2018

Computational Design Of Protein–ligand And Protein–protein Interactions

Jose Abraham Villegas
University of Pennsylvania, jahvillegas@gmail.com

Computational Design Of Protein–ligand And Protein–protein Interactions

Abstract
Central to the function of proteins is the concept of molecular recognition. Protein–ligand and protein–protein interactions make up the bulk of the chemical processes that give rise to living things. Realizing the full potential of protein design technology will therefore require an increased understanding of the design principles of molecular recognition. We have tackled problems involving molecular recognition by using computational methods to design novel protein-ligand and protein-protein interactions. Firstly, we set out to design a protein capable of recognizing lanthanide metal ions. Protein-lanthanide systems are of interest for their potential to serve as purification agents for use under biological conditions. We have designed a highly dense 6-coordinate lanthanide binding at the core of a de novo protein, and used the dynamical aspects of the protein to achieve a degree of differentiation between elements in the lanthanide series. Secondly, we investigated systems of homo-oligomeric protein complexes that self-assemble into hollow cages. We have studied the structural determinants of naturally occurring self-assembling ferritin cages and identified a single mutation that greatly increased the stability of the ferritin cage, as well as dramatically altered the overall structure of the assembly. We have also used the formulation of probabilistic protein design to arrive at novel sequences for α-helical peptides that can adjust their surfaces in accordance to different local environments. This formulation was used to identify a sequence for a peptide designed to self-assemble into a spherical particle with broken symmetry. Taken together, these efforts will lead to an increased understanding of the role of kinetics and structural plasticity in protein-ligand and protein-protein interactions.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Chemistry

First Advisor
Jeffery G. Saven

Subject Categories
Chemistry

COMPUTATIONAL DESIGN OF PROTEIN–LIGAND AND PROTEIN–PROTEIN INTERACTIONS

José Abraham Villegas

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

2018

Supervisor of Dissertation

_____

Jeffery G. Saven, Professor of Chemistry

Graduate Group Chairperson

_____

Gary A. Molander, Hirschmann-Makineni Professor of Chemistry

Dissertation Committee

David W. Christianson, Roy and Diana Vagelos Professor in Chemistry and Chemical Biology

Eric J. Schelter, Associate Professor of Chemistry

Joseph Subotnik, Professor of Chemistry

COMPUTATIONAL DESIGN OF PROTEIN-LIGAND AND PROTEIN-PROTEIN INTERACTIONS

ACKNOWLEDGMENT

I would like to first of all thank my thesis advisor Professor Jeff Saven. Witnessing your unfaltering commitment to scientific rigor has been the best training I could have hoped for. Thank you for teaching me to never compromise on obtaining a full understanding of any topic at hand.

Thanks to the members of my thesis committee, Professors David Christianson, Eric Schelter, and Joe Subotnik. Your genuine engagement with my research projects and my scientific career has been invaluable. I am indebted to you for helping me to take stock of my graduate education.

Thanks to the professors with whom I have had the privilege to collaborate. Thank you to Professor Ivan Dmochowski. Your sense of vision has allowed me to expand the way I think about science, and has pushed me to be more creative and imaginative. Thanks to Professor Darrin Pochan for your tenacity and can-do attitude. Thanks to the students who did the bulk of the experimental work, Dr. Katie Pulsipher and Nairiti Sinha, whose ingenuity and hard work have been awe-inspiring.

Thanks to my fellow lab members with whom I have shared this experience: Dr. Chris MacDermaid, Dr. Chris Lanci, Dr. Lu Gao, Dr. Chris Von Bargen, Dr. Huixi "Violet" Zhang, Dr. Matthew Eibling, Dr. Krishna Vijaydendran, Dr. Wenhao Liu, Jacquelyn Blum, and Rui "Grey" Guo. You have made this place a vibrant place to do research. I am especially grateful to Dr. Eibling for training me on the experimental techniques needed to carry out my work. Thanks to Jacque Faylo of the Christianson group, whose

work during her first-year rotation in our lab helped to get my experiments off the ground.

Thanks to Dr. José Manuel Perez Aguilar, Dr. Brenda Leonor Sanchez Gaytán, and Dr. Anne Wagner for welcoming into the Penn Chemistry community. Thanks to Dr. Osvaldo Gutierrez your constant encouragement and assurances.

Thanks to the members of the Association for Cultural Diversity in Chemistry: Ben Roose, Lukman Solola, Jennifer Matsui, Nicole Bellonzi, Jacquelyn Blum, Joo "Vicky" Jun, Shuai Zheng, and Dee Mukherjee. I have felt truly honored to have worked with you and to have gotten to know you.

Thanks to my high school chemistry teacher, Ms Diane McGanne. You opened up the world of chemistry for me. I still think of chemical equilibrium as two neighbors throwing oranges across a fence. Thanks to my high school physics teacher, Dr. Gary Reynolds. You taught me how to do numerical calculations, and to "not let anyone beat you with a calculator."

Thanks to my mother and grandmother for teaching me the importance of education and hard work. Because of you, I always push myself to do the best I can.

Thanks to Eli and Arlene Ratzabi for welcoming me into your family with open arms. Thanks to Limor, Josh, Danit, Kfir, Jude, Noa, Dean, Liam, and Mia for being so awesome.

Thanks to my sons, Mateo and Emilio. Mateo, you are the best baby one could have hoped to have less than four months before a thesis defense. Your constant smiles have the power to melt any amount of stress away. You have somehow managed to make thesis writing easier, and I hope to finally get to know you better once it is all over. Emilio, you are an absolute joy. You are so amazingly bright and happy, and I am so proud of you for what you are and what I know you will become. Most importantly, thanks to my wife Hila. Every day I am overwhelmed by the happiness you bring, and the life you have made possible. I love you dearly.

ABSTRACT

COMPUTATIONAL DESIGN OF PROTEIN–LIGAND AND PROTEIN–PROTEIN INTERACTIONS

José Abraham Villegas

Jeffery G. Saven

Central to the function of proteins is the concept of molecular recognition. Protein–ligand and protein–protein interactions make up the bulk of the chemical processes that give rise to living things. Realizing the full potential of protein design technology will therefore require an increased understanding of the design principles of molecular recognition. We have tackled problems involving molecular recognition by using computational methods to design novel protein-ligand and protein-protein interactions. Firstly, we set out to design a protein capable of recognizing lanthanide metal ions. Protein-lanthanide systems are of interest for their potential to serve as purification agents for use under biological conditions. We have designed a highly dense 6-coordinate lanthanide binding at the core of a *de novo* protein, and used the dynamical aspects of the protein to achieve a degree of differentiation between elements in the lanthanide series. Secondly, we investigated systems of homo-oligomeric protein complexes that self-assemble into hollow cages. We have studied the structural determinants of naturally occurring self-assembling ferritin cages and identified a single mutation that greatly increased the stability of the ferritin cage, as well as dramatically altered the overall structure of the assembly. We have also

used the formulation of probabilistic protein design to arrive at novel sequences for α-helical peptides that can adjust their surfaces in accordance to different local environments. This formulation was used to identify a sequence for a peptide designed to self-assemble into a spherical particle with broken symmetry. Taken together, these efforts will lead to an increased understanding of the role of kinetics and structural plasticity in protein-ligand and protein-protein interactions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

1| Introduction to Computational Protein Design

1.1. Introduction

More than a century ago, Fischer[1] carried out the first synthesis of a dipeptide. Today, advances in computational protein design have opened up immense possibilities for the design of novel protein-based substances and materials. Computational protein design aims to identify novel amino acid sequences that are likely to fold into a target structure and possess a desired functionally. The computational and experimental tools developed in the fields of biophysics and of molecular biology can be repurposed and combined to design, synthesize, and characterize new proteins with no direct evolutionary connection to existing natural proteins. For example, the field of structural biology has developed computational tools to build molecular models of secondary structure elements such as α-helices[2] and random coils.[3] These tools are essential to solving the structure of proteins by X-ray crystallography since the technique requires building 3-dimensional models to calculate electron density maps from diffraction patterns.[4, 5] These tools can be repurposed to generate models of novel structures likely to be realizable with amino acid sequence selection. The wealth of information deposited in the protein data bank[6] (PDB) can be harnessed for empirical information, such as the propensity of an amino acid to be found in a buried or exposed region of the protein, to design proteins with features that are similar to those of natural proteins.[7] When combined with a theoretical understanding of the energetics of protein folding, these tools can be harnessed to arrive at sequences that are likely to fold to a target structure.[8]

From an experimental perspective, the tools developed for the synthesis and purification of peptides and proteins allow for the convenient realization of these molecules without extensive synthetic planning and execution. The spectroscopic tools that have been developed for the study of natural proteins can be used to carry out hypothesis-driven experiments that prove or disprove the success of a given design.

Central to the function of proteins is the concept of molecular recognition.[9] Proteins are capable of recognizing a plethora of small molecules,[10] carbohydrates,[11] membrane surfaces,[12] other proteins, and copies of themselves.[13, 14] Protein-ligand and protein-protein interactions make up the bulk of the chemical processes that give rise to living things. Enzymes must bind to substrate molecules to catalyze their conversion. Metallo-proteins bind to metals to take advantage of their catalytic properties and to regulate their availability.[15] Membrane protein channels maintain osmotic balance by regulating the passage of specific ions.[16] Proteins are also able to recognize other macromolecules, such as when binding to DNA during regulation of gene expression and gene replication.[17] Protein-protein interactions are responsible for the regulation of biological processes,[18] and for the formation of micro-compartments such as vaults,[19] ferritin,[20, 21] and protein capsids.[22, 23] Therefore, an understanding of the design principles of molecular recognition will be essential to realizing the full potential of *de novo* protein design technology.

1.2. Strategies for Generation of Novel Proteins

1.2.1. Directed Evolution

Given that our knowledge of protein-function relationships remains largely incomplete, one way to arrive at novel protein sequences is to adopt the strategy of randomized mutation and fitness selection used by living organisms. This strategy has been leveraged to create novel enzymes,[24] such as a protein that catalyzes the formation of carbon-silicon bonds.[25] The need to couple the randomization of protein sequences to fitness-based sequence selection requires highly specialized technology, which must be built and optimized for every new feature of interest. This technique, while able to generate proteins with native-like catalytic activity, does not lead to an increased understanding of how protein structure dictates function. More importantly, it is not clear that the space of the protein-fitness landscape is smooth and countinous.[26] That is to say, some chemical properties might lie on isolated regions of sequence space and not accessible through step-wise changes in amino-acid sequence.

1.2.2. Rational Protein Design

The experimental realization of *de novo* designed proteins was made possible with the advent of recombinant gene expression.[27] For the first time, arbitrary sequences of amino acids could be selected and synthesized at length scales inaccessible with solid-phase peptide synthesis. Based on the sequence-structure relationships learned from naturally occurring proteins such as myohemerythrin and cytochrome c′, the first design of a *de novo* full-length protein targeted a helical-bundle fold. Helical bundles consist of a coiled-coil motif of α-helices linked together by random-coil segments. The α-helices

associate with each other by the burial of hydrophobic residues at the core and through hydrophilic interactions at solvent exposed positions. While early designed proteins were able to adopt folds with well-defined secondary structure, these did not exhibit the stability of naturally occurring proteins.[27, 28]

The problem of stability was addressed by substituting hydrophobic amino acids at core position with histidine residues capable of binding to zinc.[29] While the objective was to modulate the dynamics of *de novo* designed proteins by incorporating coordination bonds in the protein core, this was the first instance of a protein-ligand interaction in a designed protein. The realization that these simple proteins could house elements of molecular recognition led to the idea of a protein maquette, a simplified protein capable of recovering the features of complex protein-ligand systems. Protein maquettes capable of binding to macro-cyclic co-factors, such as heme, were designed.[30]

Further progress in the field of protein design was enabled by the elucidation of sequence-structure relationships coiled-coils.[31] The periodicity of the coiled-coil motif creates a repeating pattern of seven amino acid positions termed a heptad repeat, and chemical properties can be assigned at each of the residue positions. It was found that coiled-coils could be constructed by simple patterning of the amino acid sequence, by placing hydrophobic residues at positions labeled *a* and *d*. Later studies placed importance on specific residue identities at positions *a* and *d*, as well as at positions *b*, *c*, *e*, and *g* in creating higher order oligomers.[32, 33]

Naturally occurring proteins commonly make use of dynamics to carry out their function. *De novo* proteins with dynamical features, such as a co-factor dependent conformational switch, have also been designed using a rational design strategy.[34]

1.2.3. Computational Protein Design

The understanding of structure-relationships in the coiled-coil motif makes it possible to design novel proteins with back-of-the-envelope selection of amino acid sequences. But to take full advantage of the quantitative information readily available from protein biophysics and structural biology, calculations need to be carried on a computing device. The first protein to be designed using a computational model was a zinc finger mimic.[35, 36] The design strategy took advantage of the advancements made in molecular dynamics simulations of protein, of algorithms designed to identify packing of amino acids in proteins, and of statistics collected on amino acid side chains conformations in crystal structures. During the course of the next twenty years, this strategy was refined and expanded. Successes to date include the design of novel protein folds,[37] proteins capable of binding to non-biological cofactors,[38] inhibitors of protein-protein interactions,[39] novel antibodies,[40] and the design of 2-dimensional[41, 42] and 3-dimensional protein-based materials.[43-46]

1.2.4. Mixed Approach to Protein Design

The techniques outlined above need not to be mutually exclusive. In practice, a combination of these techniques can be used to arrive at optimal sequences. It is common practice for the output from a protein design program to be visually inspected and the

amino acids identities at some sites selected based on the chemical intuition of the designer. Selected sequences can be further refined through directed evolution to improve the catalytic activity of designed enzymes. Directed evolution carried out on a designed protein will typically identify amino acid mutations located away from the binding site, underscoring the deficiencies that remain in the understanding of the relationship between protein structure and molecular recognition.[47]

1.3. Elements of Computational Protein Design

1.3.1 Foldable Backbone Targets

Regions of steric overlap restrict the torsional angles of protein backbone, restricting the conformational space that proteins occupy.[48, 49] Secondary structure elements are seen to be common among proteins, and even global folds are found to be redundant.[50] In designing a novel protein, the target fold needs to be compatible with the accessible space of backbone confirmations and be compatible with a large number of possible sequences.[8] This concept is known as the foldability criterion. An analysis of crystal structures of naturally occurring proteins can serve to identify regions of conformational space that are easily accessible by sequence selection. This does not mean that protein design needs to be confined to the space of known foldable structures. However, the selection of a target backbone scaffold from proteins with commonly observed folds is likely to result in increased probability of achieving successful *de novo* designs.[51] Although it can be difficult to quantify, there is also a relationship between protein stability and foldability. Experimental data on protein thermodynamic stability and kinetic stability can help in identifying foldable backbone targets.

1.3.2. Rotamer Libraries

In X-ray crystallography, the ability to narrow down the side chain degrees of freedom allows for better quality model building, increasing the chances that the crystallographic data can be fitted to the model.[52] Toward this purpose, Richards first compiled statistics on the most commonly observed torsional angles of side chains in proteins.[53] Later improvements were made to these so-called rotamer libraries as more structural information was obtained.[54, 55] These quantitative data can also be incorporated into the computational protein design process. By increasing the resolution beyond simple hydrophilic-hydrophobic patterning, amino acids can be selected that will populate states compatible with well-packed protein cores. Amino acid side chain conformations are also important in determining electrostatic interactions, hydrogen bonds, interactions between a protein and a ligand, and interactions across protein-protein interfaces.

1.3.3. Force Field Parameters

The large molecular sizes of proteins make the energetics of interatomic interactions intractable to *ab initio* quantum calculations. By modeling the atoms of protein as classical particles, the energetics can be approximated in a computationally efficient manner.[56, 57] The choice of force constants, known in molecular dynamics as force field parameters, applied to classical equations of motion, is important for recovering experimentally determined structural features of natural proteins. By fitting to a large number of known structures, these parameters could be optimized as to provide values of the energetics that are physically meaningful.[58]

The force fields developed from the study of natural proteins can be utilized in computational protein design to calculate the energetics of proposed structures. When combined with amino acid rotamer states, the energetics of amino acid pair interactions can be calculated. The energetics can be used in the formulations of protein design algorithms to search protein-sequence space for energetically favorable sequences.

1.3.4. Solvation Models

The force fields used in molecular dynamics are parameterized for explicit solvent, where the positions of water molecules and the interactions are explicitly calculated.[56, 59] In computational protein design, the energetics of possible protein structures are usually broken down to pairwise residue interactions. The energetics of protein sequences can be enumerated by calculating the energy as the sum of one-body and two-body interactions. Explicit solvent is not typically incorporated into protein design algorithms due to the difficulty of assigning atomic positions to water molecules in the calculation of pairwise interactions.[60] Implicit-solvent models have been developed that are based on solvent accessibility, but solvent accessibility is likewise difficult to decompose into pairwise interactions.[61, 62] However, a simple dampening of the electrostatic potential can be used to mimic the dielectric environment in the interior of proteins.[63] Alternatively, an empirically derived potential taken from analyzing naturally occurring structures can be derived.[7] By setting up an amino acid propensity based on solvent accessibility, the solvent-dependent sequence features of natural proteins can be recovered in *de novo* designed proteins.

1.3.5. Algorithms for Protein Design

Once a target folded structure has been identified and the parameters of the calculation have been specified, the next step is to identify a set of sequences likely to fold into the target conformation. If one considers all 20 amino acids at each of the *N* design positions, then the number of possible sequences is 20^N. This number grows astronomically large as the number of design positions increases, so that it is not possible to enumerate through all possible sequences. This number grows even larger if side chain conformational degrees of freedom are considered. Several algorithms were developed to narrow the search space and arrive at optimized sequences in an accessible timeframe. The dead-end elimination algorithm was originally intended for repacking of side chains on known backbone structures.[64] This algorithm was expanded for use in protein design by including rotamer states for different amino acids at each design positions.[35, 36] Recently, algorithms have been developed that can guarantee that the structure with the global energy minimum can be found.[65] Monte Carlo sampling can also be used to explore the sequence space within a narrow region.[66] This becomes necessary when an ensemble of sequences is sought, aside from just the global energy minimum structure.

The explicit enumeration through sequences poses a challenge. The problem of inaccuracies in the force field must be addressed directly by altering of the force field parameters. This can entail extensive optimizing of the weights that each force field component contributes, and adapting the parameters of the solvation models.[67] The use of discrete side chain conformations, coupled with the steepness of the van der Waals

repulsive term, means that this methodology is extremely sensitive to side-chain placement.[68]

An alternative to explicit search methods is to use a formulation derived from statistical mechanics. One can account for the inaccuracies in the force field, the solvation model, and the use of discreet side chain conformations by maximizing the entropy of the ensemble of all possible sequence configurations subject to an energetic constraint, thus extracting the most dominant features of the energetics. Sequences are then ranked by their probability to fold into the desired target, rather than by the full extent of the calculated interatomic interactions.

1.4. Ongoing Challenges in Protein Design

The computational design of ligand binding proteins has been faced with low success rates.[69] There is a need for improved algorithms for searching through a large number of coordination sphere geometries.[70] Additionally, the dynamical features of *de novo* designed proteins have not been fully exploited. While proteins have been designed that can switch between different conformational states, the ability to take advantage of local structural fluctuations has not been fully realized.

Proteins also possess a degree of surface plasticity made possible by a large number of conformational degrees of freedom available to solvent-exposed side chains. Protein design algorithms typically search for deep energetic minima that will confer a high degree of certainty as to the final configuration of the side chains. The probabilistic approach, however, accounts for structural fluctuations in the amino acid side chains to

average over all possible configurations. This probabilistic approach could be further exploited to take advantage of the plasticity of protein surfaces and to design sequences compatible with a variety of local environments.

1.5. Overview of Thesis

In this thesis I will describe the work I have carried out toward advancing the field of protein design. In chapter 2, I will describe the various levels of protein design and the mathematical formulations that go into creating novel proteins. In chapter 3, I will describe the work I have done toward the design of a lanthanide-binding protein. I will describe the algorithm developed for the search through a large number of structures for the identification of a highly dense 6-coordinate binding site. I will then describe the experimental realization of the protein and discuss how the dynamical aspects of the resulting protein were exploited to achieve a degree of differentiation between the lanthanides. In chapter 4, I will discuss the work performed to understand the structural determinants of naturally occurring self-assembling ferritin cages. I will discuss how a single mutation was identified that greatly increased the stability of the cage, as well as dramatically altered the overall structure of the assembly. By studying selected ferritin mutants, the energetics of a large conformational change could be separated from the energetics of a single interface interaction. In the last chapter I will discuss the effort to use the formulation of probabilistic protein design to arrive at sequences that can adjust their surfaces in accordance to different local environments. This formulation was used to identify a sequence for a peptide designed to self-assemble into a spherical particle with broken symmetry. Taken together, these efforts will lead to an increased understanding of

the role of kinetics and structural plasticity in protein-ligand and protein-protein interactions.

2 | Computational Methods in Protein Design

2.1. Introduction

The computational design of protein-ligand and protein-protein interactions encompasses multiple levels of structural detail. The design of a protein-ligand interaction begins with the selection of an arrangement of first-shell amino acids around the ligand of interest. The spatial orientation of the coordination sphere needs to be compatible with the global fold of the protein, and the sequence must fold into the target structure and be capable of populating the target amino acid conformations at the binding site. In the case of the design of protein-protein interactions, a set of favorable relative orientations between the backbones needs to be calculated, and the sequence must allow for the population of amino acid side chains on the interfacial region that leads to the association of the proteins in the target configuration.

2.2. Generation of Binding Site Geometries for Molecular Recognition

2.2.1 Grafting of Binding Sites

The energetics of sites of molecular recognition can be difficult to determine. This problem can be sidestepped by obtaining atomic coordinates for a binding site from an experimentally realized structure. The incorporation of a predetermined binding site into a designed protein is known as grafting.[71] For example, the binding site of a metal-binding protein can be incorporated into a minimal protein model. Binding site geometries can be taken from small molecule complexes and grafted onto a protein

scaffold. In addition to experimentally realized structures, geometry optimization can be carried out in an *ab initio* calculation and the resulting structure can be grafted onto a protein scaffold.[72] The design of novel enzymes is founded on the fact that proteins catalyze reactions by stabilizing the transition state.[73] However, transition state geometries are all but non-existent in crystallographic structures.[74] Grafting can be used to incorporate the predicted transition state binding-site geometry of a protein-ligand interaction into a predetermined protein scaffold, conferring enzymatic activity to the resulting protein.

Grafting can also be used in the design of protein-protein interactions. By taking the amino acid side chains and configurations present on a known protein-protein interface, an equivalent interface can be incorporated into a designed protein.[75] The resulting protein can then be used as a competitive inhibitor of a native protein-protein interaction.[39]

2.2.2. *In situ* Determination of Binding Site Geometries

As an alternative to grafting, binding-site geometries can be constructed directly from the rotamer states available to a given set of protein backbones.[76, 77] In this case, one could obtain information on the bond lengths and angles from an experimentally realized structure, and restrict those values in the search for ligand coordination spheres. The energetics of the side chain-ligand and side chain-side chain interactions can be determined with the use of atomic potentials. While these potentials are low resolution with respect to electron configuration, they provide an indication of the extent of van der

Waals overlap, electrostatic interaction, and hydrogen bonding potential. However, one important question arises. Are the torsional angles of the side chains driven away from their equilibrium positions by the inter-ligand interactions? To determine if this is the case, the distribution conformations of ligand-free and ligand-bound side chains can be examined in X-ray crystal structures. This analysis was carried out for calcium binding sites, and it was found that the conformations of amino acids in binding pockets do not veer away from otherwise populated states.[78]

2.3. Choosing Backbone Targets from Crystal Structures

Satisfying the foldability criterion is essential to the success of protein design.[8] Since foldability is a feature that is difficult to quantify, this problem can once again be sidestepped by obtaining atomic coordinates for target backbone structures from crystal structures of natural proteins. Selecting proteins with high stability, or with commonly observed folds, can target compatibility with a large number of sequences. This strategy has been employed extensively in the design of novel enzymes and protein-based materials.[79] When this strategy is combined with grafting of binding sites, the design problem can be reduced to that of ensuring compatibility between the coordination sphere and the protein backbone.

An entire protein X-ray crystal structure need not be taken as is. Secondary structure elements such as α-helices or random loops can be selected from a sub-segment of a protein structure and stitched together to construct backbones with novel global folds.[51] Protein segments can be stitched together by performing coordinate transformations on each sub-segment and calculating their optimal relative orientations. The foldability of

such targets can be assessed by a comparison of the geometric features of the designed backbones to those of natural proteins.[80] For example, when two sub-segments are stitched together, the resulting backbone torsional angles should fall within commonly observed ranges of torsional angles in natural proteins.

2.4. Parametric Description of Coiled-coils

Global features of protein folds can also be described parametrically.[81] The dimensionality of atomic positions can be reduced to a set of geometric parameters that can then be used to reconstruct the overall structure of the backbone. Varying the values of these parameters and calculating the atomic coordinates of the backbone atoms can be used to explore other regions of conformational space. Such alternative configurations can be scored using atomic potentials, or by comparison to the distribution of geometric parameters in naturally occurring proteins. One can infer that the more populated regions of conformational space will likely be compatible with larger numbers of sequences, and that configurations that diverge from frequently observed geometric parameters would be more difficult to achieve.

It was realized by Francis Crick that the positions of the atoms along α-helices in a coiled-coil could be described using a reduced set of parameters.[2] The idea for reducing the dimensionality of protein structures stemmed from the need to fit electron density to atomic positions during X-ray crystal structure determination. Since this procedure is carried out in Fourier space, the constant conversion of atomic coordinates to reciprocal coordinates is computationally expensive. By applying Fourier transformation to the parametric equations rather than to the individual atomic coordinates, the calculation of

structure factors could be simplified. Such application would operate on the whole set of atoms simultaneously rather than on individual atoms. The parametric equations used to describe coiled-coils can also be used in *de novo* protein design to generate novel coiled-coils by simply varying the values of the geometric parameters.

The atomic coordinates of an idealized coiled-coil can be described with the following set of equations:

$$x = R_0 \cos(\omega_0 t + \phi_0) + R_1 \cos(\omega_0 t + \phi_0)\cos(\omega_1 t + \phi_1) - R_1 \cos(\alpha)\sin(\omega_0 t + \phi_0)\sin(\omega_1 t + \phi_1)$$

$$y = R_0 \sin(\omega_0 t + \phi_0) + R_1 \sin(\omega_0 t + \phi_0)\cos(\omega_1 t + \phi_1) - R_1 \cos(\alpha)\cos(\omega_0 t + \phi_0)\sin(\omega_1 t + \phi_1) \quad (2.1)$$

$$z = \frac{\omega_0 R_0}{\tan(\alpha)} t - R_1 \sin(\alpha)\sin(\omega_1 t + \phi_1)$$

where R_0 is the superhelical radius, R_1 is the minorhelical radius, ω_0 is the superhelical frequency, ω_1 is the minorhelical frequency, ϕ_0 is the superhelical phase, ϕ_1 is the minorhelical phase, α is the pitch angle, and t is the residue position.

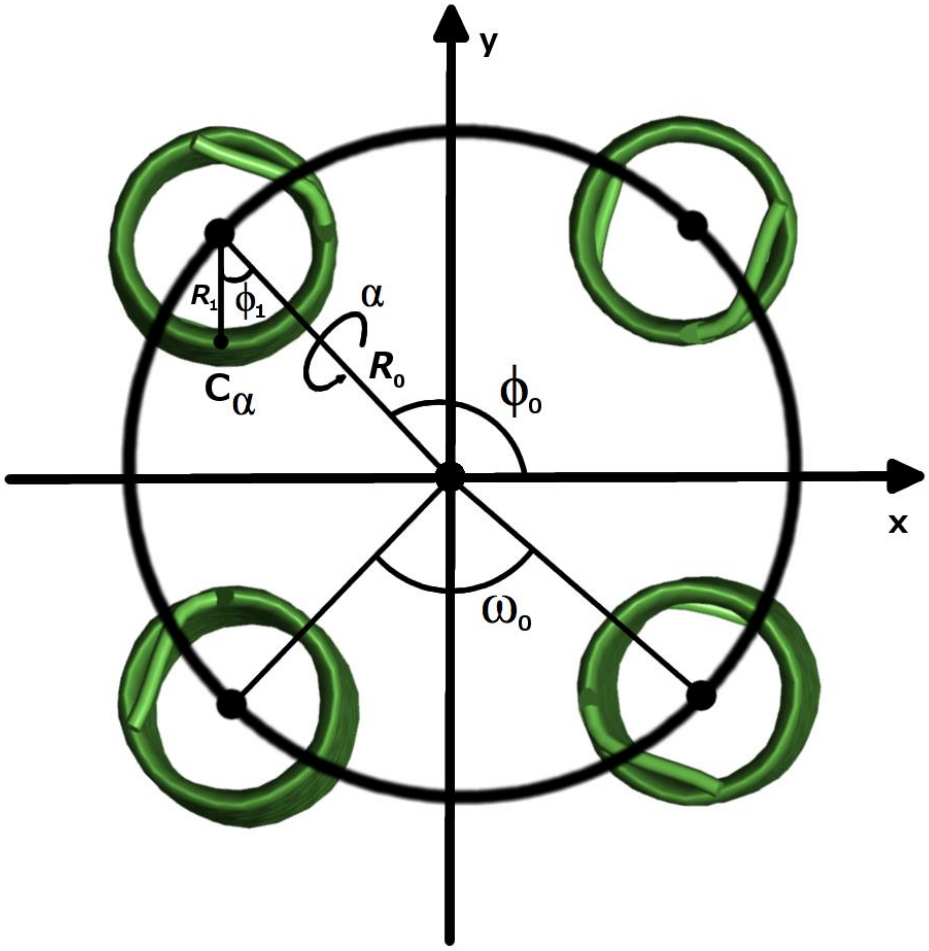

Figure 2.1: **Geometric parameters of coiled-coils.** R_0 is the superhelical radius, R_1 is the minorhelical radius, ω_0 is the superhelical frequency, ϕ_0 is the superhelical phase, ϕ_1 is the minorhelical phase, and α is the pitch angle.

In perfectly straight helices α is zero, so that the minorhelical axis is parallel to the superhelical axis. For R_0 to remain constant when α is non-zero, the minorhelical axis must wrap around the superhelical axis, resulting in curved helices. The distance it takes

for the minorhelical axis to complete one turn around the superhelical axis is called the pitch (*P*), and can be calculated from the pitch angle by:

$$P = \frac{2\pi R_0}{\tan(\alpha)} \tag{2.2}$$

Grigoryan and DeGrado added a modification to enable the displacement of the helices and decouple it from the major helical phase.[80] For each helix, the superhelical phase is modified as follows:

$$x = R_0 \cos(\omega_0 t + \phi'_0) + R_1 \cos(\omega_0 t + \phi'_0)\cos(\omega_1 t + \phi_1) - R_1 \cos(\alpha)\sin(\omega_0 t + \phi'_0)\sin(\omega_1 t + \phi_1)$$

$$y = R_0 \sin(\omega_0 t + \phi'_0) + R_1 \sin(\omega_0 t + \phi'_0)\cos(\omega_1 t + \phi_1) - R_1 \cos(\alpha)\cos(\omega_0 t + \phi'_0)\sin(\omega_1 t + \phi_1) \tag{2.3}$$

$$z = \frac{\omega_0 R_0}{\tan(\alpha)} t - R_1 \sin(\alpha)\sin(\omega_1 t + \phi_1) + \Delta z$$

$$\phi'_0 = \phi_0 + \frac{\Delta z \tan(\alpha)}{R_0}$$

where Δz is the axial displacement.

In coiled-coils with curved helices, the term “rise per residue” used to describe α-helices needs to be clarified. The rise per residue can be defined as the distance along the helical arc length that it takes to be in the plane of the next alpha carbon (termed the minor helical rise per residue) or the distance along the major helical axis that it takes to go up one residue (major helical rise per residue).

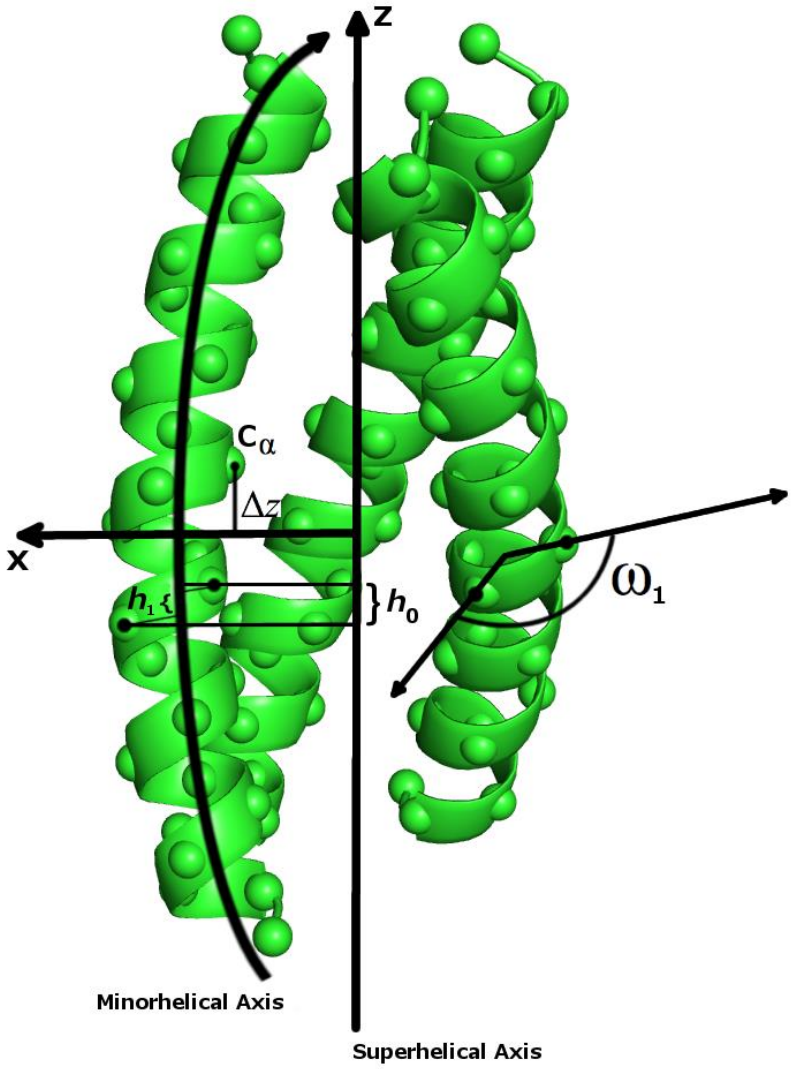

Figure 2.2: **Geometric parameters of coiled-coils.** ω_1 is the minorhelical frequency, Δz is the axial displacement, h_0 is the superhelical rise per residue, and h_1 is the minorhelical rise per residue.

The relationship between the two can be derived from the Fraser-McRae equation[81] as

$$h_0 = \frac{h_1}{\sqrt{(r R_0)^2 + 1)}} \tag{2.4}$$

where h_0 is the superhelical rise per residue, h_1 is the minorhelical rise per residue, and ρ is the radians per rise. For a left-handed α-helix, ρ can be calculated from the pitch by the relationship:

$$r = \frac{-2\rho}{P} \tag{2.5}$$

When discussing the geometry of coiled-coil peptides, it is more convenient to refer to the superhelical pitch (ϕ_0) as the positional angle with respect to a reference backbone α-carbon. The superhelical phase at each helical position can then be calculated as:

$$f'_0 = f_0 - (\frac{(L+2)h_0 r}{2}) + \Delta z r \tag{2.6}$$

where L is the number of residue positions at each helix.

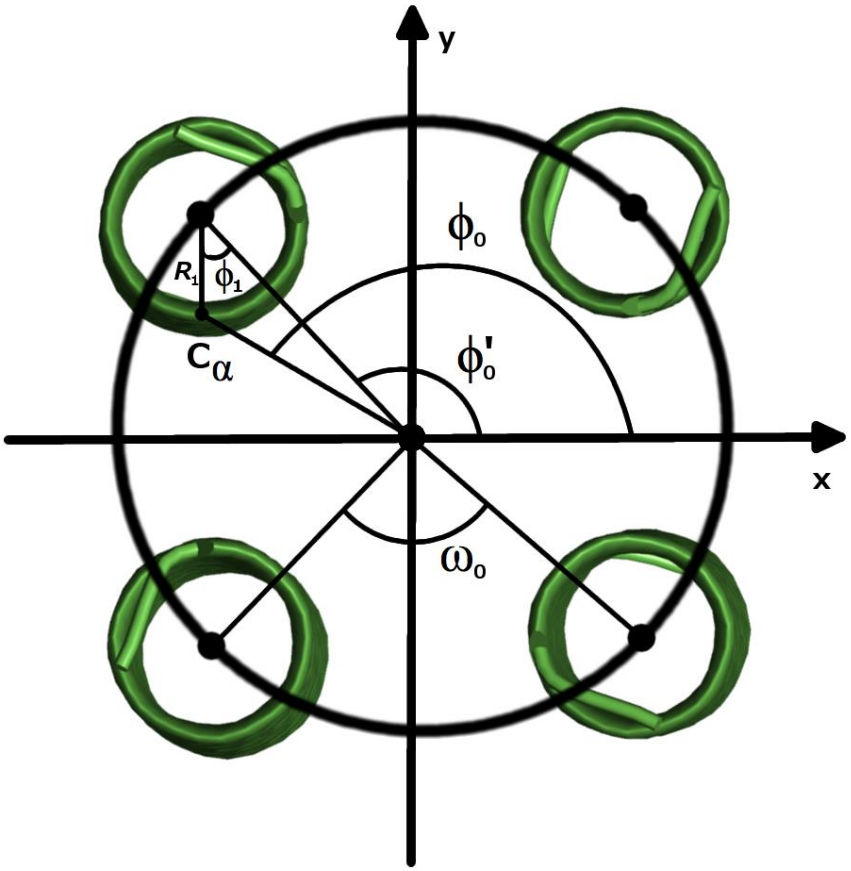

Figure 2.3: **Geometric parameters of coiled-coils.** R_1 is the minorhelical radius, ω_0 is the superhelical frequency, ϕ_0 is the superhelical phase with respect to the reference α carbone, ϕ'_0 is the minorhelical phase with respect to the minorhelical axis, and ϕ_1 is the minorhelical phase.

In their final form, the Crick equations adapted for axial offset and residue centering are:

$$x_i = R_0 \cos(\omega_0 t + \phi'_{0_i}) + R_1 \cos(\omega_0 t + \phi'_{0_i}) \cos(\omega_1 t + \phi_1) - R_1 \cos(\alpha) \sin(\omega_0 t + \phi'_{0_i}) \sin(\omega_1 t + \phi_1)$$

$$y_i = R_0 \sin(\omega_0 t + \phi'_{0_i}) + R_1 \sin(\omega_0 t + \phi'_{0_i}) \cos(\omega_1 t + \phi_1) - R_1 \cos(\alpha) \cos(\omega_0 t + \phi'_{0_i}) \sin(\omega_1 t + \phi_1) \quad (2.7)$$

$$z_i = \frac{\omega_0 R_0}{\tan(\alpha)} t - R_1 \sin(\alpha) \sin(\omega_1 t + \phi_1) + \Delta z_i r + \frac{(L+2) h_0 r}{2}$$

where *i* denotes the index of the helical position.

2.5. Fitting of Parameters to Naturally Occurring Coiled-coils

To carry out the dimensionality reduction of the atomic coordinates of a coiled-coil, one can set up an objective function that quantifies the difference in the atomic positions of the reference structure from those of the model generated by the application of the parametric equations. The root-mean-squared deviation (RMSD) between atomic positions can be calculated using the method developed by Kbasch.[82] More recently, Dill and coworkers developed a method for calculating the RMSD between two structures using quaternions.[83] This method solves the same eigenvector problem as Kabsch, but does not need to be checked for improper rotations.[84]

The objective function relates a set of Crick parameters to the RMSD between a reference structure and a model structure as follows:

$$g(\mathbf{w}) = \sqrt{\frac{\sum_{k=1}^{N} (|\mathbf{x}_k|^2 + |\mathbf{y}_k|^2 - 2\lambda_{\max}}{N}} \tag{2.8}$$

where **w** is the vector of Crick parameters to be fitted, **x** are the alpha carbon coordinates of the reference structure, **y** are the alpha carbon coordinates of the model, *N* is the number of α carbon positions, and $\lambda_{\max}$ is the maximum eigenvector of a matrix of values calculated from **x** and **y**.

To calculate $\lambda_{\max}$, a matrix correlation (*R*) is constructed from the atomic coordinates of the reference and model structures:

$$R = \sum_{k=1}^{N} x_{ik} y_{ik}, i, j = 1, 2, 3 \tag{2.9}$$

where *i* denotes the *i*th component of the vector.

From matrix R, a new matrix F is constructed as:

$$F = \begin{pmatrix} R_{11} + R_{22} + R_{33} & R_{23} - R_{32} & R_{31} - R_{13} & R_{12} - R_{21} \\ R_{23} - R_{32} & R_{11} - R_{22} - R_{33} & R_{12} + R_{21} & R_{13} + R_{31} \\ R_{31} - R_{13} & R_{12} + R_{21} & -R_{11} + R_{22} - R_{33} & R_{23} + R_{32} \\ R_{12} - R_{21} & R_{13} + R_{31} & R_{23} + R_{32} & -R_{11} - R_{22} + R_{33} \end{pmatrix} \quad (2.10)$$

λ_{max} is found by solving the eigenproblem:

$$FQ = \lambda Q \quad (2.11)$$

and λ_{max} is the largest of the eigenvalues.

In our implementation, we set up a set of C++ routines to calculate atomic coordinates of α carbons using the modified Crick equations as described above. R and F matrices are constructed from the atomic coordinates of the α carbons of the reference structure and from the model generated from the Crick equations. The MatLab API is used to interface between the C++ coiled coil routines and the MatLab routines for solving eigenproblems.

The objective function is passed to a nonlinear solver to optimize the values of the Crick parameters. The first and second derivatives with respect to each parameter are calculated using the method of finite differences:

$$\frac{\partial g}{\partial w_i} = \frac{g(\mathbf{w} + h\mathbf{w}_i) - g(\mathbf{w})}{h} \tag{2.12}$$

$$\frac{\partial^2 g}{\partial w_i \partial w_j} = \frac{\frac{\partial g(\mathbf{w} + h\mathbf{w}_j)}{\partial w_i} - \frac{\partial g(\mathbf{w})}{\partial w_i}}{h}$$

where h is an arbitrarily small value chosen to achieve downward directionality on the multi-dimensional surface.

2.6. Generation of Protein Loops

The α-helical segments of coiled-coils can be stitched together with random coil segments to form full-length single-chain constructs. Such segments can be composed of poly-glycine stretches, which are known to remain unstructured. However, the flexibility of poly-glycine loops can result in poor stability of the resulting protein.[85] To overcome this problem, loop segments can be obtained from known crystal structures.[86] A loop segment can be stitched onto two α-helices, minimizing the RMSD between stitch sites as a function of rotations and translations. This methodology requires precise matching of

the orientations of the stitch sites on the loop and on the helices. Since *de novo* protein design can explore conformations that may not have been previously observed, this strategy may result in a small number of loop candidates. Protein loops can also be constructed from smaller fragments taken from loop crystal structures.[87]

An alternative is to allow some flexibility in the conformation of the loops. Flexibility can be allowed for the backbone torsional angles at the stitch sites, or can be expanded to some or all of the sites on the loop segment. By restricting the value of the bond lengths, bond angles, and ω angles of the backbone, the dimensionality of backbone atomic coordinates can be reduced to pairs of (ϕ,ψ) angles. Values of (ϕ,ψ) angles can be extracted from structures of natural proteins and used to reconstruct foldable loop segments. To connect two helices, a poly-glycine segment is extended from one of the helices and selected (ϕ,ψ) angles are fixed to the predetermined values. The loop is then closed by manipulation of the free (ϕ,ψ) angles until the target and anchor site meet a criterion for satisfactory overlap.

A loop-closing algorithm developed by Dunbrack and co-workers is based on kinematic equations used in robotics.[3] An angle of rotation is selected and value of the rotation that will minimize the distance between the three C-terminal residue atoms on the loop (termed the anchor atoms) and the three atoms onto which the loop will be stitched (termed the target atoms). F_1, F_2, and F_3 are the position vectors of the target atoms, and M_{01}, M_{02}, and M_{03} are the initial position vectors of the anchor atoms.

A new set of anchor positions, M_1, M_2, and M_3, is calculated by minimizing the equation:

$$S = \left|\vec{O}_1 M_1 - \vec{O}_1 F_1\right|^2 - \left|\vec{O}_2 M_2 - \vec{O}_2 F_2\right|^2 - \left|\vec{O}_3 M_3 - \vec{O}_3 F_3\right|^2 \quad (2.13)$$

where O_1, O_2, and O_3 are directional vectors from the vector of rotation to each of the anchor atom positions.

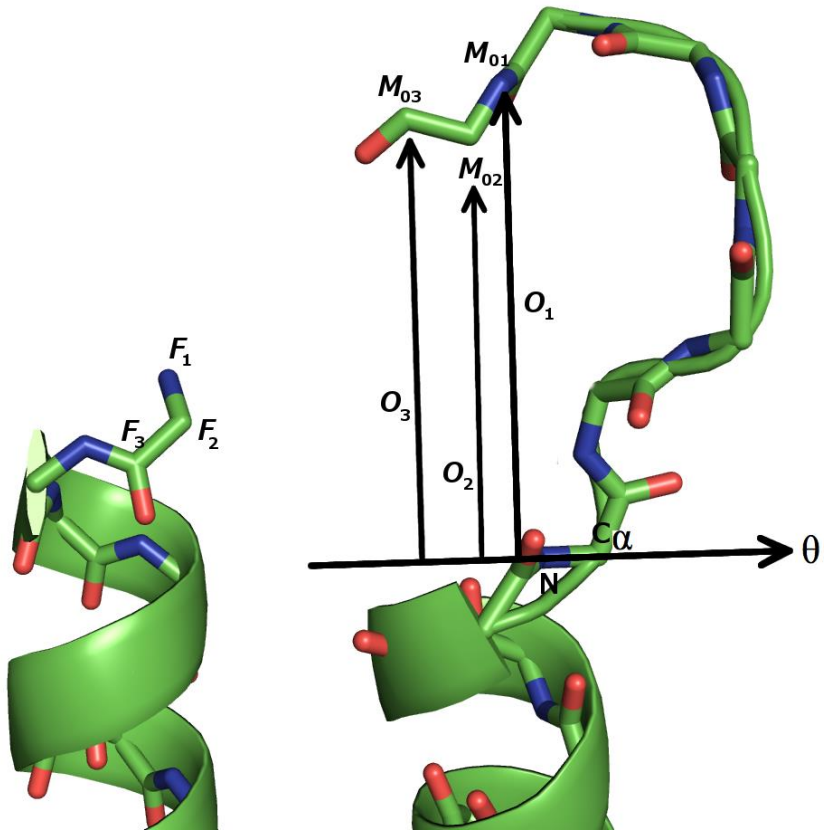

Figure 2.4: **Vector quantities in kinematic loop-closing algorithm.** F_1, F_2, and F_3 are the position vectors of the target atoms, and M_{01}, M_{02}, and M_{03} are the initial position vectors of the anchor atoms. O_1, O_2, and O_3 are directional vectors from the vector of rotation to each of the anchor atom positions, and θ is the vector along the axis of rotation. A new set of anchor positions, M_1, M_2, and M_3, is calculated that minimizes the squared distances between the anchor and target atoms.

A set of vectors is calculated as:

$$r_i = \vec{O}_i \vec{M}_{0i}$$

$$\vec{f}_i = \vec{O}_i \vec{F}_i \qquad (2.14)$$

$$\hat{s}_i = \hat{r}_i \acute{} \ \hat{q}_i$$

where $\hat{q}_i$ is the normalized vector along the axis of rotation, and $\hat{r}_i$ is the normalized vector of r_i, and i is the index of atom positions. These quantities are used to calculated the angle of rotation that will minimize the squared distances between the anchor atoms and the target atoms by the equation:

$$\tan(\alpha) = \frac{(\vec{f}_1 \cdot \hat{s}_1) r_1 + (\vec{f}_2 \cdot \hat{s}_2) r_2 + (\vec{f}_3 \cdot \hat{s}_3) r_3}{(\vec{f}_1 \cdot \hat{r}_1) r_1 + (\vec{f}_2 \cdot \hat{r}_2) r_2 + (\vec{f}_3 \cdot \hat{r}_3) r_3} \qquad (2.15)$$

This procedure yields connected protein loops where the bond lengths and angles fall within target values. However, this procedure does not guarantee that the dihedral angles will be favorable. The resulting torsional angles can be analyzed to determine if they fall

within populated regions on a Ramachandran plot. Protein loop candidates can be scored and ranked on the basis of energetic calculations carried out using atomic potentials.

2.7. Sequence Optimization

2.7.1. Calculation of Sequence Probabilities

Once a foldable target structure is chosen, the next task in to hone in on a sequence or set of sequences that are likely to fold into the desired structure. This is known as the reverse protein-folding problem. Given a set of N design positions, we can describe an ensemble Ω of 2^N possible sequences. We seek to identify a member of the ensemble that is likely to fold into the target structure, where the probability of any particular sequence is proportional to the internal energy of that sequence in the folded state. The probability of a sequence is broken down to amino acid site probabilities at each site in their respective conformations,[88]

$$W(a_1,...,a_N) = \prod_{i=1}^{N} w_i(a_i) \tag{2.16}$$

where α is the amino acid identity at site i.

This approximation is used to estimate the entropy by approximating the function:

$$S = -\sum_{protein\_states} W(protein\_states)\ln(protein\_states) \tag{2.17}$$

as:

$$S = -\sum_{i=1}^{N}\sum_{a_i=1}^{a} w_i(a_i)\ln w_i(a_i) \tag{2.18}$$

where a is the number of α amino acid degrees of freedom at site i. Conformational degrees of freedom can be accounted for, resulting in the equation:

$$S = -\sum_{i=1}^{N}\sum_{a_i=1}^{a}\sum_{r(a_i)=1}^{R} w_i(a_i, r(a_i))\ln w_i(a_i, r(a_i)) \tag{2.19}$$

where R denotes the number of rotamer states for amino acid α at site i. From here on, we shall refer to an amino acid α in a conformational state $r(\alpha)$ as a conformer.

2.7.2. Energy constraint

As in statistical mechanics, the entropy function can be maximized subject to an energetic constraint by the method of Lagrange multipliers by setting up a variational function as follows:

$$V = S - b\langle E\rangle - \sum_{i=1}^{N} (l_i(\sum_{a_i}^{q} \sum_{r(a_i)}^{R} w_i(a_i, r(a_i))) - 1) \qquad (2.20)$$

where <*E*> denotes the average energy over all conformations, and β is a quantity analogous to the inverse temperature in thermodynamics. The sum of the conformer probabilities at each site is constrained to a value 1, with each site having a corresponding Langrange multiplier λ_i. The set of conformer probabilities (**w**) and the set of Lagrange multipliers (β, λ_i) are solved for by identifying a set of values that results in:

$$0 = \vec{\nabla} V(\mathbf{w}, \beta, \lambda_i) \qquad (2.21)$$

Alternatively, the problem can be recast as an optimization problem as follows:

$$\begin{aligned} &\max S(\mathbf{w}) \\ &\text{subject to } 0 \le w_i \le 1, \\ &\sum_{a_i=1}^{g} \sum^{R} w_i(a_i, r(a_i)) = 1, \\ &\text{and } U(\mathbf{w}) = \langle E \rangle \end{aligned} \tag{2.22}$$

where the conformer probabilities are solved for using nonlinear optimization techniques.

In practice, an effective temperature β is targeted to modulate the contribution of the energy function to conformer probabilities. In this case, the objective function is optimized as:

$$\begin{aligned} &\min F(\mathbf{w}, \beta) = U(\mathbf{w}) - \frac{S(\mathbf{w})}{\beta} \\ &\text{subject to } 0 \le w_i \le 1, \\ &\text{and } \sum_{a_i=1}^{g} \sum^{R} w_i(a_i, r(a_i)) = 1 \end{aligned} \tag{2.23}$$

The result is a sequence profile generated with conformer probabilities consistent with the target values of the constraint functions.

In reality, this formulation need not be restricted to individual site probabilities. Probabilities can be calculated for any arbitrary decomposition of polymer subunits. For example, a set of covalent crosslinks can be optimized by grouping two distant sites into one conformer.

2.7.3. Energy Functions

All-atom force fields consist of atomic pairwise terms such as electrostatics and van der Waals energies, three-atom terms such as angle energies, four-atom terms such as dihedral energies, and many-atom terms such as hydrogen bonding energies. For a particular set of protein conformational degrees of freedom, these energetic terms can be grouped together into inter-conformer (one-body) energies and intra-conformer (two-body) energies. From the set of one-body and two-body energy terms, the energy of any member of the ensemble can be calculated as:

$$E(a_1, r(a_1); a_2, r(a_2), \ldots, a_N, r(a_N)) = \sum_{i=1}^{N} e_i(a_i, r(a_i)) + \sum_{i=1}^{N} \sum_{j>i}^{N} e_{ij}(a_i, r(a_i); a_j, r(a_j)) \quad (2.24)$$

where ε_i is the energy calculated from atomic positions at site i, and ε_{ij} is the energy calculated from atomic positions across conformers at sites i and j.

From a set of individual conformer probabilities, the average energy over all sequences is calculated as:

$$\langle E \rangle = \sum_{i=1}^{N} \sum_{a_i}^{q} \sum_{r(a_i)}^{R} e_i(a_i, r(a_i)) w_i(a_i, r(a_i)) + \sum_{i=1}^{N} \sum_{a_i}^{q} \sum_{r(a_i)}^{R} \sum_{j>i}^{N} \sum_{a_j}^{q} \sum_{r(a_j)}^{R} e_{ij}(a_i, r(a_i); a_j, r(a_j)) w_i(a_i, r(a_i)) w_j(a_j, r(a_j)) \quad (2.25)$$

2.7.4 Energy Constraint in the Context of Protein-protein Interactions

The entropy maximization formulation also need not be restricted to single polymer chains. Residue sites can be distributed across two or more individual backbones, and two-body energy terms across chains will account for the energetics of protein-protein interactions.

The energy function can also account for the presence of an external field. For example, when designing a protein to bind to a crystal structure model, the conformer identities at each site on the target structure are in a fixed conformation. Probabilities for the conformers on the target structure need not be calculated, so that the energy of interaction across the interface can be considered as an external field acting on the ensemble of sequences.

The energy of any one member of the ensemble can be calculated as:

$$E(a_1, r(a_1); a_2, r(a_2), ..., a_N, r(a_N)) = \sum_{i=1}^{N} e_i(a_i, r(a_i)) + \sum_{i=1}^{N} \sum_{j>i}^{N} e_{ij}(a_i, r(a_i); a_j, r(a_j)) + \sum_{i=1}^{N} \sum_{j=1}^{M} e_{ij}(a_i, r(a_i); z_j) \quad (2.26)$$

where ζ_j is the two-body interaction between a conformer and a site on a binding partner with *M* sites.

Proteins are able to form assemblies of symmetric arrangements due to the complementarity inherent in copies of identical subunits. An ensemble of sequences interacting with symmetry related elements will experience an external field energy that is dependent on the conformer probabilities of the asymmetric unit. The average energy of the asymmetric unit over all sequences becomes:

$$\begin{aligned}
\langle E \rangle = & \sum_{i=1}^{N} \sum_{a_i}^{q} \sum_{r(a_i)}^{R} e_i(a_i, r(a_i)) w_i(a_i, r(a_i)) \quad (2.27) \\
& + \sum_{i=1}^{N} \sum_{a_i}^{q} \sum_{r(a_i)}^{R} \sum_{j>i}^{N} \sum_{a_j}^{q} \sum_{r(a_j)}^{R} e_{ij}(a_i, r(a_i); a_j, r(a_j)) w_i(a_i, r(a_i)) w_j(a_j, r(a_j)) \\
& + \frac{1}{2} \sum_{i=1}^{N} \sum_{a_i}^{q} \sum_{r(a_i)}^{R} \sum_{m}^{M} e_{ij}(a_i, r(a_i); z_i^m, r(z_i^m)) w_i(a_i, r(a_i)) \\
& + \frac{1}{2} \sum_{i=1}^{N} \sum_{a_i}^{q} \sum_{r(a_i)}^{R} \sum_{m}^{M} \sum_{j^1 i}^{N} \sum_{z_j}^{q} \sum_{r(z_j)}^{R} e_{ij}(a_i, r(a_i); z_j^m, r(z_j^m)) w_i(a_i, r(a_i)) w_j(a_j, r(a_j))
\end{aligned}$$

where $r(\zeta_i^m)$ is a symmetry related element of $r(\zeta_i)$ on chain *M*. The $\frac{1}{2}$ coefficients are applied so that the energy of a symmetric assembly scales with the number of asymmetric subunits as:

$$\left\langle E_{assembly} \right\rangle = M \times \left\langle E_{assymetric\ unit} \right\rangle \tag{2.28}$$

where equivalent interactions are not double counted.

2.7.5. Environmental Energy Constraint

In folded proteins, the propensity of an amino acid to be found in a solvent exposed region is proportional to the extent of its hydrophobic character. During the course of potential energy calculations in computational protein design, the protein surface is not yet defined, making the use of solvent accessible surface area (SASA) solvation models difficult to implement. Instead, we use a statistically derived amino acid propensity based on β-carbon density to recover sequence-structure relationships in naturally occurring proteins. The potential is defined as:

$$e_e(a,r) = -T_e \ln(f(a,r)/f(r)) \tag{2.29}$$

where $-T_e$ is an effective temperature, and ρ the β-carbon density at the residue position. ρ is calculated as:

$$r(a) = \frac{n_b}{V_{sphere} - \langle V_{access}(a) \rangle} = \frac{n_b}{\frac{4}{3}\pi R_c^3 - \langle V_{access}(a) \rangle} \tag{2.30}$$

where n_β is the number of β-carbons within an 8 Å distance (R_c) from the side-chain center of mass, and $<V_{access}(\alpha)>$ is the amino acid accessible volume averaged over a set of specified rotamer conformations. The value of the potential used in the sequence calculations was determined by collecting statistics of β-carbon density for all 20 amino acids in set of 500 non-redundant globular proteins.

3| Computational Design of a Lanthanide-Binding Protein

3.1. Abstract

Lanthanide-protein complexes are of interest for their potential as imaging agents in biological applications, and as purification agents in separation technologies. Chemical screening and rational approaches have been used in the design of lanthanide-binding peptides, but the ability to systematically tune the binding-site geometry in these constructs is limited. A kinetically stable protein with a lanthanide-binding pocket situated in a region with a high density of neighboring sites would allow for the fine tuning of first-shell and second-shell coordination spheres. Four-helix bundles have long been used as model systems in which to engineer tunable metal-binding sites. We computationally designed a protein with a four-helix bundle motif that houses a lanthanide-binding site in the hydrophobic core. The protein was expressed and purified, and was shown to be highly thermodynamically and kinetically stable. Fluorescence resonance energy transfer (FRET) between a tryptophan residue and the bound lanthanide ion was used to study the thermodynamics and kinetics of binding. The non-symmetric nature of the binding pocket will make it possible to use this protein as a model system for systematically making changes to the coordination sphere on a single residue basis. The slow kinetics and the inaccessibility of the binding site will enable the construction of binding pathways to modulate the kinetic properties of the system.

3.2. Introduction

The clean energy sector is heavily dependent on the supply of rare earth elements (REEs) for their electromagnetic and optical properties.[89] For example, wind turbines make use of high-powered dysprosium and neodymium magnets, and the reduction and substitution of these elements is still an active area of investigation.[90] The increasing shift from fossil fuels to renewable energy sources is projected to give rise to an exponential rise in the demand for REEs.[91] Although a spike in the exploration of new sources of REEs has identified mineral deposits across the globe, the near monopoly held by The People's Republic of China on the REEs market is likely to make the utilization of these deposits economically challenging in the near future.[92] The absence of a geographically diversified source of these critical elements is of strategic concern to the high-tech and defense sectors.[93] Efficient and environmentally benign methods for extraction and purification of REEs would help to alleviate these supply concerns by facilitating the exploitation of domestic sources.

The 4f orbitals of elements in the lanthanide series are shielded by the larger filled 5d and 5p orbitals, and as a consequence these elements bind ligands through mostly electrostatic interactions. As a result, their separation can be difficult to achieve through chemical means. The industrial purification of rare earth elements (REEs) is most commonly carried out by solvent extraction, a process that is unsustainable due to its detrimental environmental impact.[94] The extractants in use are highly acidic in order to enable the exchange of hydrogen ions with metal ions. The resulting solution is

commonly neutralized by the addition of ammonia, which results is large amounts of contaminated wastewater.[95]

Protein-lanthanide systems are of interest for their potential to serve as purification agents for use under biological conditions.[94] Genetically encodable proteins could potentially be incorporated into lanthanide bioremediation technologies.[96-99] Additionally, protein-lanthanide systems are used in the study of biological systems since the spectroscopic properties of lanthanides make them useful as luminescent tags.[100] Peptide sequences with non-natural amino acids have been designed to bind to lanthanides,[101] but such proteins lack the ability to be genetically encodable. A short terbium-binding sequence based on only natural amino acids has also been achieved using an evolutionary approach.[102] This peptide sequence, termed the lanthanide binding tag (LBT), showed nanomolar affinity to lanthanides along with a trend in selectivity as a function of effective ionic radius. The strong affinity of LBT for lanthanide ions has made it of particular interest as a purification agent. The effects of sequence modifications to lanthanide binding affinity have been studied,[103] and the sequence has been displayed on the surfaces of bacterial cells for the sequestration of lanthanide ions.[104] Separation experiments with cell surface displayed LBT showed that while ion binding is favored for lanthanides ions of smaller radii over larger ions, a binding preference was not detected in equimolar mixtures Tb^{3+}/Dy^{3+} or Tb^{3+}/Nd^{3+}.

There is only one known case of a naturally occurring lanthanide-biding protein, a methanol dehydrogenase enzyme found in an extremophile bacterium.[105] Lanthanide ions are known bind to Ca^{2+} binding sites in other naturally occurring proteins,[106] and the

luminescent properties of terbium have been used to investigate such sites.[107] A trend in binding constants for these systems is generally observed as a function of metal effective ionic radius.[108] Peacock and coworkers designed a lanthanide-binding trimeric coiled-coil capable of encapsulating gadolinium for use as an NMR contrast agent.[109] The flexible nature of the binding site made for rapid equilibration of lanthanide binding. This system exhibited micromolar affinities for terbium and gadolinium, but no discernable thermodynamic selectivity was observed between the two. The selectivity toward other lanthanide ions was not investigated.

The protein-lanthanide systems designed to date have been based on flexible and accessible binding sites. On the other hand, the effects of protein structure rigidity on the selectivity and the kinetics of lanthanide binding have not been previously investigated. *De novo* proteins may serve as starting points for understanding and allowing for the enrichment of one lanthanide over another. The field of *de novo* protein design has seen considerable success in the design of metal complexes.[30, 110] The periodic structure of the coiled-coil motif, as well as an understanding of its sequence-structure relationships, has made it a work-horse of *de novo* protein design.[111] In a typical design, a binding pocket is engineered at the core of the protein scaffold with likely side-chain conformations that are consistent with a given target binding-site geometry. The remainder of the sequence is then chosen as to achieve the folded target structure, as well as drive the binding site amino acids to their target conformational states and away from alternate configurations. If the folding of the protein occurs cooperatively with the binding of the ligand, the kinetics of binding is rate-limited by the rate of the folding process. If protein folding is

independent from substrate binding, then the kinetics of binding will be rate-limited by the accessibility of the binding site. The ability to tune the kinetics of binding in *de novo* designed proteins could enable the exploration of kinetic selectivity as a lanthanide separation strategy.

To explore the possibility of creating protein structures capable of exhibiting selectivity over the lanthanides, we set out to design a lanthanide-binding pocket at the core of a rigid protein scaffold. We envisioned a hyper-stable protein capable of differentiating between the lanthanides on the basis of differences in atomic radii.[112] We developed an algorithm to search for high-density metal-coordination spheres, and constructed a 6-coordinate metal-binding site within a single-chain 4-helix bundle protein. The designed protein was observed to be structured in the absence of the metal, and temperature-melt experiments showed a lack of a melting point transition up to 98° C. Fluorescence resonance energy transfer (FRET) was used to assess the binding of terbium to protein. Consistent with the design, it was observed that heating of the protein was necessary to allow the metal accessibility to the buried binding site. We performed equilibrium titrations to obtain binding constants for elements across the lanthanide series, and observed an overall trend with respect to effective ionic radius that is consistent with other protein-lanthanide systems studied previously.[102, 108] Half-lives of spontaneous dissociation from the binding site were observed to be on the order of hours for various lanthanides, with a trend also observed with respect to effective ionic radius.

Since it is difficult to achieve thermodynamic selectivity of lanthanide binding, kinetic differentiation among the lanthanides provides a possible alternative for carrying out

separations. Recently, kinetic control of lanthanide separations has come under attention.[113] Here we provide proof-of-principle that the dynamical features of *de novo* designed protein could be used to kinetically differentiate between ligand sets based on small differences in physical properties.

3.3. Summary of Protein Design Methodology

Lanthanide coordination chemistry is of a different nature than the more familiar chemistry of the d-block elements. Due to their shielded f orbitals, the coordination geometries of lanthanide complexes are not restricted to the symmetry of their valence orbitals. Therefore, the design of lanthanide ligands is governed almost strictly by the electrostatic and steric interactions between the ligands.[114] The design of lanthanide complexes with high coordination numbers poses a challenge since ligands must be packed into a small volume.[115] The *in situ* generation of coordination spheres without steric overlaps requires searching through a large number of candidates.

In order to efficiently generate a large number of candidate coordination spheres, we developed a methodology to identify structures that support side chains that could form part of a lanthanide-binding site independently of other side chains in the coordination sphere. By restricting the position of the metal ion to a line, in this case the superhelical axis of the bundle (*z*-axis), side chain conformations could be independently evaluated according to their potential to coordinate to a metal ion. A library of conformations with potential points of coordination on the *z*-axis was created. This was termed the super-rotamer library, where a super-rotamer is a side chain ligand along with a metal atom

positioned to bind to the side chain. Full coordination spheres are then constructed from combinations of members of the super-rotamer library.

The crystal structure of the GCN4-pV peptide[116] was chosen as a fiducial starting point template for the design of coiled-coils capable of housing a lanthanide-binding site. The high stability of this construct, as evidenced temperature-melt experiments, led us to hypothesize that the backbone structure would compatible with a large number of sequences and would result in a highly stable protein. The anti-parallel orientations of the helices would also enable the construction of a single-chain helical-bundle by connecting the helical segments with short loop segments. The search for high-density well-packed coordination spheres necessitated considering alternate relative orientations of the helices. The dimensionality of the GCN4-pV tetramer was first reduced to a set of Crick parameters. Varying the values for the minor helical phase and then reconstructing the helices from the Crick equations generated models with alternate coiled-coiled geometries.

The set of coiled-coil candidate structures was further expanded using a combinatorial approach, where structures were generated from combinations of individual helices constructed from the Crick equations. A super-rotamer library was created at each helix, and coordination spheres were generated from combinations of super-rotamers that could coordinate a metal ion at the same point on the *z*-axis. Steric and electrostatic interactions of the candidate coordination spheres were analyzed using atomic potentials.[117]

3.4. Super-rotamer Library Construction

A super-rotamer consists of an amino-acid side chain bound to a metal ion. By constraining the position of the metal to be on the z-axis, a maximum of two super-rotamers can be generated from any one rotamer state. Given a certain metal-to-ligand distance, the position of the metal relative to the ligand must satisfy the equation

$$(x - x_c)^2 + (y - y_c)^2 + (z - z_c)^2 = R^2 \quad (3.1)$$

where (x, y, z) is the position of the metal, (x_c, y_c, z_c) is the position of the coordinating atom, and R is the ligand-to-metal distance. The cone angle is defined as the angle between the metal ion, the coordinating atom, and the antecedent atom. Super-rotamers with metal positions that do not meet a cone angle criterion can be removed from the set.

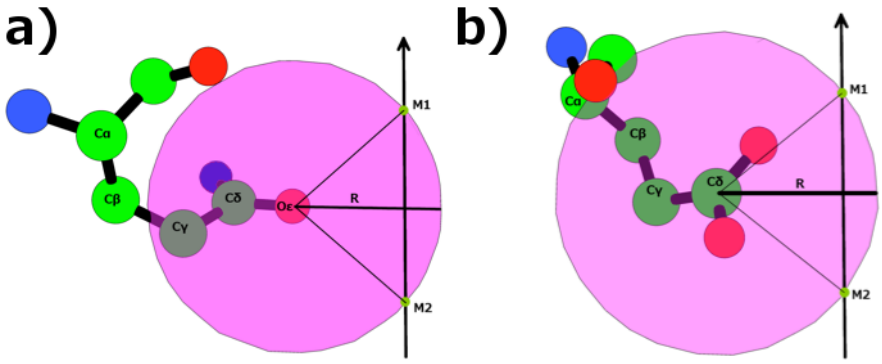

Figure 3.1: **Determination of super-rotamer metal positions.** Possible metal coordination sites for monodentate and bidentate super-rotamers. a) Glutamine side chain with monodentate coordination. b) Glutamate side chain with bidentate coordination.

Amino acid bond lengths and bond angles were generated using average values found in a set of 54 high-quality crystal structures.[118] Target ligand-to-metal distances were taken from a small molecule complex.[119]

Amino Acid Type	Binding Mode	Coordinating Atom	Antecedent Atom	Target Metal to Ligand Distance	Maximum Cone Angle	Off-rotamer Conformations
GLU	Bidentate	Cδ	Cγ	2.847 Å	20°	±20° for χ_1 in 10° increments
ASP	Bidentate	Cγ	Cβ	2.847 Å	20°	±20° for χ_1 and χ_2 in 10° increments
GLN	Monodentate	Oε	Cδ	2.359 Å	60°	None
ASN	Monodentate	Oδ	Cγ	2.359 Å	60°	None

Table 3.1: **Coordination geometry of super-rotamers.** Geometric parameters for the search of amino acids with the potential to bind a metal ion on the *z*-axis.

3.5. Calculation of Initial-value Crick Parameters

Atomic coordinates for GCN4-pV monomer were obtained from Protein Data Bank crystal structure 2B22. Coordinates for the tetrameric coiled-coil were generated using crystal symmetry transformations in PyMOL.[120] The structure was fitted to geometric parameters using the coiled-coil Crick equations by minimizing the RMSD between the reference and model structures using the IPOPT optimizer,[121] as illustrated in sections 2.4 and 2.5. The coiled-coil squareness was constrained at 0°, and the rise per residue was constrained to be the canonical value of 1.5 Å. Minimization was carried out for 3,000 steps, and the final point was used as the optimal parameter set. The RMSD between the final model and the reference structure was 0.34 Å.

Parameter	Initial Value	Lower Bound	Upper Bound	Final Value
R_0	5.0 Å	0 Å	10.0 Å	7.28 Å
ϕ_{AC}	90°	-180°	180°	35.06°
ϕ_{BD}	90°	-180°	180°	-42.70°
Z_{off}	0 Å	0 Å	10.0 Å	2.53 Å
α	90°	180°	180°	149.65°

Table 3.2: **Fitting of Crick parameters to the structure of GCN4-pV.** Initial values of optimized parameters, bounds on problem, and final parameter values.

3.6. Search Terbium-binding Motifs in Coiled-coils

For the set of coiled-coil parameters, α-carbon coordinates were calculated from the Crick equations. Coiled-coils with alternate orientations of the helices were generated by varying the minorhelical phase (ϕ_1) for each helix from -50° to +50° from the reference

structure at 5° intervals. Coiled-coils with alternate superhelical radii were considered by varying R_0 from -1 Å to + 1 Å from the reference structure at 0.1 Å intervals. The three other main-chain atoms were constructed from the α-carbon using CHARMM22 equilibrium bond lengths and angles. The 2010 Dunbrack rotamer library was used to calculate coordinates of side chain conformational states.[55] Coordinates of off-rotamer conformations are also calculated. The resulting set of side chain conformational states is used to construct the super-rotamer library.

Parameter	Minimum Value	Maximum Value	Grid Search Interval Value	Number of Intervals in Grid Search
R_0	6.28 Å	8.28 Å	0.1 Å	21
ϕ_A	-14.94°	85.06°	5.0°	21
ϕ_B	-92.7°	7.3°	5.0°	21
ϕ_C	-14.94°	85.06°	5.0°	21
ϕ_D	-92.7°	7.3°	5.0°	21
Metal z coordinate	-5.0 Å	4.99 Å	0.01 Å	1000
Coiled-coil displacement in x direction	0.0 Å	0.5 Å	0.1 Å	6
Coiled-coil displacement in y direction	0.0 Å	0.5 Å	0.1 Å	6

Table 3.3: **Coiled-coil search space.** Structural degrees of freedom for terbium-binding coiled-coil search.

A combinatorial approach is used to construct the set of possible coiled-coils from combinations of helices at each of the α-helical positions, labeled as sections A, B, C, and

D. To consider metal-positions that do not lie exactly on the superhelical axis, the coiled-coil helices are shifted so that the z-axis is positioned off-center. Twenty-one unique D_2 symmetric coiled-coil structures were constructed by varying all helical rotations simultaneously. From these structures, coordinates for each of the four distinct peptide helices were identified and recombined to create a full set of asymmetric structures. A search over all combinations of helical rotations yields 21^4, or 194,481 independent coiled coil tetramers. By scanning over values of R_0, this procedure was used to search 4,084,101 unique coiled-coil structures for metal binding at 36,000 positions (search over x and y offsets, and z positions), resulting in 147,027,636,000 candidate configurations.

For a given value of R_0 and off-center displacements, the *z*-axis from -5 Å to 5 Å is scanned at 0.01 Å intervals for candidate coordination spheres. At a given point z, the subset of super-rotamers that satisfies $|z - z_{rot}| \leq 0.05$ Å is selected, where z_{rot} is the metal position on the super-rotamer. For bidentate ligands, a rotation of the terminal χ angle is performed so that both oxygen atoms are equidistant from the metal ion. If the rotation is larger that 20°,[78, 122, 123] the super-rotamer is not used in the construction coordination spheres.

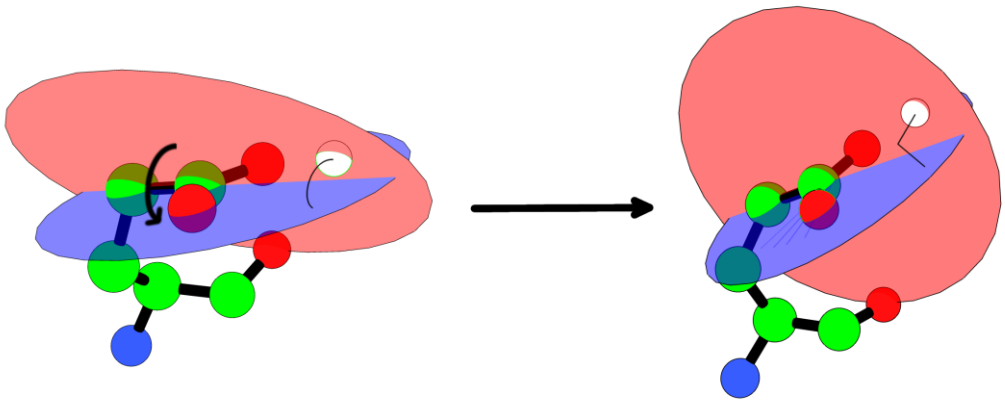

Figure 3.2: **Bidentate super-rotamer selection.** Adjustment of terminal dihedral angle for a bidentate coordinating side chain at a given metal position. Carboxylic atoms O-C-O form the blue plane, and the metal-$C_{coordinating}$-$C_{antecedent}$ form the red plane. The side chain terminal χ angle is adjusted so the the planes are perpendicular.

The ensemble (Ω_z) of all possible hexameric coiled-coils that can be created from the subset of super-rotamers is enumerated, and the energy of each member of the subset $\Omega_z(m = 3, b = 3)$ is calculated, where m is the number of monodentate ligands and b is the number of bidentate ligands. The lowest energy member of the subset $\Omega_z(m = 3, b = 3)$ is used in the construction of a structure energy landscape as a function of R_0 and (x, y, z) metal-position coordinates.

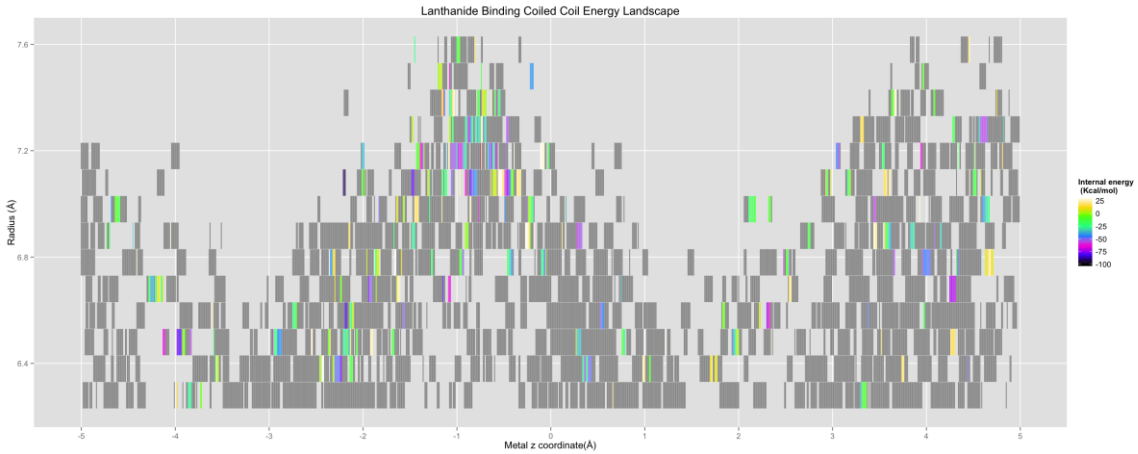

Figure 3.3: **Structure-energy landscape of terbium-binding coiled-coils.** Two-dimensional slice at x displacement = 0 and y displacement = 0.

The dihedral and Lennard-Jones terms of the AMBER84[124] force field along with a modified hydrogen bonding term are used to calculate potential energies.[125, 126] The terbium ion is not included in the energy calculations. Energy values are decomposed into one-body and two-body pairwise terms. All one-body energy terms are stored in a single vector, and all pairwise energy terms in the 194,481 combinations of coiled-coils are stored in a single matrix. Pairwise energy terms between two residues that do exist in the same coiled-coil are not calculated.

Label	Radius (Å)	x displacement (Å)	y displacement (Å)	metal z position (Å)	E_{min} (kcal/mol)
C1	6.58	0.2	0.3	1.76	-101.571
C2	6.58	0.2	0.3	1.75	-101.458
C3	6.68	0	0.3	3.74	-100.828
C4	6.68	0	0.3	3.73	-100.632
C5	6.68	0	0.3	3.72	-100.434
C6	6.68	0	0.3	3.71	-100.235
C7	6.58	0.2	0.3	1.74	-100.103
C8	6.68	0.5	0.3	4.07	-99.4731
C9	6.98	0.5	0.3	3.14	-96.901
C10	6.68	0.2	0.3	-0.47	-96.1629
C11	6.68	0.2	0.1	-0.66	-96.047
C12	6.98	0.1	0.3	-1.02	-95.9167

Table 3.4: **Results of coordination sphere search.** Lowest-energy terbium-binding coiled-coils identified in search.

3.7. Initial Sequence Optimization of Candidate Structures

Sequence optimization calculations were carried out to identify candidates that would be likely to fold into the target structure and adopt target binding site conformations. Candidate structures were subjected to rounds computational protein design. These calculations take in as input a folded structure and perform an energy sequence optimization calculation. The Lennard-Jones, electrostatic, and dihedral terms of the AMBER84 potential with a modified hydrogen bonding term were used for energy calculations. The 2002 Dunbrack rotamer library was used to build side chain conformations, with a maximum of 10 rotamers per amino acid. A pre-processing step is performed to remove any high-energy conformers.

Environmental potential values were calculated for exposed and buried sites of the GCN4-pV tetramer. These values were used to set the environmental potential targets of the design calculations. After each calculation, if any site had an amino acid type whose probability was twice as much as the next most probable type, the amino acid degrees of freedom were fixed to that type only. The iterative calculation method was repeated until no further sites displayed a high preference of one amino acid over the others.

Parameter	**Setting**
Rotamer library	Dunbrack 2002
Maximum number of rotamers	10
Force field version	AMBER84
Force field paramers	Electrostatic, Van der Waals, dihedral, hydrogen bond
β	0.5 mol/kcal
Buried beta carbon count	≤ 10
Buried environmental potential target	-7.47 kcal/mol
Exposed beta carbon count	≥ 11
Exposed environmental potential target	-4.0 kcal/mol
Amino acids allowed	Ala, Arg, Asn, Asp, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Ser, Thr, Trp, Tyr, Val

Table 3.5: **Parameters for sequence calculations.** Sequence optimization was carried out by keeping binding-site residues fixed and populating the remaining sites with amino acid rotamers.

Out of the candidates in Table 4, full sequence design calculations were carried out the lowest energy candidate at each of the local minima. The value of the environmental potential was out of range for the structure C1 due to its short radius, and therefore the structures in this local minimum were disregarded. The next structure at a different local minimum was C3, but the final sequence contained a total of 11 glycine residues. Glycine

residues are known to disrupt alpha-helix formation, therefore the structures at this local minimum were not taken into further consideration. We chose to limit the search to structures with a value of R_0 > 6.68 Å in order to avoid sequences with a large number of glycine residues, since the calculations are likely to select glycine as highly probable when sites are in close contact. The most probable sequence for C9 resulted in a construct with 8 glycine residues, while sequence optimization of C12 resulted in a most probable sequence with no glycine residues. This structure was labeled as CC-Tb and selected as the coiled-coil motif for a single-chain helical bundle.

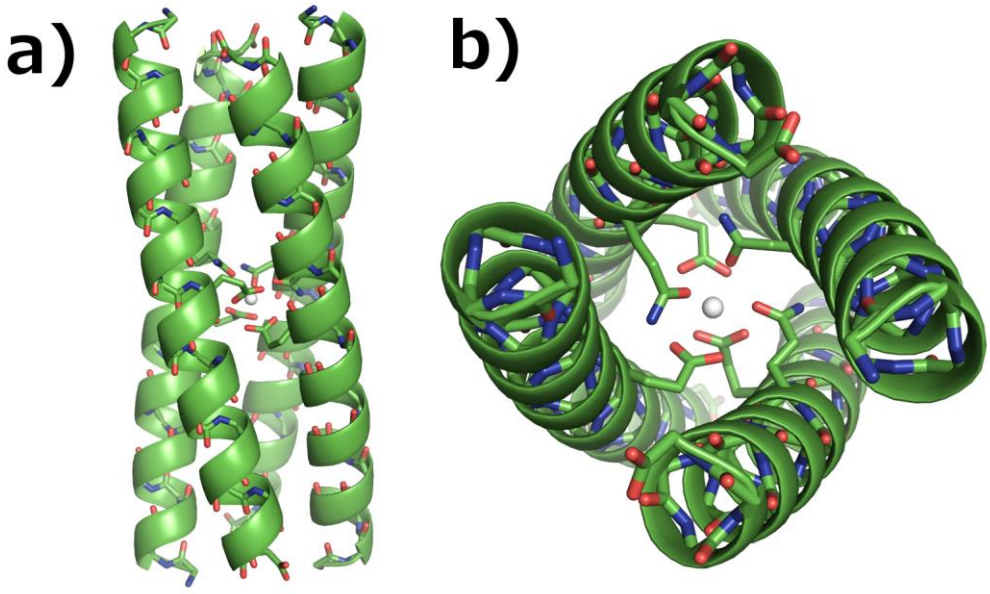

Figure 3.4: **Structure of CC-Tb.** a) Side view. b) Top view.

3.8. Loop Modeling

Based on the results from the previous calculation, the CC-Tb structure was selected for further design. Loop modeling calculations were carried out to connect the four chains of the tetramer so that the protein could be eventually expressed in *E. coli* as a single construct. The ArchDB database[127, 128] of protein loops was used to find sets of backbone torsional angles that have been observed experimentally. A cyclic coordinate descent algorithm was used to attempt to close loops between the chains.[3] Briefly, the procedure for connecting two chains via a loop is the following:

1) The C terminus of one chain is assigned as the anchor site, and the N terminus of another is assigned as the target site.

2) From the anchor site, a glycine chain of desired length is constructed and an additional site is added to the end. The algorithm aims to place this last residue in the same position as the target site by adjusting only backbone torsional angles.

3) The Arch database is accessed for experimentally observed loops of equal length. Only the subset of helix-helix loops is used.

4) Non-terminal loop angles are adjusted to match those of the selected Arch database entry.

5) One of the remaining free torsional angles is chosen at random. (Anchor site ψ, first loop site ϕ and ψ, last loop site ϕ and ψ, and target site ϕ)

6) Cyclic coordinate descent equations are solved. If the RMSD between the anchor site and the phantom anchor site is less than 0.1 Å, then the loop is considered closed.

7) Each successfully closed loop is scored using the AMBER84 potential with a modified hydrogen bonding term.

8) Sequence optimization is performed on the loop sites. All amino acids types are allowed and side chain torsional angles are taken from the 2002 Dubrack rotamer library. A maximum of 10 rotamer conformations are allowed per amino acid.

9) Average potential energy of ensemble of sequences is used to rank-order the loops.

Chain D was connected to chain A, and this segment was name chain X. Chain A was connected chain B, and chain C to chain D. These two segments, named chain Y and chain Z respectively, were calculated simultaneously so that both loops were taken from the same template structure. 2,451 helix-helix loops of length seven[129] were used for each of the connections. Loops were rank-ordered on the basis of average potential energy over all sequences. Comparison of backbone torsional angles with known Ramachandran distributions was carried out in MolProbity.[130]

Template PDB ID	Template Chain	First Residue ID	Average Energy (kcal/mol)	Ramachandran Allowed[a]
1XJA	C	141	-348.8810334	No
1YJ4	A	100	-347.6050033	No
4C9N	A	220	-347.0883921	No
2OEZ	A	31	-345.9589487	No
3GJX	A	573	-345.2512642	Yes

Table 3.6: **Loop selection for segment X.** Top 5 for loops of length 7.[a]As determined by MolProbity analysis.

Template PDB ID	Template Chain	First Residue ID	Average Energy (kcal/mol)	Ramachandran Allowed[a]
3KD3	A	23	-363.2446757	No
2CE7	C	384	-350.8490499	Yes
2IFC	C	218	-349.9256911	No
3HL0	A	179	-348.6219712	No
1HJR	A	111	-347.9766621	No

Table 3.7: **Loop selection for segments Y and Z.** Top 5 results for loops of length 7. [a]As determined by MolProbity analysis.

Based on the results of the Ramachandran analysis, template 3GJX was selected for loop X and template 2CE7 was selected for loops Y and Z.

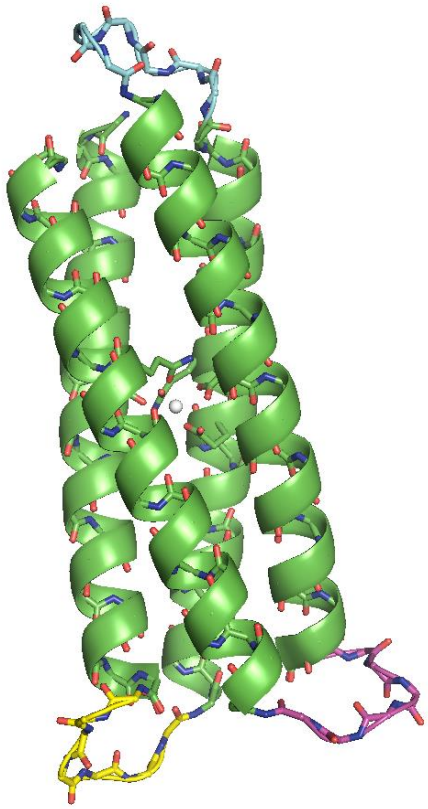

Figure 3.5: **Selected loops for scCC-Tb.** Loop X (cyan), loop Y (magenta), and loop Z (yellow).

3.9. Final Sequence Selection

Full sequence design calculations were carried out for the single chain scaffold in the same manner as indicated above for the tetrameric coiled-coil. Six rounds of iterative design calculations were carried out. The resulting probability profile was used to guide the selection of the sequence. Polar residues were avoided in the vicinity of the binding site in order to create a hydrophobic sphere around the metal binding site. This was done to prevent water accessibility to the binding pocket, as well as to prevent unwanted side

chain to metal interactions such as histidine-lanthanide bonding.[131, 132] The most probable amino acid was selected at each site, except in the following cases:

Site 26: Phenylalanine (second most probable) instead of tryptophan at site 26 to allow for only one tryptophan residue in the sequence.

Site 127: Tryptophan (second most probable) instead of histidine to be used to detect binding via Forster resonance energy transfer to terbium ion.

Site 48: Leucine (third most probable) instead of histidine to avoid histidine residue near metal-binding site. Phenylalanine (second most probable) not chosen to avoid bulky amino acid at helix-helix interface.

Site 55: Asparagine (second most probable) instead of histidine at site 55 to avoid histidine residue near metal-binding site. Polar group was allowed at this site given that glutamine (second most probable) is also polar, and phenylalanine (third most probable) is bulky.

Site 122: Leucine (third most probable) instead of histidine to avoid histidine residue near metal-binding site. Alanine (second most probable) was not chosen to avoid a cavity to the binding pocket.

Sites 31 and 103: Glycine instead of lysine to allow flexibility at α-left position on loops.

Site 73: Glycine instead of threonine to allow flexibility at α-left position on loop.

Site 14: Isoleucine instead glutamate to remove carboxylate group near metal-binding site.

The final sequence selected was:

DDDARKIIDKARNINKKALNIIQDAFKILIGSPKPSIKDVDKIIRQIL
QQESKQNDIFKKIRQEIGKVPKPGGDEARKLVEQVEKIEKSVIQLI
KQVLNHVSGTPKPNTDEVAQLLNQIINLEKQQWQLLTKIYQHM

and was named scCC-Tb. A side chain modeling calculation was carried out with the final sequence. The most probable conformation was used as the model for subsequent calculations.

3.10. Analysis of Coordination Sphere

The coordination sphere of scCC-Tb was compared to that of LBT. Analysis began by aligning the two bidentate side chains in LBT with structurally similar side chains in scCC-Tb, and proceeded by adding one side chain at a time to the alignment.

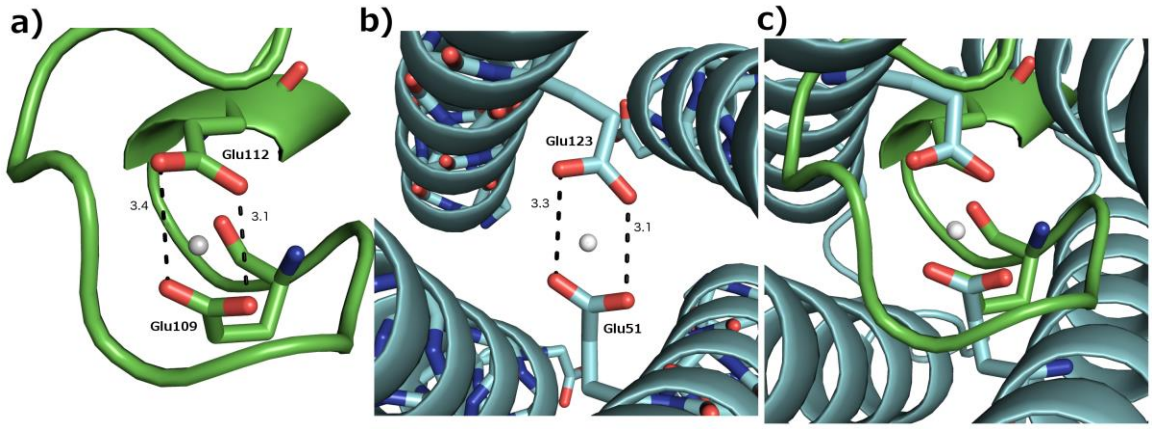

Figure 3.6: **Comparison of bidentate side chains in LBT coordination sphere with coordination sphere of scCC-Tb.** Oxygen atoms of Glu112 and Glu109 in LBT were aligned to Glu123 and Glu51 in scCC-Tb, resulting in an RMSD of 0.046 Å.

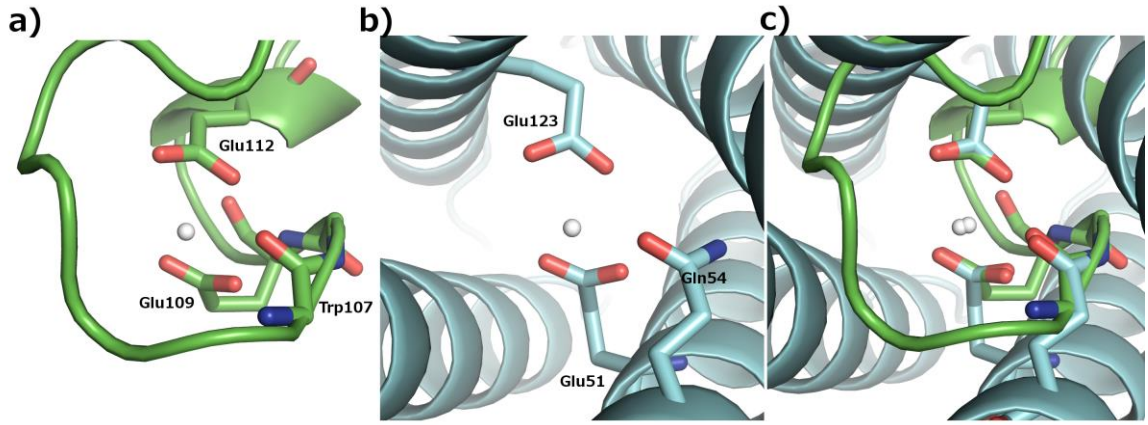

Figure 3.7: **Comparison of three side chains in LBT coordination sphere with coordination sphere of scCC-Tb.** Backbone oxygen atom of Trp107 in LBT and oxygen atom of Gln54 in scCC-Tb were added to the alignment, resulting in an RMSD of 0.267 Å.

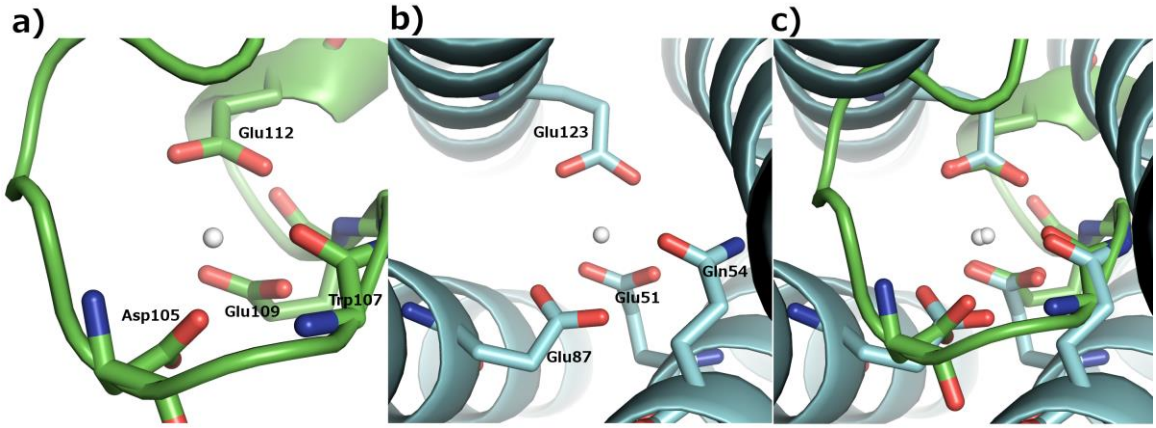

Figure 3.8: **Comparison of four side chains in LBT coordination sphere with coordination sphere of scCC-Tb.** γ carbon atom of Asp105 and δ carbon atom of Glu87 were added to the alignment, resulting in an RMSD of 0.273 Å.

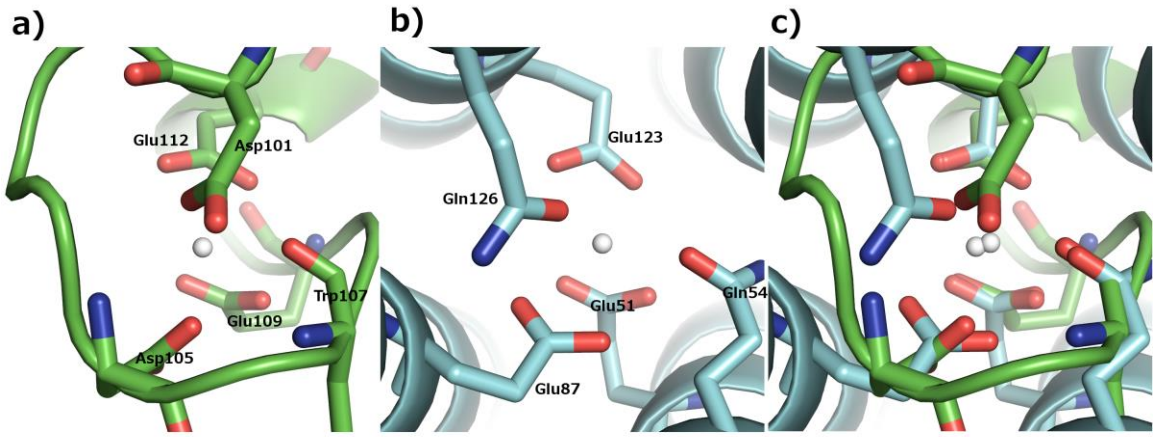

Figure 3.9: **Comparison of five side chains in LBT coordination sphere with coordination sphere of scCC-Tb.** Oxygen atom of Asp101 in LBT and oxygen atom in Gln126 were added to the alignment, resulting in an RMSD of 0.350 Å.

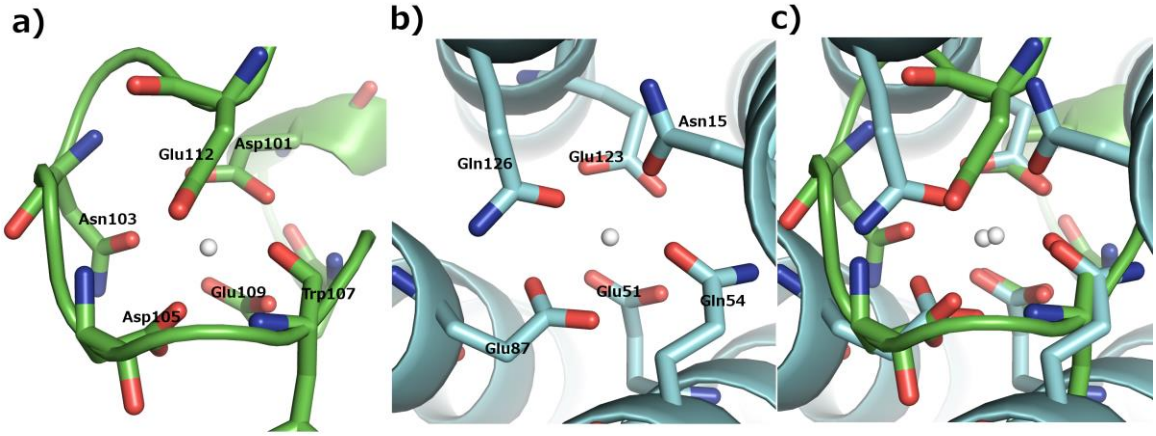

Figure 3.10: **Comparison of all side chains in LBT coordination sphere with coordination sphere of scCC-Tb.** The 6th side chain in LBT and scCC-Tb lie on opposite ends of the coordination sphere.

Five out of the six side chains in the coordination spheres of LBT and scCC-Tb are in close agreement as judged by an RMSD of 0.350 Å for the atoms indicated in Fig. 3.11. This similarity in the arrangement of atoms about the metal position was arrived at

without any input from the LBT structure. Of particular interest is the overlap between Trp107 in LBT and Gln54 in scCC-Tb, where the coordinating backbone atoms map onto coordinating side chain atoms in our design. Also of note is the fact that Glu87 in scCC-Tb adopts a bidentate coordination, but Asp105 in LBT shows monodentate coordination. This suggest that optimal high-density packing arragements side chains around a lanthanide ion are limited, even when allowing for differences in binding modes. However, Asn15 in scCC-Tb and and Asn103 in LBT do lie on opposite ends of the coordination sphere. To determine if LBT has indeed arrived at the best possible coordination sphere packing arragement possible with natural amino acids, variants of scCC-Tb could be designed that place all six side chains in the same position of the coordination sphere and the change in the lanthanide binding affinites could be measured.

3.11. Molecular Dynamics Simulation of Full Construct

In order to ascertain the viability of the design structure to bind to terbium with the predicted coordination sphere geometry, the scCC-Tb model was subjected to 50 ns of molecular dynamics simulation using the CHARMM22 all-atom force field. Force field parameters for the terbium ion were taken from those of the calcium ion, and the charge was set to +3. However, the CHARMM22 parameter file provides two different atomic radii for calcium. The structure for the lanthanide-binding tag (PDB ID 2B22 and chain A) was minimized for 10000 steps in NAMD[133] using both radii for the terbium ion, and the parameter that best fit the crystal structure side chain-to-metal distances was chosen as the one to use in the simulations. A value of R = 1.7 showed better agreement with the crystal structure for 4 out of the 6 ligands.

Amino Acid Residue	Ligating Atom	Atom to Metal Distance (Å), R = 1.367	Atom to Metal Distance (Å), R = 1.7	Atom to Metal Distance (Å), Crystal Structure
D101	Oδ	2.04	2.28	2.26
N103	Oδ	2.21	2.43	2.23
D105	Oδ	2.06	2.35	2.38
W107	O	2.33	2.50	2.30
E109	Cδ	2.52	2.74	2.81
E112	Cδ	2.57	2.79	2.83

Table 3.8: **Selection of Tb^{3+} VdW radius parameter.** Ligand-to-metal distances after 10000 steps of minimization for the lanthanide-binding tag.

The designed scCC-Tb model structure was taken as the initial configuration and placed in a 50 x 50 x 95 $Å^3$ water box with 0.15 mol/L NaCl. Forces between atoms were calculated every 1 fs, assuming rigid bonds. A 12.0 Å cutoff distance for non-bonded interactions was used with a switching function at 10.0 Å. Long-range electrostatic interactions were calculated using the particle-mesh Ewald method, with full electrostatics calculated every 2 steps. The NPT ensemble was used at a temperature of 320° K. 1000 conjugate gradient minimization steps were conducted, followed by 50 ns of simulation time. Calculations were carried out in NAMD.

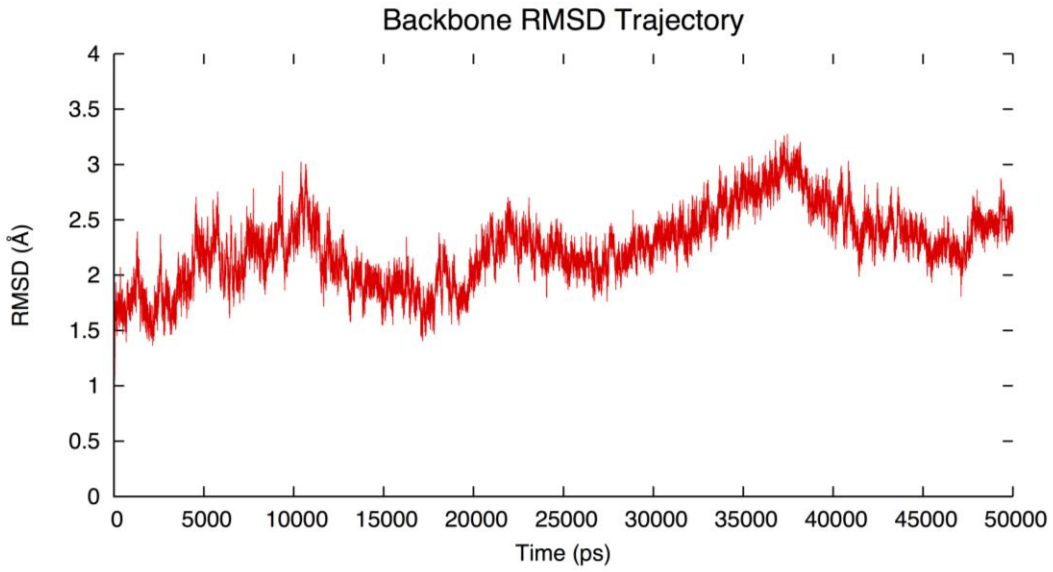

Figure 3.11: **Trajectory of scCC-Tb simulation.** Backbone root-mean-square-deviation between the initial model and the simulation trajectory.

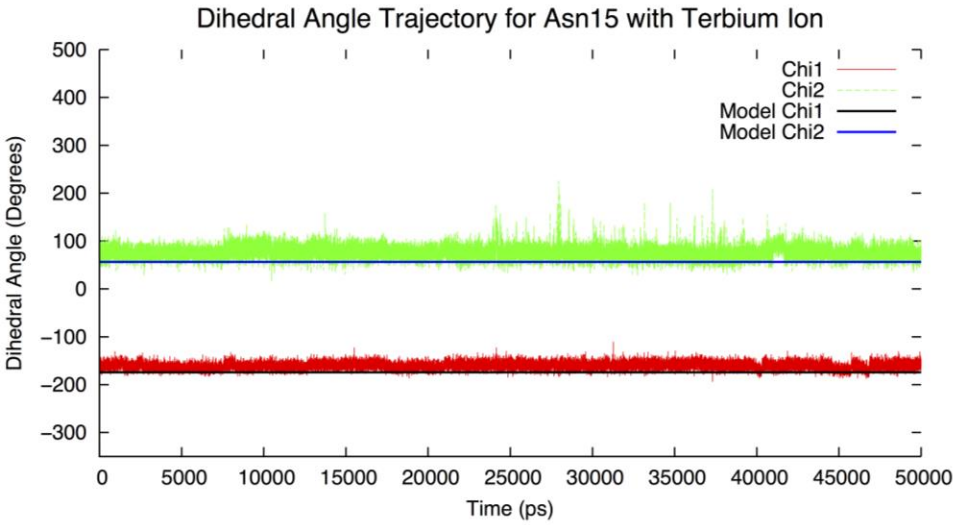

Figure 3.12: **Dihedral angle deviations of asparagine 15.** χ_1 and χ_2 torsional angles of asparagine 15 along the simulation trajectory.

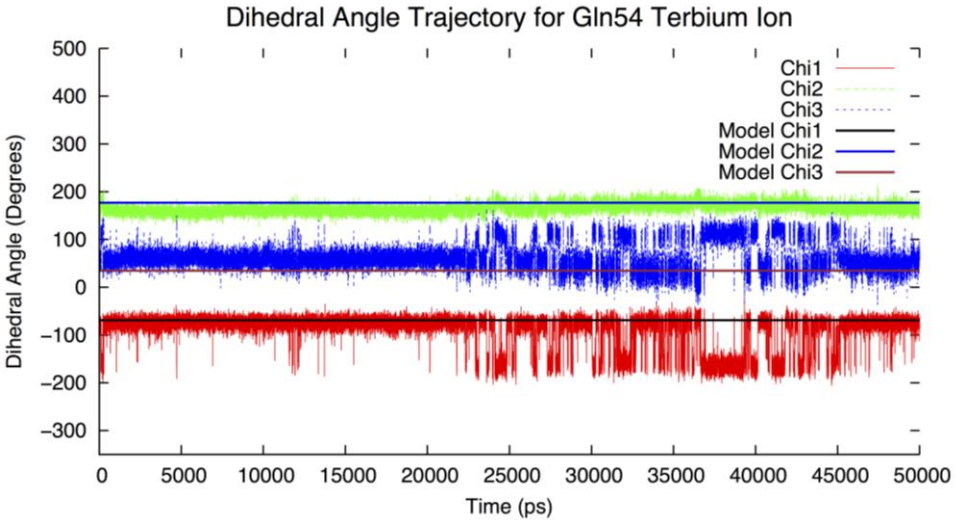

Figure 3.13: **Dihedral angle deviations of glutamine 54.** χ_1, χ_2, and χ_3 torsional angles of glutamine 54 along the simulation trajectory.

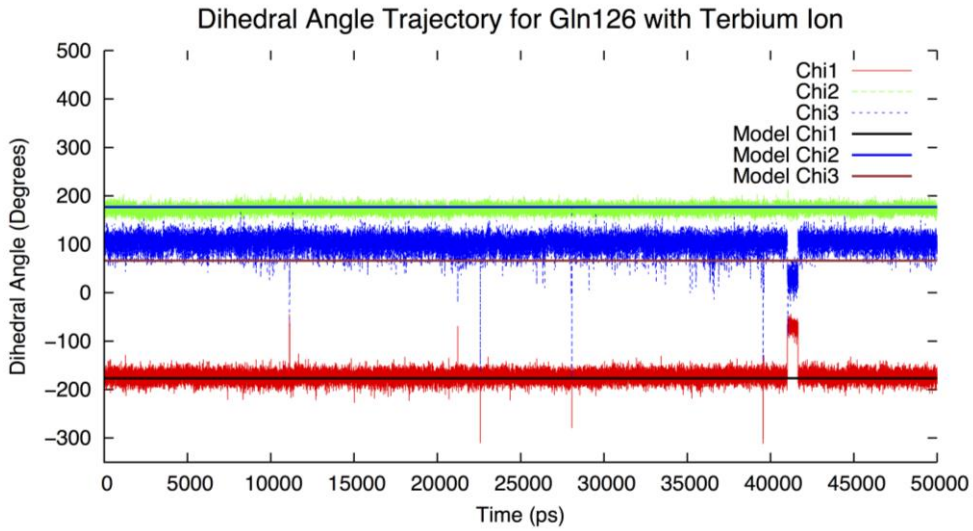

Figure 3.14: **Dihedral angle deviations of glutamine 126.** χ_1, χ_2, and χ_3 torsional angles of glutamine 126 along the simulation trajectory.

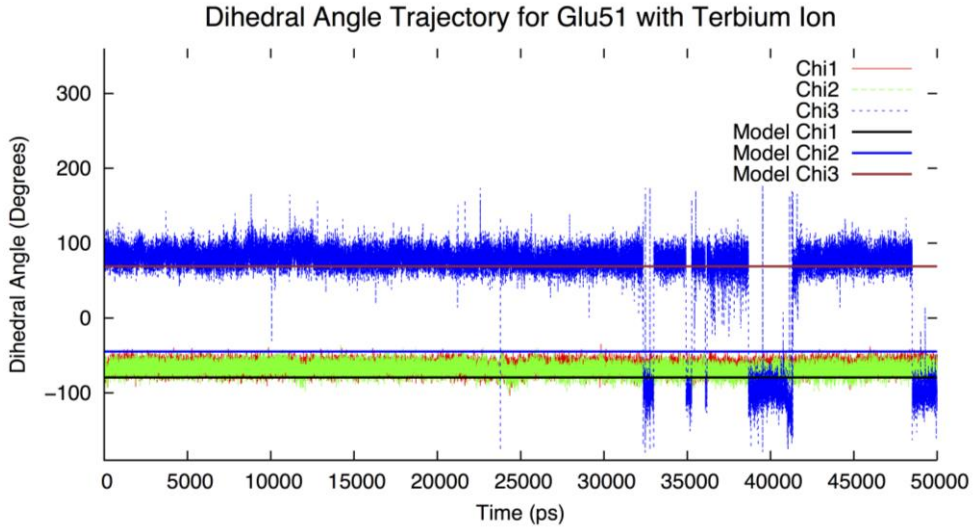

Figure 3.15: **Dihedral angle deviations of glutamate 51.** χ_1, χ_2, and χ_3 torsional angles of glutamine 51 along the simulation trajectory.

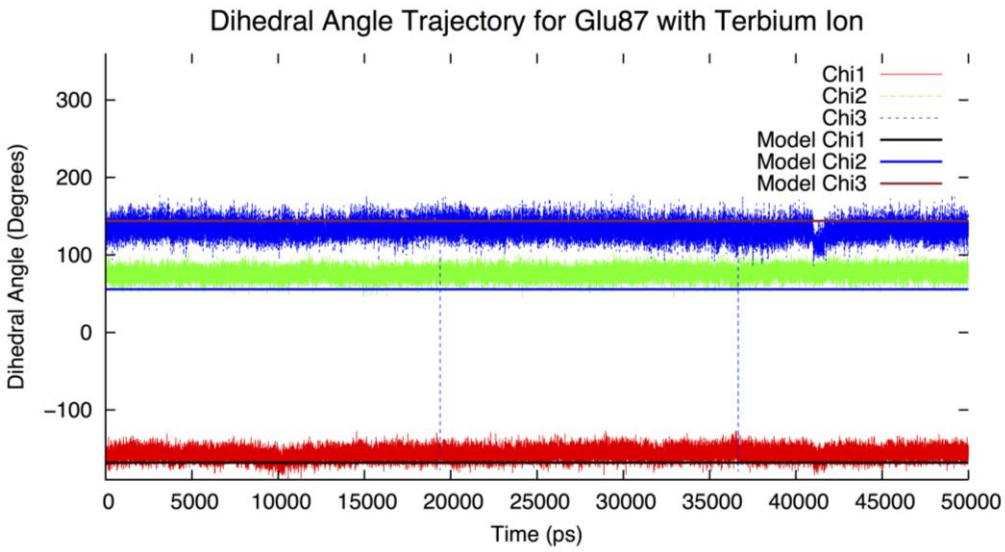

Figure 3.16: **Dihedral angle deviations of glutamate 87.** χ_1, χ_2, and χ_3 torsional angles of glutamine 87 along the simulation trajectory.

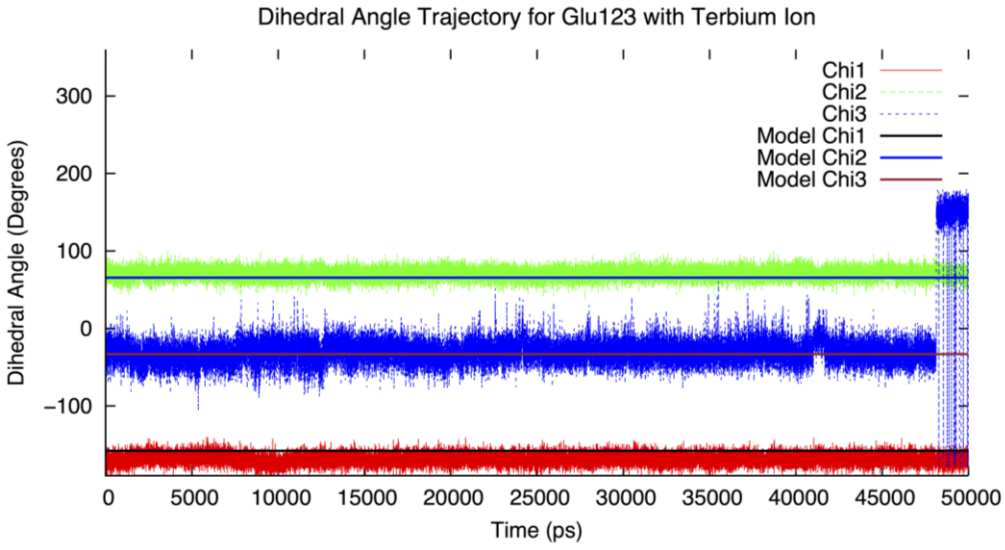

Figure 3.17: **Dihedral angle deviations of glutamate 123.** χ_1, χ_2, and χ_3 torsional angles of glutamine 123 along the simulation trajectory.

Coordination sphere side chains remained prevalently in their target conformations over the course of the simulation trajectory, leading us to hypothesize that scCC-Tb would be promising for experimental studies of binding to lanthanide ions in solution.

3.12. Experimental Characterization

3.12.1. Protein Expression and Purification.

The N terminal sequence MSSHHHHHHSSENLYFQG was added to the scCC-Tb sequence to provide for a nickel-column binding tag, and a tobacco etch virus (TEV) protease cleavage site. TEV protease was selected for its high specificity, although a trailing glycine residue would remain after cleavage. PJ414 vector with IPTG-inducible T7 promoter and ampicillin resistance was purchased from DNA2.0. Plasmids were transformed in *E. coli* BL21-CodonPlus(DE3)-RIL cells and grown overnight on agar ampicillin plates. Single colonies were incubated overnight in 10 mL of Lenox Broth (LB) medium at 37 °C while spinning at 275 RMP, with 100 μg/mL ampicillin and 50 μg/mL chloramphenicol. Cultures were transferred to 1 L of LB with antibiotics at the same concentrations. Cells were grown at 37 °C at 275 RPM and the OD600 reached ~0.6−0.8. Cells were induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at 37 °C at 275 RPM for 3 hours.

3.12.2. Attempted Purification from the Soluble Fraction

Cell pellet was combined with 5 mL/g of Bugbuster Master Mix, and 1 pellet of protease inhibitor cocktail per 10 mL of solution. Solution was rotated for 30 minutes at room temperature. Solution was spun for 15 minutes at 11,000 rpm at 4 °C, and the supernatant

was collected and filtered through a 22 μm filter. The solution was injected onto a HisPrepFF 16/10 column on an AKTA FPLC instrument (GE Healthcare). Solution was loaded onto column in HIS start buffer (20 mM Tris, 20 mM imidazole, 150 mM NaCl, pH 7.5) and eluted with 60% HIS elution buffer (20 mM Tris, 500 mM imidazole, 150 mM NaCl, pH 7.5).

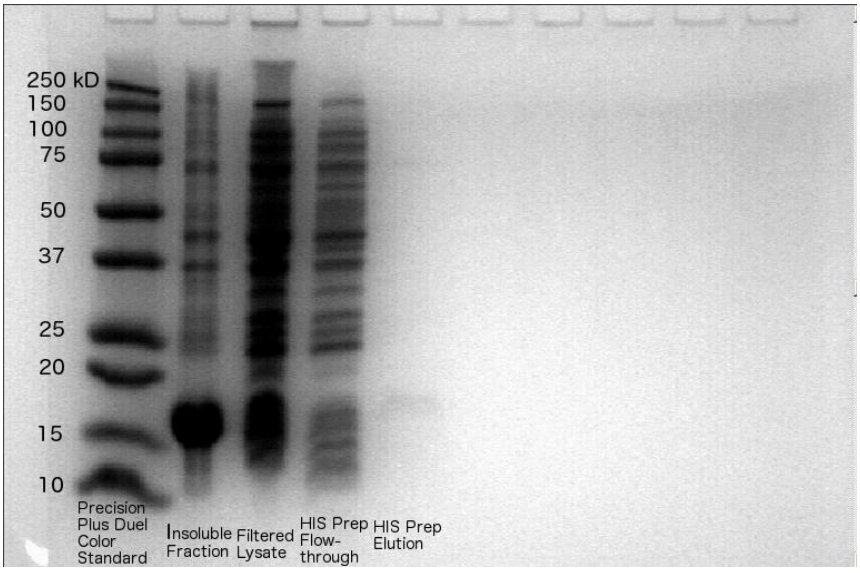

Figure 3.18: **Gel electrophoresis of scCCTb.** The expected molecular weight of the uncleaved protein is 17,918 Da. A dominant dark band at this range was observed in the sample from the insoluble fraction. Although bands at this range were observed for the soluble fraction, no overexpression relative to the other bands was observed. Minimal protein at this MW range was recovered by HIS column purification, although some separation could be observed from the column flow-through. These results indicated that indicated that protein is expressed mainly in the insoluble fraction.

Sodium docecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) indicated that the bulk of the protein was expressed in the insoluble fraction. The soluble fraction also shows a larger amount of impurities.

3.12.3. Purification from the Insoluble Fraction

Cell pellet was combined with 10 mL/g of refolding buffer (50 mM Tris, 140 mM NaCl, pH 8.0), and 1 pellet of protease inhibitor cocktail per 10 mL of solution. Solution was sonicated. Solution was spun for 15 minutes at 11,000 rpm at 4 °C, and the supernatant was discarded. The insoluble fraction was combined with 5 mL/g of denaturing buffer (8M urea, 100 mM Tris, pH 8.5) and 1 pellet of protease inhibitor cocktail per 10 mL of solution. Solution was rotated for 60 minutes at room temperature. Solution was spun for 15 minutes at 11,000 rpm at 4 °C, and the supernatant was collected. To the supernatant solution was added 7 mL/mL of refolding buffer in a drop-wise manner while stirring. The final urea concentration was 1M. Solution was again spun for 15 minutes at 11,000 rpm at 4 °C, and the supernatant was collected and filtered through a 22 μm filter. The solution was injected onto a column packed with 10 mL of nickel-loaded HIS-bind resin (EMD Millipore Corp) on FPLC. Solution was loaded onto column in HIS start buffer and eluted with 60% HIS elution buffer.

Elution peak fractions were concentrated down to less than 2 mL by transferring fractions to an Ultra-15 10K Centrifugal Filter Device (Amicon) and spinning at 6,000 RPM. Solution was combined with 10 mL of TEV protease solution filtered through a 22 μm filter, along in 1 mM 1,4-dithiothreitol (DTT) and 0.5 mM ethylenediaminetetraacetic acid (EDTA). The solution was injected onto a HIS Prep 26/10 desalting column (GE Healthcare) and buffer exchanged into a TEV cleavage buffer (50 mM Tris, 140 mM NaCl, pH 8.0). Pooled fractions were incubated at 4 °C overnight at the same concentrations of DTT and EDTA. TEV cleavage reaction product injected onto HIS

Prep 26/10 desalting column and buffer exchanged into HIS start buffer. Elution was loaded onto HIS-bind resin column in HIS start buffer and flow-through was collected. Fractions were concentrated down to less than 12 mL by transferring fractions to an Ultra-15 10K Centrifugal Filter Device and spinning at 6,000 RPM. Solution was injected onto HIS Prep 26/10 desalting column and buffer exchanged into scCC-Tb buffer (250 mM MES, 140 mM NaCl, pH 6.0). Elution was concentrated down to less than 2 mL by transferring fractions to an Ultra-15 10K Centrifugal Filter Device and spinning at 6,000 RPM. Solution was loaded onto HiLoad Superdex 16/10 column. Protein eluted as a single peak and the protein concentration was determined by UV-vis absorbance at 280 nm. Protein yield was 10.4 mg/L of LB broth.

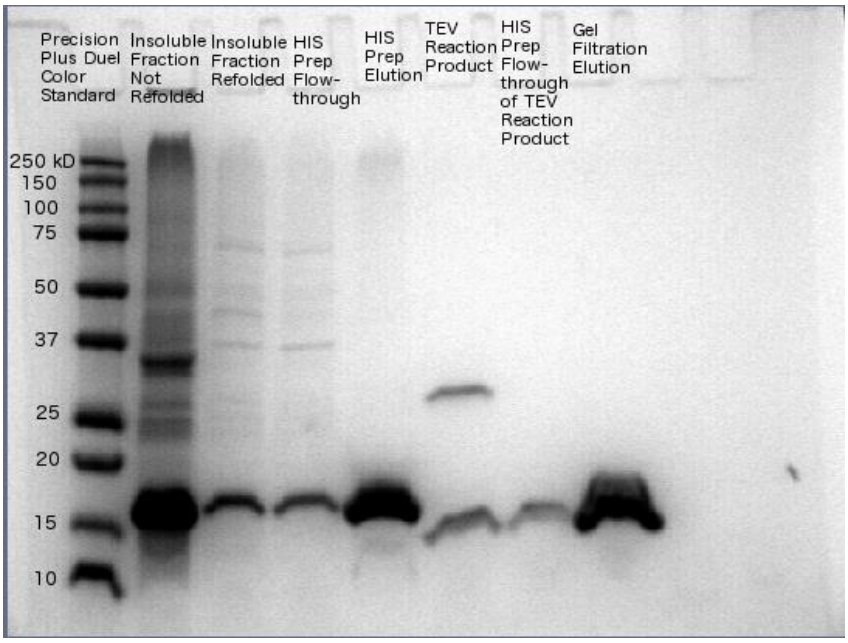

Figure 3.19: **SDS-PAGE of scCCTb purified from the insoluble fraction.** The expected molecular weight is 17,918 Da for the uncleaved protein, and 15821.45 Da for the HIS-tag cleaved protein. Protein is refolded by drop-wise addition of refolding buffer, and recovered by HIS-column purification. TEV protease is used to cleave off the HIS-tag, and removed by HIS-column purification. A drop in molecular weight is detected by the difference in migration distances of the HIS-column purified product and the TEV-cleavage reaction product. Final purification is performed by gel filtration.

Mass was confirmed by matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS). Sinapinic acid was used as a matrix, and the instrument was operated in linear positive ion mode.

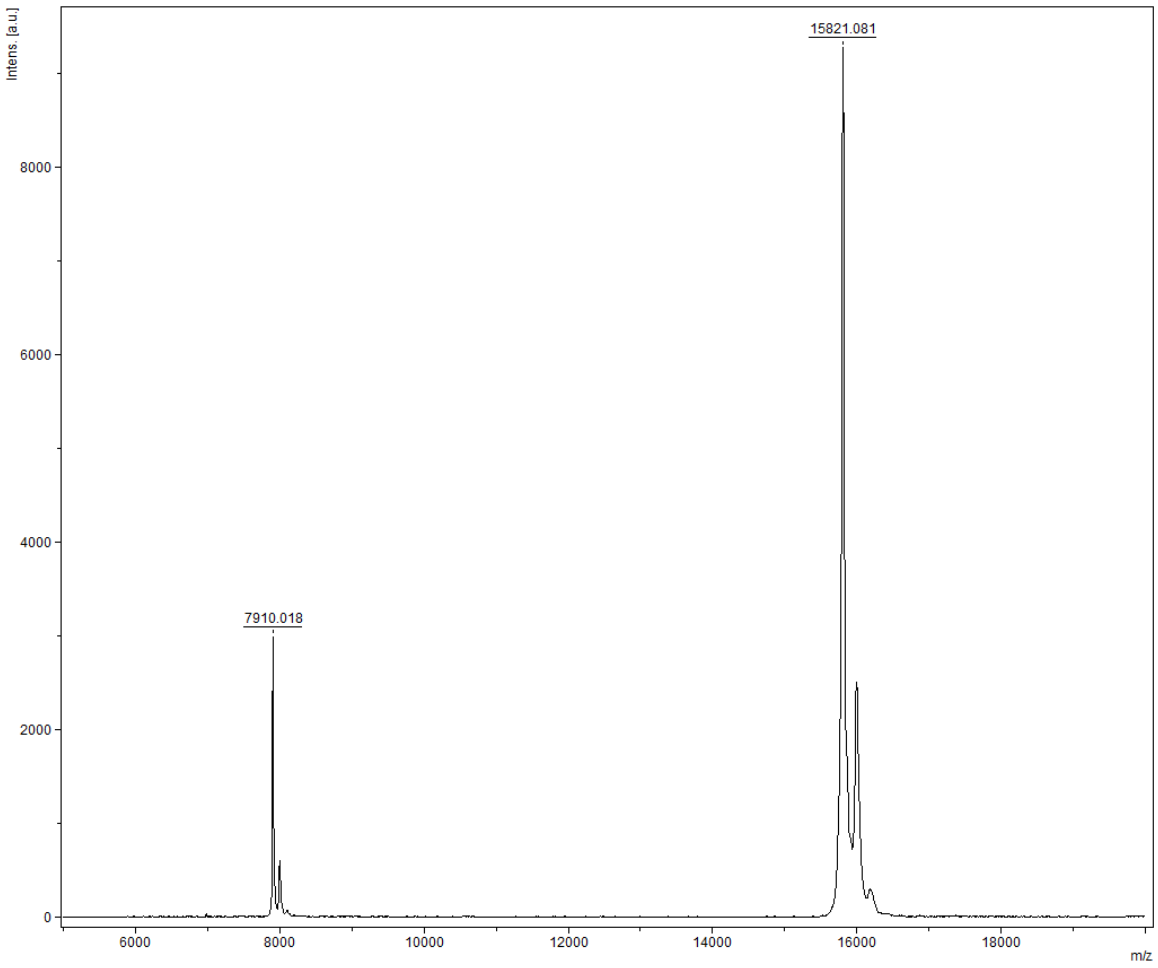

Figure 3.20: **Confirmation of HIS-tag free scCC-Tb identity by MALDI-TOF MS.** Expected molecular mass of scCC-Tb with trailing glycine residue is 15821.45 Da. M+1 and M+2 peaks confirm the identity of the protein.

3.12.4. Circular Dichroism (CD) Spectroscopy

CD measurements were collected using an Aviv 410 CD spectrometer (Aviv Biomedical, Lakewood, NJ). Temperature-melt experiments were carried out using a 1 mm quartz cuvette (Starna Cells) at a protein concentration of 15 μM in 140 mM NaCl and 250 mM MES buffer. Ellipticity at 222 nm was monitored as a function of temperature from 25 °C

to 70 °C in 1 °C steps. The temperature was equilibrated for 1 min at each step, and the signal was averaged over 300 s. A structural rearrangement of the protein was observed to occur while heating to 70° C. Temperature cycle was repeated to determine if protein had achieved a stable folded state. A second temperature ramp showed no significant structural rearrangement and no hysteresis upon cooling to 25° C.

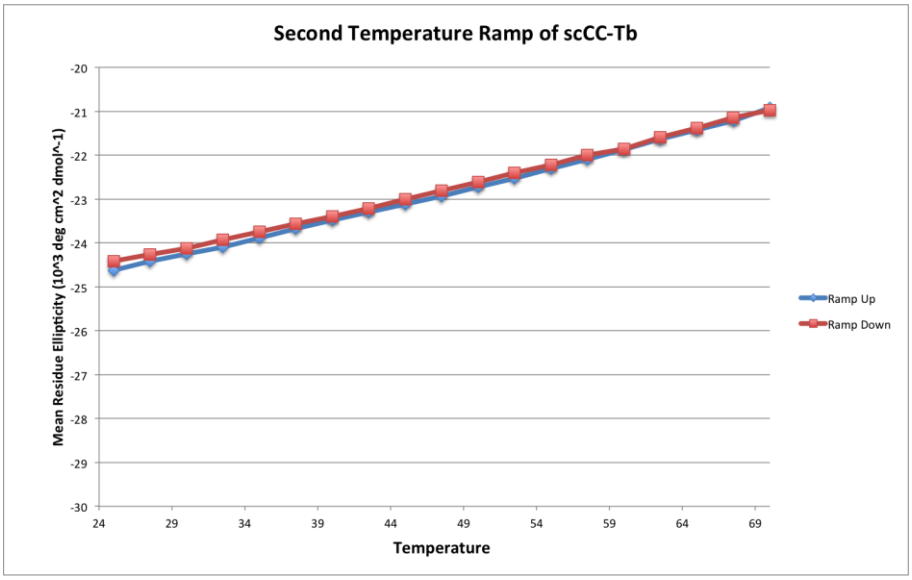

Figure 3.21: **Structural effect of temperature ramp on folded scCC-Tb.** Measurement of mean-residue ellipticity as 222 nm as a function of temperature indicates that the folded protein can withstand temperature ramps to 70° C.

A thermal melt was then carried out from 4° C to 95° C to determine is protein retained its structure after heating. No melting-point transition up to 95 °C and no significant hysteresis were observed upon cooling to 4 °C, indicating that protein is highly stable even in the absence a bound metal ion.

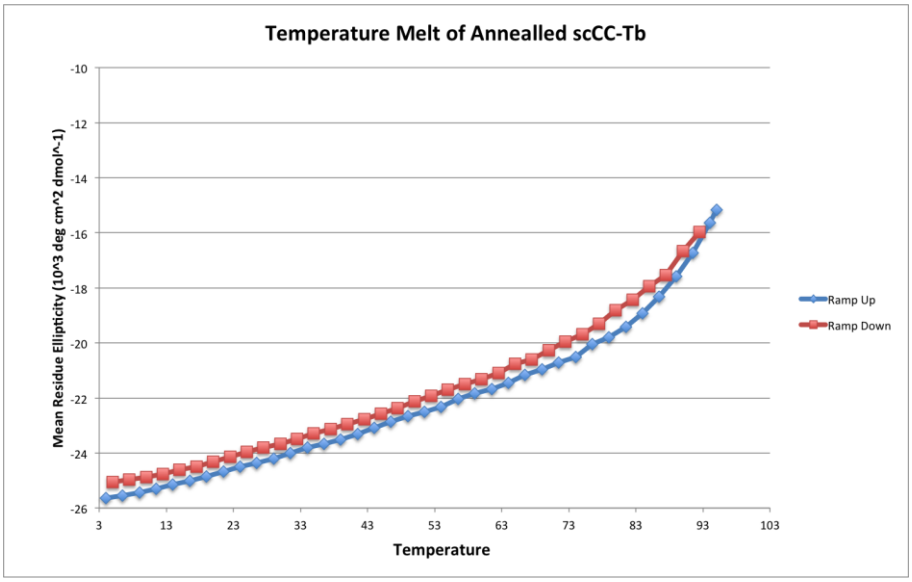

Figure 3.22: **scCC-Tb temperature melt monitored by CD spectroscopy.** Measurement of mean-residue ellipticity as 222 nm as a function of temperature indicates that the folded protein does not undergo a melting transition up to 95°C. Protein returns to fully folded state upon cooling.

Having confirmed that temperature annealing is needed for the preparation of the *apo* sample, and that 70° temperature cycles do not disrupt the structure of the folded protein, we performed CD measurements on samples of *apo* and *holo* scCC-Tb. Thermo cycling steps were carried out by incubating the samples at 70° C for 20 minutes, and cooling in a room temperature bath for 3 minutes. Samples were centrifuged for 3 minutes at 16,100 RCF at 25° C. For the preparation of the *apo* samples, two thermo cycling steps were carried out to ensure that protein reached a stable folded state. For the preparation of the *holo* sample, one thermo cycling step before addition of the metal ion was performed, and thermo cycling was repeated after addition of the metal ion.

Wavelength scans were carried out on samples of *apo* scCC-Tb and *holo* scCC-Tb using 0.1 mm quartz cuvette (Starna Cells) at a protein concentration was 50 μM in 140 mM NaCl and 250 mM MES buffer. *Holo* sample was prepared by incubating 50 μM protein with 3 equivalents of $Tb(NO_3)_3$.

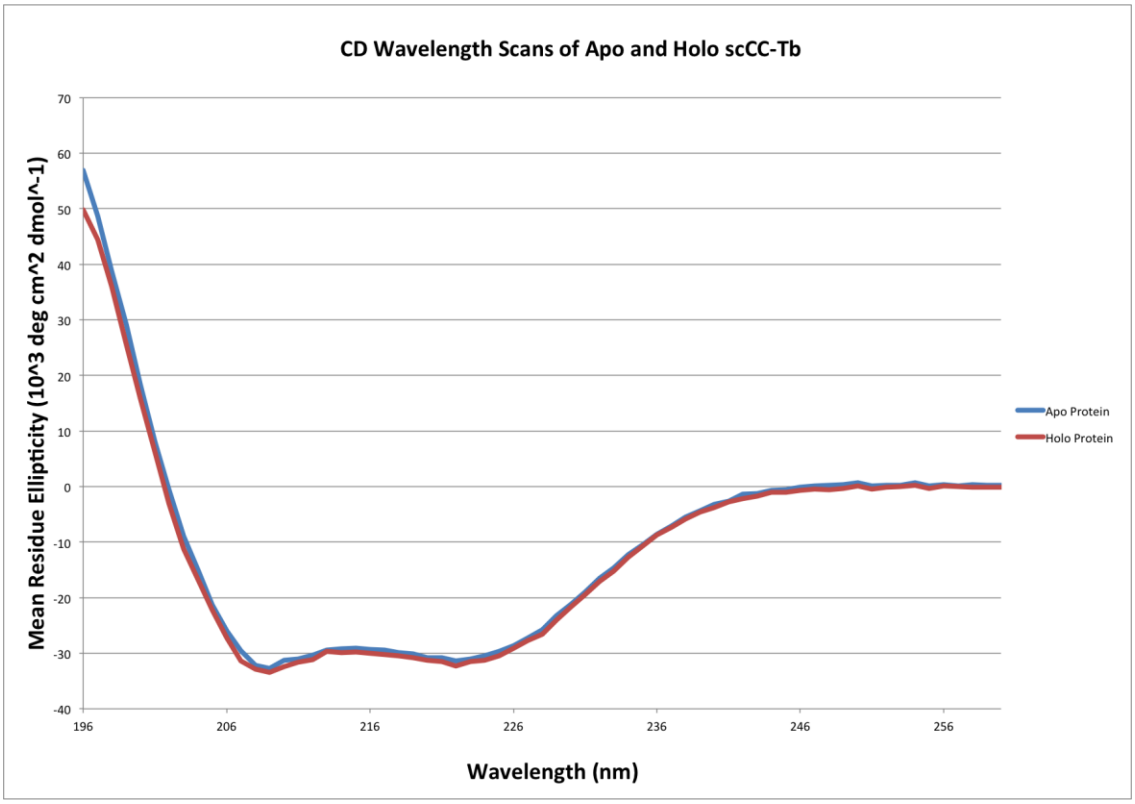

Figure 3.23: **CD wavelength scans for *apo* and *holo* scCC-Tb.** Scans from 260 nm to 196 nm show a characteristic α-helical signal, consistent with the design model. *Apo* and *holo* do not exhibit a difference in α-helical character, indicating that the protein is folded in the absence of the metal.

Wavelength scans show a characteristic α-helical signal, confirming the α-helical character of the *apo* and *holo* protein. Spectra were nearly identical for the *apo* and *holo* protein, indicating that the *apo* protein if fully folded in the absence of the metal, and that binding to the metal ion does not induce any further folding.

Thermal melts were carried out on samples of 15 μM *apo* scCC-Tb and *holo* scCC-Tb from 25° C to 95° C as indicated for above. *Holo* sample was prepared by incubating protein with 10 equivalents of $Tb(NO_3)_3$.

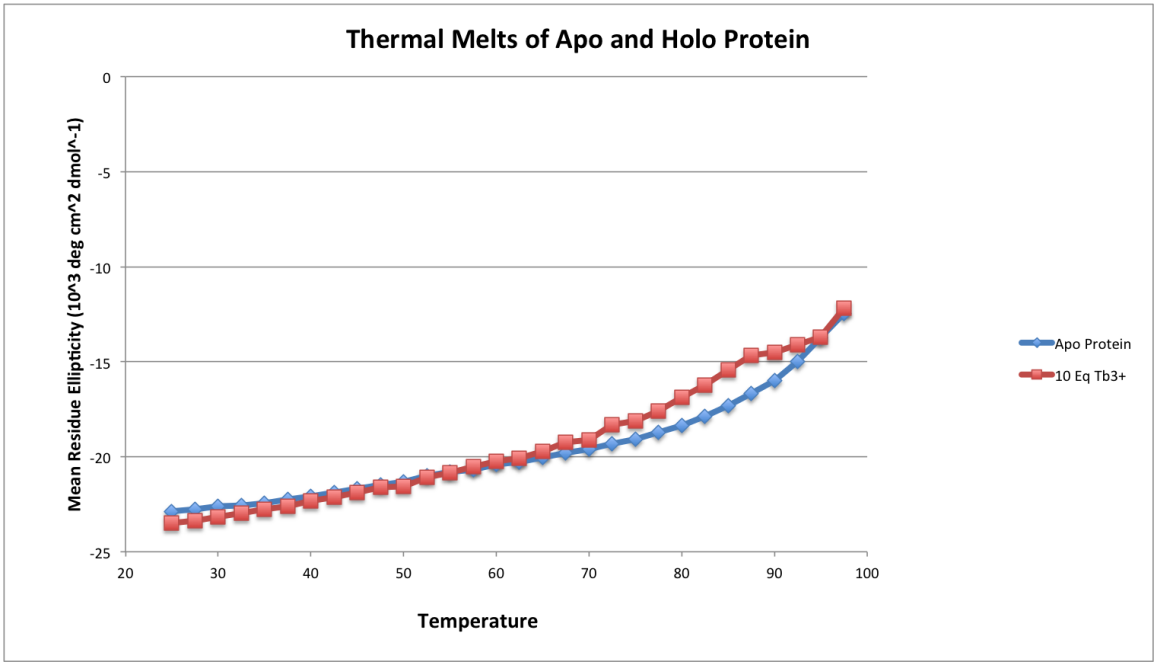

Figure 3.24: **scCC-Tb temperature melt of *apo* and *holo* forms monitored by CD spectroscopy.** Measurement of mean-residue ellipticity as 222 nm as a function of temperature indicates that the folded protein does not exhibit an appreciable change in thermal stability upon binding to the metal.

The *holo* protein does exhibit an enhancement in stability over the *apo* protein, confirming that the *apo* protein is in a fully folded and stable state.

3.12.5. Fluorescence Wavelength Scans

Fluorescence spectra were recorded on a Cary Eclipse fluorimeter spectrophotometer (Varian) using a 100 μL quartz cuvette of 1 cm path length (Starna Cells). An excitation slit width of 20 nm and an emission slit width of 5 nm were used. Samples were excited at 280 nm and fluorescence was collected at from 450 nm to 700 nm in 1 nm intervals with 5 second averaging. Experiments were carried out 25° C using a 360 nm high band-pass filter, with a PMT voltage of 800 V. Solutions of 50 μM *apo* and *holo* protein were prepared, with the *holo* sample incubated with 3 equivalents of $Tb(NO_3)_3$.

Free terbium ions exhibit low fluorescence signal due to weak photon absorption, and non-radiative decay of electronically excited states by vibrational quenching of coordinated water molecules. Fluorescence emission of terbium upon excitation of aromatic residues would indicate FRET between the protein and the metal ion, and expelling of water molecules from the terbium coordination sphere.

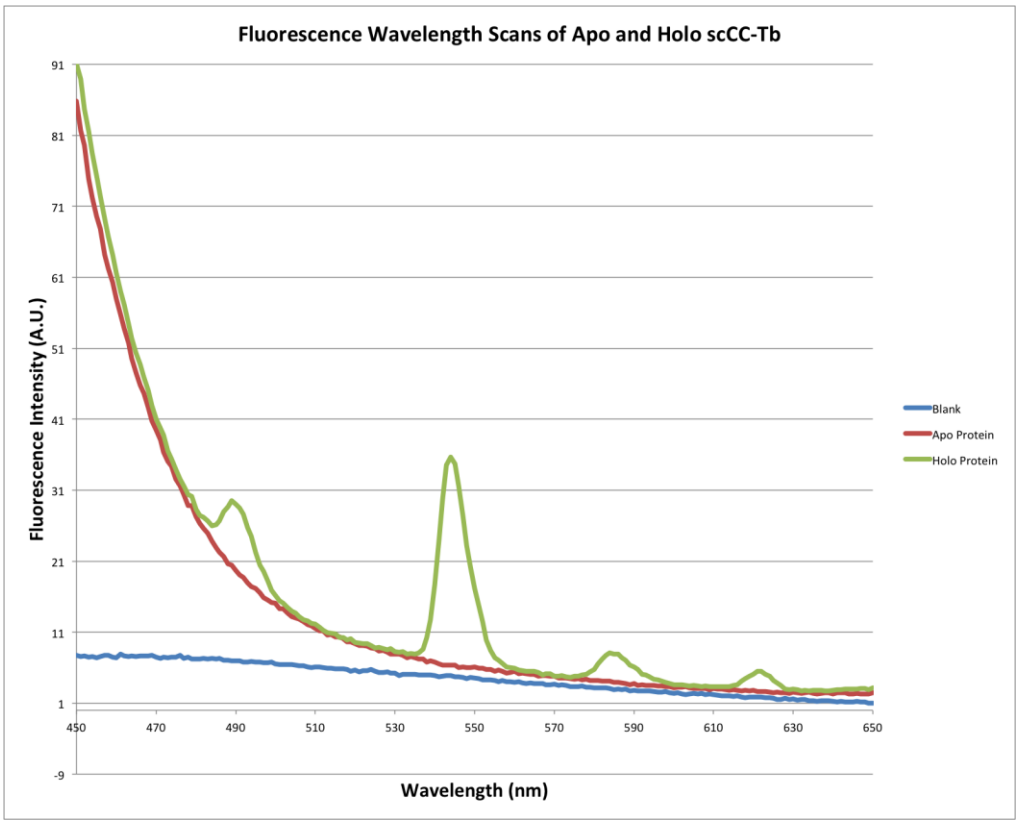

Figure 3.25. **Fluorescence wavelength scans of *apo* and *holo* scCC-Tb.** Fluorescence enhancement of terbium is observed when compared to a blank sample with no protein.

Measurements of the fluorescence of *holo* protein upon excitation at 280 nm results in a characteristic terbium emission spectrum, with enhance fluorescence observed over a protein-free sample. This result indicates that the terbium ion is binding to the protein, and that the protein is replacing water atoms in the coordination sphere of the terbium ion.

3.12.5. Fluorescence Titrations

Fluorescence spectra were recorded on a Cary Eclipse fluorimeter spectrophotometer (Varian) using a 12 μL quartz cuvette of 1.5 mm path length (Starna Cells). An excitation slit width of 20 nm and an emission slit width of 5 nm were used. Samples were excited at 280 nm and fluorescence was collected at 544 nm at 25° C using a 360 nm high band-pass filter, with a PMT voltage of 800 V. Samples were prepared by incubating 250 μL of 50 μM protein at 70° C for 20 minutes, and cooling in a room temperature bath for 3 minutes. Samples were centrifuged for 3 minutes at 16,100 RCF at 25° C. Thermo cycling step was carried out two times to ensure that protein was folded to a stable state. Terbium titration was performed by adding 0.2 equivalent aliquots of $Tb(NO_3)_3$ to 250 μL of protein sample. Thermal cycling and centrifugation were repeated after every aliquot addition of metal ion to prepare equilibrated samples.

Data was fitted to a 1:1 equilibrium binding model[134]:

$$M + L \Leftrightarrow ML \tag{3.2}$$

where M it the metal ion, and L is the ligand (in this case the protein). The expression for the dissociation constant is:

$$K_D = \frac{[M][L]}{[ML]} \tag{3.3}$$

Substituting the mass balance equations:

$$M_T = [M] + [ML] \qquad (3.4)$$
$$L_T = [L] + [ML]$$

and solving for [ML] yields the function for the formation of the metal-ligand complex as:

$$[ML] = 0.5 \times (K_D + L_T + M_T - \sqrt{(-K_D - L_T - M_T)^2 - 4L_T M_T}) \qquad (3.5)$$

The fluorescence data is fitted to the model as:

$$F = F_L + \Delta F \times [ML] \qquad (3.6)$$

where fluorescence measurement (F) changes from the fluorescence of the *apo*-protein (F_L) as a function of [ML]. ΔF accounts for the changes in fluorescence of the metal and protein from their unbound states to their bound states.

The titration data in Fig 3.24 indicate 1:1 binding of the terbium ion to the protein as observed from the transition centered at 50 μM of Tb^{3+} added. Fitting of the data to the equilibrium binding model results in a calculated K_d of 11.76 ± 5.8 with a 95% confidence interval.

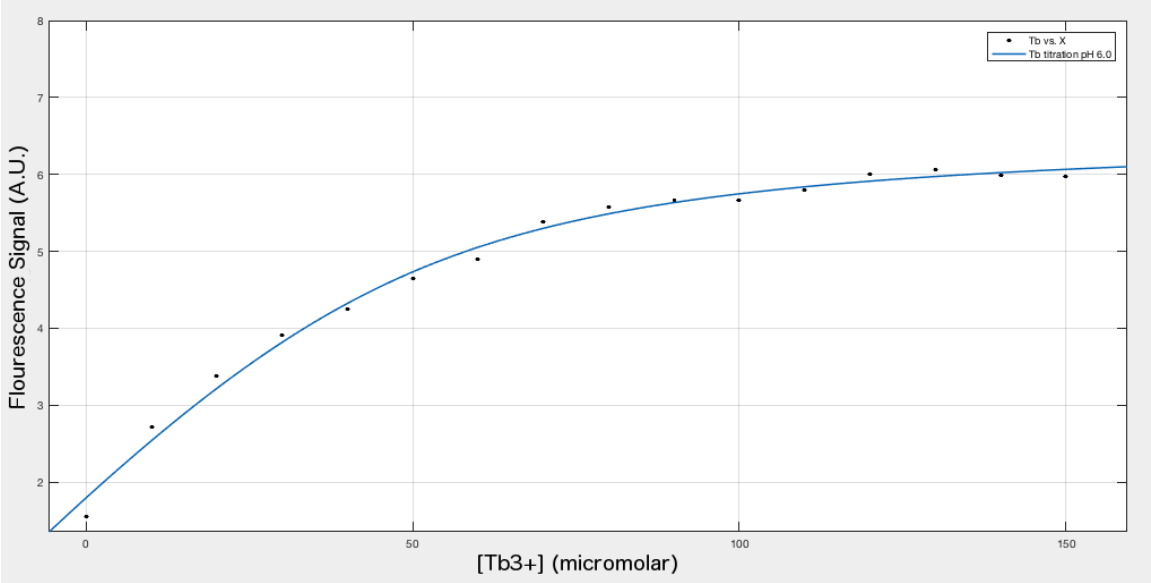

Figure 3.26: **Fluorimetric titration of Tb^{3+} binding to scCC-Tb.** Binding isotherm data was collected at pH 6, and fit to a K_d value of 12 ± 6 μM. Protein concentration was 50 μM.

Dissociation constants for other lanthanides were determined by competitive binding experiments between a terbium loaded sample and the metal of interest. Fluorescence measurements were collected as indicated above, where now the decrease in terbium fluorescence is measured as a competitive lanthanide ion is titrated into the solution. Samples were prepared by incubating 250 μL of 50 μM protein at 70° C for 20 minutes, and cooling in a room temperature bath for 3 minutes. Samples were centrifuged for 3 minutes at 16,100 RCF at 25° C. To each sample, 1.5 μL of 25 mM $Tb(NO_3)_3$ was added to obtain a 3:1 ration of metal to protein. Thermo cycling was repeated and samples were again centrifuged.

Titrations were performed by adding 1.5 equivalent aliquots of Ln^{3+} (Ln = Gd, Eu, Sm, Nd, La) to samples of 50 μM protein loaded with 150 μM terbium. Thermal cycling and

centrifugation were repeated after every aliquot addition of metal ion to prepare equilibrated samples.

Fluorescence data was fitted to an equilibrium model of two simultaneous competing reactions:

$$L + A \Leftrightarrow LA, \quad L + B \Leftrightarrow LB \tag{3.7}$$

where L is the protein, A is the terbium ion, and B is the competing lanthanum ion. The dissociation constants for each reaction are:

$$K_A = \frac{[L][A]}{[LA]}, \qquad K_B = \frac{[L][B]}{[LB]} \tag{3.8}$$

The mass balance equations are:

$$\begin{aligned} A_T &= [A] + [LA] \\ B_T &= [B] + [LB] \\ L_T &= [L] + [LA] + [LB] \end{aligned} \tag{3.9}$$

The derivation of the equation for the relation between the fluorescence signal and the formation of [LB] can be simplified by assuming a full saturation of the protein, making

[L] arbitrarily small. In the case of micromolar affinities, this approximation may not hold true. An exact solution was derived by Wang[135] as:

$$[\mathrm{LB}] = \frac{\mathrm{B_T} \times \{2\sqrt{(a^2 - 3b)}\cos(q/3) - a\}}{3K_\mathrm{B} + \{2\sqrt{(a^2 - 3b)}\cos(q/3) - a\}} \tag{3.10}$$

where:

$$\begin{aligned} q &= \arccos\frac{-2a^3 + 9ab - 27c}{2\sqrt{(a^2 - 3b)^3}} \\ a &= K_\mathrm{A} + K_\mathrm{B} + \mathrm{A_T} + \mathrm{B_T} - \mathrm{P_T} \\ b &= K_\mathrm{B}(\mathrm{A_T} - \mathrm{P_T}) + K_\mathrm{A}(\mathrm{B_T} - \mathrm{P_T}) + K_\mathrm{A}K_\mathrm{B} \\ c &= -K_\mathrm{A}K_\mathrm{B}\mathrm{P_T} \end{aligned} \tag{3.11}$$

Fitting of the competitive titration data to the equilibrium binding model results in calculated dissociation constants for other metal ions in the lanthanide series. K_A was set to the value calculated from the direct titration of terbium to protein.

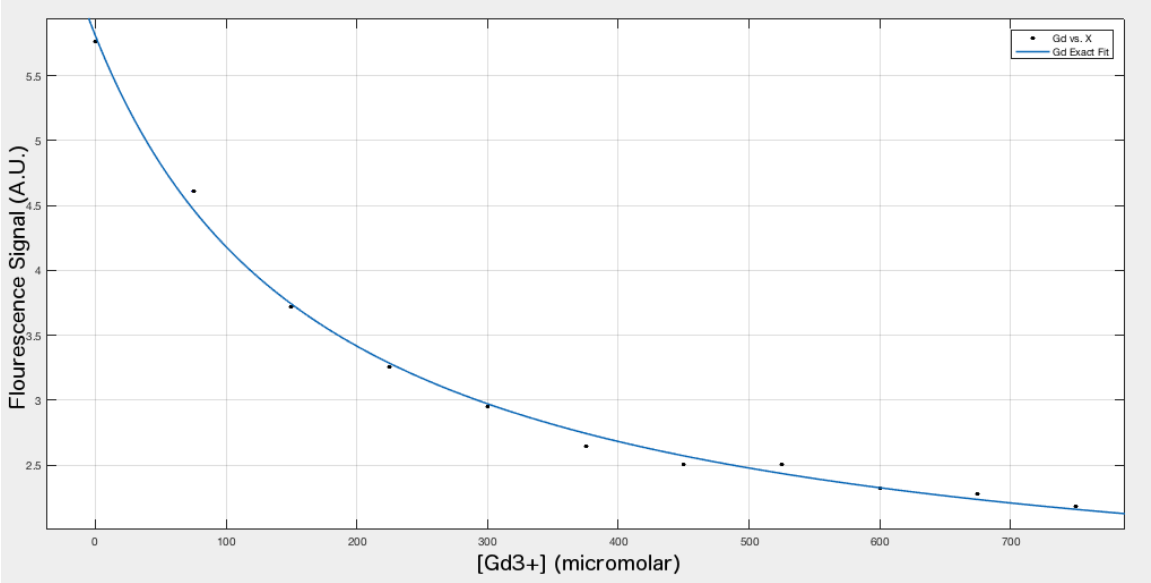

Figure 3.27: **Competitive fluorimetric titration of Gd^{3+} binding to scCC-Tb.** Binding isotherm data was collected at pH 6, and fit to a *K*d value of 13 μM.

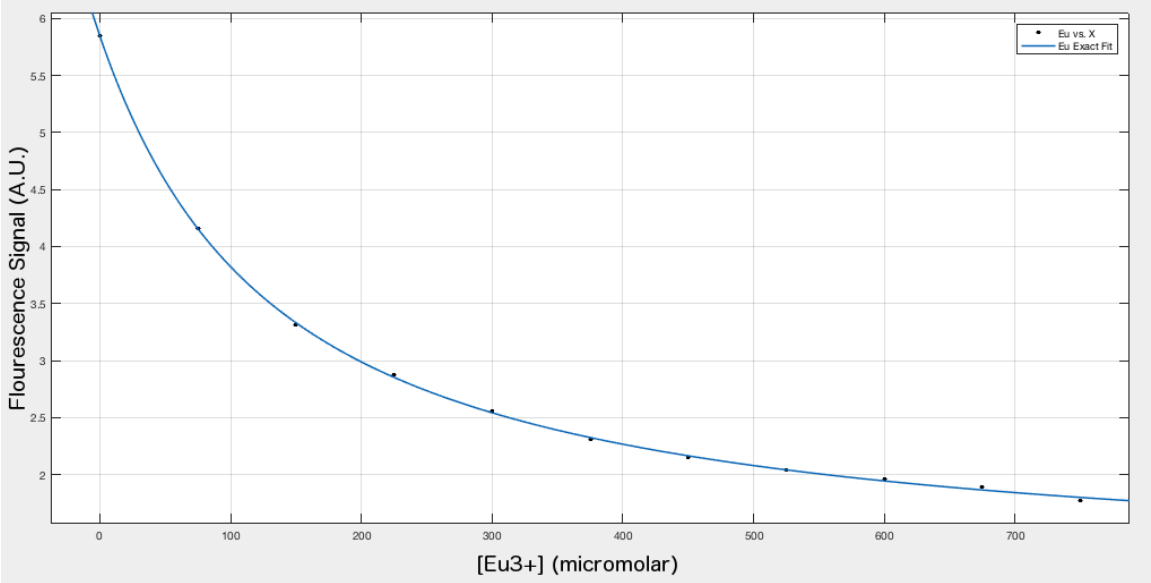

Figure 3.28: **Competitive fluorimetric titration of Eu^{3+} binding to scCC-Tb.** Binding isotherm data was collected at pH 6, and fit to a *K*d value of 9 μM.

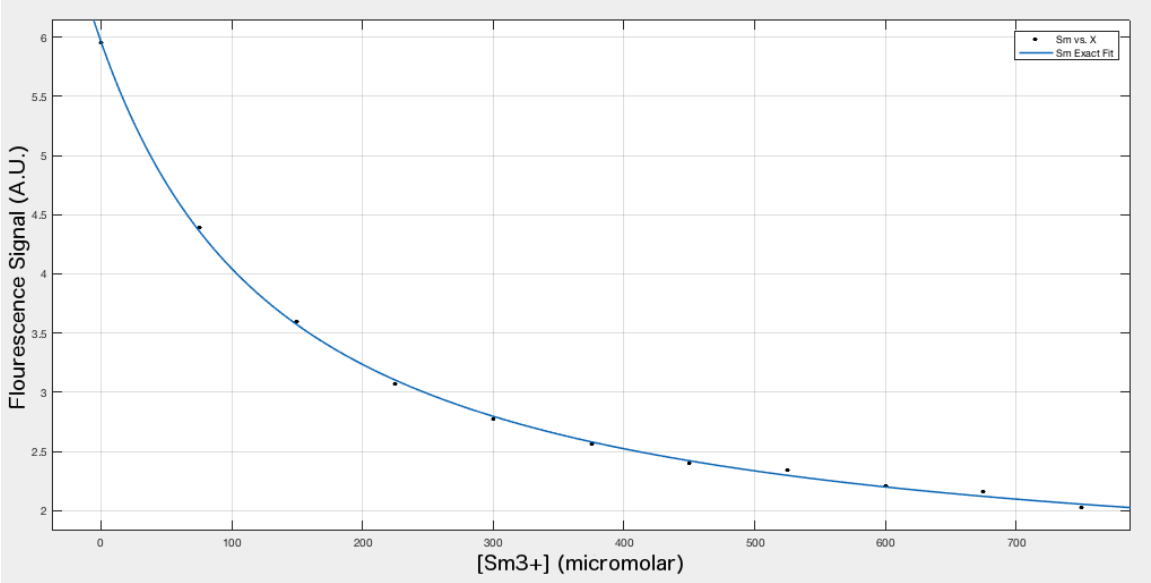

Figure 3.29: **Competitive fluorimetric titration of Sm^{3+} binding to scCC-Tb.** Binding isotherm data was collected at pH 6, and fit to a *K*d value of 10 μM.

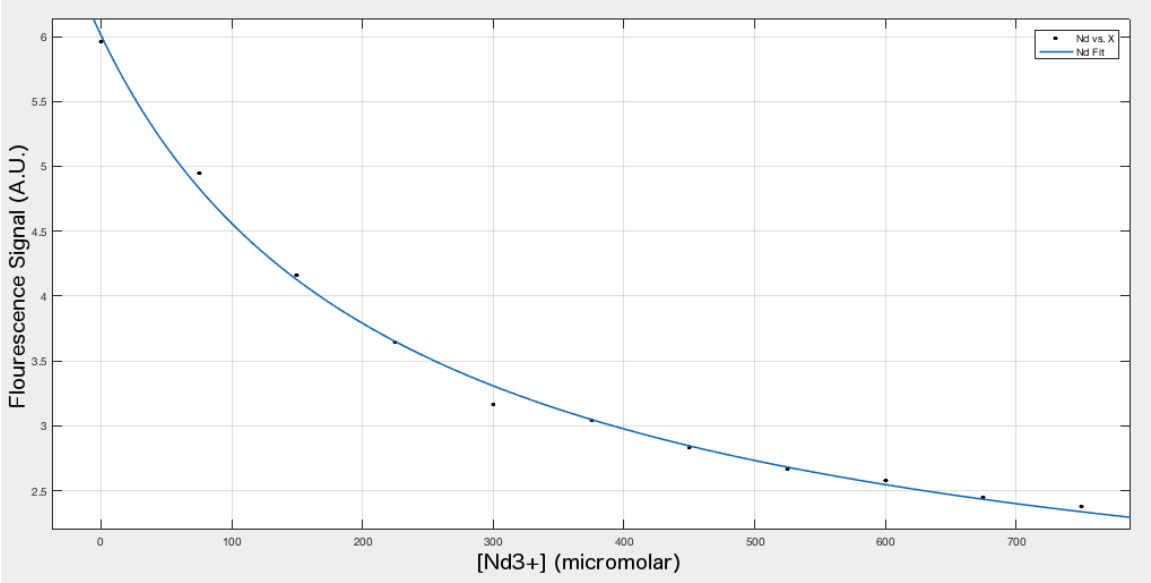

Figure 3.30: **Competitive fluorimetric titration of Nd^{3+} binding to scCC-Tb.** Binding isotherm data was collected at pH 6, and fit to a *K*d value of 18 μM.

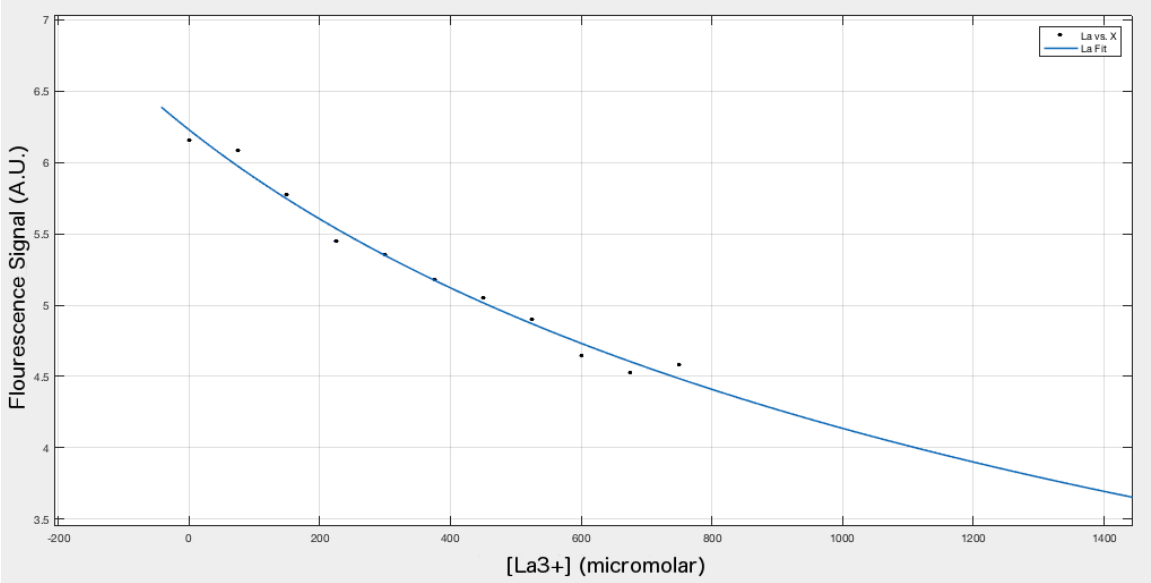

Figure 3.31: **Competitive fluorimetric titration of La^{3+} binding to scCC-Tb.** Binding isotherm data was collected at pH 6, and fit to a *K*d value of 160 μM.

Fitting of the data to the equilibrium models was carried out by nonlinear regression using the MatLab curve fitting toolbox. The values for the dissociation constants calculated for the set of lanthanide ions, along with the goodness-of-fit values, are shown in Table 3.9.

Lanthanide ion	Kd (μM)	Error with 95% confindence interval	SSE	R^2	Adjusted R^2	RMSE
Terbium	11.76	± 5.8	0.2145	0.9921	0.9909	0.1284
Gadolinium	12.88	± 3.25	0.04395	0.9966	0.9957	0.07412
Europium	9.131	± 0.546	0.003184	0.9998	0.9997	0.01995
Samarium	9.686	± 0.931	0.007616	0.9995	0.9993	0.03085
Neodymium	18.18	± 4.59	0.04052	0.9969	0.9962	0.07117
Lanthanum	159.5	± 194.7	0.05236	0.9844	0.9805	0.0809

Table 3.9: **Determination of dissociation constants for lanthanides binding to scCC-Tb.** Terbium binding isotherm was fitted to a two-state equilibrium model. Binding constants of other lanthanides were determined by fitting competitive binding isotherms to a three-state equilibrium model.

The data show a trend in dissociation constants as a function of effective ionic radius (Fig 3.30). These data indicate the designed protein exhibits selectivity for lanthanide ions of smaller radii (Tb, Gd, Eu) over the larger lanthanum ion.

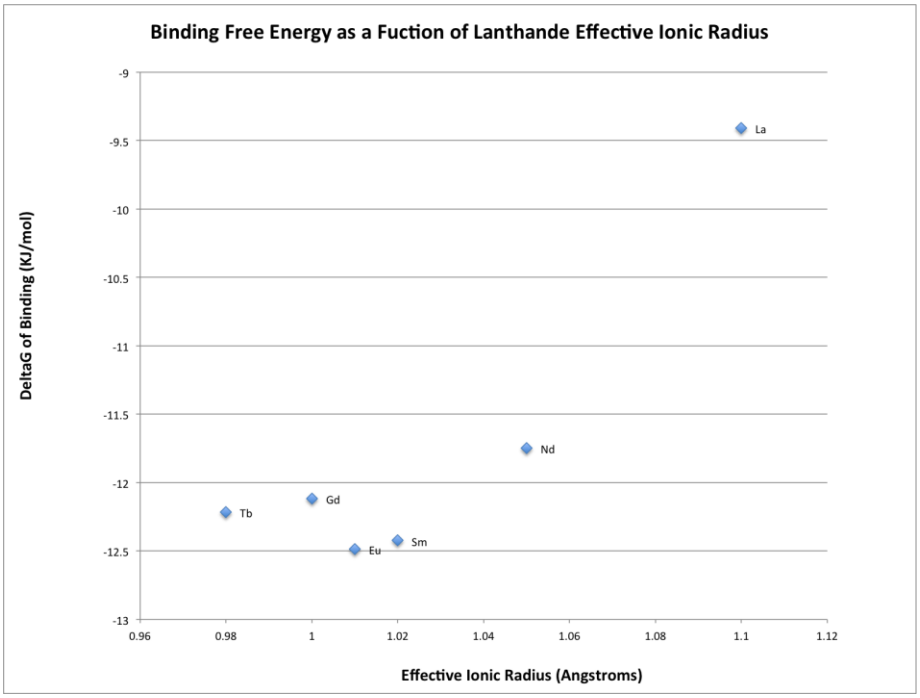

Figure 3.32: **Free energy of binding for scCC-Tb with various lanthanides as a function of effective ionic radius.** A trend in free energy is observed as a function of ion size.

3.12.6. Measurement of Rates of Dissociation

Fluorescence spectra were recorded on a Cary Eclipse fluorimeter spectrophotometer (Varian) using a 100 µL quartz cuvette of 1 cm path length (Starna Cells). Metal-bound protein samples were prepared by incubating 250 µL of 50 µM protein at 70° C for 20 minutes, and cooling in a room temperature bath for 3 minutes. Sample was centrifuged for 3 minutes at 16,100 RCF at 25° C. To each sample, 1.5 µL of 25 mM $Ln(NO_3)_3$ (Ln = Dy, Tb, Sm) was added to obtain a 3:1 ration of metal to protein. Heat cycle was repeated and samples were again centrifuged. The rates of metal dissociation were measured by adding 25 µL of 0.5 M EDTA (for a 1000:1 ration of EDTA to protein) and monitoring

the change in fluorescence signal. Fluorescence signals were collected by exciting at 280 nm, and collected at 544 nm for terbium, 573 nm for dysprosium, and 643 nm for samarium using a 360 nm high band-pass filter, with a PMT voltage of 1000 V at 20° C. In these experiments, competitive binding occurs between two ligands for the same metal. The equilibrium model is:

$$PM + L \Leftrightarrow P + L + M \Leftrightarrow LM + P \qquad (3.12)$$

where P now denotes the protein, M denotes the metal ion, and L denotes the competitive ligand (EDTA in our case). If a solution is prepared with protein and a high concentration of metal (e.g. 10 x K_d), the equilibrium in the absence of the competitive ligand will be:

$$PM \Leftrightarrow P + M \qquad (3.13)$$

and the formation of the protein-metal complex will be favored. Upon addition of an excess amount of competitive ligand, the equilibrium in equation 3.14 is shifted strongly to the right hand side, rendering the reverse reaction negligible:

$$PM + L \xrightarrow{slow} P + L + M \xrightarrow{fast} LM + P \qquad (3.14)$$

with the change in the concentration of PM given by:

$$\frac{-\,d[PM]}{dt} = k_{off}[PM] \qquad (3.15)$$

Solving the differential equation yields the rate equation as:

$$[PM] = [PM_0]e^{-k_{off}t} \qquad (3.16)$$

The change fluorescence in fluorescence is related to the decay of the protein-metal complex by equation 3.8. Data was fit to model after 180 minutes of data collection to remove non-exponential behavior caused by initial mixing period. Fitting of the data to the equilibrium models was carried out by nonlinear regression using the MatLab curve fitting toolbox.

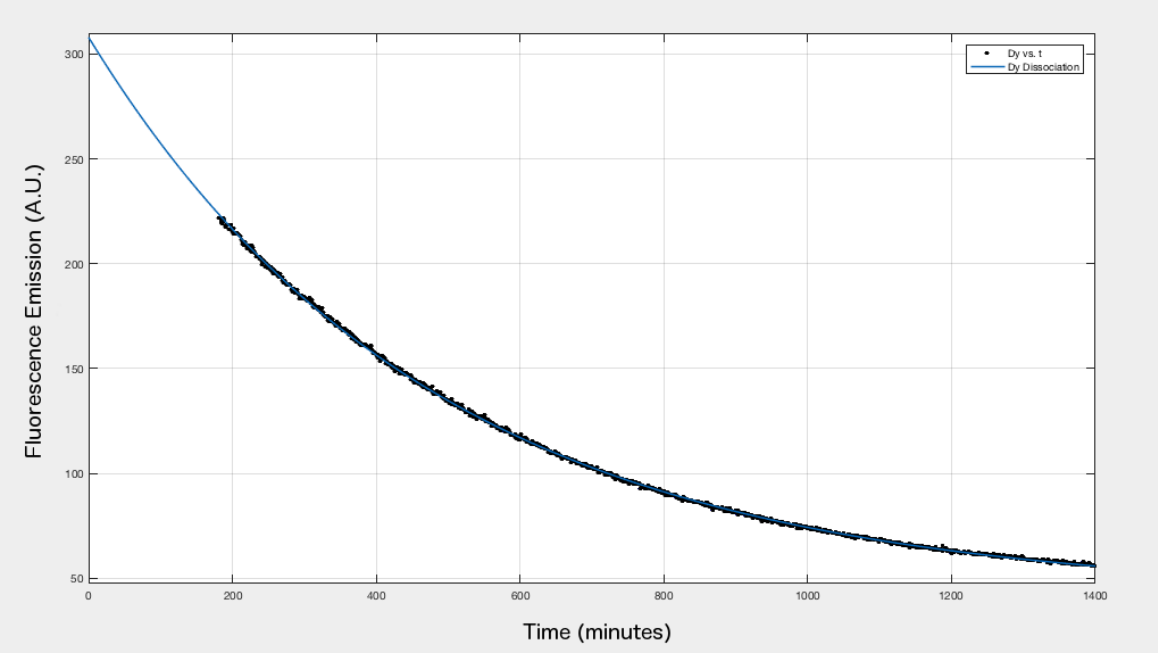

Figure 3.33: **Spontaneous dissociation of Dy^{3+} bound to scCC-Tb.** Dysprosium fluorescence emission was measure as a function of time after the addition of excess EDTA. Data was fit the exponential decay with a k_{off} of 0.002107 ± 0.000007 min^{-1} (blue line).

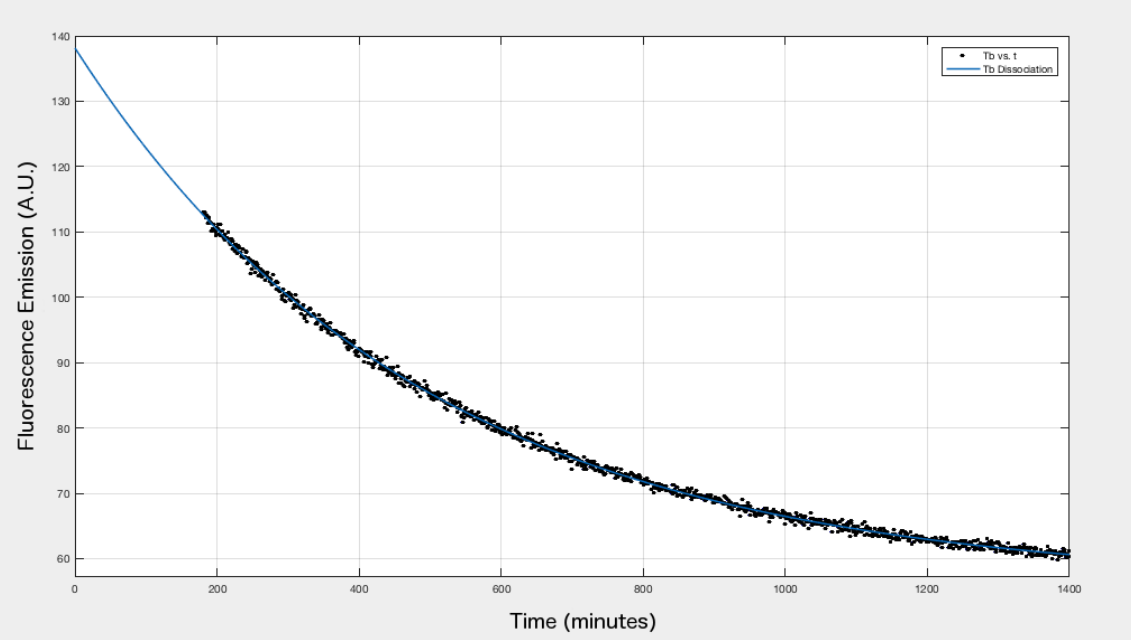

Figure 3.34: **Spontaneous dissociation of Tb^{3+} bound to scCC-Tb.** Terbium fluorescence emission was measure as a function of time after the addition of excess EDTA. Data was fit the exponential decay with a k_{off} of 0.002066 ± 0.000016 min^{-1} (blue line).

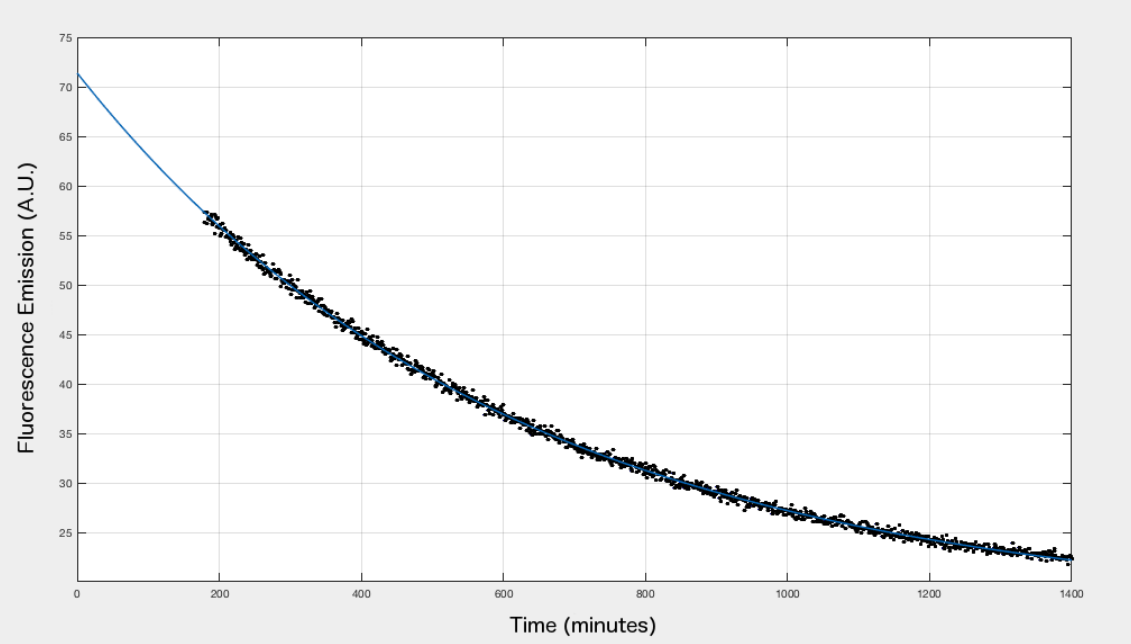

Figure 3.35: **Spontaneous dissociation of Sm^{3+} bound to scCC-Tb.** Samarium fluorescence emission was measure as a function of time after the addition of excess EDTA. Data was fit the exponential decay with a k_{off} of 0.00167 ± 0.000015 min^{-1} (blue line).

Half-lives of the decay of the metal-ligand complex were calculated by:

$$t_{1/2} = \frac{\ln(2)}{k_{off}} \tag{3.17}$$

A trend was observed for the half-lives of protein-metal complexes as a function of effective ionic radius.

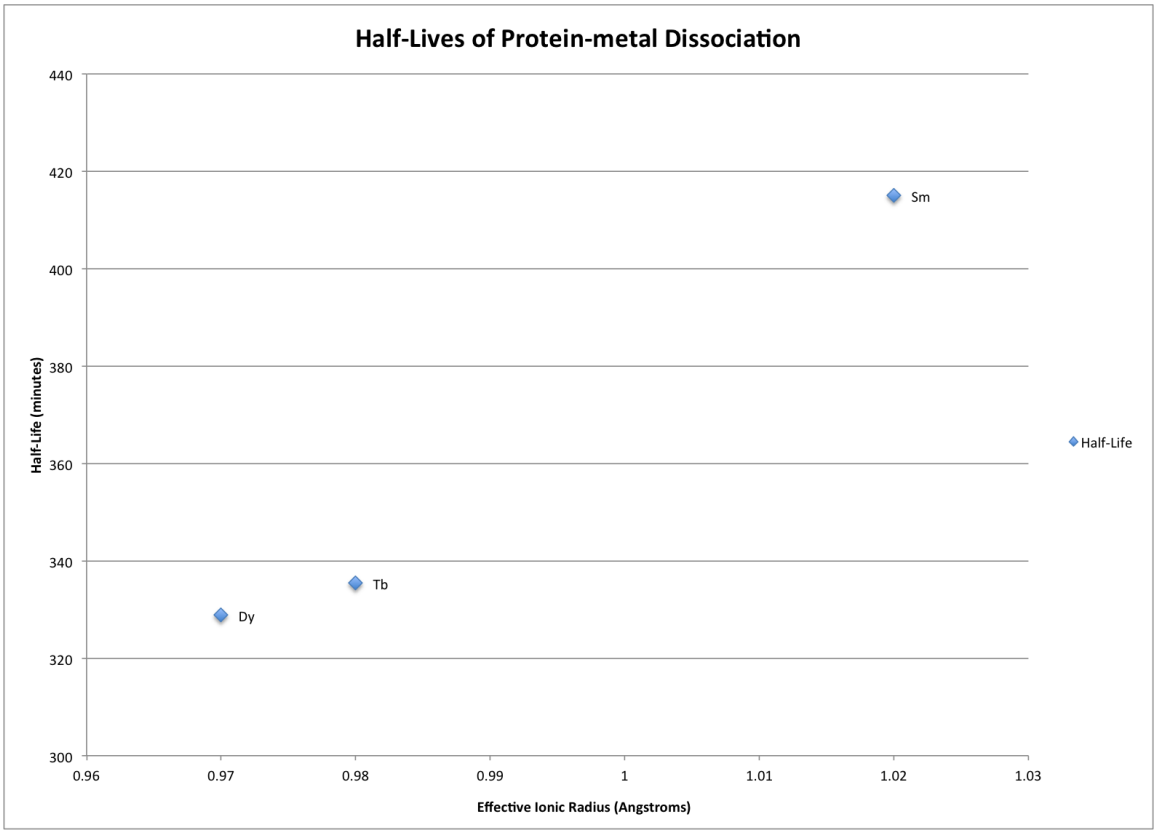

Figure 3.36: **Half-lives of dissociation for lanthanide ions bound to scCC-Tb.** Dissociation rates are dependent on effective ionic radius.

3.13. Discussion

We have used this general strategy to design a lanthanide-binding site at the core of a rigid four-helix bundle single-chain protein. Relative orientations between the helical segments of a four-helix bundle were explored to determine a configuration that would be compatible with a 6-coordiante lanthanide-binding pocket. To build a single-chain construct, the α-helical segments were connected with loop segments. To ensure that

stability of the coiled-coil motif was retained in the single-chain construct, loop segments were selected from a database of loops taken from protein X-ray crystal structures. Statistical computational protein design was used to arrive at a final sequence that would fold to the target conformation.

The resulting protein was synthesized using recombinant gene expression in *E. coli*, and recovered from the insoluble fraction. Monitoring of the 222 nm signal in CD spectroscopy indicated that the as-purified protein undergoes a structural reorganization upon heating to 70° C at pH 6.0, but is highly stable after annealing. Temperature-melt experiments to 95° C of the *apo* form showed no cooperative melting transition, and the subsequent cooling to 4° C occurred without hysteresis. Temperature-melt experiments of the *holo* did not reveal a significant difference in stability. These results are consistent with the hypothesized stability of the designed protein.

Binding of terbium to the protein was confirmed by measuring the fluorescence emission of the metal at 544 nm upon excitation of the single tryptophan residue in the sequence. It was observed that metal binding occurred in a slow manner, and heating of the solution to 70° C for 20 minutes was required to achieve full formation of the equilibrium protein-metal complex. This result is also consistent with the design of a binding in pocket in a buried hydrophobic core. The helical content and high stability of the *apo* protein and the slow kinetics of ion binding suggest that protein is fully folded in the absence of the metal ion.

The slow kinetics of ligand binding proteins with inaccessible binding sites makes it difficult to perform equilibrium experiments such as binding isotherm titrations. Several different solutions to the problem have been applied to previously designed *de novo* cofactor-binding proteins. For example, the protein DF1 was designed to bind to two iron ions, but metal binding occurred in slow manner.[136] This meant that the protein had to be unfolded and slowly refolded in the presence of the metal in order to get it to bind in an accessible time frame. A second-generation design explicitly addressed this problem by mutating a leucine residue to an alanine residue on one side of the protein to allow for the accessibility of the metal.[137] As an alternative to sequence modifications, heat can be applied to disrupt the structure and create an opening for the entrance of the ligand.[138, 139]

Folding cooperativity can also be incorporated into the design process by including predictions of folding pathways.[140] A stable core with a deep energetic well can be designed at the apolar regions of the protein core, while the polar binding site is designed by taking into account the flexibility of the backbone and choosing a sequence compatible with a low energy barrier to folding.

In our case, we were able to use heat to disrupt the protein fold enough to allow for accessibility to the binding site. High kinetic barriers to ligand binding are generally undesirable in designed proteins due to the fact that enough features can usually be designed into the binding pocket to impart ligand selectivity. In the case of lanthanide ions, the lack of differentiating chemical features makes thermodynamic selectivity difficult to achieve. Fitting of binding isotherms to fluorimetric titration data indicated that this is the case for the designed scCC-Tb protein, although a general trend in

selectivity as a function of effective ionic radius was observed.

While the difference in the free energy of binding may be small for lanthanide ions of nearly identical effective ionic radii, the difference in the energy of the transition state of unbinding may be more significant. To investigate the possibility of differentiating between lanthanide ions of nearly identical radii, we measure the spontaneous dissociation of ions from the binding site by adding excess of EDTA as a competitive binder and measuring the change in the tryptophan-to-metal FRET signal as a function of time. A protein-bound ion that spontaneously dissociates from the binding site and makes it out into bulk solution would encounter EDTA molecules with high probability and would not be able to re-bind to the protein. Fitting of the fluorescence data to a kinetic model allows for the calculation of the rates of dissociation. It was observed that the rate of metal dissociation is dependent on the ionic radius of the lanthanide ion. Such differences in dissociation could be exploited to a carry out kinetically controlled purifications of REEs.

Lanthanide Ion	Effective Ionic Radius (Å)	k_{off} (min^{-1})	Half-Life (min)
Dy^{3+}	0.97	0.002107	329.0
Tb^{3+}	0.98	0.002066	335.5
Sm^{3+}	1.02	0.00167	415.1

Table 3.10: **Observed rates of protein-metal dissociation.** Half-lives of protein-metal complex in the presence of excess EDTA increase as the effective ionic radius increases.

More importantly, the designed protein could be used as a model system for investigating the effects of structural modifications on the thermodynamics and kinetics of lanthanide

ion binding. Fine-tuning of the coordination sphere could be used to design proteins with improved selectivity towards one lanthanide element over the others. Binding pathways could be designed into the structure to fine-tune the kinetics of ion binding.[141] The atomistic control of the structural determinants of ion binding would enable the realization of lanthanide-binding protein with targeted chemical properties.

3.14. Conclusion

We have designed a four-helix bundle protein with a buried lanthanide-binding site and studied the thermodynamics and kinetics of metal ion binding. The designed protein is able to fold in the absence of the metal ion, is highly stable, and exhibits slow binding kinetics. The protein binds to lanthanide ions with micromolar affinity with a trend in thermodynamic selectivity as a function of effective ionic radius that is consistent with previously studied protein-lanthanide complexes. However, no selectivity was observed between ions in the mid-range of the lanthanide series. The slow kinetics of the system allowed for the analysis of rates of spontaneous dissociation of various lanthanides, and differences in dissociation rates were measurable for elements in the mid-range. The single-chain lanthanide-binding protein designed here will serve as a model system upon which to study the sequence-structure-function relationships of protein-lanthanide systems.

4| Ferritin- Mutational Analysis of a Naturally Occurring Self-Assembling Nanocage*

4.1. Abstract

Protein cage self-assembly enables encapsulation and sequestration of small molecules, macromolecules, and nanomaterials for many applications in bionanotechnology. Notably, wild-type thermophilic ferritin from *Archaeoglobus fulgidus* (AfFtn) exists as a stable dimer of four-helix bundle proteins at a low ionic strength, and the protein forms a hollow assembly of 24 protomers at a high ionic strength (~800 mM NaCl). This assembly process can also be initiated by highly charged gold nanoparticles (AuNPs) in solution, leading to encapsulation. These data suggest that salt solutions or charged AuNPs can shield unfavorable electrostatic interactions at AfFtn dimer−dimer interfaces, but specific "hot-spot" residues controlling assembly have not been identified. To investigate this further, we computationally designed three AfFtn mutants (E65R, D138K, and A127R) that introduce a single positive charge at sites along the dimer−dimer interface. These proteins exhibited different assembly kinetics and thermodynamics, which were ranked in order of increasing 24mer propensity: A127R < wild type < D138K ≪ E65R. E65R assembled into the 24mer across a wide range of ionic strengths (0−800 mM NaCl), and the dissociation temperature for the 24mer was 98 °C. X-ray crystal structure analysis of the E65R mutant identified a more compact,

* Adapted from Pulsipher, K.W., Villegas, J.A., et al. Thermophilic ferritin 24mer assembly and nanoparticle encapsulation modulated by interdimer electrostatic repulsion. *Biochemistry*, **2017**, *56* (28), 3596-3606.

closed-pore cage geometry. A127R and D138K mutants exhibited wild-type ability to encapsulate and stabilize 5 nm AuNPs, whereas E65R did not encapsulate AuNPs at the same high yields. This work illustrates designed protein cages with distinct assembly and encapsulation properties.

4.2. Introduction

Ferritins make up a ubiquitous family of protein cages, important for iron storage in most organisms. Maxi-ferritins share a similar overall structure of 24 tetrahelical subunits assembled to form a hollow, roughly spherical shape. The dimer is a common intermediate in ferritin self-assembly, and the 24mer/dimer distribution can be altered by single-point mutations. Orner and co-workers identified several "hot-spot" residues in *Escherichia coli* bacterioferritin and ferritin-like DNA-binding protein from starved cells (DPS) whose mutation either shut down assembly entirely[142, 143] or stabilized it.[144, 145] They accomplished this by modulating hydrophobic interactions, either by plugging water pockets with aromatic amino acids,[144, 145] by disrupting electrostatics along the dimer interface,[142] or by replacing amino acids with alanine.[146] In *E. coli* ferritin A, single-point mutations to alanine at the three-fold axes decreased 24mer cage stability.[147]

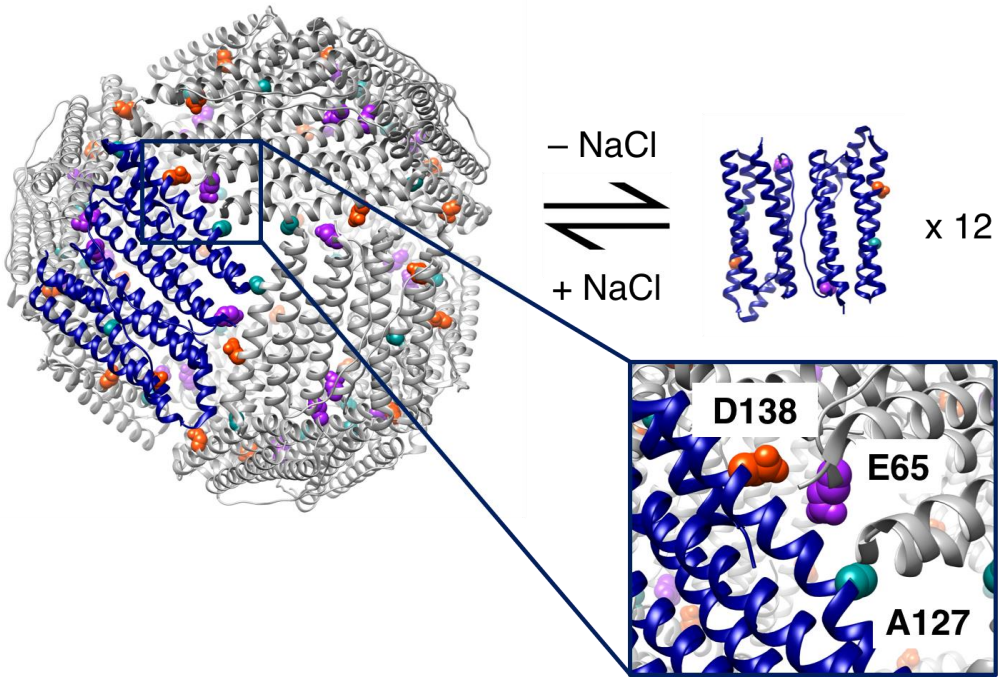

Figure 4.1: **Salt-dependent assembly for wild type AfFtn.** At high ionic strength (800 mM NaCl), the 24mer cage predominates. At low ionic strength (<200 mM NaCl), the protein disassembles into twelve dimers. The dimer is highlighted in blue in the 24mer cage on the left. Inset shows close-up of mutation positions at the trimeric interface. The crystallographic structure of AfFtn (PDB ID 1SQ3) was used to generate the figure.[19]

We investigated the self-assembly of thermophilic ferritin from the archaeon *Archaeoglobus fulgidus* (AfFtn), which has unique salt-dependent assembly not previously found in other ferritins. As shown in Fig. 4.1, at a high ionic strength (800 mM NaCl), the 24mer is the dominant assembly state. At lower ionic strengths (<200 mM NaCl), the protein disassembles into stable dimers.[21, 148] The ionic strength-dependent self-assembly of AfFtn has been used to encapsulate citrate- or bis(p-sulfonatophenyl)phenylphosphine (BSPP)-functionalized gold nanoparticles (AuNPs), rendering them more biocompatible and stable to salt-induced precipitation.[148-150] The

encapsulation process happens under mild conditions, at room temperature with gentle agitation. The resulting AfFtn−AuNP assembly maintains native protein secondary structure, subunit stoichiometry, melting temperature, and ferroxidase activity, thus highlighting unusual protein− AuNP complementarity.

The crystal structure of AfFtn [Protein Data Bank (PDB) entry 1SQ3][21] contains a trimeric interface rich in negatively charged residues. We hypothesized that electrostatic repulsion at this interface prevents subunit assembly at low ionic strengths and neutral pH, given that the estimated pI of AfFtn is 4.7. Subunit− subunit electrostatic repulsion was found to be important in governing the rate of self-assembly of *E. coli* ferritin A, which typically exists only as the 24mer except under acidic conditions.[147] As other ferritins do not feature salt-mediated self-assembly, AfFtn provides a unique opportunity to investigate the role of subunit interface electrostatics in protein cage formation. We hypothesized that decreasing the extent of electrostatic repulsion between anionic subunits by a designed amino acid substitution should promote 24mer formation at low salt concentrations. We tested this hypothesis by introducing positively charged groups at various positions along the dimer−dimer interface. As has been employed previously in the redesign of ferritin proteins, a statistical computational design strategy was used to calculate theoretical amino acid probabilities at selected sites.[151, 152] The resulting probabilities were used to guide the selection of point mutations likely to be compatible with the overall protein structure as well as the supramolecular assembly. Three single-point mutants shown in Fig. 4.2 were experimentally characterized, where each mutation replaced a single negatively charged or neutral one with a positively charged residue. The thermal stability of the mutants was investigated, and changes in the self-assembly

equilibrium, kinetics, and reversibility at different salt concentrations were explored. The capability of each mutant to encapsulate and stabilize AuNPs was also investigated. We conclude that altering electrostatic interactions between ferritin subunits provides a versatile approach to modulating protein cage assembly.

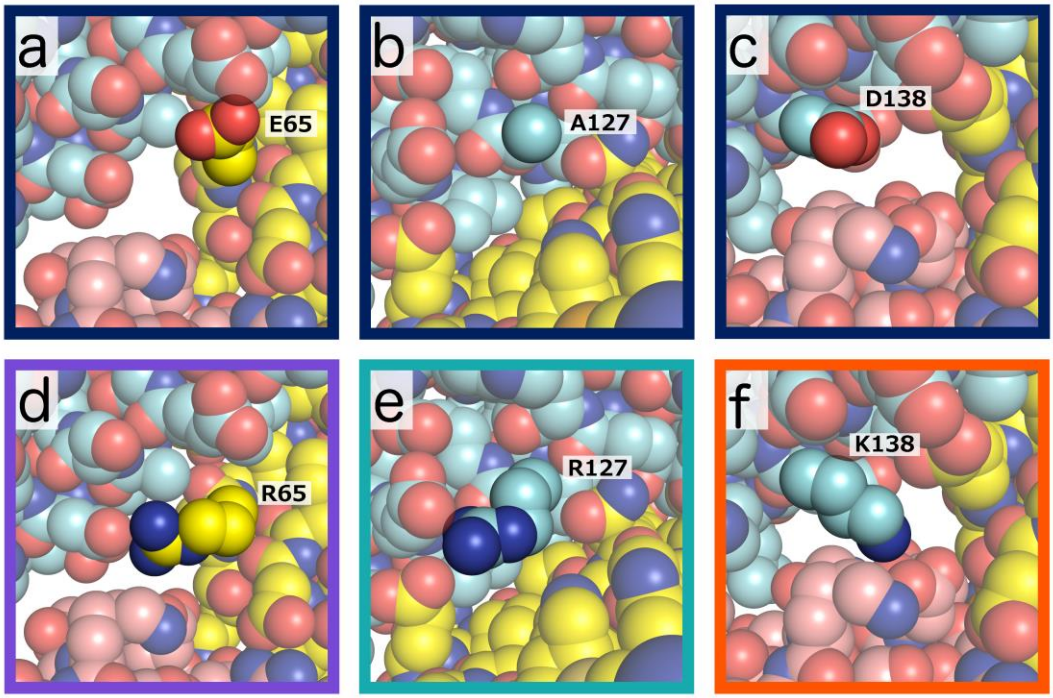

Figure 4.2: **Computationally designed mutations along the trimeric interface.** Wild-type residues (dark blue border): (a) E65, (b) A127, and (c) D138. Single-point mutations: (d) A127R (violet border), (e) D138K (teal border), and (f) E65R (orange border). Different protomers (chains) are rendered distinct colors: cyan, yellow, and pink.

4.3. Computational Design of AfFtn Mutants.

The template structure consisted of a trimer of subunits comprising chains G, H, and J from the crystallographic structure of *A. fulgidus* thermophilic ferritin 24mer (PDB entry 1SQ3[21]). Amino acid probabilities were calculated independently for sites 34, 65, 127,

131, and 138. Sites 127, 131, and 138 are situated along a helical interfacial region. Sites 34 and 65 were chosen as they form, along with site 138, the center of a carboxylate-rich pore. Eighteen natural amino acids were considered at each selected site; cysteine and proline were precluded. For each mutation calculation, residues other than the site of interest were constrained to the crystal structure conformations. Energies[58] and conformational states of mutated sites were obtained using a rotamer library.[153] The entropy-based, probabilistic formalism was used to calculate the probabilities of the amino acids and their rotamer conformations. Using CHARMM19,[58] hydrogen atoms were added, and energies were calculated using the dihedral, van der Waals, and electrostatic terms, with a nonbonded cutoff of 8 Å. β was set to 0.5 mol/kcal for these calculations.[7, 138, 139, 154] The probability of each amino acid at a mutated site was the sum of the calculated probabilities of its side-chain conformations.

In the design calculations, we sought to identify mutations that introduce a positively charged residue at the interface between adjacent protomers. We focused on the probabilities of amino acid type (a) at each site *i*, $P_i(a)$, or the ratio relative to the wild-type residue, $P_i(a)/P_i(a_{wt})$. At site 34, the wild-type Asp was the most probable residue, and as a result, this site was not selected for mutation. At site 65, Arg was the most probable amino acid [$P_{65}(R) = 0.83$], yielding the suggested mutation E65R. At site 127, Arg was the most probable amino acid [$P_{127}(R) = 0.63$], and mutation A127R was selected. At site 131, Lys was the most probable amino acid [$P_{131}(K) = 0.93$]. Upon examination of the resulting model structures, the conformations of this side chain directed the ammonium group within the protomer, and the side-chain interactions did

not span an interface with other protomers; this site was not selected for mutation. At site 138, the four most probable amino acids were Arg, Asn, Asp, and Lys, and $P_{138}(\mathrm{K})/P_{138}(\mathrm{D}) = 1.49$. Mutation D138K was selected at this site.

4.4. Experimental Verification

Experiments were carried out by members of the laboratory of Professor Ivan Dmochowski. E65R, A127R, and D138K single-point mutants were prepared using standard site-directed mutagenesis. After expression and purification, protein purity was verified by SDS−PAGE and UV−vis spectroscopy and identity was verified by MALDI-TOF MS. Cage formation in 800 mM NaCl was verified by transmission electron microscopy (TEM). Shown in Fig. 4.3 and summarized in Table 4.1, cages with approximately the same diameter as the wild type (wt) were observed in TEM micrographs, showing the mutations did not inhibit self-assembly in a high-ionic strength solution. Dynamic light scattering (DLS) also showed similar results, with all mutants having average particle diameters within 2 nm of that of the wt [D_{avg} (wt) = 13.5 nm]. Circular dichroism (CD) spectra for all three mutants showed almost no change compared to the wt, demonstrating no perturbation in secondary structure. This is unsurprising, as ferritins and ferritin-like proteins have been shown to be stable with respect to extensive mutagenesis.[152]

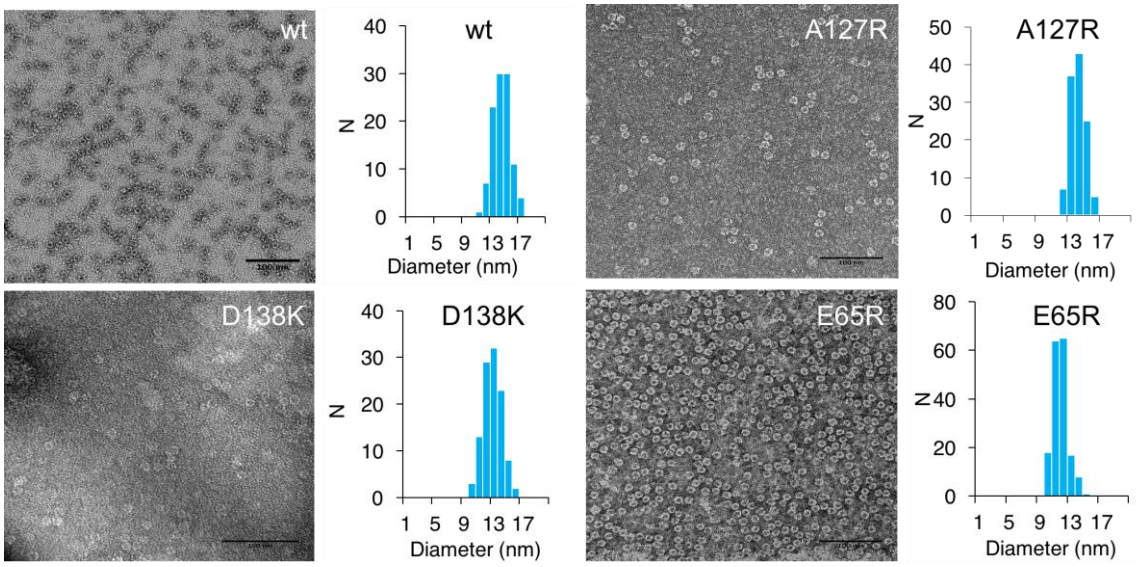

Figure 4.3: **TEM micrographs and size distributions for wt and mutant AfFtn.** Similar cage structures were observed for all samples, indicating mutations did not prevent self-assembly. Grids were stained with either 2% uranyl acetate or 2% ammonium molybdate negative stain. Particle size was measured manually using ImageJ.[155] Scale bars are 100 nm.

Sample	D_{avg} TEM (nm)	N	D_{avg} DLS (nm) [PDI]	T_m (°C)
wt	13.2 ± 1.1	101	13.5 [0.035]	84
A127R	13.3 ± 0.9	117	13.6 [0.052]	85
D138K	12.3 ± 1.3	110	14.9 [0.103]	84
E65R	11.1 ± 0.9	173	12.9 [0.059]	98

Table 4.1: **Assembled cage characterization in 800 mM NaCl**. D_{avg} is average diameter. N is number of particles that were measured manually using ImageJ to calculate average diameter. T_m was measured by CD for wt, A127R, D138K and by DSC for E65R.

The thermal stability of the mutants was investigated using CD spectroscopy and calorimetry. At 0.3 mg/mL, the temperature-dependent molar ellipticities of A127R and D138K yielded a thermal stability nearly identical to that of wt AfFtn [T_m = 84 ° C (Table 4.1)]. However, E65R did not unfold at <96 ° C, and its T_m by could not be determined by CD. Differential scanning calorimetry (DSC), which uses a pressurized, sealed sample chamber, was used to measure higher melting temperatures under aqueous conditions. By DSC at 0.5 mg/mL, a T_m of 98° C for E65R was measured; using the same technique, a T_m of 84° C for the wt was measured, which confirmed the CD-determined values. Remarkably, the thermal stability of E65R is enhanced by 14° C compared to those of the hyperthermophilic wt protein and the other two mutants. Although CD monitors changes in protein secondary structure and does not necessarily yield information about the assembly state, the single sharp transitions observed in the CD and DSC data suggest that disassembly of the 24mer may be concomitant with unfolding of the subunits under these conditions.

The enzymatic activity of the mutants was investigated using an absorbance-based ferroxidase assay.[156] An aliquot with 480 equiv of Fe^{2+} was added to each sample, and oxidation to Fe^{3+} was monitored by the increase in iron mineral absorbance at 315 nm. E65R had slightly enhanced activity compared to that of the wt, while D138K had slightly diminished activity and A127R significantly diminished activity. A127R is closest to the ferroxidase site of the protein and likely interacts with glutamates at positions 128 and 131, potentially perturbing their activity. In light of minimally affected structural features and thermal stability mentioned above, this represents a localized

disruption of the ferroxidase site compared to wt AfFtn.

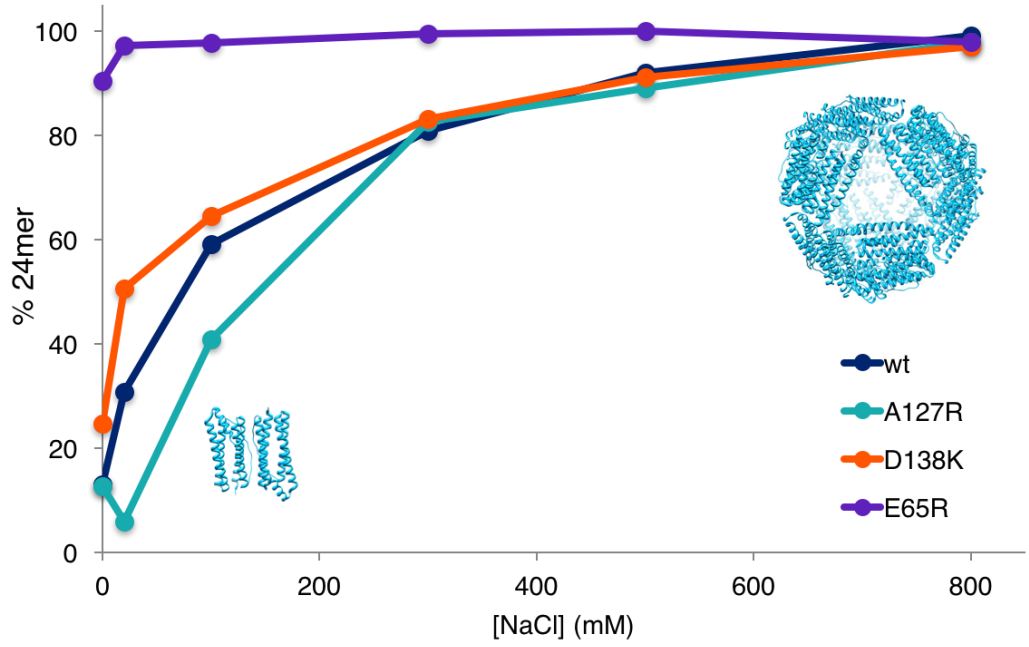

Figure 4.4: **Size exclusion chromatography to quantify the amount of 24mer present at various salt concentrations for all proteins.** Compared to wt (dark blue) at low-salt concentrations, D138K has a slightly larger percentage of the fully formed assembly (orange), and A127R has a slightly larger percentage of the dimer (teal). E65R has greater than 90% assembly at all salt concentrations tested, [NaCl] = 0-800 mM (violet).

Size exclusion chromatography (SEC), tryptophan fluorescence, DLS, and native gel electrophoresis were used to investigate the assembly state of the proteins at varying salt concentrations. SEC was performed after incubation of each protein overnight at 5 mg/mL in solutions of different salt concentrations, as seen in Fig. 4.4. The peak in the SEC trace at ~ 10 mL was attributed to the 24mer, while the peak at ~14 mL was attributed to the dimer based on column MW calibration. The area under the peaks was

used to quantify the percentage of 24mer for each protein at each salt concentration. As expected, under high-salt conditions (800 mM NaCl), all proteins showed nearly 100% 24mer. Under lower-salt conditions (<200 mM NaCl), different behaviors were observed. A127R is slightly less likely to self-assemble under low-salt conditions than the wt is, while D138K has a slightly higher propensity to assemble, with 24mer populations larger than those of the wt at 0, 20, and 100 mM NaCl. E65R shows >90% 24mer under all salt conditions tested, demonstrating a dramatic change in self-assembly equilibrium. To corroborate SEC results, Trp fluorescence was used, which is reflective of the solvation environment of Trp residues in the protein (Fig. 4.5a).

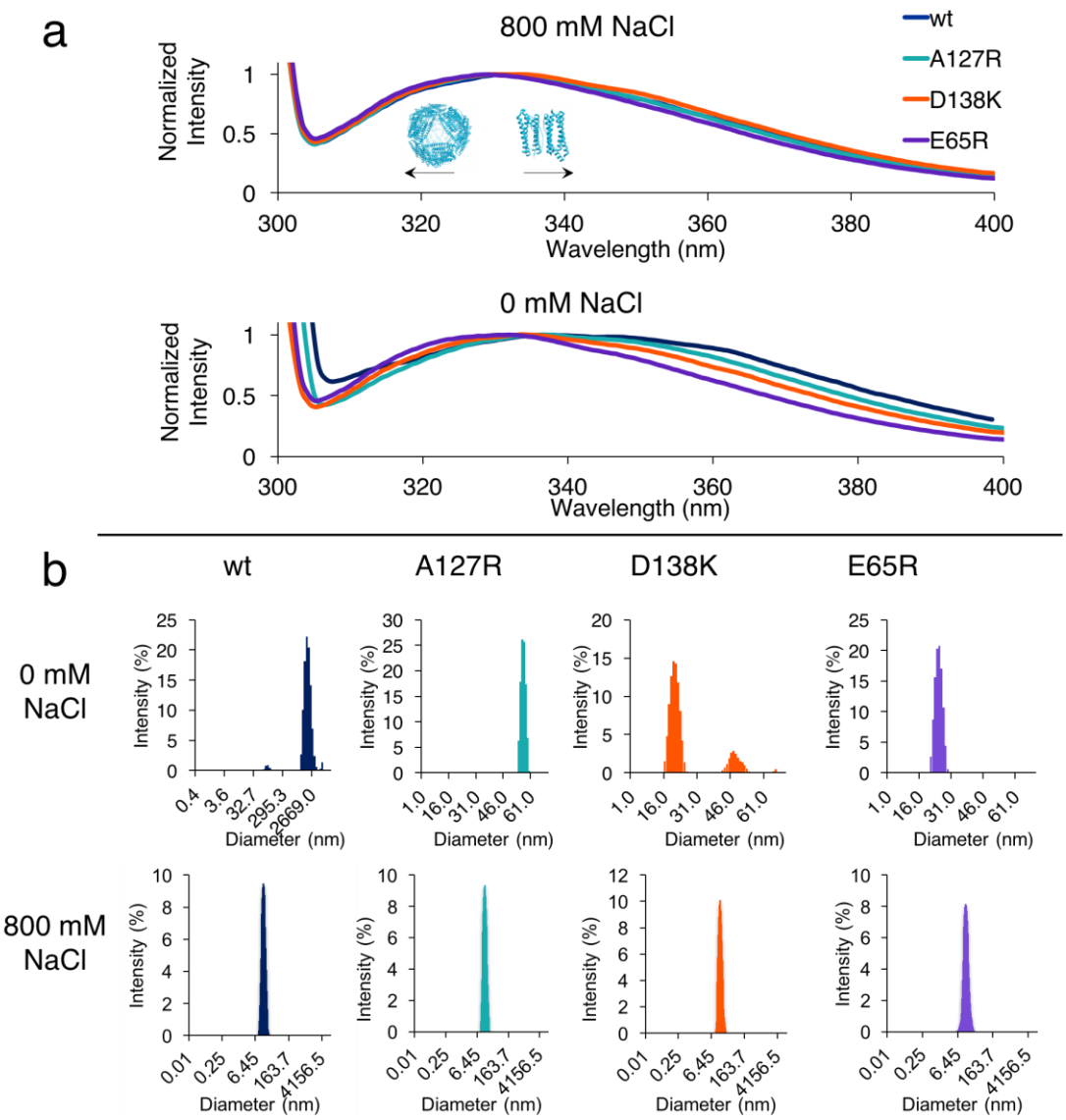

Figure 4.5: **Assembly properties of AfFtn variants. (**a) Tryptophan fluorescence results for wt and mutants. (b) Dynamic light scattering results. All proteins in 800 mM NaCl show complete 24mer assembly. At 0 mM NaCl, only E65R remains assembled, while D138K forms discrete dimers, 24mer, and some aggregate, and wt and A127R predominantly form aggregates of dimers.

AfFtn contains four Trp residues per single-chain subunit, two of which are predicted to see some change in solvation upon disassembly (Trp44 and Trp124). For the wt, in 800 mM NaCl buffer where AfFtn is completely assembled, the Trp emission maximum is at 332 nm. In 0 mM NaCl buffer, where AfFtn is disassembled into dimers, the emission red shifts to 337 nm, consistent with the Trp residues becoming more solvent exposed. The Trp fluorescence spectrum was measure for each protein in 0 and 800 mM NaCl buffer. All of the mutants exhibited fluorescence spectra similar to that of the wt at 800 mM NaCl, with similar peak shapes and maximal emission wavelengths. E65R showed a slight blue shift compared to the wt, with an emission maximum of 331 nm compared to 332 nm for wt. At 0 mM NaCl, significant changes were observed. A127R and wt had the largest red shifts, moving to 337 and 338 nm, respectively, indicating disassembly for both proteins. D138K had a red shift of only 2 nm, while E65R shifted by only 1 nm. These trends for D138K and E65R match those observed by SEC, with E65R showing minimal structural changes with a change in salt concentration and D138K showing changes smaller than those of the wt.

DLS corroborated Trp fluorescence results showing A127R with self-assembly behavior similar to that of the wt, with a decreased 24mer population under low-salt conditions (Fig. 4.5b). Like the wt, A127R 24mer was present at 800 mM NaCl and not at 0 mM NaCl. DLS also indicated that at 5 mg/mL protein and 0 mM NaCl, A127R, wt, and D138K (to a smaller extent) form a broad range of higher-molecular weight aggregates. These aggregates are too weakly associated to withstand FPLC treatment as no aggregates were observed on the sizing column under any conditions. Aggregate formation is concentration-dependent, as wt samples at 1 mg/mL had individual dimer

visible in DLS data; however, at 5 mg/mL, only aggregate was present. D138K was predominantly 24mer at 800 mM NaCl and mostly dimer (with some 24mer) at 0 mM NaCl, whereas E65R remained fully assembled at both salt concentrations. Native gel results also support the uniqueness of the E65R mutant, where E65R was the only protein to run like horse spleen apoferritin (HSAF, used as a control 24mer because of its lack of salt-mediated disassembly). The wt, D138K, and A127R all ran as smaller species.

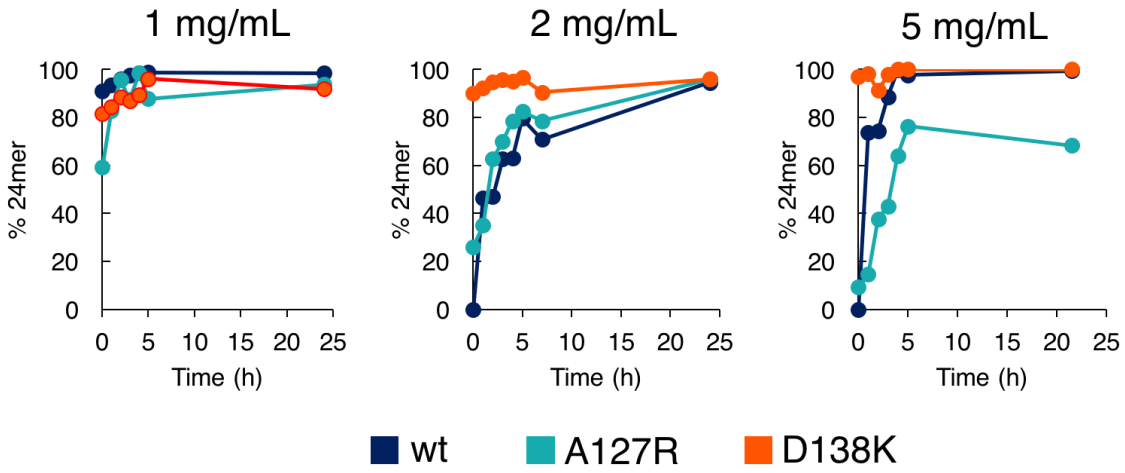

Figure 4.6: **Kinetics of assembly of AfFtn variants.** DLS was used to monitor assembly of wt and mutants, starting from dimers. Assembly rate was concentration-dependent for A127R, with faster assembly at lower protein concentrations. WT showed fastest assembly at 1 mg/mL, followed by 5 mg/mL and 2 mg/mL. D138K assembled within the time it took to take the measurement for all protein concentrations tested.

Differences in the kinetics of 24mer assembly for the wt, A127R, and D138K were investigated by monitoring particle size using DLS. Because E65R does not disassemble with a decrease in ionic strength, no changes were observed by DLS or SEC with a change in the buffer conditions. The disassembled protein in 0 mM NaCl buffer was transferred to 800 mM NaCl buffer to induce self-assembly. The rate of assembly was concentration-dependent, particularly for A127R and wt (Fig. 4.6). For wt protein at 5 mg/mL, assembly appeared to be complete within 4 h. For A127R at 5 mg/mL, the assembled 24mer population increased to ≤ 75% over 4 h and then stalled with no further assembly. At 2 mg/mL, both wt and A127R took approximately 24 h to reassemble. At 1 mg/mL, wt assembly occurred within 10 min (the time it took to prepare samples and take the DLS measurement), and A127R was >90% 24mer within 2 h. Assembly was significantly faster for D138K than for the wt and A127R at 2 and 5 mg/mL. At these concentrations, the D138K samples were >90% 24mer within 10 min. At 1 mg/mL, assembly was still quite fast, with a population of >80% 24mer within 10 min. Rapid assembly kinetics with D138K at all concentrations tested is consistent with a lack of aggregate formation, and an orderly process of subunit assembly. This is in contrast to the case for the wt and A127R, both of which started from large aggregates, particularly at 5 mg/mL. This feature should lead to more reversible and higher-yielding ferritin disassembly and assembly processes with D138K, e.g., as required for cargo loading.

A less dramatic difference among the protein disassembly kinetics was observed when moving samples from 800 to 0 mM NaCl. By DLS, A127R and the wt appeared to disassemble immediately, producing large aggregates. The disassembly of D138K was more difficult to monitor by DLS because of the relative similarity in diameters of the

24mer and dimer (and lack of aggregates). Disassembly was observed by SEC for all three proteins within 25 min. Some 24mer remained for D138K within this time frame, which matched the results for the 0 mM NaCl equilibrium measurement (Fig. 4.4).

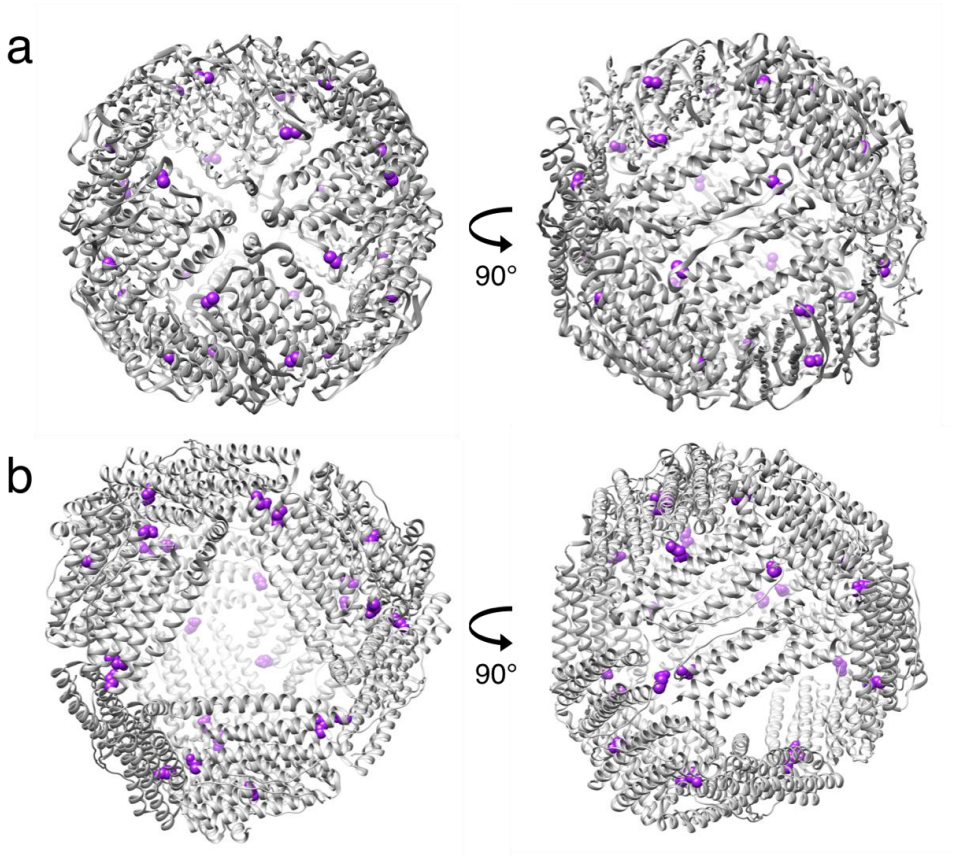

Figure 4.7: **Crystal structure of E65R assembly.** E65R exists exclusively in its 24-mer state in a closed-pore assembly. (a) Cartoon of E65R crystal structure (PDB 5V5K) with residue 65 highlighted in purple. (b) Open-pore wt AfFtn (PDB 1SQ3) with residue 65 highlighted in purple.

Additionally, a 3.08 Å resolution X-ray crystal structure of E65R was obtained (PDB entry 5V5K). Although the quality of the electron density maps does not allow side chain

conformations to be confidently modeled, the crystallography data do offer insights pertaining to the global structure of the E65R mutant. As shown in Fig 4.7, the structure of 24mer E65R shows a shift in the symmetry of the assembled cage to octahedral (as opposed to the tetrahedral structure of the wt), resulting in a lack of the large triangular pores. The 544,000 $Å^3$ volume calculated from the structure of the E65R cage using the Voss Volume Voxelator program[157] is roughly 10% smaller than the 600,000 $Å^3$ volume calculated for a poly-Ala version of the wt (PDB entry 1SQ3),[21] a finding also reflected in the TEM and DLS results (Table 4.1). These volumes correspond to outer diameters of 10.1 and 10.5 nm for E65R and the wt, respectively.

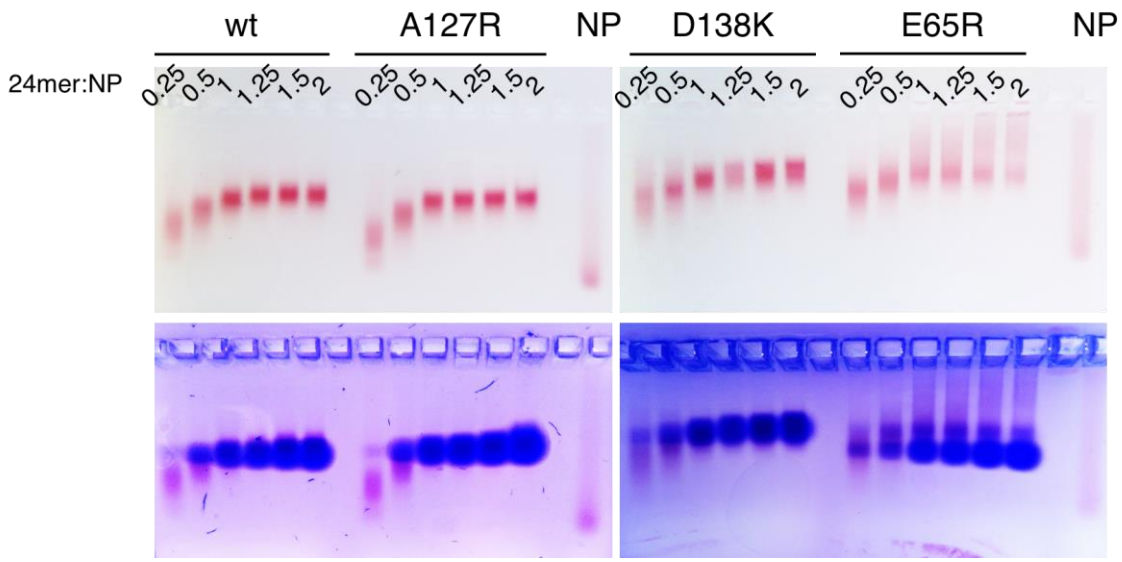

Figure 4.8: **Native gel electrophoresis showing AuNP association.** WT, A127R, and D138K at ratios of 1:1 AfFtn 24mer:AuNP, but not E65R, which shows lower propensity for disassembly.

Finally, the interaction between the assemblies of AfFtn variants with 5 nm AuNPs coated in BSPP was investigated. It was shown previously that BSPP-coated, 5 nm AuNPs are more stably encapsulated wt AfFtn than within citrate-coated AuNPs. To encapsulate the NPs, the proteins were first incubated at 0 mM NaCl overnight at 4° C to allow for maximal disassembly. AuNPs were added to the sample and for 48 h at room temperature with gentle agitation. After 48 h, the presence of AuNP did not disrupt the secondary structure of any of the proteins upon their incubation at a 1:1 AfFtn 24mer:AuNP ratio. However, by native agarose gel electrophoresis, some differences were observed among the samples. By 48 h, the wt, A127R, and D138K appeared to successfully encapsulate AuNPs as judged by cleanly overlapping blue protein and red AuNP bands, while the AuNP bands in the E65R-containing sample remained diffuse. As shown in Fig. 4.8, successful encapsulation was observed for the wt, A127R, and D138K at a 1:1 AfFtn 24mer:AuNP ratio. In contrast, the AuNP bands for the E65R-containing samples were significantly more diffuse, indicating greater variety in particle charge:mass ratio. Two bands are visible in the Coomassie-stained image, with the less intense band overlapping with the darkest part of the AuNP bands. This suggests that although some AuNPs may be encapsulated within the E65R cavity, many are not, likely because of less disassembly of the protein cage.

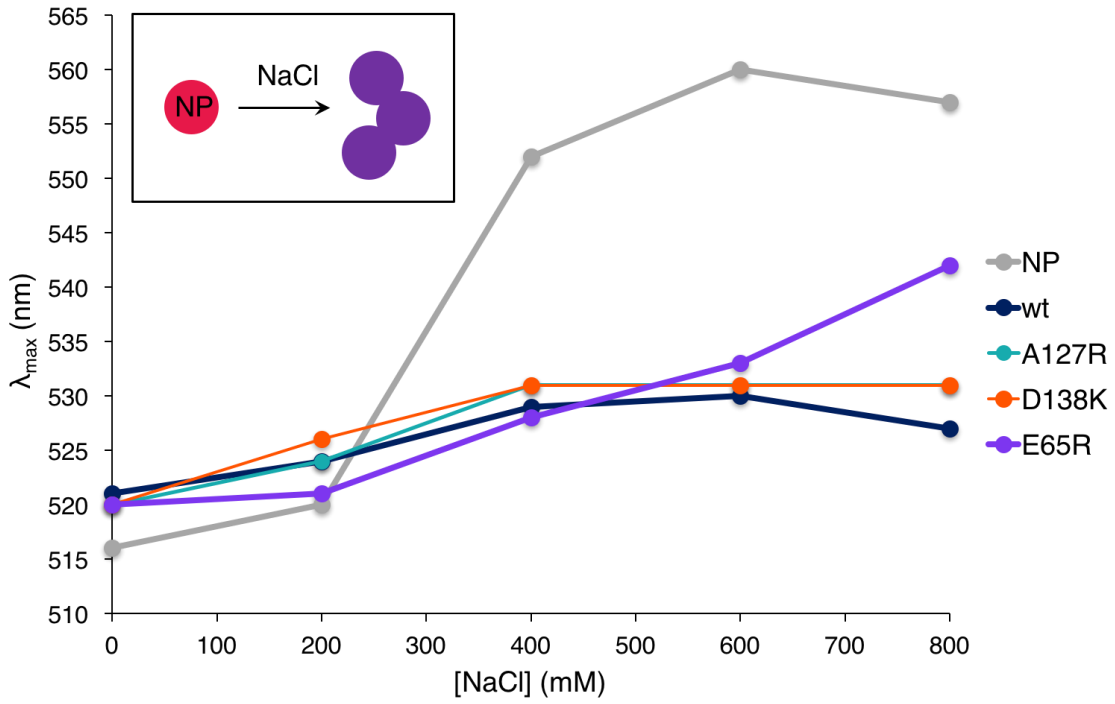

Figure 4.9: **Determination of nanoparticle passivation by AfFtn variants.** Changes in SPR peak maximum with respect to salt concentration show higher stability for AuNPs that appear associated with proteins by gel.

The ability of the protein to stabilize the AuNPs against salt-induced aggregation was also investigated. With an increase in ionic strength, electrostatically stabilized AuNPs begin to aggregate, causing a red shift in the surface plasmon resonance (SPR) peak.[158] By monitoring the SPR peak with an increasing concentration of NaCl, passivation of the AuNP surface by proteins could be observed (Fig 4.9). Bare AuNPs had the largest SPR red shift of >40 nm, from 0 to 800 mM NaCl. E65R−AuNP had the next largest shift of approximately 20 nm, while wt−AuNP, A127R− AuNP, and D138K−AuNP all had similarly small red shifts of <10 nm. This suggests that E65R does not passivate the surface of the AuNP, in agreement with the native gel results. As a 24mer, E65R likely

interacts with the AuNP surface but provides a stabilizing effect smaller than those of the wt, A127R, and D138K. By TEM several 24mer cages can be seen in contact with the AuNP surface, supporting this hypothesis.

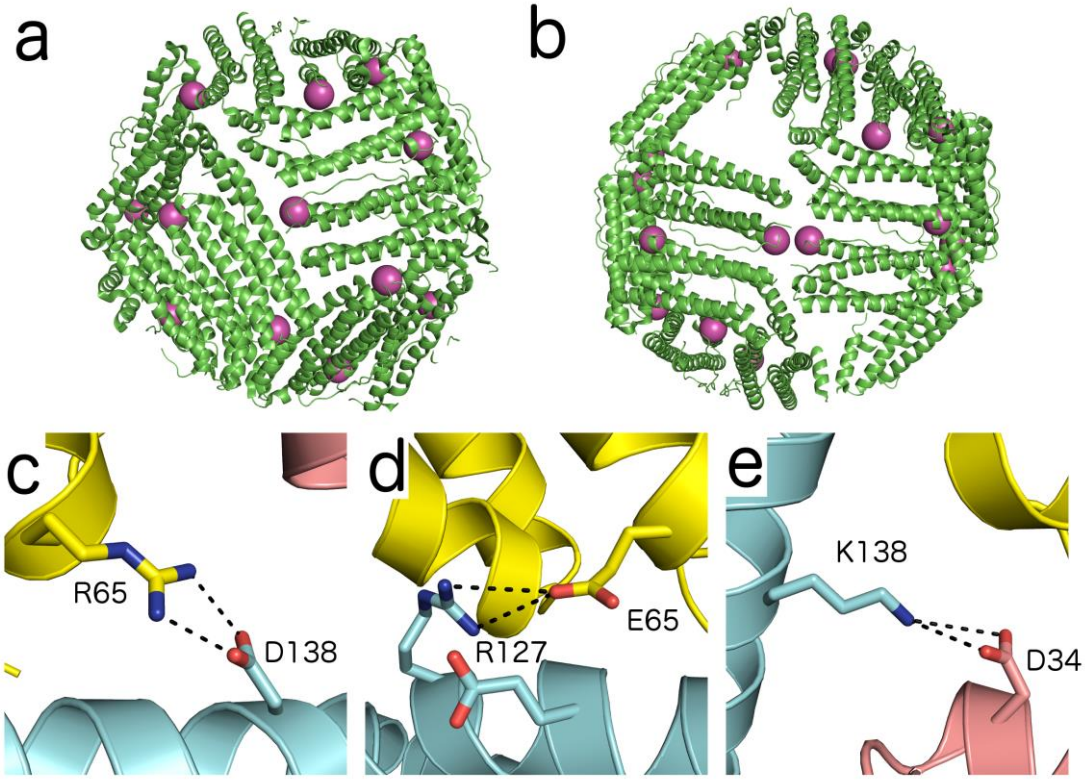

Figure 4.10: **Inter-protomer interactions in mutants of AfFtn**. Crystallographic structures of AfFtn with site 65 highlighted as purple sphere (a, b). (a) Crystallographic structure for E65R reported herein (closed pore, octahedral structure, PDB 5V5K). (b) Structure of wt AfFtn (open pore, tetrahedral structure, PDB 1SQ3). Two R65 are in close proximity at one interface. (c-e) Computationally modeled structures of mutants with most probable conformations of mutated side chains. Distinct protomers (chains) have different colors: cyan, yellow, and pink. (c) Within E65R, potential R65-D138 salt bridge. (d) Within A127R, a potential R127-E65 salt bridge within a sterically crowded local environment. (e) Within D138K, a potential K139-D34 salt bridge.

4.5. Discussion

We have designed three novel AfFtn mutants, each replacing a negative or neutral residue with one that is positively charged. None of the mutations decreased the thermal stability or hampered the ability of the protein to self-assemble into a nanocage at high ionic strengths. However, the experimental results demonstrate the profound effect a single-point mutation can have on the self-assembly of AfFtn. This is in keeping with recent literature, where *E. coli* bacterioferritin,[143] ferritin-like DNA binding protein from starved cells (Dps),[142] and bullfrog ferritin[159] were also shown to be susceptible to changes in self-assembly through minor mutagenesis. While the AfFtn mutations in which a negative residue is changed to a positive one (D138K and E65R) showed increased 24mer populations in low-ionic strength solutions, changing a neutral residue to a positive one (A127R) showed slight destabilization of the 24mer. Even between D138K and E65R there were significant differences in the favorability of 24mer assembly, with E65R remaining >90% 24mer at all salt concentrations tested and D138K disassembling under low-salt conditions. The specific interdimer location of the point mutation greatly affects self-assembly.

It is notable that E65R shows enhanced thermal stability in addition to the formation of stable 24mer assemblies at low ionic strengths. Increased thermal stability can often go hand in hand with enhanced cage stability. For example, when the self-assembly equilibrium of *E. coli* bacterioferritin[144] was shifted from a mixture of dimer and 24mer to 100% 24mer, the T_m increased by >20 ° C. Similarly, destabilization in favor of dimers has led to decreased thermal stability in mycobacterial ferritin,[160] *E. coli* ferritin A,[161]

Dps,[142] and *E. coli* bacterioferritin.[143] For wt AfFtn, the T_m under low-salt conditions was found to be essentially the same as that under high-salt conditions, indicating that the stability of the dimer is very similar to that of the assembled cage and that 24mer assembly does not increase the stability of the dimer.[149] However, in *E. coli* bacterioferritin, mutations designed to plug an interdimer water pocket with hydrophobic residues led to significantly enhanced thermal stability (ΔT_m > 20 ° C) but greater dimer population compared to that of the wt, as the geometry of the more stable dimers prevented cage formation.[145] A127R appears to favor dimer under low-salt conditions, and its thermal stability is identical to that of the wt. For D138K, its enhanced 24mer stability under low-salt conditions also does not appear to be linked to thermal stability, as it too exhibits the wt T_m. The stabilities of the individual protein subunits, their oligomeric assemblies, and the 24mer assembly are coupled and can be difficult to resolve.

At low ionic strengths, all the results support enhanced cage stability for E65R and D138K. The symmetry shift seen in the crystallographic structure of E65R is striking (Fig. 10a), but such closed-form (octahedral) structures need not exhibit enhanced cage stabilities. A double mutant of AfFtn, K150A/R151A, was previously shown to yield octahedral cage symmetry; however, this mutant maintained the salt-dependent disassembly and reassembly behavior of the wt protein.[162] The crystallographic structures of E65R and K150A/R151A are highly similar, with an α-carbon coordinate root-mean-square deviation (RMSD, as calculated using VMD[163]) of 2.19 Å for the assembled 24mer cage (see an alignment of structures in Fig. 11).

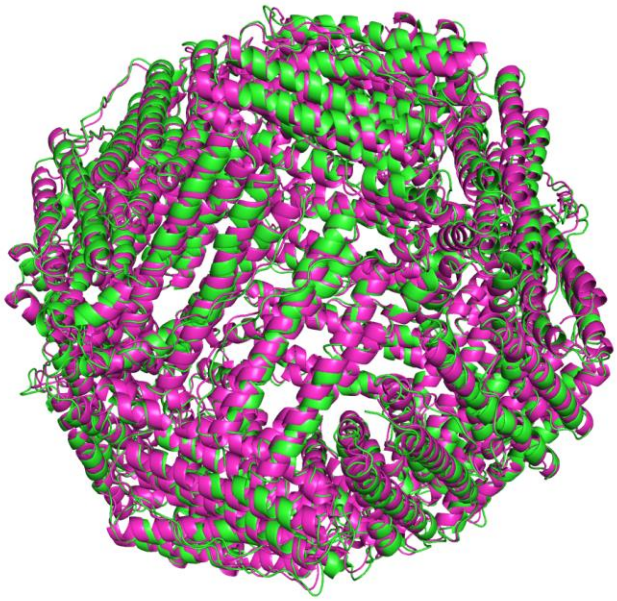

Figure 4.11: **Structural alignment of E65R and K150A/R151A.** Alignment was performed using VMD for PDB structures 5V5K (E65R) and 3KX9 (K150A/R151). Cage structure comparison, with E65R in green, and K150A/R151A aligned magenta. Quantitative comparison of the two 24mer structures yielded an α-carbon RMSD of 2.19 Å, as calculated using VMD.[163]

A trimer of subunits from the tetrahedral wild-type structure was used in the computational design of the mutants. The structure of this trimer is retained in both the open (tetrahedral, wild-type) structures and in the closed (octahedral) structures of mutants E65R and K150A/R151A. The crystallographic structures of these trimers in the wt, E65R, and K150A/R151A are similar, with α -carbon RMSD values of 1.40 Å between E65R and the wt and 1.32 Å between E65R and K150A/R151A (see the alignment of structures in Fig. 12).

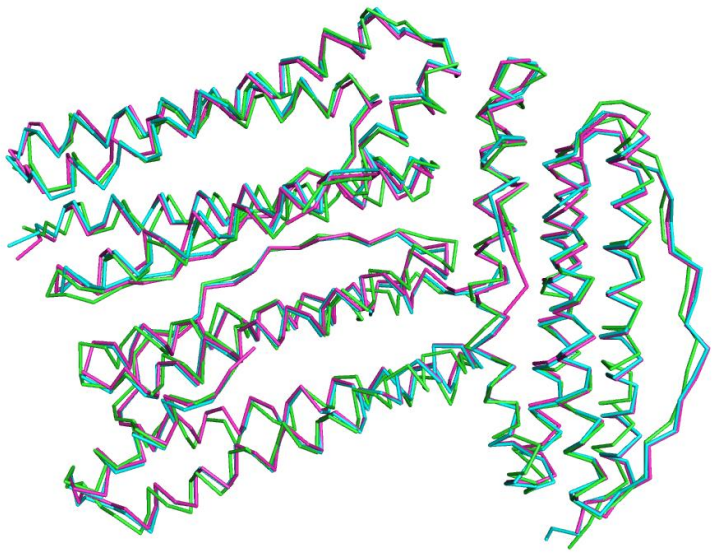

Figure 4.12. **Structural alignment of trimer of subunits from 24mer assembly.** Wild-type AfFtn (chains G, H, and J) (cyan), E65R (chains A, B, D) (green), and K150A/R151A (chains D, C, and D) (magenta). The α-carbon RMSD relative to the E65R structure was 1.40 Å for wild-type, open-pore AfFtn, 1.32 Å for K150A/R151A (chains D, C, and D).

It is remarkable that a single-point mutation led to such a dramatic change in assembly behavior. This symmetry shift can be rationalized using the structures of the 24mer and the computationally predicted side-chain conformation of R65. In the wt tetrahedral assembly, the amino acid exists in two distinct environments because of the symmetry of the cage. In one environment, R65 on one subunit is positioned directly across from R65 of a neighboring subunit (Fig. 10b), resulting in electrostatic repulsion. In the octahedral assembly, however, R65 is in only one environment. The residue is not in the proximity of an R65 residue on a separate protomer, which is consistent with octahedral assembly being preferred (Fig. 10a). Within the model of E65R, R65 and D138 form a salt bridge

at the interface present in both the tetrahedral and octahedral assemblies (Fig. 10c), which could lead to the observed enhanced cage stability. A127R was also predicted to form complementary electrostatic interactions (Fig. 10d). However, Arg127 is conformationally constrained at an interface that is more sterically crowded than the pore environment where the mutations E65R and D138K were introduced. With A127R, this introduction of a large residue at a subunit interface may reduce the stability of the 24mer at low salt concentrations relative to the wild type. Within the model of D138K, a K138−D34 salt bridge is observed between neighboring subunits, which may increase the population of 24mer relative to that of the wt (Fig. 10 e).

It is striking that we see substantial differences in self-assembly kinetics for the proteins alone, yet in the presence of AuNPs, protein assembly encapsulating the AuNP seems to be similarly fast for the wt, A127R, and D138K. A high-ionic strength solution perhaps does not model the charged AuNP surface, and thus, differences seen in protein-only assembly are not observed in the presence of AuNPs. We hypothesize that the AuNP nucleates protein assembly at its surface and may thereby increase the effective concentration of protein in solution, while favoring assembly over possible aggregation pathways. The DLS results suggest that protein concentration has a large effect on the rate of protein cage assembly, with a decreasing protein concentration in some cases leading to faster assembly. This is likely due to minimal aggregation of dimers at low concentrations, allowing assembly to occur. Reassembly could not be monitored by DLS at 0.3 mg/mL because of the low signal-to-noise ratio, but on the basis of the results from 1, 2, and 5 mg/mL samples, we would expect assembly to occur rapidly at 0.3 mg/mL for

all three proteins that disassemble.

When NPs are introduced into a biological medium, protein adsorption is rapid and evolves with time. This shell of protein on the NP surface is termed the protein corona.[164] Protein adsorption has been shown to sterically stabilize NPs against aggregation with increasing salt concentrations,[165, 166] similar to our results. A127R, D138K, and wt protein all successfully encapsulated AuNPs as seen by a native gel and prevented aggregation of AuNPs with an increasing ionic strength compared to particles without protein present. Although E65R does not encapsulate AuNPs at the same high yields as the disassembling proteins, there is still some level of AuNP stabilization, as the SPR red shift for the E65R−AuNP sample was smaller than that of bare particles. The fully assembled E65R cage may be adsorbed to the AuNP surface. It is possible that the cage dynamics of E65R are such that some AuNPs are encapsulated, as indicated by faint bands overlapping by the native gel (Fig 4.8). Such is the case for lumazine synthase from Aquifex aeolicus , a protein cage capable of encapsulating large protein cargo without first disassembling.[167] The greater 24mer stability and cargo selectivity exhibited by E65R open up the possibility of designing more specific ferritin−cargo interactions in the future. New types of cargo for ferritin encapsulation are being investigated.

4.6. Conclusion

We have shown that single-point mutations of AfFtn can have varying effects on thermal stability, assembly symmetry, and self-assembly equilibrium, kinetics, and reversibility. More dramatic charge changes such as changing negatively charged residues to positive

ones increased the stability of the 24mer at decreasing ionic strengths, while a less dramatic change, changing a neutral residue to a positive one, had a slightly destabilizing effect. The E65R mutant shows enhanced cage formation as well as thermal stability, formation of the 24mer under low-salt conditions, and self-assembly in octahedral symmetry rather than tetrahedral. The kinetics of self-assembly were also affected by mutation, with A127R showing nanocage assembly that was slower than that of the wt, and D138K assembling faster. These results corroborate earlier studies with other ferritin species, demonstrating the generality of single-point mutations along subunit interfaces dramatically affecting cage self-assembly. All mutants showed some salt stabilization of AuNPs compared to bare particles, but only mutants that retained their ability to disassemble showed full AuNP encapsulation. Enhanced control over protein cage assembly could have applications in delivery, nanomaterials separations, and controlled inorganic NP synthesis.

5| Computational Design of Self-Assembling Peptide Cages with Surface Plasticity

5.1Abstract

The computational design of protein-protein interactions allows for the potential development of symmetric assemblies of homomeric protein complexes with spherical shapes and hollow interiors. Precise control of the size and thickness of such assemblies would allow for the design of protein cages for encapsulating cargo of specific sizes, such as nanoparticles or other proteins. Rational and computational approaches have been used to design synthetic protein cages, but the lack of size tunability of the individual subunits has made it difficult to target specific sizes. Coiled-coil peptides have a periodic structure that makes them amenable to size tuning. We computationally designed a helical peptide to self-assemble into a tetrameric coiled-coil, which further assembles into a spherical cage. The design explicitly incorporated side chain conformational symmetry breaking between two α-helices in an asymmetric unit, allowing for the assembly of 48 individual α-helices in an octahedral arrangement. The designed sequence was synthesized and the kinetics of self-assembly of the peptides was controlled by dialysis from denaturing to non-denaturing conditions. Assemblies were analyzed with dynamic light scattering (DLS) and transmission electron microscopy (TEM), which revealed that the designed peptide assembles into spherical particles of the target size and shape. This computationally designed nano-cage serves to illustrate the possibility of designing

assemblies with specific dimensions using short, designed peptide sequences.

5.2. Introduction

Interactions between proteins occur extensively in living organisms and are involved in most cellular processes.[18] The strength of protein-protein interactions (PPIs) ranges from weak and transient association ($K_d \approx \mu M$) to the formation of tightly bound complexes ($K_d < \mu M$).[168, 169] Biochemical signals are passed across cells and across membranes by cascades of transient PPIs,[170] whereas strong associations hold functional protein complexes together.[171] Protein complexes can be composed of a single type of protein (homo-oligomeric) or from different types of proteins (hetero-oligomeric). Some protein complexes, such as vaults[19], virus capsids, and ferritin cages,[172] self-assemble into particles with hollow interiors.

Protein-based nano-containers could be used to encapsulate molecular cargo, and could serve as targetable drug-delivery vehicles.[173] Cage-like protein nano-structures have been also investigated for their potential to serve as templates upon which to display antigens for vaccine development.[174] Computational design has been used previously to reengineer naturally occurring proteins into homo-oligomeric and hetero-oligomeric assemblies.[44, 45, 175] However, these constructs lack size tunability given the non-periodic nature of the globular protein subunits. On the other hand, the size and length of helical peptides can be easily controlled given their periodic structure. When designed to arrange into spherical assemblies, the size of the resulting nano-cages may be potentially controlled as well.

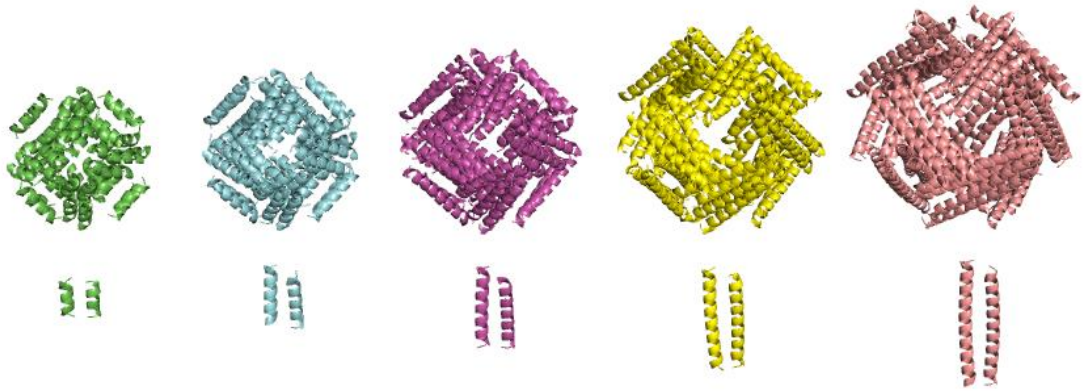

Figure 5.1: **Schematic of peptide-based nanocages.** Peptide cages of various sizes could be targeted by varying the length of the helical peptide subunits.

We set out to computationally design helical peptides that assemble to form spherical nano-cages with octahedral symmetry. We devised and implemented the concept of "sequence symmetry" between subunits in non-symmetrical assemblies in an octahedral arrangement. Side chain rotamer probability profiles are calculated independently for different chains, but amino acid probabilities are constrained so that equivalent residue positions of different chains have identical amino acid probabilities. In this manner we can access single sequences that are compatible with multiple distinct local environments.

The resulting 29-residue peptide sequences were experimentally realized and characterized. One of the sequences exhibited behavior consistent with self-assembly to form the targeted nano-cage. Dynamic light scattering (DLS) experiments confirmed that the peptides formed assemblies of the target size in solution, and transmission electron microscopy (TEM) images confirmed that spherical particles of the target size were

formed.

5.3. Computational Design of Self-Assembling Peptides

To design a peptide nano-cage, we targeted an assembly of coiled-coils in a spherical arrangement with octahedral symmetry. We selected the antiparallel tetrameric coiled-coil GCN4-pV as the backbone template upon which to construct protein-protein interfaces that would lead to self-assembly of the nano-cage. GCN4-pV is a highly stable tetramer of α-helices, which lead us to hypothesis that this structure would be amenable to substantial sequence variation. By maintaining the amino acid identities at core positions along the heptad repeat, we could ensure that the correct oligomerization state and internal symmetry were maintained. The D_2 symmetric arrangement of the tetramer means that there are three perpendicular C_2 axes, which cannot be mapped simultaneously onto the O_h point group. As a consequence, the asymmetric unit must be composed of a dimer of α-helices with one internal C_2 axis. A second C_2 axis can be mapped onto one of the C_2 axes of the O_h point group. Applying all symmetry operations results in the formation of a spherical arrangement of 12 coiled-coil subunits constructed from 24 asymmetric units. This results in a total of 48 individual α-helices that make up the nano-cage as illustrated in Fig. 5.2.

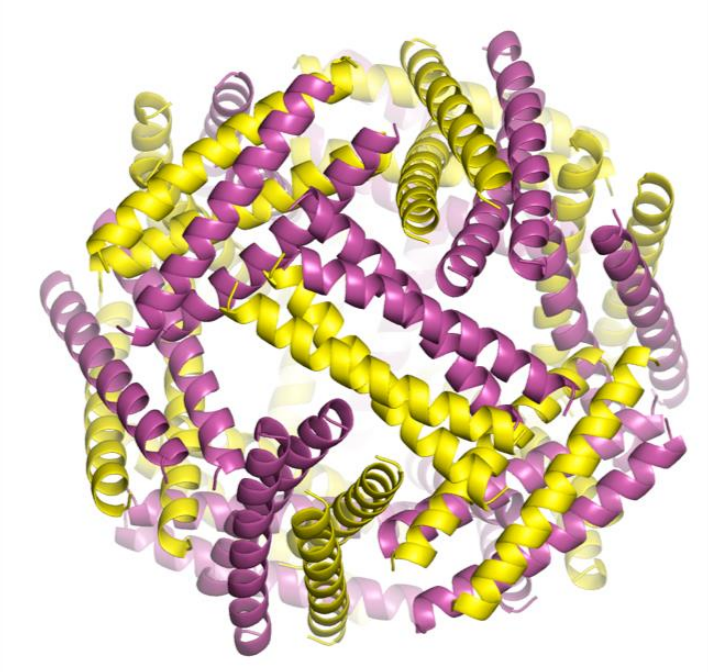

Figure 5.2: **Coiled-coil based peptide cage formation.** The coiled-coil unit (center) is composed of an asymmetric unit of two α-helices (yellow) and a C_2 symmetry related element (magenta). Formation of the octahedral cage occurs with the application of the 22 remaining symmetry operations. Cages of different sizes can be formed by translation of the coiled coil subunit in and out the plane, and rotation about the C_2 axis.

An atomic coordinate file for GCN4-pV[116] was obtained from entry in the Protein Data Bank (PDB 2B22). The structure was modified to provide an N-terminal acetyl cap and a C-terminal amide cap; each was added to the structure using PyMol. The GCN4-pV crystal structure contains one α-helix in the asymmetric unit, and the antiparallel tetramer can be generated from the symmetry operations of the $I4_122$crystal space group.

The crystal lattice information was used in PyMol to obtain one half of the coiled-coil, corresponding to antiparallel dimer. A coordinate file consisting of two chains of the tetramer was generated from the asymmetric unit (designated as chain A) and a symmetry mate (designated as chain B). The resulting structure was centered at the origin by translating each set of atomic coordinates. A sequence-structure energy landscape was generated by considering translations and rotations along a C2 symmetry axis of the tetrameric bundle (x = y at z = 0), and calculating the atomic coordinates of symmetry related elements within O_h. The 24 symmetry operations for the O_h point group were obtained from the F432 space group table in *The International Tables for Crystallography*.[176] In the F432 space group table, the x=y line at z = 0 is one of the C_2 axes in the set of transformations used to generate O_h symmetry. The symmetry related element generated by this operation from the dimeric asymmetric unit results in the formation of the full tetrameric coiled-coil as well as the octahedral cage.

Starting from the configuration shown in Fig 5.3, R is defined as the displacement along the C_2 axis, while θ is defined as the rotation about the axis. Values of R were varied from 28 Å to 43 Å at 0.1 Å intervals and values of θ were varied from 0° to 180° at 1° intervals.

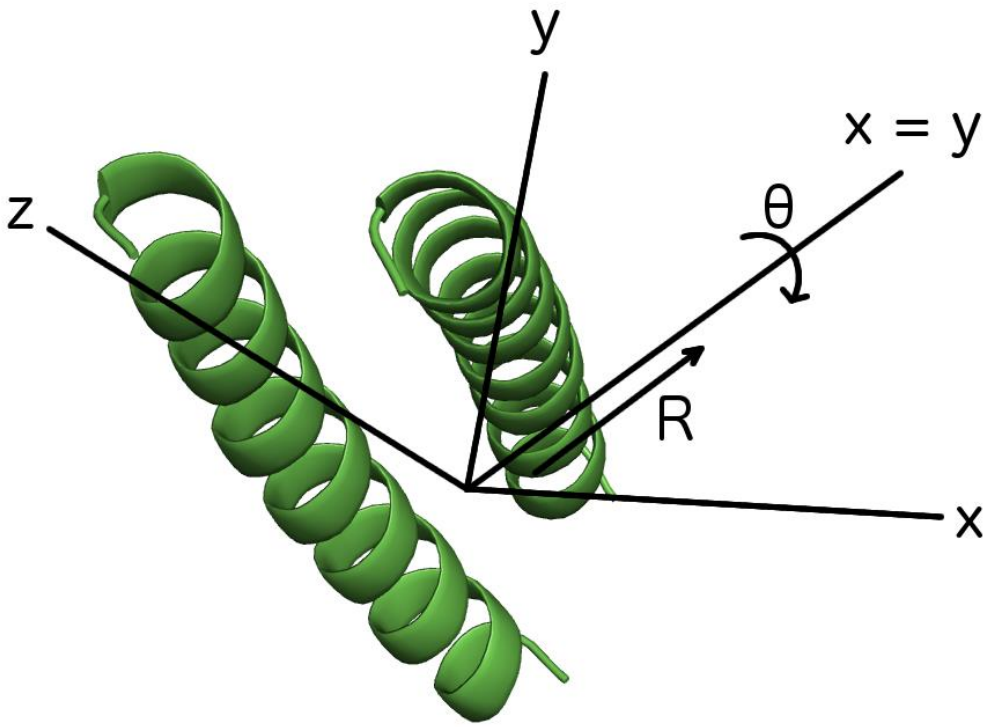

Figure 5.3: **Asymmetric unit of octahedral assembly.** Peptide asymmetric unit backbone and structural degrees of freedom.

At each point on the landscape, an ensemble with the following degrees of freedom (Table 5.1) was considered. Valine residues at the *d* and *e* positions, leucine residues at the *a* positions, and asparagine at 16 (*e* position) were retained as in the GCN4-pV sequence.

Sites	Degrees Of Freedom
1	Acetyl cap – 1 conformation.
31	Amine cap – 1 conformation.
8, 9, 15, 22, 23, 30	Valine – All conformations from Dunbrack 2002 rotamer library.
5, 12, 19, 26, 29	Leucine – All conformations from Dunbrack 2002 rotamer library.
16	Asparigine – 10 most common probable conformations from Dunbrack 2002 rotamer library.
2, 3, 4, 6, 7, 10, 11, 13, 14, 17, 18, 20, 21, 24, 25, 27, 28	All amino acids except cysteine and proline – up to the10 most common probable conformations from Dunbrack 2002 rotamer library.

Table 5.1: **Ensemble amino acid type and conformational degrees of freedom.** Amino acids and rotamer conformations allowed at each site on the alpha helix.

The resulting ensembles consisted of a total 4502 unique identity-rotamer (monomer) states across 62 sites. Virtual copies of the asymmetric unit were created in an octahedral arrangement using the 24 symmetry transformation matrices of the F_{432} crystal space group. For a given configuration of the nano-cage, monomer states having high net potential interactions (>30 kcal/mol) with backbone atoms of any subunit (plus energy of interaction with the side chain of its symmetry related element) were removed.

The site-specific type/rotamer probabilities were determined by minimizing an effective free energy for each computationally generated candidate octahedral assembly. The following parameters and constraints were used:

Parameter	Setting
Force field version	AMBER84
Force field paramers	Electrostatic, Van der Waals, dihedral, hydrogen bond
β	0.5 mol/kcal
Non-bonded cut-off	8 Å.
Distance-dependent dielectric	4rε
Pairwise energy cap	30.0 kcal/mol

Table 5.2: **Parameters of structure-energy landscape calculation.** Force field energetic terms and values of adjustable parameters.

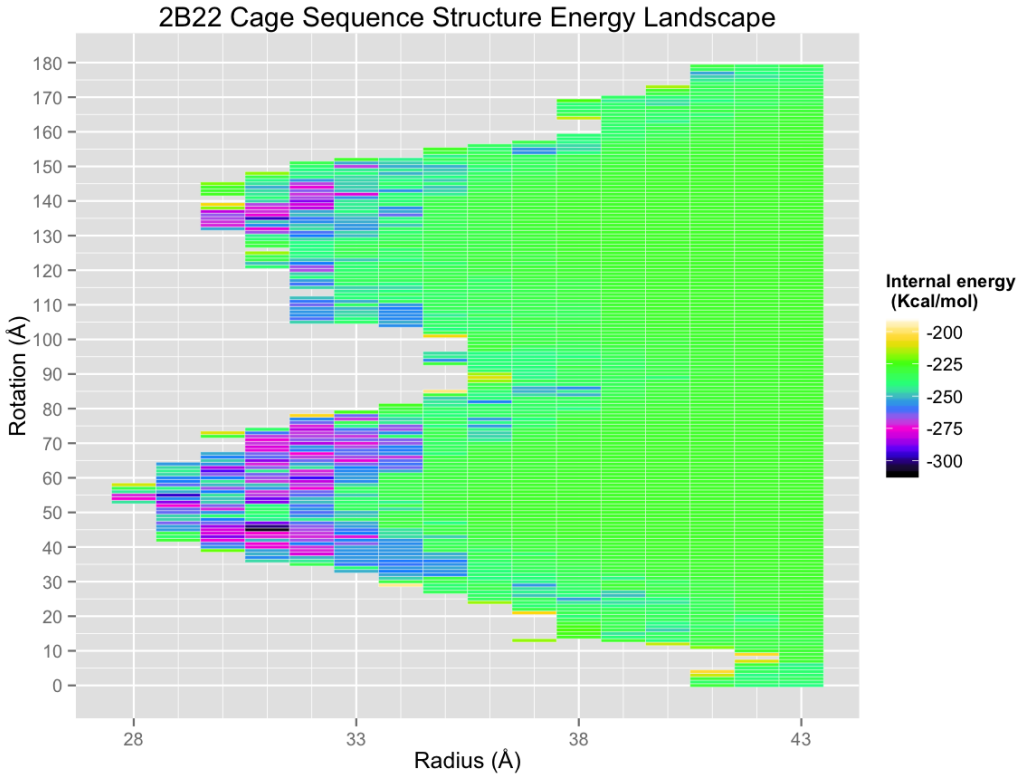

Figure 5.4: **Sequence structure energy landscape.** Local minima were selected for further design calculations.

For each candidate nano-cage structure, the calculations yield an effective internal energy, which is an average energy obtained using the calculated monomer state probabilities (Fig. 5.3). The 32 configurations on the landscape with lowest internal energies were chosen for sequence selection calculations. At each of these points, an ensemble was generated with the same amino acid degrees of freedom indicated in Table 5.1, but all conformations in the Dunbrack rotamer library were allowed. Free energy minimization was carried out as indicated above, but with the following additional

constraints. At each residue position *i* of chains A and B within the asymmetric unit, m_i types (amino acids) and c_{m_i} conformations of each amino acid (type) are permitted. We let α denote the amino acid type and r(α) the side chain rotamer state; *m* - 1 constraints were imposed to constrain that the amino acid probabilities and equivalent sites on chains A and B are equal.

$$\begin{aligned} &\min F(\mathbf{w}, b) = U(\mathbf{w}) - \frac{S(\mathbf{w})}{b} \\ &\text{subject to } 0 \le w_i \le 1, \\ &\sum_{a_i=1}^{g} \sum_{r(a_i)}^{R} w_i(a_i, r(a_i)) = 1, \\ &\text{and } \sum_{r_i(a)}^{c_{m_i}} w_i^A(a_i, r_i(a)) - \sum_{r_i(a)}^{c_{m_i}} w_i^B(a_i, r_i(a)) = 0 \end{aligned} \tag{5.3}$$

where the residue index *i* is used for equivalent sites of the two helices A and B in the asymmetric unit.

In order to guarantee that equivalent sites were populated with the same sets of amino acid types, the following trimming procedure was implemented:

1) Generate crystal lattice of untrimmed ensemble.

2) Remove high-energy states by trimming procedure outlined above.

3) Discard crystal lattice.

4) For equivalent sites on chains A and B, remove all conformers from the site on chain B of any amino acid type not permitted energetically at the corresponding

site on chain A. Conversely, remove all conformers on the site on chain A of any amino acid type not found at the corresponding site on chain B.

5) Generate crystal lattice of trimmed ensemble.

After each round of ensemble free energy minimization, each site was examined for highly probable amino acid types. If at any given site the most probable amino acid was more than twice as probable as next most probable amino acid, all other amino acid degrees of freedom were removed. Determination of the monomer state probabilities using the reduced ensemble was carried out, and the procedure was repeated until no further degrees of freedom were removed. The calculated probabilities were used to select sequences and conformations for further computational validation.

For each result, a coordinate file (PDB format) was generated with the most probable conformation of the most probable type at each site. These PDB files contained a CRYST1 header indicating an F432 space group with unit cells spaced apart by more than 999 Å. This serves to generate the octahedral cage assembly from the crystal lattice routines in analysis and visualization software.

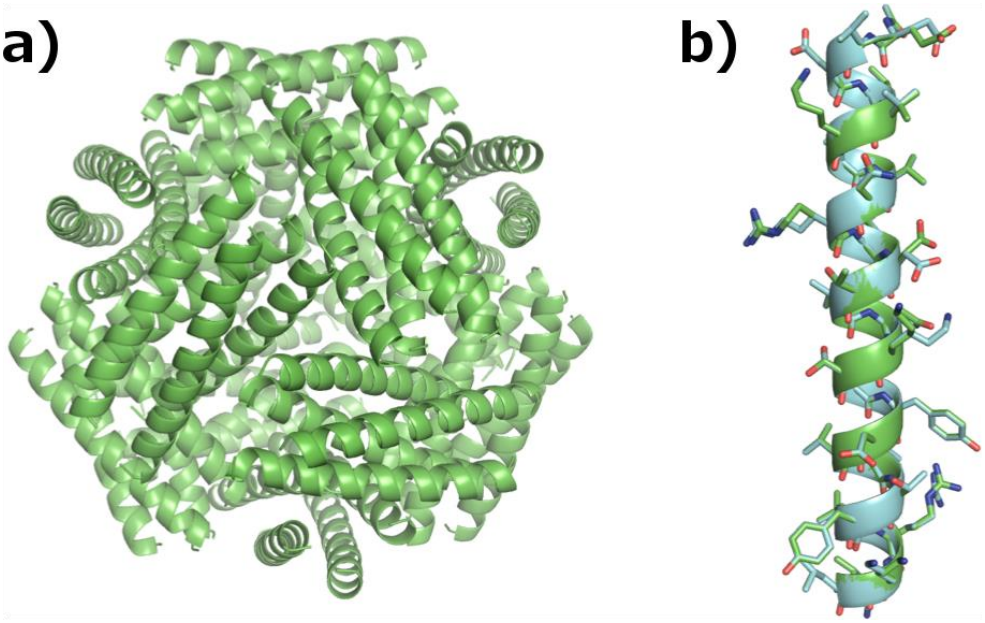

Figure 5.5: **Cage assembly of a designed peptide.** a) Structure of the peptide nano-cage at R = 30 Å and θ = 44°. Cage diameter from the outer edges is 8.6 nm. b) Overlap of the two chains in the asymmetric unit of showing differences in side chain conformations.

All 32 candidate sequences and structures were uploaded to the Proteins, Interfaces, Structures and Assemblies (PISA) server[177] and the assemblies were analyzed. To further assess the likelihood of forming the targeted peptide assemblies in solution, each candidate coiled-coil was subjected to a trimeric protein-protein docking calculation using the ClusPro web server.[178-182] A coordinate file of only a single tetrahelical coiled-coil was uploaded to the server as the receptor, the multimer docking option was selected, and the number of subunits was specified as three. The program predicts the structure of a trimer of helical bundles using only the monomer (tetrahelical coiled coil) structure. Only designs that were predicted as stable by PISA and ClusPro were considered for

experimental validation. PISA defines as stable an assembly with a calculated energy of dissociation score greater than 0 kcal/mol.

Agreement between ClusPro docking prediction and the designed cage models was judged visually. The representative structures of the most populated cluster for each set of coefficients (“balanced”, “electrostatic-favored”, “hydrophobic-favored”, and “VdW+Elec”) were visually compared to the predicted cage assembly. If each docked chain in the ClusPro prediction could be assigned to an equivalent chain in the design, then the design was judged to be stable. The backbone RMSD (atoms N, C, CA and O) between the prediction and the design was calculated. Seven candidates were selected as those having a low internal energy from the sequence calculations and low predicted PISA assembly scores.

Name	R (Å)	θ(°)	Final Ensemble Energy (kcal/mol)	PISA ΔG_{int} (kcal/mol)	Best ClusPro Prediction - Backbone RMSD (Å)	Sequence
3D-1	31	44	-281.608	-1042.4	Elec+VdW 4.99	VIDLTTVVNFLDFVNESLYHV VEWLRVLV
3D-2	30	46	-274.54	-986.2	Electrostatic -favored 1.894	FVDLTTVVNRLDYVNTSLYS VVTWLRVLV
3D-3	31	54	-257.375	-907.8	Balanced 5.02	TVDLTHVVWTLDKVNKTLY HVVTLLRILV
3D-4	30	44	-292.338	-751.7	Balanced 2.868	EVDLVKVVNRLDTVNKSLYD VVTYLRKLV
3D-5	32	138	-284.95	-726.8	Balanced 5.539	EVDLVHVVRDLDYVNKRLY YVVTWLRHLV
3D-6	31	45	-303.527	-688.4	Electrostatic -favored 4.754	DDALVTVVNRLDRVNESLYY VVEDLRKLV
3D-7	30	43	-294.916	-680.7	Electrostatic -favored 3.057	DEDLTRVVNRLDTVNKGLYD VVTYLRKLV

Table 5.3: **Self-assembling peptide candidates.** Top seven candidates were ranked by predicted PISA assembly stability, with ensemble average energy, most probable sequence, and Best ClusPro prediction.

5.4. Experimental Verification

Experiments were carried out by members of the laboratory of Professor Darrin Pochan at the University of Delaware. Peptides were synthesized by microwave-assisted solid-phase peptide synthesis (MW-SPPS) and purified on a reverse-phase high-pressure liquid instrument (RP-HPLC).

Circular dichroism (CD) spectra were collected at pH 4.5, pH 7.0, and pH 9.5. The 3D-4 peptide was observed to possess a high degree of α-helical character at pH values of 4.5 and 7.0, consistent with the design. The peptide also displayed a high degree of stability, as evidenced by a lack of a cooperative melting point transition during temperature melt experiments. Given that the theoretical isolectric point of this sequence is around pI = 9.5, a decrease in α-helical character (molar ellipticity at 222 nm) is to be expected at pH 9.5 due to a decrease in stabilizing electrostatic interactions or due to aggregation and precipitation of the peptide.

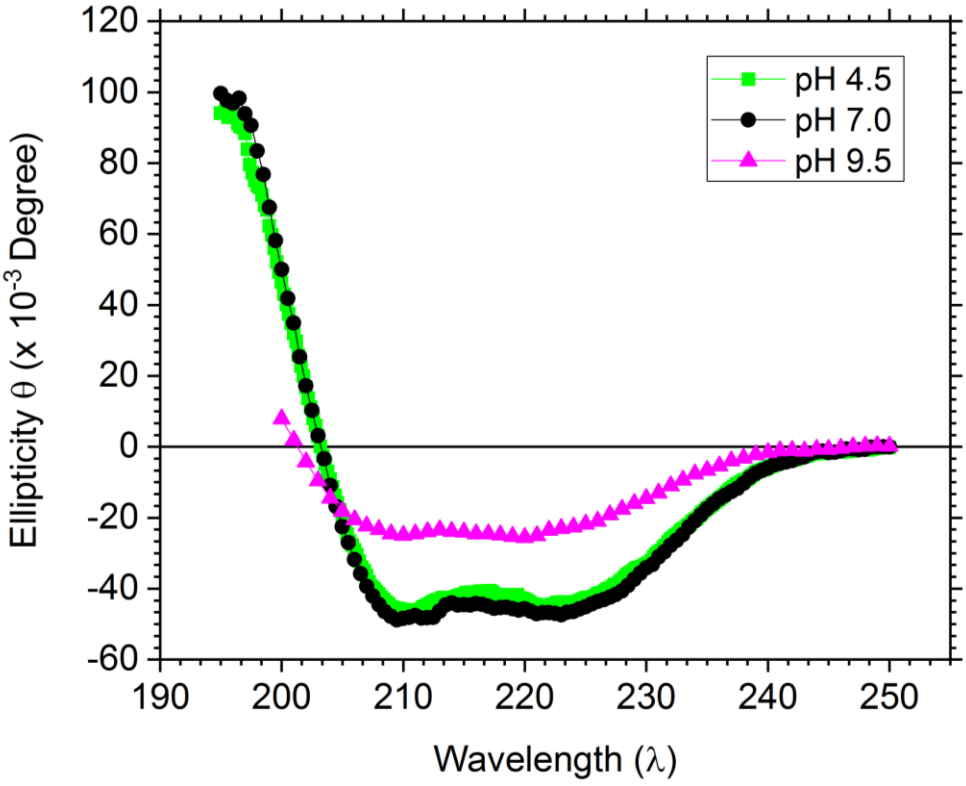

Figure 5.6: **Circular dichroism spectra of 3D-4.** Data were collected at pH 4.5, pH 7.0, and pH 9.5.

The self-assembly 3D-4 peptide was carried out in a controlled manner by solubilizing the peptide in denaturing conditions and dialyzing into selected buffer solutions. A schematic of the self-assembly protocol is shown in Fig 5.5.

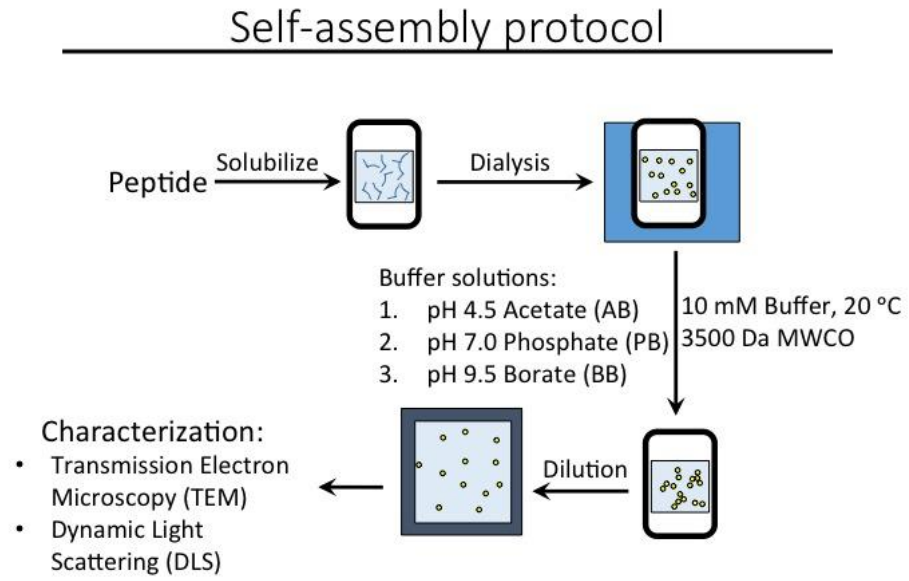

Figure 5.7: **Self-assembly protocol for 3D-4 nanocage.** Peptide is solubilized in denaturing conditions and dialyzed into a refolding buffer. Refolding of the peptide initiates self-assembly of the peptide nanocages.

In order to determine distribution of particle sizes in solution, dynamic light scattering (DLS) measurements were collected at various concentrations and for a range of pH values. Self-assembly of the peptide cages proved to be sensitive to the conditions used, with pH, ionic strength, and peptide concentration as important factors.

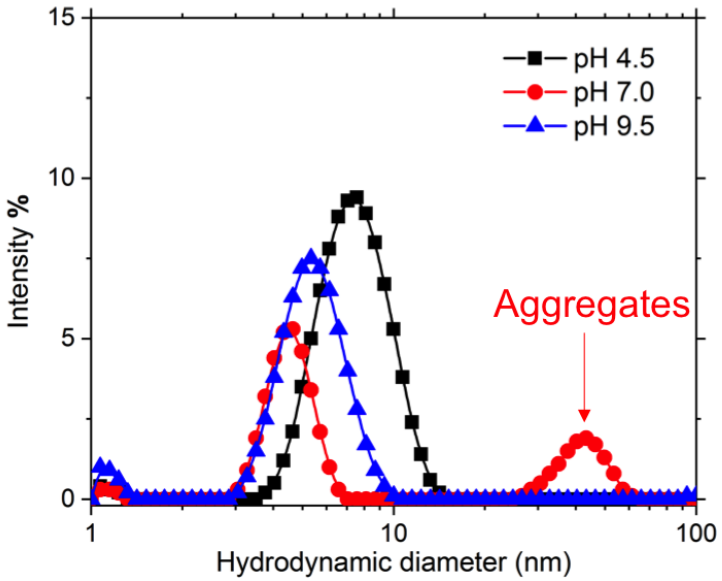

Figure 5.8: **Dynamic light scattering data for 3D-4 at varying pH.** At pH 4.5, the average particle hydrodynamic diameter is 7.6 ± 1.9 nm, compared to the expected peptide cage diameter of 8.6 nm in the design model.

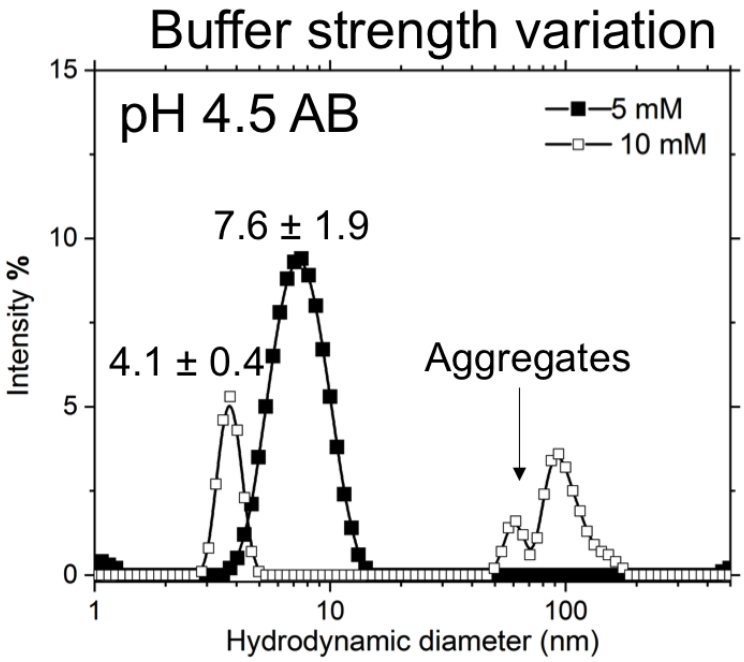

Figure 5.9: **Dynamic light scattering data for 3D-4 at varying buffer strength.** At an acetate buffer concentration of 10 nm, peptide forms larger order aggregates.

Transmission electron microscopy (TEM) was used to visualize the 3D-4 peptide nano-cages. Solutions of peptide were dialyzed from denaturing conditions to 6 pM guanidine hydrochloride (GnHCL) in pure water. Solutions of 1 mM peptide concentration were observed to form non-specific aggregates, with no well-defined particles observed. Solutions of 0.05 mM peptide concentration were observed to form well-defined spherical particles, with no non-specific aggregate formation. Most particles observed were in the range of 9 nm to 11 nm, with some larger particles of around 13 nm observed. Sonication of the solution helped to break up these higher-order structures.

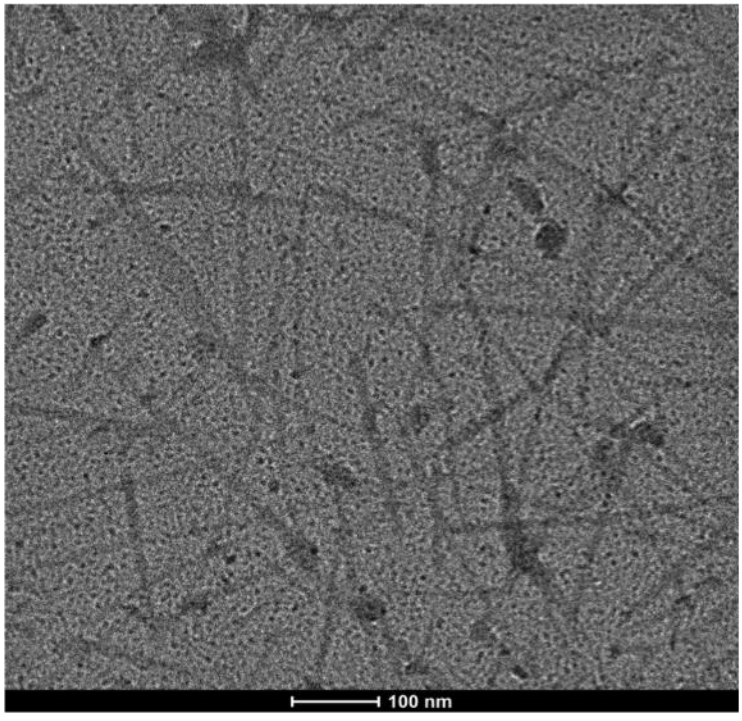

Figure 5.10: **Transmission electron microscopy (TEM) image 3D-4 at 1mM peptide concentration.** Non-specific aggregates and fibril-like structures were observed at a 1mM concentration.

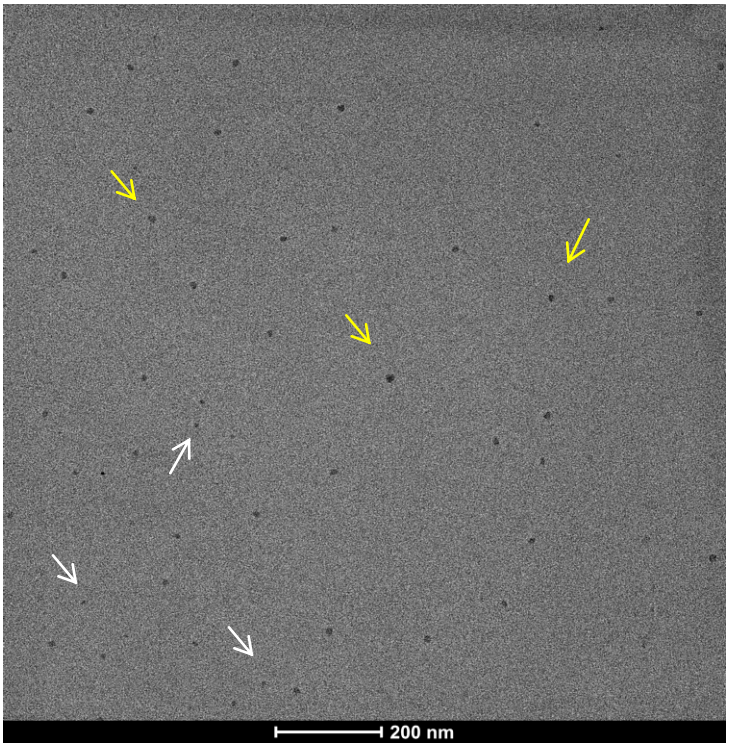

Figure 5.11: **Transmission electron microscopy (TEM) image 3D-4 at 0.05 mM peptide concentration.** At a peptide concentration of 0.05 mM, extensive particles formation is observed with no non-specific aggregation. Most particles are in the range of 9nm to 11 nm (indicated by white arrows), with some larger particles of around 13 nm (indicated by yellow particles). These larger particles could be composites of smaller particles, such as dimers or trimers.

5.5. Discussion

We have designed a set of peptides that we hypothesized would self-assemble into spherical nano-cages with octahedral symmetry. Point-group symmetry is a common feature of protein assemblies,[13] and has been used extensively in the design of novel protein-based structures.[44-46, 183, 184] This is based on the fact that symmetric arrangements can produce assemblies from subunits that interact by way of a single identical interface.[13] Icosahedral symmetry provides the highest number of chiral subunits that can

be arranged symmetrically, with 60 subunits arranged to form spherical particles.

We chose to use α-helical peptides as the subunits from which to form hollow spherical nano-cages with octahedral symmetry. However, the use of tetrameric coiled-coil subunit results in competing systems of symmetry, one for the internal symmetry of the bundle, and one for the overall symmetry of the assembly. If any of the axes of symmetry internal to the coiled coil do not align with one of the symmetry axes for the assembly, then different α-helices will occupy non-equivalent positions within that assembly. Interactions between identical molecules occupying non-equivalent positions will break the strictly symmetrical arrangement, resulting in the presence of more than one unique interface.

Symmetry breaking necessitates that the subunits interact in distinct manners depending on their positions within the assembly. Viruses have evolved proteins with interface plasticity capable of interacting in distinct ways, gaining the ability to go beyond the symmetrical limit to form large capsids without increasing the size of their genetic material.[185]

Rational and computational approaches have been used in the design of highly symmetric protein assemblies, such as homo-oligomeric and hetero-oligomeric icosahedral cages derived from naturally occurring proteins. [41, 43-46, 175, 183] Having achieved the symmetrical limit with computational protein design, an ongoing challenge has been the capacity to break the symmetry in the design of biomaterials.[184] One strategy that has been suggested for symmetry breaking in the design of protein assemblies is to allow for sequence

variations between subunits at non-identical interfaces, thus changing the oligomeric state from homomeric to heteromeric.[184, 186]

To design truly homomeric assemblies with broken symmetry requires identifying a sequence of amino acid side chains that will adopt more than one distinct conformational target. These target conformations should have energies that are nearly equivalent, so the population of one state does not outcompete the population of the other. The large number of conformational degrees of freedom available to solvent-exposed side chains, combined with the astronomically large number of sequences that are possible from the set of natural amino acids, should make it theoretically possible to access molecules with nearly degenerate energy states.

To access such sequences we devised a "sequence symmetry" constraint that could be applied to sequence optimization calculations. As shown in equation 5.3, rotamer probabilities at all sites in the asymmetric unit were calculated independently, but the sum of rotamer probabilities for each amino acid was constrained to be equal at equivalent residue positions. This resulted in identical sequence profiles for the two chains in the asymmetric unit, but different populations of likely rotamer states.

We applied the sequence symmetry constraint to the minima of a sequence-structure energy landscape composed of an asymmetric unit of two α-helices in an octahedral arrangement. The resulting models were subjected to fast Fourier-transform docking calculations[178-182] to identify subunits that were likely to assemble into the designed structures in solution. A select number of sequences were chosen for synthesis and

experimental characterization. One such sequence, termed 3D-4, had a solubility in water that made it amenable to solution experiments.

The assembly of the 3D-4 peptide in solution was sensitive to experimental conditions, a phenomenon that is well known for protein nano-cage formation.[187, 188] A high concentration of self-assembling proteins leads to run-away oligomerization, resulting in the formation of aggregates. The same is true of our designed nano-cages, which require dilute conditions for formation. In order to prevent rapid aggregate formation of the peptide subunits, the self-assembly process was carried out in a gradual manner by dialyzing from denaturing conditions to non-denaturing conditions. This process was successful in annealing the system into its target configuration, avoiding kinetic trapping in local minima.

The dialyzed solutions were shown to form oligomers of well-defined size in solution with a hydrodynamic diameter consistent with the design model. The particles were observed directly with TEM imaging, further confirming that the design of the nanocages was successful. The close agreement between the particle sizes predicted from the design models and the particle sizes observed experimentally serves to validate the hypothesis that the α-helical subunits self-assemble into nanocages by accessing distinct surface conformations.

5.6. Conclusion

We have shown that it is possible to design custom-sized homomeric peptide-based nanocages by using coiled-coil subunits of specific length and oligomerization state. The

internal symmetry of the coiled-coiled subunit can be reconciled with the overall symmetry of the assembly by allowing for symmetry breaking between α-helical segments. The self-assembly of these synthetic systems can be controlled by a gradual transitioning of the peptides from unfolded to α-helical, thus preventing the uncontrolled assembly that results from the strength of the designed protein-protein interactions between subunits. We observed that the computationally designed peptide formed assemblies of the target size in solution, and were well defined enough to visualize their spherical shapes by TEM. This methodology paves the way for the design of nano-compartments with precise dimensional control by overcoming the constraints imposed by the use non-periodic subunits and perfect point-group symmetry.

BIBLIOGRAPHY

1. Fischer, E. a. F. E., Ueber einige Derivate des Glykocolls. *Berichte der deutschen chemischen Gesellschaft* **1901,** *34* (2), 2868--2877.

2. Crick, F. H. C., THE FOURIER TRANSFORM OF A COILED-COIL. *Acta Crystallographica* **1953,** *6* (8-9), 685-689.

3. Canutescu, A. A.; Dunbrack, R. L., Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science* **2003,** *12* (5), 963-972.

4. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D-Biological Crystallography* **2004,** *60*, 2126-2132.

5. Briki, F.; Doucet, J.; Etchebest, C., A procedure for refining a coiled coil protein structure using X-ray fiber diffraction and modeling. *Biophysical Journal* **2002,** *83* (4), 1774-1783.

6. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000,** *28* (1), 235-242.

7. Kono, H.; Saven, J. G., Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *Journal of Molecular Biology* **2001,** *306* (3), 607-628.

8. Saven, J. G., Designing protein energy landscapes. *Chemical Reviews* **2001,** *101* (10), 3113-3130.

9. Janin, J., Protein-protein recognition. *Progress in Biophysics & Molecular Biology* **1995,** *64* (2-3), 145-166.

10. McFedries, A.; Schwaid, A.; Saghatelian, A., Methods for the Elucidation of Protein-Small Molecule Interactions. *Chemistry & Biology* **2013,** *20* (5), 667-673.

11. Sheriff, S.; Chang, C. Y. Y.; Ezekowitz, R. A. B., HUMAN MANNOSE-BINDING PROTEIN CARBOHYDRATE-RECOGNITION DOMAIN TRIMERIZES THROUGH A TRIPLE ALPHA-HELICAL COILED-COIL. *Nature Structural Biology* **1994,** *1* (11), 789-794.

12. Gallop, J. L.; Jao, C. C.; Kent, H. M.; Butler, P. J. G.; Evans, P. R.; Langen, R.; T McMahon, H., Mechanism of endophilin N-BAR domain-mediated membrane curvature. *Embo Journal* **2006,** *25* (12), 2898-2910.

13. Levy, E. D.; Erba, E. B.; Robinson, C. V.; Teichmann, S. A., Assembly reflects evolution of protein complexes. *Nature* **2008,** *453* (7199), 1262-U66.

14. Marsh, J. A.; Teichmann, S. A., Structure, Dynamics, Assembly, and Evolution of Protein Complexes. In *Annual Review of Biochemistry, Vol 84*, Kornberg, R. D., Ed. Annual Reviews: Palo Alto, 2015; Vol. 84, pp 551-575.

15. Waldron, K. J.; Rutherford, J. C.; Ford, D.; Robinson, N. J., Metalloproteins and metal sensing. *Nature* **2009,** *460* (7257), 823-830.

16. Skou, J. C.; Esmann, M., THE NA,K-ATPASE. *Journal of Bioenergetics and Biomembranes* **1992,** *24* (3), 249-261.

17. Lehman, I. R.; Bessman, M. J.; Simms, E. S.; Kornberg, A., ENZYMATIC SYNTHESIS OF DEOXYRIBONUCLEIC ACID .1. PREPARATION OF SUBSTRATES AND PARTIAL PURIFICATION OF AN ENZYME FROM ESCHERICHIA-COLI. *Journal of Biological Chemistry* **1958,** *233* (1), 163-170.

18. Braun, P.; Gingras, A. C., History of protein-protein interactions: From egg-white to complex networks. *Proteomics* **2012,** *12* (10), 1478-1498.

19. Kedersha, N. L.; Rome, L. H., ISOLATION AND CHARACTERIZATION OF A NOVEL RIBONUCLEOPROTEIN PARTICLE - LARGE STRUCTURES CONTAIN A SINGLE SPECIES OF SMALL RNA. *Journal of Cell Biology* **1986,** *103* (3), 699-709.

20. Andrews, S. C.; Smith, J. M. A.; Hawkins, C.; Williams, J. M.; Harrison, P. M.; Guest, J. R., OVERPRODUCTION, PURIFICATION AND CHARACTERIZATION OF THE BACTERIOFERRITIN OF ESCHERICHIA-COLI AND A C-TERMINALLY EXTENDED VARIANT. *European Journal of Biochemistry* **1993,** *213* (1), 329-338.

21. Johnson, E.; Cascio, D.; Sawaya, M. R.; Gingery, M.; Schroder, I., Crystal structures of a tetrahedral open pore ferritin from the hyperthermophilic Archaeon Archaeoglobus fulgidus. *Structure* **2005,** *13* (4), 637-648.

22. Harrison, S. C.; Olson, A. J.; Schutt, C. E.; Winkler, F. K.; Bricogne, G., Tomato bushy stunt virus at 2.9 [angst] resolution. *Nature* **1978,** *276* (5686), 368-373.

23. Abad-Zapatero, C.; Abdel-Meguid, S. S.; Johnson, J. E.; Leslie, A. G. W.; Rayment, I.; Rossmann, M. G.; Suck, D.; Tsukihara, T., Structure of southern bean mosaic virus at 2.8 [angst] resolution. *Nature* **1980,** *286* (5768), 33-39.

24. Chen, K. Q.; Arnold, F. H., TUNING THE ACTIVITY OF AN ENZYME FOR UNUSUAL ENVIRONMENTS - SEQUENTIAL RANDOM MUTAGENESIS OF SUBTILISIN-E FOR CATALYSIS IN DIMETHYLFORMAMIDE. *Proceedings of the National Academy of Sciences of the United States of America* **1993,** *90* (12), 5618-5622.

25. Kan, S. B. J.; Lewis, R. D.; Chen, K.; Arnold, F. H., Directed evolution of cytochrome c for carbon-silicon bond formation: Bringing silicon to life. *Science* **2016,** *354* (6315), 1048-1051.

26. Maynard Smith, J., Natural Selection and the Concept of a Protein Space. *Nature* **1970,** *225* (5232), 563-564.

27. Regan, L.; Degrado, W. F., CHARACTERIZATION OF A HELICAL PROTEIN DESIGNED FROM 1ST PRINCIPLES. *Science* **1988,** *241* (4868), 976-978.

28. Bryson, J. W.; Desjarlais, J. R.; Handel, T. M.; DeGrado, W. F., From coiled coils to small globular proteins: Design of a native-like three-helix bundle. *Protein Science* **1998,** *7* (6), 1404-1414.

29. Handel, T. M.; Williams, S. A.; Degrado, W. F., METAL-ION DEPENDENT MODULATION OF THE DYNAMICS OF A DESIGNED PROTEIN. *Science* **1993,** *261* (5123), 879-885.

30. Robertson, D. E.; Farid, R. S.; Moser, C. C.; Urbauer, J. L.; Mulholland, S. E.; Pidikiti, R.; Lear, J. D.; Wand, A. J.; Degrado, W. F.; Dutton, P. L., DESIGN AND SYNTHESIS OF MULTI-HEME PROTEINS. *Nature* **1994,** *368* (6470), 425-431.

31. Kamtekar, S.; Schiffer, J. M.; Xiong, H. Y.; Babik, J. M.; Hecht, M. H., PROTEIN DESIGN BY BINARY PATTERNING OF POLAR AND NONPOLAR AMINO-ACIDS. *Science* **1993,** *262* (5140), 1680-1685.

32. Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T., A SWITCH BETWEEN 2-STRANDED, 3-STRANDED AND 4-STRANDED COILED COILS IN GCN4 LEUCINE-ZIPPER MUTANTS. *Science* **1993,** *262* (5138), 1401-1407.

33. Harbury, P. B.; Kim, P. S.; Alber, T., CRYSTAL-STRUCTURE OF AN ISOLEUCINE-ZIPPER TRIMER. *Nature* **1994,** *371* (6492), 80-83.

34. Grosset, A. M.; Gibney, B. R.; Rabanal, F.; Moser, C. C.; Dutton, P. L., Proof of principle in a de novo designed protein maquette: An allosterically regulated, charge-activated conformational switch in a tetra-alpha-helix bundle. *Biochemistry* **2001,** *40* (18), 5474-5487.

35. Dahiyat, B. I.; Sarisky, C. A.; Mayo, S. L., De novo protein design: Towards fully automated sequence selection. *Journal of Molecular Biology* **1997,** *273* (4), 789-796.

36. Dahiyat, B. I.; Mayo, S. L., De novo protein design: Fully automated sequence selection. *Science* **1997,** *278* (5335), 82-87.

37. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D., Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003,** *302* (5649), 1364-1368.

38. Lehmann, A.; Saven, J. G., Computational design of four-helix bundle proteins that bind nonbiological cofactors. *Biotechnology Progress* **2008,** *24* (1), 74-79.

39. Procko, E.; Berguig, G. Y.; Shen, B. W.; Song, Y. F.; Frayo, S.; Convertine, A. J.; Margineantu, D.; Booth, G.; Correia, B. E.; Cheng, Y. H.; Schief, W. R.; Hockenbery, D. M.; Press, O. W.; Stoddard, B. L.; Stayton, P. S.; Baker, D., A Computationally Designed Inhibitor of an Epstein-Barr Viral Bcl-2 Protein Induces Apoptosis in Infected Cells. *Cell* **2014,** *157* (7), 1644-1656.

40. Baran, D.; Pszolla, M. G.; Lapidoth, G. D.; Norn, C.; Dym, O.; Unger, T.; Albeck, S.; Tyka, M. D.; Fleishman, S. J., Principles for computational design of binding antibodies. *Proceedings of the National Academy of Sciences of the United States of America* **2017,** *114* (41), 10900-10905.

41. Zhang, H. V.; Polzer, F.; Haider, M. J.; Tian, Y.; Villegas, J. A.; Kiick, K. L.; Pochan, D. J.; Saven, J. G., Computationally designed peptides for self-assembly of nanostructured lattices. *Science Advances* **2016,** *2* (9), 8.

42. Gonen, S.; DiMaio, F.; Gonen, T.; Baker, D., Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* **2015,** *348* (6241), 1365-1368.

43. Lai, Y. T.; Cascio, D.; Yeates, T. O., Structure of a 16-nm Cage Designed by Using Protein Oligomers. *Science* **2012,** *336* (6085), 1129-1129.

44. Bale, J. B.; Gonen, S.; Liu, Y. X.; Sheffler, W.; Ellis, D.; Thomas, C.; Cascio, D.; Yeates, T. O.; Gonen, T.; King, N. P.; Baker, D., Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **2016,** *353* (6297), 389-394.

45. King, N. P.; Bale, J. B.; Sheffler, W.; McNamara, D. E.; Gonen, S.; Gonen, T.; Yeates, T. O.; Baker, D., Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **2014,** *510* (7503), 103-+.

46. Lanci, C. J.; MacDermaid, C. M.; Kang, S.-g.; Acharya, R.; North, B.; Yang, X.; Qiu, X. J.; DeGrado, W. F.; Saven, J. G., Computational design of a protein crystal. *Proceedings of the National Academy of Sciences of the United States of America* **2012,** *109* (19), 7304-7309.

47. Karanicolas, J.; Com, J. E.; Chen, I.; Joachimiak, L. A.; Dym, O.; Peck, S. H.; Albeck, S.; Unger, T.; Hu, W. X.; Liu, G. H.; Delbecq, S.; Montelione, G. T.; Spiegel, C.

P.; Liu, D. R.; Baker, D., A De Novo Protein Binding Pair By Computational Design and Directed Evolution. *Molecular Cell* **2011,** *42* (2), 250-260.

48. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V., STEREOCHEMISTRY OF POLYPEPTIDE CHAIN CONFIGURATIONS. *Journal of Molecular Biology* **1963,** *7* (1), 95-&.

49. Mandel, N.; Mandel, G.; Trus, B. L.; Rosenberg, J.; Carlson, G.; Dickerson, R. E., TUNA CYTOCHROME-C AT 2.0 A-RESOLUTION .3. COORDINATE OPTIMIZATION AND COMPARISON OF STRUCTURES. *Journal of Biological Chemistry* **1977,** *252* (13), 4619-4636.

50. Chothia, C., One thousand families for the molecular biologist. *Nature* **1992,** *357* (6379), 543-544.

51. Khersonsky, O.; Fleishman, S. J., Why reinvent the wheel? Building new proteins based on ready-made parts. *Protein Science* **2016,** *25* (7), 1179-1187.

52. Wang, Q.; Canutescu, A. A.; Dunbrack, R. L., SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nature Protocols* **2008,** *3* (12), 1832-1847.

53. Ponder, J. W.; Richards, F. M., TERTIARY TEMPLATES FOR PROTEINS - USE OF PACKING CRITERIA IN THE ENUMERATION OF ALLOWED SEQUENCES FOR DIFFERENT STRUCTURAL CLASSES. *Journal of Molecular Biology* **1987,** *193* (4), 775-791.

54. Dunbrack, R. L.; Karplus, M., BACKBONE-DEPENDENT ROTAMER LIBRARY FOR PROTEINS - APPLICATION TO SIDE-CHAIN PREDICTION. *Journal of Molecular Biology* **1993,** *230* (2), 543-574.

55. Shapovalov, M. V.; Dunbrack, R. L., A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011,** *19* (6), 844-858.

56. Levitt, M.; Sharon, R., ACCURATE SIMULATION OF PROTEIN DYNAMICS IN SOLUTION. *Proceedings of the National Academy of Sciences of the United States of America* **1988,** *85* (20), 7557-7561.

57. McCammon, J. A.; Gelin, B. R.; Karplus, M., DYNAMICS OF FOLDED PROTEINS. *Nature* **1977,** *267* (5612), 585-590.

58. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.;

Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* **1998,** *102* (18), 3586-3616.

59. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER. *Journal of Chemical Physics* **1983,** *79* (2), 926-935.

60. Jiang, L.; Kuhlman, B.; Kortemme, T. A.; Baker, D., A "solvated rotainer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins-Structure Function and Bioinformatics* **2005,** *58* (4), 893-904.

61. Marshall, S. A.; Vizcarra, C. L.; Mayo, S. L., One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. *Protein Science* **2005,** *14* (5), 1293-1304.

62. Pokala, N.; Handel, T. M., Energy functions for protein design: Adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *Journal of Molecular Biology* **2005,** *347* (1), 203-227.

63. Morozov, A. V.; Kortemme, T.; Baker, D., Evaluation of models of electrostatic interactions in proteins. *Journal of Physical Chemistry B* **2003,** *107* (9), 2075-2090.

64. Desmet, J.; Maeyer, M. D.; Hazes, B.; Lasters, I., The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **1992,** *356* (6369), 539-542.

65. Simoncini, D.; Allouche, D.; de Givry, S.; Delmas, C.; Barbe, S.; Schiex, T., Guaranteed Discrete Energy Optimization on Large Protein Design Problems. *Journal of Chemical Theory and Computation* **2015,** *11* (12), 5980-5989.

66. Yang, X.; Saven, J. G., Computational methods for protein design and protein sequence variability: biased Monte Carlo and replica exchange. *Chemical Physics Letters* **2005,** *401* (1-3), 205-210.

67. Song, Y. F.; Tyka, M.; Leaver-Fay, A.; Thompson, J.; Baker, D., Structure-guided forcefield optimization. *Proteins-Structure Function and Bioinformatics* **2011,** *79* (6), 1898-1909.

68. Bolon, D. N.; Grant, R. A.; Baker, T. A.; Sauer, R. T., Specificity versus stability in computational protein design. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (36), 12724-12729.

69. Yang, W.; Lai, L. H., Computational design of ligand-binding proteins. *Current Opinion in Structural Biology* **2017,** *45*, 67-73.

70. Baker, D., An exciting but challenging road ahead for computational enzyme design. *Protein Science* **2010,** *19* (10), 1817-1819.

71. Fazelinia, H.; Cirino, P. C.; Maranas, C. D., OptGraft: A computational procedure for transferring a binding site onto an existing protein scaffold. *Protein Science* **2009,** *18* (1), 180-195.

72. Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Rothlisberger, D.; Baker, D., New algorithms and an in silico benchmark for computational enzyme design. *Protein Science* **2006,** *15* (12), 2785-2794.

73. Tantillo, D. J.; Chen, J. G.; Houk, K. N., Theozymes and compuzymes: theoretical models for biological catalysis. *Current Opinion in Chemical Biology* **1998,** *2* (6), 743-750.

74. Pearson, A. D.; Mills, J. H.; Song, Y. F.; Nasertorabi, F.; Han, G. W.; Baker, D.; Stevens, R. C.; Schultz, P. G., Trapping a transition state in a computationally designed protein bottle. *Science* **2015,** *347* (6224), 863-867.

75. Liu, X. F.; Taylor, R. D.; Griffin, L.; Coker, S. F.; Adams, R.; Ceska, T.; Shi, J. Y.; Lawson, A. D. G.; Baker, T., Computational design of an epitope-specific Keap1 binding antibody using hotspot residues grafting and CDR loop swapping. *Scientific Reports* **2017,** *7*, 11.

76. Clarke, N. D.; Yuan, S. M., METAL SEARCH - A COMPUTER-PROGRAM THAT HELPS DESIGN TETRAHEDRAL METAL-BINDING SITES. *Proteins-Structure Function and Genetics* **1995,** *23* (2), 256-263.

77. Hellinga, H. W.; Richards, F. M., CONSTRUCTION OF NEW LIGAND-BINDING SITES IN PROTEINS OF KNOWN STRUCTURE .1. COMPUTER-AIDED MODELING OF SITES WITH PREDEFINED GEOMETRY. *Journal of Molecular Biology* **1991,** *222* (3), 763-785.

78. Wang, X.; Zhao, K.; Kirberger, M.; Wong, H.; Chen, G. T.; Yang, J. J., Analysis and prediction of calcium-binding pockets from apo-protein structures exhibiting calcium-induced localized conformational changes. *Protein Science* **2010,** *19* (6), 1180-1190.

79. Huang, P. S.; Boyken, S. E.; Baker, D., The coming of age of de novo protein design. *Nature* **2016,** *537* (7620), 320-327.

80. Grigoryan, G.; DeGrado, W. F., Probing Designability via a Generalized Model of Helical Bundle Geometry. *Journal of Molecular Biology* **2011,** *405* (4), 1079-1100.

81. Lupas, A. N.; Gruber, M., The structure of alpha-helical coiled coils. *Fibrous Proteins: Coiled-Coils, Collagen and Elastomers* **2005,** *70*, 37-+.

82. Kabsch, W., SOLUTION FOR BEST ROTATION TO RELATE 2 SETS OF VECTORS. *Acta Crystallographica Section A* **1976,** *32* (SEP1), 922-923.

83. Coutsias, E. A.; Seok, C.; Dill, K. A., Using quaternions to calculate RMSD. *Journal of Computational Chemistry* **2004,** *25* (15), 1849-1857.

84. Kneller, G. R., Comment on "Using quaternions to calculate RMSD" - J. Comp. Chem. 25, 1849 (2004). *Journal of Computational Chemistry* **2005,** *26* (15), 1660-1662.

85. Nagi, A. D.; Regan, L., An inverse correlation between loop length and stability in a four-helix-bundle protein. *Folding & Design* **1997,** *2* (1), 67-75.

86. Lahr, S. J.; Engel, D. E.; Stayrook, S. E.; Maglio, O.; North, B.; Geremia, S.; Lombardi, A.; DeGrado, W. F., Analysis and design of turns in alpha-helical hairpins. *Journal of Molecular Biology* **2005,** *346* (5), 1441-1454.

87. Hu, X. Z.; Wang, H. C.; Ke, H. M.; Kuhlman, B., High-resolution design of a protein loop. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** *104* (45), 17668-17673.

88. Zou, J. M.; Saven, J. G., Statistical theory of combinatorial libraries of folding proteins: Energetic discrimination of a target structure. *Journal of Molecular Biology* **2000,** *296* (1), 281-294.

89. US Department of Energy, Critical Materials Strategy. US Department of Energy: Washington, DC, 2011; pp 1-191.

90. Pavel, C. C.; Lacal-Arantegui, R.; Marmier, A.; Schuler, D.; Tzimas, E.; Buchert, M.; Jenseit, W.; Blagoeva, D., Substitution strategies for reducing the use of rare earths in wind turbines. *Resources Policy* **2017,** *52*, 349-357.

91. Alonso, E.; Sherman, A. M.; Wallington, T. J.; Everson, M. P.; Field, F. R.; Roth, R.; Kirchain, R. E., Evaluating Rare Earth Element Availability: A Case with Revolutionary Demand from Clean Technologies. *Environmental Science & Technology* **2012,** *46* (6), 3406-3414.

92. Paulick, H.; Machacek, E., The global rare earth element exploration boom: An analysis of resources outside of China and discussion of development perspectives. *Resources Policy* **2017,** *52*, 134-153.

93. Haxel, G.; Hedrick, J. B.; Orris, G. J.; Geological Survey (U.S.), Rare earth elements

critical resources for high technology. In *USGS fact sheet 087-02* [Online] U.S. Dept. of the Interior, U.S. Geological Survey: Reston, Va., 2002; p. 4 p. http://pubs.usgs.gov/fs/2002/fs087-02/http://purl.access.gpo.gov/GPO/LPS104453.

94. Huang, X. W.; Dong, J. S.; Wang, L. S.; Feng, Z. Y.; Xue, Q. N.; Meng, X. L., Selective recovery of rare earth elements from ion-adsorption rare earth element ores by stepwise extraction with HEH(EHP) and HDEHP. *Green Chemistry* **2017,** *19* (5), 1345-1352.

95. Xiao, Y. F.; Long, Z. Q.; Huang, X. W.; Feng, Z. Y.; Cui, D. L.; Wang, L. S., Study on non-saponification extraction process for rare earth separation. *Journal of Rare Earths* **2013,** *31* (5), 512-516.

96. Texier, A. C.; Andres, Y.; LeCloirec, P., Selective biosorption of lanthanide (La, Eu) ions by Mycobacterium smegmatis. *Environmental Technology* **1997,** *18* (8), 835-841.

97. Texier, A. C.; Andres, Y.; Le Cloirec, P., Selective biosorption of lanthanide (La, Eu, Yb) ions by Pseudomonas aeruginosa. *Environmental Science & Technology* **1999,** *33* (3), 489-495.

98. Texier, A. C.; Andres, Y.; Le Cloirec, P., Selective biosorption of lanthanide (La, Eu, Yb) ions by an immobilized bacterial biomass. *Water Science and Technology* **2000,** *42* (5-6), 91-94.

99. Moriwaki, H.; Yamamoto, H., Interactions of microorganisms with rare earth ions and their utilization for separation and environmental technology. *Applied Microbiology and Biotechnology* **2013,** *97* (1), 1-8.

100. Kim, S. H.; Gunther, J. R.; Katzenellenbogen, J. A., Monitoring a Coordinated Exchange Process in a Four-Component Biological Interaction System: Development of a Time-Resolved Terbium-Based One-Donor/Three-Acceptor Multicolor FRET System. *Journal of the American Chemical Society* **2010,** *132* (13), 4685-4692.

101. Kohn, W. D.; Kay, C. M.; Sykes, B. D.; Hodges, R. S., Metal ion induced folding of a de novo designed coiled-coil peptide. *Journal of the American Chemical Society* **1998,** *120* (6), 1124-1132.

102. Nitz, M.; Sherawat, M.; Franz, K. J.; Peisach, E.; Allen, K. N.; Imperiali, B., Structural origin of the high affinity of a chemically evolved lanthanide-binding peptide. *Angewandte Chemie-International Edition* **2004,** *43* (28), 3682-3685.

103. Veliscek-Carolan, J.; Hanley, T. L.; Jolliffe, K. A., The impact of structural variation in simple lanthanide binding peptides. *Rsc Advances* **2016,** *6* (79), 75336-75346.

104. Park, D. M.; Reed, D. W.; Yung, M. C.; Eslamimanesh, A.; Lencka, M. M.; Anderko, A.; Fujita, Y.; Riman, R. E.; Navrotsky, A.; Jiao, Y. Q., Bioadsorption of Rare Earth Elements through Cell Surface Display of Lanthanide Binding Tags. *Environmental Science & Technology* **2016,** *50* (5), 2735-2742.

105. Pol, A.; Barends, T. R. M.; Dietl, A.; Khadem, A. F.; Eygensteyn, J.; Jetten, M. S. M.; Op den Camp, H. J. M., Rare earth metals are essential for methanotrophic life in volcanic mudpots. *Environmental Microbiology* **2014,** *16* (1), 255-264.

106. Pidcock, E.; Moore, G. R., Structural characteristics of protein binding sites for calcium and lanthanide ions. *Journal of Biological Inorganic Chemistry* **2001,** *6* (5-6), 479-489.

107. Allen, J. E.; McLendon, G. L., Tryptophan and tyrosine to terbium fluorescence resonance energy transfer as a method to "map" aromatic residues and monitor docking. *Biochemical and Biophysical Research Communications* **2006,** *349* (4), 1264-1268.

108. Snyder, E. E.; Buoscio, B. W.; Falke, J. J., CALCIUM(II) SITE SPECIFICITY - EFFECT OF SIZE AND CHARGE ON METAL-ION BINDING TO AN EF-HAND-LIKE SITE. *Biochemistry* **1990,** *29* (16), 3937-3943.

109. Berwick, M. R.; Lewis, D. J.; Jones, A. W.; Parslow, R. A.; Dafforn, T. R.; Cooper, H. J.; Wilkie, J.; Pikramenou, Z.; Britton, M. M.; Peacock, A. F. A., De Novo Design of Ln(III) Coiled Coils for Imaging Applications. *Journal of the American Chemical Society* **2014,** *136* (4), 1166-1169.

110. Grzyb, J.; Xu, F.; Weiner, L.; Reijerse, E. J.; Lubitz, W.; Nanda, V.; Noy, D., De novo design of a non-natural fold for an iron-sulfur protein: Alpha-helical coiled-coil with a four-iron four-sulfur cluster binding site in its central core. *Biochimica Et Biophysica Acta-Bioenergetics* **2010,** *1797* (3), 406-413.

111. Fletcher, J. M.; Boyle, A. L.; Bruning, M.; Bartlett, G. J.; Vincent, T. L.; Zaccai, N. R.; Armstrong, C. T.; Bromley, E. H. C.; Booth, P. J.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N., A Basis Set of de Novo Coiled-Coil Peptide Oligomers for Rational Protein Design and Synthetic Biology. *Acs Synthetic Biology* **2012,** *1* (6), 240-250.

112. Yu, H.; Noskov, S. Y.; Roux, B., Two mechanisms of ion selectivity in protein binding sites. *Proceedings of the National Academy of Sciences of the United States of America* **2010,** *107* (47), 20329-20334.

113. Fang, H. Y.; Cole, B. E.; Qiao, Y. S.; Bogart, J. A.; Cheisson, T.; Manor, B. C.; Carroll, P. J.; Schelter, E. J., Electro-kinetic Separation of Rare Earth Elements Using a Redox-Active Ligand. *Angewandte Chemie-International Edition* **2017,** *56* (43), 13450-13454.

114. Andraud, C.; Maury, O., Lanthanide Complexes for Nonlinear Optics: From Fundamental Aspects to Applications. *European Journal of Inorganic Chemistry* **2009,** (29-30), 4357-4371.

115. Gunnlaugsson, T.; Stomeo, F., Recent advances in the formation of luminescent lanthanide architectures and self-assemblies from structurally defined ligands. *Organic & Biomolecular Chemistry* **2007,** *5* (13), 1999-2009.

116. Deng, Y. Q.; Liu, J.; Zheng, Q.; Eliezer, D.; Kallenbach, N. R.; Lu, M., Antiparallel four-stranded coiled coil specified by a 3-3-1 hydrophobic heptad repeat. *Structure* **2006,** *14* (2), 247-255.

117. Lan, J.-H.; Shi, W.-Q.; Yuan, L.-Y.; Li, J.; Zhao, Y.-L.; Chai, Z.-F., Recent advances in computational modeling and simulations on the An(III)/Ln(III) separation process. *Coordination Chemistry Reviews* **2012,** *256* (13-14), 1406-1417.

118. Leaver-Fay, A.; O'Meara, M. J.; Tyka, M.; Jacak, R.; Song, Y. F.; Kellogg, E. H.; Thompson, J.; Davis, I. W.; Pache, R. A.; Lyskov, S.; Gray, J. J.; Kortemme, T.; Richardson, J. S.; Havranek, J. J.; Snoeyink, J.; Baker, D.; Kuhlman, B., Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. In *Methods in Protein Design*, Keating, A. E., Ed. Elsevier Academic Press Inc: San Diego, 2013; Vol. 523, pp 109-143.

119. Kameshwar, P. M.; Wadawale, A.; Ajgaonkar, V. R., Tris(N-acetylglycinato-[kappa]2O,O')triaquaterbium(III). *Acta Crystallographica Section E* **2007,** *63* (10), m2584.

120. Schrödinger, L., The PyMOL Molecular Graphics System, Version 1.3. 2015.

121. Wächter, A.; Biegler, L. T., On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* **2006,** *106* (1), 25-57.

122. Schrauber, H.; Eisenhaber, F.; Argos, P., ROTAMERS - TO BE OR NOT TO BE - AN ANALYSIS OF AMINO-ACID SIDE-CHAIN CONFORMATIONS IN GLOBULAR-PROTEINS. *Journal of Molecular Biology* **1993,** *230* (2), 592-612.

123. Harder, T.; Boomsma, W.; Paluszewski, M.; Frellsen, J.; Johansson, K. E.; Hamelryck, T., Beyond rotamers: a generative, probabilistic model of side chains in proteins. *Bmc Bioinformatics* **2010,** *11*, 13.

124. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A NEW FORCE-FIELD FOR MOLECULAR MECHANICAL

SIMULATION OF NUCLEIC-ACIDS AND PROTEINS. *Journal of the American Chemical Society* **1984,** *106* (3), 765-784.

125. Kono, H.; Doi, J., A new method for side-chain conformation prediction using a Hopfield network and reproduced rotamers. *Journal of Computational Chemistry* **1996,** *17* (14), 1667-1683.

126. Stickle, D. F.; Presta, L. G.; Dill, K. A.; Rose, G. D., HYDROGEN-BONDING IN GLOBULAR-PROTEINS. *Journal of Molecular Biology* **1992,** *226* (4), 1143-1159.

127. Espadaler, J.; Fernandez-Fuentes, N.; Hermoso, A.; Querol, E.; Aviles, F. X.; Sternberg, M. J. E.; Oliva, B., ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Research* **2004,** *32*, D185-D188.

128. Bonet, J.; Planas-Iglesias, J.; Garcia-Garcia, J.; Marin-Lopez, M. A.; Fernandez-Fuentes, N.; Oliva, B., ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Research* **2014,** *42* (D1), D315-D319.

129. Liu, H. L.; Shu, Y. C.; Wu, Y. H., Molecular dynamics simulations to determine the optimal loop length in the helix-loop-helix motif. *Journal of Biomolecular Structure & Dynamics* **2003,** *20* (6), 741-745.

130. Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D-Biological Crystallography* **2010,** *66*, 12-21.

131. Kachi-Terajima, C.; Yanagi, K.; Kaziki, T.; Kitazawa, T.; Hasegawa, M., Luminescence tuning of imidazole-based lanthanide(III) complexes Ln = Sm, Eu, Gd, Tb, Dy. *Dalton Transactions* **2011,** *40* (10), 2249-2256.

132. Sherry, A. D.; Darnall, D. W.; Birnbaum, E. R., NUCLEAR MAGNETIC-RESONANCE STUDY OF HISTIDINE-NEODYMIUM(III) COMPLEXES. *Journal of Biological Chemistry* **1972,** *247* (11), 3489-&.

133. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005,** *26* (16), 1781-1802.

134. Reddi, A. R.; Guzman, T. R.; Breece, R. M.; Tiemey, D. L.; Gibney, B. R., Deducing the energetic cost of protein folding in zinc finger proteins using designed metallopeptides. *Journal of the American Chemical Society* **2007,** *129* (42), 12815-12827.

135. Wang, Z. X., AN EXACT MATHEMATICAL EXPRESSION FOR DESCRIBING COMPETITIVE-BINDING OF 2 DIFFERENT LIGANDS TO A PROTEIN MOLECULE. *Febs Letters* **1995,** *360* (2), 111-114.

136. Lombardi, A.; Summa, C. M.; Geremia, S.; Randaccio, L.; Pavone, V.; DeGrado, W. F., Retrostructural analysis of metalloproteins: Application to the design of a minimal model for diiron proteins. *Proceedings of the National Academy of Sciences of the United States of America* **2000,** *97* (12), 6298-6305.

137. Pasternak, A.; Kaplan, S.; Lear, J. D.; DeGrado, W. F., Proton and metal ion-dependent assembly of a model diiron protein. *Protein Science* **2001,** *10* (5), 958-969.

138. Fry, H. C.; Lehmann, A.; Sinks, L. E.; Asselberghs, I.; Tronin, A.; Krishnan, V.; Blasie, J. K.; Clays, K.; DeGrado, W. F.; Saven, J. G.; Therien, M. J., Computational de Novo Design and Characterization of a Protein That Selectively Binds a Highly Hyperpolarizable Abiological Chromophore. *Journal of the American Chemical Society* **2013,** *135* (37), 13914-13926.

139. Bender, G. M.; Lehmann, A.; Zou, H.; Cheng, H.; Fry, H. C.; Engel, D.; Therien, M. J.; Blasie, J. K.; Roder, H.; Saven, J. G.; DeGrado, W. F., De novo design of a single-chain diphenylporphyrin metalloprotein. *Journal of the American Chemical Society* **2007,** *129* (35), 10732-10740.

140. Polizzi, N. F.; Wu, Y.; Lemmin, T.; Maxwell, A. M.; Zhang, S.-Q.; Rawson, J.; Beratan, D. N.; Therien, M. J.; DeGrado, W. F., De novo design of a hyperstable non-natural protein–ligand complex with sub-Å accuracy. *Nature Chemistry* **2017,** *9*, 1157.

141. Kingsley, L. J.; Lill, M. A., Substrate tunnels in enzymes: Structure-function relationships and computational methodology. *Proteins-Structure Function and Bioinformatics* **2015,** *83* (4), 599-611.

142. Zhang, Y.; Fu, J.; Chee, S. Y.; Ang, E. X. W.; Orner, B. P., Rational disruption of the oligomerization of the mini-ferritin E. coli DPS through protein-protein interface mutation. *Protein Science* **2011,** *20* (11), 1907-1917.

143. Zhang, Y.; Wang, L. J.; Ardejani, M. S.; Aris, N. F.; Li, X.; Orner, B. P.; Wang, F., Mutagenesis study to disrupt electrostatic interactions on the twofold symmetry interface of Escherichia coli bacterioferritin. *Journal of Biochemistry* **2015,** *158* (6), 505-512.

144. Ardejani, M. S.; Chok, X. L.; Foo, C. J.; Orner, B. P., Complete shift of ferritin oligomerization toward nanocage assembly via engineered protein-protein interactions. *Chemical Communications* **2013,** *49* (34), 3528-3530.

145. Ardejani, M. S.; Li, N. X.; Orner, B. P., Stabilization of a Protein Nanocage through the Plugging of a Protein-Protein Interfacial Water Pocket. *Biochemistry* **2011,** *50* (19), 4029-4037.

146. Zhang, Y.; Raudah, S.; Teo, H.; Teo, G. W. S.; Fan, R. L.; Sun, X. M.; Orner, B. P., Alanine-shaving Mutagenesis to Determine Key Interfacial Residues Governing the Assembly of a Nano-cage Maxi-ferritin. *Journal of Biological Chemistry* **2010,** *285* (16), 12078-12086.

147. Sato, D.; Takebe, S.; Kurobe, A.; Ohtomo, H.; Fujiwara, K.; Ikeguchi, M., Electrostatic Repulsion during Ferritin Assembly and Its Screening by Ions. *Biochemistry* **2016,** *55* (3), 482-488.

148. Swift, J.; Butts, C. A.; Cheung-Lau, J.; Yerubandi, V.; Dmochowski, I. J., Efficient Self-Assembly of Archaeoglobus fulgidus Ferritin around Metallic Cores. *Langmuir* **2009,** *25* (9), 5219-5225.

149. Cheung-Lau, J. C.; Liu, D.; Pulsipher, K. W.; Liu, W.; Dmochowski, I. J., Engineering a well-ordered, functional protein-gold nanoparticle assembly. *Journal of Inorganic Biochemistry* **2014,** *130*, 59-68.

150. Pulsipher, K. W.; Dmochowski, I. J., Ferritin Encapsulation and Templated Synthesis of Inorganic Nanoparticles. In *Protein Cages: Methods and Protocols*, Orner, B. P., Ed. Springer New York: New York, NY, 2015; pp 27-37.

151. Butts, C. A.; Swift, J.; Kang, S. G.; Di Costanzo, L.; Christianson, D. W.; Saven, J. G.; Dmochowski, I. J., Directing Noble Metal Ion Chemistry within a Designed Ferritin Protein. *Biochemistry* **2008,** *47* (48), 12729-12739.

152. Swift, J.; Wehbi, W. A.; Kelly, B. D.; Stowell, X. F.; Saven, J. G.; Dmochowski, I. J., Design of functional ferritin-like proteins with hydrophobic cavities. *Journal of the American Chemical Society* **2006,** *128* (20), 6611-6619.

153. Dunbrack, R. L., Rotamer libraries in the 21(st) century. *Current Opinion in Structural Biology* **2002,** *12* (4), 431-440.

154. Calhoun, J. R.; Kono, H.; Lahr, S.; Wang, W.; DeGrado, W. F.; Saven, J. G., Computational design and characterization of a monomeric helical dinuclear metalloprotein. *Journal of Molecular Biology* **2003,** *334* (5), 1101-1115.

155. Schindelin, J.; Rueden, C. T.; Hiner, M. C.; Eliceiri, K. W., The ImageJ ecosystem: An open platform for biomedical image analysis. *Molecular Reproduction and Development* **2015,** *82* (7-8), 518-529.

156. Bakker, G. R.; Boyer, R. F., IRON INCORPORATION INTO APOFERRITIN - THE ROLE OF APOFERRITIN AS A FERROXIDASE. *Journal of Biological Chemistry* **1986,** *261* (28), 3182-3185.

157. Voss, N. R.; Gerstein, M., 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Research* **2010,** *38*, W555-W562.

158. Pfeiffer, C.; Rehbock, C.; Huhn, D.; Carrillo-Carrion, C.; de Aberasturi, D. J.; Merk, V.; Barcikowski, S.; Parak, W. J., Interaction of colloidal nanoparticles with their local environment: the (ionic) nanoenvironment around nanoparticles is different from bulk and determines the physico-chemical properties of the nanoparticles. *Journal of the Royal Society Interface* **2014,** *11* (96), 13.

159. Bernacchioni, C.; Ghini, V.; Pozzi, C.; Di Pisa, F.; Theil, E. C.; Turano, P., Loop Electrostatics Modulates the Intersubunit Interactions in Ferritin. *Acs Chemical Biology* **2014,** *9* (11), 2517-2525.

160. Khare, G.; Nangpal, P.; Tyagi, A. K., Unique Residues at the 3-Fold and 4-Fold Axis of Mycobacterial Ferritin Are Involved in Oligomer Switching. *Biochemistry* **2013,** *52* (10), 1694-1704.

161. Ohtomo, H.; Ohtomo, M.; Sato, D.; Kurobe, A.; Sunato, A.; Matsumura, Y.; Kihara, H.; Fujiwara, K.; Ikeguchi, M., A Physicochemical and Mutational Analysis of Intersubunit Interactions of Escherichia coli Ferritin A. *Biochemistry* **2015,** *54* (40), 6243-6251.

162. Sana, B.; Johnson, E.; Le Magueres, P.; Criswell, A.; Cascio, D.; Lim, S., The Role of Nonconserved Residues of Archaeoglobus fulgidus Ferritin on Its Unique Structure and Biophysical Properties. *Journal of Biological Chemistry* **2013,** *288* (45), 32663-32672.

163. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling* **1996,** *14* (1), 33-38.

164. Cedervall, T.; Lynch, I.; Lindman, S.; Berggard, T.; Thulin, E.; Nilsson, H.; Dawson, K. A.; Linse, S., Understanding the nanoparticle-protein corona using methods to quantify exchange rates and affinities of proteins for nanoparticles. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** *104* (7), 2050-2055.

165. Canaveras, F.; Madueno, R.; Sevilla, J. M.; Blazquez, M.; Pineda, T., Role of the Functionalization of the Gold Nanoparticle Surface on the Formation of Bioconjugates with Human Serum Albumin. *Journal of Physical Chemistry C* **2012,** *116* (18), 10430-10437.

166. Brewer, S. H.; Glomm, W. R.; Johnson, M. C.; Knag, M. K.; Franzen, S., Probing BSA binding to citrate-coated gold nanoparticles and surfaces. *Langmuir* **2005,** *21* (20), 9303-9307.

167. Worsdorfer, B.; Pianowski, Z.; Hilvert, D., Efficient in Vitro Encapsulation of Protein Cargo by an Engineered Protein Container. *Journal of the American Chemical Society* **2012,** *134* (2), 909-911.

168. Nooren, I. M. A.; Thornton, J. M., Diversity of protein-protein interactions. *Embo Journal* **2003,** *22* (14), 3486-3492.

169. Ozbabacan, S. E. A.; Engin, H. B.; Gursoy, A.; Keskin, O., Transient proteinprotein interactions. *Protein Engineering Design & Selection* **2011,** *24* (9), 635-648.

170. Pawson, T., Specificity in signal transduction: From phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* **2004,** *116* (2), 191-203.

171. Day, E. S.; Cote, S. M.; Whitty, A., Binding Efficiency of Protein-Protein Complexes. *Biochemistry* **2012,** *51* (45), 9124-9136.

172. Zhang, Y.; Orner, B. P., Self-Assembly in the Ferritin Nano-Cage Protein Superfamily. *International Journal of Molecular Sciences* **2011,** *12* (8), 5406-5421.

173. Belletti, D.; Pederzoli, F.; Forni, F.; Vandelli, M. A.; Tosi, G.; Ruozi, B., Protein cage nanostructure as drug delivery system: magnifying glass on apoferritin. *Expert Opinion on Drug Delivery* **2017,** *14* (7), 825-840.

174. Han, J. A.; Kang, Y. J.; Shin, C.; Ra, J. S.; Shin, H. H.; Hong, S. Y.; Do, Y.; Kang, S., Ferritin protein cage nanoparticles as versatile antigen delivery nanoplatforms for dendritic cell (DC)-based vaccine development. *Nanomedicine-Nanotechnology Biology and Medicine* **2014,** *10* (3), 561-569.

175. Hsia, Y.; Bale, J. B.; Gonen, S.; Shi, D.; Sheffler, W.; Fong, K. K.; Nattermann, U.; Xu, C. F.; Huang, P. S.; Ravichandran, R.; Yi, S.; Davis, T. N.; Gonen, T.; King, N. P.; Baker, D., Design of a hyperstable 60-subunit protein icosahedron. *Nature* **2016,** *535* (7610), 136-+.

176. edited by Theo, H., *International tables for crystallography. Volume A, Space-group symmetry*. Fifth, revised edition. Dordrecht ; London : Published for the International Union of Crystallography by Kluwer Academic Publishers, 2002.: 2002.

177. Krissinel, E.; Henrick, K., Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* **2007,** *372* (3), 774-797.

178. Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J., ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Research* **2004,** *32*, W96-W99.

179. Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J., ClusPro: An automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **2004,** *20* (1), 45-50.

180. Kozakov, D.; Brenke, R.; Comeau, S. R.; Vajda, S., PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins-Structure Function and Bioinformatics* **2006,** *65* (2), 392-406.

181. Kozakov, D.; Beglov, D.; Bohnuud, T.; Mottarella, S. E.; Xia, B.; Hall, D. R.; Vajda, S., How good is automated protein docking? *Proteins-Structure Function and Bioinformatics* **2013,** *81* (12), 2159-2166.

182. Kozakov, D.; Hall, D. R.; Xia, B.; Porter, K. A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S., The ClusPro web server for protein-protein docking. *Nature Protocols* **2017,** *12* (2), 255-278.

183. King, N. P.; Sheffler, W.; Sawaya, M. R.; Vollmar, B. S.; Sumida, J. P.; Andre, I.; Gonen, T.; Yeates, T. O.; Baker, D., Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* **2012,** *336* (6085), 1171-1174.

184. Yeates, T. O., Geometric Principles for Designing Highly Symmetric Self-Assembling Protein Nanomaterials. In *Annual Review of Biophysics, Vol 46*, Dill, K. A., Ed. Annual Reviews: Palo Alto, 2017; Vol. 46, pp 23-42.

185. Crick, F. H. C.; Watson, J. D., STRUCTURE OF SMALL VIRUSES. *Nature* **1956,** *177* (4506), 473-475.

186. Atwell, S.; Ridgway, J. B. B.; Wells, J. A.; Carter, P., Stable heterodimers from remodeling the domain interface of a homodimer using a phage display library. *Journal of Molecular Biology* **1997,** *270* (1), 26-35.

187. Adolph, K.W.; Butler, P. J. G., Studies on the assembly of a spherical plant virus: I. States of aggregation of the isolated protein. *Journal of Molecular Biology* **1974,** *88* (2), 327 - 341.

188. Nguyen, H. D.; Reddy, V. S.; Brooks, C. L., Deciphering the kinetic mechanism of spontaneous self-assembly of icosahedral capsids. *Nano Letters* **2007,** *7* (2), 338-344.