Symmetric ligand binding using tunable de novo designed symmetric protein dimers

Derrick R. Hicks

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

David Baker, Chair

Philip Bradley

Justin Kollman

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Symmetric ligand binding using tunable de novo designed symmetric protein dimers

Derrick R. Hicks

Chair of the Supervisory Committee:
David Baker
Department of Biochemistry

Cyclic two-fold (C2) symmetric ligands are common in nature, synthetic chemistry, and medicine. Additionally, we can now design millions of C2 symmetric peptides with an incredible diversity of sizes, shapes, and chemistries. De novo proteins capable of binding these C2 symmetric ligands could be useful in various applications, but scaffolds and methods to do this have been lacking. To solve this problem, I created a diverse set of C2 symmetric proteins with central cavities. I first designed curved repeat protein monomers sampling a continuum of curvatures, and then docked these into homodimers, generating a very wide range of C2 cavity shapes and sizes for functionalization. In total, 77 scaffolds were experimentally characterized, and of these, 23 (30%) appear to be folded as designed based on Small Angle X-ray Scattering data. Furthermore, crystallographic data for 4 designs (2 scaffolds and 2 functionalized binders) confirms the proteins fold as expected. A third scaffold design was determined to be monomeric by crystallographic analysis. Despite its failure to form the designed homodimer, the solved

monomer was in close agreement with the design model. We believe that these diverse scaffolds provide a rich set of starting points for binding a very wide range of C2 compounds. Advantages of this conception are that the cavities can be very diverse in size, shape, and available sidechain chemistry, and as the protein hydrophobic core is separated from the pocket because the cavity lining residues are on the exterior of the monomers, functionalization to create binding interactions for specific compounds is unlikely to destabilize either the monomers or the dimer interface. Finally, we used these scaffolds to bind symmetric chlorophyll dimers, which could have applications in synthetic light-harvesting, as well as to bind a C2 symmetric peptide, which could become a platform for the creation of entirely bioorthogonal chemically induced dimers.

TABLE OF CONTENTS

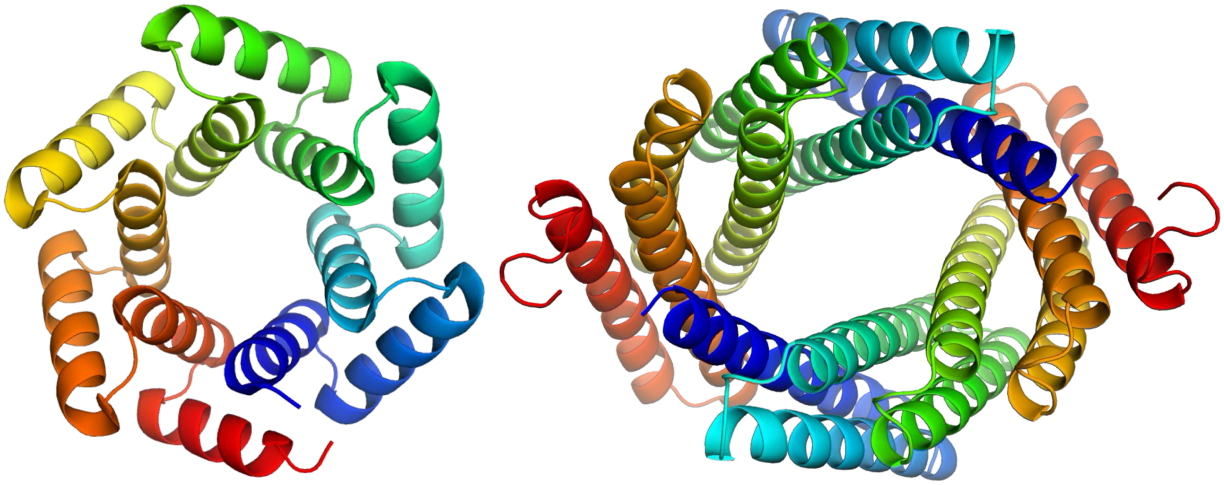


Figure 3. Original protein inspirations for this thesis. On the left is the monomeric crystal structure 4yxx (Doyle *et al.*, 2015) obtained from the Protein Data Bank (Berman *et al.*, 2000). On the right is the C2 symmetric design model TJ79C2 (Fallas *et al.*, 2017). The proteins are colored from blue (N-terminus) to red (C-terminus) with backbone cartoon representation. The images were produced using the molecular graphics program Pymol (*PyMOL*, no date). 30

2.2 MAKING A LIBRARY OF CURVED REPEAT PROTEINS 30

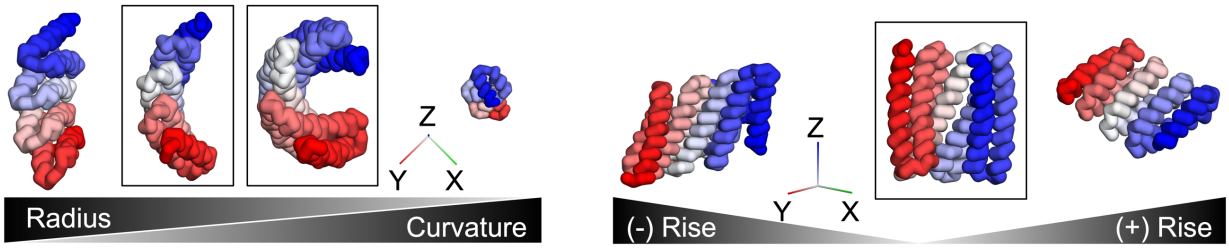


Figure 4. Super helical parameters control the shape of repeat proteins. The left side shows proteins with large to small radius and with small to large curvature. The right shows proteins from negative to positive rise. Proteins in boxes represent desired structures, having curvature and low rise. Proteins are depicted as ribbon backbones colored from blue (N-terminus) to red (C-terminus). The superhelical axis of each protein is aligned with the z-axis, and the orientation of the x, y, and z axes are shown. 33

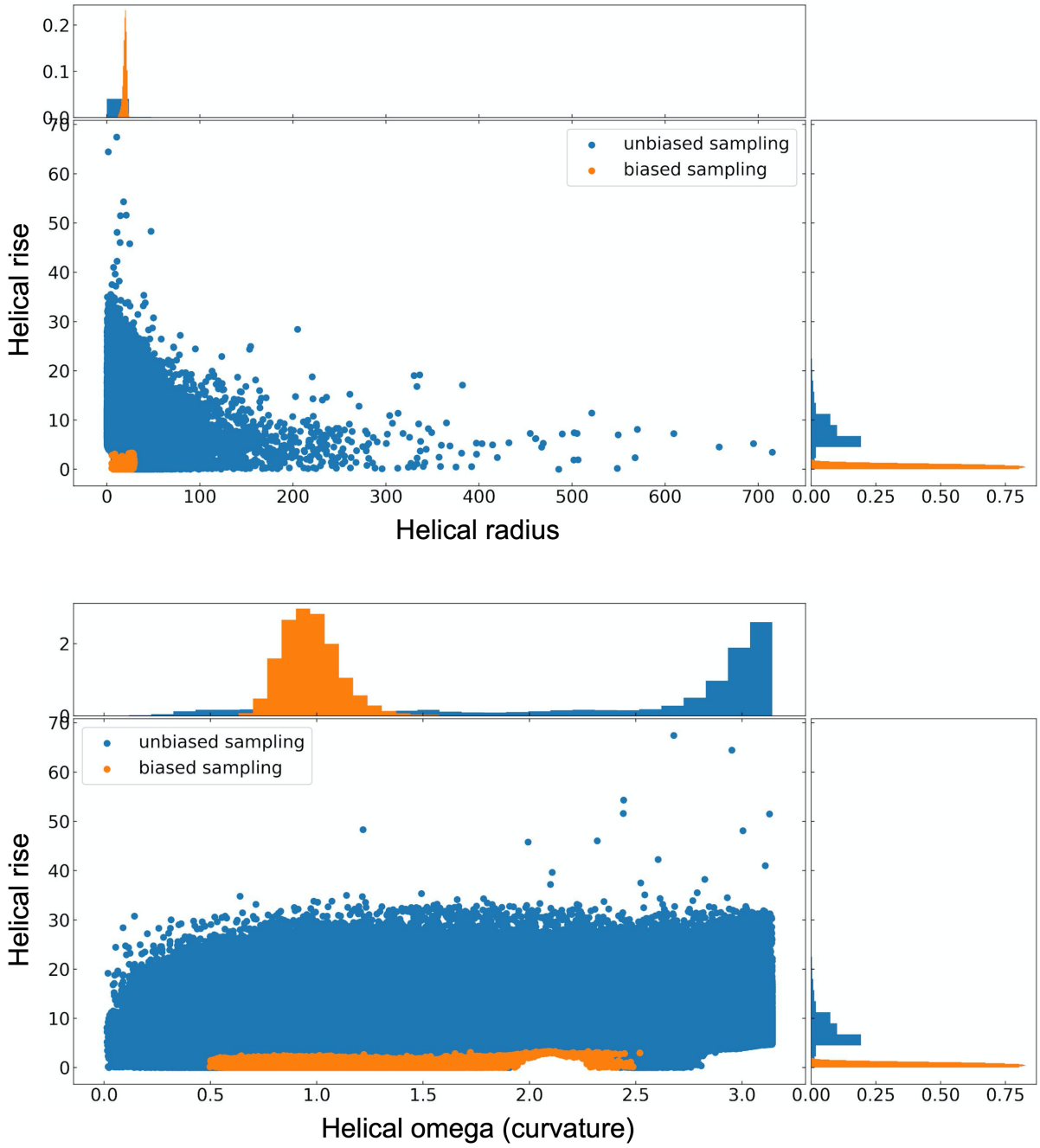


Figure 5. Scatter plot and associated histograms of helical rise vs helical radius (top) and helical rise vs helical omega (bottom) for 1 million trajectories without biased sampling (blue) or with biased sampling (orange).

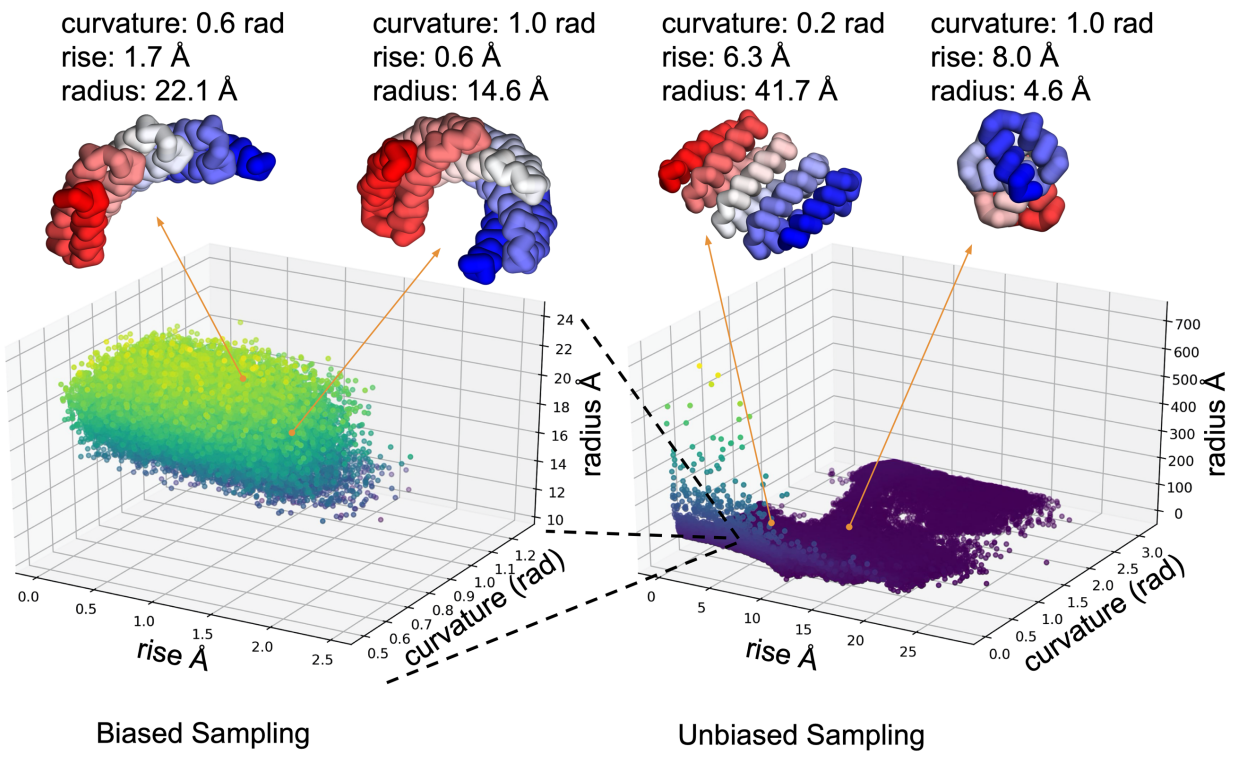


Figure 6. Three-dimensional landscape of repeat protein space for rise, radius, and omega (curvature). 1 million trajectories for repeat proteins generated using biased sampling (left) or unbiased sampling (right) plotted on a three-dimensional grid for rise, radius, and omega (curvature). Points are colored according to their radius to help visualize the three-dimensional landscape. Two proteins are shown from each of the biased and unbiased trajectories along with their helical parameters. 35

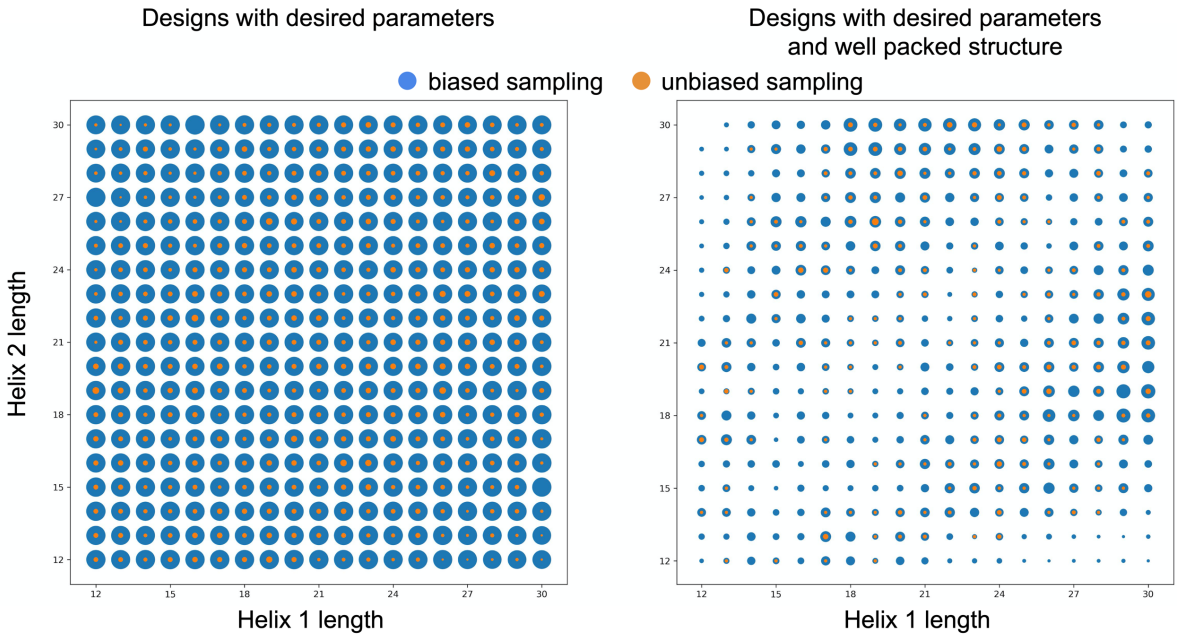


Figure 7. Scatter plot of the number of designs with desired parameters with biased sampling (blue) or without biased sampling (orange) according to the length of helix 1 and helix 2. The size of each circle is proportional to the number of designs with desired parameters (left) and being well packed (right). 1 million trajectories were attempted with biased sampling and without biased sampling. Sampling was spread out equally across all helix 1 by helix 2 combinations. 36

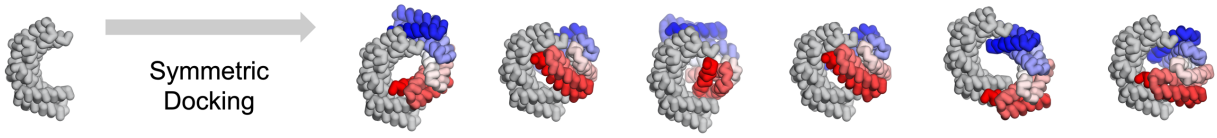


Figure 8. Symmetric docking creates diverse homodimers from a single monomer. 37

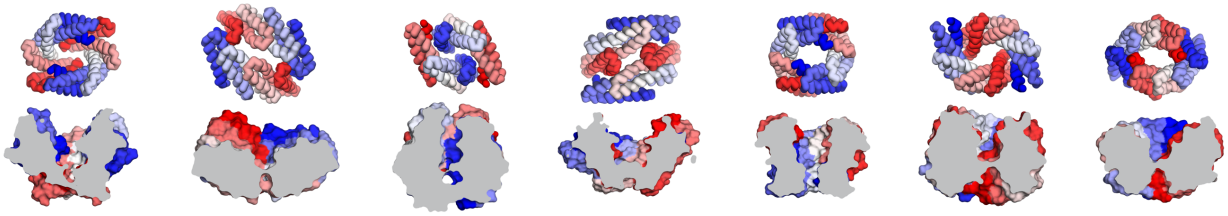


Figure 9. Diverse C2 symmetric homodimers that were generated from different monomers. The diversity of size and shape available to the central cavity is shown in a top-down view (top) and in a side view cutting through the protein (bottom). 38

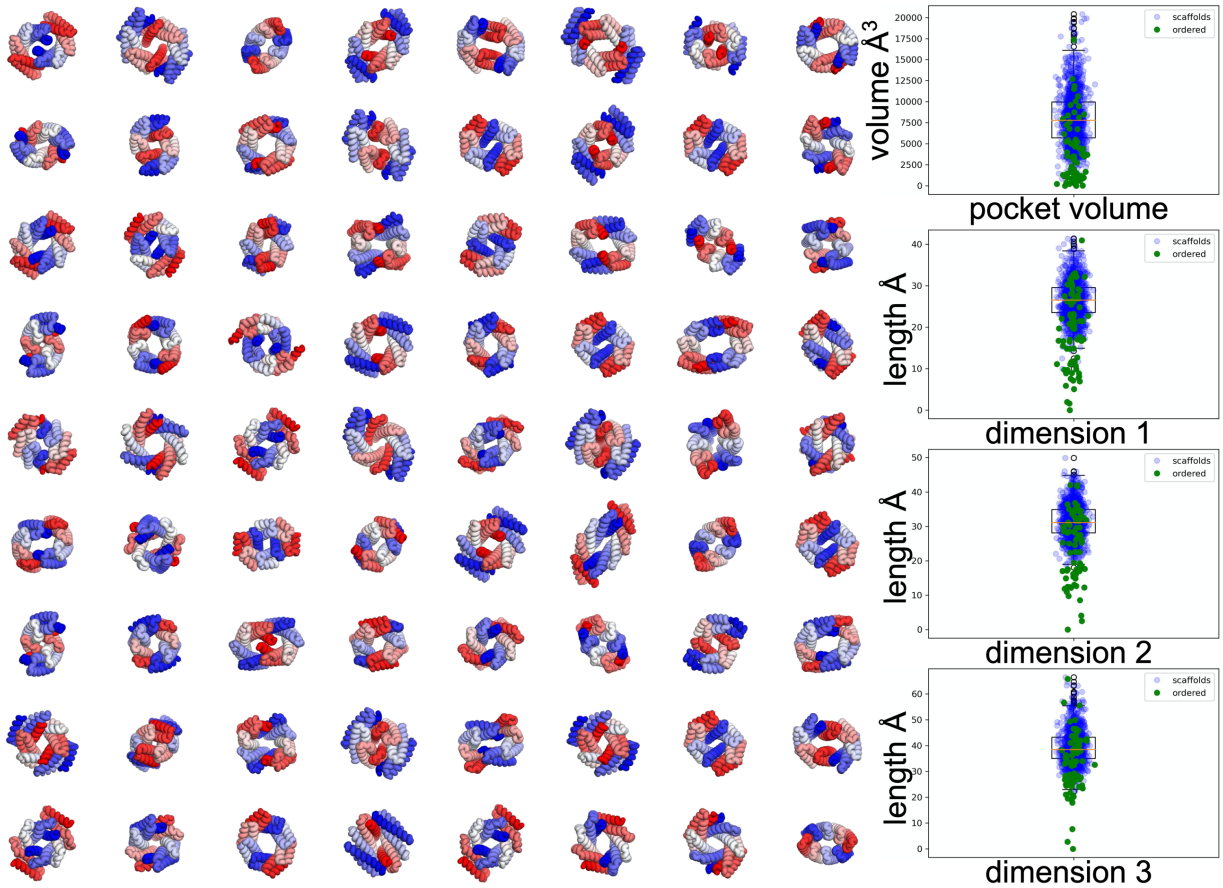
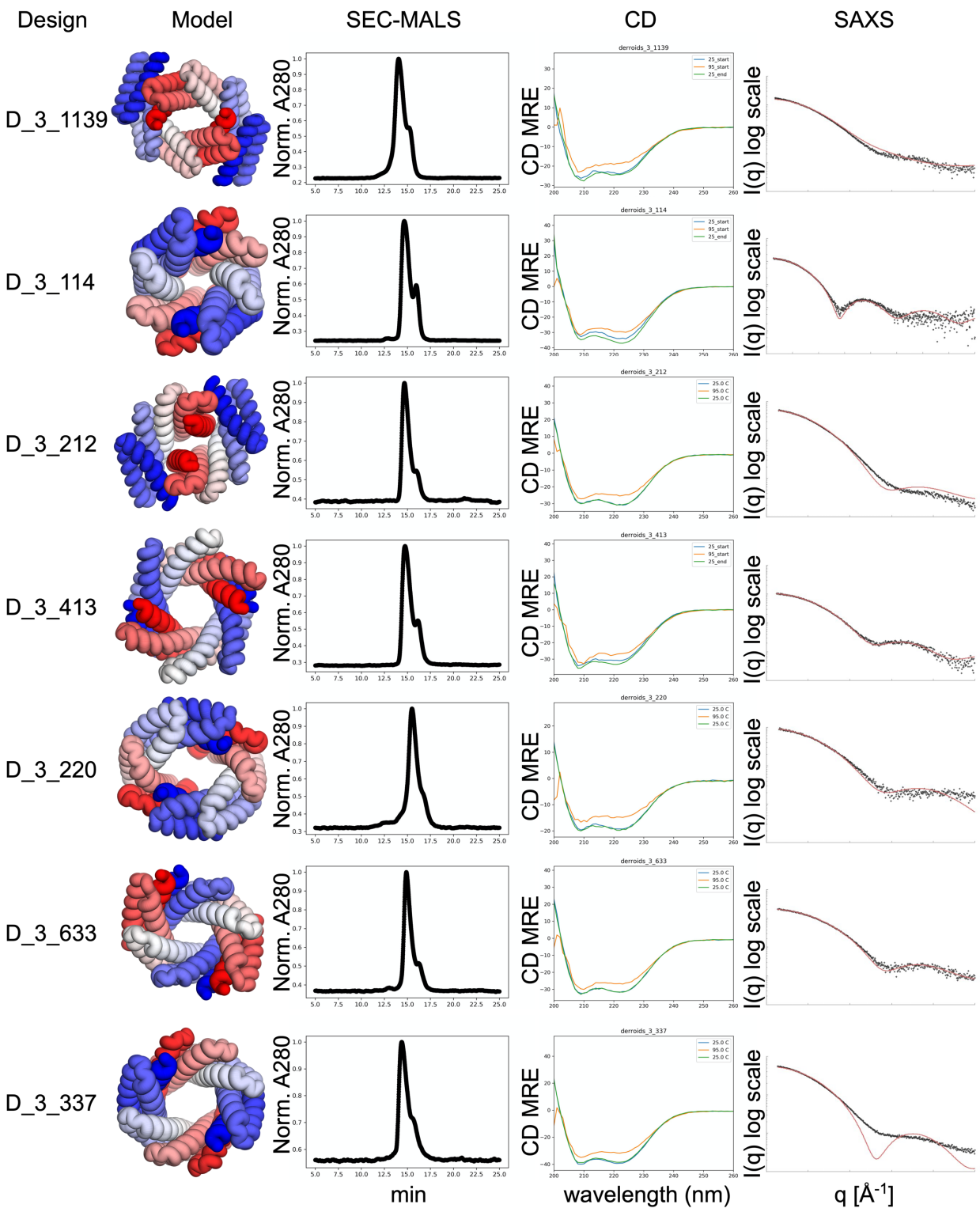


Figure 10. Designs span a diverse range of sizes, shapes, and pocket features. On the left are 72 ordered designs depicted as ribbon backbones colored from blue (N-terminus) to red (C-terminus). The right side shows boxplots plus points for four pocket features, volume, dimension 1, dimension 2, and dimension 3 for the top one thousand designs from generation 3 (boxplot and blue points) along with all ordered scaffold designs (green points). 41


Design
Model
SEC-MALS
CD
SAXS
D_3_1139
D_3_114
D_3_212
D_3_413
D_3_220
D_3_633
D_3_337
Norm. A280
CD MRE
I(q) log scale
derroids_3_1139
derroids_3_114
derroids_3_212
derroids_3_413
derroids_3_220
derroids_3_633
derroids_3_337
25_start
95_start
25_end
25.0 C
95.0 C
min
wavelength (nm)
q [Å-1]

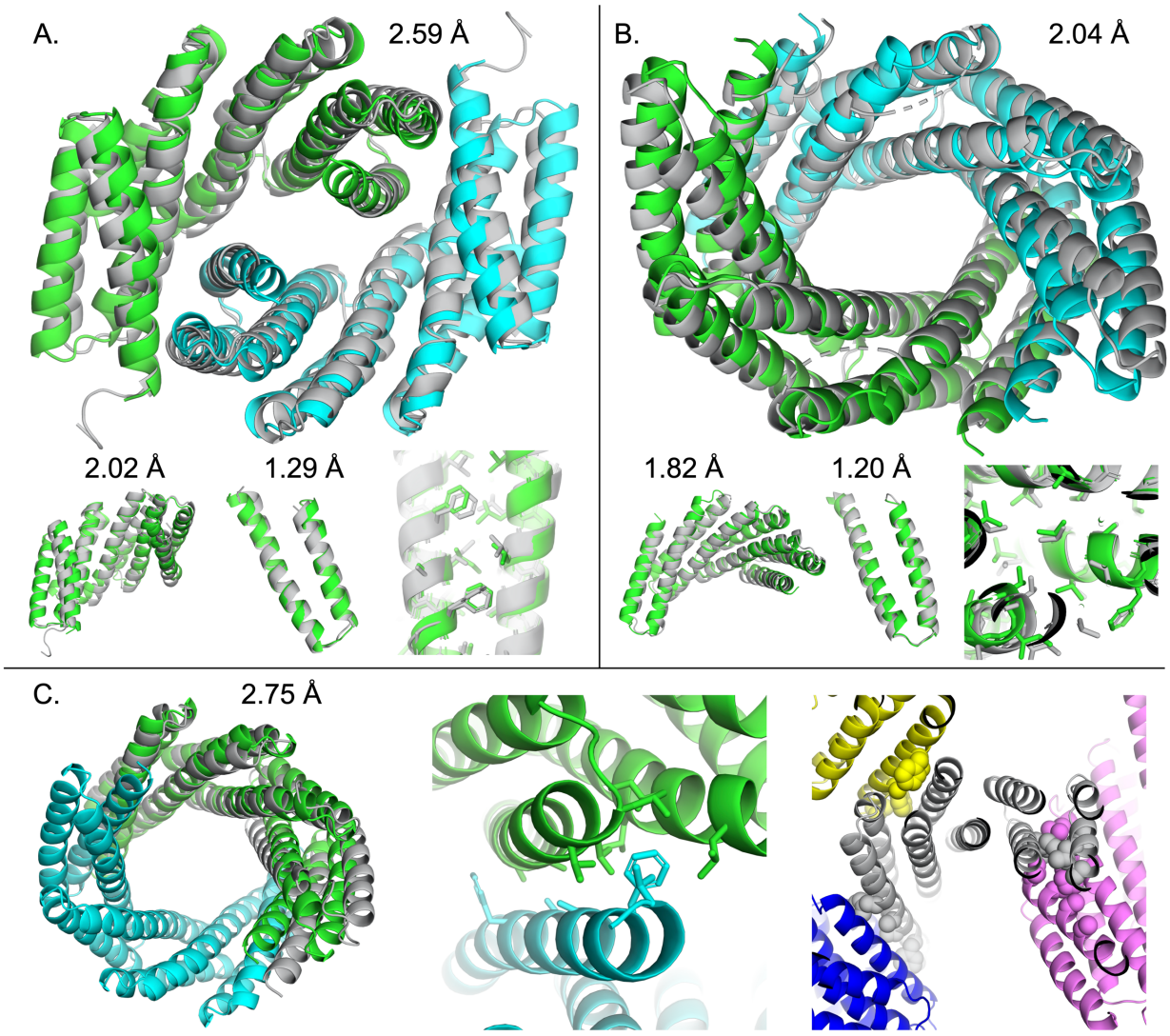


Figure 12. Crystallographic analysis of two successful designs and one failed design. Panel A shows an overlay of design D_3_212 (green and cyan) with its crystal structure (gray). The top portion shows the superposition of the homodimer, while the lower portion shows the superposition of the monomer, a repeat unit, and a section of the hydrophobic core. Associated rmsds are shown. Panel B shows an overlay of design D_3_633 (green and cyan) with its crystal structure (gray). The top portion shows the superposition of the homodimer, while the lower portion shows the superposition of the monomer, a repeat unit, and a section of the hydrophobic core. Associated rmsds are shown. Panel C shows design D_3_337 and its crystal structure. The left is an overlay of design D_3_337 (green and cyan) with its crystal structure (gray), which is monomeric, with associated rmsd. The middle shows the designed homodimer interface, with hydrophobic residues shown as sticks and the two chains colored green and cyan. On the right is the crystal structure showing the central asymmetric unit in gray and its crystal lattice neighbors colored blue, pink, and yellow. The hydrophobic residues which were intended to form the homodimer interface are shown in spheres forming key crystal contacts.

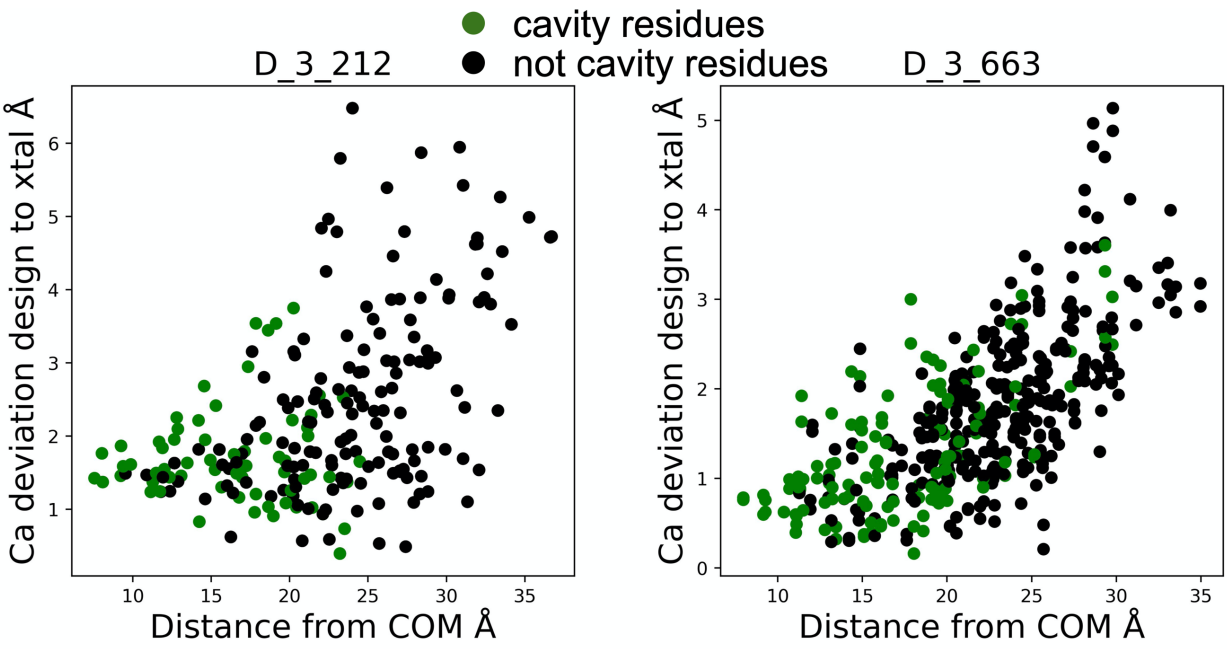
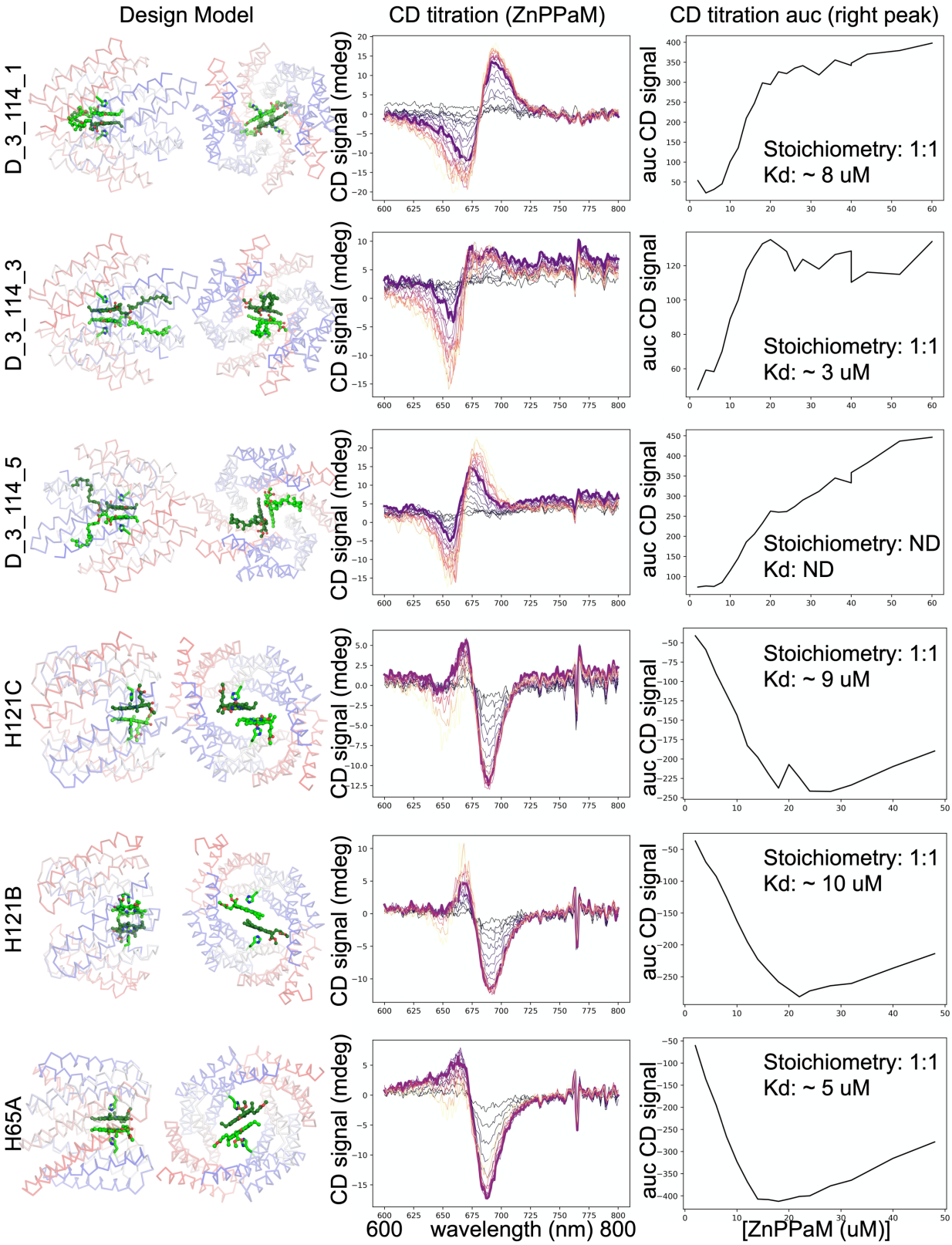
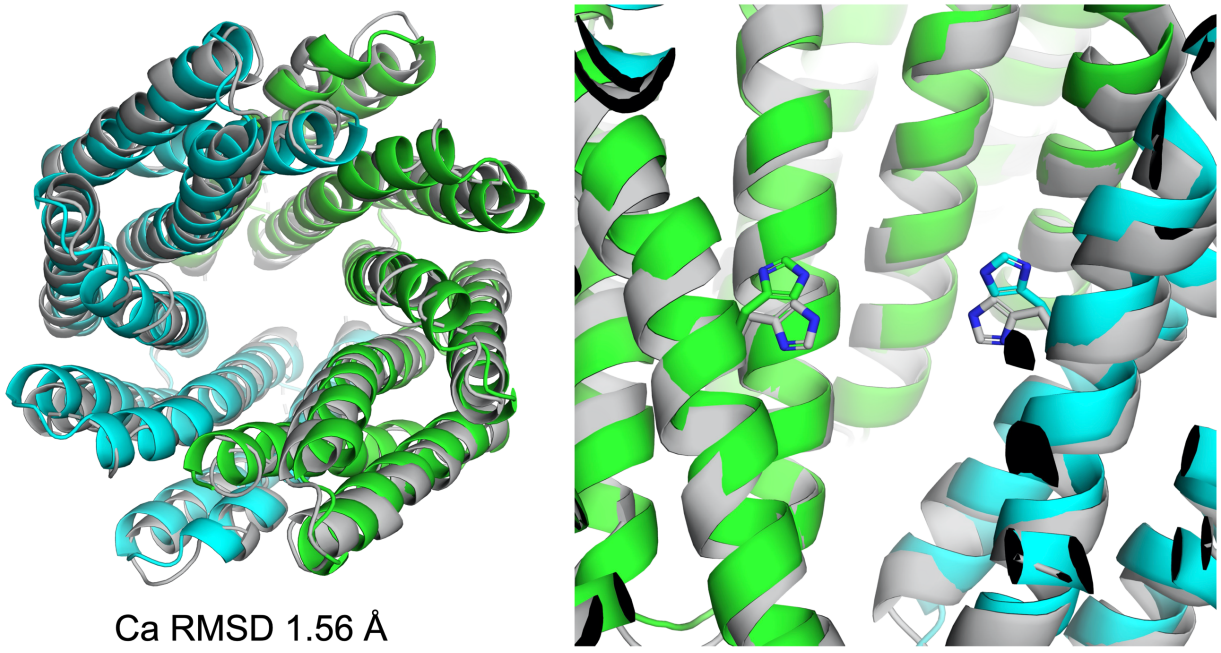


Figure 13. Scatter plot of distance from the center of mass vs Ca deviation. The plots show Ca deviations (y-axis) compared to the distance to the protein center of mass (x-axis) for design models compared to their respective crystal structures. Points for cavity lining residues are colored green, while points for the rest of the protein are colored black. The plot on the left is for design D_3_212 while the plot on the right is for design D_3_633.


Design Model
CD titration (ZnPPaM)
CD titration auc (right peak)
D_3_114_1
D_3_114_3
D_3_114_5
H121C
H121B
H65A
CD signal (mdeg)
auc CD signal
Stoichiometry: 1:1
Kd: ~ 8 uM
Stoichiometry: 1:1
Kd: ~ 3 uM
Stoichiometry: ND
Kd: ND
Stoichiometry: 1:1
Kd: ~ 9 uM
Stoichiometry: 1:1
Kd: ~ 10 uM
Stoichiometry: 1:1
Kd: ~ 5 uM
600
wavelength (nm) 800
[ZnPPaM (uM)]



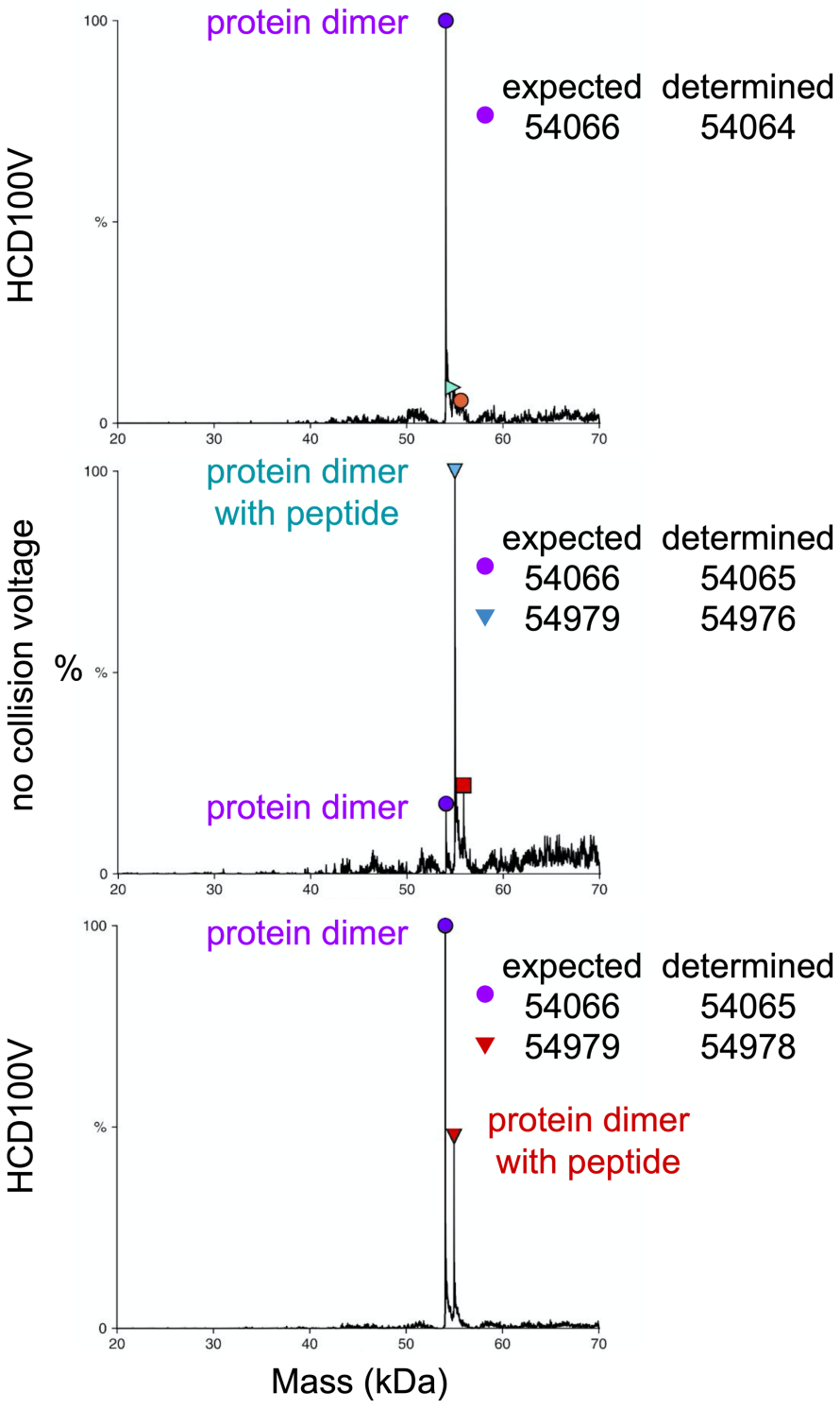

protein dimer
expected determined
54066 54064
HCD100V
protein dimer
with peptide
expected determined
54066 54065
54979 54976
no collision voltage
%
protein dimer
protein dimer
expected determined
54066 54065
54979 54978
protein dimer
with peptide
HCD100V
100
%
0
20
30
40
50
60
70
Mass (kDa)

Figure 16. Native mass spectrometry of protein-peptide complex. Design D_3_633_x8 was subjected to nMS analysis without peptide, top panel, and with peptide at 10X protein concentration, bottom two panels. The top and bottom panels show samples analyzed with all-ion fragmentation (MSMS) mode with high energy collision-induced dissociation (HCD) 100 V. The middle panel was analyzed with full MS mode (no collision voltage applied). For each panel, only deconvoluted spectra are shown. 63

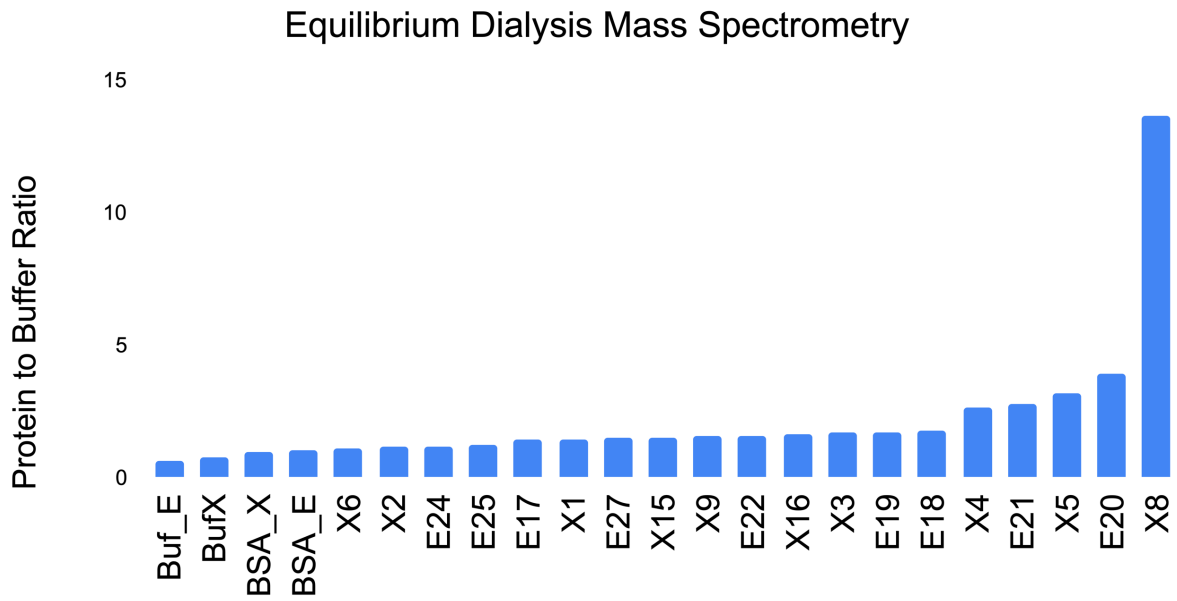
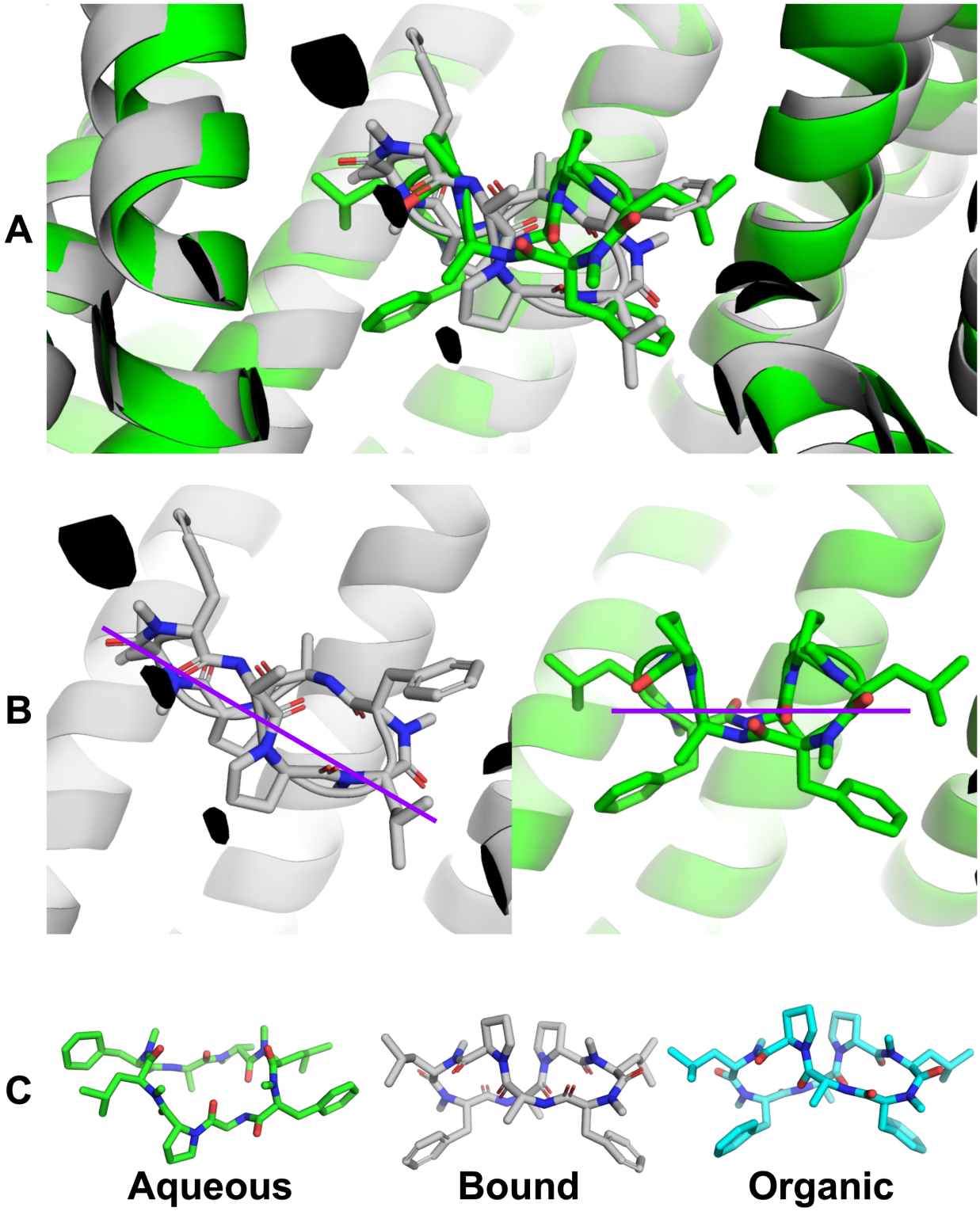


Figure 17. Equilibrium dialysis mass spectrometry of peptide binder designs. The leftmost sample is the peptide and buffer without protein. The next two samples use BSA as the protein as controls of nonspecific binding. The rest of the samples represent binder designs. Each sample name contains an E, for enantiomer, or X, for the original peptide. Because the peptide contains L and D amino acids, we used the mirror image enantiomer during design and characterization, along with the originally designed peptide. 63


A
B
C
Aqueous
Bound
Organic

ACKNOWLEDGEMENTS

To start with, I have to thank David Baker, not only for his mentorship over the last 5 years but also for his leadership in the field of protein design as a whole. For without David this field may still be trying to make airplanes by modifying owls.

It may not seem like it, but the Baker lab is fairly large, and as labs grow in size, certain problems tend to arise. I speculate that a major problem with a large lab is keeping everyone connected and collaborating. To deal with this, David has put a tremendous amount of effort into the social engineering of the lab by designing multiple weekly events for everyone to get together and just talk. This has been critical because problems are not solved well in isolation. A problem that may appear difficult for one person may be easily solved by someone else, so it is really important with a lab this size to keep everyone communicating and collaborating, and I very much appreciate the time and effort David spends making this lab feel smaller than it is. I'd like to give a special shoutout to everyone over the years that help keep this place running, and particularly everyone that has helped manage how things operate during this ongoing global pandemic.

A special thanks to the folks at the Institute for Protein Design protein production core including Michelle DeWitt and Cameron Chow for protein production and purification. Thanks to Lauren Carter for managing the IPD core, and for SEC-MALS experiments. Thanks to Alex Kang and Asim Bera for in house crystallography work. Thanks to Xinting Li for mass spectrometry. Thanks to Michelle Matsunaga, Zari Magness, Kandise Vanwormer, Austin Smith, Mike Murphy, Celine Abell, Luki Goldschmidt, Ian Haydon, Ratika Krishnamurty, Tina Nguyen, Patrick Vecchiato, Lance Stewart, and everyone else that helps make Institute for Protein Design work behind the scenes.

I'd like to thank Daniel Silva who was my main mentor in the lab during my first few years in graduate school. Daniel taught me the fundamentals of computational protein design as well as wet lab biochemistry and generally how to think about and conduct groundbreaking science. I'd also like to thank TJ Brunette, who mentored me during the initial stages of my thesis work; Benjamin Bastanta who I sat next to for years and helped me improve my computational skills and approach to designing proteins; Nate Ennist and Ralph Cacho for our collaboration on chlorophyll binding proteins; Patrick Salveson and Meerit Said for collaborating with me on peptide binders; Adam Moyer for collaborating on chlorophyll and peptide binding projects; Brian Coventry for help with developing so much code.

Mentoring people is pretty scary in my opinion, because mentees rely on mentors for their growth and success and it can feel like a huge waste of everyone's time when things to do not go well, so a super special thanks to the people that I had the opportunity to formally mentor over the years, Jeremiah Sims, Kirsten Thompson, and Isabelle Moczygemba. Despite being scared going into these, I feel like this was some of the best time I have had in science, and I think things went well for everyone, and that is largely due to the time and effort and commitment that they made to learning and working.

I would like to thank everyone at the SIBYLS beamline for all the SAXS data which has been my most important mid throughput characterization technique. SAXS Data was collected at the SIBYLS beamline which is funded by the DOE BER IDAT grant (DE-AC02-05CH11231) and NIGMS supported ALS-ENABLE (GM124169-01). This work was conducted at the Advanced Light Source (ALS), a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the Department of Energy, Office of Basic Energy Sciences, through DOE BER IDAT grant (DE-AC02-05CH11231) and NIGMS supported ALS-ENABLE (GM124169-01) and

National Institute of Health project MINOS (R01GM105404). I thank the staff at the Advanced Light Source SIBYLS beamline at Lawrence Berkeley National Laboratory, including K. Burnett, G. Hura, M. Hammel, J. Tanamachi, and J. Tainer for the services provided through the mail-in SAXS program, which is supported by the DOE Office of Biological and Environmental Research Integrated Diffraction Analysis program DOE BER IDAT grant (DE-AC02-05CH11231) and NIGMS supported ALS-ENABLE (GM124169-01) and National Institute of Health project MINOS (R01GM105404).

Thanks to the Wysocki lab for native mass spectrometry. Thanks to Madison Kennedy, Lindsey Doyle, and Barry Stoddard for our wonderful crystallography collaboration. I very much enjoyed visiting your lab and talking about my proteins.

A big thanks to the Biochemistry department and Molecular & Cellular Biology Graduate Programs which I have been a part of for the last 5+ years. Thanks to my committee members Douglas Fowler, Justin Kollman, Jesse Zalatan, and Phil Bradley for helping to guide me and keep me on track.

A very special thanks to the whole Rosetta community, which appears to be approaching 100 different labs spread across the globe, for all the scientific and non-scientific work that goes into developing and managing such a large computational project and diverse community. None of my research would be possible without this special community. It was very sad that we could not hold RosettaCon2020 in person this year!

Finally, thanks to all my friends and family inside and outside the lab for support and the many fun adventures over these historic last several years.

DEDICATION

I would like to dedicate this work to the community colleges across the USA, and Sierra College in particular, that give struggling working-class families the opportunity for college educations that would otherwise be unaffordable. Additionally, I would like to dedicate this work to the particular professors at Sierra College that helped rekindle a scientific curiosity in me that had largely been lost in my rebellious teenage years: to Michael Brelle who taught me general chemistry and analytical chemistry and first got me interested in science; to Mark Springsteel who taught me organic chemistry which had a profound impact on my understanding of the reality of our universe as well as how to study and learn in general; to Shawna Martinez who taught me botany and got me curious about the amazing nature that surrounds us daily; to Harriet Wilson who taught a grueling microbiology course that prepared me well for future classes and research; and to Charles Dailey who taught Zoology in addition to leading the science club and natural history museum at Sierra College and who was truly an amazing advocate for science education.

Chapter 1. INTRODUCTION TO PROTEINS AND PROTEIN DESIGN

Natural proteins are large biomolecules that are encoded as genes in DNA. Some famous proteins include keratin which makes hair and nails, albumin found in egg whites, hemoglobin (Muirhead and Perutz, 1963) that shuttles oxygen around the body, and antibodies that fight off infections. Proteins are produced in cells through transcription of DNA into RNA and then translation of RNA into protein. This process produces a long chain of amino acids or the primary structure of the protein. These chains of amino acids fold into local secondary structures including sheets, helices, and loops (Pauling and Corey, 1951; Pauling, Corey and Branson, 1951), which then compact into distinct 3 dimensional (tertiary) structures (Creighton, 1990) in order to hide hydrophobic (oil-like) amino acids from water (Tanford, 1978; Bryngelson *et al.*, 1995). The particular structure that a protein folds into is determined by its amino acid sequence (Anfinsen, 1973). The structure of a protein gives rise to its particular machine-like function (McLachlan, 1972), much like the structure of everyday machines that we interact with, a printing press, combustion engine, or airplane give rise to their functions. The vast array of functions which proteins perform include breaking down or building up other molecules (Schomburg and Salzmann, 1991), replicating DNA (Benkovic, Valentine and Salinas, 2001), transporting molecules into or out of cells (André, 1995), sending communication within or between cells (Hamm, 1998), and acting as scaffolding that gives structure and organization to cells (Good, Zalatan and Lim, 2011; Lynch *et al.*, 2017).

Because a protein's structure gives rise to its function, it has been essential to figure out the structure of proteins in order to learn how they work, such as early structural studies on hemoglobin (Perutz, 1978). This has traditionally been done through wet-lab experimental techniques.

However, the abundance of structural data available in the last two decades (Berman *et al.*, 2000) has allowed the development of sophisticated computational approaches that can now predict the structure of proteins from the primary sequence (Ovchinnikov *et al.*, 2017; Kryshtafovych *et al.*, 2019). These computational approaches have given rise to the field of computational protein design. By essentially running the structure prediction algorithms in reverse, we can now create novel proteins that do not exist in nature (Street and Mayo, 1999; B. Kuhlman *et al.*, 2003; Kuhlman and Bradley, 2019). The field of protein design aims to create new protein structures that can perform new and useful functions that have not arisen biologically over billions of years of evolution so that we can solve modern-day challenges that humanity faces.

1.1 The Challenges of Computational Protein Design

Computational protein design involves solving, or working around, two fundamentally difficult problems. The first problem involves developing a scoring method that can predict, to some degree, the likelihood that a particular sequence will fold to the desired structure (Jones, 1994; Schramm *et al.*, 2012; Park *et al.*, 2016). There are many approaches to scoring including statistical potentials, physics-based models, or various combinations of the two. Difficult tradeoffs must be made between the speed of calculating a score and the accuracy of a score. The second requirement is the ability to search protein space including backbone conformation, sequence identity, and rigid body orientation (quaternary structure) for proteins involving more than one polypeptide chain. This is a particularly interesting and challenging problem given the enormous size of these spaces. Consider, for example, a 100 aa protein has 99 phi and 99 psi backbone torsion angles, each of which can rotate 360°. Sampling just 3 discrete angles for each of these yields

3^198 backbone conformations, a number much larger than the number of particles in the known universe. How a protein folds to a well-defined 3-dimensional structure despite this massive conformational search space is often referred to as Levinthal's paradox (*Levinthal's Paradox*, no date). The computational search through this space for useful confirmations is just as daunting, and **the first part of my thesis research was devoted to improving our ability to search conformational space** through the introduction of nonphysical score terms capable of controlling the shape of proteins during backbone sampling.

Protein sequence space is also enormous. There are 20 standard amino acids (Bywater, 2018), so for a 100 aa protein, there are 20^100 possible sequences. Again, more than the number of particles in the known universe. To make this search more difficult, most amino acids have multiple confirmations available to them (Dunbrack and Karplus, 1994). Luckily, or unluckily, for the protein designer, most of the conformational space and the sequence space is nonsense and can essentially be ignored, leaving a much smaller but still enormous space to search through for novel proteins and functions (see figure 1 for my depiction of protein space). Finally, protein rigid body space is also quite large. For two proteins that interact asymmetrically, there are 3 translational degrees of freedom and 3 rotational degrees of freedom. Allowing just 5 Å translations sampled every 0.5Å, and 360° rotations sampled every 36° gives rise to a search space of 10^6. Ideally, for the accurate design of protein interactions, these degrees of freedom should be sampled with much higher precision, which begins to push or exceed the limits of what modern computers are capable of. To decrease the search space for sequence and rigid body orientation, I have utilized symmetry (Goodsell and Olson, 2000) where appropriate throughout my projects. That we can search through protein space and score accurately enough and fast enough to create even simple de novo proteins is quite remarkable in my opinion!

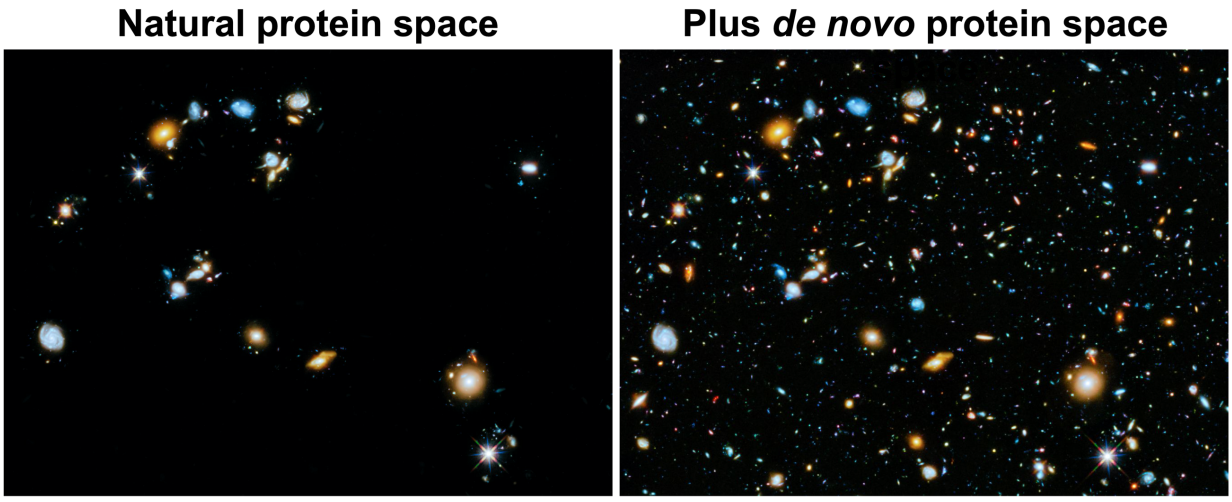


Figure 1. De novo protein design unlocks a vast number of novel proteins. Stars, galaxies, and empty space is an analogy for proteins, folds, and nonsense protein space (*Hubble Ultra Deep Field 2014*, no date).

1.2 The Promise of Computational Protein Design

While computational protein design is a relatively young field, the first fully de novo protein was created in 1997 (Dahiyat and Mayo, 1997), and the first fully de novo protein fold was created in 2003, remarkable success has already been realized. At the smallest scale, there has been some success in designing peptides, which can be synthesized to include non-canonical amino acids (Bhardwaj *et al.*, 2016; Dang *et al.*, 2017). Slightly larger miniproteins have been designed that can be screened in the tens of thousands via high throughput assays for stability and binding (Sun *et al.*, 2016; Chevalier *et al.*, 2017; Rocklin *et al.*, 2017; Cao *et al.*, 2020). Much success has been realized designing alpha-helical bundle proteins through parametric design or fragment assembly some of which include extensive buried polar networks analogous to DNA (Offer, Hicks and Woolfson, 2002; Thomson *et al.*, 2014; Brunette *et al.*, 2015; Doyle *et al.*, 2015; Boyken *et*

al., 2016; Chen *et al.*, 2019). The successful design of membrane proteins has come out of the work on helical bundles that might be able to act as transporters for ions or molecules (Akerfeldt *et al.*, 1992; Goparaju *et al.*, 2016; Joh *et al.*, 2017; Duran and Meiler, 2018; Lu *et al.*, 2018). While still highly challenging, success has been seen towards designing protein binders for various small molecules (Bender *et al.*, 2007; Christopher Fry *et al.*, 2010; Cherny *et al.*, 2012; Tinberg *et al.*, 2013; Bick *et al.*, 2017; Dou *et al.*, 2018), and **the final part of my thesis research was devoted to designing ligand binders**. Similar work has been completed to design binders for metal ions (Calhoun *et al.*, 2003; Der *et al.*, 2012; Berwick *et al.*, 2014). Various natural and non-natural protein folds have been designed that create a rich space of new proteins waiting for functionalization (Brian Kuhlman *et al.*, 2003; Figueroa *et al.*, 2013; Rämisch *et al.*, 2014; Voet *et al.*, 2014; Lin *et al.*, 2015; Huang *et al.*, 2016; Marcos *et al.*, 2017, 2018; Basanta *et al.*, 2020), and **the central part of my thesis research was devoted to designing a large de novo fold family**. Enzymes are nature's most fascinating proteins, and while protein design has largely (or entirely) failed to create proteins as catalytically active as native enzymes, there has been some success, and these artificial protein catalysts can act as starting points for laboratory evolution capable of producing enzymes more comparable to those found in nature (Broo *et al.*, 1997; Zanghellini *et al.*, 2006; Jiang *et al.*, 2008; Röthlisberger *et al.*, 2008; Faiella *et al.*, 2009; Siegel *et al.*, 2010; Der, Edwards and Kuhlman, 2012; Giger *et al.*, 2013; Burton *et al.*, 2016; 'Design and evolution of enzymes for non-natural chemistry', 2017; Watkins *et al.*, 2017; Weitzner *et al.*, 2019). Finally, at the largest scale, protein design has allowed the creation of novel virus-like particles and other materials (King *et al.*, 2012; Lanci *et al.*, 2012; Gonen *et al.*, 2015; Bale *et al.*, 2016; Hsia *et al.*, 2016; Butterfield *et al.*, 2017; Ljubetič *et al.*, 2017; Shen *et al.*, 2018; Edwardson and Hilvert, 2019; Pyles *et al.*, 2019). However, protein design is still difficult, and the full promise

of protein design is still years away. In the future, one could imagine making specific and highly active enzymes for arbitrary chemical reactions to unlock novel green chemistries, personalized binding proteins capable of targeting drugs to cancer cells, materials capable of precise patterning of molecules, DNA sequencing proteins, biosensors for toxins or novel viruses, logic circuits to control engineered cells, and treatments for pandemics such as the ongoing SARS-CoV-2 pandemic (A. Spinelli, no date).

1.3 Characterization of De Novo Designed Proteins

The goal of protein characterization is to understand the structure, function, and other relevant properties of a protein. Since de novo proteins are fundamentally similar to natural proteins, they are characterized using many of the same standard methods. Proteins are often expressed recombinantly (synthetic genes put into a host organism) in Escherichia coli (‘Recombinant protein expression in Escherichia coli’, 1999) and then first purified by immobilized metal affinity chromatography (IMAC) with a fused histidine tag that binds nickel ions (Hochuli *et al.*, 1988). Failed designs will often not express or be insoluble, which can be determined through polyacrylamide gel electrophoresis (Smith, no date). In most cases, proteins are further purified by subsequent chromatography methods based on the protein's mass (size exclusion chromatography) or charge (ion exchange chromatography), which also act as initial characterization methods (Coskun, 2016). For example, a protein should elute on a sizing column in a predictable manner based on its size and oligomeric state. Many failed designs form soluble aggregates or other undesired oligomeric states.

After purification, proteins may be subjected to mass spectrometry to determine if the protein molecular weight (MW) is correct ('Protein mass spectrometry: applications to analytical biotechnology', 1995). Unfolded or partially unfolded designs will often be proteolyzed and have lower MW than expected. Circular dichroism is usually done to determine if the protein has the expected secondary structure and monitor how stable the protein is to increased temperature or chemical denaturant (Whitmore *et al.*, 2010). Size exclusion chromatography with multi-angle light scattering (SEC-MALS) is used to more accurately determine the size of the protein in solution (Sahin and Roberts, 2012). Poorly behaving designs will often show undesired oligomeric states. Small-angle X-ray scattering (SAXS) (Lipfert and Doniach, 2007) and native mass spectrometry (Leney and Heck, 2017) can yield additional information about a protein's size, shape and oligomeric state. The highest resolution characterizations usually involve electron microscopy (Danev, Yanagisawa and Kikkawa, 2019), nuclear magnetic resonance (NMR) (Bax and Clore, 2019), or x-ray crystallography (Ilari and Savino, 2008), which can be used to evaluate atomic level accuracy of designed proteins. Finally, more specialized assays are used to determine if proteins have desired functional properties such as binding, catalysis, or other functions.

Chapter 2. Design of C2 Symmetric Homodimers with Central Cavities

Cyclic two-fold (C2) symmetric ligands, such as the chlorophyll dimer special pair, are common in nature, synthetic chemistry, and medicine. Additionally, we can now design C2 symmetric peptides with drug-like properties, which could be useful in synthetic biology applications. De novo proteins capable of binding these C2 symmetric ligands could be useful in applications ranging from chemically inducible dimerization to synthetic light-harvesting. To create a diverse library of C2 symmetric binding pockets, we first designed curved repeat protein monomers sampling a continuum of curvatures, and then docked these into C2 homodimers, generating a very wide range of C2 cavity shapes and sizes for functionalization. 77 designs were experimentally characterized, and of these, the geometry of 23 (30%) were confirmed by SAXS, and 2 of these scaffolds were shown by crystallographic analyses to be in close agreement to their design models. We believe that these diverse scaffolds provide a rich set of starting points for binding a very wide range of C2 compounds. Figure 2 shows the architecture that we envisioned, which involved curved repeat proteins docked into homodimer configurations to create a central cavity along the symmetry axis suitable for ligand binding. Advantages of this conception are that the cavities can be very diverse in size, shape, and available sidechain chemistry, and as the protein hydrophobic core is separated from the pocket because the cavity lining residues are on the exterior of the monomers, functionalization to create binding interactions for specific compounds is unlikely to destabilize either the monomers or the dimer interface.

Figure 2. Schematic showing the overall design goal, from making curved repeat protein (left) to symmetric homodimers (center) to ligand dimer binders (right). The Color gradient represents the protein chain direction from N-terminus (blue) to C-terminus (red). Arbitrary C2 symmetric ligands are shown in grey.

2.1 Protein Architectures that Inspired this Work

I was initially inspired by work on alpha-helical toroid proteins (Doyle *et al.*, 2015) and C2 symmetric homodimer proteins made from repeat proteins (Fallas *et al.*, 2017). Figure 3 shows the particular proteins from these manuscripts that seemed well suited for binding small molecules or peptides due to their closed circular architecture and size of their central cavities. In order to bind a wide range of diverse C2 symmetric molecules, I believed it would be beneficial to create a large library of C2 symmetric protein homodimers containing central cavities of diverse size and shape that can be functionalized without compromising stability.

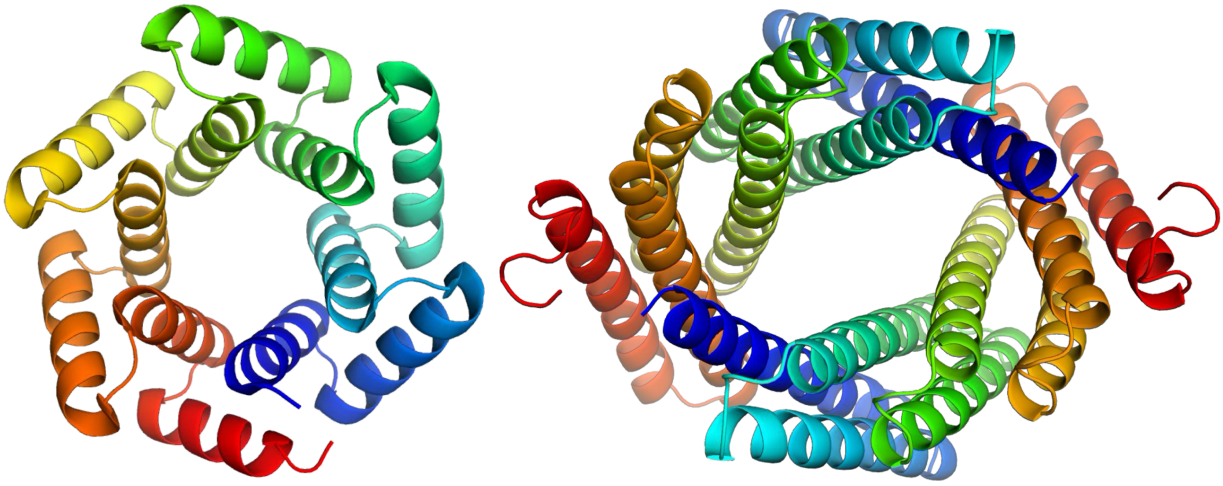


Figure 3. Original protein inspirations for this thesis. On the left is the monomeric crystal structure 4yxx (Doyle *et al.*, 2015) obtained from the Protein Data Bank (Berman *et al.*, 2000). On the right is the C2 symmetric design model TJ79C2 (Fallas *et al.*, 2017). The proteins are colored from blue (N-terminus) to red (C-terminus) with backbone cartoon representation. The images were produced using the molecular graphics program Pymol (*PyMOL*, no date).

2.2 Making a Library of Curved Repeat Proteins

In order to make a large diverse library of C2 symmetric protein homodimers containing central cavities, we first needed to create a diverse library of monomeric units to dock into various symmetric homodimer orientations. Previous work in the Baker lab had developed a pipeline to make diverse repeat proteins (Brunette *et al.*, 2015), but did not produce many proteins with the necessary shape, curved repeat proteins, to create the C2 symmetric homodimers with central cavities that I envisioned. Because of the lack of existing protein monomers, we began by generating a new set of helical repeat protein monomers with structures specifically tailored for building C2 symmetric binding pockets. We selected a range of the repeat superhelical parameters

curvature, rise, and radius that for a four repeat unit protein would approximate a half-circle, such that a dimer would approximate a full circle. Model building suggested a curvature between each repeat should be between 0.7 rad and 1.1 rad, a rise of less than 1.5 Å per repeat, and a radius between 10 Å and 22 Å. We hypothesized that when docked into dimers, these parameters would create pockets that could accommodate ligands of diverse sizes and shapes. Figure 4 illustrates how radius, curvature, and rise control the shape of repeat proteins and highlights the type of monomeric proteins we wanted to make.

Because previous research using Rosetta Monte Carlo random fragment assembly approaches to explore repeat protein space rarely sampled the range of helical parameters we were aiming for (Brunette *et al.*, 2015), we first needed to overcome this limitation. To do this, we developed methods for biasing fragment assembly towards desired regions of repeat protein parameter space; at each fragment insertion (made identically in each repeat unit) the deviation from a specified set of helical parameters (Hauser *et al.*, 2017) is computed, and the sum of these deviations is added to the coarse-grained score function used earlier. With this biased assembly protocol, we were able to focus sampling on repeat protein structures with the desired superhelical parameters (see Figure 5 and Figure 6).

During the design process, we discovered that the length of helices can control the shape of repeat proteins independent of our biased fragment assembly protocol, although to a much lesser extent. Having the length of helix 1 and helix 2 differ by 6-7 residues gives rise to curved repeat proteins at a low rate, whereas when the helices are the same length, curved repeat proteins are almost never sampled. However, with our biased fragment assembly protocol, we can achieve curved repeat proteins at all combinations of helix length. In fact, almost all trajectories with the biased fragment assembly protocol generate proteins with desired helical parameters (see figure

7). Unfortunately, many trajectories ended with unfolded or linear structures containing helices that do not pack well with each other. After filtering these poorly packed structures out, we still obtain far more curved repeat proteins at all helix combinations using our biased fragment assembly protocol (see figure 7). Obtaining a correct balance of scoring terms is critical, and our helical parameter terms are likely too strong, leading to structures that get stuck in poorly folded states so long as they satisfy the desired helical parameters. Future work should focus on properly optimizing these terms. Despite this possible limitation, we were able to generate one hundred thousand curved repeat protein backbones for subsequent design.

One hundred thousand backbones with desired helical parameters were subjected to sequence optimization using a RosettaScripts FastDesign protocol with repeat protein symmetry, which makes identical moves to each repeat unit during sequence design and minimization. Afterward, the designs were allowed to extend or shorten by up to half a repeat unit in order to have better-packed termini based on the score per residue of the terminal helix. This was done to eliminate terminal helices that make few contacts to the rest of the structure. The top twelve thousand designs based on a combination of score, packing, and sequence-structure agreement, were submitted for forward folding. Designs with a forward folding metric, the area to the left of the folding funnel from the lowest energy point to +8 rosetta energy units, of less than 25, which yielded two and a half thousand designs, were used in subsequent docking and design calculations.

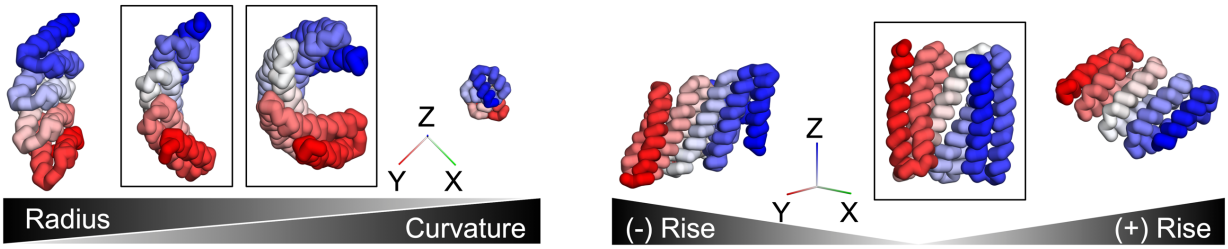


Figure 4. Super helical parameters control the shape of repeat proteins. The left side shows proteins with large to small radius and with small to large curvature. The right shows proteins from negative to positive rise. Proteins in boxes represent desired structures, having curvature and low rise. Proteins are depicted as ribbon backbones colored from blue (N-terminus) to red (C-terminus). The superhelical axis of each protein is aligned with the z-axis, and the orientation of the x, y, and z axes are shown.

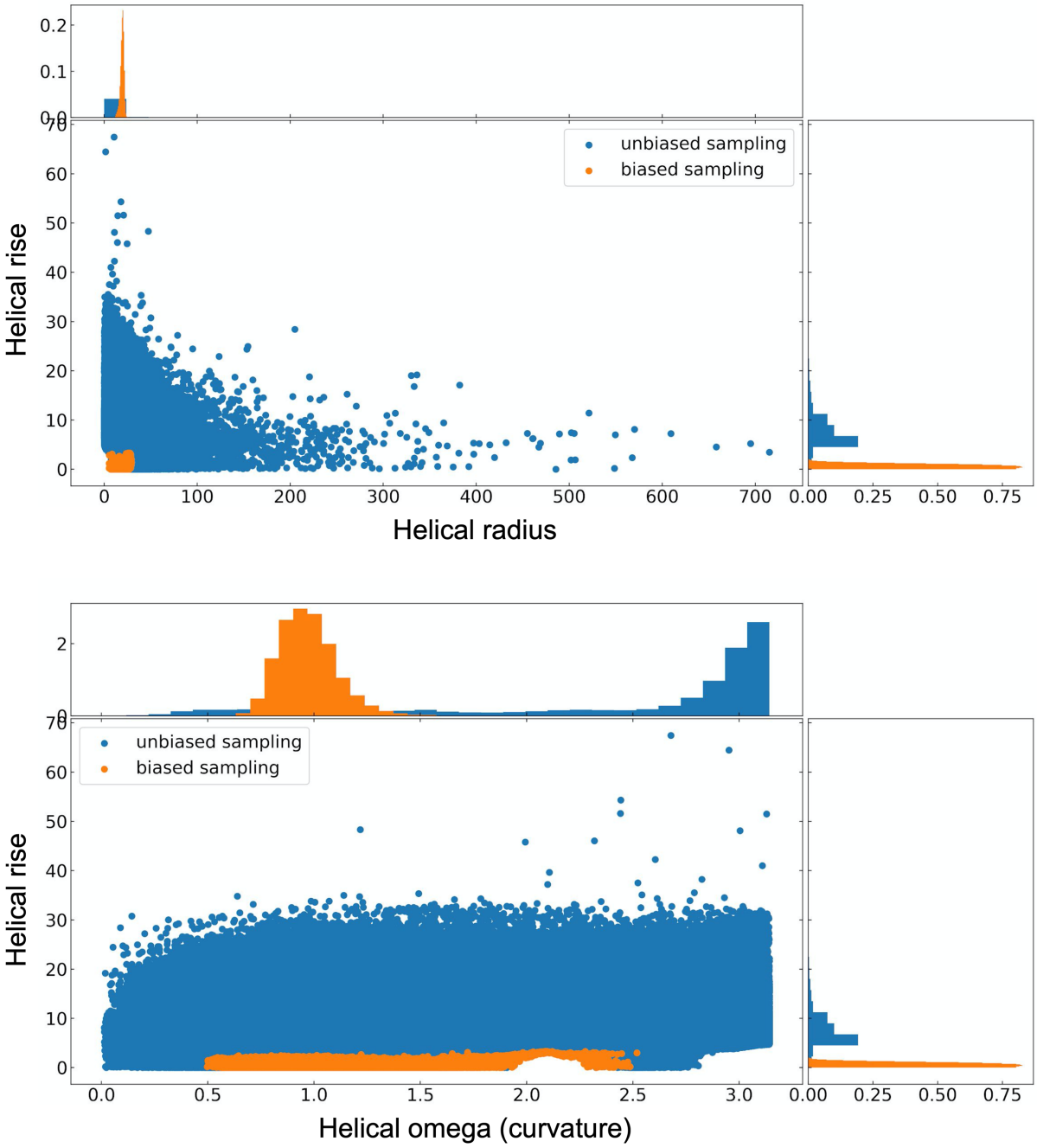


Figure 5. Scatter plot and associated histograms of helical rise vs helical radius (top) and helical rise vs helical omega (bottom) for 1 million trajectories without biased sampling (blue) or with biased sampling (orange).

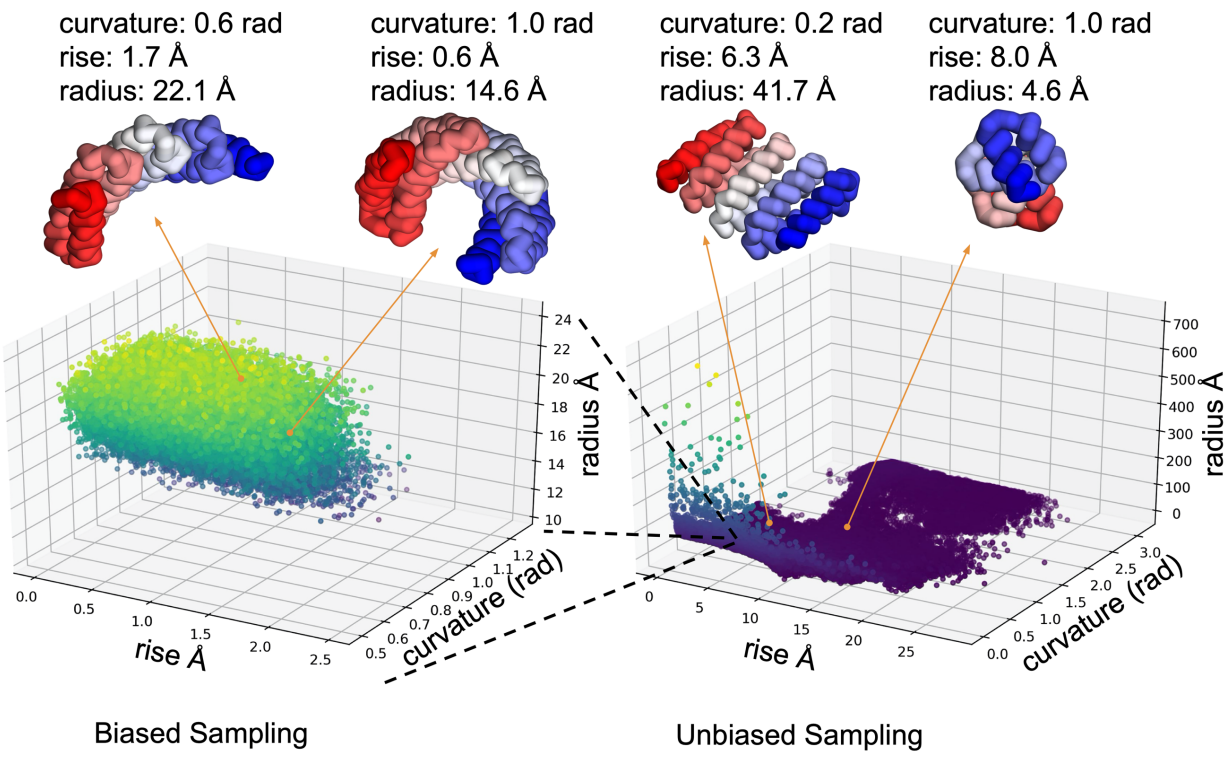


Figure 6. Three-dimensional landscape of repeat protein space for rise, radius, and omega (curvature). 1 million trajectories for repeat proteins generated using biased sampling (left) or unbiased sampling (right) plotted on a three-dimensional grid for rise, radius, and omega (curvature). Points are colored according to their radius to help visualize the three-dimensional landscape. Two proteins are shown from each of the biased and unbiased trajectories along with their helical parameters.

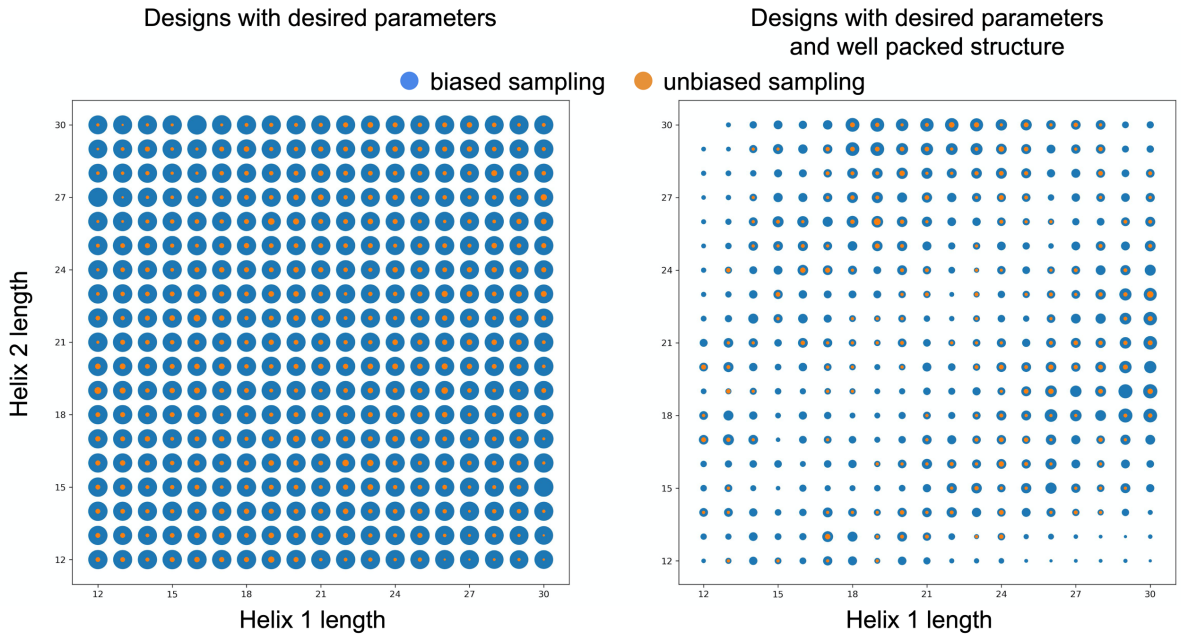


Figure 7. Scatter plot of the number of designs with desired parameters with biased sampling (blue) or without biased sampling (orange) according to the length of helix 1 and helix 2. The size of each circle is proportional to the number of designs with desired parameters (left) and being well packed (right). 1 million trajectories were attempted with biased sampling and without biased sampling. Sampling was spread out equally across all helix 1 by helix 2 combinations.

2.3 MAKING A LIBRARY OF C2 SYMMETRIC HOMODIMERS WITH CENTRAL CAVITIES

Using the two and a half thousand curved repeat proteins we had generated, we set out to create C2 symmetric homodimers with central cavities. We adapted a previous symmetric docking approach (Fallas et al. 2017) by adding a requirement that the N and C terminal helices of the monomers contact (at least one pair of residues on each terminus within 14 Å) in the dimer; this

leads to head to tail homodimers with a closed circular structure and a central cavity along the axis of symmetry. This docking protocol generated millions of docks. We subsequently removed docks with small interfaces (less than 10 contacting residues) and excessively large interfaces (greater than 24 contacts) which we thought might be poorly behaved due to the likelihood of having large surface hydrophobic interfaces before dimerization. This yielded a set of about one hundred thousand docks that we subjected to interface sequence optimization using a RosettaScripts FastDesign protocol with C2 symmetry. Figure 8 shows how a single monomer can be docked into many distinct orientations creating diverse central cavities. Figure 9 shows an example of the diversity of proteins and pockets that can be achieved by docking diverse monomers into various symmetric orientations.

The top one thousand designs were selected based on a combination of interface score, interface shape complementarity, and buried unsatisfied hydrogen bonds. Designs were randomly selected from among these for experimental characterization. Three generations of designs were tested in total, with the major difference among these generations being an overall decrease in the net charge of the proteins, in an attempt to improve crystallizability.

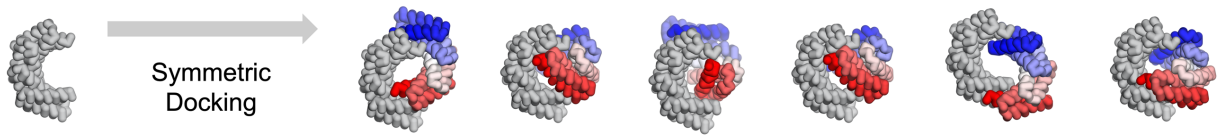


Figure 8. Symmetric docking creates diverse homodimers from a single monomer.

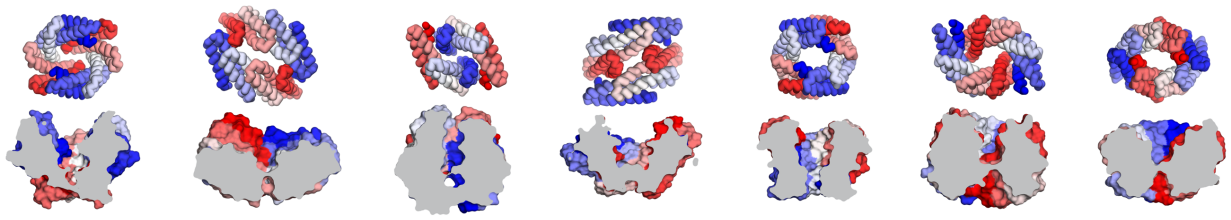


Figure 9. Diverse C2 symmetric homodimers that were generated from different monomers. The diversity of size and shape available to the central cavity is shown in a top-down view (top) and in a side view cutting through the protein (bottom).

2.4 Experimental Characterization of Proteins

Now with the ability to generate these proteins computationally, we set out to characterize a diverse set of examples spanning a range of pocket features (see Figure 10). Pocket features that we calculated include the volume, and three dimensions corresponding to the three principal axes of rotation. Figure 11 shows representative biochemical data for six successful designs and one failed design from the third generation of homodimers. Except for the failed design, Size Exclusion Chromatography coupled with Multi-Angle Light Scattering suggests that the proteins are majority monodisperse and the correct oligomerization state. Some of the proteins have a smaller secondary peak on SEC, which are likely monomer units of the design. Circular dichroism indicates that the designs are well folded helical proteins with great thermal stability. Of 17 proteins characterized by CD, only one was found to not be helical, and it was poorly expressed, had low solubility, and looked like aggregate by SAXS. Of the remaining 16 proteins, all but one appeared to remain folded at 95°C and all had nearly identical CD spectrums upon cooling back to 25°C. Finally, experimental Small Angle X-ray Scattering profiles obtained from our proteins closely matched profiles predicted for the corresponding design models, suggesting that the designed proteins are

likely the correct shape. In total, 30% of designs (23 of 77) show experimental scattering profiles that closely match predicted profiles based on their design models. The failed design, D_3_337, seems stable and helical by CD as expected. However, SEC-MALS shows this protein may be trimeric in solution. SAXS also indicates that this protein is larger than the designed dimer, potentially a trimer or larger soluble aggregate.

Figure 12 shows crystal structures for two successful designs. While there are some deviations between the experimentally determined structures and design models, particularly near the termini, the designs fold to the desired protein architecture with central cavities along the axis of symmetry. Figure 12 also shows a crystal structure we obtained for a design that failed to form a homodimer. Despite the failure to form the designed protein-protein interface, the monomer appears fairly accurate, with an rmsd of 2.75 Å, demonstrating our ability to control the shape of repeat proteins. While crystallographic analysis demonstrates D_3_337 is a monomer, SEC-MALS and SAXS indicate it may be a timer or larger soluble aggregate in solution. In the crystal structure, we see that the crystal lattice contacts are formed from the hydrophobic residues intended to form the homodimer interface. It is possible that the protein in solution forms a complex similar to the crystal lattice structure. Visual inspection of the designed interface of D_3_337 shows that it is poorly packed with only a single small cluster of hydrophobic residues (see the middle portion of panel C in figure 12).

One of the initial reasons we wished to make scaffold proteins for functionalization in the shape of ellipsoids was that we believed the circular shape would minimize deviations arising from local inaccuracies in the model by preventing lever-arm effects. We believed this would be true because deviations that would propagate and break the protein-protein interface would be unfavorable, leading to compensatory deviations that maintain the overall circular shape. Two of

our three crystal structures appear to support this idea as small local deviations do not break the overall fold. Extending from this idea, we also believed that the central cavity intended for functionalization would have the smallest deviations in the protein, and this idea appears to be supported by our data (see figure 13), corroborating the use of these proteins for binding applications.

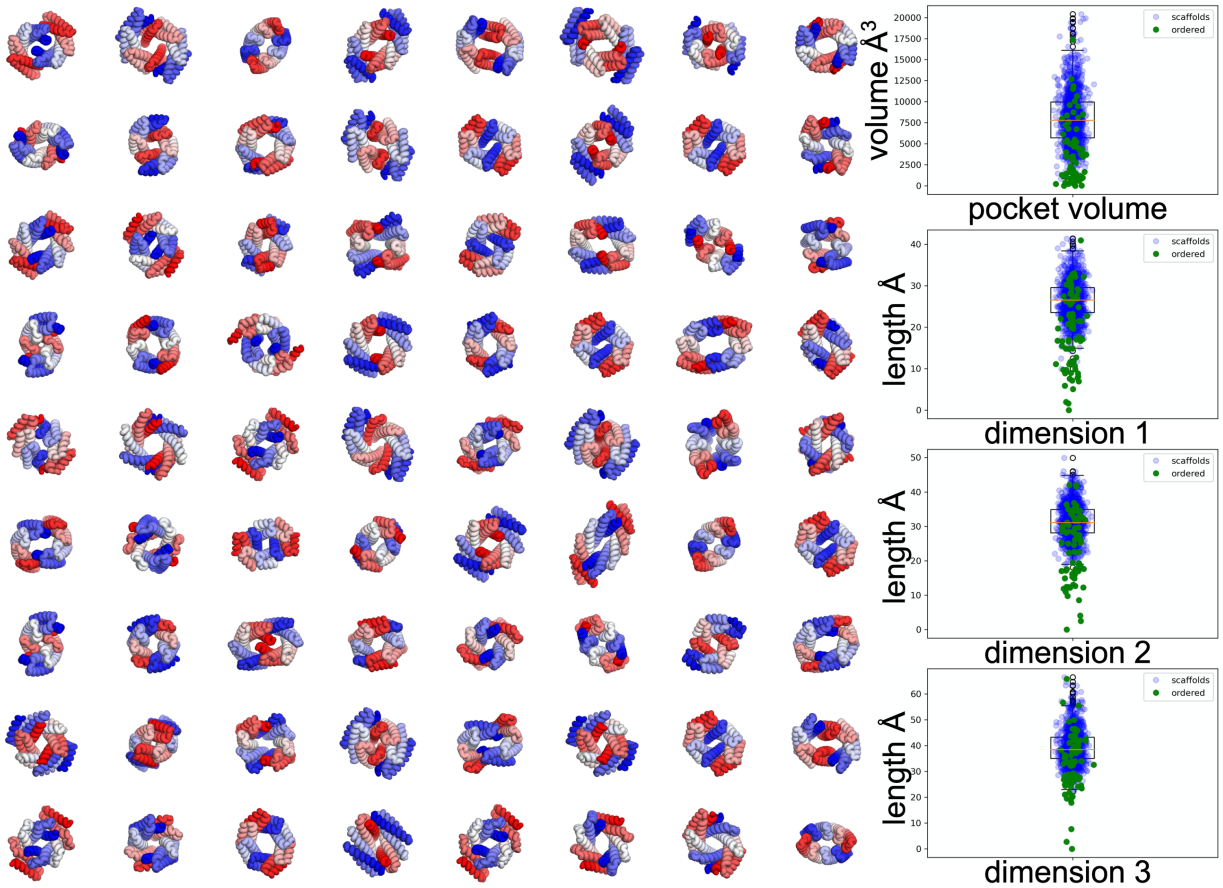


Figure 10. Designs span a diverse range of sizes, shapes, and pocket features. On the left are 72 ordered designs depicted as ribbon backbones colored from blue (N-terminus) to red (C-terminus). The right side shows boxplots plus points for four pocket features, volume, dimension 1, dimension 2, and dimension 3 for the top one thousand designs from generation 3 (boxplot and blue points) along with all ordered scaffold designs (green points).

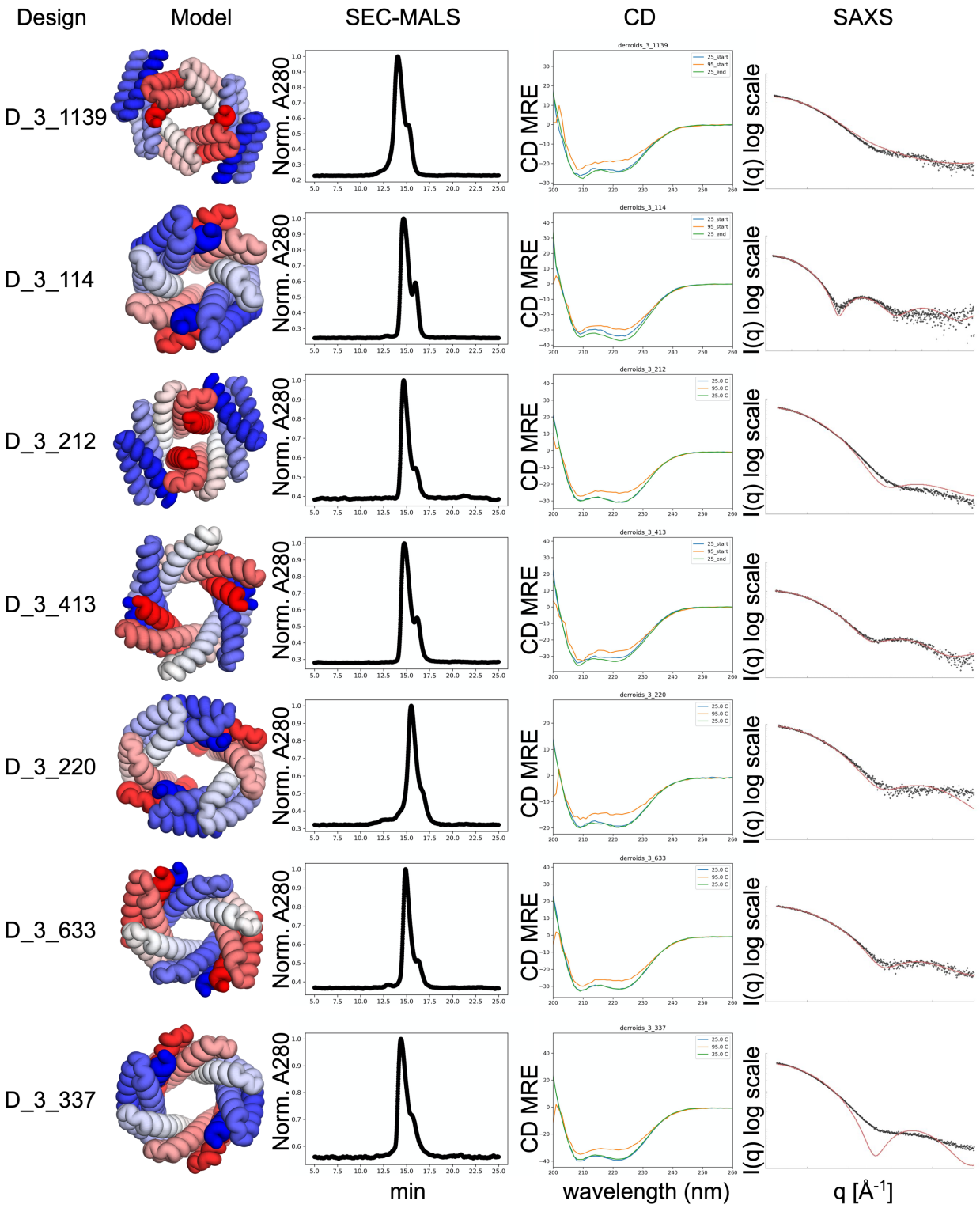

Design
Model
SEC-MALS
CD
SAXS
D_3_1139
D_3_114
D_3_212
D_3_413
D_3_220
D_3_633
D_3_337
Norm. A280
CD MRE
I(q) log scale
derroids_3_1139
derroids_3_114
derroids_3_212
derroids_3_413
derroids_3_220
derroids_3_633
derroids_3_337
25_start
95_start
25_end
25.0 C
95.0 C
25.0 C
min
wavelength (nm)
q [Å-1]

Figure 11. Representative data for 6 successful designs and 1 failed design. Design success was determined by SAXS. On the left, we show the design models depicted as ribbon backbones colored from blue (N-terminus) to red (C-terminus). Next, is shown normalized UV absorbance (A280) obtained during SEC-MALS, followed by circular dichroism scans from 200-260 nm at 25°C, 95°C and 25°C post-heating. On the right is shown predicted SAXS profiles overlaid on experimental SAXS data points for scattering vector (q) vs intensity (I).

Design	Soluble Expression	SAXS Vr	MALS ratio to dimer	Thermal CD	XTAL
D_1_1	Yes	1.2	0.5	Yes	-
D_1_2	No	-	-	-	-
D_1_3	Yes	1.9	1.1	Yes	-
D_1_4	No	-	-	-	-
D_1_5	No	3.1	-	No	-
D_1_6	Yes	1.6	1.3	Yes	-
D_1_7	No	-	-	-	-
D_1_8	No	-	-	-	-
D_1_9	No	-	-	-	-
D_1_10	No	-	-	-	-
D_1_11	No	-	-	-	-
D_1_12	No	-	-	-	-
D_1_13	Yes	0.55	1.0	Yes	-
D_1_14	Yes	1.1	0.8	Yes	-
D_2_1	No	-	-	-	-
D_2_2	Yes	-	2.7	No	-
D_2_3	No	-	-	-	-
D_2_4	No	-	-	-	-
D_2_5	No	-	-	-	-
D_2_6	No	-	-	-	-
D_2_7	No	-	-	-	-
D_2_8	No	-	-	-	-
D_2_9	No	-	-	-	-
D_2_10	Yes	1.8	0.9	Yes	-
D_2_11	No	-	-	-	-
D_2_12	Yes	2.3	0.9	Yes	-
D_2_13	Yes	1.3	1.1	Yes	-
D_2_14	No	-	-	-	-
D_2_15	No	-	-	-	-
D_2_16	Yes	1.3	0.9	Yes	-
D_2_17	Yes	1.1	0.9	Yes	-
D_2_18	Yes	0.6	0.9	Yes	-
D_2_19	No	-	-	-	-
D_2_20	No	-	-	-	-
D_2_21	No	-	-	-	-
D_3_1027	No	-	-	-	-
D_3_1029	No	3.6	-	-	-
D_3_1049	No	-	-	-	-
D_3_1061	Yes	2.1	0.8	Yes	-
D_3_1109	Yes	0.9	0.9	Yes	-
D_3_111	No	-	-	-	-
D_3_212	Yes	0.8	1.0	Yes	Scaffold
D_3_220	Yes	1.6	1.3	Yes	-
D_3_271	No	-	-	-	-
D_3_297	No	3.5	-	-	-
D_3_337	Yes	3.4	1.1	Yes	As monomer
D_3_598	No	-	-	-	-
D_3_601	No	5.9	-	-	-
D_3_633	Yes	0.7	1.1	Yes	Scaffold + Binder
D_3_636	No	-	-	-	-

D_3_840	No	-	-	-	-
D_3_87	No	-	-	-	-
D_3_881	No	-	-	-	-
D_3_904	Yes	0.8	0.9	Yes	-
D_3_946	No	-	-	-	-
D_3_972	Yes	0.7	0.8	Yes	-
D_3_1102	Yes	0.5	0.9	Yes	-
D_3_1108	No	-	-	-	-
D_3_1139	Yes	1.6	0.9	Yes	-
D_3_114	Yes	1.9	0.9	Yes	Binder
D_3_1173	No	-	-	-	-
D_3_261	Yes	2.2	1.0	Yes	-
D_3_342	No	-	-	-	-
D_3_345	No	-	-	-	-
D_3_413	Yes	0.9	1.0	Yes	-
D_3_439	Yes	2.3	1.0	Yes	-
D_3_561	No	-	-	-	-
D_3_562	Yes	6.9	1.6	Yes	-
D_3_565	No	-	-	-	-
D_3_663	No	-	-	-	-
D_3_68	Yes	2.9	0.8	Yes	-
D_3_707	No	-	-	No	-
D_3_770	Yes	-	1.2	No	-
D_3_891	Yes	0.4	1.0	Yes	-
D_3_936	Yes	TBD	1.5	Yes	-
D_3_939	No	-	-	-	-
D_3_944	No	-	-	-	-

Table 1. Summary of protein characterization. The soluble expression for each design is classified as a binary yes or no based on our ability to obtain sufficient quantities of protein for subsequent characterization. SAXS Vr is the volatility ratio described previously (Hura *et al.*, 2013). We consider designs with Vr values less than 2.5 to be successful as previously determined (Brunette *et al.*, 2015) by comparison to crystal structures. We report the SEC-MALS MW obtained for the major peak as a ratio to the expected MW of the designed homodimer. A ratio of 1 indicates a dimer. Thermal CD for each design is classified as a binary yes or no based on whether the sample appeared to be majority helical and thermostable. Each majority helical protein appeared to remain primarily folded at 95°C and their signals nearly superimposed after cooling back to 25°C. Finally, we report whether we obtained a crystal structure for the designed scaffold (one as a monomer) or for a binder derived from the scaffold. Rows with designs classified as successful by Vr are highlighted in green, while gray rows represent failed designs. Yellow rows are designs that need further characterization, and the single orange row highlights a design that failed to form the designed homodimer but crystallized as a monomer.

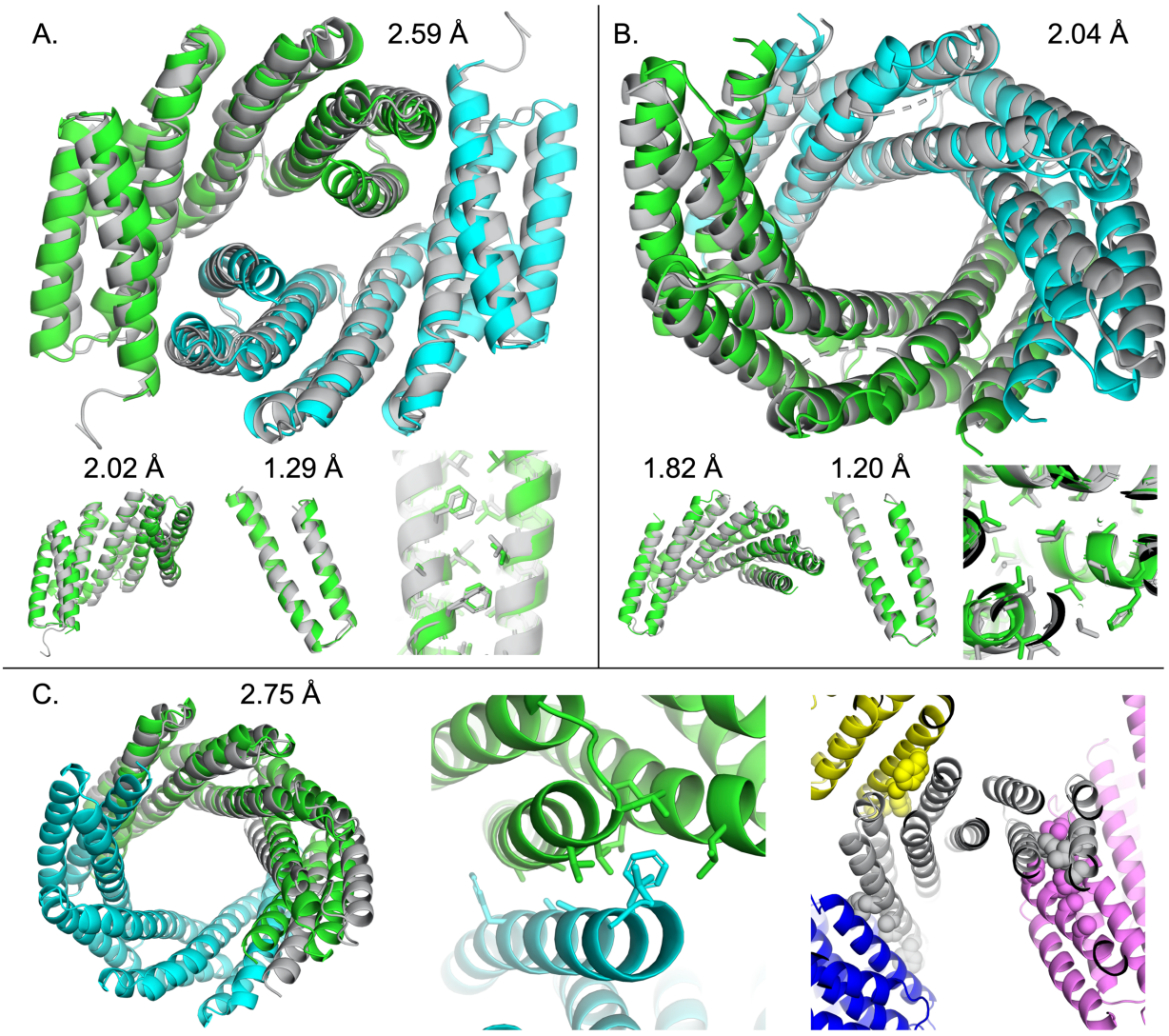


Figure 12. Crystallographic analysis of two successful designs and one failed design. Panel A shows an overlay of design D_3_212 (green and cyan) with its crystal structure (gray). The top portion shows the superposition of the homodimer, while the lower portion shows the superposition of the monomer, a repeat unit, and a section of the hydrophobic core. Associated rmsds are shown. Panel B shows an overlay of design D_3_633 (green and cyan) with its crystal structure (gray). The top portion shows the superposition of the homodimer, while the lower portion shows the superposition of the monomer, a repeat unit, and a section of the hydrophobic core. Associated rmsds are shown. Panel C shows design D_3_337 and its crystal structure. The

left is an overlay of design D_3_337 (green and cyan) with its crystal structure (gray), which is monomeric, with associated rmsd. The middle shows the designed homodimer interface, with hydrophobic residues shown as sticks and the two chains colored green and cyan. On the right is the crystal structure showing the central asymmetric unit in gray and its crystal lattice neighbors colored blue, pink, and yellow. The hydrophobic residues which were intended to form the homodimer interface are shown in spheres forming key crystal contacts.

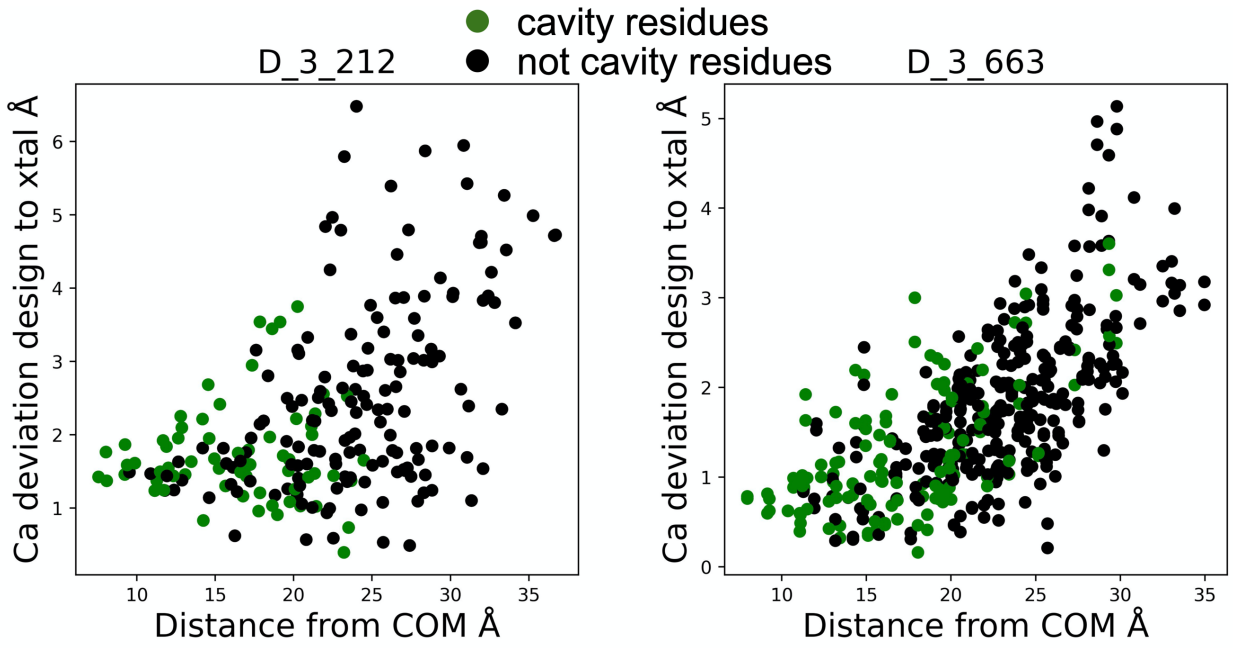


Figure 13. Scatter plot of distance from the center of mass vs Ca deviation. The plots show Ca deviations (y-axis) compared to the distance to the protein center of mass (x-axis) for design models compared to their respective crystal structures. Points for cavity lining residues are colored green, while points for the rest of the protein are colored black. The plot on the left is for design D_3_212 while the plot on the right is for design D_3_633.

2.5 DISCUSSION

We advanced the methods for creating repeat proteins by moving beyond random sampling and towards focused sampling of repeat protein conformational space. This allowed us to create a computational library of curved repeat proteins that we subsequently used to make C2 symmetric homodimer proteins with central cavities of diverse shapes ideal for binding a range of C2 symmetric ligands. We believe these proteins have several advantageous properties including thermal stability and high solubility in a range of buffer conditions. Furthermore, we designed these proteins such that the protein core is distinct from the pocket, which we believe will enhance their mutation tolerance during functionalization.

While two of the crystal structures that we solved demonstrate that our computational design models were mostly correct, they still had backbone rmsds of ~2.5 Å. Our third crystal structure showed that despite being mostly correct at the monomer level, our design failed to form the intended homodimer. Extensive sampling of the available protein ensemble around a design model, along with possible alternative conformations at the monomer and/or dimer level may improve our ability to select the lowest energy state. Ideally, this type of large-scale sampling of conformational space would be done without symmetric constraints such as repeat protein symmetry or explicit C2 symmetry. Unfortunately, this type of extensive sampling of protein conformational space is highly computationally expensive and cannot currently be applied to the thousands of designed homodimers we generated. Recent advances in the fields of structure prediction and structural refinement, particularly deep learning models, which might be able to learn something about the protein folding landscape that our current methods are missing (Senior et al. 2020), may allow us to select better design models in the future.

Chapter 3. DESIGN OF CHLOROPHYLL DIMER BINDERS

An interesting application of the designed C2 symmetric homodimers would be binding a C2 symmetric chlorophyll dimer similar to the chlorophyll special pair found in photosynthetic reaction centers that is responsible for photoinduced charge separation. This could allow for the creation of new light-harvesting systems or enzymes. It could also open up new questions in basic science in terms of how the protein environment surrounding chlorophyll impacts chlorophylls electronic properties such as absorption and emission spectrum. Using our existing library of C2 symmetric homodimers, we were able to generate designs that bind chlorophyll dimers with diverse orientations and pockets of diverse sidechain packing and chemistry. Experimental characterization indicates that the majority of chlorophyll dimer binder designs bind two chlorophyll molecules as desired. Further characterization is needed for several designs to determine their binding stoichiometries. Preliminary crystallographic analysis for one design, D_3_114_1, indicates that the protein is largely correct, and shows electron density for what appear to be two molecules of Zn pheophorbide a methyl ester (ZnPPaM), an analog of chlorophyll. However, the histidine rotamers coordinating chlorophyll appear different than designed, indicating the need to improve our design methodology.

3.1 COMPUTATIONAL DESIGN PIPELINE

We modeled in binding sites for porphyrin dimers within the dimer cavities, using a hash-based docking protocol. We first generated an ensemble of porphyrin dimers, built inverse histidine rotamers off of the metal site according to the geometries seen in crystal structures, and

then saved the 6D transform for the backbone positions in a hash table. When we searched a homodimer scaffold, we restricted the search to symmetric residue pairs, and when a backbone match was found in our hash table, we built the histidine ligand complex into the protein. Afterward, we optimized the sequence at the protein-ligand interface using a FastDesign protocol. Designs were filtered based on ligand burial, interface shape complementarity, and binding energy, and top designs were ordered for experimental characterization.

3.2 EXPERIMENTAL CHARACTERIZATION

Proteins were purified as previously described and then used for titration binding experiments monitored by circular dichroism (CD). CD spectra of chlorophyll-containing proteins from 600 to 800 nm (near the chlorin Qy transition) revealed Cotton effects indicative of excitonic coupling between the two chlorophylls or between chlorophylls and aromatic amino acid side chains (Grishina and Woody, 1994; Matile *et al.*, 1995). The intensities, positions, and shapes of the Cotton effects varied considerably between different proteins, suggesting that a wide range of chemical environments were sampled. Cotton effects were monitored during titrations of chlorophyll into protein solutions in order to determine chlorophyll-binding stoichiometries and dissociation constants, as shown in Figure 14. Of 29 designs characterized by CD titrations, 17 designs bind 2 chlorophylls per protein dimer as desired, 1 design appears to bind a single chlorophyll per protein dimer, and 11 designs either failed to bind chlorophyll or yielded inconclusive data. Dissociation constants were estimated using a simple 1 site binding equation. We note that these binding affinities are likely inaccurate due to a number of factors including potential cooperativity between the two chlorophylls binding to a protein dimer and due to

differences in CD signal obtained from a single bound chlorophyll which should have minimal cotton effects and two bound chlorophyll molecules whose interaction should lead to strong cotton effects. Despite these limitations, these binding affinities are likely accurate enough to differentiate better binders, such as H65A, from worse binders, such as D_3_114_5.

One binder, D_3_114_1 was characterized by crystallographic analysis. This protein was found to be in close agreement with the design model with a Ca rmsd of 1.56 Å. However, the histidine residues designed to coordinate chlorophyll appear to adopt a different rotamer compared to the design model (see figure 15). Extra density was seen around these histidines, presumably belonging to bound chlorophyll, but refinement was unable to unambiguously place chlorophyll into this density. CD titrations indicate that this protein binds 2 chlorophylls per protein dimer and has pronounced cotton effects due to interactions of the chlorophyll dimer.

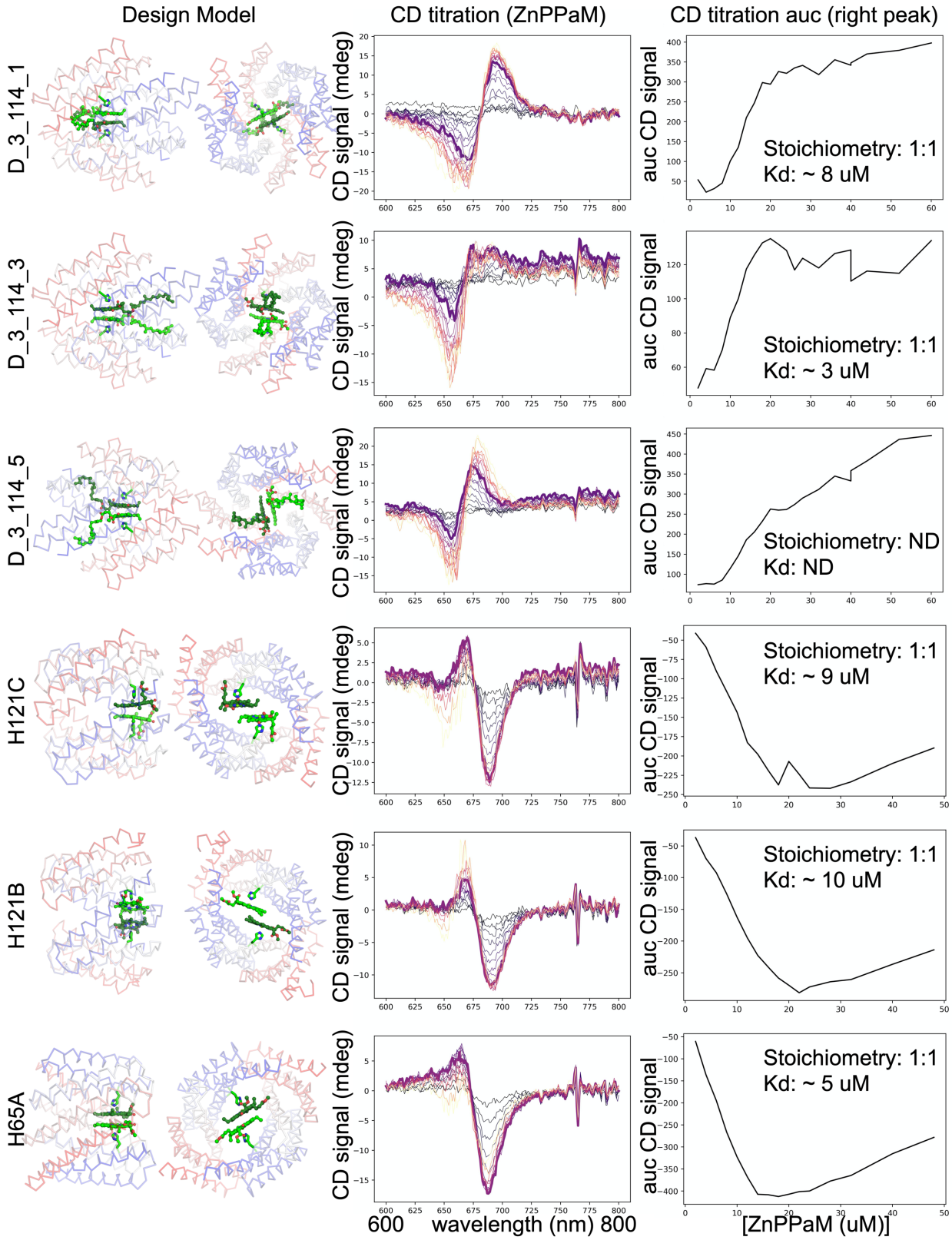

Design Model
CD titration (ZnPPaM)
CD titration auc (right peak)
D_3_114_1
D_3_114_3
D_3_114_5
H121C
H121B
H65A
CD signal (mdeg)
auc CD signal
Stoichiometry: 1:1
Kd: ~ 8 uM
Stoichiometry: 1:1
Kd: ~ 3 uM
Stoichiometry: ND
Kd: ND
Stoichiometry: 1:1
Kd: ~ 9 uM
Stoichiometry: 1:1
Kd: ~ 10 uM
Stoichiometry: 1:1
Kd: ~ 5 uM
600 wavelength (nm) 800
[ZnPPaM (uM)]

Figure 14. Circular dichroism titrations of chlorophyll binder designs. Design models are shown (left) in two orientations, first a side view of a design model and second a top-down view of the design model. The protein is shown as backbone ribbon colored blue (N-term) to red (C-term) with the bound chlorophyll and ligating histidines shown in green (oxygen is shown in red and nitrogen in blue). Next is a titration of Zn pheophorbide a methyl ester (ZnPPaM) in 20 uM of protein monomer with signal monitored by circular dichroism. CD scans are colored from black (low) to yellow (high) based on the concentration of ZnPPaM. The bold line depicts 1:1 stoichiometric concentrations of protein and ZnPPaM. Finally, the area under the curve for the rightmost peak is shown plotted against ZnPPaM concentration. The binding signal appears to saturate with 1:1 stoichiometry for all designs except D_3_114_3, which may have a weaker binding affinity than the other designs or bind less than 2 ligands per protein dimer.

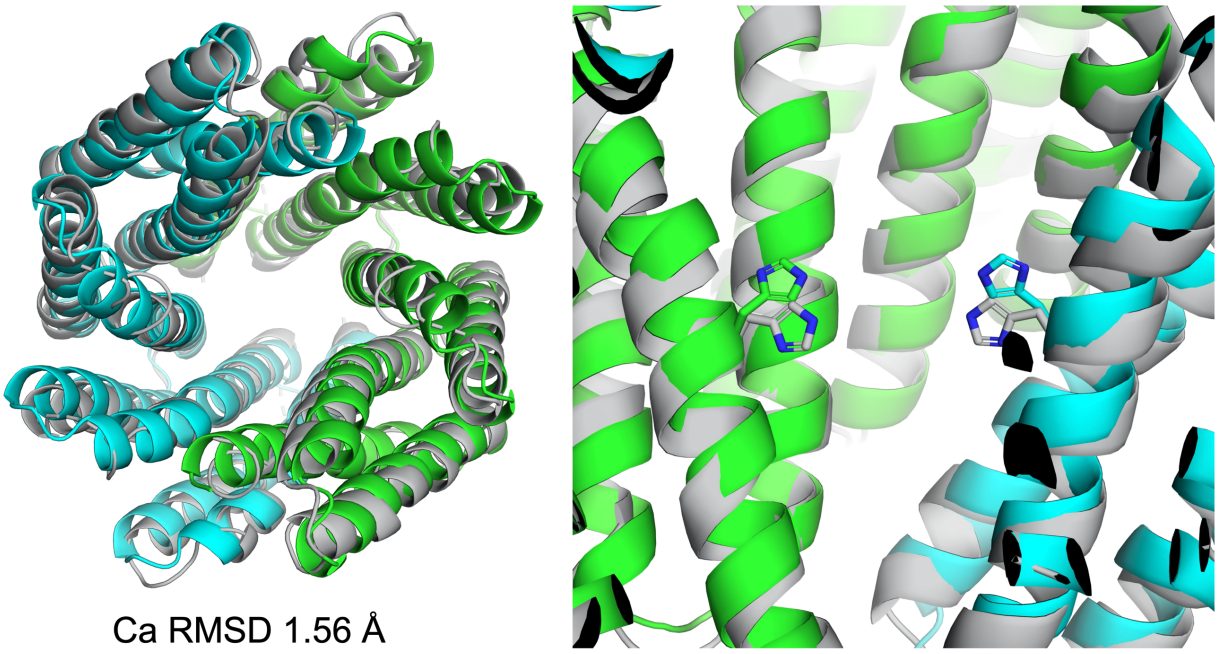


Figure 15. Crystallographic analysis of design D_3_114_1 obtained on an in-house X-ray source and solved to approximately 2.8 Å. On the left is shown a superposition of the design model (green and cyan) on the crystal structure (gray). The protein backbone appears fairly accurate with a Ca rmsd of 1.56 Å. On the right is shown the histidine pair designed to coordinate the chlorophyll dimer for the crystal structure (gray) and design model (green and cyan), with histidine nitrogens colored blue. The crystal structure shows that the histidine adopts a different rotamer compared to the design model. While there was extra density in the pocket, presumedly for the chlorophyll dimer, crystallographic refinement was unable to place the chlorophyll pair.

3.3 Discussion

Our designed C2 symmetric proteins that bind chlorophyll dimers represent an excellent starting point for future engineering applications for light-harvesting, enzyme design, or evolution. These proteins may also prove useful for fundamental research aimed at understanding the

electrochemistry of chlorophyll and chlorophyll dimers. Because the proteins are so well behaved, it should be possible to study variants that systematically perturb the electrochemistry of the binding site by changing the hydrophobicity, packing density, or charge of the pocket, or by replacing chlorophyll with various derivative molecules.

As seen previously in the design and characterization of scaffold proteins, our design methods allow us to generate novel proteins with the correct protein fold at the tertiary and quaternary levels. However, there appear to be limitations in our scoring or sampling methodology limiting the accuracy of our models to 1-2Å in terms of backbone Ca rmsd. Despite these limitations, we have been able to design functional proteins that bind chlorophyll dimers as desired. Crystallographic analysis of one chlorophyll binder design revealed differences in the rotamer of the key histidine residues responsible for coordinating the binding of chlorophyll. These differences may arise from inaccuracies in modeling the protein backbone or in inaccuracies modeling the binding site that are distinct from backbone level inaccuracies. As previously described, our design models may benefit from increased sampling of the local protein/ligand ensemble as well as of the larger protein energy landscape. Sampling of the protein energy landscape would ideally be done while eliminating or decreasing the requirement of explicit symmetry that we use during design.

A major limitation in modeling molecules like chlorophyll that contain long flexible carbon chains is the combinatorial explosion of rotamers that occurs if all possible torsions are allowed to rotate. Excluding hydrogens, the phytol/carbon chain of chlorophyll has ~15 flexible torsions. If each is given 3 possible torsions to sample, then there are ~14 million rotamers to consider for this molecule, which is orders of magnitude greater than the tens of thousands of rotamers which can be sampled in a packing trajectory expected to finish in a reasonable time (a few hours or less).

We have attempted to work around this limitation in one of three ways: first, modeling the molecule without this chain at all; second, modeling the chain as a few low energy conformations; or third, allowing a few select torsions to rotate. In addition to sampling limitations, it is also possible that the Rosetta energy function, which is mostly trained on protein crystal structures, has meaningful inaccuracies, particularly when attempting to model binding sites of large metal-containing organic molecules such as chlorophyll.

Chapter 4. DESIGN OF C2 SYMMETRIC PEPTIDE BINDERS

Another interesting application of our C2 symmetric homodimers would be binding to designed C2 symmetric peptides. These could then be converted into chemically induced dimers with more control over the properties of the protein and chemical dimerizer than ever before. These could have interesting uses in the control of engineered cell therapeutics or as basic science research tools since they would be entirely bioorthogonal. Recent advances in peptide design allow us to make thousands to millions of C2 symmetric peptides from canonical and non-canonical side chains and backbones with diverse shapes and properties. Some of these have ideal drug-like properties that make them interesting targets to bind for the creation of bioorthogonal chemically induced dimers.

Members of the Baker lab recently created and structurally validated a C2 symmetric cyclic peptide with the repeated sequence D-Ala, D-pro, L-Leu, L-Phe as well as its mirror image enantiomer L-Ala, L-pro, D-Leu, D-Phe. We decided to use our C2 symmetric homodimer protein library to try to bind this peptide. Out of 29 experimentally tested designs, one bound with low micromolar affinity. The crystallographic analysis of this design showed that the protein backbone was quite accurate, and that the peptide bound in roughly the correct location, but with different rigid body orientation. As previously discussed, this indicates the need for improved sampling and scoring methodology to improve the accuracy of computational protein design.

4.1 COMPUTATIONAL DESIGN PIPELINE

C2 symmetric peptides were generated by fragment assembly of energy minimized canonical and noncanonical residues (1-mers) into an N-mer. The N-mer is then repeated to generate a 2N-mer, which is checked for cyclization and C2 symmetry. The peptide used in this design pipeline had been previously characterized by crystallographic analysis in the lab. Because the peptide contains L and D amino acids, the crystals were grown from a racemic mixture containing the originally designed peptide along with its mirror image enantiomer. We used both versions of the peptide in subsequent design steps.

We docked both enantiomers of the peptide into the protein cavity utilizing a rifgen/rifdock pipeline. First, the peptide symmetry axis was aligned to the z-axis, and the center of mass of the peptide placed at the origin (0, 0, 0). Next, the rotamer interaction field was generated (rifgen), which attempts to enumerate hydrophobic interactions and polar, hydrogen bonding, interactions with the peptide. Inverse rotamers for all favorable interactions were saved in a hash table. Scaffold proteins, which were also centered at the origin and with their symmetry axis aligned to the z-axis, were then rotated and translated by predefined rotation matrices that enumerate a C2 symmetric grid. At each grid point, the backbone transforms from the origin to each residue in a predefined set of residues covering the potential binding site were searched in the rotamer interaction field hash table. For a particular grid point, the maximum sum of all compatible interactions was summed. Finally, the rigid body orientation for the protein scaffold and all compatible interactions were output along with the peptide.

Scaffolds with peptides docked into them were then subjected to sequence optimization of the binding site using FastDesign and C2 symmetry. Residue interactions installed during the

docking protocol were allowed to mutate to more favorable interactions if available. Top designs were selected for experimental validation based on a variety of filters that evaluate the binding energy, shape complementarity, buried unsatisfied hydrogen bonds, and packing.

4.2 EXPERIMENTAL CHARACTERIZATION

After expression and purification, designs were tested for binding by native mass spectrometry (nMS) as previously described (Pyles *et al.*, 2019). For nMS, samples were analyzed with and without peptide added to the sample at 10X concentration. For this experiment, the protein samples were at 50 uM monomer (25 uM homodimer) and the peptide was 250 uM. The samples were in TBS buffer with 5% DMSO. Native MS showed that one design, D_3_633_x8, appeared to bind the peptide with nearly 100% of the complex bound (see figure 16). The rest of the designs did not appear to bind.

Peptide binding was subsequently interrogated by equilibrium dialysis mass spectrometry as described by Thermo Fisher Scientific (*[No title]*, no date). The dialysis experiment was set up with 50 uL of protein at 100 uM on one side of the dialysis membrane and 300 ul of the peptide at 10 uM on the other side. Both sides were in TBS with 5% DMSO. After overnight equilibration at 20 °C and 250 rpm shaking, the ratio of the peptide on the protein side of the membrane compared to the buffer side of the membrane was quantified by mass spectrometry. Mass spec analysis showed the same design, D_3_633_x8, identified by nMS bound much stronger than any other designs (see Figure 17). We calculated the binding affinity of this protein to the peptide to be ~8 uM using the following equations:

ratio = ligand_on_protein_side / ligand_in_buffer

ligand_on_protein_side = 1 - ligand_in_buffer

ligand_in_buffer = 1 - fraction_ligand_bound - (1 - fraction_ligand_bound)/2

fraction_lig_bound =

(Kd+[prot]+[lig] - sqrt((Kd+[prot]+[lig])^2 - 4*[prot]*[lig])) / (2*[lig])

We had previously obtained a crystal structure of the scaffold protein used to design this binder; however, the design was made prior to obtaining this structure. The rmsd between the design and scaffold structure was 1.9 Å. We later obtained structures of the binder design with and without the bound peptide. All by all RMSDs for the binder design and the three crystal structures ranged from 1.1 Å to 2.1 Å. The peptide was bound approximately in the correct location of the protein cavity; however, it had flipped over ~180° along an axis perpendicular to the symmetry axis and tilted such that it was not bound in a symmetrical orientation (see figure 18).

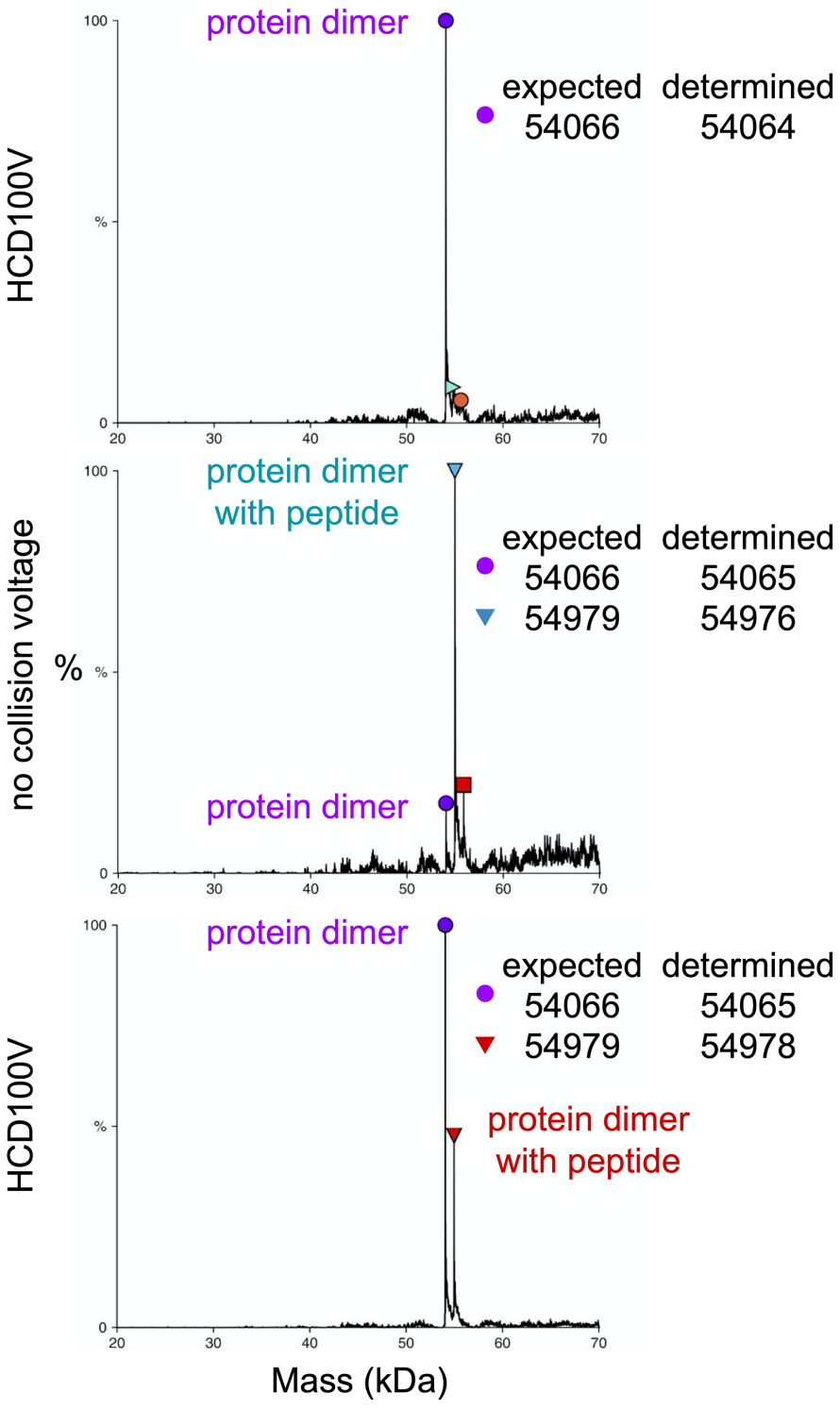

HCD100V
protein dimer
expected determined
54066 54064
no collision voltage
protein dimer
with peptide
expected determined
54066 54065
54979 54976
%
protein dimer
HCD100V
protein dimer
expected determined
54066 54065
54979 54978
protein dimer
with peptide
100
%
0
20
30
40
50
60
70
Mass (kDa)

Figure 16. Native mass spectrometry of protein-peptide complex. Design D_3_633_x8 was subjected to nMS analysis without peptide, top panel, and with peptide at 10X protein concentration, bottom two panels. The top and bottom panels show samples analyzed with all-ion fragmentation (MSMS) mode with high energy collision-induced dissociation (HCD) 100 V. The middle panel was analyzed with full MS mode (no collision voltage applied). For each panel, only deconvoluted spectra are shown.

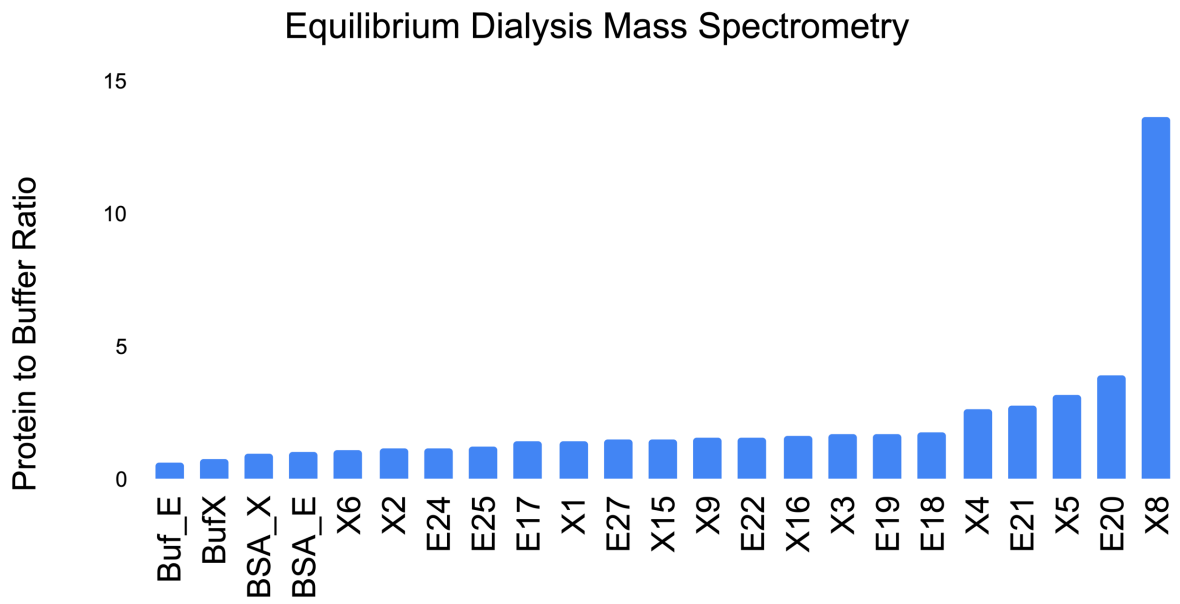


Figure 17. Equilibrium dialysis mass spectrometry of peptide binder designs. The leftmost sample is the peptide and buffer without protein. The next two samples use BSA as the protein as controls of nonspecific binding. The rest of the samples represent binder designs. Each sample name contains an E, for enantiomer, or X, for the original peptide. Because the peptide contains L and D amino acids, we used the mirror image enantiomer during design and characterization, along with the originally designed peptide.

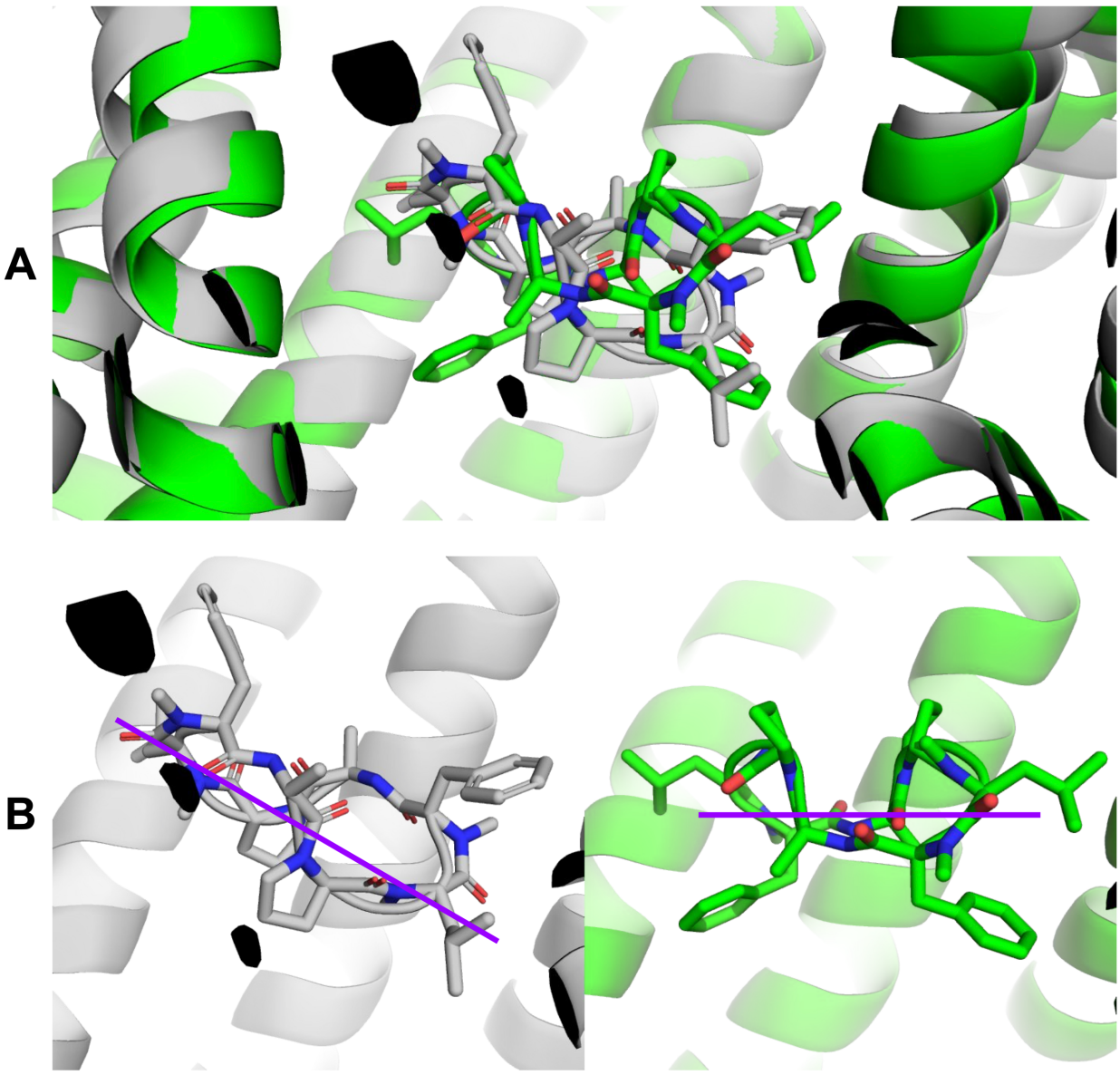

A
B
C
Aqueous
Bound
Organic

Figure 18. Crystal structures and design models for peptide binder D_3_633_x8. Panel A shows the crystal structure (gray) superimposed on the design model (green). The protein is depicted as backbone cartoon, and the peptide is depicted as sticks. Panel B shows crystal structure (left) and design model (right) with an axis (purple) drawn perpendicular to the peptides axis of symmetry. The bound peptide flipped 180° about this axis and tilted ~30° along this axis compared to the design model. Panel C shows crystal structures of the peptide determined in aqueous solution (left), in the bound protein structure (middle), and in an organic solution (right), which was the designed binding conformation.

4.3 DISCUSSION

The accuracy of our models ranged from 1 Å to 2 Å by Ca RMSD, showing that our models are close but imperfect. It is possible that enhanced sampling methods could increase the accuracy of our models. Given a fixed design sequence, enhanced sampling methods would attempt to improve the accuracy of monomer backbone and sidechain confirmations, as well as rigid body orientations of interfaces. Potential methods to accomplish this could include utilizing forward folding to obtain more accurate monomers or low energy ensembles that better reflect the true structure of our designs, as well as forward docking to improve the rigid body orientations. Unfortunately, these sampling methods are computationally expensive to run on individual inputs let alone on large computational libraries. Furthermore, it would be ideal to use monomer ensembles along with forward docking to better explore the combination of these two search spaces, however, this drastically increases the computational cost of calculations that are already time and resource consuming. It is possible in the future that improved protein refinement approaches being developed for structure prediction (Read *et al.*, 2019), such as those utilizing

deep learning techniques (Heo and Feig, 2020), could help with the refinement of designed proteins.

We bound a peptide in the correct location, but with the wrong rigid body orientation. In the crystal structure, the peptide binds in an asymmetric manner, which may indicate limitations in our sampling approaches. A major limitation in terms of designing this binder was the explicit enforcement of C2 symmetry. Because of this, we could never possibly have sampled the real binding mode. However, it is unclear why the design binds in this mode. Comparison of existing filter metrics that evaluate packing, shape complementarity, buried surface area, and energy appear very similar between the design model and the crystal structure. It is possible that this binding mode may be favored because it relieves small clashes present in the design model, potentially indicating limitations in the Rosetta energy function.

Chapter 5. METHODS

5.1 SYNTHETIC GENE CONSTRUCTS

All genes were ordered from Integrated DNA Technologies (IDT). In a few cases, genes were not synthesizable by IDT, and were instead ordered from Genscript. A His-tag containing TEV protease cleavage site and short linkers were added to the N-terminus of protein sequences. In cases in which the protein lacked a Tryptophan residue, a single Tryptophan was added to the short N-terminal linker following the TEV protease cleavage site to help with protein concentration quantification by A280. The protein sequence along with linker (GHHHHHHGSGSGENLYFQSGSGSSS or GHHHHHHGSGSGENLYFQSGWSGSSS) was reverse translated into DNA using a custom python script that attempts to maximize host-specific codon adaptation index (Sharp and Li, 1987) and IDT synthesize-ability, which includes optimizing whole gene and local GC content as well as removing repetitive sequences. Finally, a TAA stop codon was appended to the end of each gene. Genes were delivered cloned into pET-29b+ between NdeI/XhoI restriction sites.

5.2 PROTEIN EXPRESSION AND PURIFICATION

Proteins were transformed into Lemo21(DE3) E. coli from New England Biolabs (NEB) and then expressed as 0.5-liter cultures in 2-liter flasks using Studiers M2 autoinduction media with 50 ug/mL kanamycin. The cultures were either grown at 37°C for ~6-8 hours and then ~18°C overnight (~14 hours) or at 37°C the entire time ~14 hours. Cells were pelleted at 4,000g for 20 minutes, after which the supernatant was discarded. Pellets were resuspended in 30 ml lysis buffer

(25 mM Tris HCl pH 8, 300 mM NaCl, 30 mM imidazole, 10 mM lysozyme, 1 mM DNase, with Thermo Scientific Pierce protease inhibitor tablet). Cell suspensions were lysed by microfluidizer or sonication, and the lysate was clarified at 20,000g for ~45 minutes. The His-tagged proteins were bound to Ni-NTA resin (Qiagen) by batch binding or during gravity flow and washed with a wash buffer (25 mM Tris HCl pH 8, 300 mM NaCl, 30 mM imidazole). Protein was eluted with an elution buffer (25 mM Tris HCl pH 8, 150 mM NaCl, 400 mM imidazole). The His-tag was removed by TEV cleavage, followed by IMAC purification to remove TEV protease. The flowthrough was collected and concentrated prior to further purification by SEC/FPLC on a superdex 200 increase 10/300 GL column in TBS (25 mM Tris pH 8.0, 150 mM NaCl).

5.3 CIRCULAR DICHROISM

Circular dichroism spectra were measured with an AVIV Model 420 DC or Jasco J-1500 CD spectrometer. Samples were 0.25 mg/mL in TBS (25 mM Tris pH 8.0, 150 mM NaCl), and a 1-mm path length cuvette was used. The CD signal was converted to mean residue ellipticity by dividing the raw spectra by N × C × L × 10, where N is the number of residues, C is the concentration of protein, and L is the path length (0.1 cm).

5.4 SIZE EXCLUSION CHROMATOGRAPHY WITH MULTI-ANGLE LIGHT SCATTERING

Purified samples after the initial SEC run, samples were pooled then concentrated or diluted as needed to a final concentration of 2 mg/mL. 100 uL of each sample was then run through a high-performance liquid chromatography system (Agilent) using a Superdex 200 10/300 GL

column. These fractionation runs were coupled to a multi-angle light scattering detector (Wyatt) in order to determine the absolute molecular weights for each designed protein as described previously (Fallas *et al.*, 2017).

5.5 SMALL ANGLE X-RAY SCATTERING

Small-Angle X-ray Scattering (SAXS) was collected at the SIBYLS High Throughput SAXS Advanced Light Source in Berkeley, California (Dyer *et al.*, 2014). Beam exposures of 0.3 s for 10.2 s resulted in 33 frames per sample. Data was collected at low (~1 mg/mL) and high (~2-3 mg/mL) protein concentrations in SAXS buffer (25mM Tris pH 8.0, 150mM NaCl, 2% glycerol). The sibyls website (*SAXS FrameSlice*, no date) was used to analyze the data for high and low centration samples and average the best dataset. If there was obvious aggregation over the 33 frames, only the data points before aggregation arises were used in the Gunier region, otherwise, all data was included for the Gunier region. All data was used for Porod and Wide regions. The averaged file was used with scatter.jar to remove data points with outlier residuals in the Gunier region. Finally, the data was truncated at 0.25 q. This dataset was then compared to the predicted SAXS profile based on the design model using the FoxS SAXS server (*FoXS Server: Fast X-ray Scattering*, no date), and volatility ratio (Vr) was calculated to quantify how well the predicted and data matched the experimental data. Proteins with Vr of less than 2.5 were considered to be folded to the designed quaternary shape.

5.6 CRYSTALLOGRAPHY

Promising designs were screened for crystallization by wonderful collaborators, Asim Bera, Barry Stoddard, and Madision Kennedy, and X-ray structures were obtained for several designs.

D_3_212

Crystallography sample preparation, data collection, and analysis

Crystal screening was performed using Mosquito Crystal by STP Labtech. Crystals were grown in 10% PEG 20000, 20% v/v glycerol, 0.2 *M* sodium L-glutamate, 0.2 *M* DL-alanine, 0.2 *M* glycine, 0.2 *M* DL-lysine HCl, 0.2 *M* DL-serine, 0.1 *M* MES/imidazole pH 6.5. Crystals were subsequently harvested in a cryo-loop and flash frozen directly in liquid nitrogen for synchrotron data collection. Data was collected on ALS beamline 8.2.1. X-ray intensities and data reduction were evaluated and integrated using DIALS (Beilsten-Edmands J, Winter G, Gildea R, Parkhurst J, Waterman D, Evans G.*Acta Crystallogr D Struct Biol* **76**, 385-399 (01 Apr 2020). [PMID:32254063]) and merged/scaled using Pointless/Aimless in the CCP4 program suite (Winn, M. D. et al. Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. 67, 235–242 (2011).

Starting phases were obtained by molecular replacement using Phaser (McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007)) using the designed model. Sidechains were rebuilt and the model was refined with Rosetta-Phenix (Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010)). Manual rebuilding in Coot (Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol.*

Crystallogr. **60**, 2126–2132 (2004) and cycles of phenix refinement were used to build the final model. The final model was evaluated using MolProbity (Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293-315, doi:10.1002/pro.3330 (2018)). Data collection and refinement statistics are recorded in the table below.

Crystallographic Data Collection and Refinement Statistics

	D_3_212
Data collection	
Space group	$P2_12_12$
Cell dimensions	
a, *b*, *c* (Å)	58.09 97.21 36.26
α, β, γ (°)	90, 90, 90
Resolution (Å)	37.28 - 1.64 (1.70 - 1.64)[a]
No. of unique reflections	25328 (2417)
R_{merge}	0.078 (1.397)
R_{pim}	0.030 (0.527)
$I/\sigma(I)$	10.59 (1.27)
$CC_{1/2}$	0.998 (0.821)
Completeness (%)	99.04 (96.87)
Redundancy	7.7 (7.9)
Refinement	
Resolution (Å)	37.28 - 1.64 (1.70 - 1.64)
No. of reflections	25313 (2414)
R_{work} / R_{free} (%)	18.6 / 21.5 (29.4 / 33.5)
No. atoms	1991
Protein	1895

Ion /Ligand	0
Water	96
Ramachandran Favored/allowed Outlier (%)	 99.56/0.44 00.00
r.m.s. deviations	
Bond lengths (Å)	0.011
Bond angles (°)	0.970
$B_{factors}$ (Å^2)	
Protein	41.88
Water	45.10

1. Data were collected from one single crystal.
2. [a]Values in parentheses are for the highest-resolution shell.

D_3_337

Crystallography sample preparation, and data collection

Crystal screening was performed using Mosquito Crystal by STP Labtech. Crystals were grown in 10% w/v PEG 4000, 20% v/v glycerol, 0.3 M magnesium chloride, 0.3 M calcium chloride, and 0.1 M bicine/Trizma base pH 8.50. Crystals were subsequently harvested in a cryo-loop and flash frozen directly in liquid nitrogen for synchrotron data collection with additional 20% glycerol as cryoprotectant. A low-resolution data was collected on ALS beamline 8.2.1. X-ray intensities and data reduction were evaluated and integrated using DIALS (Beilsten-Edmands J, Winter G, Gildea R, Parkhurst J, Waterman D, Evans G.*Acta Crystallogr D Struct Biol* **76**, 385-399 (01 Apr 2020). [PMID:32254063]) and merged/scaled using Pointless/Aimless in the CCP4 program suite (Winn, M. D. et al. Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. 67, 235–242 (2011).

Starting phases were obtained by molecular replacement using Phaser (McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007)) using the designed model. Sidechains were rebuilt and the model was refined with Phenix (Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010)). Manual rebuilding in Coot (Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004) and cycles of phenix refinement were used to build the final model. The final model was evaluated using MolProbity (Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293-315, doi:10.1002/pro.3330 (2018)). Data collection and refinement statistics are recorded in the table below.

Crystallographic Data Collection Statistics

	D_3_337
Data collection	
Space group	$P4_12_12$
Cell dimensions	
a, *b*, *c* (Å)	90.72, 90.72, 108.70
α, *β*, *γ* (°)	90, 90, 90
Resolution (Å)	33.65 - 3.17 (3.28 - 3.17)[a]
No. of unique reflections	7677 (591)
R_{merge}	0.238 (0.788)
R_{pim}	0.066 (0.210)
$I/\sigma(I)$	12.5 (2.0)
$CC_{1/2}$	0.905 (0.939)

Completeness (%)	94.04 (73.88)
Redundancy	14.9 (14.7)
Refinement	
Resolution (Å)	33.65 - 3.17 (3.28 - 3.17)
No. of reflections	7672 (591)
R_{work} / R_{free} (%)	23.4 / 28.2 (30.3 / 34.3)
No. atoms	2070
Protein	2065
Ion /Ligand	0
Water	5
Ramachandran Favored/allowed Outlier (%)	 95.83 / 3.79 00.38
r.m.s. deviations	
Bond lengths (Å)	0.002
Bond angles (°)	0.400
$B_{factors}$ (Å^2)	
Protein	64.66
Water	22.96

D_3_633 and D_3_633_x8_apo and D_3_633_x8_halo

Crystallography sample preparation

Crystal screening was performed using Mosquito Crystal by STP Labtech. Crystals were grown in 0.2 M Zinc acetate, 0.1 M Na acetate pH 4.5, 10% PEG 3000 and 20% v/v glycerol as cryoprotectant. Crystals were subsequently harvested in a cryo-loop and flash frozen directly in liquid nitrogen for synchrotron data collection. Data was collected on ALS beamline 8.2.1.

Purified protein D_3_633_x8 with and without PJS-l-16 (peptide) were initially tested for crystallization via sparse matrix screens in 96-well sitting drops (200 nL drop volumes versus 95 µL reservoir volumes) using a mosquito crystallization robot (TTP LabTech). Crystallization conditions were then optimized with constructs that proved capable of crystallizing in larger 24-well hanging drops corresponding to initial mixtures of 1 µL well solution and 1 µL protein solution equilibrated against 1000 µL reservoirs.

The crystal of the designed protein in the absence of ligand ("D_3_633_x8_Apo") was grown from 2 M ammonium sulfate and 5% 2-propanol at a protein concentration of 216.92 µM. The crystal was transferred to a solution containing 2 M ammonium sulfate and 25% sucrose and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.2 at wavelength 400 nm and processed using program HKL20001. The crystal was found to belong to a primitive orthorhombic space group (P212121) and yielded a 2 Å resolution data set.

The crystal of the designed protein in the presence of ligand ("D_3_633_x8_PJS-l-16") was grown from 2.5 M ammonium sulfate and 4% 2-propanol at a protein concentration of 43.38 µM and ligand concentration of 86.76 µM. The crystal was transferred to a solution containing 2.5 M ammonium sulfate and 25% sucrose and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.2 at wavelength 400 nm and processed using program HKL20001. The crystal was found to belong to a primitive monoclinic space group (P21) and yielded a 2.05 Å resolution data set.

Phasing and refinement.

The structures of D_3_633, D_3_633_x8_Apo and D_3_633_x8_PJS-l-16 were solved by Molecular Replacement with Phaser2 via PHENIX3 using the coordinates of the computationally

designed structure as a search model. The structures were then built and refined using Coot4 and PHENIX5, respectively. For the PJS-1-16 bound structure, the protein and visible surrounding ligands and solvent were modeled and refined (while avoiding the modeling of any atoms within the binding site, which displayed unambiguous density for the bound circular peptide ligand from the first rounds of modeling onwards). The final rounds of model-building were focused on fitting PJS-l-16 into the unbiased density that remained in the binding pocket. PJS-l-16 energies were calculated using eLBOW6 via PHENIX3.

Final Ramachandran statistics after refinement were as follows (given as % preferred, % allowed, % outliers, respectively): D_3_633: 98.35, 1.42, 0.24; D_3_633_x8_Apo: 99.33, 0.67, 0.0; D_3_633_x8_PJS-l-16: 99.33, 0.45, 0.22.

Crystallographic Data Collection Statistics

	D_3_633
Wavelength	
Resolution range	48.05- 2.324 (2.407 - 2.324)
Space group	P 1 21 1
Unit cell	54.445 54.682 86.987 90 95.469 90
Total reflections	469049
Unique reflections	21183 (1929)
Multiplicity	4.0 (4.1)
Completeness (%)	95.19 (87.13)
Mean I/sigma(I)	15.8 (1.19)
Wilson B-factor	56.62
R-merge	0.078 (1.029)
R-meas	0.092 (1.184)
R-pim	0.047 (0.581)

CC1/2	0.658
Chi**2	1.06
Reflections used in refinement	21134 (1929)
Reflections used for R-free	1996 (182)
R-work	0.2650 (0.3159)
R-free	0.3131 (0.3723)
Number of non-hydrogen atoms	2974
macromolecules	2949
ligands	20
solvent	5
Protein residues	436
RMS(bonds)	0.002
RMS(angles)	0.43
Ramachandran favored (%)	98.35
Ramachandran allowed (%)	1.42
Ramachandran outliers (%)	0.24
Rotamer outliers (%)	0
Clashscore	4.46
Average B-factor	74.08
macromolecules	74.06
ligands	78.17
solvent	71.19
Number of TLS groups	1

Crystallographic Data Collection Statistics

	D_3_633x8_PJS-l-16
Wavelength	400
Resolution range	47.4- 2.052 (2.126 - 2.052)
Space group	P 1 21 1

Unit cell	54.717 54.562 80.478 90 96.113 90
Total reflections	1069869
Unique reflections	29557 (2789)
Multiplicity	6.3 (4.8)
Completeness (%)	99.18 (94.18)
Mean I/sigma(I)	22.05 (2.07)
Wilson B-factor	36
R-merge	0.074 (0.420)
R-meas	0.081 (0.467)
R-pim	0.031 (0.199)
CC1/2	0.943
Chi**2	0.576
Reflections used in refinement	29539 (2782)
Reflections used for R-free	1999 (187)
R-work	0.2052 (0.2517)
R-free	0.2431 (0.2915)
Number of non-hydrogen atoms	3635
macromolecules	3441
ligands	93
solvent	101
Protein residues	452
RMS(bonds)	0.002
RMS(angles)	0.4
Ramachandran favored (%)	99.33
Ramachandran allowed (%)	0.45
Ramachandran outliers (%)	0.22
Rotamer outliers (%)	0
Clashscore	2.86
Average B-factor	47.52

macromolecules	47.02
ligands	66.89
solvent	46.83
Number of TLS groups	1

Crystallographic Data Collection Statistics

	D_3_633x8_Apo
Wavelength	400
Resolution range	44- 2.0 (2.071 - 2.0)
Space group	P 21 21 21
Unit cell	54.379 55.025 146.534 90 90 90
Total reflections	1886282
Unique reflections	30453 (2906)
Multiplicity	11.8 (6.4)
Completeness (%)	99.62 (96.41)
Mean I/sigma(I)	31.36 (0.6)
Wilson B-factor	48.41
R-merge	0.059 (1.546)
R-meas	0.061 (1.677)
R-pim	0.017 (0.626)
CC1/2	0.331
Chi**2	0.45
Reflections used in refinement	30414 (2871)
Reflections used for R-free	1994 (187)
R-work	0.2329 (0.3454)
R-free	0.2814 (0.3844)
Number of non-hydrogen atoms	3329
macromolecules	3282
ligands	5

solvent	42
Protein residues	450
RMS(bonds)	0.007
RMS(angles)	0.81
Ramachandran favored (%)	99.33
Ramachandran allowed (%)	0.67
Ramachandran outliers (%)	0
Rotamer outliers (%)	0
Clashscore	6.04
Average B-factor	52.82
macromolecules	52.78
ligands	80.28
solvent	53.06

D_3_114_1

Crystallization and data collection.

Purified protein D_3_114_1 without ligand, with ZnPPaM, and with ZnCe6 were initially tested for crystallization via sparse matrix screens in 96-well sitting drops (200 nL drop volumes versus 95 µL reservoir volumes) using a mosquito crystallization robot (TTP LabTech). Crystallization conditions were then optimized with constructs that proved capable of crystallizing in larger 24-well hanging drops corresponding to initial mixtures of 1 µL well solution and 1 µL protein solution equilibrated against 1000 µL reservoirs at 937.3 µM and 468.6 µM protein concentration.

The crystal of the designed protein in the absence of ligand ("D_3_114_1_Apo") was grown from 27% PEG 3350 and 200mM KCl at a protein concentration of 937.3 µM with 2 uL of well solution and 1 uL of protein solution in the drop. The crystal was transferred to a solution containing 50% PEG 3350 and flash frozen in liquid nitrogen. Data were collected at ALS

Beamline 5.0.1 at wavelength 450 nm and processed using program HKL20001. The crystal was found to belong to orthorhombic space group (C222) and yielded a 2.65 Å resolution data set.

The crystal of the designed protein in the absence of ligand ("D_3_114_1_ZnPPaM_twinned") was grown from 27% PEG 3350 and 200mM KCl at a protein concentration of 468.6 μM and ligand concentration of 468.6 μM with 2uL of well solution and 1uL of protein solution in the drop. The crystal was transferred to a solution containing 50% PEG 3350 and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.1 at wavelength 450 nm and processed using program HKL20001. Using twinning laws gained from Xtriage PHENIX3, the crystal was found to belong to orthorhombic space group (C222) and yielded a 2.84 Å resolution data set.

The crystal of the designed protein in the presence of ligand ZnPPaM ("D_3_114_1_ZnPPaM") was grown from 23% PEG 3350 and 180mM KCl at a protein concentration of 468.6 μM and ligand concentration of 468.6 μM with 1 uL of well solution and 1 uL of protein solution in the drop. The crystal was transferred to a solution containing 50% PEG 3350 and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.1 at wavelength 400 nm and processed using program HKL20001. The crystal was found to belong to a primitive monoclinic space group (P21) and yielded a 2.95 Å resolution data set.

The crystal of the designed protein in the presence of ligand ZnCe6 ("D_3_114_1_ZnCe6") was grown from 29% PEG 3350 and 180mM KCl at a protein concentration of 431.1 μM and ligand concentration of 431.1 μM with 2 uL of well solution and 1 uL of protein solution in the drop. The crystal was transferred to a solution containing 50% PEG 3350 and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.1 at wavelength 400 nm and processed using

program HKL20001. The crystal was found to belong to a primitive monoclinic space group (P21) and yielded a 2.98 Å resolution data set.

Phasing and refinement.

The structures D_3_114_1_Apo, D_3_114_1_ZnPPaM and D_3_114_1_Ce6 were solved by Molecular Replacement with Phaser2 via PHENIX3 using the coordinates of the computationally designed structure as a search model. The structures were then built and refined using Coot4 and PHENIX5, respectively. For the ligand bound structures, the protein and visible surrounding ligands and solvent were modeled and refined (while avoiding the modeling of any atoms within the binding site, which displayed unambiguous density for the bound ligand from the first rounds of modeling onwards). The final rounds of model-building were focused on fitting the ligands into the unbiased density that remained in the binding pocket. The ligand density in each chlorophyll bound structure isn't fully resolved and cannot currently be built into the density with confidence, though there is definitely a ligand present.

Final Ramachandran statistics after refinement were as follows (given as % preferred, % allowed, % outliers, respectively): D_3_114_1_ZnPPaM_twinned: 78.48, 14.80, 6.73; D_3_114_1_Apo: still in progress; D_3_114_1_ZnPPaM:still in progress; D_3_114_1_Ce6: still in progress.

Crystallographic Data Collection Statistics

	D_3_114_1_ZnPPaM_twinned
Wavelength	450nm
Resolution range	26.72 - 2.844 (2.946 - 2.844)
Space group	C 2 2 21
Unit cell	57.284 97.151 83.684 90 90 90

Total reflections	597601
Unique reflections	5534 (434)
Multiplicity	4.9 (3.8)
Completeness (%)	95.98 (76.87)
Mean I/sigma(I)	13.79 (1.51)
Wilson B-factor	89.77
R-merge	0.082 (0.510)
R-meas	0.091 (0.581)
R-pim	0.040 (0.27)
CC1/2	0.881
Chi**2	1.007
Reflections used in refinement	5521 (432)
Reflections used for R-free	556 (40)
R-work	0.3643 (0.5174)
R-free	0.3558 (0.6813)
Number of non-hydrogen atoms	1434
macromolecules	1434
Protein residues	233
RMS(bonds)	0.007
RMS(angles)	0.95
Ramachandran favored (%)	78.48
Ramachandran allowed (%)	14.8
Ramachandran outliers (%)	6.73
Rotamer outliers (%)	0
Clashscore	22.32
Average B-factor	83.5
macromolecules	83.5

Crystallography Citations.

1. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997)
2. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007)
3. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010)
4. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010)
5. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012)
6. Electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. N.W. Moriarty, R.W. Grosse-Kunstleve, and P.D. Adams. *Acta Crystallogr D Biol Crystallogr* **65**, 1074-80 (2009)

5.7 CHLOROPHYLL BINDING TITRATIONS PERFORMED BY CD SPECTROSCOPY.

All titrations were performed in 2.0 mL aqueous buffer containing 50 mM NaCl and 10 mM Tris buffer at pH 8.0 in a 1-cm path length quartz cuvette with a magnetic stir bar. Samples were prepared with protein monomer concentrations of ~10-20 μM for CD titrations. A stock solution of 100 – 600 μM chlorophyll in dimethyl sulfoxide was titrated into the protein solution in steps of 0.1 equivalents of cofactor per protein binding site until after the binding capacity of the protein appeared to have been reached, at which point larger additions of chlorophyll were added for each step until about 3.0 equivalents had been reached. For each step, at least 10 minutes of equilibration with stirring was allowed before data collection. In CD titrations, spectra were measured in the Q-band region for chlorins (550 – 800 nm) sampling every 1.0 nm at 50 – 100 nm/min with a bandwidth of 1 nm.

5.8 PEPTIDE BINDING BY NATIVE MASS SPECTROMETRY

Native mass spectrometry (nMS) was conducted as previously described (Pyles et al., 2019). Samples were analyzed with and without peptide added to the sample at 10X concentration. The protein samples were at 50 uM monomer (25 uM homodimer) and the peptide was 250 uM. The samples were in a TBS buffer with 5% DMSO. Samples were analyzed with all-ion fragmentation (MSMS) mode with high energy collision-induced dissociation (HCD) 100 V, and with full MS mode (no collision voltage applied).

5.9 PEPTIDE BINDING EQUILIBRIUM DIALYSIS MASS SPECTROMETRY

Equilibrium dialysis mass spectrometry was conducted as described by Thermo Fisher Scientific ([No title], no date). The dialysis experiment was set up with 50 uL of protein at 100 uM on one side of the dialysis membrane and 300 ul of the peptide at 10 uM on the other side. Both sides were in TBS with 5% DMSO. After overnight equilibration at 20 °C and 250 rpm shaking, the ratio of the peptide on the protein side of the membrane compared to the buffer side of the membrane was quantified by mass spectrometry. Binding affinity was estimated using the following equations:

ratio = ligand_on_protein_side / ligand_in_buffer

ligand_on_protein_side = 1 - ligand_in_buffer

ligand_in_buffer = 1 - fraction_ligand_bound - (1 - fraction_ligand_bound)/2

fraction_lig_bound =

(Kd+[prot]+[lig] - sqrt((Kd+[prot]+[lig])^2 - 4*[prot]*[lig])) / (2*[lig])

5.10 COMPUTATIONAL METHODS

Computational code is available on github by request at:

https://github.com/drhicks/derrick_r_hicks_thesis

Please contact me at:

drhicks1 <at> uw <dot> edu or derrick.ray.hicks <at> gmail <dot> com

Code will be made publicly available after publication of relevant manuscripts.

5.11 BACKBONE GENERATION OF CURVED REPEAT PROTEIN MONOMERS

We generate repeat protein backbones using RosettaRemodel which takes a blueprint file as input and performs Monte Carlo fragment assembly using a coarse-grained energy function to accept or reject moves (made identically in each repeat). A blueprint file describes a fixed-length protein by its secondary structure assignment at each residue position. The repeating unit of the repeat proteins we generate includes helix 1, loop 1, helix 2, and loop 2. We limit our search to short “ideal” loops, having a length of 2 to 4 residues, while helix lengths range from 12 to 30 residues. Within these limits, we enumerate all possible secondary structure length combinations

and output a unique blueprint file for each combination. A RosettaScripts XML is used to run RosettaRemodel on each blueprint to create four repeat unit proteins. Remodel begins by picking 3mer and 9mer fragments for each position based on the assigned secondary structure and then performs fragment assembly with these fragments. Trajectories begin with ideal helices at helical positions and extended loops. The protocol first makes fragment insertions at loop regions in order to quickly fold the largely extended protein chain into a globular protein, before performing fragment assembly over the full length of the protein. For this work, we ran 1 million total trajectories evenly distributed over all allowable secondary structure combinations (3249 total combinations), which yielded ~300 models per secondary structure combination. We first did this with a traditional coarse-grained energy function supplemented with higher resolution, backbone only, residue pair motifs harvested from the PDB which generate better packed and designable protein models (Fallas *et al.*, 2017; Brunette *et al.*, 2020).

The aforementioned protocol failed to generate our desired shape of repeat proteins, specifically curved repeat proteins with low helical rise (< 1.5 Å per repeat unit) and significant helical curvature (between 0.7 rad and 1.1 rad) that approximated a half-circle. To overcome this limitation and develop a method to focus sampling towards arbitrary repeat protein shapes, we developed three new score terms based on the helical parameters rise, radius, and curvature (often called omega or twist). At every fragment assembly step during the Monte Carlo protocol, the helical parameters of the protein are calculated as described elsewhere (Hauser *et al.*, 2017) and then the deviation from a specified set of helical parameters is computed, and the sum of these deviations is added to the coarse-grained score function used earlier. The weight of these score terms is increased over the course of the trajectory. Furthermore, the score terms allow the user to use a linear or quadratic penalty based on the deviation or to set the penalty to 0 before or after the

desired value is reached. The ability to turn off the penalty is important if the user wants to achieve a rise = 0, in which case radius approaches infinity, which will cause scoring problems to arise for the radius term. In our study, we set the penalties to quadratic. Small scale testing showed this was appropriate, so we moved to larger-scale sampling identical to that described in the preceding paragraph, which generated an abundance of repeat protein with our desired helical parameters. We note that many trajectories got stuck in extended or spaghetti-like models that satisfied our desired helical parameters well at the expense of the other score terms. Future work should attempt to optimize the weights and functional form of these penalties with respect to the rest of the coarse-grained score function, which I believe will be project specific.

5.12 SEQUENCE DESIGN AND SELECTION OF CURVED REPEAT PROTEINS

The fragment assembly methods used to generate backbones will often create loops that are quite different (> 0.4 Å rmsd) than those found in nature by recombining fragments in novel ways, which has been found to lower folding accuracy in subsequent in silico forward folding simulations (Brunette *et al.*, 2020) and potentially during in vitro experiments too. To overcome this problem, every backbone is first subjected to a protocol that optimizes loops by replacing both unique loops in the repeat unit with all combinations of loops found in the PDB with < 0.4 Å rmsd and then propagating these changes to each of the other repeat units (Brunette *et al.*, 2020). This often results in input monomers creating several output monomers, although some inputs fail to find any loops with rmsd < 0.4 Å and are discarded at this step.

After loop optimization, backbones are subjected to sequence optimization using a RosettaScripts XML. First, explicit repeat protein symmetry is applied to the computational model

such that all subsequent steps of sequence mutation and backbone minimization will be done identically at each repeat position in the protein. The application of symmetry also reduces the design space to that of a single repeat unit plus its interface with neighboring units regardless of the number of repeat units (always 4 units in this work), which reduces the subsequent design time ~4 fold. Next, PSSM generated from the sequence of 9-mer fragments found in the PDB and having low rmsd to the design model is applied to the protein to bias the score function towards the design of structurally appropriate amino acids. This 9-mer based structural PSSM is particularly helpful in the design of loop residues, i.e., in placing prolines and glycines in highly favorable positions where either inaccuracy in the Rosetta energy function would favor other residues or the lowest energy residue is not the best residue because of negative design considerations that exist outside absolute energetics. Similarly, this PSSM also helps in placing appropriate helix capping motifs such as aspartate, asparagine, serine, or threonine. Finally, the score function is modified to include an explicit penalty for buried unsatisfied hydrogen bonds (Coventry and Baker, 2020).

FastDesign is then performed, which alternates between sequence optimization (amino acid mutations and rotamer exchanges) and backbone minimization through four cycles during which the weight of the repulsive energy term is ramped from low to high. LayerDesign is used during design to make sure the protein surface positions only mutate to polar residues. The protein core layer definition used allows mutations to small polar residues (D, H, N, Q, S, and T), but the buried unsatisfied penalty either prevents these entirely or forces them into fully satisfied polar networks which are sometimes desirable (Boyken *et al.*, 2016). Next FastRelax is run which alternates between rotamer packing and backbone minimization similar to FastDesign. These two steps are then repeated, after which various filter metrics and scores are calculated to quantify the

protein energetics, core packing, secondary structure shape complementarity, sequence-structure agreement, and buried unsatisfied hydrogen bonds.

Top designs based on protein energetics, core packing, secondary structure shape complementarity, sequence-structure agreement, and buried unsatisfied hydrogen bonds were subjected to forward folding simulations. Folding funnels with a forward folding metric of less than 25 were chosen for subsequent design steps.

After forward folding, the N-terminus and C-terminus of the proteins were optimized by replacing terminal helices making minimal contacts to the rest of the protein with better packing termini. This was done by allowing the termini to extend or shorten by up to half a repeat unit to optimize the score per residue of the protein. We attempted to maintain the protein length as close to four repeat units as possible but favored lengthening the protein by half a repeat unit over shortening the protein. In the case that both termini were better when extended by half a repeat unit, which is functionally similar to truncating by half a repeat unit, we would make two combinations of the protein, first, a protein with the N-terminus extended and C-terminus truncated and second, a protein with the N-terminus truncated and C-terminus extended.

Finally, the whole protein surface was allowed to redesign for three reasons. The first was to break up the repetitive nature of the protein sequence, which makes DNA synthesis easier. The second was to allow better electrostatic complimentary on the surface because repulsive charges were sometimes forced next to each other due to the explicit repeat symmetry used to initially design the monomers. The third was to remove surface-exposed hydrophobic residues.

5.13 C2 SYMMETRIC DOCKING

We adapted a previous symmetric docking approach (Fallas *et al.*, 2017) by adding a requirement that the N and C terminal helices of the monomers contact (at least one pair of residues on each terminus within 14 Å) in the dimer; this leads to head to tail homodimers with a closed circular structure and a central cavity along the axis of symmetry. Subsequently, we removed docks with small interfaces (less than 10 contacting residues) and excessively large interfaces (greater than 24 contacts).

5.14 C2 SYMMETRIC PROTEIN PROTEIN INTERFACE DESIGN

Interface design was conducted by running a RosettaScripts XML that is highly similar to the one used to design monomers. The critical difference being that only interface residues are allowed to mutate during interface design. Interface residues are defined as residues with Ca atoms in one protein chain being within 10 Å of Ca atoms in the other protein chain. Additionally, C2 symmetry is applied during interface design, which serves a similar function to the repeat protein symmetry used during monomer design. C2 symmetry forces chain 1 and chain 2 to maintain identical sequences (and rotamers) during sequence optimization and identical torsion angles during minimization.

A similar trajectory of FastDesign, FastRelax, FastDesign, FastRelax is run as during monomer design. Loops, as defined by DSSP, are prevented from designing. Additionally, glycine and proline residues are not allowed to mutate, and no other positions can mutate to glycine or proline. Finally, we use an amino acid composition score term along with an explicit penalty for

buried unsatisfied hydrogen bonds in order to favor the creation of small fully satisfied polar networks at the interface. This is done in the hopes of preventing the creation of large hydrophobic interfaces that might be prone to aggregation and to increase binding specificity of the interfaces. Finally, filter metrics and scores were calculated to quantify the binding energy, interface packing, interface shape complementarity, interface SASA, and buried unsatisfied hydrogen bonds across the interface.

5.15 CHLOROPHYLL BINDER DESIGN

C2 symmetric chlorophyll pairs were harvested from the PDB, and then an ensemble of C2 symmetric confirmations was sampled by perturbing the rigid body orientation of the molecule while maintaining C2 symmetry. Low energy conformations were chosen and then inverse histidine rotamers were built off of them according to the geometry of histidine relative to Mg found in the PDB (Chakrabarti, 1990). The relative transform between the two histidine backbones for all inverse rotamers was then stored in a hash table. This hash table was subsequently used for docking chlorophyll dimers into scaffolds by searching for matching backbone transforms at symmetric positions in a library of scaffold proteins. When matching sites were found, the histidine rotamers were placed into the scaffold and the chlorophyll molecules built off of the histidine. After initial placement, chlorophyll is free to rotate around an axis connecting the Mg metal of chlorophyll to the coordinating nitrogen in histidine in order to minimize clashing with the scaffold backbone.

The sequence of the binding site was then optimized using a RosettaScripts XML with mutations limited to the interface between the protein and chlorophyll. Explicit C2 symmetry was maintained during design. Proteins were filtered for experimental characterization by a combination of metrics that quantify the protein-ligand binding energy, shape complementarity, ligand burial, and geometry of histidine relative to chlorophyll.

5.16 PEPTIDE DESIGN

In the first step of C2 symmetric peptide design, low energy conformers of each of the canonical or non-canonical "residues", or 1-mers, are generated. In the second step, N 1-mers are chosen to attempt to build a C2-symmetric 2N-mer, and the rigid body transformations associated with these monomer conformations are computed. In the third step, all rigid body transforms of the resulting N-mer are computed using simple matrix multiplication of the outer-product of the chosen monomers' conformers. This outer-product typically results in ~10^6 N-mer rigid-body transforms. In the final step, N-mer conformers that result in C2-symmetric 2N-mers are identified by calculating the angle of rotation and translation about the rotation axis of the transforms. This angle must equal 180 degrees while the translation must be zero to satisfy C2 symmetry. Full-atom representations of the resulting combinations of the chosen 1-mers conformers that satisfy these criteria are built and minimized to yield a cyclic C2 symmetric peptide.

5.17 PEPTIDE BINDER DESIGN

C2 symmetric peptides were docked into C2 symmetric scaffolds using a rifgen/rifdock protocol. First, the peptide was placed at the origin (x, y, z = 0, 0, 0) and with the symmetry axis aligned to the z-axis. Next, the rifgen protocol was run which enumerates good scoring hydrophobic and hydrogen bonding interaction with the peptide by placing disembodied side chains around the peptide. The relative transform of the backbone atoms of each inverse rotamer for each low scoring sidechain interaction (many millions) were saved in a hash table. The center of mass of scaffold proteins was then placed at the origin and the protein's symmetry axis aligned to the z-axis. These scaffold proteins were then docked around the peptide by sampling C2 symmetric rigid body perturbations enumerating a high-resolution grid. At each sampling position, favorable sidechain interactions were recovered from the hash table, and the energies of all compatible interactions were summed, and the lowest energy docking positions were output.

The sequence of the binding site was then optimized using a RosettaScripts XML with mutations limited to the interface between the protein and the peptide. Explicit C2 symmetry was maintained during design. This involved deleting half of the peptide, then applying symmetry to rebuild the missing half, declaring appropriate bonds between the symmetric peptide chains, and apply distance, angle, and dihedral constraints to maintain proper peptide bond geometry for the C2 symmetric cyclic peptide. Binder designs were filtered for experimental characterization by a combination of metrics that quantify the protein-ligand binding energy, shape complementarity, ligand burial, protein-peptide hydrogen bonds, and unsatisfied hydrogen bonds in the interface.

BIBLIOGRAPHY

AKERFELDT, K. S. *ET AL.* (1992) 'TETRAPHILIN: A FOUR-HELIX PROTON CHANNEL BUILT ON A TETRAPHENYLPORPHYRIN FRAMEWORK', *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, PP. 9656–9657. DOI: 10.1021/JA00050A054.

ANDRÉ, B. (1995) 'AN OVERVIEW OF MEMBRANE TRANSPORT PROTEINS INSACCHAROMYCES CEREVISIAE', *YEAST*, PP. 1575–1611. DOI: 10.1002/YEA.320111605.

ANFINSEN, C. B. (1973) 'PRINCIPLES THAT GOVERN THE FOLDING OF PROTEIN CHAINS', *SCIENCE*, 181(4096), PP. 223–230.

A. SPINELLI, G. P. (NO DATE) 'COVID-19 PANDEMIC: PERSPECTIVES ON AN UNFOLDING CRISIS', *THE BRITISH JOURNAL OF SURGERY*. DOI: 10.1002/BJS.11627.

BALE, J. B. *ET AL.* (2016) 'ACCURATE DESIGN OF MEGADALTON-SCALE TWO-COMPONENT ICOSAHEDRAL PROTEIN COMPLEXES', *SCIENCE*, 353(6297), PP. 389–394.

BASANTA, B. *ET AL.* (2020) 'AN ENUMERATIVE ALGORITHM FOR DE NOVO DESIGN OF PROTEINS WITH DIVERSE POCKET STRUCTURES', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 117(36), PP. 22135–22145.

BAX, A. AND CLORE, G. M. (2019) 'PROTEIN NMR: BOUNDLESS OPPORTUNITIES', *JOURNAL OF MAGNETIC RESONANCE* , 306, PP. 187–191.

BENDER, G. M. *ET AL.* (2007) 'DE NOVO DESIGN OF A SINGLE-CHAIN DIPHENYLPORPHYRIN METALLOPROTEIN', *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 129(35). DOI: 10.1021/JA071199J.

BENKOVIC, S. J., VALENTINE, A. M. AND SALINAS, F. (2001) 'REPLISOME-MEDIATED DNA REPLICATION', *ANNUAL REVIEW OF BIOCHEMISTRY*, 70, PP. 181–208.

BERMAN, H. M. *ET AL.* (2000) 'THE PROTEIN DATA BANK', *NUCLEIC ACIDS RESEARCH*, 28(1), PP. 235–242.

BERWICK, M. R. *ET AL.* (2014) 'DE NOVO DESIGN OF LN(III) COILED COILS FOR IMAGING APPLICATIONS', *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 136(4), PP. 1166–1169.

BHARDWAJ, G. *ET AL.* (2016) 'ACCURATE DE NOVO DESIGN OF HYPERSTABLE CONSTRAINED PEPTIDES', *NATURE*, 538(7625), PP. 329–335.

BICK, M. J. *ET AL.* (2017) 'COMPUTATIONAL DESIGN OF ENVIRONMENTAL SENSORS FOR THE POTENT OPIOID FENTANYL', *ELIFE*, 6. DOI: 10.7554/ELIFE.28909.

BOYKEN, S. E. *ET AL.* (2016) 'DE NOVO DESIGN OF PROTEIN HOMO-OLIGOMERS WITH MODULAR HYDROGEN-BOND NETWORK-MEDIATED SPECIFICITY', *SCIENCE*, 352(6286), PP. 680–687.

BROO, K. S. *ET AL.* (1997) 'CATALYSIS OF HYDROLYSIS AND TRANSESTERIFICATION REACTIONS OFP-NITROPHENYL ESTERS BY A DESIGNED HELIX−LOOP−HELIX DIMER', *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, PP. 11362–11372. DOI: 10.1021/JA970854S.

BRUNETTE, T. J. *ET AL.* (2015) 'EXPLORING THE REPEAT PROTEIN UNIVERSE THROUGH COMPUTATIONAL PROTEIN DESIGN', *NATURE*, 528(7583), PP. 580–584.

BRUNETTE, T. J. *ET AL.* (2020) 'MODULAR REPEAT PROTEIN SCULPTING USING RIGID HELICAL JUNCTIONS', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 117(16), PP. 8870–8875.

BRYNGELSON, J. D. *ET AL.* (1995) 'FUNNELS, PATHWAYS, AND THE ENERGY LANDSCAPE OF PROTEIN FOLDING: A SYNTHESIS', *PROTEINS*, 21(3), PP. 167–195.

BURTON, A. J. *ET AL.* (2016) 'INSTALLING HYDROLYTIC ACTIVITY INTO A COMPLETELY DE NOVO PROTEIN FRAMEWORK', *NATURE CHEMISTRY*, 8(9), PP. 837–844.

BUTTERFIELD, G. L. *ET AL.* (2017) 'EVOLUTION OF A DESIGNED PROTEIN ASSEMBLY ENCAPSULATING ITS OWN RNA GENOME', *NATURE*, 552(7685), PP. 415–420.

BYWATER, R. P. (2018) 'WHY TWENTY AMINO ACID RESIDUE TYPES SUFFICE(D) TO SUPPORT ALL LIVING SYSTEMS', *PLOS ONE*, 13(10), P. E0204883.

CALHOUN, J. R. *ET AL.* (2003) 'COMPUTATIONAL DESIGN AND CHARACTERIZATION OF A MONOMERIC HELICAL DINUCLEAR METALLOPROTEIN', *JOURNAL OF MOLECULAR BIOLOGY*, 334(5), PP. 1101–1115.

CAO, L. *ET AL.* (2020) 'DE NOVO DESIGN OF PICOMOLAR SARS-COV-2 MINIPROTEIN INHIBITORS', *SCIENCE*. DOI: 10.1126/SCIENCE.ABD9909.

CHAKRABARTI, P. (1990) 'GEOMETRY OF INTERACTION OF METAL IONS WITH HISTIDINE RESIDUES IN PROTEIN STRUCTURES', *PROTEIN ENGINEERING*, 4(1), PP. 57–63.

CHEN, Z. *ET AL.* (2019) 'PROGRAMMABLE DESIGN OF ORTHOGONAL PROTEIN HETERODIMERS', *NATURE*, 565(7737), PP. 106–111.

CHERNY, I. *ET AL.* (2012) 'PROTEINS FROM AN UNEVOLVED LIBRARY OF DE NOVO DESIGNED SEQUENCES BIND A RANGE OF SMALL MOLECULES', *ACS SYNTHETIC BIOLOGY*, 1(4), PP. 130–138.

CHEVALIER, A. *ET AL.* (2017) 'MASSIVELY PARALLEL DE NOVO PROTEIN DESIGN FOR TARGETED THERAPEUTICS', *NATURE*, 550(7674), PP. 74–79.

CHRISTOPHER FRY, H. *ET AL.* (2010) 'COMPUTATIONAL DESIGN AND ELABORATION OF A DE NOVO HETEROTETRAMERIC A-HELICAL PROTEIN THAT SELECTIVELY BINDS AN EMISSIVE ABIOLOGICAL (PORPHINATO)ZINC CHROMOPHORE', *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 132(11), P. 3997.

COSKUN, O. (2016) 'SEPARATION TECHNIQUES: CHROMATOGRAPHY', *NORTHERN CLINICS OF ISTANBUL*, 3(2), P. 156.

COVENTRY, B. AND BAKER, D. (2020) 'PROTEIN SEQUENCE OPTIMIZATION WITH A PAIRWISE DECOMPOSABLE PENALTY FOR BURIED UNSATISFIED HYDROGEN BONDS', *COLD SPRING HARBOR LABORATORY*. DOI: 10.1101/2020.06.17.156646.

CREIGHTON, T. E. (1990) 'PROTEIN FOLDING', *BIOCHEMICAL JOURNAL*, 270(1), PP. 1–16.

DAHIYAT, B. I. AND MAYO, S. L. (1997) 'DE NOVO PROTEIN DESIGN: FULLY AUTOMATED SEQUENCE SELECTION', *SCIENCE*, 278(5335), PP. 82–87.

DANEV, R., YANAGISAWA, H. AND KIKKAWA, M. (2019) 'CRYO-ELECTRON MICROSCOPY METHODOLOGY: CURRENT ASPECTS AND FUTURE DIRECTIONS', *TRENDS IN BIOCHEMICAL SCIENCES*, 44(10), PP. 837–848.

DANG, B. *ET AL.* (2017) 'DE NOVO DESIGN OF COVALENTLY CONSTRAINED MESOSIZE PROTEIN SCAFFOLDS WITH UNIQUE TERTIARY STRUCTURES', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 114(41), PP. 10852–10857.

DER, B. S. *ET AL.* (2012) 'METAL-MEDIATED AFFINITY AND ORIENTATION SPECIFICITY IN A COMPUTATIONALLY DESIGNED PROTEIN HOMODIMER', *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 134(1). DOI: 10.1021/JA208015J.

DER, B. S., EDWARDS, D. R. AND KUHLMAN, B. (2012) 'CATALYSIS BY A DE NOVO ZINC-MEDIATED PROTEIN INTERFACE: IMPLICATIONS FOR NATURAL ENZYME EVOLUTION AND RATIONAL ENZYME ENGINEERING', *BIOCHEMISTRY*, 51(18), PP. 3933–3940.

'DESIGN AND EVOLUTION OF ENZYMES FOR NON-NATURAL CHEMISTRY' (2017) *CURRENT OPINION IN GREEN AND SUSTAINABLE CHEMISTRY*, 7, PP. 23–30.

DOU, J. *ET AL.* (2018) 'DE NOVO DESIGN OF A FLUORESCENCE-ACTIVATING B-BARREL', *NATURE*, 561(7724), PP. 485–491.

DOYLE, L. *ET AL.* (2015) 'RATIONAL DESIGN OF A-HELICAL TANDEM REPEAT PROTEINS WITH CLOSED ARCHITECTURES', *NATURE*, 528(7583), PP. 585–588.

DUNBRACK, R. L., JR AND KARPLUS, M. (1994) 'CONFORMATIONAL ANALYSIS OF THE BACKBONE-DEPENDENT ROTAMER PREFERENCES OF PROTEIN SIDECHAINS', *NATURE STRUCTURAL BIOLOGY*, 1(5), PP. 334–340.

DURAN, A. M. AND MEILER, J. (2018) 'COMPUTATIONAL DESIGN OF MEMBRANE PROTEINS USING ROSETTAMEMBRANE', *PROTEIN SCIENCE: A PUBLICATION OF THE PROTEIN SOCIETY*, 27(1), PP. 341–355.

DYER, K. N. *ET AL.* (2014) 'HIGH-THROUGHPUT SAXS FOR THE CHARACTERIZATION OF BIOMOLECULES IN SOLUTION: A PRACTICAL APPROACH', *METHODS IN MOLECULAR BIOLOGY* , 1091, PP. 245–258.

EDWARDSON, T. G. W. AND HILVERT, D. (2019) 'VIRUS-INSPIRED FUNCTION IN ENGINEERED PROTEIN CAGES', *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 141(24), PP. 9432–9443.

FAIELLA, M. *ET AL.* (2009) 'AN ARTIFICIAL DI-IRON OXO-PROTEIN WITH PHENOL OXIDASE ACTIVITY', *NATURE CHEMICAL BIOLOGY*, 5(12), PP. 882–884.

FALLAS, J. A. *ET AL.* (2017) 'COMPUTATIONAL DESIGN OF SELF-ASSEMBLING CYCLIC PROTEIN HOMO-OLIGOMERS', *NATURE CHEMISTRY*, 9(4), PP. 353–360.

FIGUEROA, M. *ET AL.* (2013) 'OCTARELLIN VI: USING ROSETTA TO DESIGN A PUTATIVE ARTIFICIAL (B/A)8 PROTEIN', *PLOS ONE*, 8(8), P. E71858.

FOXS SERVER: FAST X-RAY SCATTERING (NO DATE). AVAILABLE AT: HTTPS://MODBASE.COMPBIO.UCSF.EDU/FOXS/ (ACCESSED: 17 SEPTEMBER 2020).

GIGER, L. *ET AL.* (2013) 'EVOLUTION OF A DESIGNED RETRO-ALDOLASE LEADS TO COMPLETE ACTIVE SITE REMODELING', *NATURE CHEMICAL BIOLOGY*, 9(8), PP. 494–498.

GONEN, S. *ET AL.* (2015) 'DESIGN OF ORDERED TWO-DIMENSIONAL ARRAYS MEDIATED BY NONCOVALENT PROTEIN-PROTEIN INTERFACES', *SCIENCE*, PP. 1365–1368. DOI: 10.1126/SCIENCE.AAA9897.

GOOD, M. C., ZALATAN, J. G. AND LIM, W. A. (2011) 'SCAFFOLD PROTEINS: HUBS FOR CONTROLLING

THE FLOW OF CELLULAR INFORMATION', *SCIENCE*, 332(6030), PP. 680–686.

GOODSELL, D. S. AND OLSON, A. J. (2000) 'STRUCTURAL SYMMETRY AND PROTEIN FUNCTION', *ANNUAL REVIEW OF BIOPHYSICS AND BIOMOLECULAR STRUCTURE*, PP. 105–153. DOI: 10.1146/ANNUREV.BIOPHYS.29.1.105.

GOPARAJU, G. *ET AL.* (2016) 'FIRST PRINCIPLES DESIGN OF A CORE BIOENERGETIC TRANSMEMBRANE ELECTRON-TRANSFER PROTEIN', *BIOCHIMICA ET BIOPHYSICA ACTA*, 1857(5), PP. 503–512.

GRISHINA, I. B. AND WOODY, R. W. (1994) 'CONTRIBUTIONS OF TRYPTOPHAN SIDE CHAINS TO THE CIRCULAR DICHROISM OF GLOBULAR PROTEINS: EXCITON COUPLETS AND COUPLED OSCILLATORS', *FARADAY DISCUSSIONS*, (99), PP. 245–262.

HAMM, H. E. (1998) 'THE MANY FACES OF G PROTEIN SIGNALING', *JOURNAL OF BIOLOGICAL CHEMISTRY*, PP. 669–672. DOI: 10.1074/JBC.273.2.669.

HAUSER, K. *ET AL.* (2017) 'CHARACTERIZATION OF BIOMOLECULAR HELICES AND THEIR COMPLEMENTARITY USING GEOMETRIC ANALYSIS', *JOURNAL OF CHEMICAL INFORMATION AND MODELING*, 57(4), PP. 864–874.

HEO, L. AND FEIG, M. (2020) 'HIGH-ACCURACY PROTEIN STRUCTURES BY COMBINING MACHINE-LEARNING WITH PHYSICS-BASED REFINEMENT', *PROTEINS: STRUCTURE, FUNCTION, AND BIOINFORMATICS*, PP. 637–642. DOI: 10.1002/PROT.25847.

HOCHULI, E. *ET AL.* (1988) 'GENETIC APPROACH TO FACILITATE PURIFICATION OF RECOMBINANT PROTEINS WITH A NOVEL METAL CHELATE ADSORBENT', *NATURE BIOTECHNOLOGY*, PP. 1321–1325. DOI: 10.1038/NBT1188-1321.

HSIA, Y. *ET AL.* (2016) 'DESIGN OF A HYPERSTABLE 60-SUBUNIT PROTEIN DODECAHEDRON. [CORRECTED]', *NATURE*, 535(7610), PP. 136–139.

HUANG, P.-S. *ET AL.* (2016) 'DE NOVO DESIGN OF A FOUR-FOLD SYMMETRIC TIM-BARREL PROTEIN WITH ATOMIC-LEVEL ACCURACY', *NATURE CHEMICAL BIOLOGY*, 12(1), PP. 29–34.

HUBBLE ULTRA DEEP FIELD 2014 (NO DATE). AVAILABLE AT: HTTP://HUBBLESITE.ORG/CONTENTS/MEDIA/IMAGES/2014/27/3380-IMAGE (ACCESSED: 14 SEPTEMBER 2020).

HURA, G. L. *ET AL.* (2013) 'COMPREHENSIVE MACROMOLECULAR CONFORMATIONS MAPPED BY QUANTITATIVE SAXS ANALYSES', *NATURE METHODS*, 10(6), PP. 453–454.

ILARI, A. AND SAVINO, C. (2008) 'PROTEIN STRUCTURE DETERMINATION BY X-RAY CRYSTALLOGRAPHY', *BIOINFORMATICS*, PP. 63–87. DOI: 10.1007/978-1-60327-159-2_3.

JIANG, L. *ET AL.* (2008) 'DE NOVO COMPUTATIONAL DESIGN OF RETRO-ALDOL ENZYMES', *SCIENCE*, 319(5868), PP. 1387–1391.

JOH, N. H. *ET AL.* (2017) 'DESIGN OF SELF-ASSEMBLING TRANSMEMBRANE HELICAL BUNDLES TO ELUCIDATE PRINCIPLES REQUIRED FOR MEMBRANE PROTEIN FOLDING AND ION TRANSPORT', *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B: BIOLOGICAL SCIENCES*, P. 20160214. DOI: 10.1098/RSTB.2016.0214.

JONES, D. T. (1994) 'DE NOVO PROTEIN DESIGN USING PAIRWISE POTENTIALS AND A GENETIC

ALGORITHM', *PROTEIN SCIENCE: A PUBLICATION OF THE PROTEIN SOCIETY*, 3(4), P. 567.

KING, N. P. *ET AL.* (2012) 'COMPUTATIONAL DESIGN OF SELF-ASSEMBLING PROTEIN NANOMATERIALS WITH ATOMIC LEVEL ACCURACY', *SCIENCE*, 336(6085), PP. 1171–1174.

KRYSHTAFOVYCH, A. *ET AL.* (2019) 'CRITICAL ASSESSMENT OF METHODS OF PROTEIN STRUCTURE PREDICTION (CASP)—ROUND XIII', *PROTEINS: STRUCTURE, FUNCTION, AND BIOINFORMATICS*, PP. 1011–1020. DOI: 10.1002/PROT.25823.

KUHLMAN, B. *ET AL.* (2003) 'CRYSTAL STRUCTURE OF TOP7: A COMPUTATIONALLY DESIGNED PROTEIN WITH A NOVEL FOLD'. DOI: 10.2210/PDB1QYS/PDB.

KUHLMAN, B. *ET AL.* (2003) 'DESIGN OF A NOVEL GLOBULAR PROTEIN FOLD WITH ATOMIC-LEVEL ACCURACY', *SCIENCE*, 302(5649), PP. 1364–1368.

KUHLMAN, B. AND BRADLEY, P. (2019) 'ADVANCES IN PROTEIN STRUCTURE PREDICTION AND DESIGN', *NATURE REVIEWS. MOLECULAR CELL BIOLOGY*, 20(11), PP. 681–697.

LANCI, C. J. *ET AL.* (2012) 'COMPUTATIONAL DESIGN OF A PROTEIN CRYSTAL', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 109(19), PP. 7304–7309.

LENEY, A. C. AND HECK, A. J. R. (2017) 'NATIVE MASS SPECTROMETRY: WHAT IS IN THE NAME?', *JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY*, 28(1), PP. 5–13.

LEVINTHAL'S PARADOX (NO DATE). AVAILABLE AT: HTTPS://WEB.ARCHIVE.ORG/WEB/20110523080407/HTTP://WWW-MILLER.CH.CAM.AC.UK/LEVINTHAL/LEVINTHAL.HTML (ACCESSED: 9 SEPTEMBER 2020).

LIN, Y.-R. *ET AL.* (2015) 'CONTROL OVER OVERALL SHAPE AND SIZE IN DE NOVO DESIGNED PROTEINS', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 112(40), PP. E5478–85.

LIPFERT, J. AND DONIACH, S. (2007) 'SMALL-ANGLE X-RAY SCATTERING FROM RNA, PROTEINS, AND PROTEIN COMPLEXES', *ANNUAL REVIEW OF BIOPHYSICS AND BIOMOLECULAR STRUCTURE*, PP. 307–327. DOI: 10.1146/ANNUREV.BIOPHYS.36.040306.132655.

LJUBETIČ, A. *ET AL.* (2017) 'DESIGN OF COILED-COIL PROTEIN-ORIGAMI CAGES THAT SELF-ASSEMBLE IN VITRO AND IN VIVO', *NATURE BIOTECHNOLOGY*, PP. 1094–1101. DOI: 10.1038/NBT.3994.

LU, P. *ET AL.* (2018) 'ACCURATE COMPUTATIONAL DESIGN OF MULTIPASS TRANSMEMBRANE PROTEINS', *SCIENCE*, 359(6379), PP. 1042–1046.

LYNCH, E. M. *ET AL.* (2017) 'HUMAN CTP SYNTHASE FILAMENT STRUCTURE REVEALS THE ACTIVE ENZYME CONFORMATION', *NATURE STRUCTURAL & MOLECULAR BIOLOGY*, 24(6), PP. 507–514.

MARCOS, E. *ET AL.* (2017) 'PRINCIPLES FOR DESIGNING PROTEINS WITH CAVITIES FORMED BY CURVED B SHEETS', *SCIENCE*, 355(6321), PP. 201–206.

MARCOS, E. *ET AL.* (2018) 'DE NOVO DESIGN OF A NON-LOCAL B-SHEET PROTEIN WITH HIGH STABILITY AND ACCURACY', *NATURE STRUCTURAL & MOLECULAR BIOLOGY*, 25(11), PP. 1028–1034.

MATILE, S. *ET AL.* (1995) 'PORPHYRINS: POWERFUL CHROMOPHORES FOR STRUCTURAL STUDIES BY EXCITON-COUPLED CIRCULAR DICHROISM', *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, PP. 7021–

7022. DOI: 10.1021/JA00131A033.

MCLACHLAN, A. D. (1972) 'PROTEIN STRUCTURE AND FUNCTION', *ANNUAL REVIEW OF PHYSICAL CHEMISTRY*, PP. 165–192. DOI: 10.1146/ANNUREV.PC.23.100172.001121.

MUIRHEAD, H. AND PERUTZ, M. F. (1963) 'STRUCTURE OF HAEMOGLOBIN. A THREE-DIMENSIONAL FOURIER SYNTHESIS OF REDUCED HUMAN HAEMOGLOBIN AT 5-5 A RESOLUTION', *NATURE*, 199, PP. 633–638.

[NO TITLE] (NO DATE). AVAILABLE AT: HTTP://TOOLS.THERMOFISHER.COM/CONTENT/SFS/MANUALS/MAN0011571_RED_DEVICE_INSERT_UG.PDF (ACCESSED: 22 SEPTEMBER 2020).

OFFER, G., HICKS, M. R. AND WOOLFSON, D. N. (2002) 'GENERALIZED CRICK EQUATIONS FOR MODELING NONCANONICAL COILED COILS', *JOURNAL OF STRUCTURAL BIOLOGY*, 137(1-2), PP. 41–53.

OVCHINNIKOV, S. *ET AL.* (2017) 'PROTEIN STRUCTURE DETERMINATION USING METAGENOME SEQUENCE DATA', *SCIENCE*, 355(6322), PP. 294–298.

PARK, H. *ET AL.* (2016) 'SIMULTANEOUS OPTIMIZATION OF BIOMOLECULAR ENERGY FUNCTIONS ON FEATURES FROM SMALL MOLECULES AND MACROMOLECULES', *JOURNAL OF CHEMICAL THEORY AND COMPUTATION*, 12(12), PP. 6201–6212.

PAULING, L. AND COREY, R. B. (1951) 'CONFIGURATIONS OF POLYPEPTIDE CHAINS WITH FAVORED ORIENTATIONS AROUND SINGLE BONDS: TWO NEW PLEATED SHEETS', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES*, PP. 729–740. DOI: 10.1073/PNAS.37.11.729.

PAULING, L., COREY, R. B. AND BRANSON, H. R. (1951) 'THE STRUCTURE OF PROTEINS; TWO HYDROGEN-BONDED HELICAL CONFIGURATIONS OF THE POLYPEPTIDE CHAIN', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 37(4), PP. 205–211.

PERUTZ, M. F. (1978) 'HEMOGLOBIN STRUCTURE AND RESPIRATORY TRANSPORT', *SCIENTIFIC AMERICAN*, PP. 92–125. DOI: 10.1038/SCIENTIFICAMERICAN1278-92.

'PROTEIN MASS SPECTROMETRY: APPLICATIONS TO ANALYTICAL BIOTECHNOLOGY' (1995) *JOURNAL OF CHROMATOGRAPHY. A*, 705(1), PP. 21–45.

PYLES, H. *ET AL.* (2019) 'CONTROLLING PROTEIN ASSEMBLY ON INORGANIC CRYSTALS THROUGH DESIGNED PROTEIN INTERFACES', *NATURE*, 571(7764), PP. 251–256.

PYMOL (NO DATE). AVAILABLE AT: HTTPS://PYMOL.ORG/2/ (ACCESSED: 15 SEPTEMBER 2020).

RÄMISCH, S. *ET AL.* (2014) 'COMPUTATIONAL DESIGN OF A LEUCINE-RICH REPEAT PROTEIN WITH A PREDEFINED GEOMETRY', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 111(50), PP. 17875–17880.

READ, R. J. *ET AL.* (2019) 'EVALUATION OF MODEL REFINEMENT IN CASP13', *PROTEINS*, 87(12), PP. 1249–1262.

'RECOMBINANT PROTEIN EXPRESSION IN ESCHERICHIA COLI' (1999) *CURRENT OPINION IN BIOTECHNOLOGY*, 10(5), PP. 411–421.

ROCKLIN, G. J. *ET AL.* (2017) 'GLOBAL ANALYSIS OF PROTEIN FOLDING USING MASSIVELY PARALLEL

DESIGN, SYNTHESIS, AND TESTING', *SCIENCE*, 357(6347), PP. 168–175.

RÖTHLISBERGER, D. *ET AL.* (2008) 'KEMP ELIMINATION CATALYSTS BY COMPUTATIONAL ENZYME DESIGN', *NATURE*, 453(7192), PP. 190–195.

SAHIN, E. AND ROBERTS, C. J. (2012) 'SIZE-EXCLUSION CHROMATOGRAPHY WITH MULTI-ANGLE LIGHT SCATTERING FOR ELUCIDATING PROTEIN AGGREGATION MECHANISMS', IN *THERAPEUTIC PROTEINS*. HUMANA PRESS, TOTOWA, NJ, PP. 403–423.

SAXS FRAMESLICE (NO DATE). AVAILABLE AT: HTTP://SIBYLS.ALS.LBL.GOV/RAN (ACCESSED: 17 SEPTEMBER 2020).

SCHOMBURG, D. AND SALZMANN, M. (1991) 'ENZYME HANDBOOK', *ENZYME HANDBOOK*, PP. 1–1175. DOI: 10.1007/978-3-642-76729-6_1.

SCHRAMM, C. A. *ET AL.* (2012) 'KNOWLEDGE-BASED POTENTIAL FOR POSITIONING MEMBRANE-ASSOCIATED STRUCTURES AND ASSESSING RESIDUE SPECIFIC ENERGETIC CONTRIBUTIONS', *STRUCTURE* , 20(5), P. 924.

SHARP, P. M. AND LI, W. H. (1987) 'THE CODON ADAPTATION INDEX--A MEASURE OF DIRECTIONAL SYNONYMOUS CODON USAGE BIAS, AND ITS POTENTIAL APPLICATIONS', *NUCLEIC ACIDS RESEARCH*, 15(3), PP. 1281–1295.

SHEN, H. *ET AL.* (2018) 'DE NOVO DESIGN OF SELF-ASSEMBLING HELICAL PROTEIN FILAMENTS', *SCIENCE*, 362(6415), PP. 705–709.

SIEGEL, J. B. *ET AL.* (2010) 'COMPUTATIONAL DESIGN OF AN ENZYME CATALYST FOR A STEREOSELECTIVE BIMOLECULAR DIELS-ALDER REACTION', *SCIENCE*, 329(5989), PP. 309–313.

SMITH, B. J. (NO DATE) 'SDS POLYACRYLAMIDE GEL ELECTROPHORESIS OF PROTEINS', *PROTEINS*, PP. 41–56. DOI: 10.1385/0-89603-062-8:41.

STREET, A. G. AND MAYO, S. L. (1999) 'COMPUTATIONAL PROTEIN DESIGN', *STRUCTURE*, PP. R105–R109. DOI: 10.1016/S0969-2126(99)80062-8.

SUN, M. G. F. *ET AL.* (2016) 'PROTEIN ENGINEERING BY HIGHLY PARALLEL SCREENING OF COMPUTATIONALLY DESIGNED VARIANTS', *SCIENCE ADVANCES*, 2(7), P. E1600692.

TANFORD, C. (1978) 'THE HYDROPHOBIC EFFECT AND THE ORGANIZATION OF LIVING MATTER', *SCIENCE*, 200(4345), PP. 1012–1018.

THOMSON, A. R. *ET AL.* (2014) 'COMPUTATIONAL DESIGN OF WATER-SOLUBLE A-HELICAL BARRELS', *SCIENCE* , 346(6208). DOI: 10.1126/SCIENCE.1257452.

TINBERG, C. E. *ET AL.* (2013) 'COMPUTATIONAL DESIGN OF LIGAND-BINDING PROTEINS WITH HIGH AFFINITY AND SELECTIVITY', *NATURE*, 501(7466), PP. 212–216.

VOET, A. R. D. *ET AL.* (2014) 'COMPUTATIONAL DESIGN OF A SELF-ASSEMBLING SYMMETRICAL B-PROPELLER PROTEIN', *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 111(42), PP. 15102–15107.

WATKINS, D. W. *ET AL.* (2017) 'CONSTRUCTION AND IN VIVO ASSEMBLY OF A CATALYTICALLY PROFICIENT AND HYPERTHERMOSTABLE DE NOVO ENZYME', *NATURE COMMUNICATIONS*, 8(1), P. 358.

WEITZNER, B. D. *ET AL.* (2019) 'A COMPUTATIONAL METHOD FOR DESIGN OF CONNECTED CATALYTIC NETWORKS IN PROTEINS', *PROTEIN SCIENCE: A PUBLICATION OF THE PROTEIN SOCIETY*, 28(12), PP. 2036–2041.

WHITMORE, L. *ET AL.* (2010) 'PCDDB: THE PROTEIN CIRCULAR DICHROISM DATA BANK, A REPOSITORY FOR CIRCULAR DICHROISM SPECTRAL AND METADATA', *NUCLEIC ACIDS RESEARCH*, 39(SUPPL_1), PP. D480–D486.

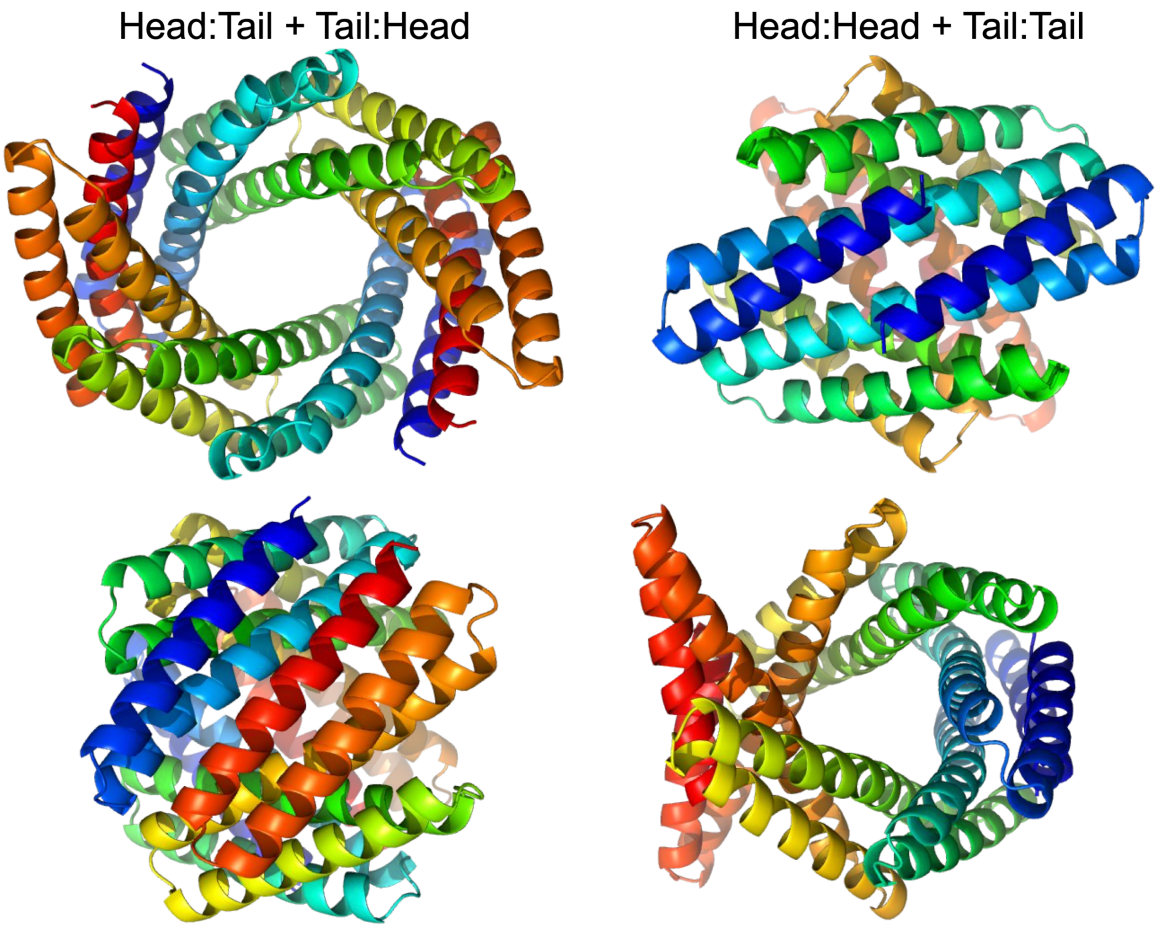
ZANGHELLINI, A. *ET AL.* (2006) 'NEW ALGORITHMS AND AN IN SILICO BENCHMARK FOR COMPUTATIONAL ENZYME DESIGN', *PROTEIN SCIENCE: A PUBLICATION OF THE PROTEIN SOCIETY*, 15(12). DOI: 10.1110/PS.062353106.

APPENDIX

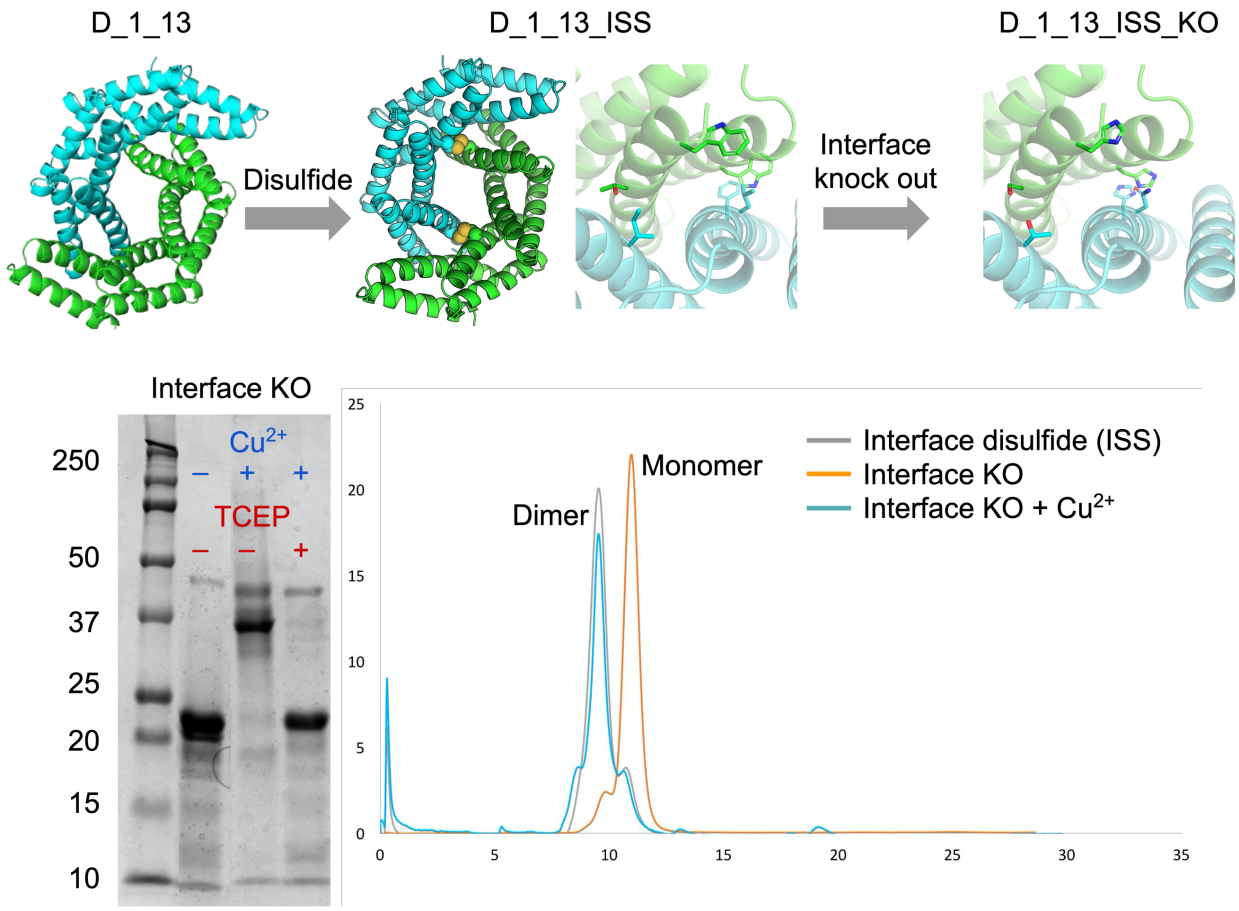
I am going to use this section to explain relevant project ideas and directions that I have excluded from the main chapters of my thesis. All of the work described in the main chapters relates to head to tail dimers produced from curved repeat proteins. A similar but different closed circular protein architecture can be created by docking curved repeat proteins into homodimers containing both head-to-head and tail to tail interfaces (see Appendix Figure 1 below). These proteins are interesting in that the axis of symmetry does not align with the central pore of the protein (like the protein described in the main chapters), but rather cuts through both protein-protein interfaces somewhat perpendicular to the axis of the main pore. I speculate that there are certain C2 symmetric molecules that would be more easily bound with this protein architecture due to the orientation of the symmetry axis. There is ongoing work to explore these types of proteins for binding.

All of the work that I have described has been related to binding C2 symmetric molecules into C2 symmetric proteins. While there are many interesting applications for this work, the majority of interesting molecules that most people would be interested in binding (including for the creation of enzyme active sites) involve binding asymmetric molecules. We are attempting to address this in two different ways. Firstly, the simplest way is to connect the C2 symmetric proteins described throughout my thesis into single-chain proteins. This would allow us to make proteins with closed circular architectures spanning a diverse range of central cavity sizes and shapes that could fit arbitrarily shaped molecules. Secondly, it is possible to dock curved repeat proteins into higher-order oligomers such that cavities are created off of the symmetry axis that could be functionalized for binding asymmetric molecules. There is ongoing work related to both of these ideas.

Additionally, we have had success converting the C2 symmetric constitutive dimers described in this thesis into monomers, by redesigning their protein-protein interfaces to be more polar. This has important implications for the possibility of converting small molecule or peptide binders into chemically induced dimers. Along these lines, we were able to introduce a symmetric disulfide bond into the scaffold design D_1_13 to create a covalent interface in design D_1_13_ISS. We subsequently redesigned select residues at the interface to more polar residues, converting the previous dimer protein into a monomer by SEC and whose symmetric disulfide bond would not form even after mixing with air for several days at millimolar protein concentrations. We later showed that we could force the disulfide bonds to form using Cu2+ as a disulfide catalyst and that this protein ran identically to its parent on SEC (see Appendix Figure 2 below), suggesting that it forms the same dimer protein of the same size and shape. As expected, reducing conditions converted this protein back to a monomer by SEC and SDS-PAGE.



Appendix Figure 1. Head-to-tail homodimers (left) compared to head-to-head + tail-to-tail homodimers (right). The top shows the proteins looking down the axis of symmetry, and below, the proteins are rotated 180° to show the side of the protein. Proteins are shown in backbone cartoon representation and colored in chainbows with the N-terminus colored blue and the C-terminus colored red.



Appendix Figure 2. A redox inducible dimer. The top panel shows the design process starting from the scaffold protein D_1_13 (shown in cartoon and colored by chains) to the interface disulfide protein D_1_13_ISS (disulfides shown in spheres with sulfur colored yellow), to the interface knockout D_1_13_ISS_KO (interface residues before and after shown as sticks and colored by atom type). The bottom panel shows experimental data for D_1_13_ISS and D_1_13_ISS_KO. On the left is an image of a stained SDS-PAGE gel. From left to right on the gel is the ladder, D_1_13_ISS_KO without Cu2+ or TCEP, next is D_1_13_ISS_KO with Cu2+ and without TCEP, and last is D_1_13_ISS_KO with Cu2+ and TCEP. Next are SEC traces overlaid for D_1_13_ISS (gray trace), D_1_13_ISS_KO (orange trace), and D_1_13_ISS_KO after incubation with Cu2+ (cyan trace).

VITA

Derrick R. Hicks grew up in the small town of Auburn California. After graduating from Placer High School in 2008, Derrick attended a local community college for 3 years, during which he obtained his Associate in Science with Honors (3.89 GPA) in each of Biological Sciences, Chemistry, and Natural Science. Derrick subsequently transferred to the University of California at Davis where he was awarded the Regents Scholarship and went on to obtain his Bachelor of Science degree with Highest Honors (3.99 GPA) in Biochemistry and Molecular Biology. During his time at UC Davis, Derrick conducted undergraduate research under the guidance of graduate student mentor Mark Lemos in the lab of Karen McDonald studying the feasibility of engineering the aquatic plant duckweed for protein production or biofuels applications. After his graduation, Derrick joined the lab of Katayoon (Katie) Dehesh at UC Davis as a research technician conducting gas chromatography-mass spectrometry to help the lab better understand plant stress response.

During his time as an undergraduate at UC Davis, Derrick grew increasingly interested in protein biochemistry. During a keystone course with Professor Enoch Baldwin on Macromolecular Structure and Function, Derrick was introduced to work on computational structure prediction and protein design being pioneered in the lab of David Baker at the University of Washington. Derrick later went on to join the University of Washington, where he published alongside his former professor, Dr. Baldwin while rotating in the lab of his future committee member, Dr. Justin Kollman on the filamentous structure of the key metabolic enzyme, CTP synthase. Derrick ultimately joined Dr. Baker's lab to conduct his thesis research related to the design of a de novo fold family of donut-like homodimer proteins with ideal properties for ligand binding functionalization, after a fruitful rotation project on the nature of "Massively parallel de novo protein design for targeted therapeutics."