Structural and mechanistic studies of the Type IIS restriction endonuclease PaqCI and *de novo* designed circular tandem repeat proteins

Madison Arza Kennedy

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Barry L. Stoddard, Chair
Philip Bradley
Raymond Monnat

Program Authorized to Offer Degree:

Biochemistry

University of Washington

Abstract

Structural and mechanistic studies of the Type IIS restriction endonuclease PaqCI and de novo designed circular tandem repeat proteins

Madison Arza Kennedy

Chair of the Supervisory Committee:
Barry L. Stoddard
Department of Biochemistry

This thesis spans the work completed on two parallel projects, to (i) study the structure and mechanism of the Type IIS restriction endonuclease PaqCI, and (ii) participate in the characterization and engineering of *de novo* designed circular tandem repeat proteins ('cTRPs') for the development of novel ligand-dependent protein dimerization systems. Through the latter project, I gained early intensive training in protein crystallography and engineering, contributing to my independent study of PaqCI. Restriction endonucleases are an essential component of innate, 'preprogrammed' phage restriction systems that protect bacteria from foreign DNA. Type II restriction endonucleases are invaluable tools in research because of their ability to identify and cleave specific DNA sequences with extremely high fidelity, as well as their unique mechanisms of cleavage. The most well-studied Type IIS enzyme, FokI, has been shown to require multimerization and engagement with multiple DNA targets for optimal cleavage activity; however, details of how it or related enzymes form a DNA-bound reaction complex have not been described at atomic resolution. Here I describe a series of crystallographic and CryoEM structures in the presence and absence of bound DNA targets that reveal aspects of DNA recognition and cleavage by the Type IIS PaqCI restriction endonuclease. The structures illustrate the enzyme's tetrameric domain organization in the absence of bound substrate and the subsequent formation of a tetrameric reaction complex poised to deliver the first of a series of double-strand breaks. Understanding the structure of the Type IIS restriction endonucleases PaqCI reveals (i) the requirement for multiple DNA targets to be pulled together for enzyme activation, (ii) that enzymatic domains are sterically constricted and can only

correctly orient for cleavage at 4/8 bases from the target site, and (iii) that the orientation of the target recognition domain on the DNA determines the motion required of the endonuclease domains to engage the cleavage site. These results bolster the dominant hypothesis that Type II restriction enzymes require the engagement of multiple unmodified targets to bias cleavage towards unprotected foreign DNA. Through these two projects I gained expertise in crystallography and cryoEM, enzymatic analyses, and protein engineering.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Barry L. Stoddard: The best PI a grad student could ever ask for. Thank you for always being flexible in science, as I jumped from project to project, and in life, as I discovered who I wanted to be.

Christal D. Sohl: The real reason I made it to graduate school. Thank you for believing in me from the beginning and entrusting me with projects bigger than I thought I could handle. You proved to me that I could truly succeed as a scientist.

Thesis Committee: Always providing the best discussions and the best feedback. Each meeting made me a better and more resilient scientist.

New England Biolabs: Yvette and Rick, this research would not have been as impactful without your input and beautiful gels.

Lab Members: Lindsey and Abbie, always providing the best smiles and the most supportive words. And every rotating and summer student who kept me inspired.

Forest: My partner and my match, you have made this journey so much easier and brighter than it would have been alone.

Ginger: TBME, you have seen me from my first day of school until my last, thank you for always encouraging me to continue and reminding me that I am smart.

Brittany & Dua: Friday nights have been and always will be a highlight of my week. Thank you for listening to endless hours of me talking on polo and in person.

Matt: From study group to grad school, we're the best structural biology duo to ever exist.

DEDICATION

Tom Arza Kennedy

To the father who supported me through every endeavor, big or small.

PART 1: PHAGE RESTRICTION AND PAQCI

Chapter 1. MOLECULAR MECHANISMS OF PHAGE RESTRICTION

The following two chapters are taken in part from my first author work published in *Nucleic Acids Research* [1] and has been expanded on for this thesis. I was responsible for the entirety of the work documented in Chapters 2, with exception to the kinetic experiments which were performed by collaborators at New England Biolabs.

1.1 PHAGE DEFENSE SYSTEMS

Bacteria have evolved a wide variety of pathways for antiviral defense, the range of which reflects the diversity of bacteriophage (phage) that a bacterium may encounter in its lifespan. The constant interactions between bacteria and phages have promoted the evolution of diverse bacterial immune systems, many that are absent from model organisms. The most notable of the defense systems are *innate* defense (systems such as restriction-modification, BREX, etc.), *adaptive* defense (typified by CRISPR-Cas systems), signaling systems (CBASS), abortive infection (through membrane damage, inhibition of protein synthesis, protein phosphorylation, and RNA cleavage), toxin-antitoxin systems, and chemical defense [2-4].

Systems that directly target foreign phage DNA for degradation or inhibition of its replication include both innate defense systems that rely on the protection of the host genome while simultaneously targeting incoming foreign DNA for degradation, [5-7] and adaptive defense systems that continuously reprogram in response to phage infections [8]. Those two well-studied 'front line' microbial defenders are augmented by (i) a variety of additional defense systems that also target foreign DNA for degradation or replication interference, (ii) additional systems that trigger abortive infections and cell death, and (iii) many newly-discovered, uncharacterized factors that are encoded within mobile genetic defense islands and are believed to also play a role in combatting phage challenges (see [9] for a recent review).

Innate or ‘preprogramed’ restriction systems include restriction-modification (RM) enzymes, as well as Dnd, Ssp, BREX, and CBASS systems, [10-14] and rely on mechanisms that protect the host genome while targeting incoming foreign DNA. The RM systems, consisting of a restriction endonuclease (‘REase’) and coupled methyltransferase (‘MTase’), are the most well-known, ubiquitous, and abundant innate

defense systems [15-17]. The REase searches for specific short DNA sequences in invading DNA duplexes and cuts within or near those sites, while the MTase chemically modifies and protects the same sequences in host DNA [18, 19]. RM systems will be further addressed in greater detail in latter sections.

DNA Degradation systems (Dnd) comprises about eight genes that work together to ensure restriction only happens when the DNA is protected by an S-modification [14]. The protein complex DndABCDE inserts sulfur into the backbone of double-stranded DNA to protect the host, while the DndFGH is the restriction component [14]. Specific double-strand cleavage occurs at rare sites where the phosphate backbone can have an attached non-bridging S atom [14]. There are five known functional proteins encoded by the *dnd* gene, but details about the S-modification activity, including DNA recognition and cleavage, have remained unclear [14].

Ssp is a phosphorothionation (PT)-sensing bacterial defense system composed of SspABCD-SspE [13]. Instead of operating on double-stranded DNA as Dnd does, it acts only on single-stranded DNA [13]. SspE senses sequence-specific PT sites and inhibits phage propagation by nicking to impair DNA replication [13]. PT is the first modification seen to the sugar-phosphate backbone where the non-bridging oxygen is replaced by sulfur [13, 14]. Unlike classic RM systems, PTs do not occur consistently at a given site in the bacterial genome, reflecting the possibility of a unique DNA-target-selection mechanism is used by Dnd and Ssp [13, 14, 20].The genes in *dnd* and *ssp* are different, establishing the potential of evolutionary diverged PT systems and the possibility of more PT systems within bacteria [13, 14, 20].

Bacteriophage exclusion system (BREX) and *Phage growth limitation* system (Pgl) are related pathways that restrict infection via the inhibition of phage DNA replication (rather than through the cleavage and degradation of phage DNA). They utilize overlapping protein machinery [4, 10, 12]; due to their similarities, Pgl has been incorporated into the BREX nomenclature as one of the six BREX subtypes, specifically ‘type 2’ [10]. The six BREX types currently known are classified by their cassette composition and organization [10]. BREX is most like the Type I RM systems as it requires the formation of a complex and a response (restriction or modification) based on the methylation state of the present DNA [10, 12]. Knowledge of the subtypes continues to grow, but the mechanism behind phage DNA targeting by BREX is still unknown.

Each BREX system comprises four to eight genes which provide the bacteria protection against a multitude of phages [10, 12]. The DNA locus of most BREX types contains a protease, a phosphatase, a DNA methylase, an ATPase, and a protein with unknown function [10]. BREX allows the phage to enter the cell, preventing replication of the phage DNA instead of invasion [10, 12]. Like classic RM systems, BREX methylates at specific locations within the host DNA, but does not degrade it [10]. The BREX methylation site is nonpalindromic such that only one strand is methylated while the other strand remains bare [10]. The methylase, pglX, was found to be essential for phage resistance which suggests that pglX is needed for defense even though BREX does not operating like a traditional RM system [10].

Pgl, or 'BREX type 2', requires a kinase (PglW), a DNA methylase (PglX/BrxX), an ATPase (PglY/BrxC), and a phosphatase (PglZ) [12]. Unlike other BREX types, Pgl undergoes lysis of the first infected bacterial cell to prevent phage from entering the remaining clones [10, 12]. Duplication of the invaded phage is allowed within the primary bacterium, but when the cell lyses, the phage that emerges cannot infect other neighboring Pgl+ hosts [12]. Thus, only one cell undergoes autolysis to prevent continued infection of other clones. This may seem altruistic, but the primary bacterium eliminates competing non-Pgl+ bacteria by releasing more Pgl- toxic phage, leaving the environmental niche open to more Pgl+ bacteria [12]. Notable structures have been solved on multiple proteins in the BREX system that point towards the growing importance of BREX as not only an anti-phage element but a regulatory element [21, 22].

To prevent the spread of the infection to neighboring bacteria, the *cyclic-oligonucleotide-based anti-phage signaling system* (CBASS) utilizes signaling pathways to kill the cell before the phage can replicate. CBASS systems are a combination of two essential proteins; the signaling oligonucleotide cyclase and the cell-killing effector [11]. These two functions do not physically interact but work together in response to each other. The phage sensing oligonucleotide cyclase protein identifies the infection (by unknown means) then produces a second messenger composed of up to three nucleotides connected to form a cyclic molecule, which activates the effector protein to promote cell death [2, 11]. Effector types include effectors with transmembrane helices, phospholipases, endonucleases, TIR domains, phosphorylase/nucleosidases, and peptidases [2, 11]. There are upwards of five thousand CBASS operons, differing primarily in operon composition, effector activity, and type of cyclic oligonucleotide created [11]. Using those differences,

CBASS systems can be classified into four main types which utilize primarily membrane damage and DNA degradation to achieve cell death before the infecting phage can reproduce. The vast diversity of CBASS operons across microbes points towards its success as an anti-phage strategy. In particular, the use of a small second messenger that can activate multiple different effectors at once reduces the time between phage sensing and cell death, becoming a critical tool against phage infection.

Contrary to innate defense systems, adaptive systems rely on immune memory. Adaptive defense systems, typified by *CRISPR-associated nucleases* [8], are continuously reprogrammed in response to past phage encounters. The CRISPR-Cas system is the first example of an adaptive immune system in bacteria. In this case, immune memory is gained by incorporating short viral-derived DNA sequences as spacers within the host genome [8]. These spacers make an RNA guide that allows the CRISPR-Cas machinery to target complementary DNA from invading phages and degrade them [2, 8].

There are three main steps of CRISPR response to phage infection including adaptation, pre-crRNA expression-processing, and interference [8]. To acquire the new spacer during adaptation, the system either uses the Cas protein to bind the target DNA and make a double-stranded break or reverse transcribes the sequence from RNA [8]. After which, that segment is incorporated into the CRISPR array between two repeats, effectively becoming a spacer [8]. The proximal repeat is duplicated when the CRISPR array is prepared [8]. The expression-processing step is performed by another Cas protein/s which transcribes the CRISPR array in one long transcript to generate CRISPR RNA [8]. The final step of interference uses the CRISPR RNA as guide RNA to recognize the spacer (or a very similar sequence) to initiate Cas nuclease-mediated cleavage and inactivation [8]. The adaption and effector module proteins are the most important pieces of these complex systems [8].

CRISPR-Cas systems can be broken down into two main classes [2, 8]. Class 1 (types I, III, and IV) is the most abundant and employs multi-subunit effectors, while class 2 (types II, V, and VI) uses a single large protein effector [8]. These two classes can be further separated into types to further classify the diversity seen in CRISPR-Cas systems, where types differ in their interference machinery, targeting mechanisms, and targeted nucleic acids. In many cases, as with restriction-modification systems, CRISPR-Cas systems (notably types I and II) can respond to and mitigate phage infection without killing the host

bacteria. In addition, CRISPR-Cas systems can operate as the last line of defense for the bacteria (particularly type III systems) to promote high probability of cell death or abortive infection [2, 8].

In abortive infection mechanisms, the cell dies before the phage can complete its replication cycle, the effect of which saves the greater colony of bacteria by preventing the spread of mature phage to the next bacterium. Abortive infection systems are genetically small enough not to inhibit the normal life cycle of a healthy bacteria and yet quick enough to arrest the cell before the invading phage can fully mature. Abortive infection includes either metabolic arrest (which can ultimately lead to the death of the cell) or direct death of the cell [2, 4]. When the metabolism of the cell is slowed, the primary defense systems, in many cases, have enough time to inactivate the infecting phage and save the host. Traditionally, abortive infection is the last line of defenses when systems such as RM have been evaded. Abortive defense systems stop the metabolism or lead to the death of the cell by destroying either the inner membrane, host DNA, or a host's protein synthesis machinery [2, 4]. The most well understood classical abortive infection systems are CBASS and the type III CRISPR-Cas system which also fit into categories of their own following the mechanism behind the recognition of phage infection [2, 4, 8, 11].

In addition to the previously discussed protein systems against invading pathogens, there are a category of defense systems known as toxin/antitoxin (TA) systems. TAs were once thought of as only a stress response of bacteria, but greater research is showing how it operates (even primarily operates) as an antiphage system [3, 23]. These systems exist in almost all prokaryote genomes as two-gene modules and are made up of protein toxins that affect the metabolism of the host and protein (sometimes RNA) antitoxins that counteracts the toxin [3, 4, 23]. In general, the protein toxin is activated by *de novo* RNA synthesis, triggering curtailment of the metabolism through mechanisms such as reduced ATP-production, translation inhibition, or DNA replication interference [3, 23]. TA systems may in fact be evolutionarily some of the first antiphage systems as research points to CRISPR-Cas systems being derived from TA systems [23]. The bacterium is at a high evolutionary pressure to keep the toxin-antitoxin system, if the bacterium loses the genetic element that contains the TA the cell would not be able to express the proteins necessary to clear the already present toxin or protect against the incoming phage and thus will die [3, 24]. This is known as 'post-segregational killing' or 'genetic addiction' and maintains environmental pressure for the bacterium to keep and pass on the TA (or related) systems [3, 24]. There are eight classes of TA systems,

separated by their mechanism. Interestingly, some now classify RM systems as toxin-antitoxin systems (though they lack any homology or shared evolutionary origin) because the restriction endonuclease acts as a pseudo 'toxin' by cutting up the incoming phage DNA and the methylase acts as an 'antitoxin' to protect the host DNA as sometimes physically separate proteins while being mobile elements, affecting global gene expression, and in some cases triggering post-segregational killing [3, 24, 25]. More research in the future will inform on the validity of this new classification. For the purposes of this paper, RM systems will remain separate from TA systems.

Bacteria can utilize more than just proteins to defend against phage infection. Chemical second messenger molecules are known to be used against other microbes that are occupying the same environmental niche as the host, and recently it was discovered that these molecules can also be used to prevent phage replication within the bacterium [26]. Chemical defense systems are widespread, used to thwart incoming phage, and best understood in *Streptomyces* [4, 26]. The bacteria can produce molecules that nonspecifically intercalate into double-stranded DNA (most common group of phages) and thereby block replication of the phage genome [26]. Many of these intercalating agents can work against many invaders, providing broad protection against double-stranded DNA phages. These small molecules are also known to be diffusible, potentially acting to protect the bacteria community from invaders by being secreted by a bacterium [4, 26].

The genes encoding these defense systems tend to reside in clusters within a bacterium's genome, creating what is known as 'defense islands' [3, 8]. The exchange of these defense islands is thought to occur through horizontal gene transfer and is often lost quickly in the bacteria population when evolutionary pressure is no longer present [2-4, 27]. The multiplicity of phage attack (owing to $\geq$ ten-fold more phage than bacteria in an ecosystem) and the inability for a bacterium to maintain all defense systems in its genome at one moment necessitates the retention of redundant defense systems [2, 23]. Moreover, it has been hypothesized that the 'effective' immune system, may not only be what a lone bacterium can encode, but the sum of defense systems contained in the entire bacteria population [4]. By and large, these islands are mobile elements whose functions and roles in bacteria are still in debate and the paradigm under which these systems are studied is shifting.

Some scientists have adopted an alternative view of the movement of defense islands, proposing they are a form of “selfish DNA” [27-29]. It is not well-defined, however, if this hypothesis completely explains the high degree of diversity yet specificity that is seen amongst these systems and associated proteins. This point is emphasized by the evolution of RM systems whose highly specific DNA recognition sequences are unlikely to appear in many phages though they are adopted by bacteria. In many cases discussed above, failure to express a unit of a multi-unit defense mechanism can lead to the premature death of the cell [2]. RM systems could then be considered symbiotic or “selfish”; promoting their own survival by forcing the cell host to keep its gene to stay alive. They can be seen as operating like homing endonucleases, viruses, or transposons by (i) defending their host (and themselves) from foreign DNA, (ii) killing other cells that have removed them, and (iii) moving between genomes [27-29]. More research on defense systems and RM enzymes are needed to understand the host-gene evolution and whether the bulk of which is driven by the need to defend the cell or to propagate genetic material.

1.2 RESTRICTION-MODIFICATION SYSTEMS

Restriction-modification (RM) systems are comprised of restriction endonuclease (‘REase’) and methyltransferase (‘MTase’) enzymes and activities that are coupled to one another genetically (and often structurally) [15-17]. Such systems operate by searching for short target sequences found in foreign invading DNA duplexes and cleaving within or near those sites while chemically modifying and protecting the same sequences in host DNA (**Figure 1.1**) [18, 19].

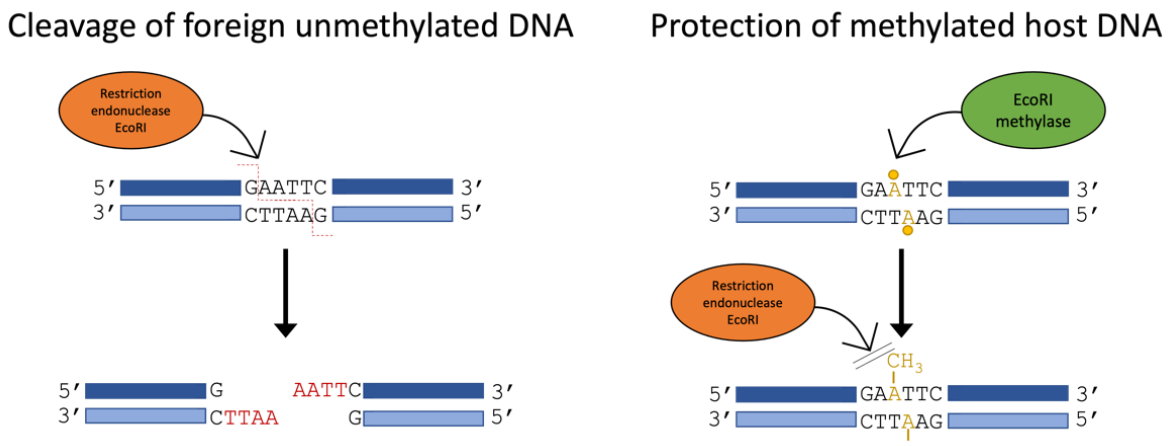

Figure 1.1. Traditional cleavage behavior by restriction endonucleases.

Restriction-modification systems distinguish between host and foreign. The restriction endonuclease (orange) searches for short target sequences found in foreign invading DNA duplexes (blue) and cleaving within or near those sites while a methylase (green) chemically modifies and protects the same sequences in host DNA. The REase can cleave non-protected DNA while the methylation blocks the binding of the REase on the host.

Classic RM systems impose enzymatic modifications, such as methylation, to individual nucleobases within target sites in the host genome, to prevent REase activity and thereby avoid degradation of self-DNA. Although methylation is the most synonymous with RM systems, many other modifications can be made that prevent the nuclease from cleaving host DNA. More than the nucleobases can be modified on DNA; in some prokaryotes sulfur molecules can be added to the phosphodiester backbone of DNA replacing a non-bridging oxygen [17].

RM systems are historically classified into four main types, defined by their subunit architectures, target search mechanisms, cofactors, and catalytic behaviors [16, 30]. However, as new enzymes are discovered and characterized, the boundaries between the Types blur, and only a few generalizations remain as hallmarks for each type [31, 32]. This work will focus on the current agreed upon type organization as of the year of publication. In Type I, II, and III RM systems, unmodified foreign DNA targets are recognized and cleaved, while corresponding targets in the host DNA are modified to prevent cleavage of the host's genome. Conversely, Type IV systems recognize and cleave foreign DNA targets that are proactively modified (using similar modifications as mentioned above as well as many not seen in host protection) by phage as a countermeasure to avoid cleavage [17].

Type I, II, and III RM systems differ primarily in their DNA cleavage mechanisms. Type I and III enzymes bind their specific target through a DNA recognition domain partnered with the methyltransferase, and then require ATP-dependent translocase subunits to bring together two bound targets to assemble active enzyme complexes capable of DNA cleavage [5, 7]. In contrast, Type II enzymes are simpler, and cleavage generally requires only magnesium as a cofactor [6]. Type II systems most often consist of separate REase and MTase enzymes, each with their own DNA recognition domain targeted to the same sequence motif, although some forms (like Type IIL and Type IIG RM enzymes) couple both activities (found either on a single protein subunit or on two closely associated subunits) to a single target recognition domain (TRD) to simultaneously bring both functions to a potential target site. The REase and MTase search for their target via localization near DNA, followed by rapid association and dissociation coupled with limited two-dimensional (2D) diffusion along the double helix [33]. For many of the most familiar Type II enzymes, the REase is a homodimer that acts separately at each individual site with each monomer cutting one DNA strand.

Type II enzymes, which are most commonly employed for molecular cloning and biotechnology applications, are further distinguished and categorized by differences between (i) their cleavage patterns (cutting either within their targets or at a fixed distance to one (or both) sides of their bound targets), (ii) their catalytic behavior (sometimes displaying non-cooperative cleavage of a single bound target, but often relying upon allosteric or cooperative cleavage mechanisms that require the simultaneous engagement of multiple target sites for maximum cleavage efficiency), and (iii) their domain organization (displaying various arrangements of protein domains involved in target recognition, cleavage and/or methylation) [6, 33, 34]. These distinctions lead to the division of Type II REases into multiple subtypes. This includes Type IIP enzymes (usually homodimers that cleave palindromic target sites), Type IIE and IIF (frequently multimeric complexes, often tetramers, that require binding of at least two target sites to cleave one site and that exhibit cooperative and/or allosteric behaviors during cleavage of those two sites), and Type IIT (heterodimers that cleave asymmetric sites) [33, 35-37]. In addition, more complex Type IIG and IIL enzymes contain both REase and MTase domains that are physically coupled and that act in a coordinated manner to either methylate or cleave host or foreign DNA, respectively [38].

Regardless of their exact type and subtype, restriction endonucleases (i) display exceptional fidelity, (ii) contain and use any one of several different active site mechanisms, and (iii) may or not require assembly of multiple bound enzyme-DNA complexes to activate cleavage. The mechanisms by which the various enzymatic domains of RM Type II systems oligomerize to instigate a productive cleavage event can differ drastically within a type or subtype [6]. The recognition of the DNA target site by the enzyme is usually facilitated by a conformational change of the protein, sometimes an additional allosteric effect, and displacement of most of the water molecules and counter ions from the protein-DNA interface [33]. In general, Type II enzymes hydrolyze the phosphodiester bond via an SN2-type mechanism utilizing active site residues, water molecules and heavy metal ions [33]. Type II restriction endonucleases utilize mainly three different catalytic motifs and mechanisms for hydrolysis of the phosphate backbone. These include PD-(D/E)xK, HNH, and GIY-YIG, as found in NotI, Eco29kI, and PacI respectively [32, 39-41]. Under optimum conditions, classic endonucleases such as EcoRI and EcoRV can distinguish between sites that vary by only one base pair conferring a 10^6 decrease in k_{cat} [42-44].

1.3 TYPE IIS RESTRICTION ENDONUCLEASES

Type IIS enzymes combine independently folded target recognition (TRD) and endonuclease (EN) domains within a single protein chain, cleaving DNA at one or both sides of their asymmetric target site [45]. They are of particular interest for biotechnology, being the source of nonspecific EN domains for various gene-targeting nuclease platforms (including zinc finger nucleases and TAL effector nucleases) [46] and also being used for a wide variety of genome mapping, sequencing, and gene assembly applications [47, 48]. Individual Type IIS enzymes display considerable variation in (i) their architecture (with the EN domain found at either the N- or C-terminal end of a protein subunit), (ii) the identity of the nuclease active site motif (usually containing HNH or PD-(D/E)xK active sites) and their cleavage mechanism, and (iii) their cleavage positions (cleaving at precise distances from the bound target, with those distances varying greatly between different enzymes) [32]. What brings this subtype together is the ability to recognize asymmetric targets and cleave double-stranded DNA on one side of the site, as seen in studies of FokI, BspMI, and MboII [30-33, 49-51].

1.4 FokI, the most well studied Type IIS protein

The most thoroughly studied Type IIS restriction endonuclease, FokI, recognizes a five-base, non-palindromic sequence (5'-GGATG-3') and cuts the top and bottom DNA strands 9 and 13 bases downstream from its target site, generating a complementary four-base 5’ overhang [52]. It is comprised of an N-terminal TRD and C-terminal EN domain that harbors a single PD-(D/E)xK catalytic motif [52]. Crystal structures of inactive FokI in the absence or presence of bound DNA illustrate similar monomeric protein conformations, with the nuclease domain loosely docked against the TRD but not positioned appropriately for DNA cleavage (indicating that additional motion of that domain is required to form the eventual reaction complex) [51, 52]. Those results, along with detailed kinetic studies [49, 53-56], implied that a multimeric assemblage of two DNA-bound enzyme subunits, along with the additional movement of the EN domains, is required for DNA cleavage. Additional studies using single-molecule force measurements [57, 58] and electron microscopy (EM) imaging [59] further indicate a mechanism in which the formation of a dimeric enzyme-DNA cleavage synapse produces a parallel arrangement of bound DNA targets and significant DNA looping. There is still much that is unclear about the details of enzymatic action by FokI and other Type IIS enzymes, even though their nuclease domains are used in the creation of zinc-finger nucleases (ZFN) and TALENs [60, 61]. Perhaps most importantly, a high-resolution view of the ultimate reaction synapse formed by FokI (or any other Type IIS enzyme) had not yet been described when my structural analysis of the Type IIS enzyme PaqCI was initiated.

Chapter 2. STRUCTURE AND MECHANISM OF THE TYPE IIS RESTRICTION ENDONUCLEASE PAQCI

2.1 INTRODUCTION

To bias cleavage towards foreign invading DNA, many restriction endonucleases (REases), including some from the Type II class (Type IIE), are thought to require multiple unmodified DNA targets to be brought together into a cooperative reaction 'synapse' before nuclease activity is licensed [62-64]. The requirement of a reaction synapse with multiple bound targets and a mechanism requiring one or more *trans*-acting endonuclease (EN) domains within that synapse is a particularly common feature of many bipartite REases with separate EN domains and target recognition domains (TRD) [62, 63, 65, 66]. However, such enzymes (spanning both the Type IIS and Type IIG REase subfamilies) span a vast array of architectures and mechanisms. Previous structural and functional studies on such enzymes, such as FokI and DrdV, revealed features that are unique to each enzyme while still cleaving at their targets with high specificity and fidelity. Despite this wealth of research, a high-resolution view of the cleavage synapse formed by a canonical Type IIS enzyme has yet to be described. To address this point, the activity and structures of the Type IIS enzyme PaqCI in the presence and absence of bound DNA were determined.

Obtaining structures of complex DNA-protein assemblages in a defined reaction state, such as a Type IIS restriction endonuclease in complex with its target site in a productive cleavage complex, using X-ray crystallography has proven to be challenging due to the dynamic equilibrium displayed by different assembly states of the oligomer (a roadblock that is true for many nucleic acid enzyme systems). CryoEM has been used with success to deconvolute different structural assemblage and reaction states for such complex enzymatic systems and is now a clearly transformative approach for obtaining structures of large protein and nucleic acid assemblages. For example, a cryoEm structure of a Type I enzyme complex, EcoR124I, was obtained bound to DNA [67], and the Stoddard lab generated multiple CryoEM structures of various mechanistic stages of the Type II restriction enzyme DrdV, which is a large protein assemblage that combines methylation and restriction [38]. The benefits to using CryoEM to solve structures of complex protein and DNA/RNA systems will be discussed in Chapter 3. Using a combination of X-ray crystallography and CryoEM can elucidate a complete understanding of a Type IIS protein.

Here, biochemical analyses and high-resolution crystallographic and CryoEM structures (in the presence and absence of bound DNA) of the Type IIS REase PaqCI are reported. While this enzyme is also a Type IIS enzyme with a PD-(D/E)xK active site motif; its domain architecture is reversed as compared to FokI (corresponding to an N-terminal EN domain). It forms a tetramer in solution in both the absence and presence of bound DNA, recognizes a longer target than FokI (5' - CACCTGC - 3'), and cleaves each strand of bound DNA much closer to the bound target site (4 and 8 bases downstream, respectively). Our analysis, therefore, present a detailed examination of an alternative, but related Type IIS enzyme while also providing a high-resolution visualization of the conformational changes that accompany DNA binding and of the enzyme trapped in a catalytically productive conformation, poised to deliver the first in a series of DNA cleavage events. Our structure explains why Type IIS enzymes such as PaqCI require binding at multiple sites to activate cleavage. The study of PaqCI provides an important basis for comparison, analysis, and additional understanding of the diversification and action of Type IIS enzymes.

Additionally, protein engineering methods (discussed in chapter 4) have proved beneficial in engineering DNA binding and editing proteins. Proteins that can edit DNA (such as zinc-finger nucleases, transcription activator-like effector nucleases, and clustered regularly interspaced short palindromic repeat-CRISPR associated proteins) are being optimized for gene editing, allowing for better manipulation of the target DNA [60, 68]. These DNA editing proteins can be broken into their respective functional parts and exchanged with each other, creating chimeras that may be more suited for the job of interest [69, 70]. This all further emphasizes the necessity for accurate structural and mechanistic knowledge of DNA editing proteins and systems as these protein tools, such as PaqCI, may soon be another standard practice in protein engineering.

2.2 DNA-FREE STRUCTURES REVEALS TETRAMERIC ORGANIZATION OF PAQCI

2.2.1 *Methods*

Enzyme identification, cloning, expression, and purification. DNA encoding the *Paucibacter aquatile* Type IIS restriction endonuclease (REase), PaqCI, was codon-optimized for *E. coli* expression and synthesized commercially by Integrated DNA Technologies (IDT, Coralville, IA) and cloned into a modified pACYC184 T7 expression vector. DNA encoding the full-length BspMI methyltransferase, M.BspMI, was cloned into a

modified pBR322 expression vector. Tagless PaqCI/pACYC184-T7 and M.BspMI/pBR322 were co-transformed into T7 Express Competent E.coli (NEB), grown at 37 °C in Terrific Broth to an A_{600} of 1.0, and then induced with 0.3 mM isopropyl 1-thio-β-D-galactopyranoside (IPTG) overnight at 19 °C. Cells were harvested by centrifugation, washed with DEAE load buffer (20 mM Tris-HCl, pH 8.5, 300 mM sodium chloride, 5 mM β-mercaptoethanol), and pelleted a second time. Pellets were flash-frozen in liquid nitrogen and stored at −80 °C.

Thawed pellets from 2 L cultures were resuspended in 100 mL of DEAE load buffer supplemented with 10 mM phenylmethylsulfonyl fluoride (PMSF), 5 mg of recombinant DNaseI (NEB), and 5 mM magnesium chloride. Lysozyme was added to 1 mg/ml, and the mixture was incubated for 15 min rocking at 4 °C. The cells were disrupted by sonication, and the lysate was cleared of debris by centrifugation at 13,000 rpm (19,685 × g) for 30 min at 4 °C. The supernatant was sterilized through a 0.45 μm syringe filter, loaded onto 2x 5 mL HiTrap DEAE columns, and flowthrough fractionated. Fractions were assayed for PaqCI endonuclease activity. Peak fractions were analyzed by SDS-PAGE, pooled, and dialyzed against Heparin load buffer (20 mM HEPES, pH 7.5, 50 mM sodium chloride, 5% glycerol, 1 mM EDTA, and 1 mM DTT). The sample was loaded onto 2x 5 mL HiTrap Heparin columns, washed with load buffer, eluted in a gradient from 50 mM to 500 mM sodium chloride, and fractionated. Peak fractions were analyzed by SDS-PAGE, pooled, concentrated, and further purified by size exclusion chromatography (SEC) using a Superdex 200 10/300 GL column (**Figure 2.1**). The sample was exchanged into a final buffer of 20 mM HEPES, pH 7.5, 150 mM sodium chloride, 5 mM magnesium chloride, and 1 mM DTT during SEC and concentrated to 10-20 mg/ml.

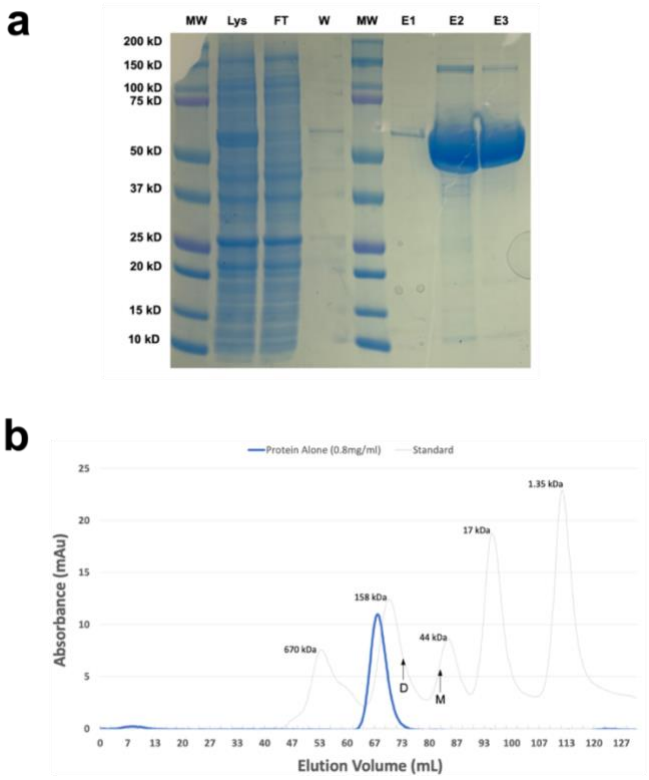

Figure 2.1. Purification and solution behavior of PaqCI.

Panel a: SDS-PAGE gel of induced cell lysate and purification via metal affinity chromatography. **Panel b:** Size exclusion chromatographic elution of final purified PaqCI used for subsequent structural analyses. The protein elution profile is dark blue; underlying elution of molecular weight standards is light dashed grey. The expected peak elution volumes for an enzyme monomer (56 kD) or dimer (112 kD) are indicated with labeled arrows. The enzyme elutes as a multimeric complex that is consistent with an enzyme tetramer.

X-ray crystallographic analyses. Crystals of the DNA-free PaqCI apo-enzyme (**Figure 2.2**) were grown using protein purified as described above. Protein samples dispensed in 1 μL drops (at concentrations ranging from 3 to 12 mg/mL in 20 mM HEPES pH 7.5, 150 mM potassium chloride, 5% glycerol v/v) were mixed with an equal volume of a crystallization solution containing 20 to 25% w/v polyethylene glycol (PEG) 3350, 0.2 M magnesium chloride hexahydrate and 0.1 M BIS-TRIS pH 5.5. Vapor phase equilibration of the resulting drops against a 1 mL reservoir of the same crystallization solution resulted in the growth of crystals with dimensions of 0.05 to 0.2 mm in each dimension within approximately 72 hours. The crystals were flash cooled in liquid nitrogen after transfer into a cryoprotective solution corresponding to elevated PEG 3350 (30% w/v) and 20% ethylene glycol. Diffraction data were collected on a Pilatus areas detector at the Advanced Light Source (ALS) synchrotron facility at beamline 5.0.1

(**Figure 2.2**). The resulting data set (**Table 2.1**) extended to 2.5 Å resolution and corresponded to a primitive tetragonal space group ($P4_32_11$) in which two copies of the protein subunit occupied the asymmetric unit.

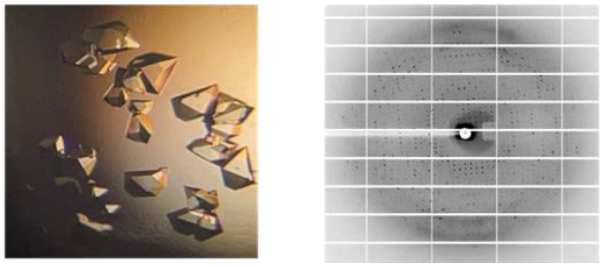

Figure 2.2. Crystallographic analysis of DNA-free PaqCI apo-enzyme.

Crystals and diffraction pattern for DNA-free PaqCI. Diffraction pattern corresponds to a 0.5 second, half-degree oscillation collected at the Advance Light Source beamline 5.0.1 on a Pilatus area detector.

Table 2.1. PaqCI crystallographic data and refinement statistics

Data Statistics	
Space group	$P4_32_12$
Wavelength(s)	0.9762 (Å)
Unit cell dimensions	a = b = 136.6 Å c = 106.4 Å
Resolution	2.5 Å (2.54-2.5)
Reflections	35387
Completeness	100.0% (100.0)
Redundancy	26.1 (25.9)
I/s(I)	43.9 (5.6)
Rmerge	0.0691 (0.422)
Rpim	0.017 (0.095)
CC1/2	1.000 (0.995)
Refinement Statistics	
PDB ID	8EM1
Resolution	2.5 Å
No. reflections	35124 (3427)
Rwork / Rfree	0.2089 / 0.2641
No. atoms	
Protein	973
Ligand/Ion	8
Water	69
B-factors	
Protein	55.46
Ligand/Ion	57.67
Waters	47.22
RMS deviations	
Bond length (Å)	0.009
Bond angles (°)	1.04

Data was processed using program HKL2000 [71]. The placement of two copies each of the N-terminal endonuclease (EN) domain and the C-terminal target recognition domain (TRD) into the asymmetric unit was then performed using the molecular replacement algorithm in program PHENIX [72]. The molecular replacement searches were conducted in two sequential steps, using independent molecular models for the EN and TRD domains that were each generated using the program AlphaFold [73]. An initial search with the model of the N-terminal domain produced a solution with excellent signal (LLG = 177.6; TFZ = 12.1) that placed two EN domains in a dimeric arrangement that appeared appropriate for eventual engagement with a DNA duplex. After fixing those domains, a second search then placed two copies of the model of the C-terminal TRD into the remaining volume of the asymmetric unit built, resulting both in strong signal (LLG 317.5; TFZ = 9.2) and distances between their termini that could easily be bridged by the peptide linker sequence connecting the two domains.

Local rebuilding of the two enzyme subunits in the asymmetric unit, including the modeling of a peptide linker in each that connects an N-terminal EN domain to its corresponding C-terminal domain) was performed using the program COOT [74], followed by refinement using the program PHENIX [72]. The final values for R_{work} / R_{free} were 0.21 / 0.26 with good geometry (**Table 2.1**).

The complete structure of the DNA-free apo-enzyme was ultimately found to correspond to a compact tetrameric assemblage, in which two additional protein subunits are generated via application of a crystallographic dyad symmetry axis. The designation of the enzyme assemblage as a tetramer agrees with (i) the solution behavior of the protein when analyzed and purified via size exclusion chromatography, and (ii) the same structure determined independently via single-particle CryoEM analysis, as described below.

CryoEM sample preparation and data collection. Samples for the PaqCI DNA-free complex was generated using purified protein as described above. The sample was filtered through a 0.22 μm centrifugal filter and loaded onto a HiLoad 16/600 Superdex 200 prep grade size exclusion column (Millipore Sigma) equilibrated in 30 mM BisTris pH 6.5, and 100 mM sodium chloride to assess oligomerization state. The oligomer peak was taken from the SEC and diluted. The complex was initially evaluated via negative stain electron microscopy (EM) for optimization of CryoEM sample preparation (**Figure 2.3**). Negative stain grids were prepared by adding the sample to glow-discharged uniform carbon film coated grids and stained with

0.75% uranyl formate as in Ohi, 2004 [75]. Data was collected semi-automatically with Leginon [76] using the Fred Hutchinson Cancer Center's (Fred Hutch) 120 kV ThermoFisher Talos 120C LaB6 microscope equipped with a Ceta camera.

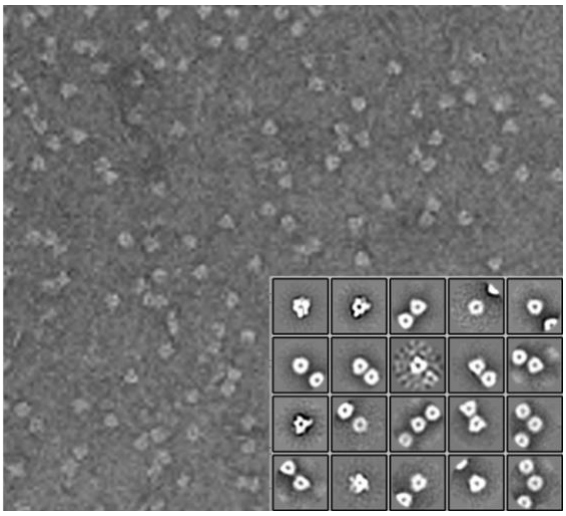

Figure 2.3. Negative Stain EM analysis of DNA-free PaqCI apo-enzyme.

Representative negative stain micrograph and class averaged 2D particle density for DNA-free PaqCI.

Vitrification conditions for CryoEM grids were screened with the optimized enzyme sample from negative stain EM (**Figure 2.4**). CryoEM grids were prepared by applying 2 to 4 µL of complex to a glow-discharged Quantifoil R1.2/1.3 or UltrAuFoil R1.2/1.3 300 mesh grid, which was blotted for 2 to 8 seconds, then plunge frozen in liquid ethane using an FEI Vitrobot Mark IV (ThermoFisher) at 4°C and 100% humidity. All data sets were collected using the Fred Hutch's 200 kV ThermoFisher Glacios X-FEG electron microscope equipped with a Gatan K3 detector. SerialEM was used for data acquisition [77]. Six second movies were collected at 0.06 seconds per frame with a per frame dose of 0.5 e-/Å2/s. The pixel size on the specimen was 0.561 Å/pixel. Collection and on-the-fly monitoring were conducted with Warp for ctf estimation, motion correction, and data quality monitoring [78].

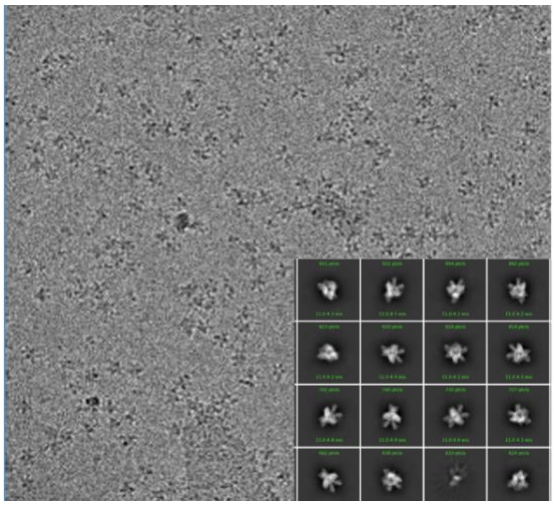

Figure 2.4. CryoEM processing of DNA-free PaqCI apo-enzyme.
Representative CryoEM micrograph and class averaged 2D particle density for DNA-free PaqCI.

To generate a map of the unbound structure, three 0° tilt and one 40° tilt data sets were collected. The first two 0° tilted CryoEM data sets for the unbound structure were collected using 0.1 mg/ml and 0.4 mg/ml complex, respectively, on Quantifoil R1.2/1.3 grids with 4 second blot times. 869 movies were collected for the first data collection, and 2,075 movies were collected for the second data collection. The final datasets for the unbound structure were collected using an UltrAuFoil 1.2/1.3 grid with 0.26 mg/ml protein blotted for 6 seconds.

Negative stain data processing. Micrographs were imported into RELION 3.1 [79] and ctf corrected using CTFFIND4 [80]. Particles were manually picked to establish 2D templates and then template picked. These particles were extracted in a 300-pixel box and binned by 4 before extraction and 2D classification to assess particle homogeneity.

CryoEM data processing. Motion-corrected data that had been binned by 2 in WARP was downloaded and imported into cryoSPARC [81]. Once in cryoSPARC, 'patch ctf' was run [81]. Each independent dataset was preliminarily processed before combining after initial particle curation. The preliminary processing included ctf estimation, exposure curation, blob picking (to generate two-dimensional (2D) templates), template picking, 2D classification, ab-initio reconstruction, and heterogeneous refinement. Templates were selected from the first round of 2D classification and used for future particle picking. Data was refined into three classes using heterogeneous refinement. One good class was selected and moved on to the next round of refinement. The process was iterated three times. Box

size remained at 64 pixels for the initial 2D classifications, then expanded to 128 pixels for the heterogeneous refinement. All data sets were combined after the final heterogeneous refinement and further processed with an additional round of 2D classification, and iterative non-uniform 3D refinement. Data was unbinned once all data sets were combined and one round of 2D classification was performed and particles were extracted with a 300-pixel box size. Before generation of the final map, global ctf, and local ctf jobs were run. The final selected particles were run through non-uniform refinement to generate the final 3D CryoEM map. Local resolution was determined using the final non-uniform refinement job. No symmetry was ever imposed during data processing.

A total of 2,118,800 putative particles were initially picked from the four data collections after template picking. In the final structure, approximately 982,000 particles were used from the 0° tilt datasets, and approximately 420,000 particles were used from the 40° tilt dataset. After processing using the pipeline described above, a total of 1,427,503 particles were used to generate a 3.0 Å (global resolution) map, which was subsequently used to confirm the tetrameric structure seen in crystallography.

CryoEM model building and refinement. The refined x-ray crystallographic model of unbound PaqCI was docked into CryoEM map, confirming the architecture of the DNA-free enzyme tetramer determined as described above. All figures were generated with PyMOL [82].

2.2.2 *Results*

Structure of the DNA-free enzyme. In the absence of bound DNA, PaqCI readily forms well-ordered diffracting crystals (**Figure 2.2**), implying that the DNA-free apo-enzyme forms a conformationally homogenous multimeric assemblage allowing it to be captured in a uniform crystal lattice. The crystal structure of the DNA-free enzyme was determined to 2.5 Å resolution (**Table 2.1**) and found to correspond to a compact enzyme tetramer (**Figure 2.5**). In that structure, the asymmetric unit corresponds to an enzyme dimer, with a larger tetrameric assemblage generated via the application of a two-fold crystallographic symmetry axis, forming a dimer-of-dimers. The overall dimensions of the tetramer are 124 Å x 118 Å x 96 Å.

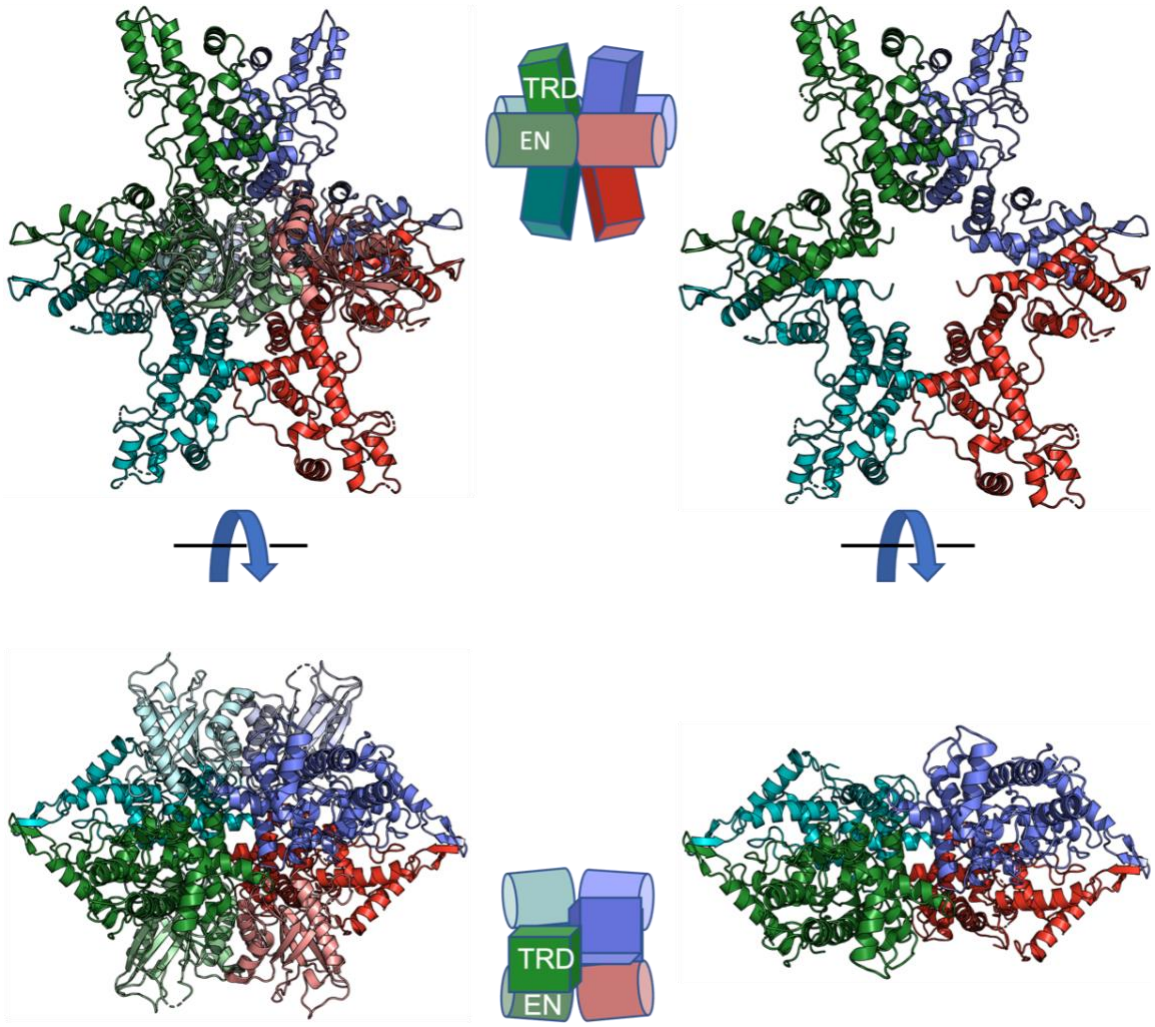

Figure 2.5. Crystallographic structural analysis of DNA-free PaqCI apo-enzyme.

Ribbon diagrams and cartoon representation of DNA-free apo-enzyme structure and domain organization. Each of the four protein subunits are colored independently and similarly in the two depictions. The endonuclease (EN) domains are colored in a lighter shade than the target recognition domains (TRDs) they are associated with. In the cartoons, the TRDs are represented by rectangles and the EN domains are shown as cylinders. The ribbon diagrams on the left illustrate the full-length protein subunits. The ribbon diagrams on the right illustrate only the C-terminal TRDs; the N-terminal EN domains are removed for clarity. The ribbon diagrams in lower panel are turned 90° around the x-axis when compared with upper panel.

In the crystallographic model, four target recognition domains (TRD) and four endonuclease (EN) domains form a starburst-like structure. The TRDs form a ring of four C-shaped subunits while two EN domains dimerize on either side of the TRD ring, sequestering their active sites against the oligomer and away from solution and DNA. Each dimer pair within the enzyme tetramer contacts the other both through contacts between the sequestered EN domains and through additional contacts between the TRDs. The conformation of each individual protein subunit brings their N- and C-termini close to one another. Across

the surface of the resulting tetrameric assemblage, basic residues are somewhat sparsely distributed, corresponding to a somewhat lower estimated pI (approximately 8.0) than is typically observed for many DNA binding proteins (**Figure 2.6**).

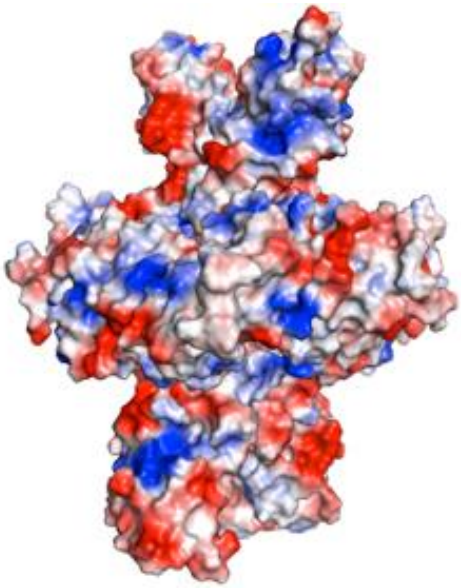

Figure 2.6. The electrostatic surface of the PaqCI tetramer.

Ribbon diagram and corresponding electrostatic surface of the DNA-free PaqCI tetramer. The enzyme complex does not present extensive regions of strongly basic, positively charged surface area, corresponding to a lower calculated pI (approximately 8.0) that necessitated formation of a stable DNA complex at neutral pH.

Unbound PaqCI forms visible particles in negative stain EM and CryoEM images (**Figure 2.3** and **Figure 2.4**), allowing the additional determination of the structure of the unbound enzyme assemblage using single-particle electron microscopy and thereby validating (in solution and at a much lower protein concentration) the size, shape, and structural features of the DNA-free apo-enzyme described above. The sample displayed a limited set of preferred orientations in both negative stain EM and CryoEM requiring the need to tilt the stage to optimize visualization of the three-dimensional structure of tetramer. The final CryoEM map for the unbound enzyme, corresponding to approximately 2.9 Å to 3.3 Å resolution, was generated from 4,284 micrographs and over one million particles. The refined crystallographic coordinates of PaqCI described above are easily docked into the CryoEM map with minimal rebuilding (**Figure 2.7**).

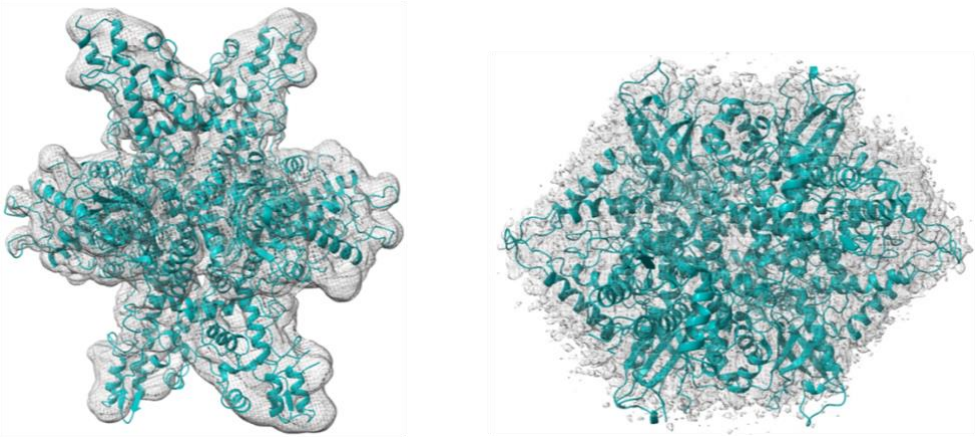

Figure 2.7. CryoEM analysis of DNA-free PaqCI.

Independently generated CryoEM electron density map, at approximately 3.0 Å global resolution, of the DNA-free apo-enzyme closely match the crystallographic model of the enzyme tetramer, validating the quaternary structure and domain organization in a solution-based analysis free of crystallographic contacts and lattice artifacts. The views seen in **Figure 2.7** are the same orientations as **Figure 2.5**.

3D variability analysis of the CryoEM DNA-free apo-enzyme tetramer map shows an oscillation between either side of the dimer pair, with correlated motions of each TRD affecting the corresponding orientation of its neighbor. This correlated motion centered across the interface appears to illustrate that the unbound tetramer samples have slightly asymmetric conformations between the two dimer pairs. In each subpopulation, the EN domains remain associated with the tetrameric core. The visualized motion agrees with the observation that the lowest resolution of the CryoEM map is at the ends of the TRDs (**Figure 2.8**).

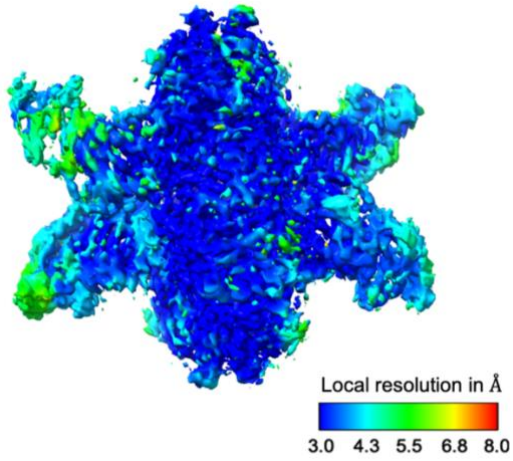

Figure 2.8. Local resolution of CryoEM DNA-free PaqCI map.

The resolution of the DNA-free PaqCI 3D CryoEM reconstruction as calculated in cryoSPARC. The image of the map was generated in UCSF Chimera.

2.3 CryoEM structure reveals four DNA duplexes bound to the PaqCI tetramer

2.3.1 *Methods*

CryoEM sample preparation and data collection. Samples for the PaqCI DNA-bound CryoEM complexes were generated using purified protein as described above in DNA-unbound methods. The sample (with DNA) was filtered through a 0.22 µm centrifugal filter and loaded onto a HiLoad 16/600 Superdex 200 prep grade size exclusion column (Millipore Sigma) equilibrated in 30 mM BisTris pH 6.5, and 100 mM sodium chloride to assess oligomerization state. The DNA-bound PaqCI complex was generated by incubating enzyme and DNA in excess at a 1:1.2 molar ratio (each monomer possessing one DNA binding site). This was done in the presence of 10 mM calcium chloride to prevent cleavage of the DNA substrate. The protein co-eluted with the DNA generating a peak centered at an estimated molecular weight of approximately 225 kD with no remnants of an unbound peak (**Figure 2.9**). The oligomer peak was taken from the SEC and diluted. The complex was initially evaluated via negative stain electron microscopy (EM) for optimization of CryoEM sample preparation. Negative stain grids were prepared by adding the sample to glow-discharged uniform carbon film coated grids and stained with 0.75% uranyl formate as in Ohi, 2004 [75]. Data was collected semi-automatically with Leginon [76] using the Fred Hutch's 120 kV ThermoFisher Talos 120C LaB6 microscope equipped with a Ceta camera.

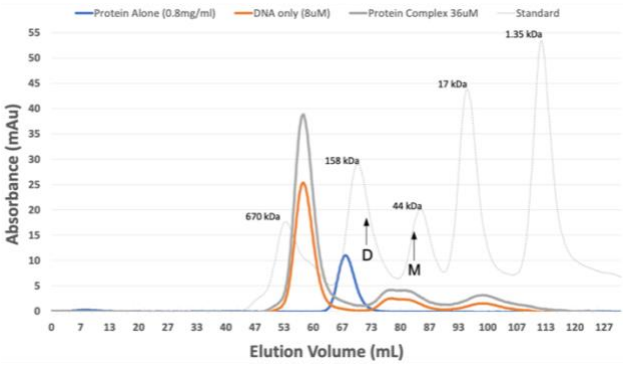

Figure 2.9. Purification of DNA-bound PaqCI.

Size exclusion chromatographic elution behavior of PaqCI enzyme in the presence of a stoichiometric excess of double-strand DNA indicates formation of a stable DNA-bound enzyme complex. The elution profiles for free protein, free DNA and a 1:1.2 stoichiometric mixture of protein and DNA are shown as bold colored lines; the elution of molecular weight standards is shown as a light dashed grey profile The expected peak elution volumes for an enzyme monomer (56 kD) or dimer (112 kD) are indicated with labeled arrows. In this experiment, the DNA (and the protein-DNA complex) elute at a significant early volume due to the length of the DNA duplex used in the experiments.

Vitrification conditions for CryoEM grids were screened with the optimized enzyme sample from negative stain EM. CryoEM grids were prepared by applying 2 to 4 µL of complex to a glow-discharged Quantifoil R1.2/1.3 or UltrAuFoil R1.2/1.3 300 mesh grid, which was blotted for 2 to 8 seconds, then plunge frozen in liquid ethane using an FEI Vitrobot Mark IV (ThermoFisher) at 4°C and 100% humidity. All data sets were collected using the Fred Hutch's 200 kV ThermoFisher Glacios X-FEG electron microscope equipped with a Gatan K3 detector. SerialEM was used for data acquisition [77]. Six second movies were collected at 0.06 seconds per frame with a per frame dose of 0.5 e-/Å2/s. The pixel size on the specimen was 0.561 Å/pixel. Collection and on-the-fly monitoring were conducted with Warp for ctf estimation, motion correction, and data quality monitoring [78].

Two data sets were collected for the DNA-bound PaqCI structure, one at 0° and one at 40° tilt that were collected on an UltrAuFoil R1.2/1.3 grid with 0.9 mg/ml complex blotted for 6 seconds. 941 movies were collected at 0° tilt, and 956 movies were collected at the 40° tilt.

Negative stain data processing. Micrographs were imported into RELION 3.1 [79] and ctf corrected using CTFFIND4 [80]. Particles were manually picked to establish 2D templates and then template picked. These particles were extracted in a 300-pixel box and binned by 4 before extraction and 2D classification to assess particle homogeneity.

CryoEM data processing. Motion-corrected data that had been binned by 2 in WARP was downloaded and imported into cryoSPARC [81]. Once in cryoSPARC, 'patch ctf' was run [81]. Each independent dataset was preliminarily processed before combining after initial particle curation. The preliminary processing the structure included ctf estimation, exposure curation, blob picking (to generate two-dimensional (2D) templates), template picking, 2D classification, ab-initio reconstruction, and heterogeneous refinement. Templates were selected from the first round of 2D classification and used for future particle picking (**Figure 2.10**). Data was refined into three classes using heterogeneous refinement. One good class was selected and moved on to the next round of refinement. The process was iterated three times. Box size remained at 64 pixels for the initial 2D classifications, then expanded to 128 pixels for the heterogeneous refinement.

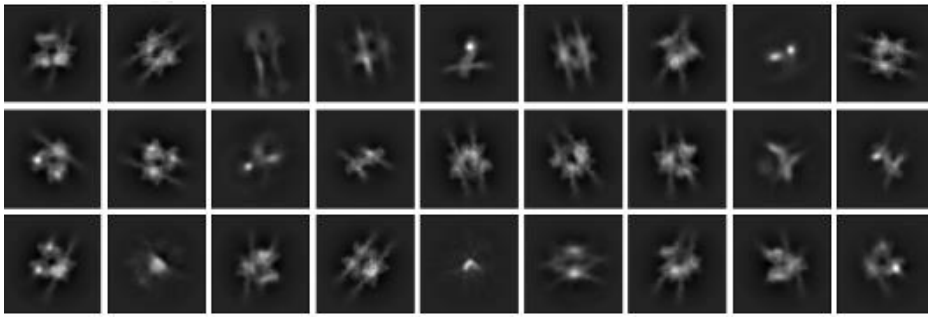

Figure 2.10. CryoEM data processing of DNA-bound PaqCI.

Class averaged 2D images of DNA-bound PaqCI complexes. The bound DNA molecules extending from the protein assemblage are readily visible.

All data sets were combined after the final heterogeneous refinement and further processed with an additional round of 2D classification, and iterative non-uniform 3D refinement. Data was unbinned once all data sets were combined and one round of 2D classification was performed and particles were extracted with a 300-pixel box size. Before generation of the final map, global ctf, and local ctf jobs were run. The final selected particles were run through non-uniform refinement to generate the final 3D CryoEM map. Local resolution was determined using the final non-uniform refinement job. No symmetry was ever imposed during data processing of either complex. A flowchart of the processing can be found in **Figure 2.11.**

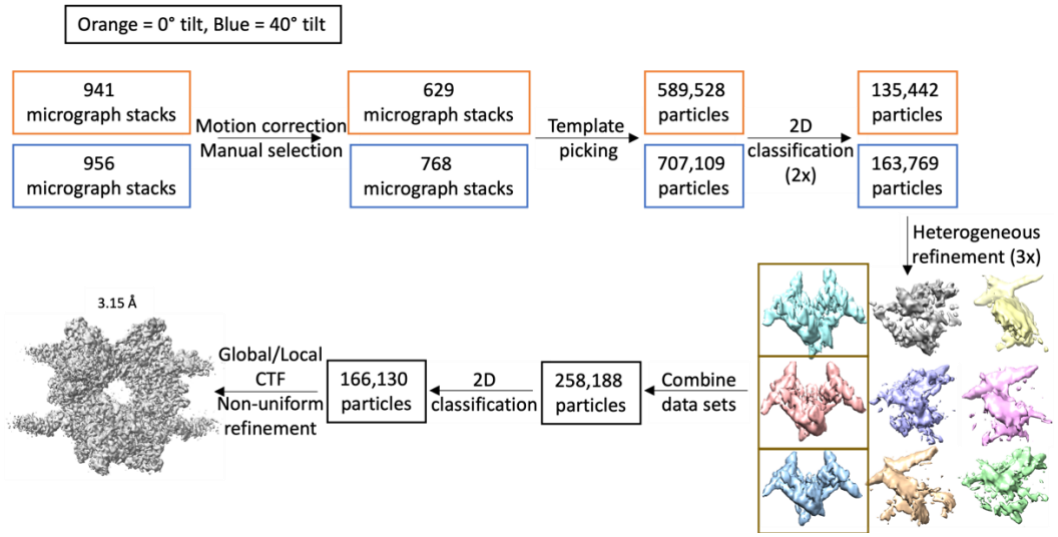

Figure 2.11. General CryoEM data processing workflow.

Heterogeneous structures in figure represent only the 40° tilt series (since this step was conducted before combining the data sets) but was repeated with the 0° tilt series. Chosen structures from the heterogeneous jobs are highlighted with a black box.

For the DNA-bound PaqCI, a total of 299,211 putative particles were template-picked from the two data collections. After processing using the pipeline described above, a total of 166,130 particles were used to generate a 3.15 Å (global resolution map). Additional methodological details of the CryoEM analysis of the DNA-bound PaqCI are provided in **Figure 2.12**.

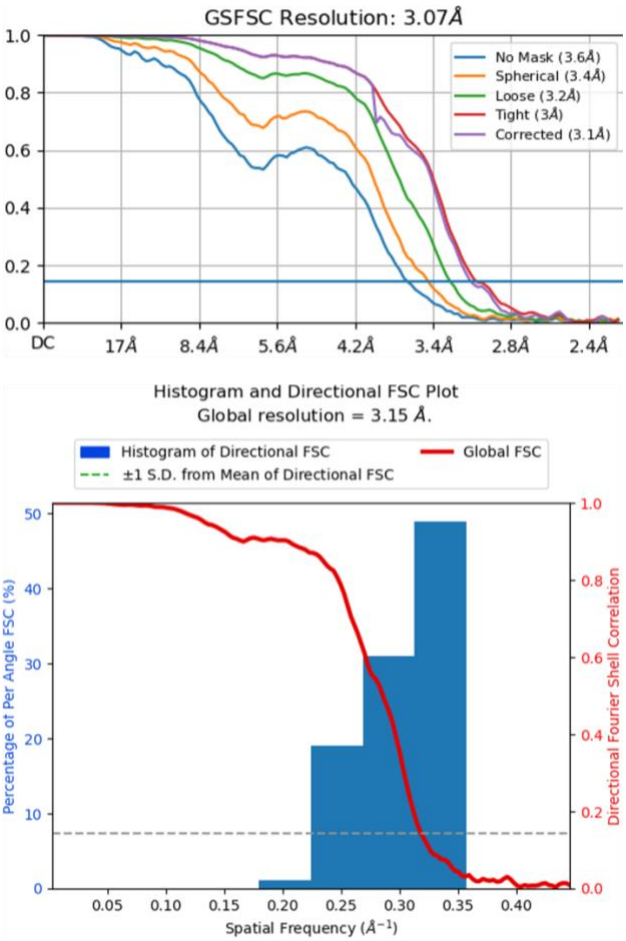

Figure 2.12. Verifying the resolution of the CryoEM map.

Gold-standard Fourier shell correlation (GSFSC) curves and fourier shell correlation (FSC) curve for the 3D reconstruction of DNA-bound PaqCI from cryoSPARC.

CryoEM model building and refinement. The structure of the DNA-bound PaqCI was built and refined using COOT [74], UCSF Chimera [83], and Phenix [72]. Fitting of the atomic model (from unbound PaqCI crystal structure) into the CryoEM map was performed using UCSF Chimera [83] and was manually adjusted in COOT [74]. First, the tetrameric TRDs were docked into density, and then the two visible EN domains were docked in the remaining densities. Once the oligomer was assembled, it was ported into Phenix for refinement using the phenix.real_space_refine application [72]. After a few rounds of refinement,

DNA oligos were added to the complex one at a time, with a round of refinement in between each addition. The final model for DNA-bound PaqCI was refined in Phenix with secondary structure and geometry restraints (**Table 2.2**). All figures and morphs were generated with PyMOL [82], and all movies were generated with UCSF Chimera [83].

Table 2.2. PaqCI CryoEM data collection, refinement, and validation statistics

Data Collection	
EM equipment	Glacios X-FEG (Thermo Fisher)
Voltage (kV)	200
Detector	Gatan K3
Pixel size (Å/pixel)	0.561
Electron dose ($e^-/Å^2$)	50
Defocus range (μm)	0.1 – 4.0
Number of collected micrographs	1,897
Number of used micrographs	1,397
Reconstruction	
Software	cryoSPARC, RELION
Number of used particles	166,130
Symmetry	C1
Resolution (Å)	3.15
Map sharpening B-factor ($Å^2$)	139.1
Refinement	
PDB/EMD ID	8EPX / EMD-28534
Software	Phenix
Cell dimensions	
a=b=c (Å)	336.6
α=β=γ (°)	90
Model composition	
Protein residues	1,663
Side chains assigned	1,663
MolProbity Score	2.27
Rms deviations	
Bonds length (Å)	0.003
Bonds Angle (°)	0.530
Ramachandran plot statistics (%)	
Most favored	94.82
Allowed	5.18
Outlier	0.00

2.3.2 *Results*

A 50 basepair double-stranded DNA construct (**Figure 2.13**) containing the enzyme's target site and sufficient downstream duplex to extend past the enzyme's cleavage site was used to generate a DNA-bound enzyme complex in the presence of calcium (which facilitates the formation of a productive cleavage complex but inhibits cleavage).

Figure 2.13. CryoEM DNA oligo.

Sequence and basepair arrangement of DNA duplexes used to form a DNA-bound enzyme complex. The duplex consists of 50 complementary basepairs and includes both the enzyme's seven basepair target site (red underlined bases) and its downstream cleavage sites on the top and bottom strands (black lines and arrows, 4 and 8 basepairs downstream from the final basepair of the target site).

The formation of the bound complex was validated via SEC analyses of the protein-DNA mixture (**Figure 2.9**) and subsequent negative stain and CryoEM images, which clearly indicated the presence of multiple bound DNA molecules extending from individual enzyme particles (**Figure 2.10**). An electron density map of the DNA-bound enzyme (**Figure 2.14**), corresponding to approximately 3.0 Å to 3.6 Å resolution, was generated from 1,291 micrographs and over 800,000 particles. The real space density was well-resolved throughout the DNA-bound tetramer, including the DNA target site (**Figure 2.15**) and the downstream cleavage site (**Figure 2.16**). The lowest resolution of the CryoEM map corresponds to the distal ends of the DNA molecules and the linker spaces between the EN and TRD domains (**Figure 2.17**).

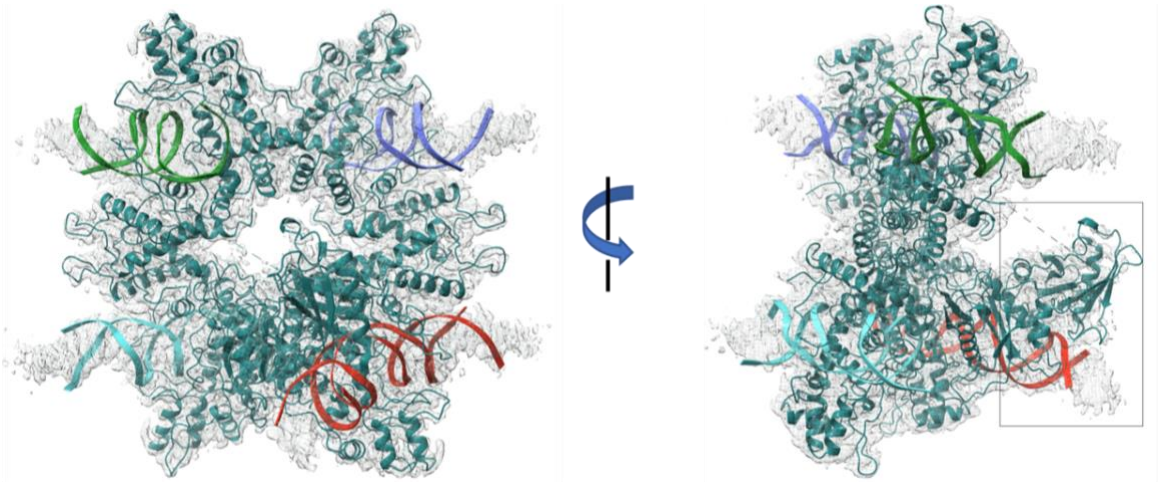

Figure 2.14. CryoEM map with PaqCI tetramer and bound DNA duplexes.

CryoEM electron density corresponding to the DNA-bound PaqCI enzyme. Each of the four double-stranded DNA oligoduplex is independently colored to match that of its bound monomer shown in **Figure 2.19**. The protein tetramer is colored in teal. The right image is rotated 90° around the y-axis and the endonuclease (EN) domains engaged for cleavage are boxed.

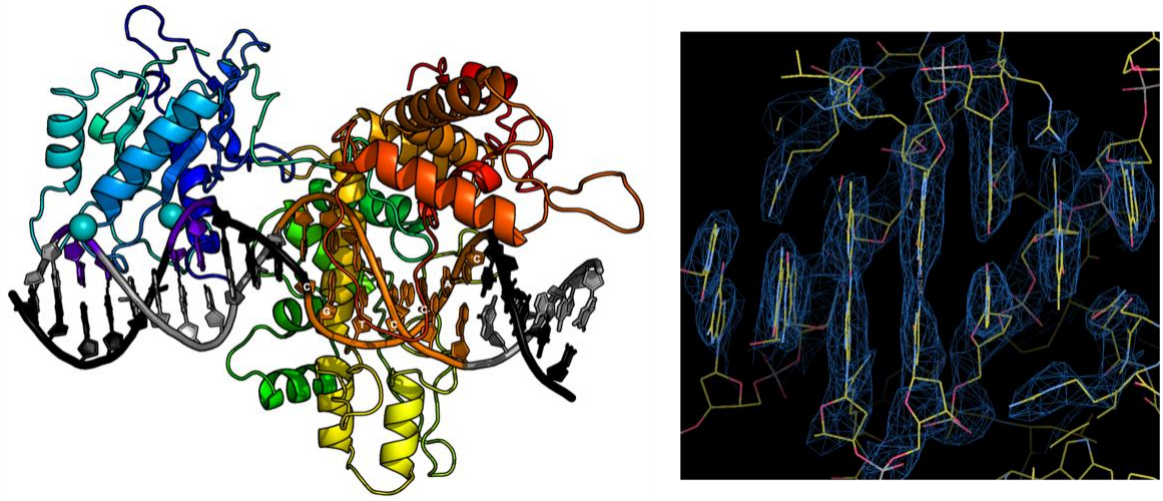

Figure 2.15. Cartoon representation and CryoEM map of target site binding by PaqCI.

Cartoon representation of *cis* DNA-bound protein subunit and CryoEM electron density across the DNA target site bases. The protein ribbon is colored in a spectrum, with the N-terminal endonuclease (EN) domain in shades of blue, and the C-terminal target recognition domain (TRD) colored in shades of yellow, orange and red. Two calcium ions are shown as cyan spheres.

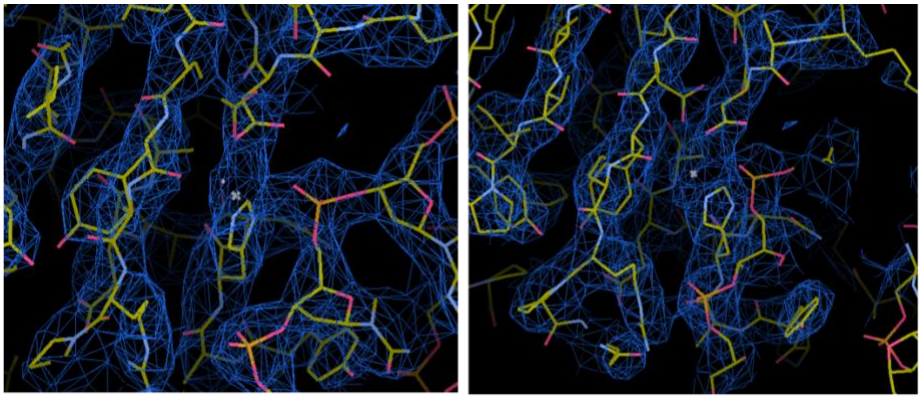

Figure 2.16. CryoEM map of target site cleavage by PaqCI.

CryoEM electron density for the cleavage regions of the *cis*-acting (left) and *trans*-acting (right) endonuclease (EN) domains in complex with their respective scissile phosphate groups on each DNA strand of the enzyme's cleavage site. A single-bound calcium ion is associated with a non-bridging oxygen of each scissile phosphate and a catalytic aspartate residue of each EN domain.

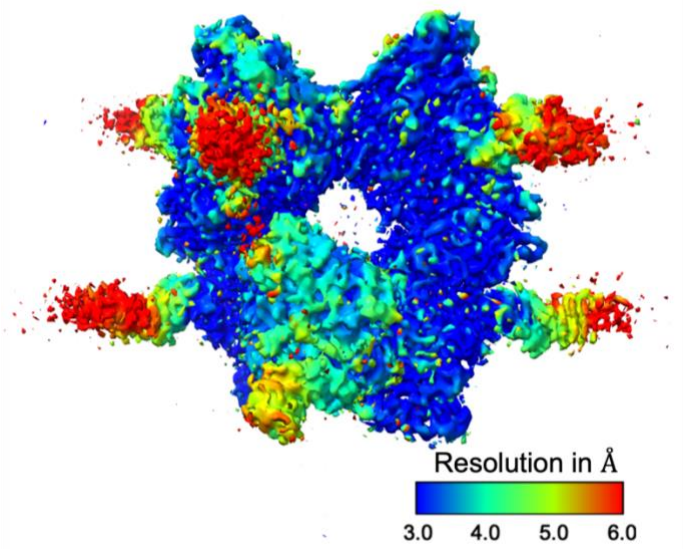

Figure 2.17. Local resolution of CryoEM DNA-bound PaqCI map.

The resolution of the DNA-bound PaqCI 3D CryoEM reconstruction as calculated in cryoSPARC. The image of the map was generated in UCSF Chimera.

3D variability analysis of the CryoEM DNA-bound tetramer map shows a reduction in overall conformational sampling. The largest motions observed within the particles correspond to the regions of the bound DNA molecules that extend away from their binding sites and the protein assemblage. The tetramer appears to remain in a relatively rigid conformation in solution once it is bound to DNA with minimal 'breathing' of the core tetrameric assembly as compared to its unbound counterpart.

Target binding. All four TRDs within the PaqCI tetramer are individually engaged with four corresponding double-stranded DNA duplexes (**Figure 2.14**) through contacts with the enzyme's target site sequence to form a complex to cleave a single DNA duplex. The binding of DNA imparts minimal rearrangement of the TRD tetrameric assembly (**Figure 2.18**). Within each TRD dimer within the larger 'dimer of dimers' tetrameric assemblage, two bound DNA molecules are oriented in a parallel arrangement relative to one another, as indicated by the numbered arrows in the cartoon diagram above the space filled structures of **Figure 2.19**. The opposing pair of bound DNA duplexes are also oriented in a parallel arrangement relative to one another. The two pairs of bound DNA duplexes (labeled 1 and 2 versus 3 and 4 in **Figure 2.19**) point in opposite directions from one another.

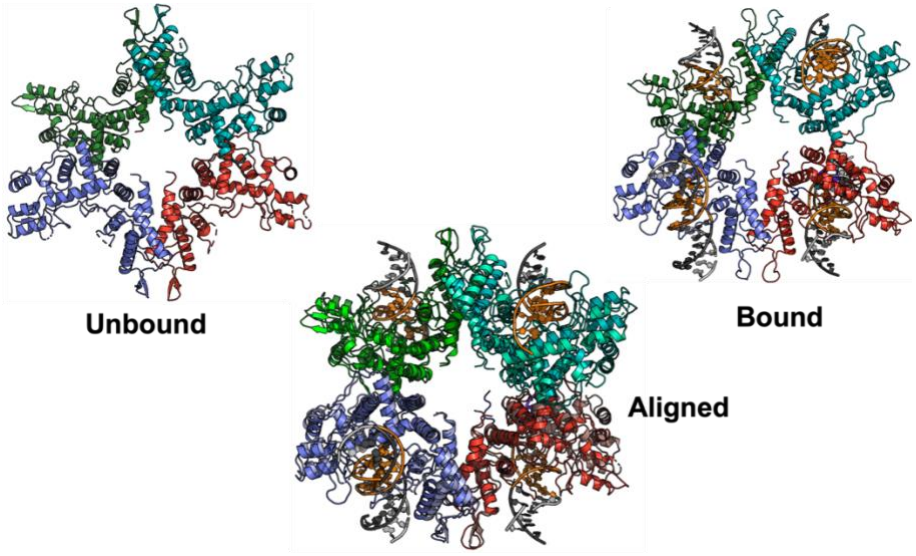

Figure 2.18. Binding of DNA by PaqCI tetramer confers minor TRD movement.

Side by side comparison and superposition of the four target recognition domains (TRDs) in the absence and presence of bound DNA indicates minimal rearrangement of the tetrameric assemblage of those domains.

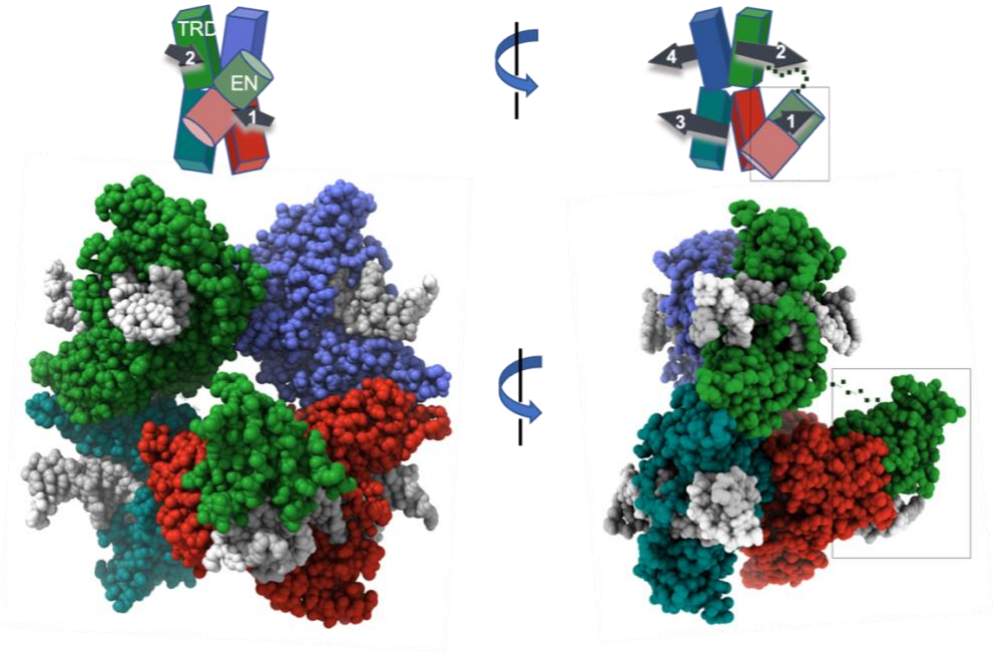

Figure 2.19. Space filling PaqCI CryoEM structure.

Cartoon representation and space filling model of DNA-bound enzyme structure and domain organization. The target recognition domains (TRDs) are represented by rectangles and the EN domains are shown as cylinders. DNA duplexes and their directionality (5' to 3') are denoted with black arrows in the cartoon. Each TRD is engaged with one DNA duplex, via contacts to bases and neighboring backbone atoms distributed across the target site. The downstream region of each bound DNA, including the sites of cleavage, extend away from the TRD tetramer. The DNA duplexes engaged to each protein dimer (DNA 1 and 2 on one side of the complex, and DNA 3 and 4 on the opposite side of the complex) are roughly parallel to one another, as indicated in the cartoon schematics. All four ENs have been released from their positions in the DNA-free apo-enzyme structure. Two ENs (from TRDs colored blue and teal bound to DNA 3 and 4) are disordered and are unobservable in the density map. The other two ENs (from TRDs colored red and green, *cis* and *trans*, bound to DNA 1 and 2) are observed to have undergone significant motions resulting in their engagement around the cleavage site on one bound DNA duplex. The two EN domains engaged for cleavage are boxed.
Size exclusion chromatography traces of DNA-bound complexes, representative negative stain EM class averaged images, and electron density surrounding the DNA cleavage site complexes are further illustrated in **Figure 2.9**.

The TRD in each enzyme monomer in the PaqCI tetramer displays small, localized conformational changes as it wraps around the target site (**Figure 2.20** and **Figure 2.15**)**.** The conformational changes within the TRDs upon double-stranded DNA binding correspond to an overall rmsd between unbound and bound states of approximately 2 Å, with the largest movement corresponding to a slight closure of α-helices and corresponding ordering of adjacent DNA-contacting loops that are unobservable in the DNA-free PaqCI enzyme (**Figure 2.20**). Multiple residues in each TRD are engaged with individual bases and backbone atoms within each target site, forming a site-specific recognition complex (**Error! Reference source not found.**). In particular, one of the three DNA-contacting loops contains four arginine residues, each of which

interacts with the backbone and/or bases of the target site in a sequence-specific manner (**Figure 2.20**). Complementing these interactions, an adjacent α-helix inserts into the major groove near the seven basepair recognition site. This helix is interrupted by a loop of 14 amino acids that is also used for DNA recognition. Seventeen residues are indicated by DNAproDB [84] to be directly involved in recognition of the target site (either via backbone or residue contacts). The DNA contacts exhibited by each subunit are similar, regardless of whether the corresponding TRD is associated with a visible DNA-bound endonuclease.

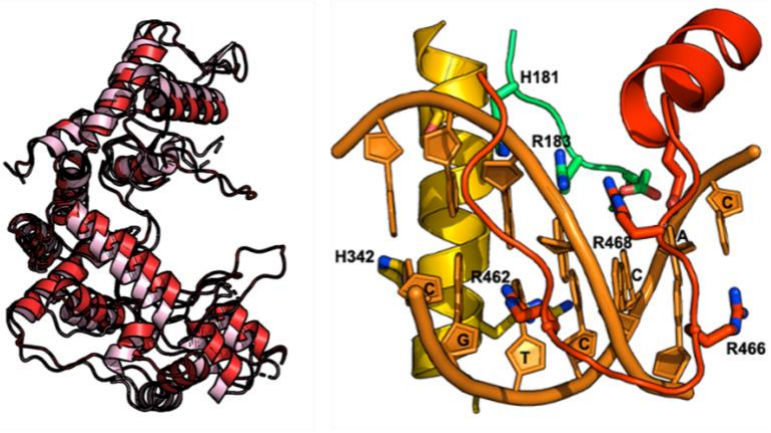

Figure 2.20. Three main loops are involved in DNA target site binding.

Superposition of DNA-free and DNA-bound TRDs (pink and red, respectively) show a slight domain closure and ordering of several DNA-contacting loops (left). A close up of one of the DNA-contacting loops of the TRD, displaying the base specific read out of the enzyme using primarily arginine residues (right).

Observation of a single pre-cleavage complex formed around one bound DNA duplex. In the DNA-bound complex, all four endonuclease domains (ENs) have dissociated from their original positions, where they were sequestered against their corresponding target recognition domains (TRDs). Two of the ENs, extending from one of the two enzyme dimers in the protein tetramer, have reformed a catalytic EN dimer around the cleavage site of one of the bound DNA duplexes (indicated by boxes in all panels of **Figure 2.19** and in **Figure 2.21**). Although both EN domains each move significantly to position themselves around their newly acquired cleavage site, the DNA-bound ENs are re-dimerized in a manner very similar to their association in the DNA-free apo-enzyme, with their protein-protein interactions largely re-established via polar contacts between α-helices of the two EN domains. The backbone rmsd between the EN domain dimer pair in the unbound and DNA-bound structures is approximately 0.8 Å (**Figure 2.22**).

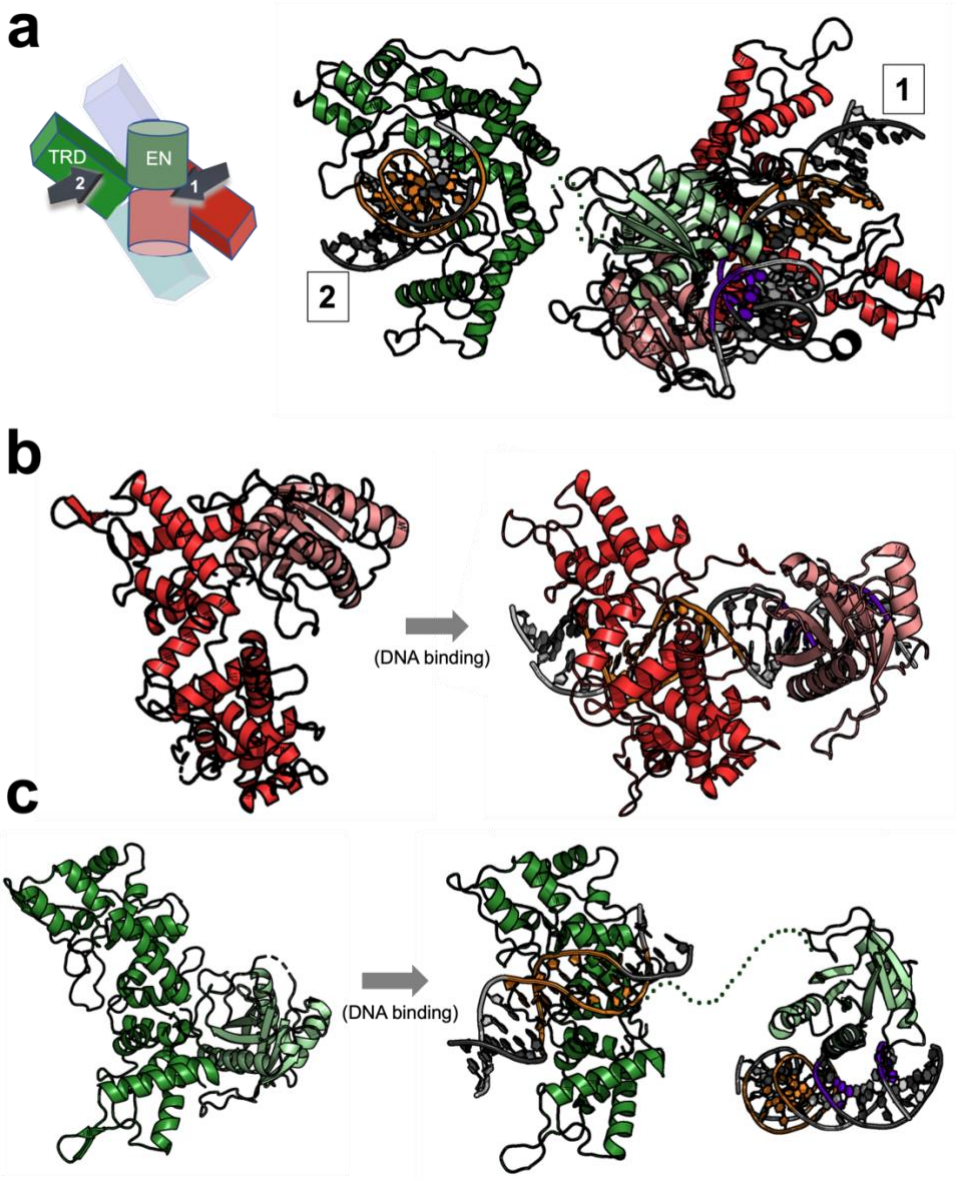

Figure 2.21. Endonuclease motions and engagement on DNA.

Panel a: Ribbon diagram and cartoon schematic of cis and trans protein subunits (as shown in **Figure 2.19**) that position their respective endonuclease (EN) domains on both strands and cleavage positions on a single bound DNA duplex. In the cartoon schematic the DNA duplexes are drawn as arrows and numbered according to their protein subunit (*cis* = 1, *trans* = 2) (as shown in **Figure 2.19**). EN domains are colored in a lighter shade than the target recognition domains (TRDs). The cleavage sites of the DNA duplex are shown in purple and the target site in orange. The 16 amino acid linker between the trans-acting EN domain and its C-terminal TRD is poorly ordered in the model and are represented in the figure as a dotted green line.
Panel b: EN domain motion of the cis-acting endonuclease domain (red subunit in **Figure 2.19**) acting on DNA 1. The EN domain (pink) swings about 180° about the axis to align with the cleavage site of DNA 1 to cut four bases from the target site while the TRD (red) stays fixed except for slight closure of the helices to bind DNA.
Panel c: EN domain motion of the trans-acting endonuclease (green subunit in **Figure 2.19**) acting on DNA 1. The TRD (green) is bound to a separate DNA duplex (DNA 2) than the one bound by the cis-acting EN and the trans-acting EN engage for cleavage (DNA 1). The EN domain (light green) reaches across the belt of TRDs to locate the cleavage site of the DNA 1 to cleave.

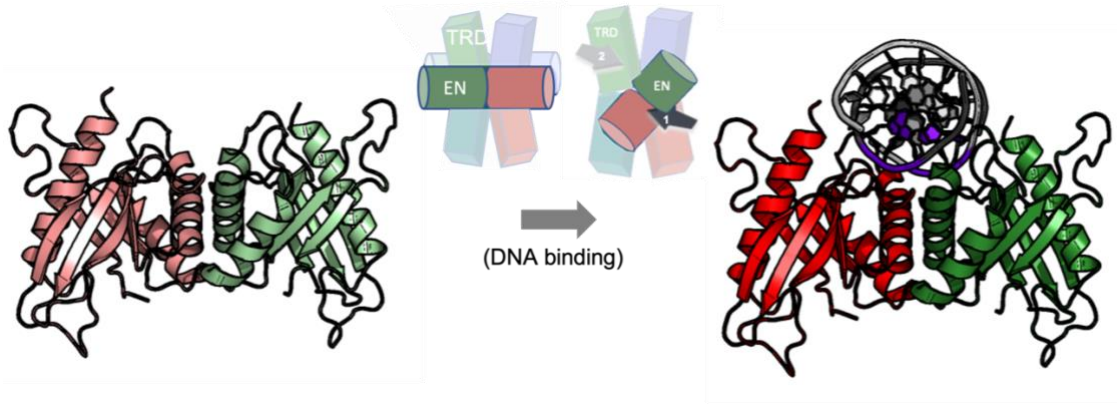

Figure 2.22. Contacts between EN domains in the enzyme tetramer in the presence and absence of a DNA cleavage site are conserved.

Endonuclease (EN) domain dimer organization in the DNA-free and DNA-bound PaqCI structures (left and right, respectively) shown as cartoon and ribbon diagrams. The dissociation of the two domains from their respective target recognition domains and subsequent re-association with the two DNA strands at the sites of cleavage result in nearly identical contacts and association between the two domains (backbone rmsd 0.8 Å).

One of the two DNA-bound EN domains (colored red in **Figure 2.19** and in **Figure 2.21**) is engaged with the DNA cleavage site in *cis* (i.e., that domain extends from the TRD that is engaged with the same DNA duplex). The other EN domain (colored green) is engaged with the opposing strand of the same cleavage site in *trans* (i.e., that domain extends across the dimer interface from a TRD bound to a different DNA duplex and target site). The remainder of this manuscript refers to the two DNA-bound EN domains as the *cis*-acting and *trans*-acting EN domains.

The *cis*-acting EN domain is positioned appropriately for cleavage of the phosphodiester located four basepairs from the target site, while the *trans*-acting EN domain (which has moved a much longer distance to re-establish contact with its partner) is positioned appropriately to cleave the phosphodiester bond eight basepairs from the target site on the opposing strand (**Figure 2.21a**). Whereas the *cis* monomer twists around the duplex by about 180° to orient the EN domain over the phosphate of interest (**Figure 2.21b**), the *trans* monomer reaches approximately 40 Å across the dimer interface, thereby reforming a new EN dimer and positioning itself appropriately for cleavage of its target phosphate. The *trans* monomer relies on a linker of 16 amino acids that connects it to its own DNA-bound TRD to reach its cleavage site (**Figure 2.21c**). The *trans* endonuclease appears to be near the maximum motion away from its TRD that would be sterically permitted by its linker.

In the complex between the DNA target site and its re-dimerized *cis*-acting and *trans*-acting ENs, each active site is positioned appropriately around the precise phosphate groups corresponding to the known cleavage pattern for the PaqCI enzyme (**Figure 2.21** and **Figure 2.22**). The EN domains each coordinate a metal ion in direct contact with the phosphate to be cleaved (**Figure 2.16**). Three conserved catalytic residues (D54, E73, and K75) complete a canonical endonuclease active site via coordination of bound metal ions and appropriate positioning for proton transfer and stabilization of the phosphoanion transition state of the hydrolysis reaction.

The opposing EN domains are disordered. The opposing two EN domains (also previously packed against their own TRDs in the DNA-free apo-enzyme), which are also released from their TRDs as a result of DNA binding, are unobservable and appear to be disordered. Even though all four DNA duplexes and their corresponding TRDs are well resolved, no class exists within the 2D classifications that display more than the two DNA-bound EN domains described above. If the unobservable EN domains are able to simultaneously associate with their own DNA target sites, that interaction would appear to be too short-lived in the trapped pre-cleavage state described here for the CryoEM analysis to resolve.

The asymmetry between the two sides of the DNA-bound tetramer (leading to the formation of a cleavage-ready complex to a single-bound DNA on only one side of the enzyme tetramer) may imply that engagement of EN domains on one bound double-stranded DNA imposes a subtle asymmetry between the two enzyme dimers that leads to a corresponding inability of the opposing EN domains to reach far enough to easily form a comparable interaction with a DNA cleavage site on the opposite side of the enzyme assemblage. This asymmetry is evident when comparing the relative conformation of the ‘DNA-cleaving' TRD dimer with the opposing TRD dimer. The two dimers in the DNA-bound structure display different subunit packing, corresponding to a backbone rmsd of approximately 3 Å between the unaligned TRDs in a superposition (**Figure 2.23**). Furthermore, the superposition of individual TRDs from opposing sides of the assemblage demonstrates that the conformational change induced by DNA binding and EN engagement involves a rigid body motion of the two TRDs within a dimer relative to one another.

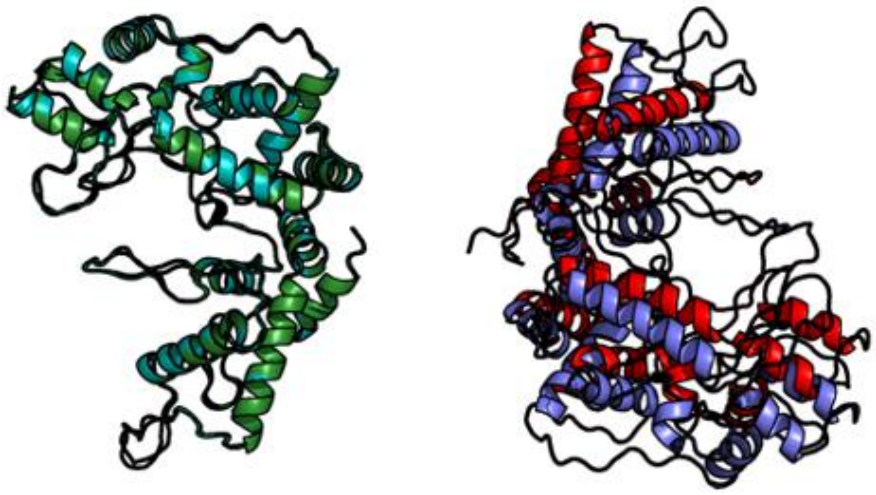

Figure 2.23. Asymmetry between DNA-cleaving and non-DNA-cleaving dimer.

Superposition of the 'cleaving' and 'non-cleaving' DNA-bound target recognition domains (colors correspond to **Figure 2.19**) indicates that an asymmetry in the TRD tetrameric structure is induced when the *cis*- and *trans*-acting endonuclease domains associate with a DNA cleavage site associated with one of the two enzyme dimers.

These observations appear to also agree with correlated motion analyses of the CryoEM maps for the unbound enzyme tetramer, which appears to indicate an oscillation across the dimer-of-dimers interface that corresponds to sampling of transient asymmetry between the two sides of the enzyme assemblage. The asymmetry between the two sides of the enzyme suggests that the enzyme cleaves in a sequential manner and is unable to be positioned to cleave two sites simultaneously.

The atomic interactions between the monomers of the DNA-free and DNA-bound tetrameric structures were analyzed using the PISA software to provide additional details of the overall interactions required for enzyme tetramerization and DNA cleavage [85]. While the tetramer is bound to DNA, the only contacts holding the quaternary structure together are in the TRDs, preserving the TRD 'ring' in both structures (**Figure 2.18**). In the unbound structure, the EN domains are held against the TRD ring via contacts between periphery EN loops and a single TRD loop on its neighboring and adjacent monomer. This locks the catalytic domain against the interior of the tetramer and prevents it from cleaving nonspecific DNA duplexes. The EN domains of the DNA-bound structures make no contact with the TRDs, but the contacts between EN domains are preserved (**Figure 2.22**).

2.4 BIOCHEMICAL ANALYSIS OF PAQCI AGREE WITH THE STRUCTURAL FINDINGS

2.4.1 *Methods*

Biochemical analyses of target site specificity and cleavage activity. Oligoduplex primers (**Table 2.3**) were synthesized by IDT. Molecular biology reagents including Q5 Hot Start High-Fidelity DNA polymerase, NEBuilder HiFi DNA assembly master mix, restriction enzymes, DNA size standards, lambda DNA, and competent cells were from New England Biolabs (NEB). Plasmid purification and nucleic acid purification clean ups were performed using Monarch DNA kits (NEB). Plasmid DNA constructs were confirmed by sequencing on an ABI 3130xl capillary machine (Applied Biosystems).

Table 2.3. Oligonucleotide primers

PaqCI_pUC19 1 For	CCT TTC GT**C ACC TGC** GTT TCG GTG ATG ACG GTG AA
PaqCI_pUC19 1 Rev	ACC GAA AC**G CAG GTG** ACG AAA GGG CCT CGT GAT ACG
PaqCI_pUC19 700HTT For	GCT TCC TCG **CAC CTG C**AC TCG CTG CGC TCG GTC
PaqCI_pUC19 700HTT Rev	GCA GCG AGT **GCA GGT G**CG AGG AAG CGG AAG AGC GCC
PaqCI_pUC19 700HTH For	GCT TCC TCG **GCA GGT G**AC TCG CTG CGC TCG GTC
PaqCI_pUC19 700HTH Rev	GCA GCG AGT **CAC CTG C**CG AGG AAG CGG AAG AGC GCC
PaqCI_pACYC184 camFOR	TAC T**GC AGG TG**C GAA GAG CAC TGG TGT CCC TGT TGA TAC CG
PaqCI_pACYC184 camREV	GTG A**GC AGG TG**C TGA GAC GAA CCA GGC GTT TAA GGG CAC CA
pUC19_addPaqCI_For	TGC TCT TCG **CAC CTG C**AG TAT ATA TGA GTA AAC TTG GTC TGA CAG TTA CCA
pUC19_addPaqCI_Rev	TCG TCT CAG **CAC CTG C**TC ACG TTC CAC TGA GCG TCA GAC C

*PaqCI recognition site in **BOLD** type
**HiFi assembly overlap underlined

PaqCI-site plasmid substrate construction. Plasmid DNA substrates were constructed containing either a single PaqCI recognition site, or two recognition sites in either head-to-head (HTH) or head-to-tail (HTT) orientation, or four recognition sites. PaqCI recognition sites were introduced at position 1 and/or position 700 of pUC19 by mutagenic PCR. The single-site construct (p1SS) was PCR amplified using PaqCI_pUC19 1 For/PaqCI_pUC19 1 Rev (see **Table 2.3**) primer pairs. Two-site constructs (p700HTT and p700HTH) were amplified as two amplicons: PaqCI_pUC19 1For/PaqCI_pUC19 700HTT Rev & PaqCI_pUC19 700HTT For/PaqCI_pUC19 1 Rev or PaqCI_pUC19 1 For/PaqCI_pUC19 HTH Rev & PaqCI_pUC19 700HTH For/PaqCI_pUC19 1 Rev (see **Table 2.3**). PCR amplicons were confirmed by agarose gel electrophoresis. Template DNA was removed by digestion with DpnI (NEB) at 37 °C for 30 minutes and purified using NEB Monarch nucleic acid purification kit following manufacturer's instructions.

Purified amplicons were assembled using NEBuilder® HiFi DNA assembly and transformed into NEB® 5α chemically competent *E. coli* according to manufacturer's instructions. Individual colonies were grown overnight in LB broth supplemented with ampicillin (100 μg/ml). Two unique four-site plasmid DNA substrates were created by inserting an approximately 850bp fragment containing the chloramphenicol gene amplified from pACYC184 flanked on both ends by PaqCI recognition sites into either the HTH or HTT 2-site plasmid substrates described above. The additional DNA was inserted 800bp downstream of the 2nd recognition site using NEBuilder HiFi DNA assembly master mix following manufacturer's instructions. Plasmid DNAs were isolated from overnight cultures, and the introduced PaqCI recognition sites were confirmed via Sanger sequencing.

PaqCI cleavage activity. One unit of enzyme is defined as the amount required to digest 1 μg of lambda DNA to completion in 1 hour at 37°C in a 50 μL volume. One unit of PaqCI is equivalent to 24 ng or 0.43 picomoles of enzyme monomer, which in a 50 μL reaction corresponds to 8.6 nM enzyme monomer concentration. The concentration of target sites in lambda phage DNA (containing 12 PaqCI sites) in the same 50 μL reaction is 7.2 nM. Digests were performed either with or without a trans-activating oligonucleotide (a short double-stranded hairpin DNA construct spanning the PaqCI recognition site, corresponding to the sequence 5'- GGA GCAGGTG AGCGAG TTTT CTCGCT CACCTGC TCC -3', that possesses the binding site (underlined) and does not extend past the point of cutting). The range of enzyme concentrations employed in units (16 units to 0.125 units) corresponds to approximately 140 nM to 1 nM enzyme monomer in the 50 μL reaction volume. For clarity we report enzyme monomer and DNA target sites concentrations because each enzyme monomer will bind one DNA target site, while acknowledging that in solution the enzyme exists as a tetramer that can bind four target sites simultaneously.

The single-site plasmid substrate (p1SS), dual-site plasmid substrates (p700HTT and p700HTH), and lambda DNA (NEB) were digested with variable concentrations (140 nM to 2.2 nM) of PaqCI, either in the presence or absence of the in trans-activating DNA oligoduplex. PaqCI enzyme was serially diluted in reaction buffer (rCutSmart™: 20 mM Tris-acetate, 10 mM magnesium acetate, 50 mM potassium acetate, 100 μg/ml recombinant albumin, pH 7.9 @ 25 °C) containing 1 μg substrate DNA per 50 μL and incubated

for 1 hour at 37 °C. Reactions with trans-activating oligoduplex were performed identically, with the activator added to the first reaction at a concentration of 40 nM activator per unit (24 ng or 8.6 nM) of PaqCI and then serially diluted with the enzyme to maintain a constant ratio of approximately 5:1 activator to enzyme. Digestion reactions were visualized on a 1% agarose gel.

The relative rate of REase activity and whether sites are cleaved in a coordinated manner was examined by cleavage of the dual-site plasmids p700HTT and p700HTH, either in the presence and absence of activating oligoduplex. Cleavage assays were performed at 37°C in rCutSmart™ buffer containing 1 µg substrate DNA (22.8 nM PaqCI sites), 2 units (48 ng or 17.2 nM) of PaqCI per 1 µg of DNA, and 80 nM activating oligoduplex (40 nM per unit of PaqCI) when indicated. Aliquots of the reaction were removed at 0.25 minute (min), 0.5 min, 1 min, 3 min, 5 min, 10 min, 30 min, and 1 hour. Reactions were terminated by adding stop solution containing 0.08% SDS (NEB Gel Loading Dye, Purple), and digestion reactions were visualized on a 1% agarose gel.

Substrates containing four target sites were digested using a ratio of 0.5:1, 1:1, 2:1 and 4:1 enzyme-binding-domains to DNA-recognition-sites ratio over a time course. To examine cleavage kinetics the four-site plasmid was pre-bound with enzyme at a 1:1 enzyme-binding-domain to recognition-site ratio in a buffer equal to CutSmart™ buffer but lacking any Mg2+ ions required for DNA cleavage to allow the DNA-enzyme complex to form prior to initiating cutting. Pre-binding was performed for 15 minutes at 37 °C in rCutSmart™ buffer lacking Mg2+ ions containing 1 µg substrate DNA in 50 µL (37 nM PaqCI sites) and 4.3 units (37 nM) of PaqCI. Cleavage was initiated by adding Mg2+ to 10 mM and aliquots were removed at 0.25 minute (min), 0.5 min, 1 min, 3 min, 5 min, 10 min, 30 min, and 1 hour.

2.4.2 *Results*

The purified enzyme runs as a large single band in an SDS-PAGE gel, with a larger faint band that may correspond to a larger persistent oligomer (**Figure 2.1**). The same protein sample behaves as a multi-subunit assemblage (appearing to correspond to a protein tetramer) when eluting from a size exclusion chromatography column (SEC) (**Figure 2.1**).

PaqCI demonstrates optimal activity at an equal to modest excess of enzyme binding sites to substrate target motifs, with an excess of enzyme to substrate leading to faster, but not more complete,

cutting (**Figure 2.24a**). Digest time points are quenched at 10 minutes and 60 minutes with protein (monomer) concentration ranging from 0.28 nM to 70.0 nM, while the DNA recognition site concentration is constant at 7.2 nM in each reaction. In the 10-minute digest, near-complete cleavage of the DNA is only seen in concentrations from 17.5 nM to 70.0 nM, with partial cleavage to no cleavage in the rest of the lower concentrations. At 60 minutes, enzyme concentrations greater than 1:1 enzyme-to-sites exhibit near-complete cleavage, while 0.5:1 enzyme-to-sites generates only slightly less cutting, while lesser amounts of enzyme result in more partial digestion.

PaqCI cleaves DNA more efficiently when a short hairpin DNA duplex containing the enzyme's target site (but not downstream DNA basepairs corresponding to the adjacent cleavage site) is added to the reaction as a *trans*-activating factor (**Figure 2.24b**). In this experiment, lambda phage DNA, with 12 PaqCI sites representing 7.2 nM sites, was digested with PaqCI from 70 nM to 1 nM in the absence or presence of the *in trans* target site oligoduplex, added at a 5:1 ratio to the PaqCI enzyme (350 nM to 5 nM activator). The presence of the *in trans* duplex target site enables essentially complete cutting of the substrate at enzyme concentrations greater than the concentration of substrate sites.

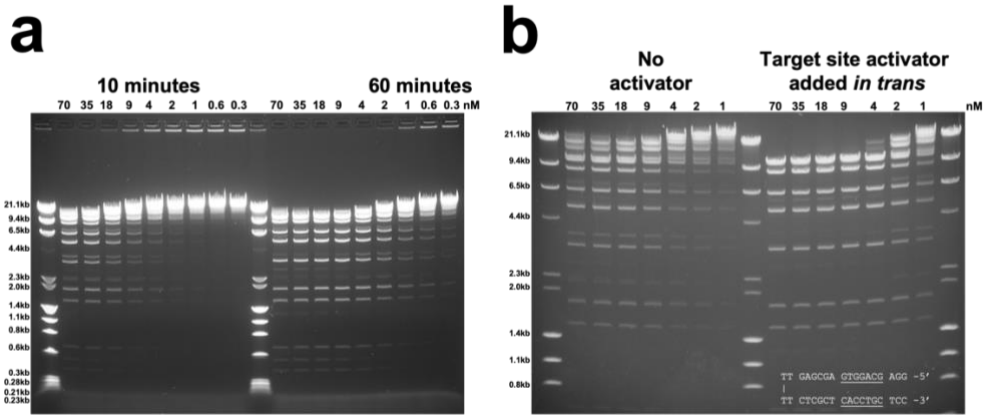

Figure 2.24. Cleavage activity of PaqCI towards lambda phage DNA.

Panel a: PaqCI demonstrates optimal activity at an equal to modest excess of enzyme binding sites to substrate target motifs, with an excess of enzyme to substrate leading to faster, but not more complete, cutting. Lambda phage DNA, with 12 PaqCI sites, are digested and quenched at various time points. Time points correspond to 10-minute and 60-minute digests, with the concentration of enzyme monomers varied from 0.3 nM to 70 nM. The DNA recognition site concentration in all digests was 7.2 nM. The molecular weight ladders Lambda-HindIII/PhiX174_HaeIII were used in panel a and b which spans 0.23 kb to 23.2 kb. This experiment was performed once. Concentrations are rounded to the nearest decimal place. **Panel b:** PaqCI cleaves DNA more efficiently when a short hairpin DNA duplex (inset in gel) containing the enzyme's target site (underlined bases), is added to the reaction as a trans-activating factor. Lambda phage DNA was digested with variable concentrations of PaqCI from 1 nM to 70nM (monomeric concentration) in the absence or presence of the trans activator. All lanes correspond to 60-minute digests. This experiment was performed once. Units were rounded to the nearest decimal place.

An additional experiment examining the cleavage of plasmid substrates (**Figure 2.25**) demonstrates that PaqCI requires the interaction between multiple bound sites for cutting. Plasmid substrates containing two target sites are cleaved more efficiently than the same plasmid containing only one target site. The single-site substrate (11.4 nM sites) is only slightly cleaved while the 2-site substrate (22.8 nM sites) is nearly completely cut. The difference in cleavage efficiency between single- or multiple-target plasmid substrates is eliminated by the presence of the activator DNA (5:1 ratio to enzyme) containing a target site added in *trans.* The enzyme's cleavage efficiency against plasmids harboring two target sites does not appear to differ significantly when the target sites (which are separated by approximately 700 bp) are positioned in a head-to-head or head-to-tail arrangement, though there may be slightly greater cutting on the head-to-tail plasmid.

In a subsequent experiment that measured a time course of plasmid substrate cleavage, PaqCI displayed sequential cleavage of the two target sites (**Figure 2**). This was demonstrated by the rapid accumulation of an initial linearized product ('Lin' in the figure, corresponding to formation of a single double-strand break) followed by subsequent, slower generation of the final products corresponding to a second double-strand break ('Lin1' and 'Lin2' in the figure). In this experiment, relaxed circular plasmid does not visibly accumulate at the early points of the time course, indicating that the cutting of top and bottom strands occurs at similar rates during the first DNA cleavage event. When the same activator DNA construct harboring the PaqCI target site is added in *trans*, similar results are observed but occur on a faster time scale (**Figure 2.26**).

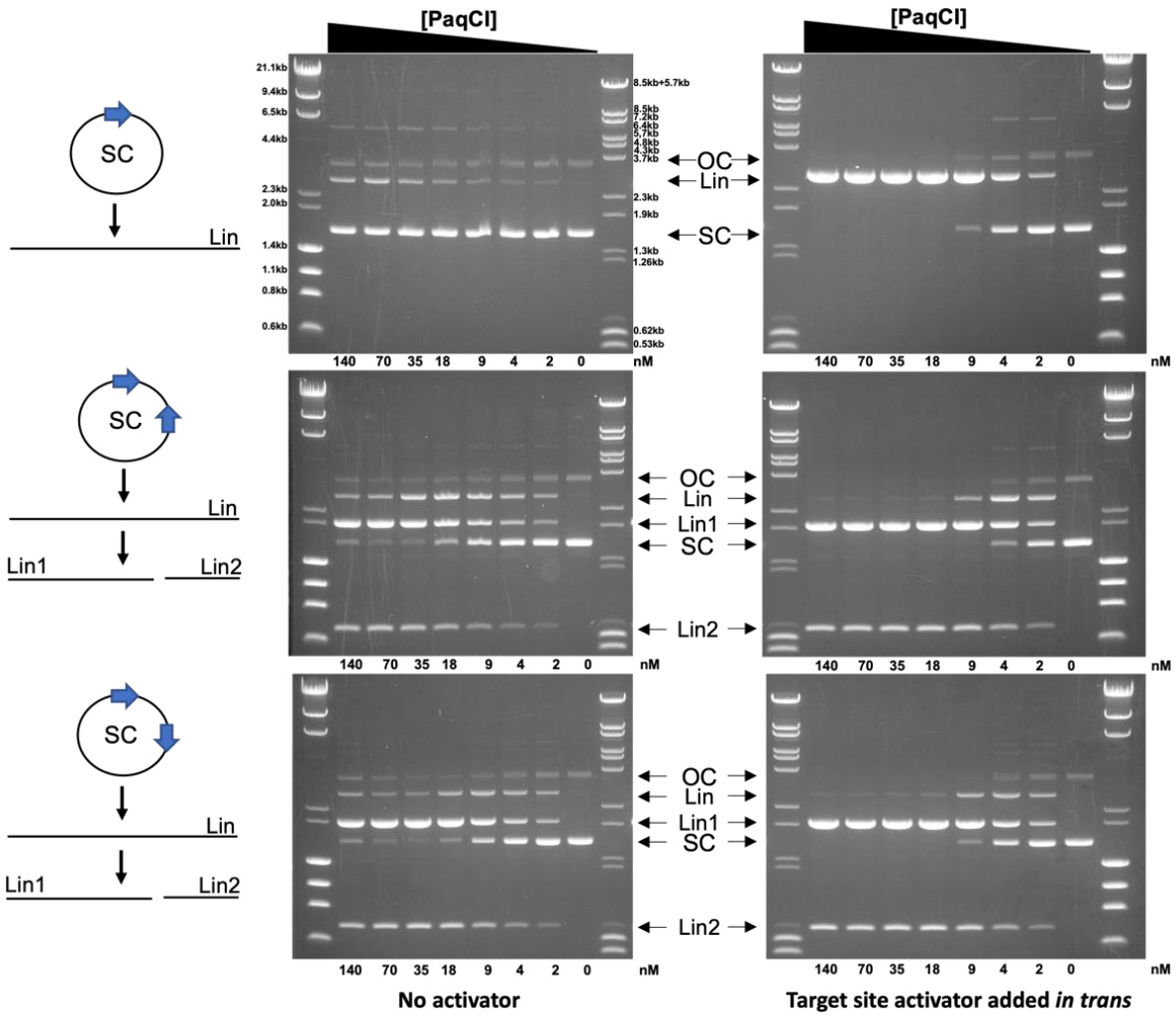

Figure 2.25. Cleavage activity of PaqCI towards plasmid substrates.

Further experiments examine the cleavage of plasmid substrates and demonstrate that PaqCI requires the interaction between multiple bound sites for cleavage. All lanes correspond to 60-minute digests at fixed DNA concentrations, with variable enzyme (monomeric concentrations). The single-site substrate (11.4 nM sites) is cleaved far less efficiently than the 2-site substrates (22.8 nM sites). The difference in cleavage efficiency between single- or multiple-target plasmid substrates is eliminated by the presence of the activator DNA (5:1 ratio to enzyme) containing a target site added in trans. The key shows open circle (OC), single-cut linear (Lin), double-cut products (Lin1, Lin2), and super-coiled (SC) species. Two molecular weight ladders were used in these experiments; (i) Lambda-HindIII/PhiX174_HaeIII which spans 0.6 kb to 23.1 kb, and (ii) Lambda-BstEII/pBR322-MspI which spans 0.5 kb to 8.5 kb. This experiment was performed once.

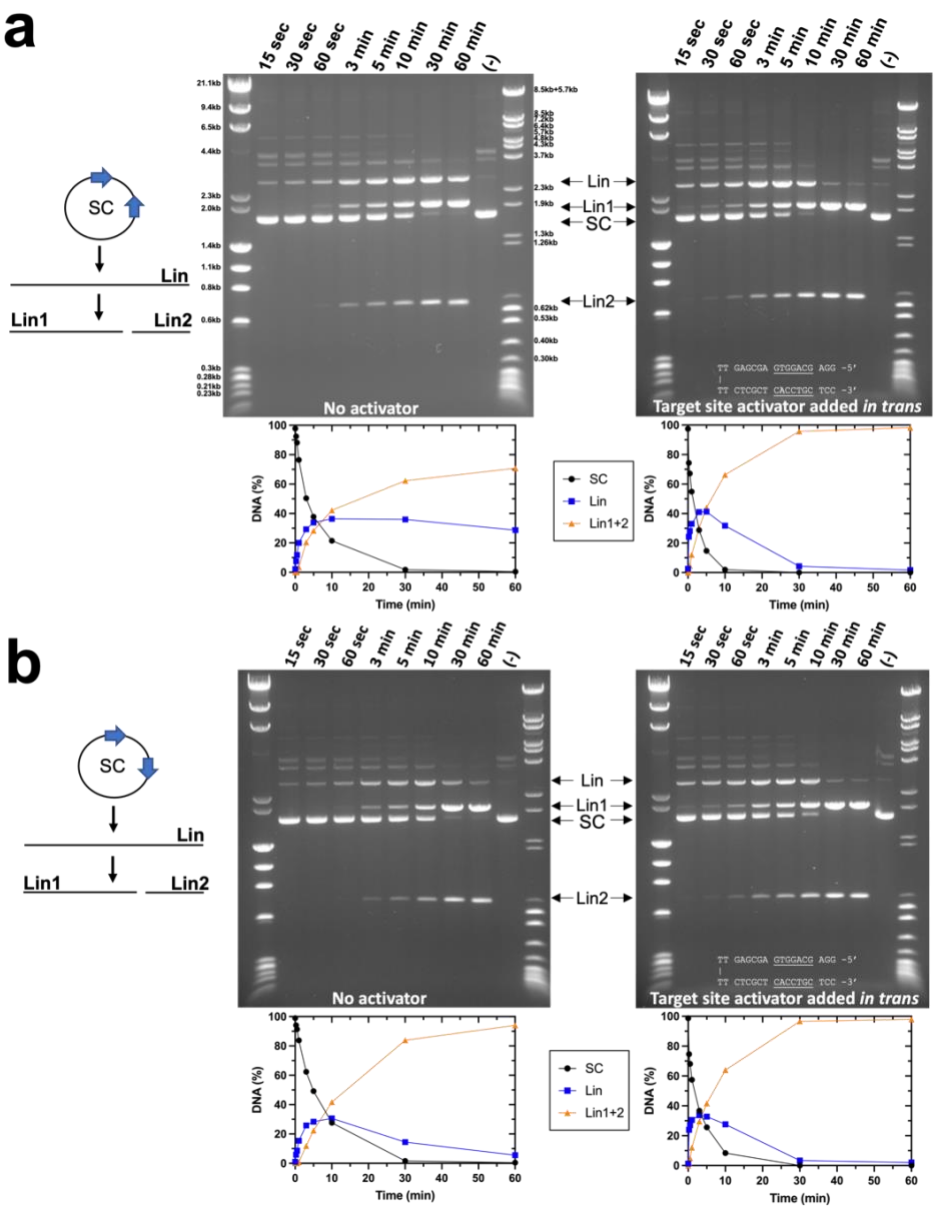

Figure 2.26. Time-dependent digestion of plasmids with two PaqCI target sites with and without added activator indicate a sequential cleavage profile.

Panel a. The no activator reaction (left gel) contained 22.8 nM DNA target sites and 17.2 nM of monomer PaqCI per 22.8 nM DNA sites incubated in 50 µl buffer. The activator reaction (right gel) contained the same proportion of DNA to PaqCI with 80 nM of added activating oligoduplex incubated in 50 µl buffer. To the left of the gel is a cartoon of the head-to-head substrate and reaction products shown in the gel. Digest time points were quenched at 0.25 minute (min), 0.5 min, 1 min, 3 min, 5 min, 10 min, 30 min, and 60 min. The mobilities of the super coil (SC), and the cleaved linear (Lin, Lin1, and Lin2) forms of the plasmid are indicated in the middle key with arrows. The intensities of each substrate and product band in the gels were quantitated using ImageJ [86] software and calculated as rough percentages to be shown graphically. The key in the middle shows supercoiled DNA substrate (purple circles, SC), linearized DNA (blue square, Lin), and double cut long and short linearized DNA (orange triangle, Lin1+2). Two molecular weight ladders were used in this experiment; (i) Lambda-HindIII/PhiX174_HaeIII which spans 0.2 kb to 23.1 kb, and (ii) Lambda-BstEII/pBR322-MspII which spans 0.3 kb to 8.5 kb. This experiment was performed once. **Panel b.** Head-to-tail reactions were quenched at the time points indicated above each lane and experiments were conducted as in Panel a. This experiment was performed once.

The structural analyses described above suggest that even when bound simultaneously to four target sites, PaqCI likely displays a random, sequential mechanism in which one double-stranded DNA at a time is cleaved within such a reaction synapse. That hypothesis was further tested in a final experiment using plasmid substrates harboring four target sites (**Figure 2.27**). Digest time points were quenched at varying time points from 15 seconds to 1 hour under the same conditions as the experiments performed in **Figure 2.26**. Each target site is placed between 700 to 900 base pairs from each other, allowing for sufficient flexibility of the plasmid DNA relative to the known persistence length of DNA [87]).

The experiment was conducted on two different plasmids, which are reflected in the corresponding plasmid schematic found on the left-hand side of **Figure 2.27**. These plasmids mimic as close to possible a head-to-head (**Figure 2.27a**) and head-to-tail (**Figure 2.27b**) configuration to compare with previous kinetic data presented in this paper. In all four experiments, the super-coiled (SC) species rapidly decreases as the single-cut linear and dual- or triple-cut intermediate fragments appear before the final fragments (frag1-4) from cutting at all four sites appear. As observed for the two-site substrate in **Figure 2.26**, there does not seem to be a major difference in the performance of the enzyme when cleaving head-to-head or head-to-tail plasmids in the absence of the activator. Prior to the formation of the complete digestion species (frag1-4), there is an equal accumulation of bands (Linear) that reflect the intermediate states of cleavage. Each band larger than the final products (frag1-4) represents a different combination of cleavage events across the four-site plasmid, associated with single-, dual-, and triple-cut substrate generated prior to the complete cleavage at all four sites. In the activator experiments, seen in the right-most gels in the figure, cleavage is remarkably similar, with a slight increase in the linear plasmid species resulting from a single cleavage event as the reaction proceeds, and more complete cutting as the reaction proceeds.

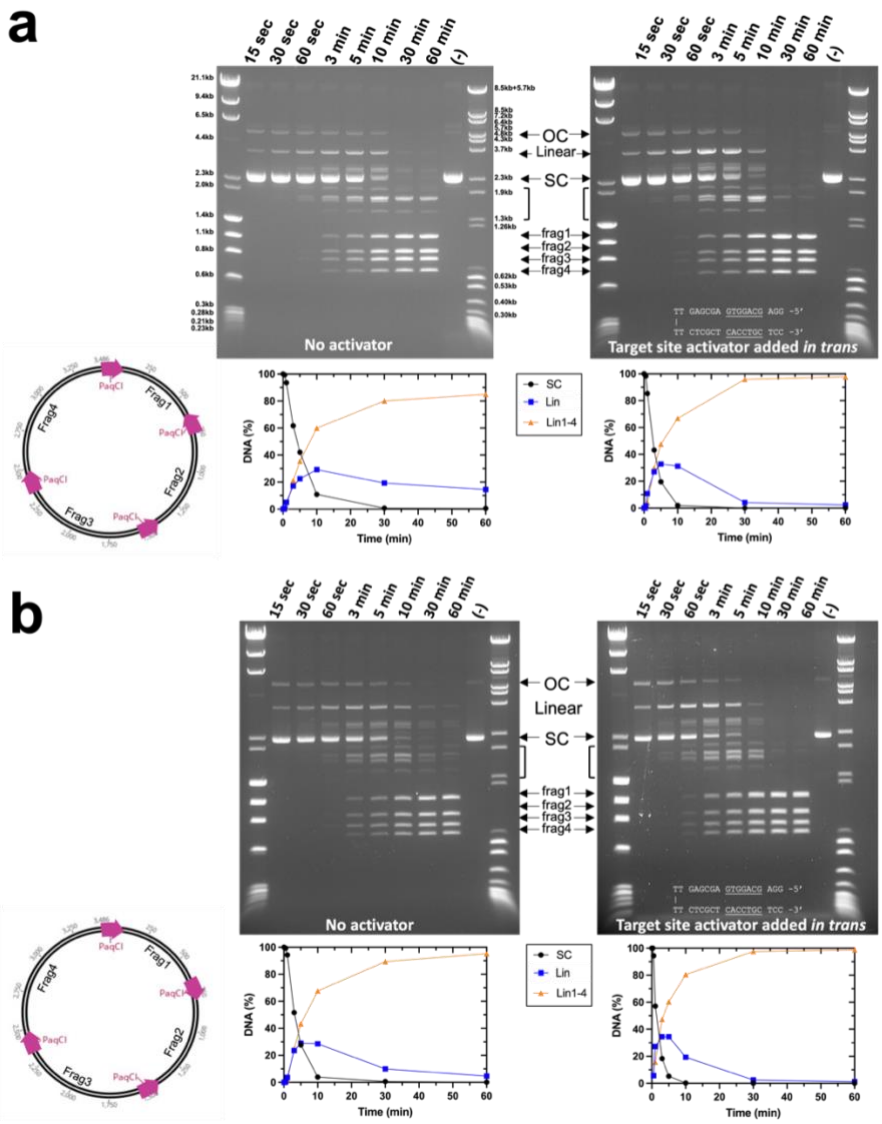

Figure 2.27. Time-dependent digestion of plasmids with four PaqCI target sites indicate a sequential cleavage at individual sites.

Panel a. Cleavage of a four-site substrate with and without an activator added in *trans*. The no activator reaction (left gel) contained 37 nM DNA target sites and 17.2 nM of monomer PaqCI per 37nM DNA target sites incubated in 50 µl buffer. The activator reaction (right gel) had the same substrate and PaqCI enzyme concentrations with 80 nM of added activating oligoduplex incubated in 50 µl buffer. Digest time points were quenched at 0.25 minute (min), 0.5 min, 1 min, 3 min, 5 min, 10 min, 30 min, and 60 minutes. The mobilities of the super coil (SC), single-cut linear (Linear), intermediates (brackets), and the complete-cleavage linear (frag1, frag2, frag3, and frag4) forms of the plasmid are indicated in the middle key with arrows corresponding to their related bands. Intermediate bands correspond to a less than fully cleaved plasmid; either 1, 2, or 3 cleavage events at any of the sites available. To the left of the gels is a cartoon of the substrate and reaction products shown in the gel. The intensities of each substrate and product band in the gels were quantitated using ImageJ [86] software and calculated as rough percentages to be shown graphically. The key in the middle shows supercoiled DNA substrate (purple circles, SC), linearized DNA (blue square, Lin), and final product linearized DNA (orange triangle, frag1-4). Two molecular weight ladders were used in this experiment; (i) Lambda-HindIII/PhiX174_HaeIII which spans 0.2 kb to 23.1 kb, and (ii) Lambda-BstEII/pBR322-MspII which spans 0.3 kb to 8.5 kb. This experiment was performed once. **Panel b.** The cleavage of an alternative four-site substrate with and without an activator added in trans. Reactions were conducted as in Panel a. This experiment was performed once.

To further investigate the sequence of cleavage events, the four-site substrate was pre-incubated at a 1:1 enzyme TRD to DNA target site ratio in the absence of Mg2+ ions to allow binding, preferably with one tetramer binding all four sites in one plasmid. Cleavage was then initiated by adding Mg2+ and time points taken (Figure 9a). Cleavage occurred more quickly following pre-binding the enzyme but produced the same fragments resulting from single, dual, triple, and complete cleavage events, indicating sequential rather than coordinated cleavage events even when all four target sites are bound in the same enzyme tetramer. The substrate is cut nearly to completion, though a small amount of partially cut substrate remains. Increasing the ratio of enzyme TRDs to sites above 1:1 resulted in an increase in the partial cut fragments (less complete cutting), with the higher 4:1 enzyme-TRD-to-sites condition having more partial cutting than 2:1 (**Figure 2.28**).

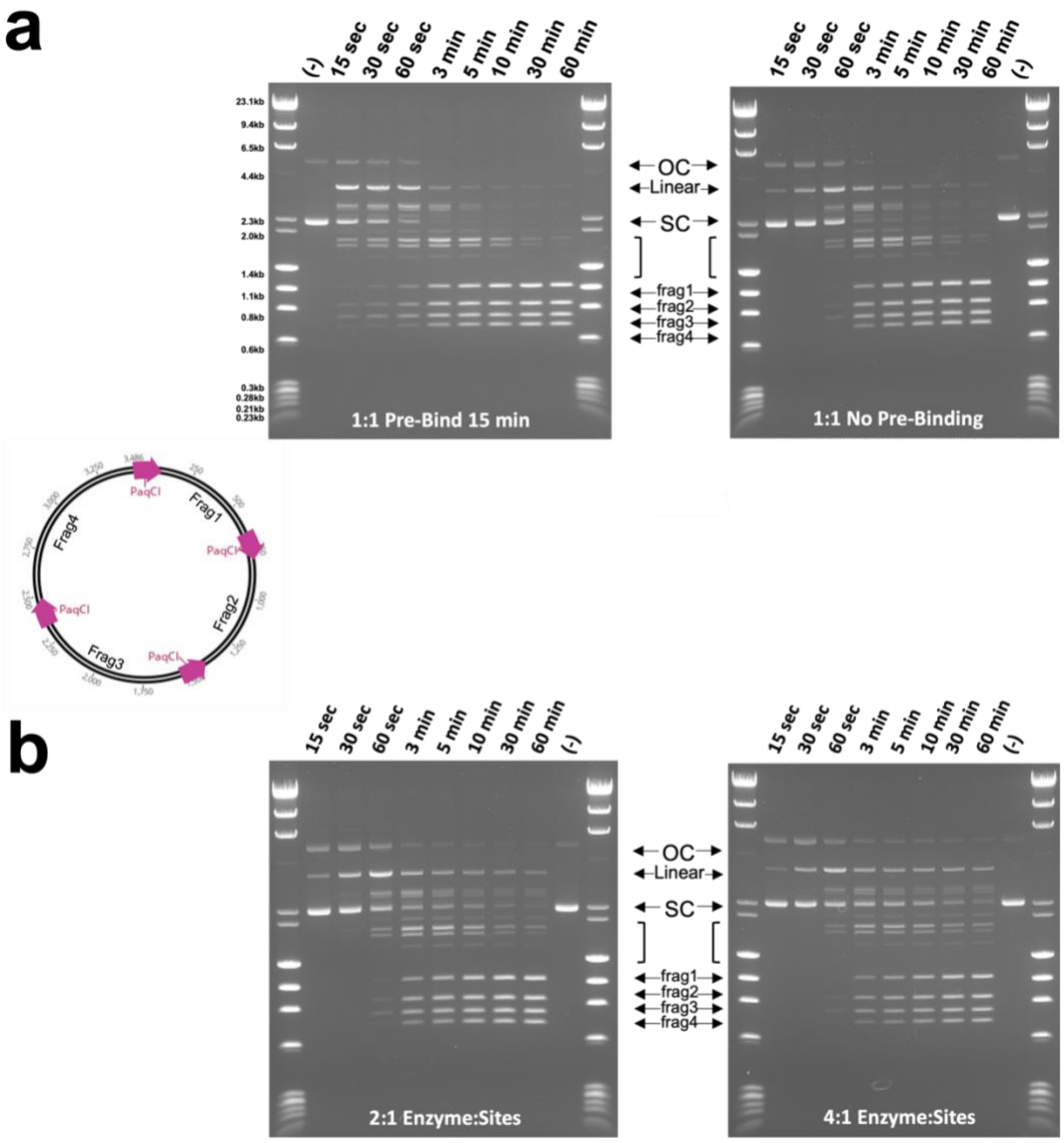

Figure 2.28. DNA pre-bound at all four sites in the PaqCI tetramer is cut sequentially.

Panel a. Cleavage of a four-site plasmid substrate either pre-bound (left) or not pre-bound (right) with a 1:1 enzyme-binding-site to DNA-sites ratio. The pre-bound reaction (left gel) contained 37 nM DNA target sites and 37 nM of PaqCI per 37 nM DNA target sites in 50 µl buffer lacking Mg^{2+} ions. The enzyme was allowed to bind for 15 minutes at 37 °C. An aliquot was removed (time 0) and the cleavage reaction was initiated by adding Mg^{2+} to 10 mM. The no pre-binding reaction (right gel) had the same substrate and PaqCI enzyme concentrations in normal buffer, with time points started upon enzyme addition. Digest time points were quenched at 0.25 minute (min), 0.5 min, 1 min, 3 min, 5 min, 10 min, 30 min, and 60 minutes. The mobilities of the super coil (SC), single-cut linear (Linear), intermediates (brackets), and the completely-cleaved linear (frag1, frag2, frag3, and frag4) forms of the plasmid are indicated in the middle key with arrows corresponding to their related bands. Intermediate bands correspond to a less than fully cleaved plasmid; single, dual, triple, and complete cleavage events. To the left of the gels is a cartoon of the substrate and reaction products shown in the gel. Molecular weight ladders were Lambda-HindIII/PhiX174_HaeIII which spans 0.2 kb to 23.1 kb. These experiments were performed once.

Panel b. Same reaction conditions and plasmid, as in Panel a right side, except that the PaqCI enzyme concentration was increased to 2:1 (74 nM enzyme to 37 nM DNA sites) or 4:1 (148 nM enzyme to 37 nM DNA sites) ratio of enzyme to DNA sites. These experiments were performed once.

2.5 MECHANISTIC INTERPRETATIONS

Size Exclusion Chromatography (SEC) and CryoEM analyses independently indicate that DNA-free PaqCI exists as a preformed tetramer, which is tasked with scanning incoming foreign DNA for the enzyme's seven basepair asymmetric target site. Like most DNA binding proteins, including REases, the enzyme multimer is assumed to find its target via a search mechanism involving subcellular localization near the host and/or foreign DNA and corresponding rapid association and dissociation along the DNA helix combined with limited episodes of one-dimensional (1D) diffusion ('sliding') [15]. However, in contrast to 'simpler' REases that have been proposed to conduct their target searches as monomers (and then to self-associate upon encountering individual target site, thereby forming a higher-order DNA-bound assemblage), PaqCI appears likely to sequentially load individually encountered target sites into the preformed enzyme multimer. How the kinetic and dynamic details of such target search mechanisms differ from one another might be an interesting future topic for detailed examination.

The PaqCI-DNA complex is a tetramer composed of a ‘dimer-of-dimers’, with each pair of subunits on opposing sides of the assemblage independently placing their pairs of bound DNA targets and corresponding cleavage sites into a parallel arrangement. The dimer-of-dimers, in the DNA-bound complex trapped in a pre-cleavage state is comprised of a "cleaving dimer" (with its two EN domains associated with a single cleavage site in *cis* and *trans*) and a "non-cleaving dimer" (with its two EN domains apparently disordered). This result implies that the DNA-bound complex might be sterically unable to simultaneously form two cleavage-competent reaction complexes, perhaps due to limitations in the extension and 'reach' of the *trans*-acting EN domain. The CryoEM structure of the DNA-bound complex indicates that the engagement of a cleavage site on one side of the complex by two EN domains appears to generate an asymmetry in the enzyme-DNA complex that might prevent the corresponding engagement of EN domains around an opposing DNA cleavage site. Correlated motion analyses of the DNA-free enzyme tetramer appear to agree with this speculation: in that analysis, the enzyme seems to display a dynamic equilibrium between two slightly asymmetric conformations that alternately tilt towards one or the other of the two enzyme dimers in the enzyme tetramer.

This observation that only one cleavage competent dimer of two EN domains positioned at one cleavage site can be formed at a time leads to the corresponding prediction that the enzyme may follow a

sequential cleavage mechanism (illustrated schematically in **Figure 2.29**), wherein each DNA within the complex of four bound target sites is cleaved in turn (although presumably in random order relative to the first cleavage event). *In vitro* kinetic studies performed on a dual-site plasmid (**Figure 2.26**) agree with the hypothesis generated from the structural studies. In both the head-to-tail and head-to-head experiments, the supercoiled (SC) plasmid is swiftly cleaved once to form the full-length linear (Lin) form. The Lin form appears at a faster rate than the large and small linear (Lin1 and Lin2) forms, implying that one site is cleaved prior to cleavage of a second site. When testing the enzyme against two different four-site plasmids (**Figure 2.27** and **Figure 2.28**) a single-cut full length linear plasmid band is generated first as the uncut supercoiled plasmid is cleaved, followed by all the linear bands which represent any combination of less than complete four-site cleavage which appear at the same relative rate suggesting that PaqCI does not prefer a particular site to cleave first and that it cleaves only one site at a time. Even with the availability of four sites in each substrate plasmid and pre-incubation with no Mg^{2+} to allow a tetramer to bind a site to each of its four TRDs, cleavage produces the same fragment pattern, without any nicked species, that indicates sequential cutting of one site at a time.

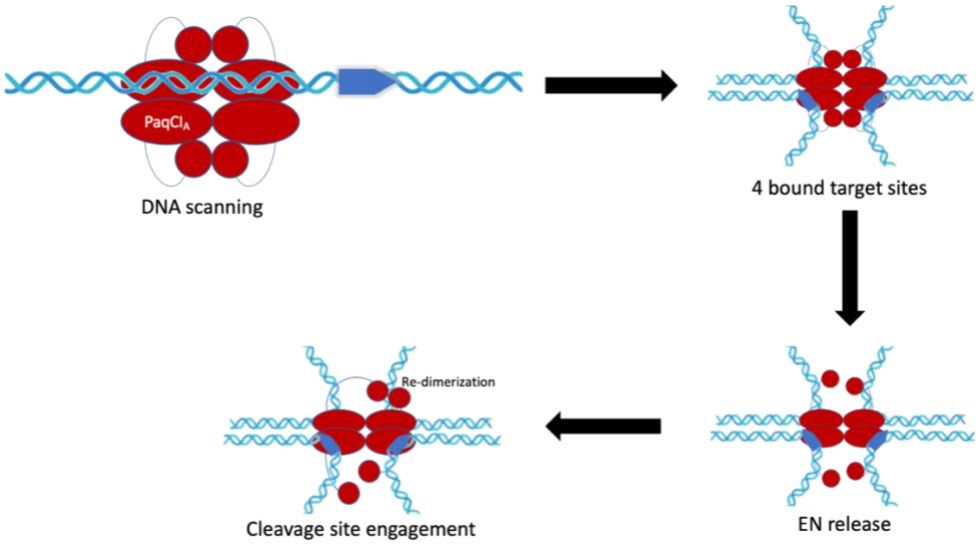

Figure 2.29. Schematic of PaqCI's mechanism as suggested by kinetic and structural data.

The PaqCI tetramer is shown in the first panel in red and a double-stranded DNA in blue with a single target site. The point of the target site reflects the direction of the asymmetric target sequence. Each monomer (i.e., PaqCI$_A$) of the oligomer is represented by an oval (TRD), a circle (EN domain), and a thin line connecting them (linker). PaqCI operates by scanning the DNA for four identical asymmetric target sites that are seven basepairs long. Once it finds four identical target sites, it brings them together in a synapse using the TRDs. The binding of the four target sites triggers the release of all the EN domains from the TRDs. As the EN domains are free in solution, two find the cleavage site on one double-stranded DNA and engage that site for cleavage.

The structure explains the very partial cutting observed on a single-site DNA substrate, where binding to a single DNA site would not activate two opposing endonuclease domains to enable cutting. At minimum an enzyme tetramer would need to bind two single-site DNAs, and the second DNA would have to be bound in the correct opposing enzyme monomer (a one in three chance), to activate two endonuclease domains for cleavage. Complete cutting of a single site substrate is observed in the presence of excess in trans recognition site oligo because when the enzyme binds the single site substrate DNA the other three binding sites in the tetramer can be occupied by the in trans activator oligo to license the release of the endonuclease domain opposite the substrate-bound monomer to form the required catalytic dimer to cut the single site substrate.

The observation that roughly 1:1 enzyme binding site to DNA target site is required for complete cleavage suggests there is little enzyme turnover. The DNA binding site remains intact after cleavage, so if the enzyme has relatively long-lived DNA binding persistence this could explain a lack of turnover. The observation that the small in trans oligo optimally stimulates cutting at roughly a 5:1 ratio of oligo to enzyme binding sites suggests the small oligo may bind much less tightly and have higher turnover than normal long DNA molecules.

The same stereochemical constraints that appear to prevent simultaneous engagement of multiple cleavage sites in the enzyme-DNA assemblage might also dictate the cleavage position of PaqCI (which cleaves the top and bottom strands 4 and 8 bases downstream from its bound target site, respectively). DNA cleavage clearly requires the re-dimerization of the EN domains around the cleavage site, which in turn dictates the four basepair separation between individual strand nicks. Modeling the possible engagement of the EN dimer either closer to, or farther away from the bound target site indicates that (i) the *cis*-acting monomer cannot easily engage DNA closer to the bound target site without imposing a significant clash with its own TRD, and (ii) the *trans*-acting monomer appears to be extended to near the limit of its potential reach across the enzyme tetramer. The combination of these two stereochemical constraints would thereby limit the association of the two EN domains to the precise phosphate groups corresponding to the enzyme's cleavage pattern.

2.6 PAQCI AS A MEMBER OF THE TYPE II ENZYME FAMILY

2.6.1 *Comparisons with BspMI and DrdV enhance understanding of the PaqCI structures*

The closely related isoschizomer BspMI (which shares 39% amino acid identity to PaqCI) behaves quite similarly: it exists as a preformed tetramer before binding its target DNA, maintains that same quaternary assemblage after binding DNA, and cleaves each bound DNA duplex in both strands without intermediates cut at a single strand of the DNA duplex [62, 66, 88]. Kinetics of BspMI also demonstrate an increase in the enzyme's cleavage efficiency when a DNA duplex containing the enzyme's recognition site is added as a *trans*-activator [62, 66]. Additionally, a single turn-over event requires two independent double-stranded cleavage events, suggesting that within a tetrameric complex harboring a dimer-of-dimer architecture, one enzyme dimer at a time cleaves its bound DNA duplexes while the other waits for its turn, just as seen with PaqCI. This could mean that PaqCI also only requires one dimer of the dimer-of-dimers to be bound at one time to cleave the DNA duplex, and the structure shown here with all four sites bound is a product of DNA saturation in the experimental preparation. However, in kinetic experiments, BspMI was reported to cleave both DNA duplex sites rapidly, with only a small accumulation of a single-cut linear species prior to cleavage at the second site [62, 66]. Acc361, another isoschizomer, produces an accumulation of a single linear species, similar to PaqCI, cleaving the dual site plasmid with two kinetically separate cleavage events, furthering the knowledge that even amongst Type IIS isoschizomers there is a high degree of variation [34, 45, 62, 66]. Additionally, BspMI, as well as FokI, exhibit some preference to the orientation of the DNA sites (head-to-head or head-to-tail). Our experiments with PaqCI suggest that there may be a slight preference for head to tail sites at the 700bp separation tested, but more kinetic analyses are necessary to definitively address this point.

The recently visualized Type IIG enzyme DrdV displays similarity to PaqCI but with significant differences in many of the underlying details of its action. DrdV also contains an N-terminal EN domain and a C-terminal TRD, recognizes and binds an asymmetric site (5’ CATGNAC 3’), cleaves downstream of its target, does so as a tetrameric enzyme assemblage engaged with multiple bound DNA targets, and relies on the combined action of *cis*- and *trans*-acting enzyme domains to generate individual double-strand breaks. However, unlike PaqCI (or FokI, described below) DrdV generates a two-base 3' overhang (cleaving

top and bottom strands 10 and 8 bases downstream of the target, respectively). Also, unlike PaqCI and FokI, it incorporates its cognate methyltransferase domain and activity into the same protein chain (located between the EN and TRD domains) and then establishes a kinetic competition between slow methylation activity at individual bound target sites relative to rapid cleavage of a multi-target enzyme assemblage, to bias the two competing outcomes towards host or foreign DNA. Finally, DrdV exists as a monomer in solution prior to DNA target binding and couples the formation of a cleavage-competent tetrameric assembly to DNA target acquisition and binding. Unlike both FokI and PaqCI, three separate DrdV subunits are engaged on a single bound DNA duplex and required for cleavage: a target-bound TRD domain from one subunit, plus two separate *trans*-acting EN domains that are assembled at the cleavage site downstream of that target. Despite those variations, DrdV still fundamentally requires *cis/trans* collaboration to enforce the requirement for multiple bound target sites for cleavage.

2.6.2 *PaqCI in the context of FokI*

FokI is the most well-studied Type IIS REase to date [49, 51-53]. While also acting as a canonical Type IIS REase, it displays significant differences as compared to PaqCI (**Figure 2.30**). Its domain organization is reversed (C-terminal EN domain for FokI, versus an N-terminal EN domain for FokI); it exists in solution primarily as a monomer in the absence of bound DNA, it recognizes a shorter five-base asymmetric target site, and it cleaves top and bottom strands farther downstream (9 and 13 bases, respectively) from its bound target site than does PaqCI (which cleaves the top and bottom strands only 4 and 8 bases away from the bound target). Whereas PaqCI accomplishes this within the context of a larger, preformed tetrameric architecture, FokI appears to instead generate a similar cleavage synapse via the transient formation of a DNA-bound dimer. Prior studies of FokI, therefore, provide an important point of comparison to understand the Type IIS REase family.

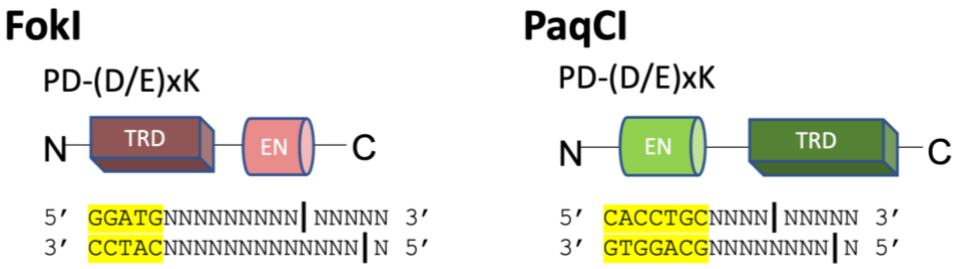

Figure 2.30. Comparison of FokI and PaqCI structural organization and activity.
Domain organization, target sites and cleavage patterns of FokI (left) and PaqCI (right).

While these studies of FokI have not generated an atomic-resolution structure of a cleavage-competent enzyme assemblage, a well-validated model (**Figure 2.31**; left) of its cleavage complex, corresponding to a transient DNA-bound enzyme dimer, has been proposed based on a variety of biophysical and structural analyses [53-59]. This model indicates that two independent enzyme monomers are bound to two separate DNA target sites via their individual TRDs. Their corresponding EN domains are wrapped around one of the DNA duplexes (thereby acting jointly in *cis* and *trans*, as observed in the structure of PaqCI). However, in that model, the FokI TRDs are not engaged in a higher-order complex with one another, and the only contact between the protein subunits involved the DNA-associated EN domains.

While the general features of how they cleave DNA (by placing *cis*-acting and *trans*-acting endonuclease domains on one bound DNA duplex at a time) are similar, the underlying details of how they sequester their EN domains in the absence of foreign DNA and the structural rearrangements that each dimer must undergo to enter similar cleavage-competent DNA-bound complexes appear to differ significantly. FokI and PaqCI share a similar EN domain dimer organization and place them in similar orientations around their cleavage sites, such that both enzymes generate four-base 5' overhangs albeit with quite different spacing between their bound targets and their respective cleavage sites (**Figure 2.31**). In both cases, the EN dimers in the DNA-free and DNA-bound states share strong structural similarity (corresponding to ~0.2 Å backbone RMSD between pre- and post-DNA binding states for PaqCI; **Figure 2.22**).

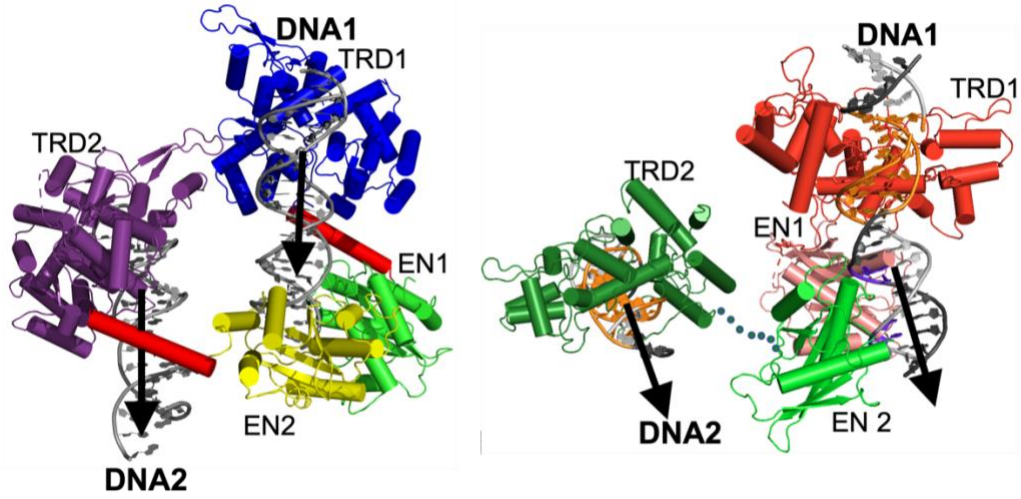

Figure 2.31. Comparison of FokI and PaqCI protein/DNA complexes.

Comparison of DNA-bound dimers visualized via single particle electron microscopy for FokI (left; figure panel figure panel generated from [59] and from the model provided by the authors) and PaqCI (right; this study). The latter enzyme-DNA complex is extracted from the larger DNA-bound PaqCI tetramer and corresponds to the dimer in that assemblage that displays visible DNA-engaged endonuclease domains.

Despite sharing a similar architecture for dimerization of the EN domains they utilize different residues on the two main α-helices involved in dimerization of the EN domain to do so. The first of the two α-helices (known as α4 in FokI [51]) makes an 'X' with its dimerizing monomer. In this interface, FokI uses an arginine and an aspartic acid to facilitate dimerization. In PaqCI, the α4 dimerization interface appears to be facilitated by a network of interactions between two histidines, a threonine and a glutamine. The second α-helix (known as α5 in FokI [51]) has minimal contact with its dimer partner in FokI's structure where as additional contacts are created in the dimer of PaqCI. The contacts made by the α5 helices of PaqCI are facilitated by an isoleucine contacting a histidine from α4. When comparing the α4 of FokI to PaqCI, there is little sequence homology except for a string of 'IGQA' and a shared tyrosine. When comparing the α5 of FokI to PaqCI, the only commonality is a tyrosine and a string of polar amino acids. In both structures, the α-helices of the EN domains that dimerize point into the major groove of the DNA and are perpendicular to it. This EN domain dimerization motif is common amongst other Type II enzymes, such as BamHI, and is notable considering the significant differences in topology between these proteins (**Figure 2.32**) [51].

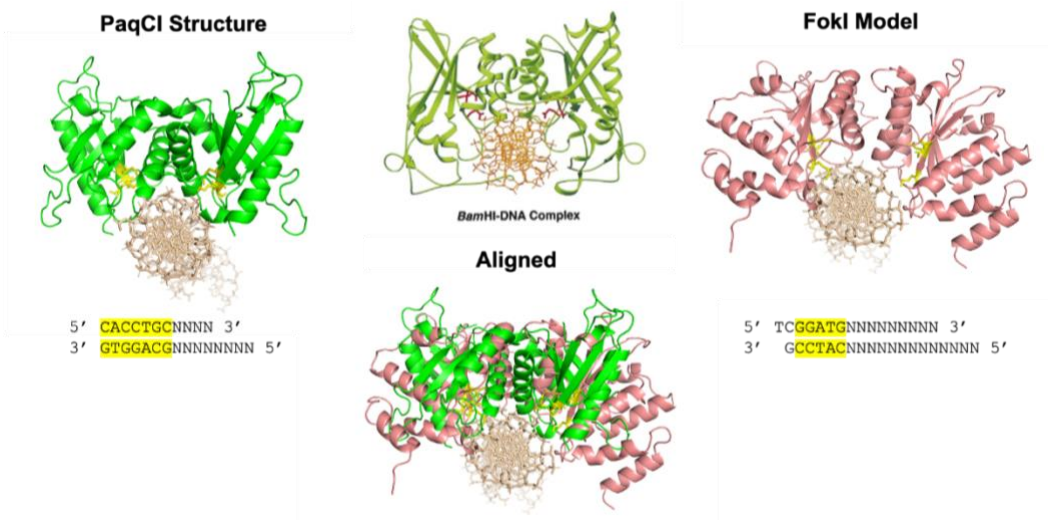

Figure 2.32. EN domains of PaqCI and FokI dimerize over DNA in similar ways.

Both PaqCI and FokI dimerize using similar α-helices which point into the major groove of the DNA duplex and are roughly perpendicular to it. These enzymes mirror their Type II family member BamHI in this dimerization conformation.

To dimerize over the cleavage site, PaqCI and FokI each rely on a flexible 16-residue linker between the EN and TRD domains. This peptide region allows the *trans*-acting PaqCI EN domain to extend almost 40 Å across the tetramer interface to re-engage with its *cis*-acting EN partner at the DNA cleavage site. In contrast, FokI does not appear to require the full length of its linker to contact its cleavage site, only needing to extend a predicted distance of about 30 Å [59]. Furthermore, both appear to rely on mechanisms in which an enzyme dimer comprised of two DNA-bound subunits orients two DNA duplexes in a parallel arrangement, leading to transient dimerization of two EN domains in *cis*- and *trans*- on a single bound DNA cleavage site. However, whereas PaqCI accomplishes this enzyme-DNA arrangement and EN dimerization in the cleavage complex within the context of it larger, preformed tetrameric architecture, FokI appears to instead generate a similar cleavage synapse via the transient formation of a smaller DNA-bound dimer.

FokI is prevented from cleaving non-specific DNA strands by sequestering its C-terminal EN domain containing a single PD-(D/E)xK catalytic motif loosely against the N-terminal TRD [51, 52]. The TRDs of PaqCI and FokI have a rmsd of about 24 Å. The TRD of FokI has an extra three α-helices compared to PaqCI that are used to dock the EN domain when FokI is not bound to a target site. These three additional α-helices in the TRD barely contact the DNA, implying that their primary role is to exclude

the EN domain from the DNA [51, 52]. In contrast, PaqCI conserves the dimerization of the EN domains and sequesters each EN dimer at the center of the quaternary TRD ring and away from solution.

In both cases, the greatest structural difference post-DNA binding is seen in the linkers between domains and the closure of the TRD around the DNA. PaqCI utilizes four loops (176-186, 282-296, 323-337, 457-469) and one helix (338-354) for recognition of the target site. Some residues are directly involved in the recognition while others are near the site and may be facilitating additional recognition contacts through secondary methods. The loop 323-337 was predicted to be the primary mode of recognition used by PaqCI based on consensus sequences and mapping of lysine residues (using N-hydroxysuccinimido (NHS)-Biotin and mass spectrometry). Upon binding DNA, glutamine residues of the loop are inserting themselves within the DNA helix whereas the predicted lysines (K334 and K458) are either unstructured or not close enough to the DNA to make direct contact. Binding to the target site by FokI is facilitated by two particular loops that grip the DNA duplex in both the major and minor grooves, causing little change in the bend of the DNA helix [52]. FokI also places the ends of two helices and a loop in the major groove of the DNA but based on predicted direct contacts employs different amino acids to accomplish the job.

In the FokI dimer, the TRDs face each other (DNA binding 'C' portion of the TRDs facing inward) instead of being back-to-back, as observed in the PaqCI tetramer. When bound, the double-stranded DNA substrates in FokI are, at their closest, ~8 Å apart from each other, whereas PaqCI's substrate is ~50 Å apart (**Figure 2.31**). This difference potentially changes the orientation of the EN, TRD, and double-stranded DNA molecules in the cleavage dimer and may contribute to the distinctive cleavage distances and the variation in search mechanisms.

There appear to be few or no residues that establish base-specific contacts with the DNA in the EN domains in either PaqCI or FokI. In PaqCI, there are a few contacts between the EN domain and DNA, outside of the catalytic residues (D54, E73, K75), that may be playing a role in DNA recognition (Y3, Q78, E79, H82) to ensure that the proper contacts are made with the DNA backbone before cleavage. Therefore, these proteins likely rely on stereochemical constraints to define their cleavage sites and patterns relative to the bound target sites. Since the TRD is essential for DNA recognition and binding, we can hypothesize that the cleavage location is controlled by (i) the requirement for the EN dimer to re-associate around the cleavage site, coupled with (ii) the necessity that the *cis*-acting EN domain avoids steric clash with its own

TRD and the *trans*-acting domain not be required to move beyond the reach of its linker peptide [49, 52, 53]. This hypothesis is satisfactorily supported by an examination of the DNA-bound structure of PaqCI (for which it does not appear that the EN dimer can be positioned in an alternative location other than over the scissile phosphates). In contrast, it is more difficult to confidently assign a structural mechanism that would limit FokI to cleavage at a strongly preferred distance of 9 and 13 bases from its bound target site. Potentially, when the EN domain is released, the linker region becomes structured and might thereby act as a 'measuring rod', allowing the enzyme to cleave at 9 and 13 each time it binds its target site [53, 59].

The mechanistic features of FokI and PaqCI described above are similar in terms of the topology of DNA bound enzyme-dimers and the relative orientation of the bound DNA duplexes and disposition of their *cis*- and *trans*-acting EN domains, but differ in terms of the quaternary assemblage of each enzyme and the exact stereochemical constraints on formation of the ultimate cleavage complex. They each offer multiple point of control over cleavage events to prevent off-target events: first, the release of the EN domain is dependent on target specific DNA binding, and second proper orientation of the two catalytic domains over the correct phosphodiester bonds requires dimerization [49].

For FokI, neither cleavage event (9 or 13) has a noticeable lag time, meaning that the two strands must be cleaved simultaneously rather than sequentially, and this can happen in either order, even though there is a *cis* and *trans* monomer, either can cut first [54, 55, 59].

In other known Type IIS enzymes, an enzyme monomer, with both a TRD and EN domain, can be forced to nick DNA in its monomeric state, but it usually cleaves the top strand, whereas FokI *cis* nicks the bottom strand when forced [54, 55, 59]. PaqCI appears to exist as a tetramer whether bound to DNA substrate or not. When activated, the cis monomer is poised to cleave the top strand. It is not yet known if PaqCI can be forced into nicking the DNA strand and whether it will prefer the top strand. This preference seen in Type IIS enzymes may be primarily due to sterics and helical rotation. Further studies are necessary to determine the operation of PaqCI in this capacity.

Many type II enzymes need to interact with at least two sites before they can cleave DNA. These enzymes, therefore, trap loops in the DNA and have a higher affinity for sites on the same strand over sites on adjacent strands due to the location of the same strand target site generally being at a closer physical distance [54]. In FokI, the bound monomer and free monomer only associate weekly, so the dimer is short-

lived, adding an energetic preference to two bound monomers forming the complex [49, 53]. When FokI binds to its target site and dimerizes, the stability of that complex hinges on the position of the two target sites and whether the cleft of the helix is facing the same direction [49, 53]. For the two EN domains to come together and bind, they must be on the same side of the DNA, which may be difficult depending on the location of the target sites within the DNA. Due to this, the two monomers must bind the substrate simultaneously and trap the DNA in an advantageous loop [49, 53, 54]. FokI engages and cleaves DNA with two target sites on the same strand of DNA faster than sites on different strands [49, 53, 54].. Thus, if the two sites are on a single plasmid, a twist of the DNA would be required to form a functional loop, and that twist would favor the parallel orientation of the DNA [57]. FokI's organization of target sites in parallel is not dependent on whether these sites are in head-to-head, tail-to-tail, or head-to-tail orientation [57, 58]. Parallel orientation is preferred when the sites are on the same DNA strand, which is governed by other rules, such as tension and torsion, whereas when the two sites are on separate strands, the only preference is towards which arrangement that will provide the lowest free energy [57, 58]. The 5 Å CryoEM model bolstered the hypothesis that FokI binds to DNA in a parallel manner (**Figure 2.31**) [59]. Although the FokI dimer conformation with target sites on separate strands has not been solved structurally, it has been shown kinetically [57]. This brings up questions about supercoiling the tension and angles naturally found in DNA [59]. The binding of FokI to the DNA in parallel is due to the energetics of the loop formation on the same strand (which is easier to locate).

Within the cell, the DNA concentration can vary *in vivo* (supercoiling) contributing to inaccurate predictions for proper concentration of DNA in *in vitro* kinetic experiments. The resulting crowding, condensation, and enhancement increases the multimerization, tangling, and complex formation of DNA [89]. In the cell, it is estimated that 40% of the space is crowded by molecules other than water, increasing the activity of DNA [89]. The average DNA concentration within the nucleoid of *E. coli* is an estimated 30-100 mg/ml and RNA around 100 mg/ml while proteins are generally crowding the cell at 250 mg/ml, all of which can affect the behavior of the DNA and the searching RM systems [89]. Interactions with surrounding proteins and RNA along with sequence specific looping changes the sequence specific local DNA concentration. All this information begins to explain why the *in vivo* kinetics vary from the *in vitro* kinetics, creating more questions for the field.

PaqCI and FokI both appear to generate a cleavage synapse with DNA in parallel arrangement [57-59]. If this is true, the directionality of the DNA seen in the PaqCI structure may be different if a plasmid DNA was used to solve the structure instead of strands. This necessary looping is an interesting conundrum for PaqCI since it acts as a tetramer during target site binding and cleavage. Perhaps this is another mode of regulation to bias cleavage towards foreign DNA. In either case, the unwinding of the DNA post-cleavage likely upsets the binding of the EN domains and signals the enzyme to disengage and reengage a new cleavage site. Further experiments using atomic force spectroscopy and magnetic tweezers would be necessary to determine the specific DNA looping mechanism of PaqCI.

Chapter 3. HIGHLIGHTS OF CRYOEM CONTRIBUTIONS TO THE STUDY OF NUCLEIC ACID ENZYMES

The impact of CryoEM on structural and mechanistic knowledge of complex enzymatic systems is ever growing. Last year alone over four thousand CryoEM structures were deposited to the PDB, more than any year before, representing about 30% of all structures deposited that year [90]. This new strategy has been helpful for determining the structure of enzymes that work with or upon DNA and RNA substrates and reactants, many of which were elusive to past methods. There are now 1,282 deposited single particle EM structures of protein bound to DNA in the PDB as of spring 2023 (search qualifications: >=1DNA, >=1Protein, EM single particle), changing the landscape of structural knowledge.

In general, the reason why many others were never able to crystallize active complexes bound to DNA/RNA is because they display a combination of conformational heterogeneity, sampling of asymmetric spaces or flexibility, and dynamic exchange between multiple assembly states (sometimes even with other proteins in addition to the DNA/RNA). In contrast, these DNA/RNA systems proved to be ideal for CryoEM, with multiple intermediate reaction states being visible and solvable from a single well-performed data collection with adequate numbers of micrographs and individual particles and movies to structurally determine all the states.

CryoEM has been transformative in terms of answering questions and providing mechanistic insights that previously would never be possible. In this short chapter (which I intend to eventually expand and publish as a review article), I provide anecdotal summaries of two nucleic acid enzyme systems and questions for which recent CryoEM studies have proven equally changing in terms of answering long-standing questions or controversies about mechanism or biology.

Type IIS R(M) systems. A set of structures have recently been determined by the Stoddard lab using CryoEM of multi-domain and/or multifunctional restriction/modification enzymes that were unable to be crystallized or solved before the invention of CryoEM. These enzymes act through a mechanism in which they must pull together multiple DNA target sites into a precise reaction "synapse" to license DNA cleavage activity. The first enzyme, DrdV, contains both a methyltransferase and an endonuclease domain within a single subunit, and therefore sets up a kinetic competition between those two activities as it accurately

modifies host DNA and cleaves foreign DNA as covered in an earlier chapter [38]. The second enzyme, PaqCI, is an endonuclease only, but also pulls requires engagement of multiple target sites and assembly of a multimeric protein-DNA complex before cleavage can occur [1]. Both posed challenges for traditional crystallographic methods and remained unsolvable until CryoEM. For DrdV, there was a challenge with it not only being a large complex, but the complex existed in multiple oligomeric states at one moment in time. Whereas for PaqCI, it binds four DNA substrates in an asymmetric manner and seems to require longer DNA tails.

Type I RM systems. In addition to DrdV and PaqCI, another restriction modification (RM) system has recently been solved using CryoEM; EcoR124I [67]. EcoR124I is a Type I RM system found to exist in multiple different assemblies of methylase, restriction endonuclease, and target recognition domain [67]. Type I RM systems generally recognize DNA target sequences with two specific regions separated by a nonspecific DNA spacer [5, 67]. When in contact with an unmethylated target sequence, an ATP-dependent translocation across the DNA is triggered while the enzyme remains bound to the original target site, creating a loop. Despite extensive study, the mechanism behind the regulation and inhibition of Type I RM systems remained unclear until structural studies such as the one addressed here [67] was completed. As well as capturing the complex in multiple assembly states ($R_2M_2S_1$, $R_1M_2S_1$, M_2S_1), Goa et al. solved the structures in complex with Ocr and ArdA, two DNA mimics that inhibit activity but does not prevent conformational transitions [67]. These structures provide key information into not only the assembly of a Type I RM system, but into the unique regulation of the system in bacteria. These sophisticated systems regulate the translocation across the DNA and methylation or cleavage of the target DNA through different structural conformations. Without CryoEM, the structure and function of these enzymes would not be fully understood.

Retroviral Intasomes. A structural view of retroviral intasomes was previously impossible with structural methods until the advent of CryoEM. In general, when a retrovirus like HIV invades a host, viral-encoded integrase enzymes (which enter the cell as part of the viral particle) integrate retroviral reverse transcripts into the host cell genome [91]. Once host DNA is made, an integrase hydrolyzes the 3’ end of the DNA that is adjacent to the phylogenetically conserved CpA’s. The integrase then uses the created CA-3’-OHs to cleave the host phosphodiesters, which inevitably joins the viral 3’ end to the cleaved 5’

phosphates via SN2 chemistry. The unattached viral 5' end that remains after the insertion gets inserted by the gap repair machineries of the cell. Retroviral intasomes are the series of nucleoprotein complexes (including integrases) that perform the hydrolysis and strand transfer reactions of the DNA incorporation steps.

Already employed in the clinic are integrase strand transfer inhibitors (INSTIs) which inhibits the second of the two activities described above [92]. INSTIs have been pivotal in the treatment of retroviruses and are now used worldwide to treat new infections and assist in other drug regiments. Sadly, as these INSTIs are used more often in the clinic, there is a growth of drug resistant retroviruses that evade the function of the INSTIs, leading to the need of new classes/types of antiviral drugs that target a different part of the intasome system.

The first structure of an intasome was of the cleaved synaptic complex of the prototype foamy virus (PFV) solved by x-ray crystallography [93]. This intasome is the second of four intasome complexes in the PFV integration pathway. This structure was successful because the PFV integrase is very well behaved in solution and the tetrameric intasome formed upon addition of the DNA resisted challenges with high salt, which otherwise would have disrupted non-productive integrase-DNA complexes. Not long after, the structure of the remaining PFV intasome complexes were solved [94, 95]. With diffracting crystals, scientists were able to expose the ground state intasome to divalent metals, therefor characterizing the integrase 3' processing and strand transfer reactions [95].

Despite these accomplishments, what worked for PFV integrases has failed for all other retroviral integrases attempted. This is due, in large part, to the fact that active intasome complexes can be built from four, eight, or 16 copies of the integrase enzyme [96]. As discussed above, as complexes increase in size, the heterogeneity also increases, obliterating the chances to grow well-diffracting crystals. Whereas single particle CryoEM allows for the characterization of multiple assembly states and can document the dynamic exchange between these states. Already EM research has reflected the large-scale heterogeneity exhibited throughout retroviral intasomes in studies regarding SIVrcm (the epidemiological forebearer of HIV-1 integrase) and HIV-1 [97, 98]. Additionally, the EM images of SIVrcm revealed a 'stacking' of intasomes subunits resulting in a long string of intasomes, which was a structural feature previously unknown but essential for solving high-resolution structures of not only the wild-type SIVrcm intasome but the INSTI

resistant intasome [97]. CryoEM has also proved sufficient for solving the structure of the HIV-1 integrase active site and intasome complex [98, 99]. With greater structural understanding of these high-level complexes, scientists can better design drugs to target this system to counteract retroviral infections.

PART 2: ENGINEERING AND FUNCTIONALIZATION OF CIRCULAR TANDEM REPEAT PROTEINS

Chapter 4. DESIGNED TANDEM REPEAT PROTEINS

4.1 PROTEIN ENGINEERING AND DE NOVO DESIGNED PROTEINS

The creation of novel proteins with unique folds, assemblies, and mechanisms has long been a focus of protein engineering [100] but investigators have often been confounded in their goal to design proteins accurately, consistently, and efficiently due to the barriers of existing technology. With the modern advancement of computational algorithms and new techniques for screening large libraries of protein designs, scientists are now able to construct and screen multitudes of protein designs at an unprecedented rate. Protein engineering strategies can enhance a protein's kinetic properties, substrate recognition, and stability. Already, engineered proteins have been a boon to the scientific community and are being used for a multitude of applications within biotechnology and clinical settings [101, 102].

Current state of the art protein engineering methods include rational design (also termed 'structure-guided' or 'structure-based' design), directed evolution (which is also largely guided by structural information), and *de novo* design [100] each harnessing computational and laboratory strategies to generate new proteins. Structure-based design of proteins is one of the most widely used protein engineering strategy and primarily involves site-directed mutagenesis based on structural knowledge of the given protein [100]. Locations for mutations are identified based on the protein's native structure (solved by x-ray crystallography or CryoEM, primarily) and mechanism which reduces the number of protein variants that need to be expressed, purified, and screened, enhancing the probability of a beneficial mutation. If there is not sufficient structural or mechanistic knowledge of the protein, scientists can use directed evolution to generate the new protein variants. Directed evolution is a cyclical process involving random mutagenesis (instead of specific point mutations) and selection until the desired protein traits are achieved [100]. Unfortunately, this generates a much larger library of variants to screen than for structure-based design. To overcome this screening problem, automatic and high-throughput strategies have been developed, though the process is still time intensive. Many labs have been successful in employing both

structure-based and directed evolution approaches in protein engineering, emphasizing the benefit of combinatorial approaches to optimizing protein structure and function.

On the other hand, *de novo* design strictly relies on computational methods to develop synthetic proteins from scratch without any native structural data, using only knowledge of inputted sequences and structures of naturally existing proteins. These *de novo* designed proteins are then further optimized with structure-based and random strategies. Additionally, these computational approaches of *de novo* design can be used to further hone proteins developed from the previous two approaches and narrow down libraries to screen. Protein engineering can be most successful when multiple strategies are applied.

4.2 DESIGNING THE PROTEIN AND LIGAND INTERFACE

Protein engineering has been essential for the development of new ligand binding proteins [103, 104], but it is difficult to engineer highly specific ligand binding proteins that are sensitive to one particular ligand over another. To facilitate proper binding of a ligand to a protein, interactions such as hydrogen bonds, electrostatic attractions, and van der Waals attractions need to be leveraged to ensure proper specificity while maintaining function of the overall protein. This intricate balance is an important feature in biology and nature, as seen in immune recognition, gene expression, cellular metabolism and signaling.

Previous state of the art processes to design specific ligand binders involved immunizing animals to generate specific antibodies against target antigens or conducting lab directed evolution against the ligand and protein interface, all resulting in low affinity binders. Both processes lacked control and specification over the ligand/protein interface. It is tempting to use strictly computational approaches when designing specific ligand binders, but many computational methods have a poor ability to accurately calculate binding affinities at all backbone and sidechain orientations possible on one protein [105-107]. This issue is compounded by the necessity of sampling not only the protein conformation and the protein sequence during the design process, which requires a considerable amount of computing power on its own, but also sample the conformations of the ligand [108].

As an example of these issues, the following series of structural and biochemical studies from the Baker and Stoddard labs will be summarized to illustrate how computational approaches can be used for the design of *de novo* ligand binding proteins. The proteins explored in these papers (i) employ enough

flexibility in the structure to allow binding of the ligand without disassembling the protein, (ii) have similar shape to the ligand to allow for synergy between the pore and the ligand, and (iii) begin with pre-organization that allows the entropic penalties of binding a ligand to be reduced.

Computationally designed specific protein binders against the small molecule steroid known as digoxigenin (DIG) have been created and validated by co-crystalizing structures (which confirmed that the structure matched the computational design), and shown to be specific for DIG over all other related steroids [104]. To create a binding site suitable for the steroid, amino acids were first identified, and placed surrounding the ligand in empty space, and then placed on to a pre-designed protein scaffold [104]. With that model, the interactions were further optimized to accommodate the ligand. Yeast surface display [109] and flow cytometry was used to screen the binder library and identify levels of affinities [104]. To further evaluate the function of the final design and improve the methodology, next-generation sequencing was used to create a binding fitness map, bringing greater understanding to the amino acid level contribution to binding affinity [104, 110-112]. By harnessing natural factors of protein/ligand binding, the group isolated three key features to accurate ligan binding design as outlined in the paragraph above. These key features have been used by many scientists, including those involved in the work of this thesis, since this paper's [104] inception to engineer unique protein and ligand interactions.

Even with the great progress in designing the interface, the computer can still miss important conformations that the protein can sample when binding the ligand of interest. Such as was seen in Dou et al., 2017 and Day et al., 2018 [113, 114] papers. In both, the desired ligand was bound successfully, but in a confirmation and orientation not predicted by the computational design. In the former, the ligand steroid hormone 17a-hydroxylprogesterone (17-OHP) was used to computationally design a specific protein binding site [114]. Following the methodology of Tinberg et al. [104], the protein had an extended, nonpolar, shape that reflected the structure of 17-OHP with plenty of hydrogen binding residues to accommodate the polar parts of the steroid [114]. Despite the extensive hydrogen binding and unique shape of the protein, a co-crystal structure showed 17-OHP rotated 180° around a pseudo-two-fold axis from its predicted design location [114]. Additionally, the crystal structure reveled multiple binding modes within the designed pocket while still contacting all the designed residues in the engineered pore [114]. Such inaccuracies could extrapolate in other design scenarios, leading to deleterious effects if applied to clinical or biotechnological

settings. Additional rounds of optimization via mutagenesis and binding selection increased the ability of the protein to bind 17-OHP in one conformation. The group was able to determine that the divergence from the design protein came from (i) an inability to sample changes in the protein backbone upon binding which then modifies the distribution of sidechain rotameric states and (ii) an underestimation of the energetic cost of removing the solvent from the different chemical groups of 17-OHP [114].

In the latter paper, off target binding was exhibited despite following the three key design principles [113]. The protein was designed to bind tetrahydrocannabinol (THC) but was found to also bind 25-hydroxy-cholecalciferol (25-D3) [113]. Through crystallography, it was found that the difference in specificity arose from the adjustment of the protein's conformation corresponding to a change in the orientation of an α-helix and a loop which flanked the ligand pore [113]. This also showed that with a few additional rounds of design, a protein capable of binding a ligand with high affinity is achievable [113]. These errors occur due to sampling limitations and inaccuracies in current energy functions in the design programs [113]. To combat these issues, groups are working to improve the solvation models and the sampling methods. Lessons from these two papers point to the unpredictability of protein binding and the inability to sample the entire backbone's conformation.

Once a scaffold is assembled the shape of the cavity can be made to match a ligand, and in the case of Dou et al. 2018, a fluorogenic compound know as (Z)-4-(3,5- difluoro-4-hydroxybenzylidene)-1,2-dimethyl-1H-imidazol-5(4H)-one (DFHBI) [103]. Following the lessons of the previous works, the DFHBI was first placed in the model rigidly, and then the surrounding amino acids were mutated in a hierarchical grid-based search method to achieve better shape and chemical composition to host the ligand [103]. The resulting designs match their experimental counterparts (tested via x-ray crystallography) and were shown to bind with high specificity the DFHBI ligand [103]. The methodology illustrated the success that can be found by custom building the backbone of the protein for the ligand of interest. In the future, these custom ligand binders can be used as sensors and catalysts that operate with high specificity and affinity [103].

4.3 CIRCULAR TANDEM REPEAT PROTEINS (CTRPS)

A class of proteins called tandem repeat proteins ('TRPs') have existed in nature and are made up of modular units of repeating protein sequence resulting typically in repeated structure that form interacting

surfaces [115-118]. The modular components can be made up of α-helical bundles, β-sheets, or mixed topologies of both [115-118]. The core architecture of TRPs is controlled by the internal geometry and local packing of the assembled repeats. The fact that these proteins are made up of engineerable modular units with few constrains to the number of repeated modules that can be designed to assemble makes them perfect for *de novo* computational design [115-118]. The modular nature allows for rapid diversification of binding surfaces via recombination of these optimizable building blocks [125, 126]. Even when engineered, these TRPs retain the same assets as their naturally evolved counterparts. They have relatively few constraints on their length, size, and flexibility, allowing them to sample a wide variety of molecular space. The repeating surfaces can be used to bind a multitude of substrates ranging from peptides to small molecules to DNA to RNA and are associated with scaffolding proteins [119-124].

In nature, these modular repetitions of structures play an important role as macromolecular binding and scaffolding domains, enzymes, and building blocks for fibrous materials, and scientists can use computational design to highjack these properties for similar and new applications [117]. Two main types of TRPs have been computationally explored; either an extended linear shape (with N- and C-termini far from each other) or closed circular shape (with N- and C-termini making contact) [128-131]. The latter being coined as circular tandem repeat proteins (‘cTRPs’).

This thesis will focus on cTRPs. Within the following chapter multiple styles of cTRPs will be shown along with a small reflection of the capabilities of these computationally designed rings. In general, cTRPs are made from repeating two-helix bundles that can tolerate a wide array of sizes and symmetries (**Figure 4.1**, taken directly from the corresponding paper [129]), many of which have been characterized by the Stoddard lab [128, 129, 135]. There are many benefits to working with cTRPs for protein engineering. The closed circular structure offers stabilization through interactions between the first and last repeats of the oligomer, which removes the need for capping repeats to maintain solubility [129]. In particular, Doyle et al. developed a new approach to TRP design (**Figure 4.2**, taken directly from the corresponding paper [129]) that is embedded in the Rosetta molecular modeling package [136] and takes advantage of existing *de novo* design methods [137]. This approach is used in other labs when creating cTRP designs and was essential to the design of the cTRPs showcased in this thesis.

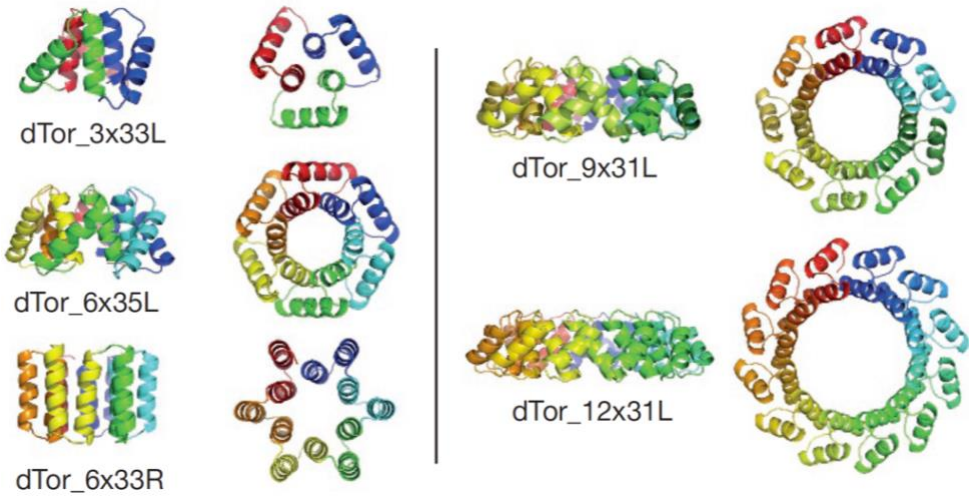

Figure 4.1. Designed monomeric repeat architectures. [129]

Side and top views of a representative design model from each family are shown in cartoon representation colored from blue to red as the chain proceeds from the N to the C terminus [129]. Reproduced with permission from Springer Nature.

Creating a protein that can dimerize in the presence of a particular ligand (peptide or small molecule) provides the ability to create novel chemically induce dimerization systems, drug delivery systems, or alternative energy sources. Due to the multiplicity of cTRPs subunits, they display significant avidity effects that can be harnessed to create tunable pockets to bind unique ligands, as shown in this thesis. Despite the favorable aspects described above, cTRPs have been noted to have trouble accurately stoichiometrically assembling [129] when broken down into homo or hetero sub parts, solutions to which are discussed in Chapter 5.1 and Chapter 5.3.

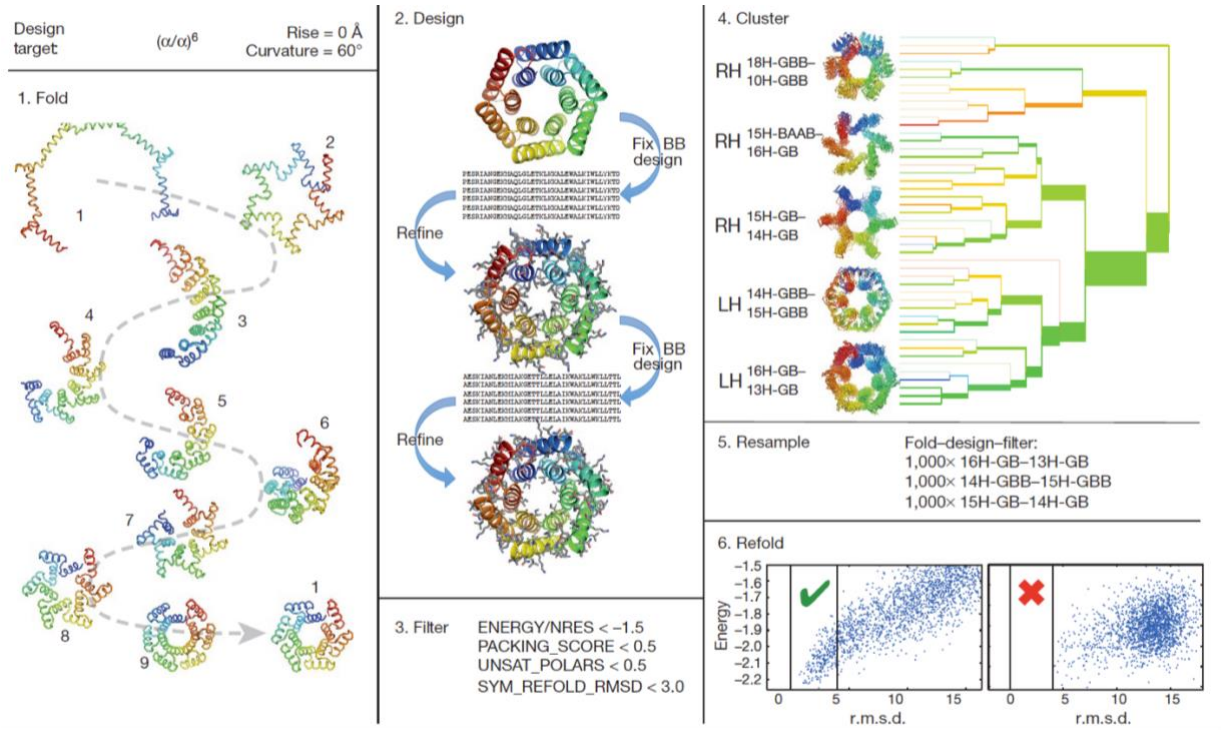

Figure 4.2. Overview of the repeat module design process. [129]

Given a design target consisting of secondary structure types (α /α in this example), repeat number (6), and desired inter-repeat geometry (rise and curvature), the main steps of the design methodology are highlighted. (1) symmetric fragment assembly to generate starting backbone conformations; (2) all-atom sequence design and structure relaxation; (3) filtering to eliminate designs with suboptimal per-residue energy, poor packing, buried unsatisfied polar atoms, or low sequence–structure compatibility; (4) clustering to identify recurring packing arrangements; (5) intensified sampling of architectures identified in the clustering step; (6) final design assessment by large-scale re-prediction of the designed structure starting from the designed sequence; rmsd, root mean squared deviation [129]. Reproduced with permission from Springer Nature.

It has long been the goal to *de novo* design a C2 symmetric dimer that can monomerically dimerize with identical interactions to a ligand of interest. Despite the vast array of C2 symmetric proteins that exist in nature, they are not easily re-engineered to bind a particular ligand due to the many requirements necessary to bind with high specificity. These C2 symmetric proteins must have a large inner cavity that is the correct size and shape for the molecule of interest but still come together reliably in a larger oligomer. As illustrated above, there are many cTRPs which can be made to have C2 symmetry and equipped with large pore spaces capable of accommodating a ligand without ruining the stability of the protein. With the combined effort of many labs, these cTRPs have successfully been designed to specifically bind different C2 symmetric cyclic peptides and chlorophyll dimers as will be discussed in Chapter 5.2.

4.4 HETERO-OLIGOMERIC CTRPS

Pseudosymmetric hetero-oligomers are found throughout nature, and they may have evolved from homo-oligomers that were once composed of the same subunits. In this evolution, gene duplication may have created divergence, which led to specialization and diversification until alike subunits each had a unique specialized function [165, 166]. Hetero-oligomeric proteins have found wide use in nature, becoming assembly centers where each unique subunit can recruit other unique, secondary proteins [167-170].

As with designing specific ligand binding, designing hetero-oligomer cTRPs with unique subunits and specific structural symmetry remains a challenge. Pairing tunable pores capable of binding specific ligands with distinct protein monomer subunits that can reliably oligomerize into one cTRP, would create a construct that not only binds a unique substrate but can carry unique cargo on each monomer in the oligomer.

Scientists have been working to develop strategies to design psedosymmetric hetero-oligomers, and considerable progress has been made [101, 131, 171-179], but to date, there is no systematic path to psedosymmetric structures. Because of the compounding issues with designing such specific hetero-oligomeric cTRPs, specificity must come from the amino acid sidechain identities between monomers [180]. With the combined effort of many labs, a pathway to design and functionalize hetero-oligomeric cTRPs have successfully been accomplished [180, 181], as will be further discussed in Chapter 5.3.

By combining specific ligand binding and hetero-oligomeric design strategies, we can look towards creating the ultimate chemically induced, functional, cargo-carrying *de novo* designed protein, the applications of which reach beyond even current conjectures described in this thesis.

Chapter 5. DESIGN AND CRYSTALLOGRAPHIC ANALYSIS OF ENGINEERED CTRPS

I solved over 35 structures of *de novo* cTRPs during my PhD, 14 of which are included in manuscripts either *in preparation* and/or *submitted* or published [180-183]. These 14 solved structures will be discussed in Chapters 5.2 and 5.3. Along with solving structures, I was responsible for designing cTRPs using Rosetta which is described in Chapter 5.1 [135].

5.1 DESIGN OF FUNCTIONALISED CIRCULAR TANDEM REPEAT PROTEINS WITH LONGER REPEAT TOPOLOGIES AND ENHANCED SUBUNIT CONTACT SURFACES

This section's written material (methods and results) is taken verbatim with permission from its author from a paper which I coauthored [135]. An initial series of cTRPs were designed to self-assemble into a 24-repeat oligomer [128]. There were inadequate contacts shared between the subunits, as such, the larger structure failed to assemble. To combat this issue, a new generation of computationally designed cTRP proteins were generated that incorporated either increased repeat size (longer secondary structural elements) or disulfide staples (cystines) to ensure proper assembly of the cTRP and improve both the energetics and the stoichiometric control of self-association. These new cTRPs are known as thick cTRPs ('tcTRPs').

This study details the *de novo* design of the tcTRPs, their structural features via X-ray crystallography and CryoEM, and their ability to be functionalized with additional folded protein domains [135]. For this work, I designed the stabilizing disulfide 'stapled' constructs (tcTRP24$_8$SS and tcTRP24$_8$SS-Cap) using Rosetta and assisted in protein expression and purification, the methods of which are described in the following sections. I was not involved with the tcTRP9 variants nor the functionalization aspects of the paper and thus those sections were removed from this thesis. The functionalization of the tcTRPs with SARS-CoV-2 spike and SH2 can still be found in the original publication [135] and the success of these structures and their impact on the field will be summarized in the conclusion of this chapter.

5.1.1 *Methods*

Constructs and nomenclature. The construct names, sequences, and figures for all constructs described here, with exception to the disulfide proteins, are provided in the Supplementary Table S1 of the published paper [135]. The constructs described in this article (and listed in that Table) are referred to as 'tcTRPs' ('thick circular Tandem Repeat Proteins'). I designed two of the constructs which are described in (**Table 5.4**). Following the nomenclature for previous cTRP constructs described and used in a prior study [128], the exact size and assemblages are further annotated as 'tcTRP24', where the underlying tcTRP24 scaffold contain a total of 24 repeats. For scaffolds that are assembled from smaller identical protein subunits, the constructs are annotated as tcTRP24x, where each subunit contains 'x' repeats. For example, 'tcTRP24$_8$' (read as 'tcTRP24 sub8') refers to a particle containing a total of 24 repeats, assembled from the trimerization of protein subunits containing eight repeats each. Some tcTRP24 constructs contain disulfide staples between protein subunits and are named 'tcTRP24xSS'.

Table 5.4. Disulfide tcTRP construct designs

tcTRP24$_8$ SS	
DNA sequence	ATGGCTAGCAGCCATCATCATCATCATCATAGCAGCGGCCTGGTGCCGCGCGGCAGCTCCATGGGTAATAGC GAACTGGCGGCGCGTTGCCTGATTATCCTGTTTCAGCAACTGGTGGAACTGGCGCGTCTGGCGATTGAGAGC GGCGATGAGGAACTGCTGCGTCGTGTGAGCGAGTGGCTGGAGGAAGTTATTAAAGACATGCGTCGTGTGGTT GAGCAAGCGCTGCGTGAGGGTAACAGCGAGCTGGCGGCGCGCATCCTGATCATTCTGTTCCAGCAACTGGTG GAGCTGGCGCGCCTGGCGATCGAAAGCGGCGACGAAGAGTTATTACGTCGTGTGAGCGAATGGCTGGAGGAA GTTATCAAAGATATGCGTCGTGTGGTTGAACAAGCGCTGCGCGAAGGTAACAGCGAACTGGCGGCGCGCATT CTGATCATTCTGTTTCAGCAACTGGTTGAGCTGGCGCGTCTGGCGATCGAGAGCGGTGACGAAGAGTTACTT CGTCGTGTGAGCGAGTGGTTAGAGGAAGTTATCAAGGATATGCGTCGTGTTGTGGAGCAAGCGCTGCGCGAA GGCAATAGCGAGCTGGCGGCGCGTATTCTGATTATCCTGTTCCAACAACTGGTGGAGTTAGCTCGTCTGGCG ATTGAAAGCGGCGACGAAGAATTATTACGTCGTGTGAGCGAGTGGCTAGAGGAAGTTATTAAGGACATGCGT CGTGTTGTTGAACAAGCGCTGCGTGAAGGCAACTCTGAGCTGGCGGCGCGTATCCTGATTATTCTGTTCCAA CAACTGGTTGAGTTAGCTCGCCTGGCGATTGAGAGCGGTGACGAAGAATTACTTCGTCGTGTGAGCGAGTGG CTTGAGGAAGTTATTAAAGATATGCGTCGTGTTGTTGAGCAGGCGCTGCGCGAAGGCAACAGTGAATTAGCT GCGCGTATTCTGATTATTCTGTTTCAACAACTGGTGGAGCTTGCTCGCCTGGCGATCGAGAGCGGCGACGAA GAGCTTTTACGTCGTGTGAGCGAGTGGCTCGAGGAAGTTATTAAGGATATGCGTCGTGTGGTGGAGCAGGCG CTGCGCGAGGGCAATTCTGAGTTAGCTGCGCGCATTTTAATTATCTTATTCCAGCAACTGGTTGAGTTAGCG CGCCTGGCGATTGAAAGCGGTGACGAAGAGCTTCTTCGTCGTGTGAGCGAGTGGCTAGAGGAAGTTATAAAG GATATGCGTCGTGTCGTTGAGCAGGCGCTGCGTGAAGGCAATAGCGAACTGGCGTGCCGTATTCTGATCATC CTTTTTCAGCAACTGGTGGAGTTAGCGCGTCTGGCGATCGAAAGCGGTGATGAAGAGCTGCTGCGTCGTGTT AGCGAGTGGCTGGAAGAAGTTATCAAGGACATGCGTCGTGTGGTGGAACAGGCGCTGCGTGAGGGTTAA
Protein sequence	MASSHHHHHHSSGLVPRGSSMGNSELAARCLIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVV EQALREGNSELAARILIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQALREGNSELAARI LIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQALREGNSELAARILIILFQQLVELARLA IESGDEELLRRVSEWLEEVIKDMRRVVEQALREGNSELAARILIILFQQLVELARLAIESGDEELLRRVSEW LEEVIKDMRRVVEQALREGNSELAARILIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQA LREGNSELAARILIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQALREGNSELACRILII LFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQALREG*
tcTRP24$_8$ SS-Cap	
DNA sequence	ATGGCTAGCAGCCATCATCATCATCATCATAGCAGCGGCCTGGTGCCGCGCGGCAGCTCCATGGGTGGTAAC AGCGAGGAAGCGGCGCGTAAGCTGATCATTCTGTTCCAGCAACTGGTGGAGAAAGCGCGTAAGGCGATTGAA AGCGGCGACGAGGAAGAGCTGCGTCGTGTGAGCGAGGAACTGGAGGAAGTTATCAAGGATATGCGTCGTGTG GTTGAACAGGCGCTGCGTGAGGGTAACAGCGAACTGGCGGCTCGTATCCTGATTATTCTGTTTCAGCAACTG GTTGAGCTGGCGCGTCTGGCGATAGAGAGCGGTGATGAAGAGTTATTACGTCGTGTGAGCGAATGGCTGGAG GAAGTTATTAAAGACATGCGTCGTGTTGTGGAGCAGGCGCTGCGCGAAGGTAACAGCGAACTGGCGGCACGT ATCCTGATAATTCTGTTTCAACAACTGGTGGAGCTGGCGCGTCTGGCGATCGAGAGCGGTGATGAAGAATTA TTACGTCGTGTGAGCGAGTGGCTGGAGGAAGTTATAAAAGACATGCGTCGTGTGGTGGAGCAGGCGCTGCGT GAAGGTAACAGCGAACTGGCGGCGCGTATCCTGATAATTCTGTTCCAACAACTGGTTGAGCTGGCGCGTCTG GCGATAGAAAGCGGTGATGAAGAGCTTTTACGTCGTGTGAGCGAGTGGCTGGAGGAAGTTATAAAGGACATG CGTCGTGTTGTTGAGCAGGCGCTGCGCGAGGGTAACAGCGAACTGGCGGCCCGTATCCTGATAATTCTGTTC CAGCAACTGGTGGAGCTGGCGCGTCTGGCGATAGAATCTGGTGATGAAGAGTTACTTCGTCGTGTGAGCGAG TGGCTGGAGGAAGTTATAAAGGATATGCGTCGTGTTGTGGAACAGGCGCTGCGCGAGGGTAACAGCGAACTG GCGGCACGTATCCTGATCATTCTGTTTCAGCAACTGGTTGAGCTGGCGCGTCTGGCGATAGAAAGTGGTGAT GAAGAGCTTCTTCGTCGTGTGAGCGAGTGGCTGGAGGAAGTTATAAAGGATATGCGTCGTGTAGTGGAGCAG GCGCTGCGAGAGGGTAACAGCGAACTGGCGGCACGTATCCTGATTATTCTGTTCCAGCAACTGGTGGAGCTG GCGCGTCTGGCGATAGAAAGTGGGGATGAAGAGTTACTCCGTCGTGTGAGCGAGTGGCTGGAGGAAGTTATA AAGGATATGCGTCGTGTCGTGGAGCAGGCGCTGCGGGAGGGTAACAGCGAACTGGCGGCACGTATCCTGGAG ATTCTGTTCCAGCAACTGGTTGAGCTGGCGCGTCTGGCGAAGGAAAGTGGCGATGAAGAGTTACTACGTCGT GTGAGCGAGTGGCTGGAGGAAGTTAAGAAGGATATGCGTCGTGTAGAGGAACAGGCGAAACGCGAGGGTATT ATATAA
Protein sequence	MASSHHHHHHSSGLVPRGSSMGGNSEEAARKLIILFQQLVEKARKAIESGDEEELRRVSEELEEVIKDMRRV VEQALREGNSELAARILIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQALREGNSELAAR ILIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQALREGNSELAARILIILFQQLVELARL AIESGDEELLRRVSEWLEEVIKDMRRVVEQALREGNSELAARILIILFQQLVELARLAIESGDEELLRRVSE WLEEVIKDMRRVVEQALREGNSELAARILIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQ ALREGNSELAARILIILFQQLVELARLAIESGDEELLRRVSEWLEEVIKDMRRVVEQALREGNSELAARILE ILFQQLVELARLAKESGDEELLRRVSEWLEEVKKDMRRVEEQAKREGII*

Computational protein design. Protein design simulations were conducted exactly as described previously [128]. That approach corresponds to a geometry-guided repeat computational strategy implemented in the Rosetta package [136] with additional *de novo* design elements [137]. Key features

include the application of parametric symmetrization of backbone and side-chain conformations applied across all repeats (such that computational complexity scales only with repeat length); a pseudo-energy term that optimizes the inter-repeat geometry; clustering and resampling protocols that allow intensified exploration of promising topologies; and an in silico validation step that assesses sequence-structure compatibility by attempting to re-predict the designed structure given only the designed sequence.

In this work, two additional modifications of the previously described approach were implemented: the 'Ref2015' energy function [184] was used for all protein design and structure recapitulation calculations, and the range of allowed helix lengths was increased to 20−45 residues. Initial simulations explored helical linkers of length 1−5 residues with unconstrained backbone torsion angles. Clustering analysis of low-energy designs from these simulations revealed convergence on a 2-residue, antiparallel connection with backbone conformation 'GB' (one residue in a left-handed α-helical conformation and one residue in an extended conformation). A subsequent round of designs focused on 'GB' linkers was conducted to enhance sampling in this low-energy region of conformational space.

The identification of residue positions for the incorporation of disulfide staples into the tcTRP24 [115] trimer was performed by utilizing the Rosetta 'Disulfidize Mover' routine [185]. Each edge helix involved in the trimerization was selected and corresponding residues scanned. The distance between adjoining beta-carbons was used to determine potential residues; once identified they were mutated to cystine residues and tested through rotamer optimization and energy minimization.

Protein expression and purification. All constructs encoding tcTRPs described in this study were designed and ligated into an in-house pET15HE expression vector33 or a commercially available pET28b expression vector and sequence verified. The coding sequence and the corresponding translated protein sequences, including the N-terminal poly-histidine affinity tag and thrombin cleavage site preceding the first tcTRP repeat, are provided in the Supplementary Table S1 of the published paper [135].

Plasmids were transfected into BL21(DE3)-RIL Escherichia coli cells (Agilent Technologies) and plated on LB medium augmented with 100 $\mu g m L^{-1}$ ampicillin. Protein was expressed via a previously described autoinduction protocol [186]. Briefly, 1 L of ZYP-5052 media containing 100 $\mu g m L^{-1}$ ampicillin was inoculated with individual transformants, shaken at 37 °C for 8 h followed by 16 °C for 24 h. Expression cultures were pelleted by centrifugation and stored at −20 °C until purification.

Frozen cell pellets were thawed at room temperature and resuspended in 100 mL of 1× phosphate-buffered saline (PBS; 137mM NaCl, 10mM Na_2HPO_4, 2.7mM KCl, pH 7.4.). PMSF was added to a final concentration of 0.5 µM. Cells were lysed via sonication and centrifuged in an SS34 rotor at 16,000 rpm for 20 min at 4 °C to remove cell debris. The supernatant was passed through a 5 µm filter, added to 2mL of nickel-NTA metal affinity resin (Invitrogen) equilibrated with 1× PBS, and then incubated on a rocker platform at 4 °C for 1 h. After loading onto a gravity-fed column, the resin was washed twice with 25 mL of PBS containing 25mM Imidazole. The protein was then eluted from the column by three additions of 5 mL 1× PBS containing 300mM Imidazole. Fractions containing the eluted protein were pooled, concentrated, and buffer exchanged into 1× PBS. The sample was then filtered through a 0.2 µm filter and run over a size exclusion column (Cytiva HiLoad 16/60 Superdex 200) equilibrated in either 1× PBS or 20mM Tris pH 7.5 + 150mM NaCl.

Cryogenic electron microscopy (CryoEM) visualization of tcTRP24$_8$ and tcTRP24$_8$SS. Both purified proteins were screened with negative-stained transmission electron microscopy (TEM) using a 120 KV JOEL1400 electron microscope equipped with a 16 megapixel (4k × 4k) GATAN RIOL CMOS detector. The samples were prepared by depositing 4 µL of purified proteins at approximately 40 nM to the surface of a glow-discharged uniform carbon-coated grid. The particles were allowed to adsorb to the carbon film for ~ 1 min and washed three times with 20 µL of water and once with a drop of 0.7% uranyl formate followed by staining for 25 s with a 40 µL droplet of uranyl formate solution. Excess stains were wicked away with filter paper and the grids were air-dried overnight prior to analysis.

The tcTRP24$_8$SS particles were further analyzed by CryoEM. Samples were prepared by applying an aliquot of 3 µL protein sample of tcTRP24$_8$SS to a glow-discharged Quantifoil1.2/1.3 holey carbon grid, blotted with filter paper for 5 s and plunge-cooled in liquid ethane using an FEI Vitrobot Mark IV. Cryo-EM micrographs were collected on a 200 kV Glacios microscope (FEI) equipped with a Gatan K2 Summit direct detection camera. The microscope was operated at a calibrated magnification of 37,000×, yielding a pixel size of 1.16 Å on micrographs with an accumulated dosage of 60 e−/A2S. In total, 627 movies were collected from two screening sessions, including 82 at a tilt angle of 45°.

All data preprocessing, 2D classification, and 3D model generation and refinement, as well as post refinement polishing, were performed using the software package CryoSPARC2 [81]. For each movie stack,

the frames were aligned for beam induced motion correction using Patch-motion-correction. Patch-CTF was used to estimate the contrast transfer function (CTF) parameter. A new ring-shape algorithm with inner/outer diameters of 100/120 was used for automated blob picking. After inspection and local motion correction, 627,763 particles were accepted for reference-free 2D classification. Two consecutive runs of 2D classification/selection were used to root out false positive and bad (overlapping) particles. A total of 121,426 particles in 20 classes were used for ab initio 3D reconstruction.

It is obvious from the selected classes that there were at least two populations of particles with different diameters. Three models were requested for ab initio 3D reconstruction. Results from 3D reconstruction showed multiple circular-disk particles with different diameters. The proportion of the three 3D classes varied with the number of consecutive 2D classification/select and images selected. Multiple trials were performed with different particle picking protocols and particle diameters. All approaches yielded similar results.

5.1.2 *Results*

Design, expression, and solution behavior of tcTRP9 and tcTRP24. We previously developed an approach to geometry guided repeat protein design that was described in [120, 129, 187] and is implemented in the Rosetta molecular modeling package [136, 137]. Key features include symmetry of backbone and side-chain conformations extended across all repeats (allowing computational complexity to scale with repeat length rather than protein length); a pseudo-energy term that favors the desired inter-repeat geometry; clustering and resampling stages that allow intensified exploration of promising topologies; and an in silico validation step that assesses sequence-structure compatibility by attempting to re-predict the designed structure given only the designed sequence. When applying this approach to the design of closed tandem repeat proteins with increased thickness (‘tcTRPs’ or ‘thick(er) circular tandem repeat proteins’), we identified designs with repeats corresponding to right-handed helical bundles as displaying favorable predicted folding energetics for their designed backbone topologies, and well-formed energy funnels when subjecting their sequences to unbiased fold predictions in Rosetta. We ultimately selected a small number of individual designs corresponding to two different tcTRP sizes and symmetries (containing 9 repeats and 24 repeats, respectively) to examine and validate using biophysical and structural approaches.

The smaller of the two designed tcTRPs ('tcTRP9') was initially expressed and purified both as a monomeric construct containing 513 residues (MW = 56.6 kDa) composed of nine equivalent repeats of 57 residues each, and as a smaller protein chain ('tcTRP9_3') containing three repeats (172 residues; MW=18.9 kDa) that was intended to assemble into a trimeric tcTRP particle with the same architecture and dimensions as its single chain parent (with an exterior diameter of ~60 Å and an interior pore ~15 Å across). The larger of the designed tcTRPs ('tcTRP24'; **Figure 5.1**) was generated solely from smaller subunits containing either six repeats ('tcTRP24_6') or eight repeats (tcTRP24_8') that were intended to assemble into tetrameric or trimeric tcTRP particles, each with an overall architecture and dimensions corresponding to the original design of a single-chain tcTRP containing 24 repeats in total (with an exterior diameter of ~100 Å and an interior pore ~ 60 Å across). The tcTRP24_8 was ultimately chosen for the disulfide staple design.

Generation and analysis of tcTRP24_8. Having demonstrated the accuracy and behavior of a designed thick(er) cTRP (tcTRP) harboring nine repeats and a relatively acute radius of curvature between sequential repeats, we next decided to generate and examine a similar tcTRP harboring 24 repeats, corresponding to a considerably larger internal and external diameter and a shallower curvature between consecutive repeats. Because the large size of these constructs (and the corresponding large number of repeats within their designed structures) precluded generating monomeric, single-chain constructs of 24 linear repeats, we decided to immediately test our designs using smaller subunits with fewer number of repeats that were again intended to self assemble to form full-sized tcTRP toroidal particles.

Designs corresponding to subunits containing either 6 (tcTRP24_6) or 8 (tcTRP24_8) repeats, that were respectively intended to assemble into a full-sized particle with 24 repeats via tetramerization or trimerization were found to both express at high levels. The tcTRP24_8 construct was slightly better behaved in subsequent purification attempts and was therefore used for further experiments. The protein was purified to relative homogeneity using a three-column protocol.

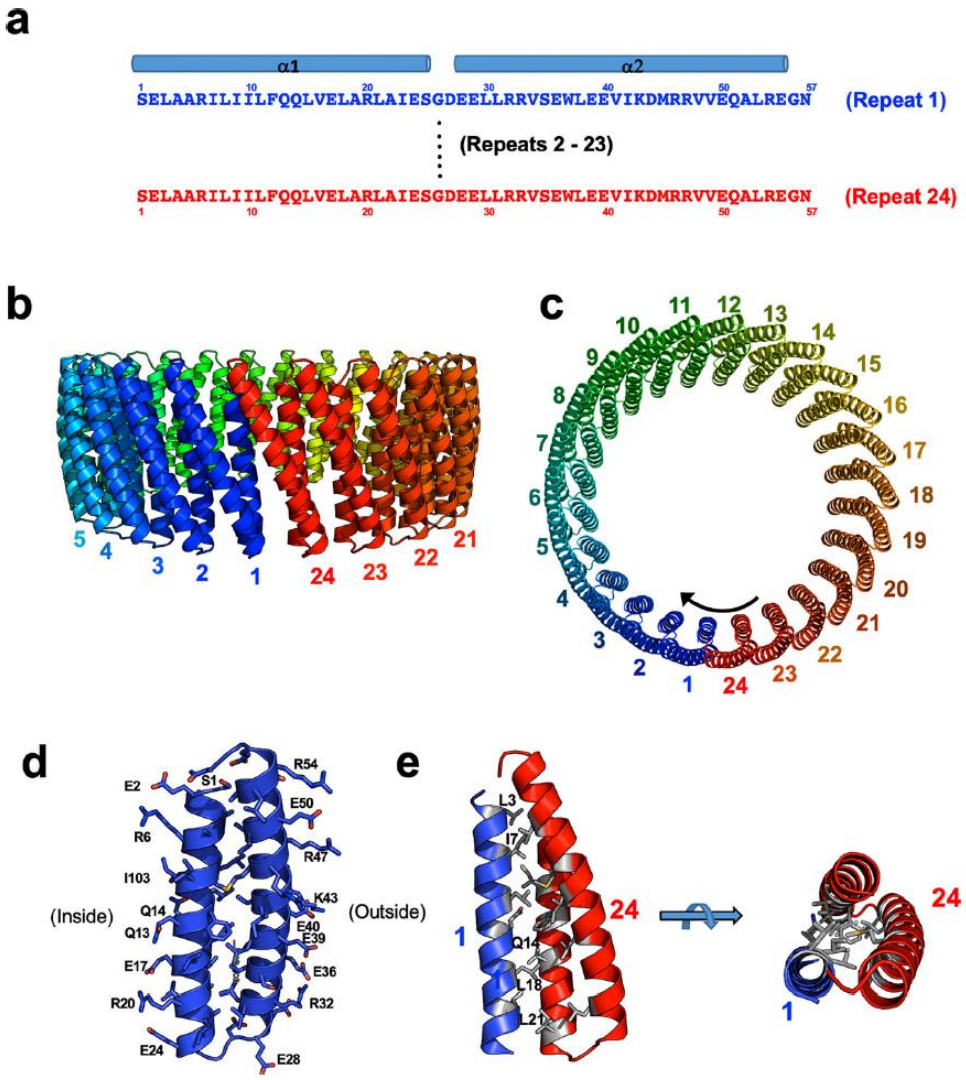

Figure 5.1. Design of tcTRP24.

a) Sequence and secondary structure of individual repeats, which each consist of an anti-parallel two-helix bundle connected by short turns composed of 'GD' sequences. b, c) Twenty-four consecutive repeats (colored blue at the N-terminal repeat and progressing to red at the C-terminal repeat in all figure panels) form a closed cylindrical structure, with an exterior diameter of ~100 Å and an interior pore diameter of ~60 Å. d) The structural composition of the two-helix bundle corresponding to each designed repeat is chemically similar to that of the tcTRP9 design (**Error! Reference source not found.**), again placing mostly charged glutamic acid and arginine residues on the exterior of the particle, and a network of charged and hydrophilic residues on its interior surface. e) The interface between individual repeats.

The tcTRP24$_8$ construct eluted from the final SEC column over a wide range of elution volumes (**Figure 5.2**), corresponding to a heterogeneous population of sizes ranging from greater than 670 kDa to a final peak corresponding to a mass slightly greater than 158 kDa (near the expected mass of individual tcTRP particles composed of three subunits with a total of 24 repeats). Fractions from the SEC elution were subsequently subjected to TEM imaging of negative stained specimen (**Figure 5.3**, left panel). The resulting

images contained mixtures of individual open ('c'-shaped) and closed (circular) toroidal particles with the diameter of the latter corresponding to the expected dimension of the designed protein. Those particles were interspersed with fibrous assemblages that appeared to correspond to elongated chains of protein subunits, forming spiral assemblages of variable lengths. The thickness of the fibers was similar to the diameter of the neighboring rings.

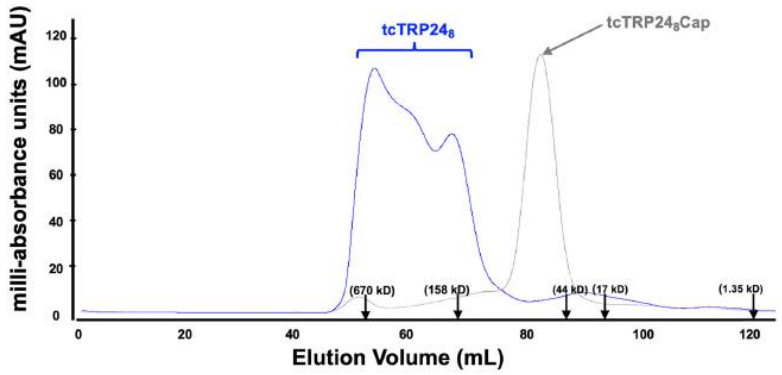

Figure 5.2. Assembly behavior of tcTRP24$_8$.

Size-exclusion chromatographic (SEC) elution profiles of the original 'tcTRP248' homotrimeric construct (blue), and the 'tcTRP24$_8$Cap' construct (gray), which harbors a pair of disrupting mutations at the subunit interface that prevents self-association. The original construct displays a wide range of apparent masses, with the final peak corresponding to the approximately expected elution and mass for the desired homotrimeric species.

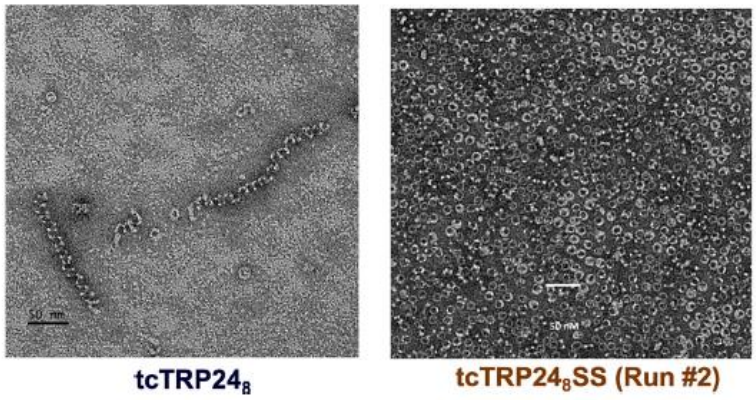

Figure 5.3. Differences in Assembly behavior between tcTRP24$_8$ and tcTRP24$_8$SS.

Negative stain EM image of tcTRP24$_8$ (left) and of tcTRP24$_8$SS ('Run #2'; right). The original construct (lacking disulfide staples; left panel) displays a mixture of correctly sized trimeric particles (small rings on the micrograph) and long extended fibers. The resized stapled construct (right panel) displays a uniform distribution of tcTRP particles.

Based on those images, we reasoned that the tcTRP24$_8$ design might be forming a mixture of desired circular particles, along with partially closed assemblages with an exposed interface surface at their termini that might promote the incorporation of additional protein subunits (leading to a mixture of

unassembled subunits, closed toroidal particles, and the growth of fibers until free protein was exhausted). We therefore attempted to promote the closed toroidal form of this construct by incorporating a pair of cysteine residues at neighboring positions across each subunit interface so that a disulfide staple might lock the protein into the desired closed topology, as described previously [128, 188]. Based on the tcTRP24$_8$ design, an alanine at position five and an isoleucine at position seven (respectively located within the N- and C-terminal repeats of each subunit) were mutated to cysteine residues, with the intention of installing potential disulfide bonds within each interface of the intended tcTRP24$_8$ trimeric particle ('tcTRP24$_8$SS'). This construct was again well-expressed in E. coli and purified using the same three-column protocol including a terminal size exclusion chromatography (SEC) step. Like the original tcTRP24$_8$ construct, the protein again eluted over a range of volume and corresponding masses extending from approximately 150 to > 670 kDa. However, the peak corresponding to the mass of the desired trimer appeared to be greater in relative height (**Figure 5.4**, 'tcTRP24$_8$SS Run #1').

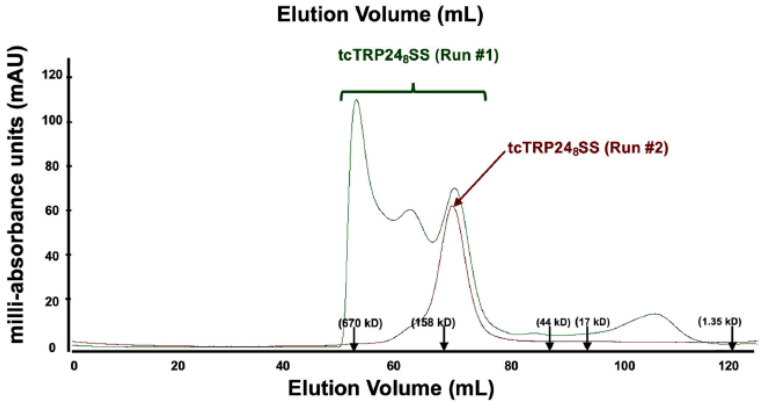

Figure 5.4. Assembly behavior of tcTRP24$_8$SS.

SEC elution profiles of the 'tcTRP248SS' homotrimeric construct directly from the initial purification out of bacterial cells ('Run#1', green), and after collecting, concentrating, and re-running the tail of the original, properly sized fractions ('Run#2', brown). The original prep (Run#1) still forms a distribution of masses after purification, albeit with the earliest and latest eluting peaks increased in relative percentage of material. However, collection of the late-arriving fractions corresponding to the predicted mass of the intended homotrimer and a second run over the same column now results in a single uniform peak corresponding to the intended mass of the designed homotrimeric construct.

Reasoning that the designed disulfide bond might only form after cell lysis and exposure to a non-reducing environment (and during that time, would compete with misfolding and fiber formation), we isolated the slowest eluting half of the final peak (corresponding to the expected elution of a correctly formed trimeric particle), concentrated the protein, and subjected it to a second size exclusion step and analysis (**Figure**

5.4, ‘tcTRP24$_8$SS Run #2’). In that second run, the protein now ran as a single uniform peak at a volume and predicted mass corresponding to the desired trimeric assemblage. To further test that conclusion, fractions of tcTRP24$_8$SS from this second SEC elution were again subjected to transmission electron microscopy using negative stained specimens (**Figure 5.3**, right panel) and found to correspond to individual toroidal particles with dimensions close to the computational design of a 24-repeat toroid.

A negative control construct (‘tcTRP24$_8$Cap’), containing bulky side-chain substitutions in the protein interface that were intended to block self-assembly through steric clashes, was observed to elute at a later volume and smaller corresponding mass, further validating the assembly of the tcTRP24$_8$ and tcTRP24$_8$SS constructs (**Figure 5.2**). The structure of the isolated tcTRP24$_8$SS construct from the final SEC run was then further examined by CryoEM single particle analysis (**Figure 5.5** and **Table 5.5**). The resulting density map at a resolution of approximately 6.9 Å, at a Gold Standard Fourier Shell Correlation (GSFSC) of 0.143 between the two half maps, clearly indicated the presence of a close circular shaped particle with two concentric layers of helices lining its periphery. The topology of the map is consistent with the trimeric tcTRP24$_8$SS toroidal assembly and is closely superimposed on the original design model (**Figure 5.5**). In the final electron density map, three strong features extending from equivalent positions around the tcTRP particle correspond to N-terminal poly-histidine affinity tags and linkers extending from the N-terminus of each protein subunit (an unbiased feature that further validated the trimeric assemblage of the particle). Those atoms are clearly disordered and are not modeled into the density.

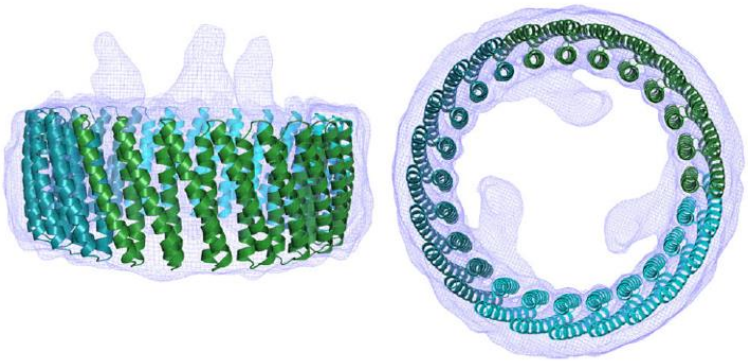

Figure 5.5. CryoEM single-particle reconstruction of tcTRP24$_8$SS.

Two different views of the electron density map superposed with the original model. The three-strong features extending from equivalent positions around the tcTRP particle correspond to N-terminal poly-histidine affinity tags and linkers extending from the N-terminus of each protein subunit (an unbiased feature that further validated the trimeric assemblage of the particle). Those atoms are clearly disordered and are not modeled into the density.

Table 5.5. tcTRP24$_8$SS CryoEM data collection, refinement, and validation statistics

Data Collection and Processing	
Magnification	38 Kx
Voltage (kV)	200
Electron dose (e^-/Å^2)	40
Defocus range (μm)	-0.8 to -2.3
Pixel size (Å/pixel)	1.16
Symmetry	C3
Initial particle images (no.)	627,763
Final particle images (no.)	121,426
Map resolution (Å)	6.9
FSC threshold	0.143
Map resolution range (Å)	30-6.9
Refinement	
Initial model used	tcTRP24$_8$SS
Model resolution (Å)	6.9
FSC threshold	0.143
Model resolution range (Å)	30-6.9
Map sharpening B-factor (Å^2)	-418.6
Model composition	
Protein residues	1,368
Ligands	None
B-factors (Å^2)	
Protein	20.0
Ligands	None
MolProbity Score	1.6
Clashscore	1.05
Rms deviations	
Bonds length (Å)	0.01
Bonds Angle (°)	1.2
Ramachandran plot statistics (%)	
Most favored	98.32
Allowed	1.68
Outlier	0.00

Inspection of the images from the selected 2D class averages of particles used in the reconstruction indicated that the protein assemblage, even with designed disulfide staples in place at the interfaces between protein subunits, displayed significant conformational flexibility or ‘breathing’ that are manifested in slight deviations from circularity and small but significant variations in the diameter of the particles, which severely limited the resolution of the 3D reconstruction.

In the same CryoEM analyses, we also noted that while the vast majority of the 2D class averages of the closed circular tcTRP24$_8$SS particles contained 24 repeats as designed, a small fraction of particles (less than 5%) contained more or less than 24 repeats, indicating the presence of ‘n +/− 1’ and/ or ‘n +/− 2’ repeats in a small fraction of the underlying tcTRP subunits. One possible explanation for this observation could be occasional recombination events within the repetitious tcTRP24$_8$ coding sequence, leading to the insertion or deletion of individual DNA repeats from the expression vector. Using PCR, we compared the length of the coding sequence inserts in the original expression vector with plasmid populations recovered from cells after extended culture growth and expression but did not see any indication of abnormal length inserts accumulating during cell growth and expression. Thus, the basis for the small percentage of tcTRP24$_8$ particles harboring more or less than 24 repeats is unclear.

5.2 De novo design of protein homodimers containing tunable C2 symmetric protein pockets that bind ligands

The bulk of this section’s written material is taken verbatim from Hicks et al. 2022 [183] and a submitted manuscript Ennist et al. 2023 [182], in which I am a coauthor, with explicit permission from the authors and further expanded upon to include unpublished data. I am responsible for the bulk of the x-ray crystallographic work in these projects. Additional unpublished data sets are also presented in this section and reflect the ongoing work to design tunable C2 symmetric pockets to bind unique ligands.

An overarching goal of ongoing cTRP design is to equip the scaffold proteins with the ability to oligomerize only when in the presence of a unique ligand. This pursuit comes with many challenges; the proteins need to have (i) a tunable and highly specific pore, (ii) stable monomeric states, and (iii) mutable helices to accommodate a variety of ligand shapes and sizes [183]. The Baker lab developed an approach that builds on previous strategies for *de novo* design of α-helical repeat proteins [129, 189] to create nanocages and chemically induced protein switches, and proteins that bind specific mineral surfaces [131, 190, 191]. This strategy has been expanded for cTRPs in which the previously circular cTRP pore was adapted to be elliptical, broadening the range of C2 symmetric ligands possible within a given pore [183].

To create C2 symmetric proteins with cavities large enough for a ligand, a diverse library of monomeric proteins to sample different symmetric orientations was made. This screening method had been

done before but did not yield elliptical structures with enough space in the middle to dock a ligand [189]. To overcome this barrier, a new generation of designs were envisioned that were made up of four helical four-unit repeat monomers aimed to build C2 symmetric proteins when oligomerized. Additional calculations and design considerations were applied to create the generation of proteins described in Hicks' 2022 paper [183]. The central pore was optimized to bind C2 symmetric ligands by employing α-helices with minimal rise across the superhelical axis and favoring the ends where the two monomers connect [183]. A final parent C2 symmetric homodimer design architecture was chosen for modification for ligand specifications (**Figure 5.6**). The pore can be lined with different sidechain functional groups and the inside is separate from the exterior of the proteins, making it unlikely that customization of the inner proteins will destabilize the oligomer [183]. Greater description of this protein design work can be found in their associated cited papers [182, 183].

Figure 5.6. Design Pipeline of C2 Symmetric Ligand Binders. [183]

Schematic of design pipeline from curved repeat protein (Left) to symmetric homodimers (Center) to C2 symmetric ligand binders (Right). Color gradient represents the protein chain direction from N terminus (blue) to C terminus (red). Hypothetical C2 symmetric ligands are shown in gray. [183]

Just as these cTRPs can be redesigned to accommodate a single symmetric circular peptide, these circular proteins can be modified to bind naturally occurring small molecules, in this case chlorophyll. These chlorophyll dimers are known as a "special pair" which accept excitation energy and initiate an electron-transfer cascade while harvesting light [182]. The behavior of the photosynthetic proteins can fine tune many aspects of the photosynthetic process, such as the absorption and fluorescence spectra and the energy transfer [138-141]. To develop new synthetic energy conversion technologies, we need to know more about the intricate function of these special pair systems. Various studies have been performed on bacteriochlorophylls [142-147] and chemists have attempted to develop small molecule special pair-like structures [148-152], but they do not fully capture the native function of these special pairs [153-164]. To study these molecules further in an assembly that matches their special pair geometries it is essential to

separate them from the complex photosystems found in nature, which is possible to do by trapping them inside the C2 symmetric cTRPs as shown in this chapter [182].

cTRPs in this thesis have been optimized to fit two different types of unique C2 ligands include cyclic peptides (1 per cTRP dimer) and chlorophyll analogues (2 per cTRP dimer) [182, 183]. This thesis describes the two different cyclic peptides that have been successfully used as ligands for the C2 symmetric homodimers that have yet to be published. In the first successful design, the peptide was well defined in the pore of the cTRP, but it was in a confirmation that had not been seen in any solvent explored. The second successful design showed the peptide bound in its expected conformation, with little differences from the designed structure.

The chlorophyll bound x-ray crystallographic structures shows that one C2 cTRP dimer binds two chlorophylls in a geometry where one chlorophyll matches the native special pairs while the other is positioned in a new conformation [182]. Harnessing all that was learned previously in designing ligand binding proteins [129, 131, 183, 189], the team was able to design a homodimer with perfect C2 symmetry that bound a chlorophyll dimer such that the axis was like the native reaction center [192, 193].

5.2.1 *Methods*

C2 symmetric protein design. Proteins were designed using the Rosetta macromolecular modeling suite [194]. Backbones were generated with RosettaRemodel using a coarse-grained energy function supplemented with scoring terms that bias the trajectories toward desired helical parameters. Monomers were subsequently designed with a FastDesign protocol with repeat symmetry enforced and top scoring monomers with docked into C2 symmetric homodimer geometries using the Rosetta app sicdock as previously published [131] with the added requirement that the proteins form closed circular architectures. Interfaces were then designed using a FastDesign protocol with C2 symmetry enforced, and top scoring designs were ordered for experiment characterization. Additional computational methods can be found in associated papers [155, 156, 183].

Computational placement of the chlorophyll special pair into symmetric protein scaffolds. Identifying residue positions capable of accommodating the chlorophyll special pair which is scalable to millions of potential scaffolds was achieved by utilizing a motif-hash based method [131] specifically adopted for the

histidine-chlorophyll dimer motif inspired by the special pair of purple bacteria, P865. However, the number of example structures of the histidine-chlorophyll dimer motif found in the PDB is not acceptable for effectively populating a motif hash table. Therefore, generating additional structural examples of the symmetric histidine-chlorophyll dimer complex was generated *de novo*.

The conformer generation was achieved using the NeRF algorithm (available on github at https://github.com/atom-moyer/nerf) which translates internal molecular coordinates to global molecular coordinates. Various conformers were generated by varying the internal coordinates such as the relative positioning of the chlorophyll groups, the dihedral of ligation by the histidine residue, and the rotamer of the histidine sidechain. The full complex was duplicated along the C2 axis to create the symmetric complex. If the relative orientations of the chlorophylls were varied, clashes between the rings and their substitutions were evaluated and filtered. The full process of *de novo* motif generation was repeated for ligation with the epsilon and delta nitrogen of the imidazole ring.

Once the *de novo* conformers were generated, the 6-D transformation that defines the relative orientation of the N-CA-C atoms of the ligating histidine residues were hashed using a method described previously [131]. The hashed 6-D transformation was used as a key in a multi value hash table (https://github.com/atom-moyer/getpy), and the associated value was a vector that defined the information necessary to rebuild the histidine-chlorophyll complex, which nitrogen from the histidine was used for ligation and the internal coordinates of the histidine rotamer.

During evaluation of design scaffolds, the 6-D transformation of each symmetric residue pair across chains was evaluated and hashed using the same method used to hash the *de novo* conformers described above. That allowed the identification of symmetric residue pairs which have similar 6-D transformations to the potentially acceptable ligation geometries. If a matching 6-D transformation was found, the histidine-chlorophyll complex was rebuilt from the associated value in the hash table, and the complex was evaluated in the context of the protein. If the chlorophylls did not clash with the backbone atoms of the protein, the placement was accepted and passed into the protein design process.

A python package and example scripts which generate the *de novo* hash tables and place the histidine-chlorophyll complexes into symmetric proteins can be found here: https://github.com/atom-moyer/stapler.

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 [195] 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.

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 (~24 hours) or at 37 °C the entire time ~14 hours. Cells were pelleted at 4,000g for 10 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, 1mM PMSF, 0.75% CHAPS, 1 mM DNase, 10mM Lysozyme, with Thermo Scientific Pierce protease inhibitor tablet). Cell suspensions were lysed by microfluidizer or sonication, and the lysate was clarified at 20,000g for ~30 minutes. The His-tagged proteins were bound to Ni-NTA resin (Qiagen) 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, 300 mM NaCl, 300 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, 300 mM NaCl).

An expression plasmid (“pET15b D_3_633_x8”) encoding construct D_3_633_x8 (preceded by an N-terminal histidine affinity tag and a TEV protease cleavage site) was sequence verified and transformed into BL21(DE3) RIL Escherichia coli cells (Agilent Technologies) for protein expression. Individual colonies

were (grown overnight on lysogeny broth (LB) plates augmented with kanamycin at 100 μg mL−1) were picked and transferred to separate 10 mL aliquots of LB–kanamycin media and shaken overnight at 37 °C. Individual 10 mL aliquots of overnight cell cultures were added to individual 1 liter flasks LB medium plus kanamycin, which were then shaken at 37 °C until the cells reached an optical density (OD) at 600 nm of 0.6–0.8. Isopropyl-β-d-thiogalactoside (IPTG) was then added to each flask to a final concentration of 1 mM to induce protein expression. The flasks were shaken overnight at 16 °C, and then pelleted by centrifugation and stored at − 20 °C until purification. Cell pellets from 4 liters of cell culture were resuspended in 50 mL of PBS solution (140 mM NaCl, 2.5 mM KCl, 10 mM NaHPO4, 2 mM KH2PO4) containing 20 mM imidazole (pH 8.0) and 1 mM phenylmethylsulfonylfloride (PMSF). Cells were lysed via sonication and centrifuged to remove cell debris. The supernatant was passed through a 5 μm and a 0.45 μm filter. After loading onto a gravity-fed nickel-NTA metal affinity resin (Invitrogen) column, the resin was washed with 50 mL of the same buffer used for lysis, and the protein was eluted from the column with 30 mL of PBS containing 300 mM imidazole (pH 8.0). The purified protein was exchanged into 30 mM Tris (pH 8) and 150 mM NaCl while also concentrating to a volume of 1 mL. The protein was further purified via size-exclusion chromatography using HiLoad 16/60 Superdex 200 column (GE). All samples were tested for purity via SDS–polyacrylamide gel electrophoresis. The final protein sample was concentrated to 12.6 mg mL−1 or 216.92 μM for crystallization.

In additional experiments, proteolytic removal of the N-terminal His-tag did not lead to successful diffracting crystals. In those experiments, the His tag was removed by digestion with AcTEV protease (Millipore Sigma) for 2 hours at room temperature. Following the digestion, the protein solution was incubated on a rocker platform at 4 °C for 1 hour after adding 3 mL of resuspended nickel-NTA metal affinity resin (Invitrogen). The cleaved protein was then separated from AcTEV and remaining uncleaved protein via elution from a gravity-fed column.

Protein-chlorophyll sample preparation. Zn pheophorbide a methyl ester (ZnPPaM) was purchased from Frontier Scientific Inc. ZnPPaM stock solutions were prepared in dimethyl sulfoxide (DMSO) or methanol to concentrations between 200 μM and 1 mM. ZnPPaM concentrations were determined using mass measurements and by using the known absorptivity of Zn pheophytin a, which has a similar absorbance spectrum and an extinction coefficient of 77,300 $M^{-1}cm^{-1}$ in 80% acetone/20% deionized water

(Jones et al., 1976). Ultraviolet/visible (UV/vis) absorbance spectra were collected using a Jasco V-750 spectrophotometer with a 1 nm bandwidth and 400 nm/min scanning speed. Protein-ZnPPaM complexes were prepared by slowly adding freshly prepared ZnPPaM stock solution to protein solution in aqueous buffer at room temperature and incubating samples for several hours. Unbound ZnPPaM was removed by centrifugation to pellet precipitated ZnPPaM, sterile filtration using a 0.22 μm syringe filter, and/or running a PD-10 desalting column purification (Sephadex^{TM} G-25 M resin, Cytiva Life Sciences).

D_3_633 crystallography sample preparation, phasing, and refinement. Crystal screening was performed using Mosquito Crystal by STP Labtech and monitored by JANSi UVEX imaging system. 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 (**Figure 5.7**). 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 (**Table 5.6**). The structure of D_3_633, was solved by Molecular Replacement with Phaser [196] via PHENIX [72] using the coordinates of the computationally designed structure as a search model. The structures were then built and refined using Coot [74] and PHENIX [72], respectively. Final Ramachandran statistics after refinement were as follows (given as % preferred, % allowed, % outliers, respectively): D_3_633: 98.35, 1.42, 0.24.

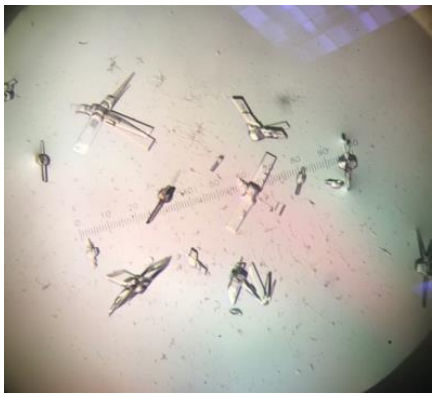

Figure 5.7. Crystals of dimer D_3_633.

Crystals of the parent peptide ligand binder, D_3_633, appeared to be flat rods growing from a single point of nucleation.

D_3_633_8x crystallography sample preparation, phasing, and refinement. Purified protein D_3_633_x8 with and without PJS-I-16 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") 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 HKL2000 [71] (**Table 5.6**). The crystal was found to belong to a primitive orthorhombic space group ($P2_12_12_1$) and yielded a 2 Å resolution data set.

The ligand 'PJS-l-16' (a C2 symmetric circular peptide) was added to the protein construct D_3_633_x8 at twice the concentration of the protein dimer and 1.7% final DMSO and incubated for 5 minutes before adding to crystal trays. 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 HKL2000 [71] (**Table 5.6**). The crystal was found to belong to a primitive monoclinic space group ($P2_1$) and yielded a 2.05 Å resolution data set.

The structures of D_3_633_x8 and D_3_633_x8 PJS-l-16 were solved by molecular replacement with Phaser [196] via PHENIX [72] using the coordinates of the computationally designed structure as a search model. The structures were then built and refined using Coot [74] and PHENIX [72], 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 (**Figure 5.8**). PJS-l-16 energies were calculated using eLBOW [197] via PHENIX [72]. Final Ramachandran statistics after refinement were as follows (given as % preferred, % allowed, % outliers, respectively): D_3_633_x8: 99.33, 0.67, 0.0; D_3_633_x8 PJS-l-16: 99.33, 0.45, 0.22 (**Table 5.6**).

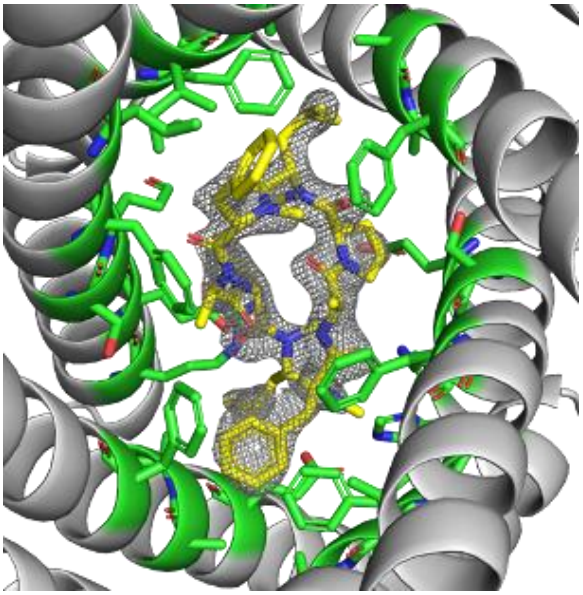

Figure 5.8. Ligand PJS-1-16 fits well within the unbiased density.

PJS-1-16 (yellow) fits within the density (grey mesh) from the x-ray crystallographic data. All The entire protein (grey cartoon and green sticks) and all water molecules were modeled before anything was placed within the density of the pore.

MYS13 and SRH7 crystallography sample preparation, phasing, and refinement. Purified proteins MYS13 and SRH7 with and without their ligands 70H and 16I, respectively, 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. Crystals grew in the presence of the ligands.

The crystal of the designed protein MYS13 was grown from 100 mM calcium chloride and 18% PEG 3350 at a protein dimer concentration of 271 µM. The ligand 70H was added to the protein MYS13 at a concentration twice the protein dimer and 1.2% final DMSO and incubated for 20 minutes before being added to crystal trays. The crystal was transferred to a solution containing 25% PEG 3350 and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.1 and processed using program HKL2000 [71] (**Table 5.6**). The crystal was found to belong to the space group P 1 and yielded a 2.3 Å resolution data set.

The crystal of the designed protein SRH7 was grown from 100 mM sodium acetate pH 4.5 and 8% PEG 4000 at a protein dimer concentration of 153 μM. The ligand PJS-l-16 was added to the protein SRH7 at a concentration twice the protein dimer and 3% final DMSO and incubated for 20 minutes before being added to crystal trays. The crystal was transferred to a solution containing 25% PEG 4000 and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.1 and processed using program HKL2000 [71] (**Table 5.6**). The crystal was found to belong to the space group P 2_1 2_1 2_1 and yielded a 2.75 Å resolution data set.

The structures of MYS13 and SRH7 PJS-l-16 were solved by molecular replacement with Phaser [196] via PHENIX [72] using the coordinates of the computationally designed structure as a search model. The structures were then built and refined using Coot [74] and PHENIX [72], respectively. Despite having ligand present in the crystallization solution, the entire ligand 70H could not be seen in the density of MYS13, thus the name MYS13 lacks the peptide. Residual density was present, but not enough to build in the ligand with a high degree of confidence. For the SRH7 bound to PJS-l-16, the same steps were followed as with the D_3_633_8x structure; the protein and solvent were modeled and refined first (while avoiding the modeling of any atoms within the binding site) then built into the unambiguous density for the bound circular peptide ligand, with the last step being to add waters. PJS-l-16 energies were calculated using eLBOW [197] via PHENIX [72]. Final Ramachandran statistics after refinement were as follows (given as % preferred, % allowed, % outliers, respectively): MYS13: 99.7, 0.3, 0.0; SRH7 PJS-l-16: 98.11, 1.57, 0.31 (**Table 5.6**).

Table 5.6. Cyclic peptide binder crystallographic data and refinement statistics

Design	D_3_633	D_3_633_8x	D_3_633_8x PJS-I-16	MYS13	SRH7 PJS-I-16
Data Statistics					
Space group	P 1 2_1 1	P 2_1 2_1 2_1	P 1 2_1 1	P 1	P 2_1 2_1 2_1
Unit cell dimensions	a=54.44 Å, b=54.68 Å, c=86.98 Å, α=γ=90°, β=95.5°	a=54.38 Å, b=55.03 Å, c=146.53 Å, α=γ=β=90°	a=54.72 Å, b=54.56 Å, c=80.48 Å, α=γ=90°, β=96.1°	a=47.85 Å, b=59.29 Å, c=60.56 Å, α=88.19°, γ=66.83°, β=88.26°	a=33.43 Å, b=79.80 Å, c=105.94 Å, α=γ=β=90°
Resolution (Å)	48.05 - 2.32 (2.407 - 2.32)	44.00 – 2.0 (2.071 – 2.0)	47.4 – 2.05 (2.126 – 2.05)	43.9 – 2.3 (2.382 – 2.3)	39.9 – 2.75 (2.848 – 2.75)
Reflections	21183 (1929)	30453 (2906)	29557 (2789)	26355 (2363)	7762 (700)
Completeness (%)	95.19 (87.13)	99.62 (96.41)	99.18 (94.18)	96.97 (86.88)	98.75 (91.03)
Redundancy	4.0 (4.1)	11.8 (6.4)	6.3 (4.8)	3.3 (2.7)	11.6 (9.2)
I/s(I)	15.8 (1.19)	31.36 (0.6)	22.05 (2.07)	15.5 (1.5)	23.44 (2.5)
Rmerge	0.078 (1.029)	0.056 (1.546)	0.074 (0.42)	0.064 (0.481)	0.113 (0.497)
Rpim	0.047 (0.581)	0.017 (0.626)	0.031 (0.199)	0.041 (0.331)	0.035 (0.163)
CC1/2	0.658 (1.06)	0.998 (0.331)	0.995 (0.943)	0.998 (0.75)	1.021 (0.977)
Refinement Statistics					
PDB ID	7RKC	Unpublished	Unpublished	Unpublished	Unpublished
Resolution (Å)	2.32	2.0	2.05	2.3	2.75
Rwork / Rfree	26.5 / 31.3 (31.6 / 37.2)	23.3 / 28.1 (34.5 / 38.4)	20.5 / 24.3 (25.2 / 29.2)	22.3 / 24.6 (27.8 / 31.9)	25.5 / 30.4 (33.9 / 35.9)
No. atoms					
Protein	2949	450	452	682	322
Ligand/Ion	20	5	93	31	66
Water	5	450	101	0	3
B-factors					
Protein	74.06	52.78	47.52	57.25	70.02
Ligand/Ion		80.28	66.89		56.20
Waters	71.19	53.06	46.83	57.00	52.18
RMS deviations					
Bond length (Å)	0.002	0.007	0.002	0.002	0.001
Bond angles (°)	0.43	0.81	0.40	0.41	0.36

Chlorophyll dimer crystallography sample preparation, phasing, and refinement. X-ray crystallography for SP1 and SP2: Crystals of SP1 and SP2 were grown using protein purified as described above. Protein samples dispensed in 1 µL drops at purification concentrations were mixed with equal volume of a crystallization solution and set in hanging drops (refer to **Table 5.7** for conditions). Vapor phase equilibration of the resulting drops against a 1 mL reservoir of the same crystallization solution resulted in

growth of crystals (**Figure 5.9**). The crystals were flash cooled in liquid nitrogen. Diffraction data were collected on a Pilatus area detector at the Advanced Light Source (ALS) synchrotron facility at beamline 5.0.2 for SP1-ZnPPaM and SP2-ZnPPaM protein assemblies. Diffraction data were collected on a Rigaku HyPix-6000HE hybrid photon counting detector at the Fred Hutch for SP2. The resulting data sets (**Table 5.7**) extend to 2.0 Å, 2.4 Å, and 2.5 Å resolution for SP1-ZnPPaM, apo-state SP2, and SP2-ZnPPaM, respectively. The asymmetric units of the SP1-ZnPPaM and apo-state SP2 structures each contained one complete dimer (two copies of a protein subunit), and the SP2-ZnPPaM structure had two dimers in the asymmetric unit.

Data were processed using HKL2000 [71] or Aimless [198]. The placement of subunits was determined using the molecular replacement algorithm in program PHENIX [72]. Local rebuilding of all constructs was performed using the program COOT [74], followed by refinement in PHENIX [72]. For the ZnPPaM-bound structures, the protein was built and refined completely with waters (excluding waters from the binding site) and other chemicals before manually fitting ZnPPaM into the density that remained. ZnPPaM energies were calculated using eLBOW [197]. The final values for R_{work} / R_{free} are notated in **Table 5.7**.

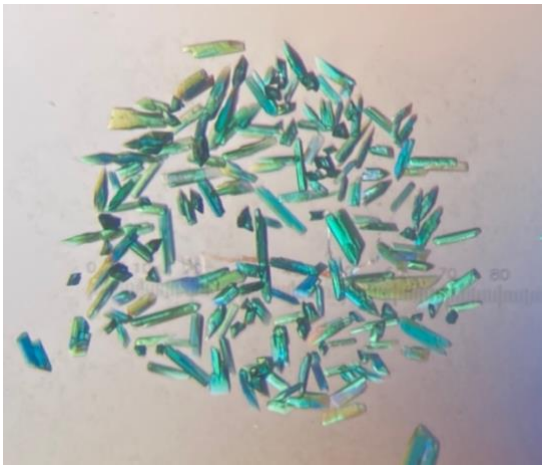

Figure 5.9. Green SP1 crystals.

Crystals of the chlorophyll binders, SP1 and SP2, grew in rods and were various shades of light green, demonstrating the binding of the ligand.

Table 5.7. Chlorophyll binder crystallographic data and refinement statistics

Design	**SP1-ZnPPaM**	**SP2 (apo-state)**	**SP2-ZnPPaM**
Well solution	32% (w/v) PEG-3350, 200 mM lithium sulfate, 100 mM BisTris pH 6.5	24% (w/v) PEG-3350, 140 mM KCl	30% (w/v) PEG-3350, 100 mM ammonium sulfate
Data Statistics			
Space group	P 4_1	P 2_1	P 1
Unit cell dimensions	a=b=52.35 Å, c=173.72 Å, a=b=c=90°	a=54.13 Å, b=76.1 Å, c=63.23 Å, b=99.06°, a=c=90°	a=52.8 Å, b=54.9 Å, c=89.3 Å, a=87.83°, b=84.06°, c=69.45°
Resolution (Å)	36.2 - 2.0 (2.07 - 2.0)	28.89 - 2.4 (2.46 - 2.4)	49.26 - 2.07 (2.13 - 2.07)
Reflections	31367	19927	31294
Completeness (%)	99.8 (99.5)	98.2 (98.2)	92.3 (60.2)
Redundancy	13.0 (13.2)	48.7 (49.7)	3.5 (3.4)
I/s(I)	32.4 (5.4)	50.1 (22.7)	7.5 (0.9)
Rmerge	0.066 (0.429)	0.082 (0.213)	0.067 (1.297)
Rpim	0.019 (0.121)	0.012 (0.030)	0.043 (0.834)
CC1/2	0.999 (0.967)	1.01 (0.84)	0.98 (0.97)
Refinement Statistics			
PDB ID	7UNJ	7UNH	7UNI
Resolution (Å)	2.0	2.4	2.5
Rwork / Rfree	0.196 (0.219) / 0.237 (0.291)	0.209 (0.223) / 0.263 (0.322)	0.201 (0.258) / 0.251 (0.321)
No. atoms			
Protein	3203	3656	6684
Ligand/Ion	107	8	206
Water	65	189	9
Wilson B-factor	38.31	22.70	52.90
RMS deviations			
Bond length (Å)	0.006	0.002	0.007
Bond angles (°)	0.76	0.37	0.86
Ramachandran distribution (favored% / allowed% / outlier%)	99.54 / 0.23 / 0.23	99.38 / 0.62 / 0.0	97.7 / 2.09 / 0.21

Protein structure alignment. Protein crystal structures were compared to Rosetta design models by aligning corresponding backbone Cα atoms and calculating RMSDs using TM-align (Zhang & Skolnick, 2005). (B)Chl special pair geometries were compared using the align function in The PyMOL Molecular Graphics System, Version 2.5.2, Schrödinger, LLC. To facilitate comparison of the geometries of special pairs composed of different (B)Chl types, omit unimportant conformational differences such as rotameric states of peripheral substituents, and neglect differences in the Mg(II) vs. Zn(II) positions, only the atoms of the tetrapyrrole rings were considered in pairwise special pair structural alignments. These atoms

included the 4 pyrrole nitrogen atoms, 16 pyrrole carbon atoms, and 4 methine bridge carbons from each (B)Chl monomer, giving 48 atoms per (B)Chl dimer that were used for structural comparisons. Corresponding atoms were aligned in PyMOL and the RMSD over all 48 atom pairs was calculated. Native BChl a special pairs used for comparison to the SP1 protein came from 5 different species of purple photosynthetic bacteria, including *Rhodobacter sphaeroides, Rhodopseudomonas palustris, Thermochromatium tepidum, Gemmatimonas phototrophica,* and *Thiorhodovibrio* strain 970. The PDB IDs of the nine X-ray crystal and cryo-EM structures containing the native special pairs used for comparison to SP1 were: 7PIL, 7VNY, 6Z27, 6Z02, 6Z5S, 3WMM, 5Y5S, 7O0U, and 7C9R [199-206].

5.2.2 *Results*

The Hicks team was able to generate 100,000 curved repeat protein backbones for preliminary ligand binding design. The top 12,000 designs were submitted for Ab Initio folding simulations, with a final 2,500 designs submitted for docking and design calculations. This group of 2,500 were used to design ellipsoidal central pores capable of holding a ligand. After further optimization, only 44 of the designs were able to be expressed and purified. Of those, only 36 of the proteins were helical and hyperstable. To understand these designs further, three were chosen to crystalize; D_3_337 (7RMY), D_3_212 (7RMX), and D_3_633 (7RKC). The D_3_633 design (the object of further optimization) has repeating helix lengths of 24 and 29 amino acid that surround a central cavity with maximum height and width of 28 Å and 37 Å and a volume of 4,062 $Å^3$. D_3_633 forms the desired ellipsoidal architecture that has a cavity large enough for binding of a small molecule with minimal deviations from the design (Ca rmsd 2.04 Å) (**Figure 5.10**). The design D_3_633 was chosen to develop further to bind a C2 symmetric cyclic peptide. The following results have yet to be published but demonstrate the successful optimization of the C2 dimerization system.

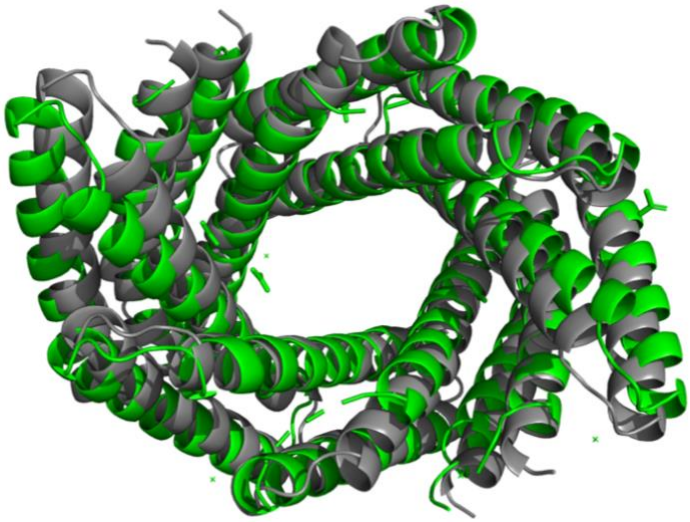

Figure 5.10. Alignment of crystal structure and Rosetta model of D_3_633.

Rosetta design model of D_3_633 in gray. Crystal structure at 2.32 Å of D_3_633 in green. Crystal structure aligns with Rosetta design model 2.04 Å Cα RMSD.

The C2 symmetric cyclic peptide, PJS-I-16, was chosen from an existing screen of C2 molecules due to its already limited ability to bind D_3_633. PJS-I-16 is a 1077 Da cyclized peptide made up of phenylalanine, leucine, proline, and alanine with a few additional modifications to the backbone (**Figure 5.11**). This peptide was designed to be passively cell-permeable, based on limiting the number of polar atoms exposed to solvent. At the backbone, the peptide internally satisfies every NH with a carbonyl within the macrocycle and when that is not possible, the α-nitrogen is alkylated. At the side-chain level, only hydrophobic amino acids were allowed. The peptide is very flexible, and its conformation is highly dependent on its environment and its neighbors and contacts. Before the protein was engineered to bind PJS-I-16, the peptide was structurally studied independently to optimize the mutations made to the cTRP. Structural results showed that the peptide formed a circle shape in H_2O-ACN whereas it formed a squished ellipsoid in EtOAc (**Figure 5.12**).

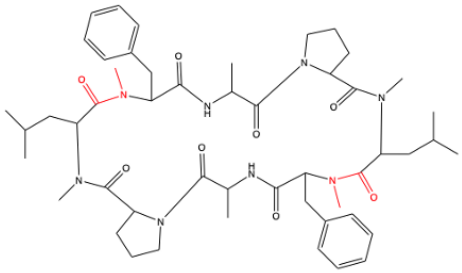

Figure 5.11. Structure of cyclic ligand PJS-I-16.

ChemDraw representation of the cyclized peptide. It is made up of four amino acids (Phe, Leu, Pro, and Ala) which repeat twice.

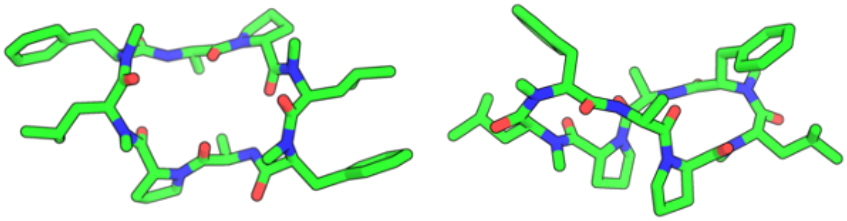

Figure 5.12. PJS-I-16 structure in different solvents.

The structure of PJS-I-16 was solved in two different solvents. At left is the structure of PJS-I-16 when in H2O-ACN, forming a circle shape. At right is the structure of PJS-I-16 when in EtOAc, forming a collapsed ellipsoid shape.

The peptide was docked into D_3_633 by superimposing the symmetry aces of the peptide and the protein and sampling the two remaining rigid body degrees of freedom (the translation along and rotation around the symmetry axes). Each successful docking attempt was then used for further design of the protein interface to maximize interactions with the ligand. By only changing eight residues in the pore area of each monomer the new design known as ‘D_3_633_8x’ was able to bind to a cyclic peptide with a great deal of affinity (somewhere between mid-nanomolar to low micromolar) and specificity to confirmation.

Crystal structures were solved of D_3_633_8x with and without the peptide. D_3_633_8x without the ligand added forms the desired ellipsoidal architecture with a cavity large designed for a C2 symmetric peptide with minimal deviations from the design (Ca rmsd 1.992 Å) (**Figure 5.13**). When the ligand is added, D_3_633_8x binds a single peptide in the designed pore, and the crystal structure exhibiting even fewer deviations from the design (Ca rmsd 1.717 Å) (**Figure 5.14**).

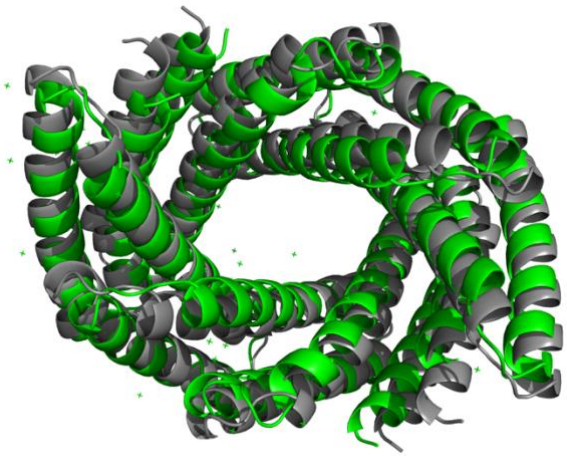

Figure 5.13. Alignment of crystal structure and Rosetta model of D_3_633_8x without ligand.

Rosetta design model of D_3_633_8x in gray. Crystal structure at 2.0 Å of D_3_633_8x (without the peptide) in green. Crystal structure aligns with Rosetta design model 1.992 Å Cα RMSD.

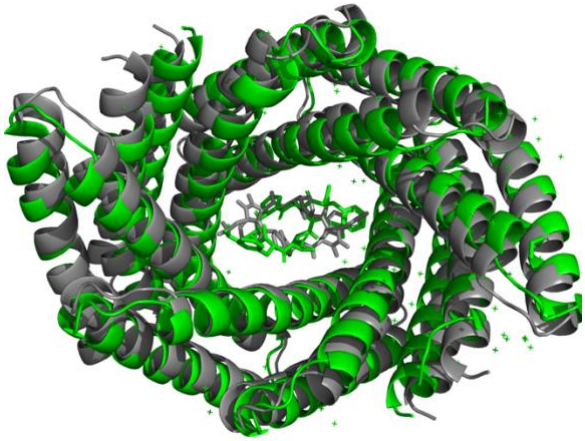

Figure 5.14. Alignment of crystal structure and Rosetta model of D_3_633_8x with ligand.

Rosetta design model of D_3_633_8x with peptide in gray. Crystal structure at 2.05 Å of D_3_633_8x with peptide in green. Crystal structure aligns with Rosetta design model 1717 Å Cα atom RMSD.

The peptide binding mode and contacts in the cTRP cavity is very different than predicted via the first round of computational docking and subsequent design. The peptide is flipped 180° around the x-axis in the pore as seen in **Figure 5.14** and has lost its C2 symmetry (**Figure 5.15**). The conformation of the peptide inside the cTRP resembled the EtOAc conformation except the omega bonds between the leucine and phenylalanine are not both cis, one of the bonds is in the trans conformation. The density of this conformation was undeniable (**Figure 5.8**), even when forced, the density would not accommodate both

bonds being cis. It is odd that a C2 symmetric peptide would not bind symmetrically into a protein pore that is also C2 symmetric, there is something that we are not yet appreciating in the design and optimization of the protein/ligand interface. The computational algorithm failed to sample the unique combination of protein conformation and bound peptide conformation that corresponds to the actual binding mode. But, the peptide is about the same size as the volume of the cavity, and is chemically somewhat 'appropriate' for association with the residues that line that pocket, so the program produced a solution that was reasonable, but not precisely correct.

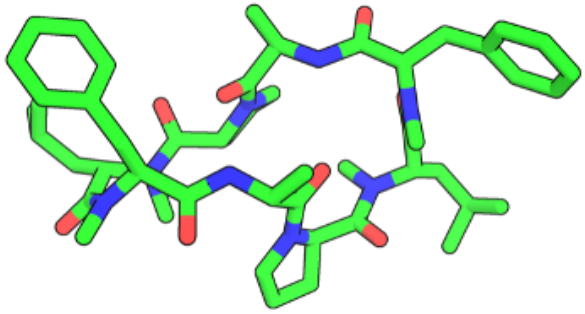

Figure 5.15. Peptide structure inside cTRP is no longer C2 symmetric.

The peptide structure while bound to the C2 symmetric cTRP looks very close to the EtOAc structure, but one NMeLeu-NMePhe omega bond is in the trans confirmation rather than the cis confirmation as expected.

When one considers the combinatorial, computational complexity of trying to accurately predict the binding mode, it's difficult to get it right the first time. The algorithm must sample the overall pocket dimensions (depends on accurate modeling of the protein backbone and helical packing), the protein side chain rotamers throughout the pocket, the backbone conformation of the peptide, and the side chain rotamers extending from the peptide. With lessons learned from the designing of D_3_633_8x, the team move on to new designs to adapt to an alternate cyclic peptide and the improvements to the design method increased the accuracy of the designs.

Two protein designs were chosen (MYS13 and SRH7), and their resulting crystal structures were closer to the indented size and shape of the C2 symmetric cTRP than D_3_633_8x with the RMSD for both being under 1 Å. A pipeline like the one used for D_3_633 was utilized to engineer MYS13 and SRH7 to bind cyclic peptides. MYS13 was optimized to bind a six-residue cyclic peptide, known as 70H, made up of repeating alanine, phenylalanine, and modified tryptophan, forming a ring shape in the center of the protein.

SRH7 was optimized to bind PJS-I-16. Unfortunately, MYS13 was unable to be crystalized in the presence of its peptide (**Figure 5.16**). In a few data sets, there was an unidentified density in the pore of the cTRP, but none were resolved enough to accurately build any part of the peptide into. SRH7, however, lead to a successful peptide bound structure (**Figure 5.17**).

Figure 5.16. Alignment of crystal structure and Rosetta model of MYS13.

Rosetta design model of MYS13 with peptide in gray. Crystal structure at 2.05 Å of MYS13 without peptide in green. Crystal structure aligns with Rosetta design model 0.856 Å Cα atom RMSD.

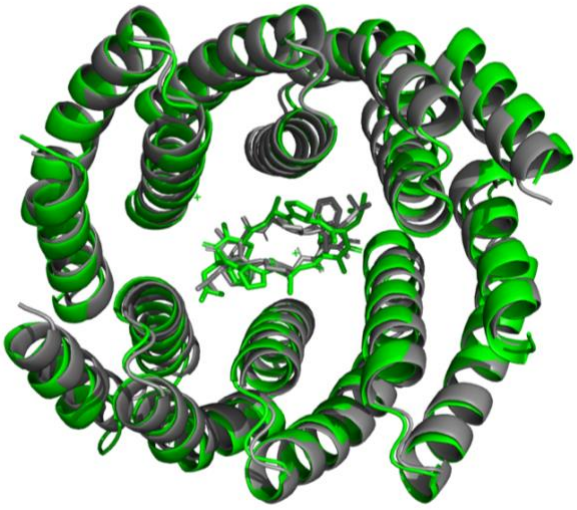

Figure 5.17. Alignment of crystal structure and Rosetta model of SRH7 with PJS-I-16.

Rosetta design model of SRH7 with peptide PJS-I-16 in gray. Crystal structure at 2.05 Å of SRH7 with peptide 16I in green. Crystal structure aligns with Rosetta design model 0.849 Å Cα atom RMSD.

As seen above, binding a small molecule as a dimer is a computational challenge, because the binding interface involves not just the protein but also the second small molecule, which has an independent set of rotational and translational degrees of freedom. In many cases, however, it is important that the molecule be bound in a particular way to convey the mechanism intended. Engineering proteins to bind photosynthetic proteins, such as chlorophyll, is a first step in developing new synthetic energy conversion technologies, including renewable fuel production, but we need to know more about their intricate function for accurate development of these solar-to-energy technologies. To control these degrees of freedom, we sought to optimize homodimers from the previous research to bind a C2-symmetric chlorophyll (Chl) pair such that the C2 symmetry axes of the protein and chromophore are coincident, similar to native reaction centers, which can have true C2 symmetry [192, 193] or pseudo-C2 symmetry. C2 symmetry ensures that the two bound Chl molecules will have near-degenerate site energies, improving the resonance between pigment transitions necessary to create delocalized states [207]. For Chl dimer protein scaffolds, we chose hyperstable C2-symmetric repeat protein dimers containing symmetric pockets with tunable sizes and geometries [129, 131, 183, 189]. In this dimeric repeat protein architecture (**Figure 5.18**), the hydrophobic core is independent from the small molecule binding site, enabling full customization for binding with little impact on the overall protein structure. Several thousand C2-symmetric homodimers that sample a wide range of superhelical curvature, rise, and radius parameters have been generated [131, 183].

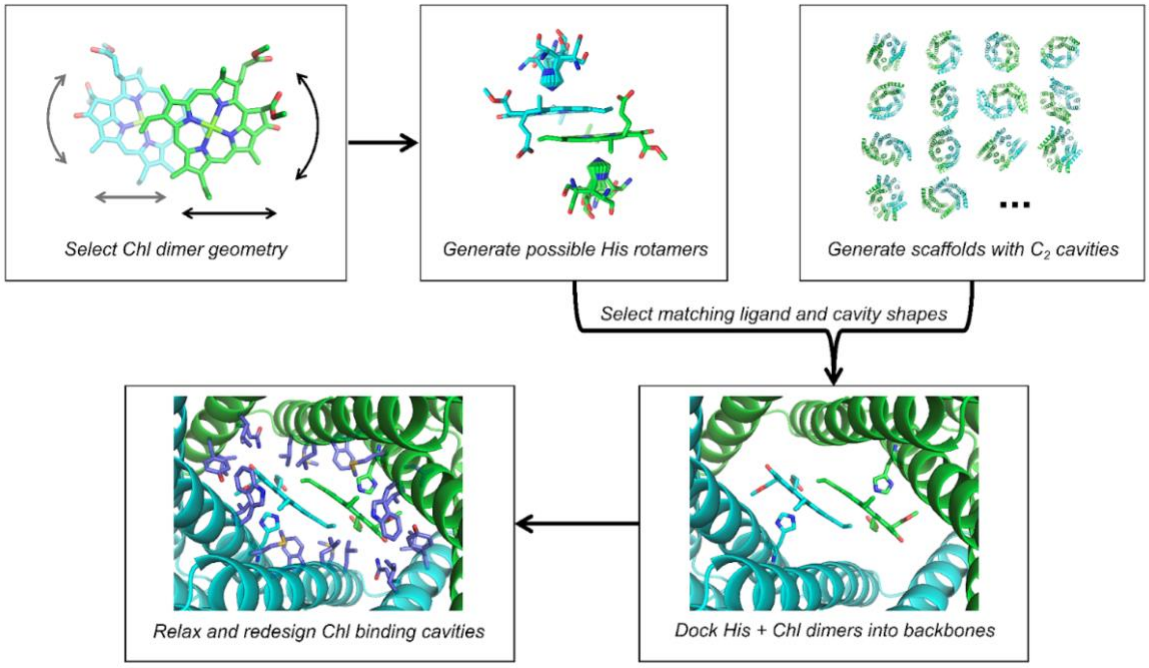

Figure 5.18. Design Pipeline of Chlorophyll Ligand Binders. [182]

Computational design of Chl special pair proteins begins with selection of a Chl dimer geometry and generation of inverse His rotamers. His-Chl dimers are docked into designed homodimers, and the Chl binding pockets are redesigned using Rosetta FastDesign. Adapted from Figure 1c from the accompanying paper [182].

To probe the effect of geometry on Chl-Chl coupling, we set out to design a range of C2-symmetric dimers that hold two closely interacting Chl molecules in varied geometries including the arrangement found in native special pairs. In native proteins, (B)Chls typically have a pentacoordinate central Mg(II) or Zn(II) ('ZnPPaM' ligand) ion with a histidine (His) Nε atom as the axial ligand. For each chosen special pair geometry, we built a His rotamer interaction field and stored the possible His-Chl interaction geometries in a hash table (**Figure 5.18**). For each geometrically compatible C2 scaffold, we cycled through His-Chl rotamers from the hash table, aligned them to the scaffold C2-symmetry axis, and searched for matches of the His N-Cα-C backbone atoms to the backbone atoms of the residues lining the binding cavity. Scaffolds for which the His N-Cα-C backbone atoms aligned with corresponding atoms in the protein backbone, and which could accommodate the Chl dimer without clashes, were redesigned using symmetric Rosetta FastDesign to optimize hydrophobic packing and hydrogen bonding around the Chls [136, 194, 208-210] (**Figure 5.18**). Designs were filtered based on the Rosetta full-atom energy, the solvent-accessible surface area of the Chl dimer (DSasa), His rotamers, and His Nε-metal ligation geometry. We selected 43 designs

based upon 13 unique scaffolds for experimental characterization (see paper [182] for amino acid sequences). The protein monomer sizes range from 20.6 to 28.4 kDa (179 to 261 amino acids). We refer to these 48 designs as Chl Special Pair proteins, or SP for brevity.

Following SP protein expression in E. coli, SDS-PAGE gels showed that all 48 designs were present in the soluble fractions of lysates. Proteins were purified by Ni-NTA and size-exclusion chromatography (SEC). All SEC traces exhibited protein absorption at the elution volume expected for homodimer formation. Of 20 designs investigated by Small Angle X-ray Scattering (SAXS) in the apo-state, 15 had SAXS profiles suggesting a 3-dimensional shape consistent with the design model.

Based on the results of SEC, SAXS, and spectroscopic experiments [182], we selected promising candidates for X-ray crystallographic structure determination. We solved the crystal structures of 3 designs, SP1, SP2, and SP3x (a close relative of SP3 spoken of in the corresponding paper [182]), and found that all three had protein backbone conformations that matched the corresponding design models to within 1.7 Å Cα RMSD.

The x-ray crystal structure of SP1 was solved in the ZnPPaM-bound state to 2.0 Å resolution, revealing a special pair geometry closely matching that of purple photosynthetic bacteria (**Figure 5.19, Figure 5.20a-f**). The rotameric state of the Zn-ligating His121 is identical to that in the design model, and several hydrophobic and T-stacking interactions form as designed. Hydrogen bonds to the ring E ketone group, shown to be important for modulating special pair redox potentials [211], form with Gln10 in both ZnPPaM molecules. Alignment of the tetrapyrrole rings of the SP1-ZnPPaM dimer to nine native BChl a special pairs from different species of purple bacteria gave RMSDs of 0.23-0.28 Å [199, 200, 202-204, 206]. For comparison, the special pairs of two crystal structures of the same Thermochromatium tepidum LH1-RC complex deviate from one another by 0.22 Å RMSD across the tetrapyrrole rings (PDB IDs: 3WMM and 5Y5S) [200, 206]. The RMSD between the ZnPPaM dimer in the SP1 crystal structure and its design model is 0.25 Å.

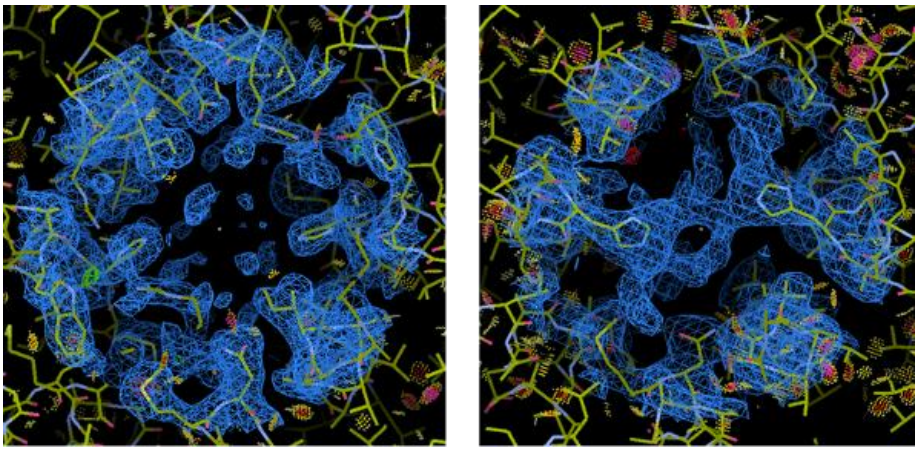

Figure 5.19. Comparable density between SP1 unbound and ligand bound.

Electron density of SP1 shows clear presence of the ligand (right) in comparison to the unbound structure (left). Crystal structure was solved first by placing all protein atoms and waters, avoiding the clear electron density in the center of the pore. Then the ligand was built into the structure last. One can clearly see the two histidine residues which coordinate the metal of the ligand.

SP2 was intended to assemble a ZnPPaM dimer with a conformation significantly different from native special pairs to investigate the effect of dimer geometry. The SP2 crystal structure was solved in both the apo-state and the ZnPPaM-bound state to 2.4 and 2.5 Å resolution, respectively. The apo- and holo-state amino acid backbones both agree with the SP2 design model to within 1.4 Å RMSD (**Figure 5.20g**). The holo-state crystal structure has two copies of the SP2 dimer in the asymmetric unit; alignment of the two ZnPPaM dimers shows their binding geometries are equivalent, with an RMSD of 0.22 Å over the tetrapyrrole rings. The ZnPPaM molecules are ligated by His178 as in the SP2 design model. After alignment of the crystal structure and design model protein backbones, the corresponding tetrapyrrole rings are approximately coplanar. Despite the accuracy of the protein backbone design, the crystal structure shows the ZnPPaM molecules are rotated and translated relative to the design model (**Figure 5.20h**). Compared to the apo-state crystal structure, the SP2 binding cavity widens by ~1.6 Å in the presence of ZnPPaM; this expansion provides the extra volume needed for the ZnPPaM molecules to adopt their unexpected conformation. While the ZnPPaM dimer in SP2 differs from the design model, the crystal structure nevertheless satisfies the objective of creating a non-native dimer geometry.

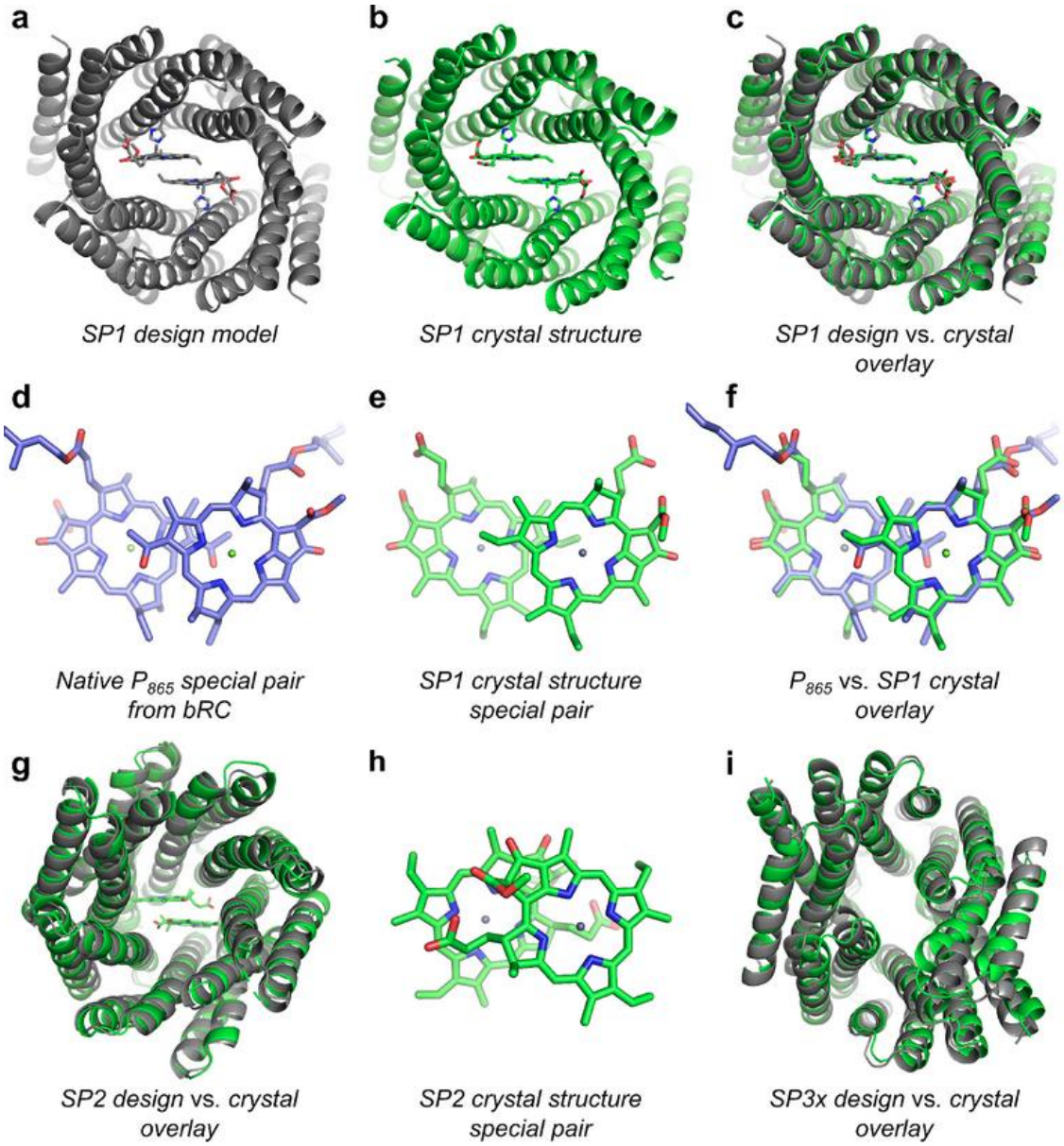

Figure 5.20. Crystal structures of designed SP proteins. [182]

(a) Rosetta design model of SP1. (b) SP1 crystal structure at 2.0 Å resolution with ZnPPaM molecules bound (PDB ID: 7UNJ). (c) SP1 Rosetta design model (gray) aligns to SP1 crystal structure (green) with 1.6 Å Cα atom RMSD. (d) The BChl a special pair from a 2.5 Å resolution cryo-EM structure of the purple bacterial RC-LH1 complex (Rhodobacter sphaeroides) (PDB ID: 7PIL) (Qian et al., 2021). (e) The ZnPPaM dimer from the SP1 crystal structure shown in panel (b). (f) The ZnPPaM dimer of the SP1 crystal structure (green) aligns with the native purple bacterial special pair (blue) to 0.23 Å RMSD across corresponding atoms of the tetrapyrrole rings. (g) Rosetta design model of SP2 (gray) aligns to the holo-state SP2 crystal structure (green, PDB ID: 7UNI) with 1.4 Å Cα atom RMSD. (h) The ZnPPaM dimer in the SP2 crystal structure (green) deviates from the predicted dimer geometry (not shown; 3.5 Å RMSD). (i) Rosetta design model of SP3x (gray) aligns to apo-state SP3x crystal structure (green, PDB ID: 8EVM) with 1.6 Å Cα atom RMSD. Adapted from Figure 3 from the accompanying paper [182].

5.3 De novo design of self-assembling protein hetero-oligomers

This section's written material is taken verbatim from a paper in preprint on bioRxiv Kibler et al. 2023 [212] and a paper in preparation Lee et al. 2023 [181], in which I am a coauthor, with explicit permission from the authors and expanded upon to reflect my contributions. I performed almost all the x-ray crystallographic work for these projects. Additional description of the protein design process can be found in the associated papers [180, 181].

To being to tackle the challenge of developing self-assembling protein hetero-oligomers, the Kibler and Lee teams broke down the problem into smaller steps [180]. By exploiting hydrogen bond networks ("HBNets") in the inter-subunit faces, specific unique contacts can be made. Interfaces that utilize HBNets have greater specificity than hydrophobic interactions alone, so using HBNets can increase the on-target binding of the oligomer [173]. Already there has been success at using HBNets to create heterodimeric interfaces and that knowledge was leveraged for this project [174]. To design hetero-oligomeric cTRPs a stepwise strategy to create unique HBNets at each interface was developed [180]. For additional information regarding the design pathway, refer to the primary citation [180].

Kibler et al. symmetrically re-designed the interfaces of the parent homo-oligomeric cTRPs, and then recombined experimentally validated designs to form pseudosymmetric hetero-oligomers [180]. This method generated stable hetero-oligomers that have up to three unique subunits that oligomerize with specificity to their associated partner. These are validated according to size exclusion chromatography (SEC), native mass spectrometry (nMS), and SEC-Multi-Angle Light Scattering (SEC-MALS) [180]. Additional small-angle X-ray scattering (SAXS) and negative-stain electron microscopy (EM) experiments confirmed the design models [180]. Crystal structures were also solved, and the designs confirmed.

This generated approach can be used to design a greater range of psedosymmetric oligomers. In the future, this strategy could be used to control cell signaling pathways or to develop advanced protein materials. As proof of concept, these psedosymmetric trimers were used to create enormous nanocages [181].

5.3.1 *Methods*

Construction of synthetic genes. All synthetic genes were ordered from either Integrated DNA Technologies Inc. (Coralville, IA, USA) (IDT) or Genscript Inc. (Piscataway, NJ, USA). Genes ordered from IDT were reverse translated and codon optimized using Domesticator (https://github.com/rdkibler/domesticator_3). Protein tags were added to aid in purification (His6-tag) or identification (EHEE_rd2_0005 (aka EHEE)9, superfolderGFP, or mScarlet-I) were separated from the designed protein with a tobacco etch virus protease (TEVp) cleavage site (ENLYFQG). In many cases, stop codons were introduced at the end of the inserted gene to prevent incorporation of the vector's built-in tags. Cloning into pET29b+ or its derivatives is always done at the NdeI/NcoI site.

For expression of BGLs, protomer sequences were extracted and the n-terminal tag "MGHHHHHHGSENLYFQGWS" was added. Codon-optimized genes with a 3' stop codon were cloned into pET29b+ by IDT. See Supplemental Table 1.1 in associated paper for full amino acid sequences. For co-expression of hetBGLs, all three proteins were arranged sequentially off the same transcript, separated by ribosome binding sequences (RBSs) (first: TAAGAAGGAGATATCATCATG; second: TAAAGAAGGAGATATCATATG). One protomer received superfolderGFP, another received EHEE, and the third received nothing. One protomer (with either superfolderGFP for the EHEE) received a His6-tag. See Supplemental Table 1.3 of associated paper for full amino acid sequences. These were ordered from IDT.

Transformation and expression. For single plasmid expressions, plasmids (100ng) were transformed into chemically competent E. coli expression strain BL21(DE3)Star (Invitrogen) for protein expression following manufacturer's protocol, with the exception of using 10ul competent cells per reaction. Following transformation and recovery, the entire transformation products were used to inoculate 1 mL Luria-Bertani (LB) medium containing 100 ug/mL kanamycin and grown at 37°C with shaking at 225 rpm overnight. 500ul of overnight cultures were diluted into 50 mL TBM-5052 supplemented with 100 ug/mL kanamycin in 250 mL baffled flasks, and incubated at 37°C with shaking at 225 rpm for 18-24 hours.

Immobilized metal affinity chromatography (IMAC). Cultures were harvested by centrifugation at 4000 rcf for 10 minutes, culture supernatant decanted, and pellets resuspended to 30 mL in Lysis buffer. 300ul PMSF (100mM in 100% EtOH) is added immediately prior to sonication at 70% power for 5 minutes.

“Lysate” fractions are saved, and then lysates were clarified by ultracentrifugation at 14,000 rcf at 12°C for at least 30 minutes and applied to 1.5 mL Ni-NTA resin (Qiagen) pre-equilibrated with Lysis buffer and packed into Econo-Pac columns (Bio-Rad) for gravity chromatography. The columns were washed twice with 15 mL Wash buffer and eluted with 10 mL Elution buffer.

Size-exclusion chromatography (SEC). Samples were concentrated using 10k MWCO spin concentrators and were purified using a Superdex 200 10/300 increase column (Cytiva) in SEC buffer using an ÄKTA pure system (Cytiva). SEC traces were also used to qualitatively determine homogeneity and quantitatively measure total yield by A280 absorbance integrated over the collected fractions using Unicorn (Cytiva).

TEVp cleavage. Purification and mass tags were buffer exchanged into TEV buffer and cleaved with TEVp at a ratio of 1 mg TEV per 100 mg substrate for 24-72h at room temperature. After TEV cleavage, samples were exchanged into Lysis buffer and passed over a Ni-NTA gravity column and washed with 10ml lysis buffer. Flowthrough was collected, concentrated using 10k MWCO spin concentrators, and purified once again by SEC. Anion exchange chromatography (AEC) TEVp-cleaved samples of hetBGL03-15-18 intended for crystallization remained contaminated with GFP due to an oversight that meant GFP did not have a His6-tag. To remove the GFP, samples were exchanged into AEC buffer A and loaded onto a HiScreen Q FF (Cytiva) column pre-equilibrated in AEC buffer A using an ÄKTA pure system. A gradient of AEC buffer B was applied with pauses at 15% (150 mM NaCl) and 25% (250mM NaCl) to elute the GFP and hetBGL03-15-18, respectively. Separation was measured by differential absorbance at 480 nm (GFP absorbance) and 280 nm, and SDS-PAGE.

SDS-PAGE. Samples were diluted 1:1 with 2x Laemmli Sample Buffer (Bio-Rad) without Beta-mercaptoethanol and 15ul were loaded onto AnykD™ Criterion™ TGX™ Precast Midi Protein Gels (Bio-Rad). Ladder was 10ul of Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad). Gels were run at 300V for 18 minutes, then stained using an eStain™ L1 Protein Staining System (Genscript). Stained gels were imaged using a ChemidocXRS+ (Bio-Rad).

Crystallography sample preparation, phasing, and refinement. Crystals of BGL06, BGL14_styr, BGL15, BGL18, and hetBGL03-15-18 were grown using protein purified as described above and TEVp cleaved and optionally AEC purified. Protein samples dispensed in 1 uL drops at purification concentrations

were mixed with equal volume of a crystallization solution and set in hanging drops (refer to **Table 5.8** for conditions). Vapor phase equilibration of the resulting drops against a 1 mL reservoir of the same crystallization solution resulting in growth of crystals. The crystals were flash cooled in liquid nitrogen after transfer into a cryoprotective solution (refer to **Table 5.8** for conditions). Diffraction data were collected on a Pilatus areas detector at the Advanced Light Source (ALS) synchrotron facility at beamline 5.0.2 for BGL06, BGL14_styr, BGL15, and BGL18. Diffraction data were collected on a Rigaku HyPix-6000HE hybrid photon counting detector at the Fred Hutch for hetBGL03-15-18. The resulting data sets (**Table 5.8**) extend to 2.1 Å, 3.0 Å, 3.3 Å, 3.0 Å, and 2.1 Å resolution for BGL06, BGL14_styr, BGL15, BGL18, and hetBGL03-15-18, respectively. Most data had complete trimers within the asymmetric unit (three copies of a protein subunit), with exceptions to BGL18 and BGL15 which had 2 and 4 trimers in the asymmetric unit, respectively. Data was processed using the program HKL2000 [71] or Aimless [198]. The placement of subunits was determined using the molecular replacement algorithm in program PHENIX [72]. Local rebuilding of all constructs was performed using the program COOT [74], followed by refinement using the program PHENIX [72]. The final values for Rwork / Rfree are notated in **Table 5.8**.

Table 5.8. Hetero-oligomer crystallographic data and refinement statistics

Design	**BGL06**	**BGL14_styr**	**BGL15**	**BGL18**	**hetBGL03-15-18**
Crystallography					
Well solution	20% (w/v) PEG-1000 phosphate-citrate pH 4.2, 200 mM Li2SO4	30% (v/v) PEG-400, Tris pH 8.5, 200 mM MgCl2	30% (v/v) PEG-400, CAPS pH 10.5	20% (w/v) PEG-3350, 0.1 M Bis-Tris pH 6.5, 0.2 M MgCl2	15% (w/v) PEG-3350, 0.1 M Magnesium formate dihydrate
Cryoprotectant	All flash-frozen in well solution plus 20% ethylene glycol				
Data Statistics					
Space group	$P2_1$	$P2_1$	$P3_2$	$P2_1$	$P3_221$
Unit cell dimensions	a=58.2 Å, b=78.9 Å, c=58.7 Å, b=118.3°	a=105.8 Å, b=83.2 Å, c=119.8 Å, b=93.0°	a=b=97.8 Å, c=133.9 Å	a=58.9 Å, b=61.7 Å, c=73.7 Å, β=100.2°	a=b=59.3 Å, c=260.3 Å
Resolution (Å)	2.1 (2.14-2.1)	3.3 (3.39-3.33)	3.0 (3.05-3.0)	2.1 (2.18-2.1)	3.0 (3.05-3.0)
Reflections	27320	30779	28352	30567	10082
Completeness (%)	99.8 (99.7)	99.8 (100)	99.1 (96.8)	98.7 (99.6)	88.4 (53.6)
Redundancy	6.4 (6.7)	6.4 (6.4)	10.3 (9.8)	9.9 (9.9)	8.4 (3.1)
I/s(I)	56.2 (16.2)	11.0 (1.3)	10.9 (1.3)	27.1 (6.0)	28.5 (1.05)
Rmerge	0.061 (0.132)	0.084 (1.682)	0.192 (1.224)	0.057 (0.354)	0.110 (0.660)
Rpim	0.026 (0.054)	0.033 (0.659)	0.063 (0.412)	0.019 (0.118)	0.038 (0.397)
CC1/2	1.002 (0.998)	0.999 (0.524)	0.999 (0.742)	1.000 (0.969)	1.003 (0.664)
Refinement Statistics					
PDB ID	8E0L	8E12	8E0M	8E0N	8E0O
Resolution (Å)	2.1 Å	4.0 Å	3.0 Å	2.1 Å	3.0 Å
Rwork / Rfree	0.210 / 0.255	0.280 / 0.328	0.207 / 0.259	0.238 /0.262	0.282 / 0.321
No. atoms					
Protein	3571	10690	7569	3689	3074
Water	110	0	13	55	1
Wilson B-factor	35.65	141.65	63.77	26.36	100.57
RMS deviations					
Bond length	0.002 Å	0.002 Å	0.002 Å	0.001 Å	0.002 Å
Bond angles	0.38°	0.52°	0.52°	0.35°	0.656°
Ramachandran Distribution (Favored%/Allowed%/ Outlier%)	99.38 / 0.62 / 0.0	97.69 / 2.05 / 0.26	99.31 / 0.69 / 0.0	98.99 / 1.01 / 0.0	98.32 / 1.26 / 0.42

Crystals of BGL17_A31 were grown using protein purified as described above and TEV cleaved. Protein samples dispensed in 1 μL drops at purification concentrations were mixed with equal volume of a

crystallization solution and set in hanging drops with 100mM ammonium citrate tribasic pH 7.0 and 10% w/v polyethylene glycol 3350. Vapor phase equilibration of the resulting drops against a 1 mL reservoir of the same crystallization solution resulting in growth of crystals. The crystals were flash cooled in liquid nitrogen after transfer into a cryoprotective solution of well solution plus 20% ethylene glycol. Diffraction data were collected on a Pilatus areas detector at the Advanced Light Source (ALS) synchrotron facility at beamline 5.0.1. The resulting data set (**Table 5.9**) extend to 4.5 Å with one copy of the protein in the asymmetric unit. The greater homotrimer can be generated via application of a crystallographic symmetry axis.

Data was processed using program HKL2000 [71]. The placement of subunits was determined using the molecular replacement algorithm in program PHENIX [72]. Local rebuilding was performed using the program COOT [74], followed by refinement using the program PHENIX [72]. There was a shift in the α-helices from the design to the crystal structure that required a rigid body fit of two helical bundles at a time to build into the density. The final values for Rwork / Rfree were 0.301 / 0.344 with good geometry (**Table 5.9**).

Table 5.9. BGL17_A31 crystallographic data and refinement statistics

Data Statistics	
Publication name	LK031
Space group	I23
Unit cell dimensions	a=b=c=153.47 Å, a=b=c=90°
Resolution (Å)	4.5 (4.5-4.66)
Reflections	3609
Completeness (%)	97.4 (94.5)
Redundancy	18.7 (13.4)
I/s(I)	30.25 (2)
Rmerge	0.087 (0.821)
Rpim	0.021 (0.230)
CC1/2	0.996 (0.804)
Refinement Statistics	
PDB ID	8FLX
Resolution (Å)	4.5
Rwork / Rfree	0.301 / 0.344
No. atoms	
Protein	342
Water	0
Wilson B-factor	247.83
RMS deviations	
Bond length	0.002 Å
Bond angles	0.50°
Ramachandran Distribution (Favored%/Allowed%/Outlier%)	93.82 / 5.88 / 0.29

5.3.2 *Results*

From a parent design made up of three identical monomers, BGL0 (tcTRP9_sub3 from [135]), 20 modified designs were created, expressed, and purified for further testing. Additional information on the designing of the 20 cTRPs can be found in the associated paper [180].

We obtained crystal structures of BGL06 from the HR-C set (**Figure 5.21a**), BGL14 and BGL15 from the HR-N set (**Figure 5.21b,c**), and BGL18 from the NM set (**Figure 5.21d**). At the backbone level, the crystal structures show good agreement with the design models (1.2 Å, 1.3 Å, 0.7 Å, and 1.4 Å CA-RMSD, respectively), as well as to the parental BGL0 crystal structure (TM-scores 0.94, 0.82, 0.81, and 0.97, respectively, to PDB 6XR2). Sidechains at the interfaces of BGL06, BGL14, and BGL18 are well enough resolved to determine that HBNets in all three structures are formed correctly. BGL06 has an

interchain bidentate HBNet involving Gln14 and Gln146, each backed up with second shell interactions from Ser18 and Ser125, respectively (**Figure 5.21a**). The string of hydrogen bonds in the BGL14 interface alternates between chains, with a fully satisfied Asn13 in the very middle which donates a hydrogen bond to His153. BGL15 has three copies of the ring in the unit cell, but the low resolution of the dataset (4.0 Å) obscured details about the HBNet residue sidechains. BGL18 has two copies of the ring in the unit cell and all six interfaces were resolved and highly similar to each other. The other copy of the trimer in the unit cell is 1.6Å CA-RMSD to the design model. It has a generally polar interface with a meandering HBNet spanning Asn10, Thr17, Asn126 (which may be donating a hydrogen bond to the backbone oxygen of Asn10), Thr152, Ser155, and Ser159.

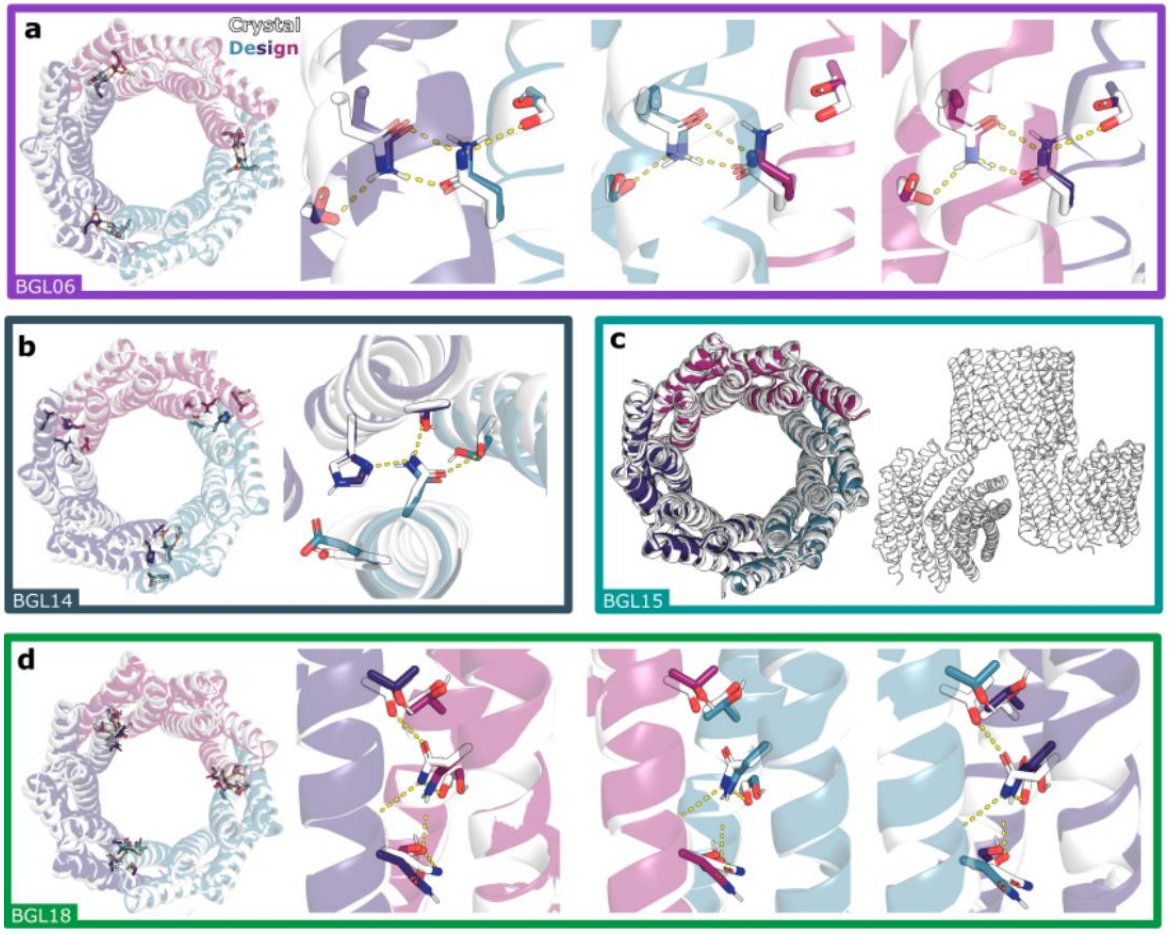

Figure 5.21. Crystal structures of BGL homotrimers. [180]

(a,b,d) Crystal structures (white) of four different BGL homotrimers are superimposed onto the design models which are colored in teal, purple, and magenta. (left) The top-down view of the full ring with HBNets highlighted and visible through transparent cartoon models. The backbones of all crystal structures match well with the design models (<1.4 Å RMSD over all Cα atoms). (Right) Close up view(s) of each HBNet well enough resolved to build full side chains. (a) BGL06 (2.1 Å resolution; RMSD = 1.2 Å over all Cα atoms; PDB: 8E0L) and (d) BGL18 (3.0 Å resolution; RMSD = 1.4 Å over all Cα atoms; PDB: 8E0N) were both well enough resolved to model all three interfaces. BGL18 contained two copies of the full trimer in the crystal unit cell, which are structurally similar to each other (RMSD = 0.3 Å over all Cα atoms between copies). (b) One interface of BGL14 (3.0 Å resolution; RMSD = 1.3 Å over all Cα atoms; PDB: 8E12) was well resolved. (c) BGL15 (4.0Å resolution, PDB: 8E0M) had three copies of the trimer in the unit cell and all three were close to the design model (0.63 Å, 0.64 Å, and 0.66 Å RMSD over Cα atoms). See **Table 5.8** for crystallographic details. Adapted from Figure 3 from the accompanying paper [180].

We tested two different strategies for choosing interfaces for recombination. First, to attempt to maximize binding specificity, we constructed a set of 6 recombinants (called “Set 1”) with one interface from each of HR-N, HR-C, and NM. Second, to generate heterotrimers with nearly identical subunit backbone structures, we recombined the original design BGL0 with validated homo-oligomers from the less-perturbed

NM set (BGL17, BGL18, and BGL19). From the four homo-oligomers we enumerated all eight possible combinations (called “Set 2”).

Both sets of designed heterotrimers (total 14) were ordered as tricistronic constructs with TEVp-cleavable GFP or miniprotein27 tags on two of the chains to differentiate their masses by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and nMS, and the third chain having a His6-tag for affinity co-purification (**Figure 5.22a**). We found that 11/14 co-purified with three distinct bands on SDS-PAGE, with 9 showing roughly stoichiometric ratios by gel densitometry (**Figure 5.22b**). The SEC profiles of eight (2 from Set 1 and 6 from Set 2) were largely monodisperse and had the expected retention volume (**Figure 5.22c**). nMS analyses of these eight designs showed that each was the intended ABC species (consisting of the three unique chains: A, B, and C) without any of the many possible off-target combinations of subunits (AAB, BBB, etc; **Figure 5.22d**). Each subunit of a heterotrimer from Set 1 has a different number of helices due to the use of interfaces from the HR-N set with interfaces from the HR-C or NM sets.

To make them more like the subunits of heterotrimers from Set 2, which have the same number of helices in each subunit, we deleted the loop of the N-terminal outer helix in the HR-N-derived interface and reconnected it at the C-terminus of the next protomer (essentially converting the HR-N interface to HR-C); 2/3 such rearrangements remained clean heterotrimers, which brought the number of validated hetBGLs to 10.

We solved the crystal structure of hetBGL03-15-18, a heterotrimer from Set 1 composed of the interfaces extracted from BGL03, BGL15, and BGL18, at 2.1 Å resolution (**Figure 5.22e**). The structure superimposes to the design model with 1.2 Å CA-RMSD. The three distinct subunits have an average pairwise sequence identity of 62.3% but high structural similarity (aligned CA-RMSD of 1.5 Å), showing that this method can produce highly (pseudo)symmetric hetero-oligomers. The BGL18-derived HBNet is very similar to that in the homotrimer crystal structure, and the HBNet of BGL15, which was not resolved in the homotrimer crystal structure, is confirmed to be in the correct conformation, with the addition of a hydrogen bond between Ser140 and a crystallographic water. The previously unsolved BGL03 interface differs from the design model due to fraying of the C-terminal outer helix likely due to reduction of hydrophobic packing stemming from polar HBNet residues placed too close to the C-terminus; placement of HBNet residues too close to the termini was avoided in future designs.

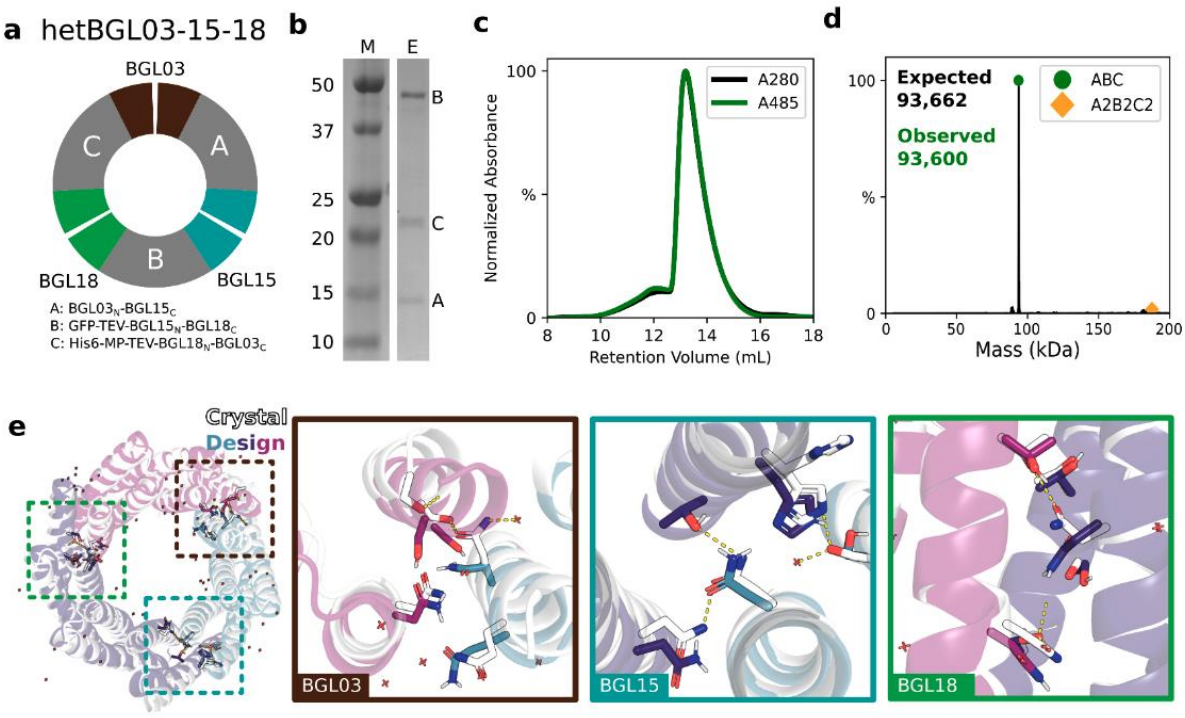

Figure 5.22. Characterization of hetBGL heterotrimers. [180]

(a-e) Results of hetBGL03-15-18. (a) (top) Schematic showing the result of recombination between BGL03, BGL15, and BGL18. (bottom) The expression constructs for each chain as expressed tricistronically. “GFP” is superfolderGFP, included to add a large amount of mass and provide an additional spectroscopic identification, and “MP” is EHEE_rd2_000527, a stable and small protein included to add a small amount of mass. (b) SDS-PAGE of IMAC elution. M: protein ladder. E: IMAC elution. Protein size affects band staining, so it is difficult to judge stoichiometry from the band darkness/size, but the correct trend in stain density is observed for equal stoichiometry. (c) SEC trace overlaying normalized absorbances at A280 and A485. (d) Mass deconvoluted nMS data showing the intended major species (green circle, ABC) and a minor species with twice the mass, likely to be a dimer of trimers (orange diamond, A2B2C2). (e) (left) Crystal structure in white (2.1 Å resolution, PDB: 8E0O) and design model in teal, purple, and magenta overlaid to show global agreement (CA-RMSD: 1.2 Å, TM-score: 0.96). Dashed boxes indicate approximate locations of zoomed-in views of the interfaces derived from BGL03, BGL15, and BGL18, showing side chains of HBNet residues and their inferred hydrogen bonds as yellow dashes. Adapted from Figure 3 from the accompanying paper [180].

The pseudosymmetric proteins can also be used to design protein nanocages, and BGL17_A31 is a step in that direction. Already high triangulation (T) number cages have been designed using psedosymmetric heterotrimers [213], but the specificity of the T number and stoichiometry was not consistent. Until recently, it was not possible to design a heterotrimer with three distinct monomers that oligomerizes consistently [171, 180]. cTRPs meant for nanocage designs were equipped with arms radiating from the central pore structure.

The design of BGL17_A31 started from the same BGL family [135, 171, 180]. Initially, over 2000 different armed BGLs were designed using HelixFuse [214] and Rosetta [136]. From those designs, 39 of

them were chosen to experimentally characterize. After solubility testing, size-exclusion chromatography (SEC) to test for monodispersity, and negative-stain electron microscopy to validate the size and shape of the design, 22 constructs remained. Only one of the constructs was crystallized, BGL17_A31 and interestingly was closer to the AlphaFold design than the design model. The RMSD between the design model and the crystal structure was 4.322 Å (**Figure 5.23**). It demonstrates the successful design of a hetero-oligomeric armed structure. Additional cryoEM characterization was used to validate the other structures and nanocage formations which can be found in its accompanying paper that is in preparation Lee et al.

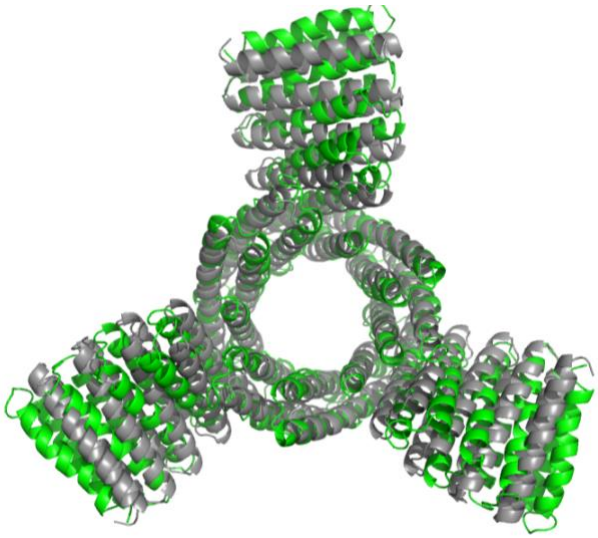

Figure 5.23. Alignment of BGL17_A31 crystal structure with Rosetta model.

Rosetta design model of BGL17_A31 in gray. Crystal structure at 4.5 Å of BGL17_A31 in green. Crystal structure aligns with Rosetta design model 4.322 Å Cα atom RMSD.

5.4 CONCLUSIONS

The parent *de novo* designed circular tandem repeat proteins (cTRPs) [128, 129] were initially designed following naturally occurring tandem repeat proteins [215-217] which are known to specifically bind a wide range of dynamic ligands such as DNA, peptides, and unique structural motifs. As shown in the research discussed, cTRPs have the advantage of being highly soluble, thermally stable, easily assembled (with and without the addition of protein staples [128]), extendable to accommodate different size pores, and able to tolerate the addition of other functional protein domains and ligands. The application of these proteins are endless, from high-avidity and/or high-affinity binding partners [128, 156, 180, 183], to cell stimulatory

factors [128], to display scaffolds for vaccine development [191, 218]. It is not hard to imagine placing other functional proteins to these highly stable cTRPs for things such as bioremediation [219] or utilizing the cTRPs' assembling properties to create a chain reaction of sequential protein activities [220, 221]. The ability of the cTRPs to accommodate other functional enzymes was proven possible due to the additional work performed by Hallinan et al [135]. In this work, the scientists augmented the trimer cTRP design with a three-fold arrangement of an anti-SARS CoV-2 VHH domain [222] showing the ability of these laden cTRPs to act as scaffolds despite carrying cargo attached to the N- or C-terminus of each subunit, even in some cases performing better in target binding affinity and biological activity assays [135]. The self-assembly, driven by disulfides or larger specific interfaces, further equips these cTRPs to be applied to different biotechnological problems.

Along with carrying cargo, these cTRPs have been optimized to contain a ligand within their pore. The new cTRP designs have proven that they can be designed to selectively bind to one specific ligand of interest over others. Sometimes, the protein forces the ligand into a conformation not previously seen, however, with additional design consideration, the cTRP can bind the peptide in the desired geometry. The cTRP ligand binders are C2 symmetric homodimers with central pores that can be optimized to accommodate vast differences in shape and charge (due to the pores separation from the protein core), allowing for many options for ligands and additional functionalization [131, 182, 183]. This is of particular interest for forming chemically inducible dimerization or oligomerization systems amongst other applications discussed in the introduction. Just like previous cTRPs, these C2 dimers have high thermal stability and solubility even when exposed to a wide range of buffer conditions. The design strategy spoken of in this thesis and in the companion papers, can be expanded to include creating binding pockets of higher order symmetries to accommodate more than just C2 symmetric molecules [131, 182, 183].

With the end goal of creating ligand responsive, enzyme carrying, and oligomerizing scaffolds, *de novo* hetero-oligomers were created. The main difficulty of designing hetero-oligomeric proteins the needs for a different interface at each juncture of the oligomer to have it assemble correctly. Kibler, Lee, and their teams have successfully designed some of the first *de novo* hetero-oligomers by harnessing the power of hydrogen bond networks. These constructs can then be further optimized for pseudosymmetric nanomaterials or diverse hubs for target clustering in cell signaling.

Utilizing all this knowledge can create unique and highly specified functional proteins. Take for example the ligand binders. The ligand binding interface can first be optimized in the homo-oligomeric state and then those modifications can be transferred to the hetero-oligomeric scaffold, which can then be optimized for assembly in only the presence of the specific ligand. Additionally, once those states are optimized, cargo can be tethered to the oligomer and further enhance the ability of the designed protein, effectively bringing the cargo to the location of the unique ligand to execute its function. The applications of these protein systems are endless and have benefits from clinical settings to biotechnology.

REFERENCES

1. Kennedy, M.A., et al., *Structures, activity and mechanism of the Type IIS restriction endonuclease PaqCI.* Nucleic Acids Res, 2023.
2. Lopatina, A., N. Tal, and R. Sorek, *Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy.* Annu Rev Virol, 2020. **7**(1): p. 371-384.
3. Jurenas, D., et al., *Biology and evolution of bacterial toxin-antitoxin systems.* Nat Rev Microbiol, 2022. **20**(6): p. 335-350.
4. Bernheim, A. and R. Sorek, *The pan-immune system of bacteria: antiviral defence as a community resource.* Nat Rev Microbiol, 2020. **18**(2): p. 113-119.
5. Loenen, W.A., et al., *Type I restriction enzymes and their relatives.* Nucleic Acids Res, 2014. **42**(1): p. 20-44.
6. Pingoud, A., G.G. Wilson, and W. Wende, *Type II restriction endonucleases--a historical perspective and more.* Nucleic Acids Res, 2014. **42**(12): p. 7489-527.
7. Rao, D.N., D.T. Dryden, and S. Bheemanaik, *Type III restriction-modification enzymes: a historical perspective.* Nucleic Acids Res, 2014. **42**(1): p. 45-55.
8. Koonin, E.V. and K.S. Makarova, *Orgins and evolution of CRISPR-Cas systems.* Philosophical transactions of the Royal Society of London, 2019. **Series B**(Biological sciences 374): p. 20180087.
9. Egido, J.E., et al., *Mechanisms and clinical importance of bacteriophage resistance.* FEMS Microbiol Rev, 2022. **46**(1).
10. Goldfarb, T., et al., *BREX is a novel phage resistance system widespread in microbial genomes.* EMBO J, 2015. **34**(2): p. 169-83.
11. Millman, A., et al., *Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems.* Nat Microbiol, 2020. **5**(12): p. 1608-1615.
12. Sumby, P. and M.C. Smith, *Genetics of the phage growth limitation (Pgi) system of Streptomyces coelicolor A3(2).* Molecular microbiology, 2002. **44**: p. 489-500.
13. Xiong, X., et al., *SspABCD-SspE is a phosphorothioation-sensing bacterial defence system with broad anti-phage activities.* Nat Microbiol, 2020. **5**(7): p. 917-928.
14. Xu, T., et al., *A novel host-specific restriction system associated with DNA backbone S-modification in Salmonella.* Nucleic Acids Res, 2010. **38**(20): p. 7133-41.
15. Halford, S.E., *An end to 40 years of mistakes in DNA-protein association kinetics?* Biochem Soc Trans, 2009. **37**(Pt 2): p. 343-8.
16. Loenen, W.A., et al., *Highlights of the DNA cutters: a short history of the restriction enzymes.* Nucleic Acids Res, 2014. **42**(1): p. 3-19.
17. Loenen, W.A. and E.A. Raleigh, *The other face of restriction: modification-dependent enzymes.* Nucleic Acids Res, 2014. **42**(1): p. 56-69.
18. Anantharaman, V., L. Aravind, and E.V. Koonin, *Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins.* Curr Opin Chem Biol, 2003. **7**(1): p. 12-20.
19. Bujnicki, J.M., *Understanding the evolution of restriction-modification systems: clues from sequence and structure comparisons.* Acta Biochim Pol, 2001. **48**(4): p. 935-67.
20. Wang, L., et al., *DNA phosphorothioate modification-a new multi-functional epigenetic system in bacteria.* FEMS Microbiol Rev, 2019. **43**(2): p. 109-122.
21. Luyten, Y.A., et al., *Identification and characterization of the WYL BrxR protein and its gene as separable regulatory elements of a BREX phage restriction system.* Nucleic Acids Res, 2022. **50**(9): p. 5171-5190.
22. Shen, B.W., et al., *Structure, substrate binding and activity of a unique AAA+ protein: the BrxL phage restriction factor.* Nucleic Acids Res, 2023.
23. Song, S. and T.K. Wood, *A Primary Physiological Role of Toxin/Antitoxin Systems Is Phage Inhibition.* Front Microbiol, 2020. **11**: p. 1895.
24. Mruk, I. and I. Kobayashi, *To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems.* Nucleic Acids Res, 2014. **42**(1): p. 70-86.
25. Makarova, K.S., Y.I. Wolf, and E.V. Koonin, *Comparative genomics of defense systems in archaea and bacteria.* Nucleic Acids Res, 2013. **41**(8): p. 4360-77.
26. Kronheim, S., et al., *A chemical defence against phage infection.* Nature, 2018. **564**(7735): p. 283-286.
27. Kobayashi, I., *Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution.* Nucleic Acids Research, 2001. **29**(18): p. 3742-3756.
28. Naito, T., K. Kusano, and I. Kobayashi, *Selfish behavior of restriction-modification systems.* Science, 1995. **267**(5199): p. 897-9.
29. Edgell, D.R., *Selfish DNA: homing endonucleases find a home.* Curr Biol, 2009. **19**(3): p. R115-7.
30. Roberts, R.J., et al., *A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes.* Nucleic Acids Res, 2003. **31**(7): p. 1805-12.
31. Roberts, R.J., et al., *REBASE--a database for DNA restriction and modification: enzymes, genes and genomes.* Nucleic Acids Res, 2015. **43**(Database issue): p. D298-9.
32. Orlowski, J. and J.M. Bujnicki, *Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses.* Nucleic Acids Res, 2008. **36**(11): p. 3552-69.
33. Pingoud, A., et al., *Type II restriction endonucleases: structure and mechanism.* Cell Mol Life Sci, 2005. **62**(6): p. 685-707.
34. Mucke, M., D.H. Kruger, and M. Reuter, *Diversity of type II restriction endonucleases that require two DNA recognition sites.* Nucleic Acids Res, 2003. **31**(21): p. 6079-84.
35. Pingoud, A. and A. Jeltsch, *Structure and function of type II restriction endonuclease.* Nucleic Acids Res, 2001.

36. Shen, B.W., et al., *Structure, subunit organization and behavior of the asymmetric Type IIT restriction endonuclease BbvCI.* Nucleic Acids Res, 2019. **47**(1): p. 450-467.
37. Pingoud, V., et al., *Evolutionary relationship between different subgroups of restriction endonucleases.* J Biol Chem, 2002. **277**(16): p. 14306-14.
38. Shen, B.W., et al., *Coordination of phage genome degradation versus host genome protection by a bifunctional restriction-modification enzyme visualized by CryoEM.* Structure, 2021. **29**(6): p. 521-530 e5.
39. Aravind, L., K.S. Makarova, and E.V. Koonin, *SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories.* Nucleic Acids Res, 2000. **28**(18): p. 3417-32.
40. Ibryashkina, E.M., et al., *Type II restriction endonuclease R.Eco29kI is a member of the GIY-YIG nuclease superfamily.* BMC Struct Biol, 2007. **7**: p. 48.
41. Kuhlmann, U.C., et al., *Structural parsimony in endonuclease active sites: should the number of homing endonuclease families be redefined?* Federation of European Biochemical Societies, 1999. **463**: p. 2.
42. Jen-Jacobson, L., et al., *Structual adaptation in the interaction of EcoRI endonunclease with methylated GAATTC sites.* The EMBO, 1996. **15**(11): p. 2870-2882.
43. Taylor, J.D. and S.E. Halford, *Discrimination between DNA sequences by the EcoRV restriction endonuclease.* Biochemistry, 1989. **28**(15): p. 6198-207.
44. Thielking, V., et al., *Accuracy of the EcoRI restriction endonuclease: binding and cleavage studies with oligodeoxynucleotide substrates containing degenerate recognition sequences.* Biochemistry, 1990. **29**(19): p. 4682-91.
45. Szybalski, W., et al., *Class-IIS restriction enzymes--a review.* Gene, 1991. **100**: p. 13-26.
46. Shamshirgaran, Y., et al., *Tools for Efficient Genome Editing; ZFN, TALEN, and CRISPR.* Methods Mol Biol, 2022. **2495**: p. 29-46.
47. Drmanac, R., et al., *Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays.* Science, 2010. **327**(5961): p. 78-81.
48. Marillonnet, S. and S. Werner, *Assembly of Complex Pathways Using Type IIs Restriction Enzymes.* Methods Mol Biol, 2019. **1927**: p. 93-109.
49. Bitinaite, J., et al., *FokI dimerization is required for DNA cleavage.* Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10570-5.
50. Kaczorowski, T., P. Skowron, and A.J. Podhajska, *Purification and characterization of the FokI restriction endonuclease.* Gene, 1989. **80**(2): p. 209-16.
51. Wah, D.A., et al., *Structure of FokI has implications for DNA cleavage.* Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10564-9.
52. Wah, D.A., et al., *Structure of the multimodular endonuclease FokI bound to DNA.* Nature, 1997. **388**(6637): p. 97-100.
53. Vanamee, E.S., S. Santagata, and A.K. Aggarwal, *FokI requires two specific DNA sites for cleavage.* J Mol Biol, 2001. **309**(1): p. 69-78.
54. Catto, L.E., et al., *Protein assembly and DNA looping by the FokI restriction endonuclease.* Nucleic Acids Res, 2006. **34**(6): p. 1711-20.
55. Catto, L.E., et al., *Dynamics and consequences of DNA looping by the FokI restriction endonuclease.* Nucleic Acids Res, 2008. **36**(6): p. 2073-81.
56. Sanders, K.L., et al., *Targeting individual subunits of the FokI restriction endonuclease to specific DNA strands.* Nucleic Acids Res, 2009. **37**(7): p. 2105-15.
57. Laurens, N., et al., *DNA looping by FokI: the impact of twisting and bending rigidity on protein-induced looping dynamics.* Nucleic Acids Res, 2012. **40**(11): p. 4988-97.
58. Rusling, D.A., et al., *DNA looping by FokI: the impact of synapse geometry on loop topology at varied site orientations.* Nucleic Acids Res, 2012. **40**(11): p. 4977-87.
59. Vanamee, E.S., J. Berriman, and A.K. Aggarwal, *An EM view of the FokI synaptic complex by single particle analysis.* J Mol Biol, 2007. **370**(2): p. 207-12.
60. Gaj, T., C.A. Gersbach, and C.F. Barbas, 3rd, *ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering.* Trends Biotechnol, 2013. **31**(7): p. 397-405.
61. Tovkach, A., V. Zeevi, and T. Tzfira, *Expression, purification and characterization of cloning-grade zinc finger nuclease.* J Biotechnol, 2011. **151**(1): p. 1-8.
62. Bath, A.J., et al., *Many type IIs restriction endonucleases interact with two recognition sites before cleaving DNA.* J Biol Chem, 2002. **277**(6): p. 4024-33.
63. Embleton, M.L., V. Siksnys, and S.E. Halford, *DNA cleavage reactions by type II restriction enzymes that require two copies of their recognition sites.* J Mol Biol, 2001. **311**(3): p. 503-14.
64. Halford, S.E., A.J. Welsh, and M.D. Szczelkun, *Enzyme-mediated DNA looping.* Annu Rev Biophys Biomol Struct, 2004. **33**: p. 1-24.
65. Gowers, D.M., S.R. Bellamy, and S.E. Halford, *One recognition sequence, seven restriction enzymes, five reaction mechanisms.* Nucleic Acids Res, 2004. **32**(11): p. 3469-79.
66. Gormley, N.A., A.L. Hillberg, and S.E. Halford, *The type IIs restriction endonuclease BspMI is a tetramer that acts concertedly at two copies of an asymmetric DNA sequence.* J Biol Chem, 2002. **277**(6): p. 4034-41.
67. Gao, Y., et al., *Structural insights into assembly, operation and inhibition of a type I restriction-modification system.* Nat Microbiol, 2020. **5**(9): p. 1107-1118.
68. Klug, A., *The discovery of zinc fingers and their applications in gene regulation and genome manipulation.* Annu Rev Biochem, 2010. **79**: p. 213-31.
69. Li, T., et al., *TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain.* Nucleic Acids Res, 2011. **39**(1): p. 359-72.
70. Chandrasegaran, S. and J. Smith, *Chimeric restriction enzymes: what is next?* Biol Chem, 1999. **380**(7-8): p. 841-8.
71. Otwinowski, Z. and W. Minor, *Processing of X-ray diffraction data collected in oscillation mode.* Macromolecular Crystallography, Pt A, 1997. **276**: p. 307-326.

72. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 213-21.
73. Varadi, M., et al., *AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.* Nucleic Acids Res, 2022. **50**(D1): p. D439-D444.
74. Emsley, P., et al., *Features and development of Coot.* Acta Crystallographica, 2007. **D66**: p. 486-501.
75. Ohi, M., et al., *Negative Staining and Image Classification - Powerful Tools in Modern Electron Microscopy.* Biol Proced Online, 2004. **6**: p. 23-34.
76. Suloway, C., et al., *Automated molecular microscopy: the new Leginon system.* J Struct Biol, 2005. **151**(1): p. 41-60.
77. Mastronarde, D.N., *Automated electron microscope tomography using robust prediction of specimen movements.* J Struct Biol, 2005. **152**(1): p. 36-51.
78. Tegunov, D. and P. Cramer, *Real-time cryo-electron microscopy data preprocessing with Warp.* Nat Methods, 2019. **16**(11): p. 1146-1152.
79. Scheres, S.H., *RELION: implementation of a Bayesian approach to cryo-EM structure determination.* J Struct Biol, 2012. **180**(3): p. 519-30.
80. Rohou, A. and N. Grigorieff, *CTFFIND4: Fast and accurate defocus estimation from electron micrographs.* J Struct Biol, 2015. **192**(2): p. 216-21.
81. Punjani, A., et al., *cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination.* Nat Methods, 2017. **14**(3): p. 290-296.
82. Schrödinger, L., *The PyMol Molecular Graphics System. 1.2r3pre edn (PyMOLThe PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC, 2020).*
83. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
84. Sagendorf, J.M., et al., *DNAproDB: an expanded database and web-based tool for structural analysis of DNA-protein complexes.* Nucleic Acids Res, 2020. **48**(D1): p. D277-D287.
85. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* J Mol Biol, 2007. **372**(3): p. 774-97.
86. Schneider, C.A., W.S. Rasband, and K.W. Eliceiri, *NIH Image to ImageJ: 25 years of image analysis.* Nat Methods, 2012. **9**(7): p. 671-5.
87. Manning, G.S., *The persistence length of DNA is reached from the persistence length of its null isomer through an internal electrostatic stretching force.* Biophys J, 2006. **91**(10): p. 3607-16.
88. Wang, Y., S. Ran, and G. Yang, *Single molecular investigation of DNA looping and aggregation by restriction endonuclease BspMI.* Sci Rep, 2014. **4**: p. 5897.
89. Hildebrandt, E.R. and N.R. Cozzarelli, *Comparison of recombination in vitro and in E. coli cells: measure of the effective concentration of DNA in vivo.* Cell, 1995. **81**(3): p. 331-40.
90. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
91. Nowotny, M., *Retroviral integrase superfamily: the structural perspective.* EMBO Rep, 2009. **10**(2): p. 144-51.
92. Brooks, K.M., et al., *Integrase Inhibitors: After 10 Years of Experience, Is the Best Yet to Come?* Pharmacotherapy, 2019. **39**(5): p. 576-598.
93. Hare, S., et al., *Retroviral intasome assembly and inhibition of DNA strand transfer.* Nature, 2010. **464**(7286): p. 232-6.
94. Maertens, G.N., S. Hare, and P. Cherepanov, *The mechanism of retroviral integration from X-ray structures of its key intermediates.* Nature, 2010. **468**(7321): p. 326-9.
95. Hare, S., G.N. Maertens, and P. Cherepanov, *3'-processing and strand transfer catalysed by retroviral integrase in crystallo.* EMBO J, 2012. **31**(13): p. 3020-8.
96. Engelman, A.N. and P. Cherepanov, *Retroviral intasomes arising.* Curr Opin Struct Biol, 2017. **47**: p. 23-29.
97. Cook, N.J., et al., *Structural basis of second-generation HIV integrase inhibitor action and viral resistance.* Science, 2020. **367**(6479): p. 806-810.
98. Passos, D.O., et al., *Structural basis for strand-transfer inhibitor binding to HIV intasomes.* Science, 2020. **367**(6479): p. 810-814.
99. Li, M., et al., *Mechanisms of HIV-1 Integrase Resistance to Dolutegravir and Potent Inhibition of Drug Resistant Variants.* bioRxiv, 2023.
100. Sinha, R. and P. Shukla, *Current Trends in Protein Engineering: Updates and Progress.* Curr Protein Pept Sci, 2019. **20**(5): p. 398-407.
101. Mohan, K., et al., *Topological control of cytokine receptor signaling induces differential effects in hematopoiesis.* Science, 2019. **364**(6442).
102. Silva, D.A., et al., *De novo design of potent and selective mimics of IL-2 and IL-15.* Nature, 2019. **565**(7738): p. 186-191.
103. Dou, J., et al., *De novo design of a fluorescence-activating beta-barrel.* Nature, 2018. **561**(7724): p. 485-491.
104. Tinberg, C.E., et al., *Computational design of ligand-binding proteins with high affinity and selectivity.* Nature, 2013. **501**(7466): p. 212-216.
105. Ashtawy, H.M. and N.R. Mahapatra, *A comparative assessment of ranking accuracies of conventional and machine-learning-based scoring functions for protein-ligand binding affinity prediction.* IEEE/ACM Trans Comput Biol Bioinform, 2012. **9**(5): p. 1301-13.
106. Ross, G.A., G.M. Morris, and P.C. Biggin, *One Size Does Not Fit All: The Limits of Structure-Based Models in Drug Discovery.* J Chem Theory Comput, 2013. **9**(9): p. 4266-4274.
107. Ballester, P.J., A. Schreyer, and T.L. Blundell, *Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity?* J Chem Inf Model, 2014. **54**(3): p. 944-55.
108. MacDonald, J.T. and P.S. Freemont, *Computational protein design with backbone plasticity.* Biochem Soc Trans, 2016. **44**(5): p. 1523-1529.
109. Chao, G., et al., *Isolating and engineering human antibodies using yeast surface display.* Nat Protoc, 2006. **1**(2): p. 755-68.

110. Fowler, D.M., et al., *High-resolution mapping of protein sequence-function relationships.* Nat Methods, 2010. **7**(9): p. 741-6.
111. McLaughlin, R.N., Jr., et al., *The spatial architecture of protein function and adaptation.* Nature, 2012. **491**(7422): p. 138-42.
112. Whitehead, T.A., et al., *Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing.* Nat Biotechnol, 2012. **30**(6): p. 543-8.
113. Day, A.L., et al., *Unintended specificity of an engineered ligand-binding protein facilitated by unpredicted plasticity of the protein fold.* Protein Eng Des Sel, 2018. **31**(10): p. 375-387.
114. Dou, J., et al., *Sampling and energy evaluation challenges in ligand binding protein design.* Protein Sci, 2017. **26**(12): p. 2426-2437.
115. Hannan, A.J., *Tandem repeats mediating genetic plasticity in health and disease.* Nat Rev Genet, 2018. **19**(5): p. 286-298.
116. Jernigan, K.K. and S.R. Bordenstein, *Tandem-repeat protein domains across the tree of life.* PeerJ, 2015. **3**: p. e732.
117. Kajava, A.V., *Tandem repeats in proteins: from sequence to structure.* J Struct Biol, 2012. **179**(3): p. 279-88.
118. Paladin, L., et al., *RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures.* Nucleic Acids Res, 2021. **49**(D1): p. D452-D457.
119. Wang, X., et al., *Modular recognition of RNA by a human pumilio-homology domain.* Cell, 2002. **110**(4): p. 501-12.
120. Mak, A.N., et al., *The crystal structure of TAL effector PthXo1 bound to its DNA target.* Science, 2012. **335**(6069): p. 716-9.
121. Deng, D., et al., *Structural basis for sequence-specific recognition of DNA by TAL effectors.* Science, 2012. **335**(6069): p. 720-3.
122. Barkan, A., et al., *A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins.* PLoS Genet, 2012. **8**(8): p. e1002910.
123. Reichen, C., S. Hansen, and A. Pluckthun, *Modular peptide binding: from a comparison of natural binders to designed armadillo repeat proteins.* J Struct Biol, 2014. **185**(2): p. 147-62.
124. Wierenga, R.K., *The TIM-barrel fold: a versatile framework for efficient enzymes.* FEBS Lett, 2001. **492**(3): p. 193-8.
125. Andrade, M.A., C. Perez-Iratxeta, and C.P. Ponting, *Protein repeats: structures, functions, and evolution.* J Struct Biol, 2001. **134**(2-3): p. 117-31.
126. Grove, T.Z., A.L. Cortajarena, and L. Regan, *Ligand binding by repeat proteins: natural and designed.* Curr Opin Struct Biol, 2008. **18**(4): p. 507-15.
127. Javadi, Y. and L.S. Itzhaki, *Tandem-repeat proteins: regularity plus modularity equals design-ability.* Curr Opin Struct Biol, 2013. **23**(4): p. 622-31.
128. Correnti, C.E., et al., *Engineering and functionalization of large circular tandem repeat protein nanoparticles.* Nat Struct Mol Biol, 2020. **27**(4): p. 342-350.
129. Doyle, L., et al., *Rational design of alpha-helical tandem repeat proteins with closed architectures.* Nature, 2015. **528**(7583): p. 585-8.
130. Park, K., et al., *Control of repeat-protein curvature by computational protein design.* Nat Struct Mol Biol, 2015. **22**(2): p. 167-74.
131. Fallas, J.A., et al., *Computational design of self-assembling cyclic protein homo-oligomers.* Nat Chem, 2017. **9**(4): p. 353-360.
132. Bandyopadhyay, S., K. Chandramouli, and M.K. Johnson, *Iron-sulfur cluster biosynthesis.* Biochem Soc Trans, 2008. **36**(Pt 6): p. 1112-9.
133. Erickson, J. and D. Kempf, *Structure-based design of symmetric inhibitors of HIV-1 protease.* Arch Virol Suppl, 1994. **9**: p. 19-29.
134. Oie T., M.G.M., Christoffersen R. E.,, *Structural characterization of a special-pair chlorophyll dimer model of P700.* Int. J. Quantum Chem., 1982. **22**: p. 157–171.
135. Hallinan, J.P., et al., *Design of functionalised circular tandem repeat proteins with longer repeat topologies and enhanced subunit contact surfaces.* Commun Biol, 2021. **4**(1): p. 1240.
136. Leaver-Fay, A., et al., *ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.* Methods Enzymol, 2011. **487**: p. 545-74.
137. Koga, N., et al., *Principles for designing ideal protein structures.* Nature, 2012. **491**(7423): p. 222-7.
138. Croce, R. and H. van Amerongen, *Natural strategies for photosynthetic light harvesting.* Nat Chem Biol, 2014. **10**(7): p. 492-501.
139. Mirkovic, T., et al., *Light Absorption and Energy Transfer in the Antenna Complexes of Photosynthetic Organisms.* Chem Rev, 2017. **117**(2): p. 249-293.
140. Romero, E., V.I. Novoderezhkin, and R. van Grondelle, *Quantum design of photosynthesis for bio-inspired solar-energy conversion.* Nature, 2017. **543**(7645): p. 355-365.
141. Sener, M., et al., *Forster Energy Transfer Theory as Reflected in the Structures of Photosynthetic Light-Harvesting Systems.* Chemphyschem, 2011. **12**(3): p. 518-531.
142. Bednarczyk, D., et al., *Fine Tuning of Chlorophyll Spectra by Protein-Induced Ring Deformation.* Angew Chem Int Ed Engl, 2016. **55**(24): p. 6901-5.
143. Chang, M.C., et al., *Spectroscopic characterization of the light-harvesting complex of Rhodospirillum rubrum and its structural subunit.* Biochemistry, 1990. **29**(2): p. 421-9.
144. Ferretti, M., et al., *The nature of coherences in the B820 bacteriochlorophyll dimer revealed by two-dimensional electronic spectroscopy.* Phys Chem Chem Phys, 2014. **16**(21): p. 9930-9.
145. Pieper, J., et al., *Excitonic energy level structure and pigment-protein interactions in the recombinant water-soluble chlorophyll protein. II. Spectral hole-burning experiments.* J Phys Chem B, 2011. **115**(14): p. 4053-65.
146. Srivastava, A., et al., *Accurate prediction of mutation-induced frequency shifts in chlorophyll proteins with a simple electrostatic model.* J Chem Phys, 2021. **155**(15): p. 151102.
147. Visschers, R.W., et al., *Fluorescence polarization and low-temperature absorption spectroscopy of a subunit form of light-harvesting complex I from purple photosynthetic bacteria.* Biochemistry, 1991. **30**(23): p. 5734-42.

148. Boxer, S.G. and G.L. Closs, *Covalently Bound Dimeric Derivative of Pyrochlorophyllide a - Possible Model for Reaction Center Chlorophyll.* Journal of the American Chemical Society, 1976. **98**(17): p. 5406-5408.
149. Kobuke, Y. and H. Miyaji, *Supramolecular Organization of Imidazolyl-Porphyrin to a Slipped Cofacial Dimer.* Journal of the American Chemical Society, 1994. **116**(9): p. 4111-4112.
150. McCleese, C., et al., *Excitonic Interactions in Bacteriochlorin Homo-Dyads Enable Charge Transfer: A New Approach to the Artificial Photosynthetic Special Pair.* J Phys Chem B, 2018. **122**(14): p. 4131-4140.
151. Sharma, V.K., et al., *Dimeric Corrole Analogs of Chlorophyll Special Pairs.* J Am Chem Soc, 2021. **143**(25): p. 9450-9460.
152. Wasielewski, M.R., M.H. Studier, and J.J. Katz, *Covalently linked chlorophyll a dimer: A biomimetic model of special pair chlorophyll.* Proc Natl Acad Sci U S A, 1976. **73**(12): p. 4282-6.
153. Curti, M., et al., *Engineering excitonically coupled dimers in an artificial protein for light harvesting via computational modeling.* Protein Science, 2023. **32**(3).
154. Ennist, N.M., et al., *Maquette Strategy for Creation of Light- and Redox-Active Proteins*, in *Photosynthesis and Bioenergetics*. 2017. p. 1-33.
155. Ennist, N.M., et al., *Rational design of photosynthetic reaction center protein maquettes.* Front Mol Biosci, 2022. **9**: p. 997295.
156. Ennist, N.M., et al., *De novo protein design of photochemical reaction centers.* Nat Commun, 2022. **13**(1): p. 4937.
157. Farid, T.A., et al., *Elementary tetrahelical protein design for diverse oxidoreductase functions.* Nat Chem Biol, 2013. **9**(12): p. 826-833.
158. Fry, H.C., et al., *Computational Design and Elaboration of a de Novo Heterotetrameric alpha-Helical Protein That Selectively Binds an Emissive Abiological (Porphinato)zinc Chromophore.* Journal of the American Chemical Society, 2010. **132**(11): p. 3997-4005.
159. Kodali, G., et al., *Design and engineering of water-soluble light-harvesting protein maquettes.* Chem Sci, 2017. **8**(1): p. 316-324.
160. Moser, C.C., et al., *De Novo Construction of Redox Active Proteins.* Methods Enzymol, 2016. **580**: p. 365-88.
161. Pirro, F., et al., *Allosteric cooperation in a de novo-designed two-domain protein.* Proc Natl Acad Sci U S A, 2020. **117**(52): p. 33246-33253.
162. Polizzi, N.F., et al., *De novo design of a hyperstable non-natural protein-ligand complex with sub-A accuracy.* Nat Chem, 2017. **9**(12): p. 1157-1164.
163. Rabanal, F., W.F. DeGrado, and P.L. Dutton, *Toward the synthesis of a photosynthetic reaction center maquette: A cofacial porphyrin pair assembled between two subunits of a synthetic four-helix bundle multiheme protein.* Journal of the American Chemical Society, 1996. **118**(2): p. 473-474.
164. Wahadoszamen, M., et al., *The role of charge-transfer states in energy transfer and dissipation within natural and artificial bacteriochlorophyll proteins.* Nature Communications, 2014. **5**.
165. Lang, D., et al., *Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion.* Science, 2000. **289**(5484): p. 1546-50.
166. Levy, E.D. and S. Teichmann, *Structural, evolutionary, and assembly principles of protein oligomerization.* Prog Mol Biol Transl Sci, 2013. **117**: p. 25-51.
167. Fromont-Racine, M., et al., *Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins.* Yeast, 2000. **17**(2): p. 95-110.
168. Khusial, P., R. Plaag, and G.W. Zieve, *LSm proteins form heptameric rings that bind to RNA via repeating motifs.* Trends in Biochemical Sciences, 2005. **30**(9): p. 522-528.
169. Wilusz, C.J. and J. Wilusz, *Lsm proteins and Hfq: Life at the 3 end.* Rna Biology, 2013. **10**(4): p. 592-601.
170. Williams, G.J., et al., *Structure of the heterotrimeric PCNA from Sulfolobus solfataricus.* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2006. **62**(Pt 10): p. 944-8.
171. Bermeo, S., et al., *De novo design of obligate ABC-type heterotrimeric proteins.* Nat Struct Mol Biol, 2022. **29**(12): p. 1266-1276.
172. Bolon, D.N., et al., *Specificity versus stability in computational protein design.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12724-9.
173. Boyken, S.E., et al., *De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity.* Science, 2016. **352**(6286): p. 680-7.
174. Chen, Z., et al., *Programmable design of orthogonal protein heterodimers.* Nature, 2019. **565**(7737): p. 106-111.
175. Dawson, W.M., et al., *Coiled coils 9-to-5: rational de novo design of alpha-helical barrels with tunable oligomeric states.* Chem Sci, 2021. **12**(20): p. 6923-6928.
176. Lombardi, A., J.W. Bryson, and W.F. DeGrado, *De novo design of heterotrimeric coiled coils.* Biopolymers, 1996. **40**(5): p. 495-504.
177. Reinke, A.W., R.A. Grant, and A.E. Keating, *A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering.* J Am Chem Soc, 2010. **132**(17): p. 6025-31.
178. Sahtoe, D.D., et al., *Reconfigurable asymmetric protein assemblies through implicit negative design.* Science, 2022. **375**(6578): p. 283-+.
179. Wicky, B.I.M., et al., *Hallucinating symmetric protein assemblies.* Science, 2022. **378**(6615): p. 56-61.
180. Kibler, R.D., et al., *Stepwise design of pseudosymmetric protein hetero-oligomers.* bioRxiv, 2023.
181. Lee, S., et al., *A general design route to four component T=4 nanocages with tetrahedral, octahedral, and icosahedral symmetry*. 2023.
182. Ennist, N., et al., *De novo design of energy transfer proteins housing excitonically coupled chlorophyll special pairs.* 2023.
183. Hicks, D.R., et al., *De novo design of protein homodimers containing tunable symmetric protein pockets.* Proc Natl Acad Sci U S A, 2022. **119**(30): p. e2113400119.
184. Park, H., et al., *Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules.* J Chem Theory Comput, 2016. **12**(12): p. 6201-6212.
185. Marcos, E., et al., *De novo design of a non-local beta-sheet protein with high stability and accuracy.* Nat Struct Mol Biol, 2018. **25**(11): p. 1028-1034.

186. Studier, F.W., *Protein production by auto-induction in high density shaking cultures.* Protein Expr Purif, 2005. **41**(1): p. 207-34.
187. Bradley, P., *Structural modeling of TAL effector-DNA interactions.* Protein Sci, 2012. **21**(4): p. 471-4.
188. Mejias, S.H., et al., *Controlled nanometric fibers of self-assembled designed protein scaffolds.* Nanoscale, 2014. **6**(19): p. 10982-8.
189. Brunette, T.J., et al., *Exploring the repeat protein universe through computational protein design.* Nature, 2015. **528**(7583): p. 580-4.
190. Foight, G.W., et al., *Multi-input chemical control of protein dimerization for programming graded cellular responses.* Nat Biotechnol, 2019. **37**(10): p. 1209-1216.
191. Ueda, G., et al., *Tailored design of protein nanoparticle scaffolds for multivalent presentation of viral glycoprotein antigens.* Elife, 2020. **9**.
192. Gisriel, C., et al., *Structure of a symmetric photosynthetic reaction center-photosystem.* Science, 2017. **357**(6355): p. 1021-1025.
193. Chen, J.H., et al., *Architecture of the photosynthetic complex from a green sulfur bacterium.* Science, 2020. **370**(6519).
194. Leman, J.K., et al., *Macromolecular modeling and design in Rosetta: recent methods and frameworks.* Nat Methods, 2020. **17**(7): p. 665-680.
195. Sharp, P.M. and W.H. Li, *The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.* Nucleic Acids Res, 1987. **15**(3): p. 1281-95.
196. McCoy, A.J., et al., *Phaser crystallographic software.* J Appl Crystallogr, 2007. **40**(Pt 4): p. 658-674.
197. Moriarty, N.W., R.W. Grosse-Kunstleve, and P.D. Adams, *electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation.* Acta Crystallogr D Biol Crystallogr, 2009. **65**(Pt 10): p. 1074-80.
198. Evans, P.R. and G.N. Murshudov, *How good are my data and what is the resolution?* Acta Crystallogr D Biol Crystallogr, 2013. **69**(Pt 7): p. 1204-14.
199. Cao, P., et al., *Structural basis for the assembly and quinone transport mechanisms of the dimeric photosynthetic RC-LH1 supercomplex.* Nat Commun, 2022. **13**(1): p. 1977.
200. Niwa, S., et al., *Structure of the LH1-RC complex from Thermochromatium tepidum at 3.0 A.* Nature, 2014. **508**(7495): p. 228-32.
201. Qian, P., et al., *2.4-A structure of the double-ring Gemmatimonas phototrophica photosystem.* Sci Adv, 2022. **8**(7): p. eabk3139.
202. Qian, P., et al., *Cryo-EM structure of the monomeric Rhodobacter sphaeroides RC-LH1 core complex at 2.5 A.* Biochemical Journal, 2021. **478**(20): p. 3775-3790.
203. Selikhanov, G., et al., *Novel approaches for the lipid sponge phase crystallization of the Rhodobacter sphaeroides photosynthetic reaction center.* IUCrJ, 2020. **7**(Pt 6): p. 1084-1091.
204. Swainsbury, D.J.K., et al., *Structures of Rhodopseudomonas palustris RC-LH1 complexes with open or closed quinone channels.* Sci Adv, 2021. **7**(3).
205. Tani, K., et al., *Cryo-EM structure of a Ca(2+)-bound photosynthetic LH1-RC complex containing multiple alphabeta-polypeptides.* Nat Commun, 2020. **11**(1): p. 4955.
206. Yu, L.J., et al., *Structure of photosynthetic LH1-RC supercomplex at 1.9 A resolution.* Nature, 2018. **556**(7700): p. 209-213.
207. Reppert, M., *Bioexcitons by Design: How Do We Get There?* J Phys Chem B, 2023. **127**(9): p. 1872-1879.
208. DiMaio, F., et al., *Modeling symmetric macromolecular structures in Rosetta3.* PLoS One, 2011. **6**(6): p. e20450.
209. Fleishman, S.J., et al., *RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite.* PLoS One, 2011. **6**(6): p. e20161.
210. Maguire, J.B., et al., *Perturbing the energy landscape for improved packing during computational protein design.* Proteins, 2021. **89**(4): p. 436-449.
211. Lin, X., et al., *Specific alteration of the oxidation potential of the electron donor in reaction centers from Rhodobacter sphaeroides.* Proc Natl Acad Sci U S A, 1994. **91**(22): p. 10265-9.
212. Kibler, R.D., et al., *Stepwise design of pseudosymmetric protein hetero-oligomers.* 2023.
213. Dowling, Q., et al., *Hierarchical design of pseudosymmetric protein nanoparticles*. 2023.
214. Hsia, Y., et al., *Design of multi-scale protein complexes by hierarchical building block fusion.* Nat Commun, 2021. **12**(1): p. 2294.
215. Moore, R., A. Chandrahas, and L. Bleris, *Transcription activator-like effectors: a toolkit for synthetic biology.* ACS Synth Biol, 2014. **3**(10): p. 708-16.
216. Nishanth, M.J. and B. Simon, *Functions, mechanisms and regulation of Pumilio/Puf family RNA binding proteins: a comprehensive review.* Mol Biol Rep, 2020. **47**(1): p. 785-807.
217. Perez-Riba, A. and L.S. Itzhaki, *The tetratricopeptide-repeat motif is a versatile platform that enables diverse modes of molecular recognition.* Curr Opin Struct Biol, 2019. **54**: p. 43-49.
218. Antanasijevic, A., et al., *Structural and functional evaluation of de novo-designed, two-component nanoparticle carriers for HIV Env trimer immunogens.* PLoS Pathog, 2020. **16**(8): p. e1008665.
219. Tran, H.M., et al., *Nanomaterials for Treating Bacterial Biofilms on Implantable Medical Devices.* Nanomaterials (Basel), 2020. **10**(11).
220. Bauler, P., et al., *Channeling by Proximity: The Catalytic Advantages of Active Site Colocalization Using Brownian Dynamics.* J Phys Chem Lett, 2010. **1**(9): p. 1332-1335.
221. Castellana, M., et al., *Enzyme clustering accelerates processing of intermediates through metabolic channeling.* Nat Biotechnol, 2014. **32**(10): p. 1011-8.
222. Wrapp, D., et al., *Structural Basis for Potent Neutralization of Betacoronaviruses by Single-Domain Camelid Antibodies.* Cell, 2020. **181**(5): p. 1004-1015 e15.

VITA

I have developed a unique combination of communication, writing, and analytical thinking skills, creative problem-solving ability, through my formal PhD training in biochemistry at the University of Washington (UW) in the Biophysics, Structure and Design Program (BPSD) and the Fred Hutchinson Cancer Center (Fred Hutch)

As a trained structural biologist in the Stoddard lab, I have become adept in x-ray crystallography, cryo-electron microscopy (EM), and protein design and engineering. I spearheaded research on the restriction endonuclease, PaqCI, elucidating the mechanism of the tetrameric DNA cleaving enzyme. This research became my second first author paper as well as the jewel of my thesis. Also during this time, I have been a part of and supported the coordination of flourishing collaborations with research project partners, such as the Baker lab. I solved over 14 structures for the Baker lab, where I honed my structure determination skills and deepened my understanding of protein design and engineering. My educational experiences have taught me how to manage my time efficiently to complete personal and project goals. I have successfully written and submitted multiple fellowships, many of which I have been awarded. I have also participated in the composition of manuscripts that have led to publication.

In addition to the rigorous training of my PhD program, I have been working closely with my thesis advisor, Dr. Barry Stoddard, to understand the scientific research industry and participate in the development, review, and editing of grants and papers. To supplement my scientific education and prepare myself for a future career in science writing and management, I've earned the UW's certificate in writing and will receive the certificate in editing this summer.

As a leader, I am on the organizing committee for the monthly community EM meeting, which brings together expert scientists, trainees, and professors across institutes and universities to establish a thriving EM community in Seattle. Each month we coordinate the hosting of two speakers whose research covers a wide range of EM applications.