeGrado, "De novo design of a single-chain diphenylporphyrin metalloprotein.," *J. Am. Chem. Soc.*, vol. 129, pp. 10732–40, Sept. 2007.

90. C. J. Lanci, C. M. MacDermaid, S.-g. Kang, R. Acharya, B. North, X. Yang, X. J. Qiu, W. F. DeGrado, and J. G. Saven, "Computational design of a protein crystal.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, pp. 7304–9, May 2012.

91. T. Mori, M. Kyotani, and K. Akagi, "Helicity-controlled liquid crystal reaction field using nonbridged and bridged binaphthyl derivatives available for synthesis of helical conjugated polymers," *Macromolecules*, vol. 41, pp. 607–613, 2008.

92. T. Verbiest, S. Sioncke, A. Persoons, L. Vyklický, and T. J. Katz, "Electric-field-modulated circular-difference effects in second-harmonic generation from a chiral liquid crystal," *Angew. Chemie - Int. Ed.*, vol. 41, pp. 3882–3884, 2002.

93. G. Koekcelberghs, T. Verbiest, M. Vangheluwe, L. De Groof, I. Asselberghs, I. Picard, K. Clays, A. Persoons, and C. Samyn, "Influence of monomer optical purity on the conformation and properties of chiral, donor-embedded polybinaphthalenes for nonlinear optical purposes," *Chem. Mater.*, vol. 17, pp. 118–121, 2005.

94. T. Verbiest, S. V. Elshocht, M. Kauranen, A. Persoons, C. Nuckolls, and T. Katz, "Enhancement of nonlinear optical properties through supramolecular chirality," *Nonlinear Opt. '98. Mater. Fundam. Appl. Top. Meet. (Cat. No.98CH36244)*, 1998.

95. B. Göhler, V. Hamelbeck, T. Z. Markus, M. Kettner, G. F. Hanne, Z. Vager, R. Naaman, and H. Zacharias, "Spin selectivity in electron transmission through self-assembled monolayers of double-stranded DNA.," *Science*, vol. 331, pp. 894–897, 2011.

96. T. Yamamoto, T. Fukushima, A. Kosaka, W. Jin, Y. Yamamoto, N. Ishii, and T. Aida, "Conductive one-handed nanocoils by coassembly of hexabenzocoronenes: Control of morphology and helical chirality," *Angew. Chemie - Int. Ed.*, vol. 47, pp. 1672–1675, 2008.

97. F. Di Maria, P. Olivelli, M. Gazzano, A. Zanelli, M. Biasiucci, G. Gigli, D. Gentili, P. D'Angelo, M. Cavallini, and G. Barbarella, "A successful chemical strategy to induce oligothiophene self-assembly into fibers with tunable shape and function," *J. Am. Chem. Soc.*, vol. 133, pp. 8654–8661, 2011.

98. Y. Sawada, S. Furumi, A. Takai, M. Takeuchi, K. Noguchi, and K. Tanaka, "Rhodium-catalyzed enantioselective synthesis, crystal structures, and photophysical properties of helically chiral 1,1'-bitriphenylenes," *J. Am. Chem. Soc.*, vol. 134, pp. 4080–4083, 2012.

99. K. Watanabe, I. Osaka, S. Yorozuya, and K. Akagi, "Helically $\pi$-stacked thiophene-based copolymers with circularly polarized fluorescence: High dissymmetry factors enhanced by self-ordering in chiral nematic liquid crystal phase," *Chem. Mater.*, vol. 24, pp. 1011–1024, 2012.

100. S. M. Bachilo, M. S. Strano, C. Kittrell, R. H. Hauge, R. E. Smalley, and R. B. Weisman, "Structure-assigned optical spectra of single-walled carbon nanotubes.," *Science*, vol. 298, pp. 2361–2366, 2002.

101. P. Deria, L. E. Sinks, T.-H. Park, D. M. Tomezsko, M. J. Brukman, D. A. Bonnell, and M. J. Therien, "Phase transfer catalysts drive diverse organic solvent solubility of single-walled carbon nanotubes helically wrapped by ionic, semiconducting polymers," *Nano Lett.*, vol. 10, pp. 4192–4199, 2010.

102. Y. K. Kang, O.-S. Lee, P. Deria, S. H. Kim, T.-H. Park, D. A. Bonnell, J. G. Saven, and M. J. Therien, "Helical wrapping of single-walled carbon nanotubes by water soluble poly(p-phenyleneethynylene)," *Nano Lett.*, vol. 9, pp. 1414–1418, 2009.

103. B. A. Larsen, P. Deria, J. M. Holt, I. N. Stanton, M. J. Heben, M. J. Therien, and J. L. Blackburn, "Effect of solvent polarity and electrophilicity on quantum yields and solvatochromic shifts of single-walled carbon nanotube photoluminescence," *J. Am. Chem. Soc.*, vol. 134, pp. 12485–12491, 2012.

104. J. Park, P. Deria, and M. J. Therien, "Dynamics and transient absorption spectral signatures of the single-wall carbon nanotube electronically excited triplet state," *J. Am. Chem. Soc.*, vol. 133, pp. 17156–17159, 2011.

105. M. R. Rosario-Canales, P. Deria, M. J. Therien, and J. J. Santiago-Avils, "Composite electronic materials based on poly(3,4-propylenedioxythiophene) and highly charged poly(aryleneethynylene)-wrapped carbon nanotubes for supercapacitors," *ACS Appl. Mater. Interfaces*, vol. 4, pp. 102–109, 2012.

106. P. Deria, C. D. Von Bargen, J. H. Olivier, A. S. Kumbhar, J. G. Saven, and M. J. Therien, "Single-handed helical wrapping of single-walled carbon nanotubes by chiral, ionic, semiconducting polymers," *J. Am. Chem. Soc.*, vol. 135, pp. 16220–16234, 2013.

107. H. J. Gao, Y. Kong, D. X. Cui, and C. S. Ozkan, "Spontaneous Insertion of DNA Oligonucleotides into Carbon Nanotubes," *Nano Lett.*, vol. 3, no. 4, pp. 471–473, 2003.

108. R. Johnson, A. Johnson, and M. Klein, "Probing the structure of dna-carbon nanotube hybrids with molecular dynamics," *Nano Lett.*, vol. 8, no. 1, pp. 69–75, 2008.

109. R. R. Johnson, A. Kohlmeyer, A. T. C. Johnson, and M. L. Klein, "Free Energy Landscape of a DNA-Carbon Nanotube Hybrid Using Replica Exchange Molecular Dynamics," *Nano Lett.*, vol. 9, no. 2, pp. 537–541, 2009.

110. D. Roxbury, A. Jagota, and J. Mittal, "Sequence-Specific Self-Stitching Motif of Short Single-Stranded DNA on a Single-Walled Carbon Nanotube," *J. Am. Chem. Soc.*, vol. 133, no. 34, pp. 13545–13550, 2011.

111. M. V. Karachevtsev and V. A. Karachevtsev, "Peculiarities of Homooligonucleotides Wrapping around Carbon Nanotubes: Molecular Dynamics Modeling," *J. Phys. Chem.*, vol. 115, pp. 9271–9279, July 2011.

112. D. Roxbury, J. Mittal, and A. Jagota, "Molecular-Basis of Single-Walled Carbon Nanotube Recognition by Single-Stranded DNA," *Nano Lett.*, vol. 12, no. 3, pp. 1464–1469, 2012.

113. Y. H. Xie and A. K. Soh, "Investigation of Non-Covalent Association of Single-Walled Carbon Nanotube with Amylose by Molecular Dynamics Simulation," *Mater. Lett.*, vol. 59, no. 8, pp. 971–975, 2005.

114. M. J. Yang, V. Koutsos, and M. Zaiser, "Interactions between Polymers and Carbon Nanotubes: A Molecular Dynamics Study," *J. Phys. Chem. B*, vol. 109, no. 20, pp. 10009–10014, 2005.

115. C. Caddeo, C. Melis, L. Colombo, and A. Mattoni, "Understanding the helical wrapping of poly(3-hexylthiophene) on carbon nanotubes," *J. Phys. Chem. C*, vol. 114, no. 49, pp. 21109–21113, 2010.

116. M. Foroutan and A. T. Nasrabadhi, "Investigation of the interfacial binding between single-walled carbon nanotubes and heterocyclic conjugated polymers," *J. Phys. Chem. B*, vol. 114, pp. 5320–5326, 2010.

117. S. S. Tallury and M. A. Pasquinelli, "Molecular dynamics simulations of flexible polymer chains wrapping single-walled carbon nanotubes," *J. Phys .Chem. B*, vol. 114, pp. 4122–4129, 2010.

118. A. Furmanchuk, J. Leszczyski, S. Tretiak, and S. Kilina, "Morphology and optical response of carbon nanotubes functionalized by conjugated polymers," *J. Phys. Chem. C*, vol. 116, pp. 6831–6840, 2012.

119. H. Yang, Y. Chen, Y. Liu, W. S. Cai, and Z.-S. Li, "Molecular Dynamics Simulation of Polyethylene on Single Wall Carbon Nanotube," *J. Chem. Phys.*, vol. 127, no. 9, p. 094902, 2007.

120. J. Liu, X.-L. Wang, L. Zhao, G. Zhang, Z.-Y. Lu, and Z.-S. Li, "The Absorption and Diffusion of Polyethylene Chains on the Carbon Nanotube: The Molecular Dynamics Study," *J. Polym. Sci., Part B: Polym. Phys.*, vol. 46, no. 3, pp. 272–280, 2008.

121. M. Naito, K. Nobusawa, H. Onouchi, M. Nakamura, K. Yasui, A. Ikeda, and M. Fujiki, "Stiffness-and Conformation-Dependent Polymer Wrapping onto Single-Walled Carbon Nanotubes," *J. Am. Chem. Soc.*, vol. 130, no. 49, pp. 16697–16703, 2008.

122. F. A. Lemasson, T. Strunk, P. Gerstel, F. Hennrich, S. Lebedkin, C. Barner-Kowollik, W. Wenzel, M. M. Kappes, and M. Mayor, "Selective Dispersion of Single-Walled Carbon Nanotubes with Specific Chiral Indices by Poly( N-decyl-2,7-carbazole)," *J. Am. Chem. Soc.*, vol. 133, pp. 652–655, Feb. 2011.

123. J. Gao, M. A. Loi, E. J. F. de Carvalho, and M. C. dos Santos, "Selective Wrapping and Supramolecular Structures of Polyfluorene–Carbon Nanotube Hybrids," *ACS Nano*, vol. 5, pp. 3993–3999, May 2011.

124. G. Zuo, Q. Huang, G. Wei, R. Zhou, and H. Fang, "Plugging into Proteins: Poisoning Protein Function by a Hydrophobic Nanoparticle.," *ACS Nano*, vol. 4, pp. 7508–7514, Dec. 2010.

125. C. Ge, J. Du, L. Zhao, L. Wang, Y. Liu, D. Li, Y. Yang, R. Zhou, Y. Zhao, Z. Chai, and C. Chen, "Binding of Blood Proteins to Carbon Nanotubes Reduces Cytotoxicity.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108, pp. 16968–16973, Oct. 2011.

126. Y. Liu, C. Chipot, X. Shao, and W. Cai, "Solubilizing carbon nanotubes through noncovalent functionalization. insight from the reversible wrapping of alginic acid around a single-walled carbon nanotube," *J. Phys. Chem. B*, vol. 114, pp. 5783–5789, 2010.

127. Y. Liu, C. Chipot, X. Shao, and W. Cai, "Free-energy landscape of the helical wrapping of a carbon nanotube by a polysaccharide," *J. Phys. Chem. C*, vol. 115, no. 5, pp. 1851–1856, 2011.

128. C. D. Von Bargen, C. M. MacDermaid, O.-S. Lee, P. Deria, M. J. Therien, and J. G. Saven, "Origins of the helical wrapping of phenyleneethynylene polymers about single-walled carbon nanotubes.," *J. Phys. Chem. B*, vol. 117, pp. 12953–12965, 2013.

129. T. Rodinger and R. Pomès, "Enhancing the accuracy, the efficiency and the scope of free energy simulations," *Curr. Opin. Struct. Biol.*, vol. 15, pp. 164–170, 2005.

130. T. Simonson, G. Archontis, and M. Karplus, "Free energy simulations come of age: Protein-ligand recognition," *Acc. Chem. Res.*, vol. 35, pp. 430–437, 2002.

131. W. F. Van Gunsteren, X. Daura, and A. E. Mark, "Computation of free energy," *Helv. Chim. Acta*, vol. 85, pp. 3113–3129, 2002.

132. C. Chipot and D. A. Pearlman, "Free Energy Calculations. The Long and Winding Gilded Road," 2002.

133. D. L. Beveridge and F. M. DiCapua, "Free energy via molecular simulation: applications to chemical and biomolecular systems.," *Annu. Rev. Biophys. Biophys. Chem.*, vol. 18, pp. 431–492, 1989.

134. C. Jarzynski, "Equilibrium free energy differences from nonequilibrium measurements: a master equation approach," *Phys Rev E*, p. 29, 1997.

135. J. Henin and C. Chipot, "Overcoming free energy barriers using unconstrained molecular dynamics simulations," *J. Chem. Phys.*, vol. 121, no. 7, pp. 2904–2915, 2004.

136. M. J. Mitchell and J. A. McCammon, "Free energy difference calculations by thermodynamic integration: Difficulties in obtaining a precise value," *J. Comput. Chem.*, vol. 12, pp. 271–275, 1991.

137. H. Nymeyer, S. Gnanakaran, and A. E. García, "Atomic Simulations of Protein Folding, Using the Replica Exchange Algorithm," 2004.

138. C. Chipot, "Frontiers in free-energy calculations of biological systems," 2014.

139. C. Kurland and J. Gallant, "Errors of heterologous protein expression.," *Curr. Opin. Biotechnol.*, vol. 7, pp. 489–493, 1996.

140. S.-g. Kang, *Probabilistic Computational Protein Design: Advances in Methodology and the Incorporation of Non-Biological Molecular Components*. PhD thesis, University of Pennsylvania, 2009.

141. O. Conchillo-Solé, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura, and S. Ventura, "AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides.," *BMC Bioinformatics*, vol. 8, p. 65, 2007.

142. D. L. Minor and P. S. Kim, "Measurement of the beta-sheet-forming propensities of amino acids.," *Nature*, vol. 367, pp. 660–663, 1994.

143. C. N. Pace and J. M. Scholtz, "A helix propensity scale based on experimental studies of peptides and proteins.," *Biophys. J.*, vol. 75, pp. 422–427, 1998.

144. A. Senes, D. C. Chadi, P. B. Law, R. F. S. Walters, V. Nanda, and W. F. DeGrado, "Ez, a Depth-dependent Potential for Assessing the Energies of Insertion of Amino Acid Side-chains into Membranes: Derivation and Applications to Determining the Orientation of Transmembrane and Interfacial Helices," *J. Mol. Biol.*, vol. 366, pp. 436–448, 2007.

145. C. A. Schramm, B. T. Hannigan, J. E. Donald, C. Keasar, J. G. Saven, W. F. Degrado, and I. Samish, "Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions," *Structure*, vol. 20, pp. 924–935, 2012.

146. S. Betz, R. Fairman, K. O'Neil, J. Lear, and W. Degrado, "Design of two-stranded and three-stranded coiled-coil peptides.," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 348, pp. 81–88, 1995.

147. K. T. O'Neil, R. H. Hoess, D. P. Raleigh, and W. F. DeGrado, "Thermodynamic genetics of the folding of the B1 immunoglobulin-binding domain from streptococcal protein G," *Proteins Struct. Funct. Genet.*, vol. 21, pp. 11–21, 1995.

148. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," 1983.

149. T. Lazaridis and M. Karplus, "Effective energy function for proteins in solution," *Proteins Struct. Funct. Genet.*, vol. 35, pp. 133–152, 1999.

150. B. Roux and T. Simonson, "Implicit solvent models," in *Biophys. Chem.*, vol. 78, pp. 1–20, 1999.

151. J. P. Linge, M. A. Williams, C. A. E. M. Spronk, A. M. J. J. Bonvin, and M. Nilges, "Refinement of protein structures in explicit solvent," *Proteins Struct. Funct. Genet.*, vol. 50, pp. 496–506, 2003.

152. S. Plimpton, "Fast Parallel Algorithms for Short-Range Molecular Dynamics," *J. Comp. Phys.*, vol. 117, pp. 1–19, 1995.

153. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD.," *J. Comput. Chem.*, vol. 26, pp. 1781–802, Dec. 2005.

154. B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *J. Chem. Theory Comput.*, vol. 4, pp. 435–447, 2008.

155. B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, pp. 1545–1614, 2009.

156. W. Wang, *Computational Design of Proteins and Gene Libraries*. PhD thesis, University of Pennsylvania, 2004.

157. A. Leaver-Fay, M. J. O'Meara, M. Tyka, R. Jacak, Y. Song, E. H. Kellogg, J. Thompson, I. W. Davis, R. A. Pache, S. Lyskov, J. J. Gray, T. Kortemme, J. S. Richardson, J. J. Havranek, J. Snoeyink, D. Baker, and B. Kuhlman, "Scientific benchmarks for guiding macromolecular energy function improvement," *Methods Enzymol.*, vol. 523, pp. 109–143, 2013.

158. D. B. Gordon, S. A. Marshall, and S. L. Mayo, "Energy functions for protein design.," *Curr. Opin. Struct. Biol.*, vol. 9, pp. 509–513, 1999.

159. Z. Li, Y. Yang, J. Zhan, L. Dai, and Y. Zhou, "Energy functions in de novo protein design: current challenges and future prospects.," *Annu. Rev. Biophys.*, vol. 42, pp. 315–35, 2013.

160. L. G. Dunfield, A. W. Burgess, and H. A. Scheraga, "Energy parameters in polypeptides. 8. Empirical potential energy algorithm for the conformational analysis of large molecules," *J. Phys. Chem.*, vol. 82, pp. 2609–2616, Nov. 1978.

161. W. L. Jorgensen, "Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water," *J. Am. Chem. Soc.*, vol. 103, pp. 335–340, Jan. 1981.

162. D. F. Stickle, L. G. Presta, K. A. Dill, and G. D. Rose, "Hydrogen bonding in globular proteins.," *J. Mol. Biol.*, vol. 226, pp. 1143–59, Aug. 1992.

163. B. Cartledge, ed., *Nuclear Power*. New York: Oxford University Press, 1993.

164. J. M. Popic, B. Salbu, T. Strand, and L. Skipperud, "Assessment of radionuclide and metal contamination in a thorium rich area in Norway.," *J. Environ. Monit.*, vol. 13, pp. 1730–1738, 2011.

165. D. J. Sapsford, R. J. Bowell, J. N. Geroni, K. M. Penman, and M. Dey, "Factors influencing the release rate of uranium, thorium, yttrium and rare earth elements from a low grade ore," *Miner. Eng.*, vol. 39, pp. 165–172, 2012.

166. C. W. Abney, S. Liu, and W. Lin, "Tuning amidoximate to enhance uranyl binding: A density functional theory study," *J. Phys. Chem. A*, vol. 117, pp. 11558–11565, 2013.

167. R. Pardoux, S. Sauge-Merle, D. Lemaire, P. Delangle, L. Guilloreau, J. M. Adriano, and C. Berthomieu, "Modulating Uranium binding affinity in engineered calmodulin EF-hand peptides: Effect of phosphorylation," *PLoS One*, vol. 7, 2012.

168. L. Zhou, M. Bosscher, C. Zhang, S. Ozçubukçu, L. Zhang, W. Zhang, C. J. Li, J. Liu, M. P. Jensen, L. Lai, and C. He, "A protein engineered to bind uranyl selectively and with femtomolar affinity.," *Nat. Chem.*, vol. 6, pp. 236–41, Mar. 2014.

169. X. Du and T. E. Graedel, "Global in-use stocks of the rare earth elements: A first estimate," *Environ. Sci. Technol.*, vol. 45, pp. 4096–4101, 2011.

170. I. TABUSHI, Y. KOBUKE, and T. NISHIYA, "Extraction of uranium from seawater by polymer-bound macrocyclic hexaketone," 1979.

171. M. KANNO, "Present Status of Study on Extraction of Uranium from Sea Water," 1984.

172. A. C. Sather, O. B. Berryman, and J. Rebek, "Selective recognition and extraction of the uranyl ion," *J. Am. Chem. Soc.*, vol. 132, pp. 13572–13574, 2010.

173. A. E. V. Gorden, J. Xu, K. N. Raymond, and P. Durbin, "Rational design of sequestering agents for plutonium and other actinides," *Chem. Rev.*, vol. 103, pp. 4207–4282, 2003.

174. H. L. Jung, Z. Wang, J. Liu, and Y. Lu, "Highly sensitive and selective colorimetric sensors for uranyl (UO 2 2+): Development and comparison of labeled and label-free DNAzyme-gold nanoparticle systems," *J. Am. Chem. Soc.*, vol. 130, pp. 14217–14226, 2008.

175. M. Carboni, C. W. Abney, S. B. Liu, and W. B. Lin, "Highly porous and stable metal-organic frameworks for uranium extraction," *Chem. Sci.*, vol. 4, pp. 2396–2402, 2013.

176. L. Le Clainche and C. Vita, "Selective binding of uranyl cation by a novel calmodulin peptide," *Environ. Chem. Lett.*, vol. 4, pp. 45–49, 2006.

177. S. V. Wegner, H. Boyaci, H. Chen, M. P. Jensen, and C. He, "Engineering a uranyl-specific binding protein from NikR.," *Angew. Chem. Int. Ed. Engl.*, vol. 48, pp. 2339–2341, 2009.

178. W. F. DeGrado, C. M. Summa, V. Pavone, F. Nastri, and A. Lombardi, "De novo design and structural characterization of proteins and metalloproteins.," *Annu. Rev. Biochem.*, vol. 68, pp. 779–819, 1999.

179. Y. Lu, N. Yeung, N. Sieracki, and N. M. Marshall, "Design of functional metalloproteins.," *Nature*, vol. 460, pp. 855–862, 2009.

180. R. J. Radford, J. D. Brodin, E. N. Salgado, and F. A. Tezcan, "Expanding the utility of proteins as platforms for coordination chemistry," 2011.

181. S. D. Khare, Y. Kipnis, P. J. Greisen, R. Takeuchi, Y. Ashani, M. Goldsmith, Y. Song, J. L. Gallaher, I. Silman, H. Leader, J. L. Sussman, B. L. Stoddard, D. S. Tawfik, and D. Baker, "Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis," 2012.

182. J. D. Brodin, X. I. Ambroggio, C. Tang, K. N. Parent, T. S. Baker, and F. A. Tezcan, "Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays," 2012.

183. A. F. A. Peacock, "Incorporating metals into de novo proteins," 2013.

184. L. A. Solomon, G. Kodali, C. C. Moser, and P. L. Dutton, "Engineering the assembly of heme cofactors in man-made proteins," *J. Am. Chem. Soc.*, vol. 136, pp. 3192–3199, 2014.

185. C. M. MacDermaid, *Computational design of proteins and protein crystals*. PhD thesis, University of Pennsylvania, 2012.

186. F. H. Allen, "The Cambridge Structural Database: A quarter of a million crystal structures and rising," *Acta Crystallogr. Sect. B Struct. Sci.*, vol. 58, pp. 380–388, 2002.

187. R. D. Lins, E. R. Vorpagel, M. Guglielmi, and T. P. Straatsma, "Computer simulation of uranyl uptake by the rough lipopolysaccharide membrane of Pseudomonas aeruginosa," *Biomacromolecules*, vol. 9, pp. 29–35, 2008.

188. F. H. C. Crick, "The packing of $\alpha$-helices: simple coiled-coils," 1953.

189. G. Grigoryan and W. F. Degrado, "Probing designability via a generalized model of helical bundle geometry.," *J. Mol. Biol.*, vol. 405, pp. 1079–100, Jan. 2011.

190. R. D. B. Fraser and T. MacRae, *Conformation in fibrous proteins and related synthetic polypeptides*. Academic Press LTD, 1974.

191. V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: All-atom structure validation for macromolecular crystallography," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, pp. 12–21, 2010.

192. O. D. Testa, E. Moutevelis, and D. N. Woolfson, "CC+: A relational database of coiled-coil structures," *Nucleic Acids Res.*, vol. 37, 2009.

193. G. Renger, "Coupling of electron and proton transfer in oxidative water cleavage in photosynthesis," *Biochim. Biophys. Acta - Bioenerg.*, vol. 1655, pp. 195–204, 2004.

194. M. H. V. Huynh and T. J. Meyer, "Proton-coupled electron transfer," 2007.

195. J. Barber, "Photosynthetic energy conversion: natural and artificial.," *Chem. Soc. Rev.*, vol. 38, no. 1, pp. 185–196, 2009.

196. R. a. Marcus and N. Sutin, "Electron transfers in chemistry and biology," *Biochim. Biophys. Acta - Rev. Bioenerg.*, vol. 811, pp. 265–322, 1985.

197. G. L. Closs and J. R. Miller, "Intramolecular long-distance electron transfer in organic molecules.," *Science*, vol. 240, pp. 440–447, 1988.

198. M. R. Wasielewski, "Photoinduced electron transfer in supramolecular systems for artificial photosynthesis," *Chem. Rev.*, vol. 92, pp. 435–461, 1992.

199. A. C. Benniston and A. Harriman, "Charge on the move: how electron-transfer dynamics depend on molecular conformation.," *Chem. Soc. Rev.*, vol. 35, pp. 169–179, 2006.

200. S. S. Skourtis, D. H. Waldeck, and D. N. Beratan, "Fluctuations in biological and bioinspired electron-transfer reactions.," *Annu. Rev. Phys. Chem.*, vol. 61, pp. 461–485, 2010.

201. Y. K. Kang, P. M. Iovine, and M. J. Therien, "Electron transfer reactions of rigid, cofacially compressed, $\pi$-stacked porphyrin-bridge-quinone systems," 2011.

202. N. P. Redmore, I. V. Rubtsov, and M. J. Therien, "Synthesis, excited-state dynamics, and reactivity of a directly-linked pyromellitimide - (Porphinato)zinc(II) complex," *Inorg. Chem.*, vol. 41, pp. 566–570, Feb. 2002.

203. N. P. Redmore, I. V. Rubtsov, and M. J. Therien, "Synthesis, electronic structure, and electron transfer dynamics of (aryl)ethynyl-bridged donor-acceptor systems," *J. Am. Chem. Soc.*, vol. 125, pp. 8769–8778, 2003.

204. T. L. Benanti and D. Venkataraman, "Organic solar cells: An overview focusing on active layer morphology," 2006.

205. S. Günes, H. Neugebauer, and N. S. Sariciftci, "Conjugated polymer-based organic solar cells," 2007.

206. P. V. Kamat, "Meeting the clean energy demand: Nanostructure architectures for solar energy conversion," *J. Phys. Chem. C*, vol. 111, pp. 2834–2860, 2007.

207. J. Roncali, P. Leriche, and A. Cravino, "From one- to three-dimensional organic semiconductors: In search of the organic silicon?," *Adv. Mater.*, vol. 19, pp. 2045–2060, 2007.

208. P. Heremans, D. Cheyns, and B. P. Rand, "Strategies for increasing the efficiency of heterojunction organic solar cells: material selection and device architecture.," *Acc. Chem. Res.*, vol. 42, pp. 1740–1747, 2009.

209. K. Kalyanasundaram and M. Graetzel, "Artificial photosynthesis: Biomimetic approaches to solar energy conversion and storage," 2010.

210. N. Mataga, S. Taniguchi, H. Chosrowjan, A. Osuka, and N. Yoshida, "Ultrafast charge transfer and radiationless relaxations from higher excited state (S2) of directly linked Zn-porphyrin (ZP)-acceptor dyads: Investigations into fundamental problems of exciplex chemistry," *Chem. Phys.*, vol. 295, pp. 215–228, 2003.

211. N. Yoshida, T. Ishizuka, K. Yofu, M. Murakami, H. Miyasaka, T. Okada, Y. Nagata, A. Itaya, H. S. Cho, D. Kim, and A. Osuka, "Synthesis of directly linked zinc(II) porphyrin-imide dyads and energy gap dependence of intramolecular electron transfer reactions," *Chem. - A Eur. J.*, vol. 9, pp. 2854–2866, 2003.

212. B. Apostolovic, M. Danial, and H.-A. Klok, "Coiled coils: attractive protein folding motifs for the fabrication of self-assembled, responsive and bioactive materials.," *Chem. Soc. Rev.*, vol. 39, pp. 3541–3575, 2010.

213. J. H. Fuhrhop, "Porphyrin assemblies and their scaffolds," *Langmuir*, vol. 30, pp. 1–12, 2014.

214. D. N. Woolfson, "The design of coiled-coil structures and assemblies," 2005.

215. G. Grigoryan and A. E. Keating, "Structural specificity in coiled-coil interactions," 2008.

216. I. L. Alberts, K. Nadassy, and S. J. Wodak, "Analysis of zinc binding sites in protein crystal structures.," *Protein Sci.*, vol. 7, pp. 1700–1716, Aug. 1998.

217. S. A. Kang, P. J. Marjavaara, and B. R. Crane, "Electron transfer between cytochrome c and cytochome c peroxidase in single crystals," *J. Am. Chem. Soc.*, vol. 126, pp. 10836–10837, 2004.

218. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *J. Chem. Phys.*, vol. 21, pp. 1087–1092, 1953.

219. U. Hobohm and C. Sander, "Enlarged representative set of protein structures.," *Protein Sci.*, vol. 3, pp. 522–524, 1994.

220. A. a. Canutescu and R. L. Dunbrack, "Cyclic coordinate descent: A robotics algorithm for protein loop closure.," *Protein Sci.*, vol. 12, pp. 963–72, May 2003.

221. W. Boomsma and T. Hamelryck, "Full cyclic coordinate descent: solving the protein loop closure problem in Calpha space.," *BMC Bioinformatics*, vol. 6, p. 159, 2005.

222. D. J. Mandell, E. A. Coutsias, and T. Kortemme, "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling.," 2009.

223. D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. L. Dunbrack, "Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model," *PLoS Comput. Biol.*, vol. 6, 2010.

224. B. W. Matthews, "Solvent content of protein crystals.," *J. Mol. Biol.*, vol. 33, pp. 491–497, 1968.

225. M. Wu, P. Kaur, H. Yue, A. Clemmens, and D. H. Waldeck, "Charge density effects on the aggregation properties of poly(p-phenylene-ethynylene)-based anionic polyelectrolytes," *J. Phys. Chem. B*, vol. 112, pp. 3300–3310, 2008.

226. Y. Xia, X. Deng, L. Wang, X. Li, X. Zhu, and Y. Cao, "An extremely narrow-band-gap conjugated polymer with heterocyclic backbone and its use in optoelectronic devices," *Macromol. Rapid Commun.*, vol. 27, no. 15, pp. 1260–1264, 2006.

227. C. Hoven, R. Yang, A. Garcia, A. J. Heeger, T.-Q. Nguyen, , and G. C. Bazan, "Ion motion in conjugated polyelectrolyte electron transporting layers," *J. Am. Chem. Soc.*, vol. 129, no. 36, pp. 10976–10977, 2007.

228. S. W. T. III, G. D. Joly, , and T. M. Swager, "Chemical sensors based on amplifying fluorescent conjugated polymers," *Chem. Rev.*, vol. 107, no. 4, pp. 1339–1386, 2007.

229. J. Tour, "Molecular electronics. synthesis and testing of components," *Acc. Chem. Res.*, vol. 33, no. 11, pp. 791–804, 2000.

230. J. Kim and T. Swager, "Control of conformational and interpolymer effects in conjugated polymers," *Nature*, vol. 411, no. 6841, pp. 1030–1034, 2001.

231. H. B. Uwe, "Poly(aryleneethynylene)s," *Macromol. Rapid Commun.*, vol. 30, no. 9-10, pp. 773–805, 2009.

232. T. Andrew and T. Swager, "Structure-property relationships for exciton transfer in conjugated polymers," *J. Polym. Sci., Part B: Polym. Phys.*, vol. 49, no. 7, pp. 476–498, 2011.

233. S. Maskey, F. Pierce, D. Perahia, and G. S. Grest, "Conformational study of a single molecule of poly para phenylene ethynylenes in dilute solutions," *J. Chem. Phys.*, vol. 134, no. 24, p. 24906, 2011.

234. G. Jeschke, M. Sajid, M. Schulte, N. Ramezanian, A. Volkov, H. Zimmermann, and A. Godt, "Flexibility of shape-persistent molecular building blocks composed of p-phenylene and ethynylene units," *J. Aam. Chem. Soc.*, vol. 132, pp. 10107–10117, 2010.

235. P. Kaur, H. Yue, M. Wu, M. Liu, J. Treece, and D. H. Waldeck, "Solvation and aggregation of polyphenylethynylene based anionic polyelectrolytes in dilute solutions," *J. Phys. Chem. B*, vol. 111, pp. 8589–8596, 2007.

236. Z. Zhang, Y. Che, R. A. Smaldone, M. Xu, B. R. Bunes, J. S. Moore, and L. Zang, "Reversible dispersion and release of carbon nanotubes using foldable oligomers," *J. Am. Chem. Soc.*, vol. 132, pp. 14113–14117, 2010.

237. S. Iijima and T. Ichihashi, "Single-shell carbon nanotubes of 1-nm diameter," *Nature*, vol. 363, no. 6430, pp. 603–605, 1993.

238. D. Bethune, C. Kiang, M. Devries, G. Gorman, R. Savoy, J. Vazquez, and R. Beyers, "Cobalt-catalyzed growth of carbon nanotubes with single-atom-layerwalls," *Nature*, vol. 363, no. 6430, pp. 605–607, 1993.

239. H. Chen, Q. Xue, Q. Zheng, J. Xie, and K. Yan, "Influence of nanotube chirality, temperature, and chemical modification on the interfacial bonding between carbon nanotubes and polyphenylacetylene," *J. Phys. Chem. C*, vol. 112, pp. 16514–16520, 2008.

240. J. A. Misewich, R. Martel, P. Avouris, J. C. Tsang, S. Heinze, and J. Tersoff, "Electrically induced optical emission from a carbon nanotube fet," *Science*, vol. 300, pp. 783–786, 2003.

241. D. Tasis, N. Tagmatarchis, A. Bianco, , and M. Prato, "Chemistry of carbon nanotubes," *Chem. Rev.*, vol. 106, pp. 1105–1136, 2006.

242. Y.-P. Sun, K. Fu, Y. Lin, and W. Huang, "Functionalized Carbon Nanotubes: Properties and Applications," *Acc. Chem. Res.*, vol. 35, pp. 1096–1104, Dec. 2002.

243. C. A. Dyke and J. M. Tour, "Covalent Functionalization of Single-Walled Carbon Nanotubes for Materials Applications," *J. Phys. Chem. A*, vol. 108, no. 51, pp. 11151–11159, 2004.

244. B. R. Goldsmith, J. G. Coroneus, V. R. Khalap, A. A. Kane, G. A. Weiss, and P. G. Collins, "Conductance-Controlled Point Functionalization of Single-Walled Carbon Nanotubes," *Science*, vol. 315, pp. 77–81, Jan. 2007.

245. J. García-Lastra, K. Thygesen, M. Strange, and Á. Rubio, "Conductance of Sidewall-Functionalized Carbon Nanotubes: Universal Dependence on Adsorption Sites," *Physical Review Letters*, vol. 101, p. 236806, Dec. 2008.

246. A. López-Bezanilla, F. Triozon, S. Latil, X. Blase, and S. Roche, "Effect of the Chemical Functionalization on Charge Transport in Carbon Nanotubes at the Mesoscopic Scale," *Nano letters*, vol. 9, pp. 940–944, Mar. 2009.

247. S. Deng, Y. Zhang, A. H. Brozena, M. L. Mayes, P. Banerjee, W.-A. Chiou, G. W. Rubloff, G. C. Schatz, and Y. Wang, "Confined Propagation of Covalent Chemical Reactions on Single-Walled Carbon Nanotubes," *Nature Communications*, vol. 2, pp. 382–, July 2011.

248. J. Chen, H. Liu, W. A. Weimer, M. D. Halls, D. H. Waldeck, and G. C. Walker, "Noncovalent engineering of carbon nanotube surfaces by rigid, functional conjugated polymers," *J. Am. Chem. Soc.*, vol. 124, pp. 9034–9035, 2002.

249. M. S. Arnold, A. A. Green, J. F. Hulvat, S. I. Stupp, and M. C. Hersam, "Sorting Carbon Nanotubes by Electronic Structure using Density Differentiation," *Nat. Nanotechnol.*, vol. 1, pp. 60–65, Oct. 2006.

250. Y.-L. Zhao and J. F. Stoddart, "Noncovalent Functionalization of Single-Walled Carbon Nanotubes.," *Acc. Chem. Res.*, vol. 42, pp. 1161–1171, Aug. 2009.

251. T. Premkumar, R. Mezzenga, and K. E. Geckeler, "Carbon Nanotubes in the Liquid Phase: Addressing the Issue of Dispersion.," *Small*, vol. 8, pp. 1299–1313, May 2012.

252. M. J. O'Connell, S. M. Bachilo, C. B. Huffman, V. C. Moore, M. S. Strano, E. H. Haroz, K. L. Rialon, P. J. Boul, W. H. Noon, C. Kittrell, and et al., "Band gap fluorescence from individual single-walled carbon nanotubes," *Science*, vol. 297, pp. 593–596, 2002.

253. S. Bandow, A. M. Rao, K. A. Williams, A. Thess, R. E. Smalley, and P. C. Eklund, "Purification of single-wall carbon nanotubes by microfiltration," *J. Phys. Chem. B*, vol. 101, no. 44, pp. 8839–8842, 1997.

254. M. J. O'Connell, P. Boul, L. M. Ericson, C. Huffman, Y. Wang, E. Haroz, C. Kuper, J. Tour, K. D. Ausman, and R. E. Smalley, "Reversible water-solubilization of single-walled carbon nanotubes by polymer wrapping," *Chem. Phys. Lett.*, vol. 342, pp. 265–271, 2001.

255. M. Panhuis, A. Maiti, A. B. Dalton, A. van den Noort, J. N. Coleman, B. McCarthy, and W. J. Blau, "Selective interaction in a polymer-single-wall carbon nanotube composite.," *J. Phys. Chem. B*, vol. 107, pp. 478–482, 2003.

256. A. Star, J. C. P. Gabriel, K. Bradley, and G. Gruner, "Electronic detection of specific protein binding using nanotube fet devices," *Nano Lett.*, vol. 3, pp. 459–463, 2003.

257. A. Star, J. F. Stoddart, D. Steuerman, M. Diehl, A. Boukai, E. W. Wong, X. Yang, S. W. Chung, H. Choi, and J. R. Heath, "Preparation and properties of polymer-wrapped single-walled carbon nanotubes," *Angew. Chem., Int. Ed.*, vol. 40, pp. 1721–1725, 2001.

258. V. Zorbas, A. Ortiz-Acevedo, A. B. Dalton, M. M. Yoshida, G. R. Dieckmann, R. K. Draper, R. H. Baughman, M. Jose-Yacaman, and I. H. Musselman, "Preparation and characterization of individual peptide-wrapped single-walled carbon nanotubes," *J. Am. Chem. Soc.*, vol. 126, pp. 7222–7227, 2004.

259. G. Grigoryan, Y. H. Kim, R. Acharya, K. Axelrod, R. M. Jain, L. Willis, M. Drndic, J. M. Kikkawa, and W. DeGrado, "Computational design of virus-like protein assemblies on carbon nanotube surfaces," *Science*, vol. 332, no. 1071-1076, 2011.

260. S. S. Karajanagi, H. Yang, P. Asuri, E. Sellitto, J. S. Dordick, and R. S. Kane, "Protein-Assisted Solubilization of Single-Walled Carbon Nanotubes," *Langmuir*, vol. 22, pp. 1392–1395, Feb. 2006.

261. C. Staii and A. T. Johnson, "DNA-decorated Carbon Nanotubes for Chemical Sensing," *Nano Lett.*, vol. 5, no. 9, pp. 1774–1778, 2005.

262. H. Cathcart, V. Nicolosi, J. M. Hughes, W. J. Blau, J. M. Kelly, S. J. Quinn, and J. N. Coleman, "Ordered dna wrapping switches on luminescence in single-walled nanotube dispersions," *J. Am. Chem. Soc.*, vol. 130, pp. 12734–12744, 2008.

263. X. Tu, S. Manohar, A. Jagota, and M. Zheng, "Dna sequence motifs for structure-specific recognition and separation of carbon nanotubes," *Nature*, vol. 460, pp. 250–253, 2009.

264. M. Zheng, A. Jagota, M. S. Strano, A. P. Santos, P. Barone, S. G. Chou, B. A. Diner, M. S. Dresselhaus, R. S. McLean, G. B. Onoa, and et al., "Structure-based carbon nanotube sorting by sequence-dependent dna assembly," *Science*, vol. 302, pp. 1545–1548, 2003.

265. M. Zheng, A. Jagota, E. D. Semke, B. A. Diner, R. S. McLean, S. R. Lustig, R. E.

Richardson, and N. G. Tassi, "Dna-assisted dispersion and separation of carbon nanotubes," *Nat. Mater.*, vol. 2, pp. 338–342, 2003.

266. R. Bandyopadhyaya, E. Nativ-Roth, O. Regev, and R. Yerushalmi-Rozen, "Stabilization of Individual Carbon Nanotubes in Aqueous Solutions," *Nano Lett.*, vol. 2, pp. 25–28, Jan. 2002.

267. N. Minami, Y. Kim, K. Miyashita, S. Kazaoui, and B. Nalini, "Cellulose derivatives as excellent dispersants for single-walled carbon nanotubes as demonstrated by absorption and photoluminescence spectroscopy," *Appl. Phys. Lett.*, vol. 88, pp. 093123–093126, 2006.

268. Y. Liu, P. Liang, H.-Y. Zhang, and D.-S. Guo, "Cation-Controlled Aqueous Dispersions of Alginic-Acid-Wrapped Multi-Walled Carbon Nanotubes," *Small*, vol. 2, pp. 874–878, July 2006.

269. M. Numata, M. Asai, K. Kaneko, A.-H. Bae, T. Hasegawa, K. Sakurai, and S. Shinkai, "Inclusion of cut and as-grown single-walled carbon nanotubes in the helical superstructure of schizophyllan and curdlan ($\beta$-1,3-glucans)," *J. Am. Chem. Soc.*, vol. 127, pp. 5875–5884, 2005.

270. A. Ikeda, K. Nobusawa, T. Hamano, and J.-i. Kikuchi, "Single-walled carbon nanotubes template the one-dimensional ordering of a polythiophene derivative," *Org. Lett.*, vol. 8, pp. 5489–5492, 2006.

271. J.-Y. Hwang, A. Nish, J. Doig, S. Douven, C.-W. Chen, L.-C. Chen, and R. J. Nicholas, "Polymer structure and solvent effects on the selective dispersion of single-walled carbon nanotubes," *J. Am. Chem. Soc.*, vol. 130, pp. 3543–3553, 2008.

272. Z. Zhang, Y. Che, R. A. Smaldone, M. Xu, B. R. Bunes, J. S. Moore, and L. Zang, "Reversible dispersion and release of carbon nanotubes using foldable oligomers," *J. Am. Chem. Soc.*, vol. 132, no. 40, pp. 14113–14117, 2010.

273. S. Velayudham, C. H. Lee, M. Xie, D. Blair, N. Bauman, Y. K. Yap, S. A. Green, and H. Liu, "Noncovalent functionalization of boron nitride nanotubes with poly(p-phenylene-ethynylene)s and polythiophene," *ACS Appl. Mater. Interfaces*, vol. 2, no. 1, pp. 104–110, 2010.

274. Z. Yang, Z. Wang, X. Tian, P. Xiu, and R. Zhou, "Amino Acid Analogues Bind to Carbon Nanotube via 1-1 Interactions: Comparison of Molecular Mechanical and Quantum Mechanical Calculations.," *J. Chem. Phys.*, vol. 136, p. 025103, Jan. 2012.

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

276. J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the

cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, no. 3, pp. 327–341, 1976.

277. H. C. Andersen, "Rattle: A velocity version of the shake algorithm for molecular dynamics calculations," *J. Chem. Phys.*, vol. 52, no. 1, pp. 24–34, 1982.

278. T. Darden, D. York, and L. Pedersen, "Particle mesh ewald: An n.log(n) method for ewald sums in large systems," *J. Chem. Phys.*, vol. 98, pp. 10089–10093, 1993.

279. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, pp. 926–935, 1983.

280. W. Humphrey, A. Dalke, and K. Schulten, "Vmd: Visual molecular dynamics," *J. Mol. Graphics Modell.*, vol. 14, no. 1, pp. 33–38, 1993.

281. S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks, "Constant pressure molecular dynamics simulation: The langevin piston method," *J. Chem. Phys.*, vol. 103, pp. 4613–4621, 1995.

282. O.-S. Lee and J. G. Saven, "Simulation studies of a helical m-phenylne ethylene foldamer," *J. Phys. Chem. B*, vol. 108, pp. 11988–11994, 2004.

283. D. C. Spellmeyer, P. D. J. Grootenhuis, M. D. Miller, L. F. Kuyper, and P. A. Kollman, "Theoretical investigations of the rotational barrier in anisole: An ab initio and molecular dynamics study," *J. Phys. Chem.*, vol. 94, pp. 4483–4491, 1990.

284. H. Lee, R. M. Venable, A. D. MacKerell Jr., and R. W. Pastor, "Molecular dynamics studies of polyethylene oxide and polyethylene glycol: Hydrodynamic radius and shape anisotropy," *Biophys. J.*, vol. 95, pp. 1590–1599, 2008.

285. Robert R. Johnson, "Nanotube Builder 1.0: A plug-in to generate carbon nanotubes within Visual Molecular Dynamics".

286. A. D. MacKerell Jr., D. Bashford, M. Bellott, J. R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, and et al., "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, 1998.

287. E. Darve and A. Pohorille, "Calculating free energies using average force," *J. Chem. Phys.*, vol. 115, no. 20, pp. 9169–9183, 2001.

288. D. Rodriguez-Gomez, E. Darve, and A. Pohorille, "Assessing the efficiency of free energy calculation methods," *J. Chem. Phys.*, vol. 120, no. 8, pp. 3563–3578, 2004.

289. J. Henin, G. Fiorin, C. Chipot, and M. L. Klein, "Exploring multidimensional free energy landscapes using time-dependent biases on collective variables," *J. Chem. Theory Comput.*, vol. 6, no. 1, pp. 35–47, 2010.

290. Y. Nievergelt, "Fitting helices to data by total least squares," *Comput.-Aided Geom. Des.*, vol. 14, no. 8, pp. 707–718, 1997.

291. C. Chiu, G. R. Dieckmann, and S. O. Nielsen, "Molecular dynamics study of a nanotube-binding amphiphilic helical peptide at different water/hydrophobic interfaces," *J. Phys. Chem. B*, vol. 112, pp. 16326–16333, 2008.

292. A. Shrake and J. A. Rupley, "Environment and exposure to solvent of protein atoms. lysozyme and insulin," *J. Mol. Biol.*, vol. 79, pp. 351–364, 1973.

293. P. J. Flory, *Principles of Polymer Chemistry.* Cornell University Press, 18 ed., December 1953.

294. L. D. Hall, G. Orpen, M. Pilkington, and J. D. Wallis, "Molecular Distortions in Crystalline 2-phenylethynylbenzoic acid," *J. Chem. Crystallogr.*, vol. 31, no. 2, pp. 97–103, 2001.

295. M. Pilkington, J. D. Wallis, G. T. Smith, and J. A. K. Howard, "Geometry Distorting Intramolecular Interactions to an Alkyne Group in 1-(2-aminophenyl)-2-(2-nitrophenyl)ethyne: a Joint Experimental-Theoretical Study," *J. Chem. Soc., Perkin Trans. 2*, no. 9, p. 1849, 1996.

296. K. Okuyama, T. Hasegawa, M. Ito, and N. Mikami, "Electronic Spectra of Tolan in a Supersonic Free Jet: Large-Amplitude Torsional Motion," *J. Phys. Chem.*, vol. 88, pp. 1711–1716, Apr. 1984.

297. S. Saebø, J. Almlöf, J. E. Boggs, and J. G. Stark, "Two Approaches to the Computational Determination of Molecular Structure: the Torsional Angle in Tolane and the Effect of Fluorination on the Structure of Oxirane," *J. Mol. Struct. (Theochem.)*, vol. 200, pp. 361–373, Oct. 1989.

298. S. J. Greaves, E. L. Flynn, E. L. Futcher, E. Wrede, D. P. Lydon, P. J. Low, S. R. Rutter, and A. Beeby, "Cavity Ring-Down Spectroscopy of the Torsional Motions of 1,4-Bis(phenylethynyl)benzene," *J. Phys. Chem. A*, vol. 110, pp. 2114–2121, Feb. 2006.

299. S.-g. Kang, G. Zhou, P. Yang, Y. Liu, B. Sun, T. Huynh, H. Meng, L. Zhao, G. Xing, C. Chen, and et al., "Molecular mechanism of pancreatic tumor metastasis inhibition by gd@c-82(oh)(22) and its implication for de novo design of nanomedicine," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, pp. 15431–15436, 2012.

300. J. Onuchic, Z. LutheySchulten, and P. Wolynes, "Theory of Protein Folding: The Energy Landscape Perspective," *Annu. Rev. Phys. Chem.*, vol. 48, pp. 545–600, 1997.

301. H. Lammert, P. G. Wolynes, and J. N. Onuchic, "The Role of Atomic Level Steric Effects and Attractive Forces in Protein Folding," *Proteins-Structure Function And Bioinformatics*, vol. 80, pp. 362–373, FEB 2012.

302. J. Nelson, J. Saven, J. Moore, and P. Wolynes, "Solvophobically Driven Folding of Nonbiological Oligomers," *Science*, vol. 277, pp. 1793–1796, SEP 19 1997.

303. D. Hill, M. Mio, R. Prince, T. Hughes, and J. Moore, "A Field Guide to Foldamers," *Chem. Rev.*, vol. 101, pp. 3893–4011, DEC 2001.

304. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, . Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, "Gaussian09 Revision D.01." Gaussian Inc. Wallingford CT 2009.

305. M. Nishizaka, T. Mori, and Y. Inoue, "Axial chirality of donor-donor, donor-acceptor, and tethered 1,1 '-binaphthyls: A theoretical revisit with dynamics trajectories," *J. Phys. Chem. A*, vol. 115, no. 21, pp. 5488–5495, 2011.

306. M. Nishizaka, T. Mori, and Y. Inoue, "Fitting helices to data by total least squares," *J. Phys. Chem. Lett.*, vol. 1, no. 12, pp. 1809–1812, 2010.

307. L. Di Bari, G. Pescitelli, and P. Salvadori, "Conformational study of 2,2'-homosubstituted 1,1'-binaphthyls by means of uv and cd spectroscopy," *J. Am. Chem. Soc.*, vol. 121, pp. 7998–8004, 1999.

308. P. Lustenberger, E. Martinborough, T. M.Denti, and et al., "Geometrical optimisation of 1,1'-binaphthalene receptors for enantioselective molecular recognition of excitatory amino acid derivatives," *J. Chem. Society-Perkin Trans.*, vol. 4, pp. 747–761, 1998.

309. Robert R. Johnson, "Nanotube Builder 1.0: A plug-in to generate carbon nanotubes within Visual Molecular Dynamics".