The Architecture and Assembly Principles of Bacterial Protein Nanocompartments with a Two-Component Shell

by

Cassandra Ann Dutcher

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biological Chemistry)
in the University of Michigan
2024

Doctoral Committee:

Assistant Professor Tobias W. Giessen, Chair
Associate Professor Nicole M. Koropatkin
Associate Professor Allen P. Liu
Professor E. Neil G. Marsh
Professor Janet L. Smith

Cassandra Ann Dutcher

dutcherc@umich.edu

ORCID iD: 0000-0002-7285-4932

© Cassandra Ann Dutcher 2024

Dedication

To my parents, Greg and Brenda, and my husband, Brandon. Thank you for your unwavering love and support, as well as the occasional motivational push.

Acknowledgments

I am unbelievably grateful to all those who have provided me with guidance and support throughout my PhD. First and foremost, I would like to thank my advisor, Dr. Tobias Giessen, for your mentorship over the last five years, which has been instrumental in helping me grow as an independent scientist. A big thank you to the rest of the Giessen Lab – Dr. Michael Andreas, Dr. Jesse Jones, Dr. Robert Benisch, Natalia Ubilla, and Seokmu Kwon – for your endless encouragement and helpful feedback. I would like to especially thank Mike Andreas for your many contributions to this work and the countless hours that you have spent teaching me new lab techniques.

Thank you to all of the mentors I have had along the way; you have played an instrumental role in my growth as a scientist, and I have learned valuable skills from all of you. I am very fortunate to have received four years of great mentorship in Dr. Cheryl Kerfeld's lab, and I am especially grateful to Dr. Aiko Turmo and Dr. Bryan Ferlez. I also extend my gratitude to the entire Computational and Automated Biology Team at Conagen, who all taught me so much in such a short span of time, with a special thanks to my internship mentor, Dr. Curtis Walton.

I would also like to acknowledge my friends, who have been among my most ardent supporters. To the friends I have made during my graduate student tenure – Dr. Balaji Banoth, Dr. Mehdi Eshraghi, Ruth Gauer, Claire Griffith, Cara Loomis, Dr. Ruby Miller, Brandon Nichols, and Natalia Ubilla (again!) – I am unbelievably grateful for your friendship which has

carried me through both good times and bad. I would also like to thank my friends at Michigan State University who have supported me from the very beginning of my academic journey, especially Madeline Bresson, Andy Jurasek, Sean McGuire, Keenan O'Dea, and Damien Sheppard. A very special thanks goes to Ashley Cross: I am incredibly grateful to have found such a great friend in you so early in life.

Lastly, I would like to express my profound gratitude to my entire family for your love and support. To the family I have gained through marriage, especially Bob, Lorie, Evan, and Connor Llewellyn, and Sue and Rus Bailey, thank you for always checking in and cheering me on. I am very grateful to my extended Dutcher family for all their love, especially my Aunt Pam and Uncle Jim. Thank you to my brother, Gregory, for always challenging me, and my sister, Rachael, for always encouraging me. To my parents, Greg and Brenda Dutcher: You have been my most enthusiastic cheerleaders throughout my life, and I am endlessly grateful. None of my accomplishments would have been attainable without you, always encouraging me to persevere and helping me push through the toughest times. Lastly, and above all, I owe the biggest thanks to my husband, Brandon, who never stopped believing that I was capable of everything that I set my mind to. I could not have done this without you.

Table of Contents

Dedication ii

Acknowledgments iii

List of Tables viii

List of Figures ix

Abstract xv

Chapter 1 Introduction 1

1.1 Encapsulin nanocompartments 1

1.1.1 Genetic distribution and classification 2

1.1.2 Shell structure 5

1.1.3 Enzymatic cargoes and their implications in prokaryotic metabolism 10

1.1.4 Targeting peptides and cargo loading mechanisms 14

1.2 Geosmin Synthase 17

1.2.1 Geosmin: An earthy smelling terpenoid 18

1.2.2 Enzymatic reaction, structure, and stability 20

1.2.3 Geosmin synthase as a putative encapsulin cargo 23

1.3 Thesis overview and scope 26

1.4 References 27

Chapter 2 A Two-Component Quasi-Icosahedral Protein Nanocompartment with Variable Shell Composition and Irregular Tiling 36

2.1 Introduction 36

2.2 Materials and Methods 39

2.2.1 Bioinformatic and phylogenetic analysis 39

2.2.2 Molecular biology and cloning 40

2.2.3 Protein expression and purification 41

2.2.4 Negative stain transmission electron microscopy 42

2.2.5 Native PAGE 42

2.2.6 Dynamic light scattering (DLS) and thermal ramp experiments 43

2.2.7 Cryo-electron microscopy (cryo-EM) 43

2.2.8 Protein crosslinking assays 47

2.2.9 Protein identification via mass spectrometry 47

2.2.10 Mass spectrometric analysis of crosslinked proteins 48

2.3 Results 49

2.3.1 Two-component encapsulin operons are diverse and widespread in bacteria 49

2.3.2 Heterologous expression of individual shell components highlights differential assembly behavior 52

2.3.3 Sl-Enc1 and Sl-Enc2 co-assemble into compositionally variable two-component shells 53

2.3.4 Single particle cryo-EM analysis confirms mixed shell assembly and suggests irregular tiling of two-component shells 56

2.3.5 Crosslinking analysis of two-component shells confirms heterogeneous two-fold interactions 64

2.4 Discussion 66

2.5 Supplementary Information 72

2.6 References 88

Chapter 3 Encapsulation of Geosmin Synthase Within a Two-Component Protein Shell in *Myxococcus xanthus* 95

3.1 Introduction 95

3.2 Materials and Methods 98

3.2.1 Molecular biology and cloning 98

3.2.2 Protein expression and purification 99

3.2.3 GeoA Solubility Test 100

3.2.4 Negative stain transmission electron microscopy 101

3.2.5 Dynamic light scattering (DLS) and thermal ramp experiments 101

3.2.6 Protein identification via mass spectrometry 102

3.2.7 Cryo-electron Microscopy 102

3.2.8 Geosmin synthase activity assays 105

3.2.9 Gas chromatography mass spectrometry (GC-MS) analysis 106

3.3 Results 107

3.3.1 Heterologous production of encapsulin shell components demonstrates differential assembly behavior and ability to form mixed shells 107

3.3.2 The M. xanthus two-component encapsulin's cargo is geosmin synthase 110

3.3.3 Single particle cryo-EM analysis of Mx encapsulin shells 112

3.3.4 Cargo loading is mediated at the three-fold axis of symmetry 116

3.3.5 Geosmin synthase activity is similar between encapsulins with different shell compositions 119

3.4 Discussion 122

3.5 Supplementary Information 126

3.6 References 129

Chapter 4 Conclusions and Future Directions 135

4.1 Concluding remarks 135

4.2 Future work 138

4.2.1 Future experiments with the M. xanthus two-component encapsulin 138

4.2.2 Proposed studies of other two-component encapsulins 140

4.3 References 143

List of Tables

Supplemental Table 2-1: DNA sequences of gBlock gene fragments. 72

Supplemental Table 2-2: DNA sequences of primers used to construct plasmids. 74

Supplemental Table 2-3: Primer pairs used to construct the plasmids used in this study. . 75

Supplemental Table 2-4: Protein sequences used in this work 76

Table 3-1: Buffer conditions used for geosmin synthase solubility test. 101

Supplemental Table 3-1: DNA sequences of gene fragments. 126

Supplemental Table 3-2: DNA sequences of PCR primers used to construct plasmids. 128

Supplemental Table 3-3: Primer pairs used to construct the plasmids in this study. 128

Supplemental Table 3-4: Protein sequences used in this study. Underlined sequences represent expression tags that were kept or removed based on the plasmid design. 129

List of Figures

Figure 1-1: Standard encapsulin operon and loading schematic. Encapsulin "Enc" shell proteins (blue) and cargo proteins (yellow) co-assemble to form cargo-loaded encapsulin shells. A targeting peptide on the cargo's terminus (orange) interacts with the shell, mediating cargo encapsulation. 1

Figure 1-2: Operon structure of the four encapsulin families. Predicted encapsulin shell genes (Enc) are blue, cargo genes are yellow, cyclic nucleotide-binding domains (CBDs) are light blue, and neighboring accessory proteins are gray. Figure adapted from T. W. Giessen, 2022 (17). 3

Figure 1-3: Encapsulin shell structure. Encapsulin shell protomers exhibiting the HK97-fold are colored by region. Assembled encapsulin shells of varying triangulation numbers, with pentamers colored in dark blue and hexamers colored in light blue. The sole Family 2 encapsulin is from *A. baumannii* (PDB: 8T6R) with a cysteine desulfurase (CD) cargo. Family 1 encapsulins shown: *T. maritima* (PDB: 3DKT) with a ferritin-like protein (Flp) cargo, *M. xanthus* (PDB: 8TK7) with a dye-decoloring peroxidase (DyP) cargo, and *Q. thermotolerans* (PDB: 6NJ8) with an iron-mineralizing encapsulin-associated Firmicute (IMEF) cargo. 6

Figure 1-4: Encapsulin Family 2B protomer and shell structure from *S. griseus*. (**A**) Encapsulin shell protomer (PDB ID: 9BHU) exhibiting the HK97-fold and a cyclic nucleotide-binding domain (CBD). Regions are labeled by color. (**B**) Assembled encapsulin shell. The HK97-fold (blue) forms the icosahedral shell structure, with the CBDs (purple) displayed on the shell's exterior. (**C**) Interior and exterior views of the shell's two-fold pores. 8

Figure 1-5: Encapsulin cargoes and their biochemistry. (**A**) Ferritin-like proteins (Flps) and iron-mineralizing encapsulin-associated Firmicute (IMEF) cargoes exhibit ferroxidase activity in which ferrous iron is oxidized to ferric iron. Ferric iron is insoluble and forms precipitates known as ferrihydrite. (**B**) Dye-decolorizing peroxidases (DyPs) catalyze the oxidation of many substrates, including most anthraquinone dyes. (**C**) Cysteine desulfurase (CD) catalyzes the conversion of L-cysteine to L-alanine and forms a persulfide intermediate bound to CysD. It is predicted that the continued activity of CysD leads to the creation of long poly-persulfide chains, resulting in the formation of sulfur crystals. (**D**) 2-Methylisoborneol synthase (2-MIBS) catalyzes the conversion of 2-methylgeranyl pyrophosphate (2-MeGPP) to 2-methylisoborneol (2-MIB). 12

Figure 1-6: Reaction scheme for the conversion of farnesyl pyrophosphate to geosmin. Geosmin synthase's N-terminal domain (GeoA-NTD) first catalyzes the cyclization of farnesyl pyrophosphate to germacradienol. Then, the C-terminal domain (GeoA-CTD) catalyzes the cyclization-fragmentation reaction of germacradienol to geosmin. 20

Figure 1-7: Solved structure of the ligand-bound *S. coelicolor* geosmin synthase's N-terminal domain. (**A**) Crystal structure of the GeoA N-terminal domain (PDB ID: 5DZ2) coordinating three Mg^{2+} ions (green) in complex with alendronate (yellow) (100). The two conserved metal binding motifs are colored dark pink. (**B**) Zoomed view of the NTD active site. The four residues from the metal binding motifs responsible for coordinating the Mg^{2+} ions are displayed and labeled. 21

Figure 1-8: Predicted structures of *S. coelicolor* geosmin synthase. (**A**) AlphaFold model of the full-length geosmin synthase. N-terminal domain (orange) and C-terminal domain (green) are connected by a long flexible linker containing a single α-helix (yellow). Metal binding motifs (blue) are positioned away from each other. (**B**) Structural alignment of the ligand-bound geosmin synthase crystal structure (pink, PDB ID: 5DZ2) (100) and the AlphaFold model of the GeoA C-terminal domain (green). 22

Figure 2-1: The HK97-fold and distribution and diversity of Family 2B two-component encapsulin systems. (**A**) Two-component Family 2B encapsulin operon types. TC: terpene cyclase, PT: polyprenyl transferase, XK: xylulose kinase, CD: cysteine desulfurase, CBD: cyclic nucleotide (cNMP)-binding domain. (**B**) Phylogenetic distribution of two-component encapsulin systems. **(C**) Phylogenetic tree of all identified shell proteins encoded in two-component Family 2B encapsulin operons. Outer ring: Enc1 vs Enc2. Inner ring: phylum. Nodes colored by cargo type. (**D**) Domain organization of the two *Streptomyces* lydicus (Sl) shell proteins highlighting two conserved sequence differences found in all Enc1 and Enc2 shell proteins (diagnostic motif 1 and 2). (**E**) Comparative sequence analysis of Enc1 and Enc2 shell proteins. HK97: HK97-domain. 50

Figure 2-2: Sequence alignment of Sl-Enc1 and Sl-Enc2. A sequence alignment of Sl-Enc1 and Sl-Enc2 generated via Clustal Omega is shown. Sequences are colored by domain. 52

Figure 2-3: Separate analysis of Sl-Enc1 and Sl-Enc2. (**A**) SDS-PAGE analysis of individually expressed and purified Sl-Enc1 and Sl-Enc2. M: molecular weight marker. (**B**) Negative stain TEM micrographs of purified Sl-Enc1 and Sl-Enc2. (**C**) Native PAGE analysis of purified Sl-Enc1 and Sl-Enc2. 53

Figure 2-4: Analysis of mixed Sl-Enc1-Sl-Enc2 shells. (**A**) Negative stain TEM micrograph of purified Sl-Enc1 majority mixed shells. (**B**) Representative native PAGE analysis of an Sl-Enc1 majority mixed sample. (**C**) SDS-PAGE analysis of three replicate Sl-Enc1 majority mixed samples. The Sl-Enc1 to Sl-Enc2 ratio based on gel densitometry is shown on the right for individual purification replicates. (**D**) Dynamic light scattering (DLS) analysis of individually purified Sl-Enc1 (left) and Sl-Enc2 (middle) as well as of an Sl-Enc1 majority mixed sample (right). PDI: polydispersity index. 55

Figure 2-5: SDS-PAGE analysis of Sl-Enc1 majority and Sl-Enc2 majority samples subjected to cryo-EM analysis. (**A**) SDS-PAGE gel lane of protein sample used for cryo- EM data collection resulting in a symmetry-averaged Sl-Enc1 shell map. (**B**) SDS-PAGE gel lane of protein sample used for cryo-EM data collection resulting in a symmetry- averaged Sl-Enc2 shell map. 57

Figure 2-6: Structural analysis of mixed shells via single particle cryo-EM. (**A**) Protomer structures of Sl-Enc1 and Sl-Enc2. Structures colored by domain. (**B**) Homogeneous Sl-Enc2 shell with T=1 icosahedral symmetry. HK97-domain colored green. CBD colored pink. (**C**) Exterior and interior views of the two- (left), five- (middle), and three-fold (right) interactions found in the homogeneous Sl-Enc2 shell. (**D**) Two- (left), five- (middle), and three-fold (right) pores of the homogeneous Sl-Enc2 shell shown as electrostatic surface. Both the exterior and interior are shown with pore diameters highlighted. 58

Figure 2-7: Comparison of CBD orientations within Sl-Enc1 and Sl-Enc2. Structural alignment of the Sl-Enc1 and Sl-Enc2 proteins highlighting the different CBD tilts relative to the HK97-domain. A relative tilt between the Sl-Enc2 and Sl-Enc1 CBDs of ca. 23° can be observed. CBDs shown as surface, HK97-domain as ribbon. 60

Figure 2-8: Protomer interaction analysis in mixed shells via 3D classification. (**A**) Schematic of a T=1 icosahedral protein shell highlighting the four unique interactions of a given protomer. Interactions are labelled by symmetry; orange: interaction across the 2-fold pore; green: interaction at the 3-fold pore; pink: 2-fold P-domain interaction; blue: 5-fold A-domain interaction. (**B**) 3D classification masks used to investigate all possible protomer interactions in mixed shells. (**C**) Observed protomer arrangements across the four unique interactions outlined in (A). Color scheme based on respective mask color. Sl-Enc1 shown in light green. Sl-Enc2 shown in dark green. (**D**) The diagnostic regions (diagnostic motif 1) within the Sl-Enc1 and Sl-Enc2 protomers used to unambiguously identify them. As an example, Sl-Enc1 from the Sl-Enc1-Sl-Enc1 2-fold P-domain interaction and Sl-Enc2 from the Sl-Enc2-Sl-Enc1 2-fold P- domain interaction is shown as cryo-EM density and atomic models with the respective sequence stretches shown below. 61

Figure 2-9: Two-fold pore variability and visualization of mixed shell heterogeneity. (**A**) Two-fold pore Sl-Enc1-Sl-Enc1 interactions observed via 3D classification analysis highlighting two resolved (top) or unresolved (bottom) N-arms (red). **(B**) Electrostatic surface representation of the two distinct two-fold pore Sl-Enc1-Sl-Enc1 arrangements. Top: two resolved N-arms; Bottom: unresolved N-arms. Left: exterior view. Right: interior view. (**C**) Color labelling scheme for the four unique interactions and protomer arrangements found across them used in subsequent panels. (**D**) Back-mapping of sub-particles contributing to the different classes identified for the 2-fold (pore) interaction via 3D classification onto a sample micrograph highlighting that all classes, representing unique protomer arrangements, can generally be found in each individual encapsulin particle. Each colored dot represents a single sub-particle sorted into one of the highlighted classes during 3D classification. (**E**) Back-mapping of sub- particles for the 2-fold (P-domain) interaction. (**F**) Back-mapping of sub-particles for the 3-fold interaction. (**G**) Back-mapping of sub-particles for the 5-fold interaction. (**H**) Five sample particles (1-5) from the sample micrograph shown above are highlighted with all the distinct protomer arrangements shown using the color scheme of panel (A). All identified protomer arrangements across the four unique interactions can generally be found within each individual encapsulin particle. As a subset of 10,000 particles was used for this analysis, not all interactions that exist within each particle will be present. 63

Figure 2-10: Site-specific crosslinking analysis of mixed shells. (**A**) Schematic of our crosslinking strategy based on CBD cysteine mutants (Sl-Enc1: R195C; Sl-Enc2: G198C). (**B**)

SDS-PAGE analysis of a control crosslinking experiment using native mixed shells highlighting that no major higher molecular weight bands are present in the post crosslinking sample. (**C**) SDS-PAGE analysis of crosslinked mixed samples containing either one CBD cysteine (Sl-Enc1: R195C) or two CBD cysteines (Sl-Enc1: R195C and Sl-Enc2: G198C). The higher molecular weight bands observed in post crosslinking samples are outlined. (**D**) Mass spectrometric identification of crosslinks present in the higher molecular weight crosslinked bands. The identified crosslinks for each band are shown..... 65

Figure 2-11: Location of the diagnostic motifs within the Sl-Enc1 and Sl-Enc2 protomers. The two conserved differences between Enc1 and Enc2 shell proteins are highlighted. CBDs not shown for visual clarity..... 68

Supplemental Figure 2-1: Analysis of Sl-Enc2 majority shells. (**A**) Negative stain TEM micrograph of an Sl-Enc1 + His-Sl-Enc2 mixed shell sample purified via affinity purification and size exclusion chromatography. (**B**) Native PAGE analysis of an Sl-Enc1 majority mixed sample. M: molecular weight marker. (**C**) SDS-PAGE of purified Sl-Enc1 + His-Sl-Enc2. Outlined region on SDS-PAGE gel was excised and prepared for mass spectrometric analysis for protein identification. Total spectrum counts of Sl-Enc1 and Sl-Enc2 from mass spectrometry analysis are shown. 77

Supplemental Figure 2-2: Thermal ramp data for Sl-Enc1 and Sl-Enc2. (**A**) Thermal ramp data for Sl-Enc1, Sl-Enc2, Sl-Enc1 + Sl-Enc2. (B) SLS results for SL-Enc1, Sl-Enc1, and Sl-Enc1 + Sl-Enc2. (**C**) Average T_m and T_{agg} for Sl-Enc1, Sl-Enc2, and Sl-Enc1 + Sl-Enc2. 77

Supplemental Figure 2-3: Majority Sl-Enc1 cryo-EM data processing workflow. Cryo-EM workflow showing gold-standard Fourier shell correlation (FSC) curves for the final symmetry- averaged (I) consensus map as well as a local resolution analysis..... 78

Supplemental Figure 2-4: Majority Sl-Enc2 cryo-EM data processing workflow. Cryo-EM workflow showing gold-standard Fourier shell correlation (FSC) curves for the final symmetry- averaged (I) consensus map as well as a local resolution analysis..... 79

Supplemental Figure 2-5: Cryo-EM 3D classification analysis workflow for the 2-fold (pore) interaction. 80

Supplemental Figure 2-6: Cryo-EM 3D classification analysis workflow for the 2-fold (P-domain) interaction...... 81

Supplemental Figure 2-7: Cryo-EM 3D classification analysis workflow for the 3-fold interaction...... 82

Supplemental Figure 2-8: Cryo-EM 3D classification analysis workflow for the 5-fold interaction...... 83

Supplemental Figure 2-9: Identified high quality classes for the 2-fold (pore) 3D classification analysis. Left: Cryo-EM densities and atomic models for the five identified classes representing all possible interaction permutations. The class shown on top represents

the Sl-Enc1/Sl-Enc1 interaction with resolved N-arm. Right: Diagnostic regions located at the edge of the A-domain used to distinguish between Sl-Enc1 and Sl-Enc2..... 84

Supplemental Figure 2-10: Identified high quality classes for the 2-fold (P-domain) 3D classification analysis. Left: Cryo-EM densities and atomic models for the four identified classes representing all possible interaction permutations. Right: Diagnostic regions located at the edge of the A-domain used to distinguish between Sl-Enc1 and Sl-Enc2..... 85

Supplemental Figure 2-11: Identified high quality classes for the 3-fold 3D classification analysis. Left: Cryo-EM densities and atomic models for the four identified classes representing all possible interaction permutations. Right: Diagnostic regions located at the edge of the A- domain used to distinguish between Sl-Enc1 and Sl-Enc2. 86

Supplemental Figure 2-12: Identified high quality classes for the 5-fold 3D classification analysis. Left: Cryo-EM densities and atomic models for the four identified classes representing all possible interaction permutations. Right: Diagnostic regions located at the edge of the A- domain used to distinguish between Sl-Enc1 and Sl-Enc2. 87

Supplemental Figure 2-13: Cysteine mutant shell assembly and crosslinking mass spectrometry. (**A**) Negative stain TEM micrograph of a mixed shell sample containing His-Sl-Enc1 and Strep-Sl-Enc2. (**B**) Negative stain TEM micrograph of a mixed shell sample containing mutant His-Sl-Enc1(R195C) and Strep-Sl-Enc2. (**C**) Negative stain TEM micrograph of a mixed shell sample containing mutant His-Sl-Enc1(R195C) and mutant Strep-Sl-Enc2 (G198C). (**D**) SDS-PAGE analysis of crosslinked double-cysteine mutant sample used for mass spectrometric analysis. M: molecular weight marker. XL-S: Crosslinked His-Sl-Enc1(R195C) + Strep-Sl-Enc2 (G198C) shells. 88

Figure 3-1: *Myxoccocus xanthus* Family 2B two-component encapsulin operon with a geosmin synthase (GeoA) cargo and heterologous production. (**A**) A two-component encapsulin operon containing a GeoA cargo identified in *M. xanthus*. (**B**) Negative stain TEM micrographs of individually expressed and purified Mx-Enc1 and Mx-Enc2..... 108

Figure 3-2: Analysis of mixed Mx-Enc1 and Mx-Enc2 shells. (**A**) Negative stain TEM micrograph of purified mixed shells with singly His-tagged Mx-Enc1 (top row) and a micrograph of mixed shells with singly His-tagged Mx-Enc2 (bottom row). (**B**) SDS-PAGE of purified mixed encapsulin shell samples. M: Molecular weight marker E: Encapsulin shell. The SDS-PAGE bands shown in the box were excised and subjected to tryptic digestion and subsequent LC-MS analysis. Peptide-spectrum matches (PSMs) from LC-MS for each shell component are listed. (**C**) Dynamic light scattering (DLS) analysis of His-Mx-Enc1 + Mx-Enc2 shells (top row) and Mx-Enc1 + His-Mx-Enc2 shells. PDI: Polydispersity index..... 109

Figure 3-3: Cargo loading of geosmin synthase into Mx-Enc shells. (**A**) SEC of mixed Mx-Enc1/Mx-Enc2 shells with GeoA (Superose 6 column). The major peak (MP) is consistent with the size of T=1 encapsulin shells. (**B**) SDS-PAGE of SEC MP showing bands for GeoA and Mx-Enc1/Mx-Enc2. (**C**) Negative stain TEM micrograph of cargo-loaded mixed shells. (**D**) SEC of Mx-Enc1 shells with GeoA (Superose 6 column). (**E**) SDS-PAGE of SEC MP showing

bands for GeoA and Mx-Enc1. (**F**) Negative stain TEM micrograph of cargo-loaded Enc1 shells. 111

Figure 3-4: Structural analysis of Mx encapsulin shells via single particle cryo-EM. (**A**) Protomer structures of Sl-Enc1 and Sl-Enc2. Structures are colored by domain. (**B**) Homogeneous Mx-Enc1 shell with T=1 icosahedral symmetry. The HK97-fold is colored purple, and the CBD is colored pink. (**C**) Exterior view of the five-fold pore in the homogeneous Mx-Enc1 shell. (**D**) Exterior view of the three-fold pore in the homogenous Mx-Enc1 shell. (**E**) Exterior and interior views of the two-fold pore in the homogenous Mx-Enc1 shell. Pore diameters are highlighted. 113

Figure 3-5: Structural comparison of two-component encapsulin protomers. Structural alignment of the following protomers (left to right): (**A**) Mx-Enc1 with Mx-Enc2; (**B**) Mx-Enc1 with Sl-Enc2; (**C**) Mx-Enc2 with Sl-Enc1. For the Sl-Enc1 and Sl-Enc2 protomers, the AlphaFold predictions for the CBDs were docked into the HK97-fold domain models. (**D**) Sl-Enc1/Sl-Enc1 interactions at the two-fold pore. N-arm is shown in blue. (**E**) Predicted Mx-Enc2/Mx-Enc2 interactions at the two-fold pore. N-arm is shown in blue. 115

Figure 3-6: GeoA cargo loading into Mx shells. (**A**) AlphaFold predicted structure of *M. xanthus* GeoA with the N-terminal domain (NTD) in dark orange, the C-terminal domain (CTD) in blue, and the intergenic and C-terminal regions in yellow. The predicted targeting peptide sequences for *M. xanthus* GeoA are compared to the overall consensus sequences. (**B**) Operon structure showing the tested GeoA mutant with a truncated C-terminus and mutated TP in the intergenic region, co-expressed with two His-tagged shell proteins. Underneath the operon diagram is the corresponding SEC chromatogram, with an SDS-PAGE gel showing the protein composition of the SEC's major peak (MP). (**C**) Cut-away view of an Mx-Enc1 shell with extracted cryo-EM densities for the TP. 118

Figure 3-7: Geosmin synthase solubility test. SDS-PAGE was used to analyze the relative solubility of N-terminally His-tagged GeoA after cell lysis and centrifugation, showing the pre-lysis cell resuspension (CR), lysis supernatant (S), and lysis pellet (P) fractions. Lysis buffer conditions #1-8 used to resuspend the cell pellets are listed in Table 3-1. The yellow arrow points to the overexpressed GeoA band (86.9 kDa). 120

Figure 3-8: Geosmin synthase enzymatic activity. (**A**) Reaction scheme for the conversion of farnesyl pyrophosphate (FPP) to geosmin. The N-terminal domain (NTD) catalyzes the first step in the reaction, and the C-terminal domain (CTD) catalyzes the second step. (**B**) GC-MS analysis of GeoA enzymatic activity. Incubation of GeoA with FPP yielded geosmin, with its characteristic mass fragment at m/z +112 identified as the most abundant ion. The top chromatogram is from a geosmin synthase standard, and the bottom is from Mx-Enc1_Mx-Enc2_GeoA incubated with FPP. (**C**) Measurement of geosmin synthase activity for Mx-Enc1_GeoA and Mx-Enc1_Mx-Enc2_GeoA at two different substrate concentrations. Measurements were performed in triplicate, and error bars represent the standard deviation. .. 121

Abstract

Encapsulin nanocompartments are a newly discovered class of prokaryotic subcellular organelles that compartmentalize dedicated enzymes within a protein shell. Encapsulins play important and diverse roles in cellular metabolism and have been shown to act as storage containers for essential nutrients, improve the microbial cell's resistance to oxidative stress, and facilitate isoprenoid biosynthesis. Their specific biological functions depend on the encapsulated cargo enzyme, which is targeted to the shell's interior via a targeting peptide or domain. The encapsulin shell is constructed from many copies of a single protomer containing the HK97-like fold, which self-assembles to form icosahedral structures of a defined shape and size. Thousands of new encapsulin operons in prokaryotic genomes have recently been identified, including a new operon class containing two shell genes, termed "two-component encapsulins." All previously studied encapsulins have only been shown to use a single protein component for shell assembly. The overarching aim of this dissertation was to determine the two-component encapsulin structures and the basis for mixed shell assembly behavior to better understand the role of an additional shell protein in overall encapsulin function and regulation.

To address this aim, the structural characterization of a two-component encapsulin shell from *Streptomyces lydicus* was performed in the absence of enzymatic cargo, as described in Chapter 2. I establish the differential assembly behavior of the two shell components from *S. lydicus* and demonstrate their ability to form mixed encapsulin shells. Using cryo-EM, the structures of both shell proteins were determined. Subsequent 3D-classification analysis of cryo-

EM data and crosslinking studies show that the two-component mixed encapsulin shell uses an irregular tiling pattern with a variable shell composition. Chapter 3 describes the structural and biochemical characterization of a two-component encapsulin from *Myxoccocus xanthus*, specifically in the presence of its enzymatic cargo, a terpene cyclase. I again demonstrate the differential assembly behavior of the two shell components and confirm their ability to form mixed shells in the presence and absence of cargo. Cryo-EM analysis enabled the determination of both shell protomer structures and showed that cargo loading is mediated at the three-fold axis of symmetry. I further identify two strongly conserved targeting peptides on the terpene cyclase cargo and demonstrate their importance in facilitating cargo loading. This work also establishes that the terpene cyclase cargo is a geosmin synthase and demonstrates that its enzymatic activity is comparable between encapsulins with different shell protein compositions.

Altogether, this work presents the first demonstration of an encapsulin to incorporate two shell components and also represents the first known HK97-fold assembly to utilize two components with a random tiling pattern. These results provide critical insight into the structural basis for mixed shell assembly in two-component encapsulins.

Chapter 1 Introduction

1.1 Encapsulin nanocompartments

Intracellular organization is a fundamental feature of all biological systems (1). Though prokaryotes are defined by their lack of a nucleus and general absence of lipid-bound organelles, they have been shown to use protein-based methods of compartmentalization for spatial organization (2, 3). The localization of enzymes in a distinct reaction space can confer a number of distinct advantages for catalysis and, in turn, for the cell. Enclosed spaces can enable the co-localization of multiple enzymes in a pathway to ensure efficient catalysis of a desired product. Further, the presence of a physical barrier can prevent enzymatic intermediates from leaving the reaction space and exclude unwanted proteins or molecules from impacting the reaction(s) within (4). Enclosed reaction spaces may also provide a more optimized environment for a specific redox state or pH.

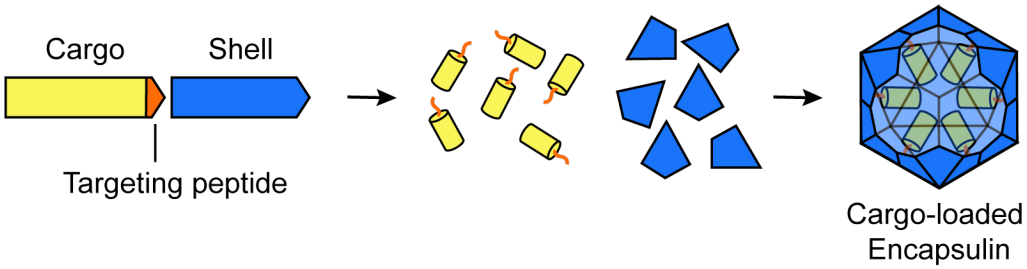


Figure 1-1: Standard encapsulin operon and loading schematic. Encapsulin "Enc" shell proteins (blue) and cargo proteins (yellow) co-assemble to form cargo-loaded encapsulin shells. A targeting peptide on the cargo's terminus (orange) interacts with the shell, mediating cargo encapsulation.

Encapsulins are a newer class of protein nanocompartment found in many prokaryotic phyla that compartmentalize specific metabolic processes (5, 6). Encapsulins generally use only two protein components, a single shell protein and cargo enzyme, originating from the same

operon (Figure 1-1). The encapsulin shell protomers utilize the HK97-fold and self-assemble to form an icosahedral shell with a defined size ranging from 18 to 42 nm in diameter (7-9). Cargo enzymes are enclosed within the encapsulin shell and targeted to the shell interior via a targeting peptide (TP) or targeting domain (TD) located on the cargo's N- or C-terminus (10-13). Encapsulins sequester a variety of cargo enzymes including ferroxidases, peroxidases, desulfurases, and terpene cyclases. The biological function of encapsulins is dependent on the cargo enzyme. They have been found to serve as iron or sulfur storage compartments (8, 9, 14), facilitate oxidative stress resistance (15), and for other cargoes, may simply provide an optimized and/or regulated reaction space (16, 17). Over the last seven years, the mining of bacterial and archaeal genomes has resulted in the discovery of a plethora of putative encapsulin operons in novel prokaryotic phyla with new putative cargo enzymes and containing other novel genetic and structural features (18-20). As a result, many recent studies have focused on the characterization of these newly identified encapsulins. Though much progress has been made in characterizing new encapsulins during this period, only a small subset of encapsulin systems have been biochemically or structurally characterized. Moreover, even less is understood about how some encapsulins contribute to the overall fitness of their host organisms. In this section, I will discuss encapsulin nanocompartment biology and review the existing literature regarding their distribution across prokaryotic phyla, protein structure, and biochemical function.

1.1.1 Genetic distribution and classification

Using large-scale computational searches, over 6,000 putative encapsulin operons have been identified in 31 bacterial phyla and 4 archaeal phyla (19). They are classified into four

distinct families based on the phylogeny of shell proteins and the encapsulin operon structure (19, 20) (Figure 1-2).

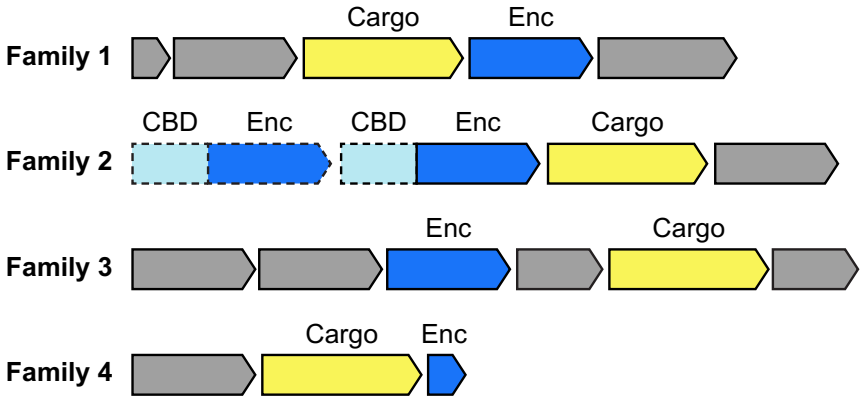


Figure 1-2: Operon structure of the four encapsulin families. Predicted encapsulin shell genes (Enc) are blue, cargo genes are yellow, cyclic nucleotide-binding domains (CBDs) are light blue, and neighboring accessory proteins are gray. Figure adapted from T. W. Giessen, 2022 (17).

Family 1 is the most widespread encapsulin class, as operons are found in 31 prokaryotic phyla and comprise 40% of all identified encapsulin operons to date (19). The general operon structure is very well conserved between the different cargo types and generally consists of the cargo gene located directly upstream of the encapsulin shell gene (PF04454) (Figure 1-2). Nearly all cargoes found in Family 1 encapsulin operons are dye-decoloring peroxidases (DyPs) or in the ferritin-like protein (Flp) superfamily. DyPs, which are named for their ability to oxidize anthraquinone dyes (21), are most commonly found in Actinobacteria and Proteobacteria and comprise about a third of all Family 1 cargoes. Cargoes in the Flp superfamily are predicted to share a common structure – a four-helix bundle fold – and encompass seven types of enzymes: Flp, rubrerythrin, CoQ7, Mn catalase, hemerythrin, bacterioferritin, and iron-mineralizing encapsulin-associated firmicute (IMEF) (17). From the Flp superfamily, only Flp, hemerythrin, and IMEF encapsulin cargoes have been studied.

Family 2 (PF19307) is the largest class of encapsulins and varies from Family 1 in its sequence similarity and a differing operon structure (19, 20). This is the only family that is divided into two subfamilies based on the absence (Family 2A) or presence (Family 2B) of a putative cyclic nucleotide-binding domain insertion (CBD) in the encapsulin shell (Figure 1-2) (19). Accessory proteins are observed in select Family 2A and 2B operons, some of which are predicted to contribute to the biological function of the encapsulin system (14, 22). Nearly all of Family 2A operons contain a cysteine desulfurase as the putative cargo (97%), with a minority of operons containing unidentified or terpene cyclase cargoes. Family 2A cysteine desulfurase operons are found in 12 bacterial phyla and are most abundant in Proteobacteria. The majority of Family 2B operons constitute the majority of Family 2 encapsulins (63%) and are found in 9 phyla with 4 putative cargo types: terpene cyclases, polyprenyl transferases, xylulose kinases, and cysteine desulfurases. The cNMP-binding domains that are the unifying feature of Family 2B encapsulins are most commonly found in transcriptional regulators, ion channels, and protein kinases (23-25) and can bind the cyclic nucleotides and second messengers cAMP, cGMP, and c-di-GMP (26). The overall role of the CBDs in encapsulin function and their ligands is thus far unknown.

Family 2B is also the only class where encapsulin operons encoding two discrete shell proteins are found. Although between 15% and 20% of encapsulin-containing organisms possess multiple encapsulin operons within their genomes, no encapsulin system has been experimentally shown to utilize protein components outside of a single operon (19). Prior to this dissertation work, no two-component encapsulin has been experimentally characterized, nor has any encapsulin been shown to utilize more than one shell protein. These two-component operons are associated with all cargo types reported for Family 2B operons and are predominantly found in

species of Actinobacteria, although such operons have been reported in four other bacterial phyla.

Encapsulin families 3 and 4 are much smaller, each accounting for only 2% of all known encapsulin operons (19). Family 3 operons are distinguished by the large biosynthetic gene clusters that surround the encapsulin shell genes and have been identified in Actinobacteria, Proteobacteria, and Chloroflexi. Family 4 encapsulins are the most structurally distinct class due to the deletion of several conserved regions of the HK97-fold in the shell protomer, leaving only the A-domain. A-domain encapsulin operons are found exclusively in thermophilic anaerobes within the archaeal phylum Euryarchaeota and the bacterial phylum Bacteroidetes. Four putative cargoes have been identified for Family 4 encapsulins, which are a hydrogenase, deoxyribose-phosphate aldolase (DeoC), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and osmotically inducible protein C (OsmC). In all operons, the A-domain is located directly downstream from the cargo gene. No encapsulins have been experimentally characterized from either family.

1.1.2 Shell structure

The HK97 protein fold plays an important role in disease and metabolism across all domains of life. The HK97-fold was named after the virus from whose major capsid protein the structure was first determined, the Hong Kong 97 bacteriophage (27). This protein fold is found in the major capsid proteins of icosahedral viruses that infect archaea and bacteria of the Caudovirales order (28, 29), eukaryotic viruses of the order Herpesvirales (30), and in the shells of encapsulin nanocompartments. Encapsulins are defined by their use of the HK97-fold to assemble protein nanocompartments which load enzymatic cargo. The encapsulin shells' use of

the HK97-fold is suggestive of an evolutionary relationship with viruses, and encapsulins may even be viral-derived (17, 19). They are the only non-viral structures that utilize the HK97-fold and, as such, have some structural differences that enable them to optimally serve their specific purpose as prokaryotic organelles rather than infectious agents (8, 17, 19).

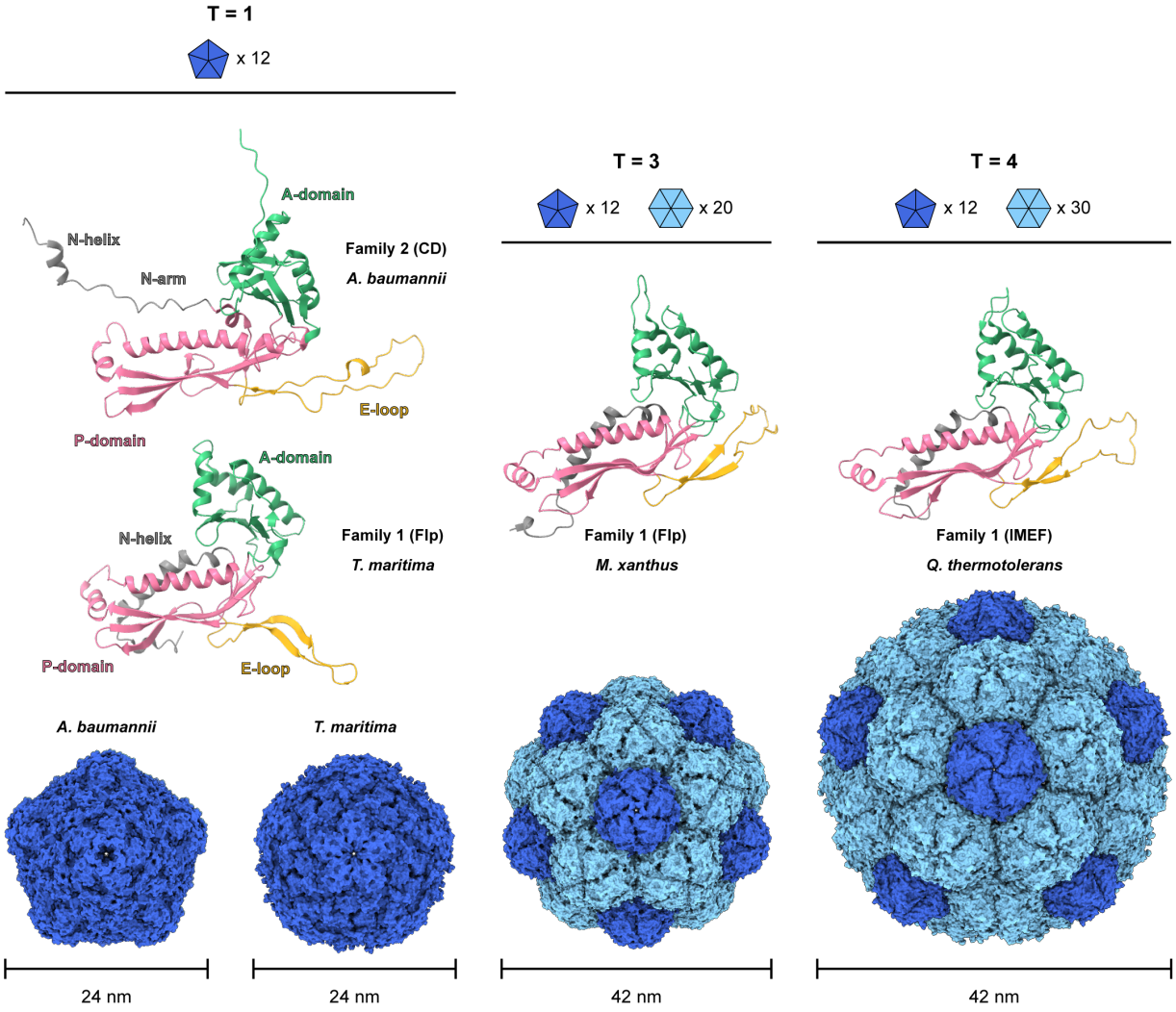


Figure 1-3: Encapsulin shell structure. Encapsulin shell protomers exhibiting the HK97-fold are colored by region. Assembled encapsulin shells of varying triangulation numbers, with pentamers colored in dark blue and hexamers colored in light blue. The sole Family 2 encapsulin is from *A. baumannii* (PDB: 8T6R) with a cysteine desulfurase (CD) cargo. Family 1 encapsulins shown: *T. maritima* (PDB: 3DKT) with a ferritin-like protein (Flp) cargo, *M. xanthus* (PDB: 8TK7) with a dye-decoloring peroxidase (DyP) cargo, and *Q. thermotolerans* (PDB: 6NJ8) with an iron-mineralizing encapsulin-associated Firmicute (IMEF) cargo.

Encapsulins are the smallest known icosahedral assemblies to utilize the HK97-fold (Figure 1-3), ranging in diameter from 18 nm with a triangulation number (T) of T=1 (60

protomers) to 42 nm (T=4, 240 protomers). Despite the significant disparity in assembly size, the structures of all HK97-fold proteins contain three highly conserved domains: the axial domain (A-domain), peripheral domain (P-domain), and extended loop (E-loop) (Figure 1-3). The A-domain is triangular in shape and is responsible for mediating contacts at the five-fold and six-fold axes of symmetry. The P-domain, as its name implies, is located on the periphery of the pentamers and hexamers formed by the A-domain and mediates protein-protein interactions at the two-fold and three-fold symmetry axes. The E-loop is a long and flexible region that generally interacts with the P-domains of adjacent subunits and is thought to be the primary determinant of encapsulin triangulation number and shell size based on its length and conformation (8, 17). Furthermore, for a given protomer, the A-domain and E-loop can flexibly adopt different conformations to accommodate the assembly of both hexamers and pentamers (9, 17), which are necessary for the complete assembly of T=3 and larger icosahedral shells (31, 32). For both encapsulins and viruses, the E-loop is also flexible enough to accommodate a variety of insertion domains (33, 34) while maintaining the structural integrity of the icosahedral shell or capsid. Aside from modest differences in E-loop length, the only insertion domains found in encapsulins to date are the CBDs present in all Family 2 systems, which are around 120 residues in length (~20 kilodaltons) (16, 19).

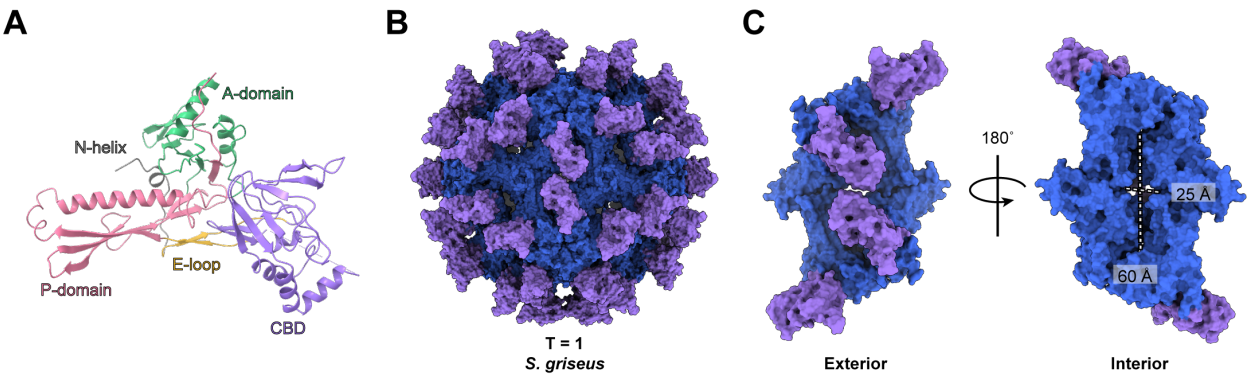


Figure 1-4: Encapsulin Family 2B protomer and shell structure from *S. griseus*. (**A**) Encapsulin shell protomer (PDB ID: 9BHU) exhibiting the HK97-fold and a cyclic nucleotide-binding domain (CBD). Regions are labeled by color. (**B**) Assembled encapsulin shell. The HK97-fold (blue) forms the icosahedral shell structure, with the CBDs (purple) displayed on the shell's exterior. (**C**) Interior and exterior views of the shell's two-fold pores.

Currently, only one structure has been solved for a Family 2B encapsulin containing a CBD insertion, which was isolated from *Streptomyces griseus* and encapsulates a terpene cyclase cargo (Figure 1-4, A and B) (16). The CBDs are displayed externally and are positioned overtop the two-fold pores, which, at 60 x 25 Å in dimension, are the largest natively occurring encapsulin pores reported to date (Figure 1-4C). Researchers also noted that the cryogenic electron microscopy (cryo-EM) densities for the CBDs had visibly worse resolution than the HK97-fold, indicating a greater degree of flexibility compared to the rest of the protomer. As predicted, the CBD was found to adopt a β-barrel structure that is very similar to the cAMP-binding domain of the *E. coli* catabolite activator protein (CAP) (35). However, CBDs were not experimentally found to bind cAMP, and a further comparative analysis of the cAMP binding pocket in CAP demonstrated that significant differences in hydrophobicity likely contribute to this inability to bind cAMP. Due to their positioning over the major encapsulin pores and positioning flexibility, CBDs may play a role in regulating the encapsulin function by modulating the accessibility of these pores. Since the encapsulin's pores are the only way for molecules to enter or exit a fully assembled shell, regulating the accessibility of the largest pores would have a significant impact on enzyme activity by controlling substrate availability. Though

cAMP does not bind to the *S. griseus* encapsulin CBD, it is still plausible that other ligands may bind to the CBD and induce a conformational change analogous to cAMP binding of *E. coli* CAP.

Aside from the CBDs found only in Family 2B, there are some minor differences between encapsulin shell structures based on their family assignment. Structures of Family 1 shells have the broadest range in size, with T=1, T=3, and T=4 shells (7-9), while all three solved structures of Family 2 systems are T=1 (14, 16, 20). However, there are too few examples to know whether this difference in shell size distribution is a true feature or a product of a small sample size. Though no data is currently available for Family 4 encapsulins, they are predicted to be the most structurally distinct class based on the genetic absence of the P-domain and E-loop – features common to all other HK97 structures – and may very likely not assemble into icosahedral shells at all (17, 19). Family 1 and 2 encapsulins also use different mechanisms of cargo loading and facilitate interactions with cargo at different regions on the shell protomers, though this is discussed more in-depth in Section 1.1.4, titled *Targeting Peptides and Cargo Loading Mechanisms.* More notably, Family 2A encapsulin protomers contain an extended N-terminal arm (N-arm) that ends with a short helix, akin to those found in HK97-fold viral capsids (14, 20, 33). N-terminal helices are present in Family 1 encapsulins, though the extended N-arm is absent. The N-arms in the two Family 2A encapsulins are relatively long and followed by a disordered and unresolved N-terminal extension. They are positioned at the two-fold symmetry axis where they directly interact, obstructing what would otherwise be a large two-fold pore similar to that observed in the Family 2B encapsulin shell (16).

Encapsulins are very stable and homogenous assemblies, generally forming shells of a defined shape and size. The differences in T number and overall shell size observed between

HK97 structures are primarily attributed to subtle sequence and structural conformation variations in the shell protomer, and for viruses, the addition of scaffolding domains or proteins. For encapsulins specifically, protomers will nearly always form shells of a consistent size and shape, regardless of the presence of cargo. The single known exception to this rule is an Flp encapsulin from *Myxococcus xanthus*, which has been shown to assemble both T=1 and T=3 shells in the absence of cargo (36). When both the cargo and shell are expressed, homogenous T=3 shells are observed. Other encapsulins have been shown to have very dynamic pores, as is the case for two separate encapsulins from *Haliangium ochraceum* and *Acidipropionibacterium acidipropionici*. Two distinct states of the five-fold pore were resolved from cryo-EM data of the *H. ochraceum* encapsulin, with pores measured at 5 Å and 15 Å in diameter (37). This work is the first evidence that pores may not be as static as previously thought, though the dynamics are likely specific to each type of encapsulin. Furthermore, this change in pore size is significant because it represents a dramatic difference in the capacity for metabolite flux through the shell. Even larger dynamic changes have been observed in the five-fold pores of the *A. acidipropionici* encapsulin shell. These pores have similarly been observed in two discrete states, with diameters of 5 Å and 20 Å, as visualized with cryo-EM (38). Researchers also determined that this change was highly correlated with the pH of the buffer solution, with more acidic conditions (pH 3.0) resulting in a greater proportion of open pores.

1.1.3 Enzymatic cargoes and their implications in prokaryotic metabolism

Encapsulins are found to sequester a diverse group of enzymes that enable them to serve different functions for their host cell. Only a handful of encapsulin cargoes have been confirmed and studied, all of which are from Family 1 and Family 2. To study the physiological role of

encapsulins, scientists have used a combination of *in vitro* biochemical characterization techniques as well as *in vivo* studies conducted with the native microbes.

The first recognized encapsulin to be biochemically characterized was from *Thermotoga maritima*, whose cargo is a ferritin-like protein (8). Since then, many other Flp encapsulins, including one IMEF encapsulin, have been studied (9, 37, 39, 40). Flps and other cargoes exhibiting ferroxidase activity are only found in Family 1 encapsulins and catalyze the oxidation of ferrous iron (Fe^{2+}) to ferric iron (Fe^{3+}), resulting in the formation of large precipitates called ferrihydrites (Figure 1-5A) (9). It is predicted that the largest encapsulin, a T=4 shell with an IMEF cargo, can store up to 23,000 iron atoms. Studies in *Myxococcus xanthus* have shown that when the encapsulin shell protein was deleted from the genome, the cells suffered from defective sporulation and were much more susceptible to oxidative damage from hydrogen peroxide (39, 41). These combined data support the hypothesis that encapsulins with cargoes that perform ferroxidase activity serve as iron storage organelles, though it is currently unknown if or how the cell can remobilize the ferric iron for later use.

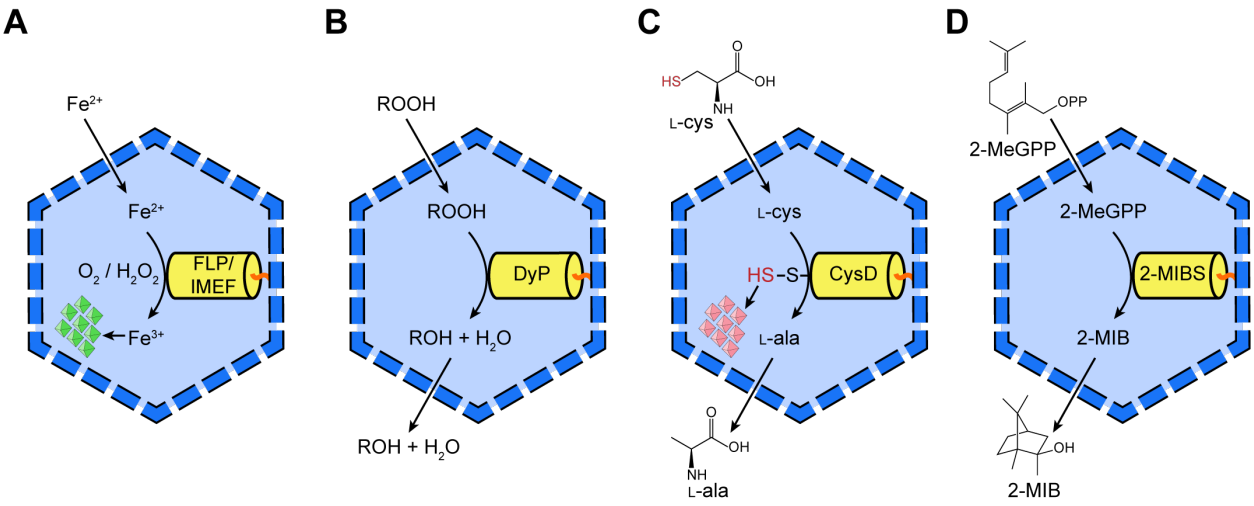


Figure 1-5: Encapsulin cargoes and their biochemistry. (**A**) Ferritin-like proteins (Flps) and iron-mineralizing encapsulin-associated Firmicute (IMEF) cargoes exhibit ferroxidase activity in which ferrous iron is oxidized to ferric iron. Ferric iron is insoluble and forms precipitates known as ferrihydrite. (**B**) Dye-decolorizing peroxidases (DyPs) catalyze the oxidation of many substrates, including most anthraquinone dyes. (**C**) Cysteine desulfurase (CD) catalyzes the conversion of L-cysteine to L-alanine and forms a persulfide intermediate bound to CysD. It is predicted that the continued activity of CysD leads to the creation of long poly-persulfide chains, resulting in the formation of sulfur crystals. (**D**) 2-Methylisoborneol synthase (2-MIBS) catalyzes the conversion of 2-methylgeranyl pyrophosphate (2-MeGPP) to 2-methylisoborneol (2-MIB).

The most abundant cargo type in Family 1 is the promiscuous DyP (Figure 1-5B), which catalyzes the heme-dependent oxidation and degradation of its substrate while simultaneously reducing hydrogen peroxide to water. Encapsulin-associated DyPs have been shown to break down lignin and dyes in addition to other standard peroxidase substrates, but their native substrates remain unknown (42, 43). All solved structures of DyP-loaded encapsulins are T = 1 (12, 44), though surprisingly, different oligomeric states of the enzyme have been observed. The DyPs of encapsulins from *Klebsiella pneumoniae* and *Brevibacterium linens* were found to form a dimer of trimers inside the shell (12, 44). For *K. pneumoniae*, the cargo loading efficiency was also very low likely due to the size of the complex, with an estimated 12 DyP monomers (two hexamers). Interestingly, another study using cryo-EM found the encapsulated DyP hexamers of *Mycobacterium smegmatis* to interact, forming a dodecameric (12-mer) structure (45). A number of benefits of DyP encapsulation have been proposed, though further investigation is still required to better understand the role of DyP encapsulins in prokaryotic physiology. Due to

differences in the oligomeric state sometimes observed between DyPs inside and outside of the shell, it is possible that the shell plays a role in stabilizing the enzyme's oligomeric state for optimal enzyme activity (12). A study of the DyP encapsulin from the pathogen *Mycobacterium tuberculosis* showed that they significantly improve the ability of cells to survive in conditions of low pH and high reactive oxygen species (ROS), suggesting an overall function in oxidative stress resistance (15).

In addition to serving as iron-storage organelles, encapsulins have also been found to store elemental sulfur (14). Family 2A encapsulins accomplish this through the compartmentalization of a cysteine desulfurase (Figure 1-5C), which catalyzes the pyridoxal-5′-phosphate (PLP) conversion of L-cysteine to L-alanine and forms a persulfide intermediate bound to the enzyme (46, 47). In the cyanobacterium *Synechococcus elongatus* and the proteobacterium *Acinetobacter baumannii,* both the encapsulin shell and cysteine desulfurase cargo are upregulated during sulfur starvation conditions, suggesting a role in sulfur storage (14, 20). Further analysis of the cargo-loaded encapsulins showed the presence of electron-dense puncta which were determined to be elemental sulfur, likely stored as long poly-persulfide chains that have formed sulfur crystals. In this way, the shell would function as a barrier to prevent many cellular reducing agents from accessing the poly-persulfide, successfully enabling the storage of large quantities of elemental sulfur.

To date, only one Family 2B encapsulin has been biochemically characterized. Recent work has established the terpene cyclase 2-methylisoborneol synthase (2-MIBS) as an encapsulin cargo (Figure 1-5D) (16, 48), which catalyzes the Mg^{2+}-dependent conversion of 2-methyl GPP to 2-methylisoborneol (2-MIB) (49, 50). Though 2-MIBS encapsulin operons have only been identified in Actinobacteria (19), 2-MIB is also produced in species of Myxobacteria and

Cyanobacteria (49, 51, 52), demonstrating that encapsulation is not a requirement for 2-MIB production.

Enzymatic activity assays and GC-MS analysis of an encapsulated 2-MIBS from *Streptomyces griseus* confirmed 2-MIBS enzyme activity and 2-MIB formation, although no significant difference was observed in enzyme activity between encapsulated and unencapsulated 2-MIBS. This is the first encapsulin to be biochemically characterized from Family 2B and as such, its encapsulin shell protomer contains the characteristic cyclic nucleotide-binding domain (CBD) which was found to be displayed on the shell's exterior. Surprisingly, cyclic adenosine monophosphate (cAMP), a prolific second messenger essential for many biochemical processes, was not found to bind to the shell protomer's CBDs and had no impact on enzyme activity.

Ultimately, the cellular advantage of encapsulating 2-MIBS remains to be determined, since encapsulation is not a requirement for 2-MIB production (49, 51). One possibility is that the encapsulin shell provides a method of regulating 2-MIB production and that the binding of an appropriate ligand to the CBDs will induce a conformational change that alters the size of the encapsulin shell pores. This change in pore size would contribute to the shell's permeability and therefore would impact the ability of 2-MIBS to access substrate, potentially inhibiting the synthesis of 2-MIB. However, the biological function of the molecule 2-MIB is still unknown, further obscuring the benefit of enzyme encapsulation and potential regulation.

1.1.4 Targeting peptides and cargo loading mechanisms

The study of cargo loading mechanisms is essential to the understanding of encapsulin biology, as it is the sole determinant of which protein(s) are targeted to the shell's interior. Further, this knowledge can enable researchers to load non-native cargoes into the protein shell,

thus engineering novel encapsulins for uses in targeted drug delivery, vaccines, catalysis, and more (53-55). Though only encapsulins from two families have been experimentally characterized, distinct mechanisms of cargo loading and targeting peptide structure are observed within each family.

For all Family 1 encapsulins, a short TP is located on the C-terminus of the encapsulin cargo, separated from the globular domain with a flexible linker between 10 and 50 residues in length (10). Targeting peptides interact with a hydrophobic patch between the N-terminal helix and P-domain. As determined from analyses of several cyro-EM datasets, the TP regions that interact with the shell rigidly are usually 7-12 residues long (10). These sequences generally consist of two to three hydrophobic residues, such as isoleucine, leucine, or valine, spaced one to two residues apart with a glycine, serine, or both (8, 9, 36, 37). Positively charged residues, such as arginine and lysine, will sometimes follow the consensus targeting sequence and are seen to interact less rigidly with nearby negatively charged patches on the shell protomers. TPs interact exclusively with a single protomer and do not cross any axes of symmetry on the protein shell, indicating that, in principle, the maximal cargo loading corresponds to the number of protomers that comprise the shell. However, experimental data has demonstrated that maximal cargo loading is highly dependent upon the cargo's size and oligomerization state due to the limited space within the shell (10). In all predicted Family 1 operons, the gene encoding for the cargo enzyme directly neighbors the encapsulin gene (19). Since the encapsulin shells self-assemble spontaneously and do not require cargo loading for assembly, the close proximity of the two genes likely helps facilitate interactions between the TPs and shell protomers immediately after translation, promoting efficient cargo loading.

Rather than using short, discrete targeting peptides to load their cargo, all so-far studied Family 2 encapsulins require targeting domains instead. Aside from xylulose kinases, all predicted Family 2 cargoes have long, disordered, and unannotated regions located at their termini (19). These disordered regions have been experimentally shown to serve as TDs for cysteine desulfurase and 2-methylisoborneol synthase cargoes, while this remains unconfirmed for the geosmin synthases and polyprenyl transferases. Cysteine desulfurases have the longest disordered regions of any Family 2 cargo, which are around 225 residues in length and located at the N-terminus (19). The *Synechococcus elongatus* cysteine desulfurase encapsulin (Family 2A) was found to require the full-length TD for optimal cargo loading and interacts with the shell lumen around the three-fold pores (20). Interactions between the TDs and shell protomers at the three-fold pore were also observed for another cysteine desulfurase encapsulin originating from *Acinetobacter baumannii 118362* (14). For the *A. baumannii* encapsulin, additional cryo-EM density was found at the A-domain and was not connected to the density observed at the three-fold axis. Both TD densities were found to interact with primarily nonpolar and hydrophobic surface patches on the shell protomer, though the resolution of the cryo-EM data was too low to determine either of the TD sequences.

For a 2-MIBS cargo (Family 2B), cryo-EM density was again observed at the three-fold axes, though the density quality was too poor to assign sequences. To determine if one or more specific regions of the 2-MIBS TD are responsible for cargo loading, the 120-residue domain was divided into five regions based on the predicted secondary structure, and the effects of the systematic deletion of these regions on cargo loading efficiency were subsequently investigated. The results of this experiment show that the most efficient cargo loading requires the full NTD, although select regions were found to have more of an effect on cargo loading than others.

Altogether, the cumulative structural data from three cargo-loaded Family 2 encapsulins suggests that multiple regions of the TDs interact with the shell and are required for effective cargo loading. Further studies are needed to better understand the mechanism of cargo loading for Family 2 encapsulins, including more structural analyses to determine the specific interactions between the TDs and shell.

1.2 Geosmin Synthase

Terpenoids comprise the largest and most structurally diverse families of natural products, containing over 76,000 characterized compounds (56, 57). Terpenoids are very versatile and have been utilized in industry as fragrances and flavoring agents, and even have medicinal uses (58, 59). Terpenoids, which are also known as isoprenoids, are all built from the same basic five-carbon (C_5) isoprene units: dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP) (56, 60). These isoprene units are first joined linearly to generate longer C_{5n} isoprenoid diphosphates like the monoterpenoid (C_{10}) geranyl pyrophosphate (GPP), sesquiterpenoid (C_{15}) farnesyl pyrophosphate (FPP), and diterpenoid (C_{20}) geranylgeranyl pyrophosphate (GGPP). Subsequently, these molecules undergo enzyme-catalyzed modifications, such as cyclization and rearrangement reactions, to yield a large and structurally diverse expanse of terpenoid products (56, 60, 61). Terpene cyclases (TCs) are a class of enzymes largely responsible for generating the incredible diversity of terpenoid compounds found in nature by catalyzing these cyclization reactions. In this section, I will discuss the putative encapsulin cargo and terpene cyclase, geosmin synthase.

1.2.1 Geosmin: An earthy smelling terpenoid

Geosmin is among the first terpenoids to be isolated and characterized from *Streptomyces* (62), an Actinobacterial genus renowned for its wide phylogenetic spread and production of a plethora of terpenoid natural products (63, 64). Geosmin, whose name originates from the Greek words “ge” for “earth” and osme” for odor” (62), is a strong odorant that is responsible for the characteristic earthy smell of wet soil (65) and is a significant contributor to the musty flavors found in contaminated drinking water, fish, and wine (66-69). The removal of geosmin from potable water and downstream products remains a challenge to this day, and water treatment methods to eradicate geosmin in a cost-effective manner are still in development (70). The difficulty in removing geosmin largely stems from the human capacity to detect geosmin at extremely low concentrations, with geosmin having a taste and odor threshold in the parts per trillion (71).

Geosmin is produced in many species within the prokaryotic phyla of Actinobacteria (72-78), Myxobacteria (76, 79-82), and Cyanobacteria (76, 83-85). Generally, the geosmin contamination of water and aquaculture products is attributed to its production in Cyanobacteria, while the distinct scents of wet earth originate from the soil-dwelling species Actinobacteria and Myxobacteria (86). Since geosmin is a volatile compound, the first geosmin-producing bacterial strains were identified using gas chromatography-mass spectrometry (GC-MS) to analyze the organic extracts or distillates of cell cultures. Recent advancements in DNA sequencing technologies have resulted in a rapid rise in the quantity of sequenced bacterial genomes. This progress has accelerated the identification of bacteria that produce geosmin, thereby expanding the list of known geosmin-producing organisms. Geosmin synthase (GeoA) is the only known

enzyme that can synthesize geosmin (87-89), and the most recent bioinformatic analyses have identified more than 1,062 GeoA homologs across several bacterial phyla (90).

While the exact biological function of geosmin remains unknown, current evidence shows geosmin can serve as an effective signaling molecule to other organisms. One study reported that the scent of geosmin and 2-methylisoborneol – another terpenoid with a strong earthy scent – in *Streptomyces coelicolor* triggers an attraction in springtails, which consequently serve as vectors for the distribution of *Streptomyces* spores (91). Another study conducted with *Drosophila melanogaster* demonstrated that, when given the choice, flies always avoided the odor of geosmin and geosmin-producing *S. coelicolor* cultures when making feeding and breeding decisions (92). Similarly, the nematode *Caenorhabditis elegans* has been shown to consistently avoid *Streptomyces* colonies and their cell-free lysates, indicating an aversion to the secretion products of *Streptomyces* (93). When the impact of geosmin on *C. elegans* was specifically investigated, researchers found that though geosmin is not toxic, the worms generally avoided feeding on *S. coelicolor* or *Myxococcus xanthus* colonies that produced geosmin. Furthermore, worms with defective chemosensation and consequently unable to detect geosmin would more readily feed on *S. coelicolor* or *M. xanthus* colonies, resulting in the death of both the worms and the bacteria (94).

In every instance identified, geosmin functions as a signaling molecule: it acts as a deterrent to organisms that might harm it and as an attractant to those that assist in its proliferation. The hundreds of published works on geosmin, especially concerning its removal from water sources and food products, suggest that geosmin is a warning signal to humans as well. Other secondary metabolites found in geosmin-producing microbes can be toxic to humans (95) and it is likely that human consumption of such microbes would be mutually harmful.

1.2.2 Enzymatic reaction, structure, and stability

Geosmin synthase is a bifunctional enzyme that catalyzes the Mg^{2+}-dependent conversion of farnesyl pyrophosphate (FPP) to geosmin (Figure 1-6) (87, 89, 96, 97). The N-terminal domain (NTD) of GeoA catalyzes the cyclization of FPP to form inorganic pyrophosphate (PP_i), germacradienol (major product), and germacrene D (minor product) (96, 98, 99). Then, germacradienol is bound to the active site of the C-terminal domain (CTD), where the CTD subsequently catalyzes a proton-dependent cyclization reaction, a retro-Prins-type fragmentation, to produce geosmin and acetone (89). The NTD alone is incapable of producing geosmin either directly from FPP or germacradienol (87), and similarly, the CTD is unable to catalyze the cyclization of FPP (89). This demonstrates that each domain catalyzes a distinct reaction and that both are essential for geosmin synthesis. Both reactions are metal-dependent and specifically require Mg^{2+}, though Mn^{2+} can serve as a substitute divalent cation (90). GeoA does not require additional cofactors or substrates to facilitate these two reactions (89).

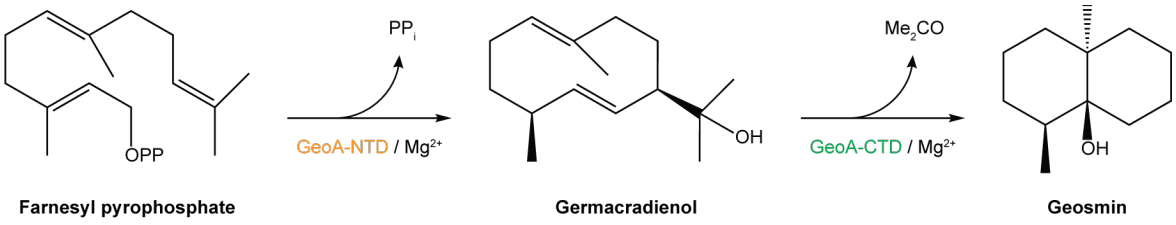


Figure 1-6: Reaction scheme for the conversion of farnesyl pyrophosphate to geosmin. Geosmin synthase's N-terminal domain (GeoA-NTD) first catalyzes the cyclization of farnesyl pyrophosphate to germacradienol. Then, the C-terminal domain (GeoA-CTD) catalyzes the cyclization-fragmentation reaction of germacradienol to geosmin.

Though no structure of full-length geosmin synthase has been determined, two structures have been solved for the NTD of *S. coelicolor* geosmin synthase using x-ray crystallography: one structure with the unliganded domain, and the other structure in complex with Mg^{2+} and alendronate, a bisphosphonate drug (100). The NTD of GeoA has an α-helical fold that is

characteristic of class I terpene synthases (56) (Figure 1-7A). In the ligand-bound protein structure, three Mg^{2+} ions in complex with alendronate are coordinated by two metal-binding motifs that are widely conserved across all terpene cyclases (56, 57). The aspartate-rich motif ($\underline{D}_{86}$DHFLE$_{91}$) contributes a single aspartate (underlined) to the coordination of the Mg^{2+} ions, while the other motif, known as the NSE triad, contributes an asparagine, serine, and glutamine ($\underline{N}_{229}$DLF$\underline{S}$YQR$\underline{E}_{237}$) (Figure 1-7B). For bacterial TCs, the consensus sequence of the longer metal-binding motif is NxxxSxxxE, which is consistent with the sequence found in the *S. coelicolor* GeoA. Different conformations of these two motifs were observed between the structures of the ligand-bound and unbound enzyme, whereupon ligand binding, the helix containing the aspartate-rich motif moved closer to the helix containing the NSE triad. These two motifs were also found in the CTD sequence ($\underline{D}_{455}$DYYP$_{459}$ and $\underline{N}_{598}$DVF$\underline{S}$YQK$\underline{E}_{606}$) and similarly predicted to coordinate Mg^{2+} ions.

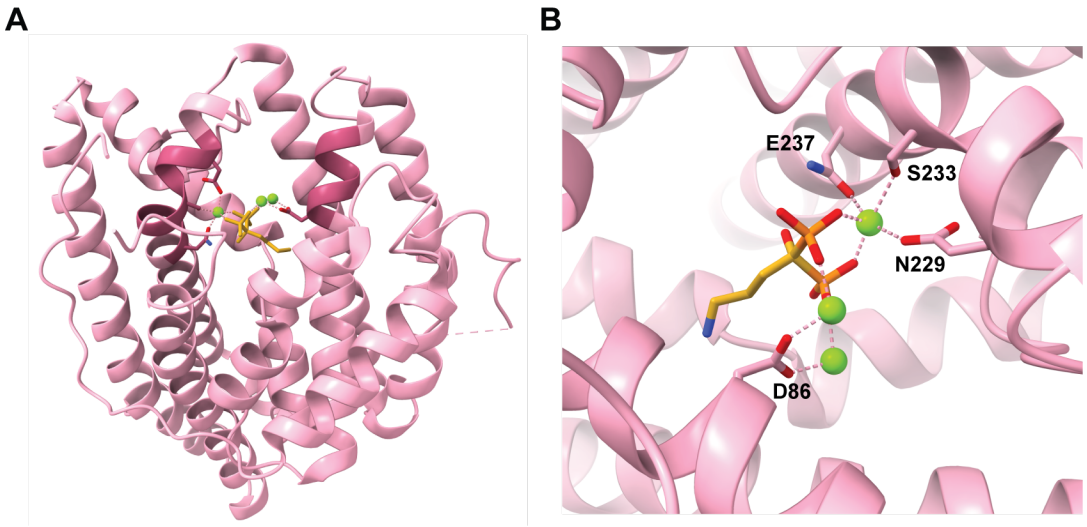


Figure 1-7: Solved structure of the ligand-bound *S. coelicolor* geosmin synthase's N-terminal domain. (**A**) Crystal structure of the GeoA N-terminal domain (PDB ID: 5DZ2) coordinating three Mg^{2+} ions (green) in complex with alendronate (yellow) (100). The two conserved metal binding motifs are colored dark pink. (**B**) Zoomed view of the NTD active site. The four residues from the metal binding motifs responsible for coordinating the Mg^{2+} ions are displayed and labeled.

Since only the structure of the NTD has been solved, the oligomeric state of the full-length geosmin synthase is unknown. The NTD-Mg^{2+}_3-alendronate complex was found to crystallize as a dimer in the asymmetric unit with the active sites antiparallel relative to each other, but these results have not been verified with the full-length protein in solution using methods such as native PAGE, analytical size-exclusion chromatography (SEC), or dynamic light scattering (DLS). The CTD is predicted to adopt an α-helical fold, similar to that of the NTD, with the two domains connected by a long, flexible linker (Figure 1-8, A and B). Current evidence suggests that substrates are not channeled between the two active sites via a substrate tunnel (89, 90); otherwise, the specific interactions between the two domains remain unknown.

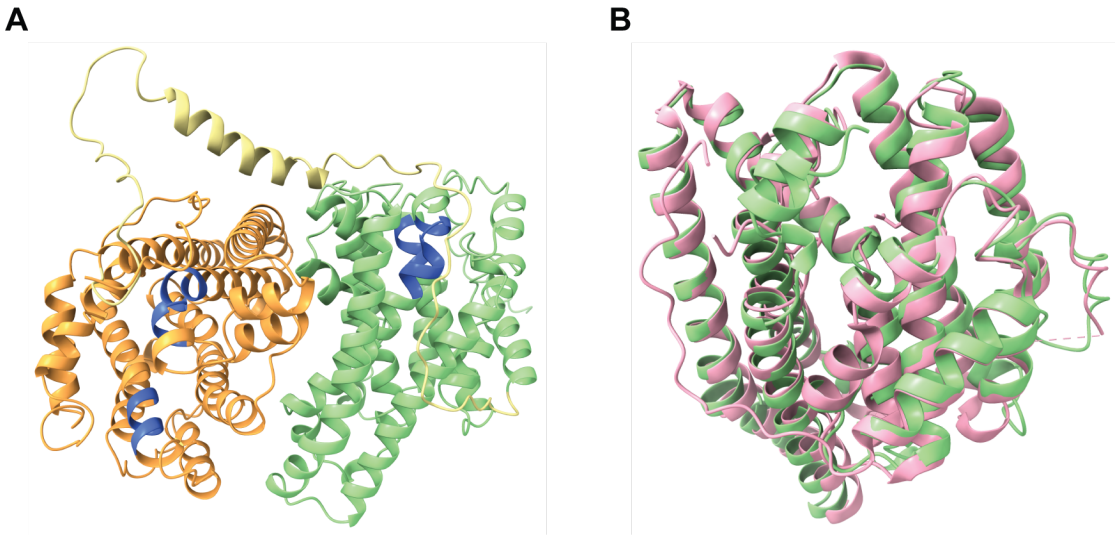


Figure 1-8: Predicted structures of *S. coelicolor* geosmin synthase. (**A**) AlphaFold model of the full-length geosmin synthase. N-terminal domain (orange) and C-terminal domain (green) are connected by a long flexible linker containing a single α-helix (yellow). Metal binding motifs (blue) are positioned away from each other. (**B**) Structural alignment of the ligand-bound geosmin synthase crystal structure (pink, PDB ID: 5DZ2) (100) and the AlphaFold model of the GeoA C-terminal domain (green).

Many geosmin synthase homologs, originating from all three bacterial phyla in which *geoA* genes are found, have been heterologously expressed, purified, and biochemically characterized. One notable feature shared by most GeoA enzymes is their low *in vitro* solubility. Three homologs of GeoA from different organisms (*Streptomyces coelicolor*, *Streptomyces*

avermitilis, and *Nostoc punctiforme* PCC 73102) that were expressed in *E. coli* were insoluble and necessitated purification from inclusion bodies (87, 99) or co-expression and purification with the chaperone protein combination GroES/GroEL (98). Only one study has reported the production of soluble, full-length GeoA, which was C-terminally His-tagged, from the Actinobacterium *Allokutzneria albata* (90). However, soluble protein could be obtained for the *S. coelicolor* enzyme after removing 36 C-terminal residues and subsequently adding a 2-residue linker followed by a 6x His-tag (100). Significant *in vitro* insolubility is not a universal feature of TCs. 2-MIBS, a single-domain bacterial TC that produces an earthy-smelling compound akin to geosmin, has no significant *in vitro* solubility issues and is generally purified from the soluble fraction after cell lysis, regardless if it is expressed alone or alongside encapsulin shell genes (16, 101). Interestingly, fusicoccadiene synthase, another bifunctional TC with a similar domain structure to that of GeoA, was reported to be soluble in solution in its full-length form and is predicted to form a hexameric complex (102). Ultimately, it is unknown why geosmin synthase is significantly more challenging to solubilize and subsequently purify. The multifunctional nature of GeoA or the long, flexible inter-domain linker may contribute to misfolding, hampering its ability to be heterologously expressed and purified.

1.2.3 Geosmin synthase as a putative encapsulin cargo

Recent bioinformatic studies have identified 118 putative geosmin synthase genes in Family 2B encapsulin operons (19), though no geosmin synthase has been demonstrated to be encapsulated within an encapsulin shell prior to this work. Family 2B encapsulin shell genes were initially misannotated as Crp/Fnr transcription factors due to the presence of the cyclic nucleotide-binding insertion domains (17) and continuously mislabeled as such in subsequent

analyses of geosmin synthase operons (103). However, like many other encapsulin cargoes, not all geosmin synthase genes are in encapsulin operons. Bioinformatic analyses in 2023 identified 1,062 homologs of geosmin synthase (90), greatly outnumbering the total predicted geosmin synthase encapsulin operons and implying that less than 10% of GeoA enzymes from sequenced genomes are encapsulin-associated. Most GeoA encapsulin operons are found in Proteobacteria and Cyanobacteria (19), while most geosmin synthases in Actinobacterial species are not encapsulin-associated. One phylogenetic analysis of geosmin synthases from 2012 observed two distinct clusters, with one cluster containing only actinobacterial enzymes and the other containing enzymes from Cyanobacteria and Myxobacteria (76). Although unknown to researchers at the time, the identified phylogenetic clusters are strongly correlated with the presence or absence of neighboring encapsulin genes and suggest that encapsulated geosmin synthases may have conserved characteristics distinct from those that are unencapsulated.

A sequence alignment of encapsulin-associated geosmin synthases has revealed two highly conserved motifs rich in glycine, proline, and alanine that are thought to be targeting peptides (19). Both consensus sequences are found within regions of predicted disorder, one at the C-terminus and the other in the long, flexible linker connecting the two TC domains, and are highly similar to each other. Furthermore, the consensus sequences of the putative TPs in geosmin synthases closely resemble the consensus sequence identified in the targeting domain of 2-MIBS cargoes. While this peptide was not found to be the sole determinant of cargo loading in the *S. griseus* 2-MIBS encapsulin, its deletion resulted in fewer cargoes loaded into the shell (16). Similar hydrophobic regions are found in the disordered N-terminal domains of the two structurally characterized cysteine desulfurase encapsulins, as well as the short targeting peptides

found in Family 1 encapsulins (10, 14, 20). The presence of these conserved motifs in geosmin synthases supports the hypothesis that they are a widespread encapsulin cargo.

The benefit of encapsulating geosmin synthase within a protein shell is not immediately clear. Previous studies have shown that organisms whose genomes contain geosmin synthases can produce geosmin regardless of whether they have neighboring encapsulin genes, demonstrating that encapsulation is not a strict requirement for geosmin production. Though many encapsulins have been demonstrated to serve as storage organelles for the products of their enzymatic cargoes, this hypothesis seems unlikely for geosmin based on its proposed function as a secreted signaling molecule. One advantage of encapsulation lies in the co-localization of GeoA with FPP, its initial substrate, and germacradienol, the enzymatic intermediate and CTD's substrate, in a confined space. The two catalytic domains of geosmin synthase each catalyze a distinct reaction step, and current evidence suggests there is not a substrate channel connecting the two active sites for the direct transfer of the NTD reaction product (germacradienol) to the CTD (89, 90). In the absence of a substrate channel, the encapsulin shell may serve to minimize the escape of germacradienol before it can be converted to geosmin, thereby increasing the overall efficiency of the reaction. Separately, encapsulation could act as a means to stabilize the protein structure by minimizing aggregation, given that GeoA has shown a tendency to aggregate when heterologously expressed and purified. A third possibility lies in the ability of the shell to confer additional regulation to GeoA via its CBDs, considering that GeoA is only found in Family 2B operons. It is also possible that encapsulation serves multiple or all of these functions simultaneously. Ultimately, upon confirmation that GeoA is an encapsulin cargo, further investigation is needed to understand the biochemical and physiological benefits of encapsulating geosmin synthase.

1.3 Thesis overview and scope

A comprehensive search through all available microbial proteomes resulted in the recent discovery of 6,133 putative encapsulin operons spread across 35 prokaryotic phyla (19). This phylogenetic analysis demonstrated the existence of two novel encapsulin families, a number of new putative encapsulin cargoes, and "two-component" encapsulin operons that contain two shell genes. All previously characterized encapsulins have been shown to require only a single shell protein, and furthermore, no encapsulin has demonstrated the ability to incorporate more than one protein component in its shell. Even among the broad class of viruses that share the same protein fold as encapsulin shell proteins, only one virus structure has been found to require two separate shell components that fulfill different structural roles (104-106). The overall aim of this thesis is to determine the encapsulin shell structures and the underlying principles guiding shell assembly for two-component encapsulins. I have accomplished this by completing the biochemical and structural characterization of two separate two-component encapsulin systems that utilize different operon organizations and are found in separate bacterial phyla.

In Chapter 2, I report the structural characterization of a two-component encapsulin shell from *Streptomyces lydicus* in a xylulose kinase-type operon. I demonstrate the differential assembly behavior of the two shell components as well as their ability to preferentially form mixed shells. The structures of both shell proteins were determined with cryo-EM, demonstrating that the encapsulin proteins contain the HK97-fold common to all encapsulins in addition to a CBD. Subsequent 3D-classification analysis of cryo-EM data and crosslinking studies show that the two-component mixed encapsulin shell uses an irregular tiling pattern with a variable shell composition.

In Chapter 3, I showcase the structural and biochemical characterization of a two-component encapsulin from *Myxococcus xanthus,* which encapsulates a terpene cyclase. Here, I again demonstrate the differential assembly behavior of the two shell components and further confirm their ability to form mixed encapsulin shells. Both mixed shells and single-protomer shells were demonstrated to facilitate the efficient cargo loading of a geosmin synthase cargo enzyme. Cryo-EM analysis was performed on cargo-loaded mixed shells and cargo-loaded single-protomer shells, which enabled the determination of both shell protomer structures and a resolved targeting peptide-shell interaction. Furthermore, two key targeting peptide regions on geosmin synthase were identified to be important for cargo loading. Finally, we established that the geosmin synthase cargo can synthesize geosmin within a protein shell and demonstrated that the enzymatic activity of encapsulated geosmin synthase is comparable between encapsulins with different shell protein compositions.

Altogether, this work serves as the first demonstration of an encapsulin to incorporate two shell components and also represents the first known HK97-fold assembly to preferentially utilize two components with a random tiling mechanism.

1.4 References

1. Y. Diekmann, J. B. Pereira-Leal, Evolution of intracellular compartmentalization. *Biochemical journal*. **449,** 319-331 (2013).
2. D. Murat, M. Byrne, A. Komeili, Cell biology of prokaryotic organelles. *Cold Spring Harbor perspectives in biology*. **2,** a000422 (2010).
3. E. Cornejo, N. Abreu, A. Komeili, Compartmentalization and organelle formation in bacteria. *Current opinion in cell biology*. **26,** 132-138 (2014).
4. T. W. Giessen, P. A. Silver, Encapsulation as a strategy for the design of biological compartmentalization. *J Mol Biol*. **428,** 916-927 (2016).

5. J. A. Jones, T. W. Giessen, Advances in encapsulin nanocompartment biology and engineering. *Biotechnol Bioeng*. **118,** 491-505 (2021).

6. R. J. Nichols, C. Cassidy-Amstutz, T. Chaijarasphong, D. F. Savage, Encapsulins: molecular biology of the shell. *Crit Rev Biochem Mol Biol*. **52,** 583-594 (2017).

7. F. Akita, K. T. Chong, H. Tanaka, E. Yamashita, N. Miyazaki, Y. Nakaishi, M. Suzuki, K. Namba, Y. Ono, T. Tsukihara, A. Nakagawa, The crystal structure of a virus-like particle from the hyperthermophilic archaeon Pyrococcus furiosus provides insight into the evolution of viruses. *J Mol Biol*. **368,** 1469-1483 (2007).

8. M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology 2008 15:9*. **15,** (2008-08-31).

9. T. W. Giessen, B. J. Orlando, A. A. Verdegaal, M. G. Chambers, J. Gardener, D. C. Bell, G. Birrane, M. Liao, P. A. Silver, Large protein organelles form a new iron sequestration system with high storage capacity. *Elife*. **8,** (2019).

10. J. A. Jones, R. Benisch, T. W. Giessen, Encapsulin cargo loading: progress and potential. *J Mater Chem B*. **11,** 4377-4388 (2023).

11. W. J. Altenburg, N. Rollins, P. A. Silver, T. W. Giessen, Exploring targeting peptide-shell interactions in encapsulin nanocompartments. *Sci Rep*. **11,** 4951 (2021).

12. J. A. Jones, M. P. Andreas, T. W. Giessen, Structural basis for peroxidase encapsulation inside the encapsulin from the Gram-negative pathogen Klebsiella pneumoniae. *Nat Commun*. **15,** 2558 (2024).

13. S. Kwon, M. P. Andreas, T. W. Giessen, Structure and heterogeneity of a highly cargo-loaded encapsulin shell. *J Struct Biol*. **215,** 108022 (2023).

14. R. Benisch, M. P. Andreas, T. W. Giessen, A widespread bacterial protein compartment sequesters and stores elemental sulfur. *Science Advances* (2024-02-02).

15. K. A. Lien, K. Dinshaw, R. J. Nichols, C. Cassidy-Amstutz, M. Knight, R. Singh, L. D. Eltis, D. F. Savage, S. A. Stanley, A nanocompartment system contributes to defense against oxidative stress in. *Elife*. **10,** (2021).

16. M. P. Andreas, T. W. Giessen, The biosynthesis of the odorant 2-methylisoborneol is compartmentalized inside a protein shell. *bioRxiv* 2024.2004. 2023.590730 (2024).

17. T. W. Giessen, Encapsulins. *Annu Rev Biochem*. **91,** 353-380 (2022).

18. T. W. Giessen, P. A. Silver, Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nature microbiology*. **2,** 1-11 (2017).

19. M. P. Andreas, T. W. Giessen, Large-scale computational discovery and analysis of virus-derived microbial nanocompartments. *Nat Commun*. **12,** 4748 (2021).

20. R. J. Nichols, B. Lafrance, N. R. Phillips, D. R. Radford, L. M. Oltrogge, L. E. Valentin-Alvarado, A. J. Bischoff, E. Nogales, D. F. Savage, Discovery and characterization of a novel family of prokaryotic nanocompartments involved in sulfur metabolism. *Elife*. **10,** (2021).

21. S. J. Kim, M. Shoda, Purification and characterization of a novel peroxidase from Geotrichum candidum Dec 1 involved in decolorization of dyes. *Applied and Environmental Microbiology*. **65,** 1029-1035 (1999).

22. R. Benisch, T. W. Giessen, Structural and biochemical characterization of an encapsulin-associated rhodanese from Acinetobacter baumannii. *bioRxiv* (2024).

23. M. Gomelsky, M. Y. Galperin, Bacterial second messengers, cGMP and c-di-GMP, in a quest for regulatory dominance. *The EMBO journal*. **32,** 2421-2423 (2013).

24. H. M. Berman, L. F. Ten Eyck, D. S. Goodsell, N. M. Haste, A. Kornev, S. S. Taylor, The cAMP binding domain: an ancient signaling module. *Proceedings of the National Academy of Sciences*. **102,** 45-50 (2005).

25. N. Kannan, J. Wu, G. S. Anand, S. Yooseph, A. F. Neuwald, J. C. Venter, S. S. Taylor, Evolution of allostery in the cyclic nucleotide binding module. *Genome biology*. **8,** 1-13 (2007).

26. M. A. Matilla, F. Velando, D. Martín-Mora, E. Monteagudo-Cascales, T. Krell, A catalogue of signal molecules that interact with sensor kinases, chemoreceptors and transcriptional regulators. *FEMS Microbiology Reviews*. **46,** fuab043 (2022).

27. W. R. Wikoff, L. Liljas, R. L. Duda, H. Tsuruta, R. W. Hendrix, J. E. Johnson, Topologically linked protein rings in the bacteriophage HK97 capsid. *Science*. **289,** 2129-2133 (2000).

28. M. K. Pietilä, P. Laurinmäki, D. A. Russell, C.-C. Ko, D. Jacobs-Sera, R. W. Hendrix, D. H. Bamford, S. J. Butcher, M. K. Pietilä, P. Laurinmäki, D. A. Russell, C.-C. Ko, D. Jacobs-Sera, R. W. Hendrix, D. H. Bamford, S. J. Butcher, Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. *Proceedings of the National Academy of Sciences*. **110,** (2013-6-25).

29. M. Krupovic, E. V. Koonin, Multiple origins of viral capsid proteins from cellular ancestors. *Proceedings of the National Academy of Sciences*. **114,** E2401-E2410 (2017).

30. M. L. Baker, W. Jiang, F. J. Rixon, W. Chiu, Common ancestry of herpesviruses and tailed DNA bacteriophages. *J Virol*. **79,** 14967-14970 (2005).

31. J. E. Johnson, J. A. Speir, Quasi-equivalent viruses: a paradigm for protein assemblies. *J Mol Biol*. **269,** 665-675 (1997).

32. D. L. Caspar, A. Klug, Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol*. **27,** 1-24 (1962).

33. M. M. Suhanovsky, C. M. Teschke, Nature's favorite building block: Deciphering folding and capsid assembly of proteins with the HK97-fold. *Virology*. **479-480,** 487-497 (2015).

34. R. L. Duda, C. M. Teschke, The amazing HK97 fold: versatile results of modest differences. *Current opinion in virology*. **36,** 9-16 (2019).

35. N. Popovych, S.-R. Tzeng, M. Tonelli, R. H. Ebright, C. G. Kalodimos, Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proceedings of the National Academy of Sciences*. **106,** 6927-6932 (2009).

36. E. Eren, B. Wang, D. C. Winkler, N. R. Watts, A. C. Steven, P. T. Wingfield, Structural characterization of the Myxococcus xanthus encapsulin and ferritin-like cargo system gives insight into its iron storage mechanism. *Structure*. **30,** 551-563.e554 (2022).

37. J. Ross, Z. Mciver, T. Lambert, C. Piergentili, J. E. Bird, K. J. Gallagher, F. L. Cruickshank, P. James, E. Zarazúa-Arvizu, L. E. Horsfall, K. J. Waldron, M. D. Wilson, C. L. Mackay, A. Baslé, D. J. Clarke, J. Marles-Wright, Pore dynamics and asymmetric cargo loading in an encapsulin nanocompartment. *Sci Adv*. **8,** eabj4461 (2022).

38. J. A. Jones, M. P. Andreas, T. W. Giessen, Exploring the extreme acid tolerance of a dynamic protein nanocage. *Biomacromolecules*. **24,** 1388-1399 (2023).

39. C. A. Mchugh, J. Fontana, D. Nemecek, N. Cheng, A. A. Aksyuk, J. B. Heymann, D. C. Winkler, A. S. Lam, J. S. Wall, A. C. Steven, E. Hoiczyk, A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *EMBO J*. **33,** 1896-1911 (2014).

40. B. J. Lafrance, C. Cassidy-Amstutz, R. J. Nichols, L. M. Oltrogge, E. Nogales, D. F. Savage, The encapsulin from Thermotoga maritima is a flavoprotein with a symmetry matched ferritin-like cargo protein. *Scientific reports*. **11,** 22810 (2021).

41. D. Kim, J. Choi, S. Lee, H. Hyun, K. Lee, K. Cho, Mutants defective in the production of encapsulin show a tan-phase-locked phenotype in Myxococcus xanthus. *journal of microbiology*. **57,** 795-802 (2019).

42. M. Ahmad, J. N. Roberts, E. M. Hardiman, R. Singh, L. D. Eltis, T. D. Bugg, Identification of DypB from Rhodococcus jostii RHA1 as a lignin peroxidase. *Biochemistry*. **50,** 5096-5107 (2011).

43. R. Rahmanpour, T. D. Bugg, Assembly in vitro of Rhodococcus jostii RHA1 encapsulin and peroxidase DypB to form a nanocompartment. *FEBS J*. **280,** 2097-2104 (2013).

44. R. M. Putri, C. Allende-Ballestero, D. Luque, R. Klem, K.-A. Rousou, A. Liu, C. H.-H. Traulsen, W. F. Rurup, M. S. Koay, J. R. CastóN, Structural characterization of native

and modified encapsulins as nanoplatforms for in vitro catalysis and cellular uptake. *ACS nano*. **11,** 12796-12804 (2017).

45. Y. Tang, A. Mu, Y. Zhang, S. Zhou, W. Wang, Y. Lai, X. Zhou, F. Liu, X. Yang, H. Gong, Cryo-EM structure of Mycobacterium smegmatis DyP-loaded encapsulin. *Proceedings of the National Academy of Sciences*. **118,** e2025658118 (2021).

46. R. Hidese, H. Mihara, N. Esaki, Bacterial cysteine desulfurases: versatile key players in biosynthetic pathways of sulfur-containing biofactors. *Applied microbiology and biotechnology*. **91,** 47-61 (2011).

47. M. Das, A. Dewan, S. Shee, A. Singh, The multifaceted bacterial cysteine desulfurases: From metabolism to pathogenesis. *Antioxidants*. **10,** 997 (2021).

48. M. P. Andreas, T. W. Giessen, Heterologous expression and purification of encapsulins in *Streptomyces coelicolor*. *MethodsX*. **9,** 101787 (2022).

49. J. S. Dickschat, T. Nawrath, V. Thiel, B. Kunze, R. Müller, S. Schulz, Biosynthesis of the off-flavor 2-methylisoborneol by the myxobacterium Nannocystis exedens. *Angewandte Chemie International Edition*. **46,** 8287-8290 (2007).

50. C.-M. Wang, D. E. Cane, Biochemistry and molecular genetics of the biosynthesis of the earthy odorant methylisoborneol in Streptomyces coelicolor. *Journal of the American Chemical Society*. **130,** 8908-8909 (2008).

51. S. Giglio, W. Chou, H. Ikeda, D. Cane, P. Monis, Biosynthesis of 2-methylisoborneol in cyanobacteria. *Environmental science & technology*. **45,** 992-998 (2011).

52. Z. Wang, Y. Xu, J. Shao, J. Wang, R. Li, Genes associated with 2-methylisoborneol biosynthesis in cyanobacteria: isolation, characterization, and expression in response to light. *PLoS one*. **6,** e18665 (2011).

53. H. Moon, J. Lee, J. Min, S. Kang, Developing genetically engineered encapsulin protein cage nanoparticles as a targeted delivery nanoplatform. *Biomacromolecules*. **15,** 3794-3801 (2014).

54. P. Lagoutte, C. Mignon, G. Stadthagen, S. Potisopon, S. Donnat, J. Mast, A. Lugari, B. Werle, Simultaneous surface display and cargo loading of encapsulin nanocompartments and their use for rational vaccine design. *Vaccine*. **36,** 3622-3628 (2018).

55. M. C. Jenkins, S. Lutz, Encapsulin nanocontainers as versatile scaffolds for the development of artificial metabolons. *ACS Synthetic Biology*. **10,** 857-869 (2021).

56. D. W. Christianson, Structural and chemical biology of terpenoid cyclases. *Chemical reviews*. **117,** 11570-11648 (2017).

57. J. D. Rudolf, C.-Y. Chang, Terpene synthases in disguise: enzymology, structure, and opportunities of non-canonical terpene synthases. *Natural product reports*. **37,** 425-463 (2020).

58. D. Cox-Georgian, N. Ramadoss, C. Dona, C. Basu, Therapeutic and medicinal uses of terpenes. *Medicinal plants: from farm to pharmacy* 333-359 (2019).

59. S. Serra, Recent developments in the synthesis of the flavors and fragrances of terpenoid origin. *Studies in natural products chemistry*. **46,** 201-226 (2015).

60. J. D. Rudolf, T. A. Alsup, B. Xu, Z. Li, Bacterial terpenome. *Natural product reports*. **38,** 905-980 (2021).

61. J. S. Dickschat, Bacterial terpene cyclases. *Natural product reports*. **33,** 87-110 (2016).

62. N. Gerber, H. Lechevalier, Geosmin, an earthy-smelling substance isolated from actinomycetes. *Applied microbiology*. **13,** 935-938 (1965).

63. E. A. Barka, P. Vatsa, L. Sanchez, N. Gaveau-Vaillant, C. Jacquard, H.-P. Klenk, C. Clément, Y. Ouhdouch, G. P. Van Wezel, Taxonomy, physiology, and natural products of Actinobacteria. *Microbiology and molecular biology reviews*. **80,** 1-43 (2016).

64. D. A. Hopwood, Streptomyces in nature and medicine: the antibiotic makers. 2007: Oxford University Press.

65. R. G. Buttery, J. A. Garibaldi, Geosmin and methylisoborneol in garden soil. *Journal of agricultural and food chemistry*. **24,** 1246-1247 (1976).

66. C. S. Tucker, Off-flavor problems in aquaculture. *Reviews in Fisheries Science*. **8,** 45-88 (2000).

67. T. P. Heil, R. C. Lindsay, Volatile compounds in flavor-tainted fish from the upper Wisconsin River. *Journal of Environmental Science & Health Part B*. **23,** 489-512 (1988).

68. R. T. Lovell, I. Y. Lelana, C. E. Boyd, M. S. Armstrong, Geosmin and musty-muddy flavors in pond-raised channel catfish. *Transactions of the American Fisheries Society*. **115,** 485-489 (1986).

69. P. Darriet, M. Pons, S. Lamy, D. Dubourdieu, Identification and quantification of geosmin, an earthy odorant contaminating wines. *Journal of Agricultural and Food Chemistry*. **48,** 4835-4838 (2000).

70. S. Mustapha, J. Tijani, M. Ndamitso, A. Abdulkareem, D. Shuaib, A. Mohammed, A critical review on geosmin and 2-methylisoborneol in water: sources, effects, detection, and removal techniques. *Environmental Monitoring and Assessment*. **193,** 1-34 (2021).

71. N. N. Gerber, Volatile substances from actinomycetes: their role in the odor pollution of water. *CRC critical reviews in microbiology*. **7,** 191-214 (1979).

72. N. N. Gerber, Geosmin, an earthy-smelling substance isolated from actinomycetes. *Biotechnology and Bioengineering*. **9,** 321-327 (1967).

73. C. E. Schöller, H. Gürtler, R. Pedersen, S. Molin, K. Wilkins, Volatile metabolites from actinomycetes. *Journal of agricultural and food chemistry*. **50,** 2615-2621 (2002).

74. C. P. Dionigi, D. F. Millie, A. M. Spanier, P. B. Johnsen, Spore and geosmin production by Streptomyces tendae on several media. *Journal of agricultural and food chemistry*. **40,** 122-125 (1992).

75. J. S. Dickschat, T. Martens, T. Brinkhoff, M. Simon, S. Schulz, Volatiles released by a Streptomyces species isolated from the North Sea. *Chemistry & biodiversity*. **2,** 837-865 (2005).

76. C. A. Citron, J. Gleitzmann, G. Laurenzano, R. Pukall, J. S. Dickschat, Terpenoids are widespread in actinomycetes: a correlation of secondary metabolism and genome data. *ChemBioChem*. **13,** 202-214 (2012).

77. P. Rabe, C. A. Citron, J. S. Dickschat, Volatile terpenes from actinomycetes: a biosynthetic study correlating chemical analyses to genome data. *ChemBioChem*. **14,** 2345-2354 (2013).

78. C. A. Citron, L. Barra, J. Wink, J. S. Dickschat, Volatiles from nineteen recently genome sequenced actinomycetes. *Organic & biomolecular chemistry*. **13,** 2673-2683 (2015).

79. A. Ahearne, H. Albataineh, S. E. Dowd, D. C. Stevens, Assessment of evolutionary relationships for prioritization of myxobacteria for natural product discovery. *Microorganisms*. **9,** 1376 (2021).

80. W. Trowitzsch, L. Witte, H. Reichenbach, Geosmin from earthy smelling cultures of Nannocystis exedens (Myxobacterales). *FEMS Microbiology Letters*. **12,** 257-260 (1981).

81. S. Schulz, J. Fuhlendorff, H. Reichenbach, Identification and synthesis of volatiles released by the myxobacterium Chondromyces crocatus. *Tetrahedron*. **60,** 3863-3872 (2004).

82. J. S. Dickschat, S. C. Wenzel, H. B. Bode, R. Müller, S. Schulz, Biosynthesis of volatiles by the myxobacterium Myxococcus xanthus. *ChemBioChem*. **5,** 778-787 (2004).

83. G. Izaguirre, C. J. Hwang, S. W. Krasner, M. J. Mcguire, Geosmin and 2-methylisoborneol from cyanobacteria in three water supply systems. *Applied and environmental microbiology*. **43,** 708-714 (1982).

84. C. Höckelmann, P. G. Becher, S. H. V. Reuß, F. Jüttner, Sesquiterpenes of the geosmin-producing cyanobacterium Calothrix PCC 7507 and their toxicity to invertebrates. *Zeitschrift für Naturforschung C*. **64,** 49-55 (2009).

85. S. Suurnäkki, G. V. Gomez-Saez, A. Rantala-Ylinen, J. Jokela, D. P. Fewer, K. Sivonen, Identification of geosmin and 2-methylisoborneol in cyanobacteria and molecular detection methods for the producers of these compounds. *Water research*. **68,** 56-66 (2015).

86. F. JüTtner, S. B. Watson, Biochemical and ecological control of geosmin and 2-methylisoborneol in source waters. *Applied and environmental microbiology*. **73,** 4395-4406 (2007).

87. D. E. Cane, R. M. Watt, Expression and mechanistic analysis of a germacradienol synthase from Streptomyces coelicolor implicated in geosmin biosynthesis. *Proceedings of the National Academy of Sciences*. **100,** 1547-1551 (2003).

88. B. Gust, G. L. Challis, K. Fowler, T. Kieser, K. F. Chater, PCR-targeted Streptomyces gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proceedings of the National Academy of Sciences*. **100,** 1541-1546 (2003).

89. J. Jiang, X. He, D. E. Cane, Biosynthesis of the earthy odorant geosmin by a bifunctional Streptomyces coelicolor enzyme. *Nature Chemical Biology*. **3,** 711-715 (2007).

90. H. Xu, J. S. Dickschat, A Detailed View on Geosmin Biosynthesis. *ChemBioChem*. **24,** e202300101 (2023).

91. P. G. Becher, V. Verschut, M. J. Bibb, M. J. Bush, B. P. Molnár, E. Barane, M. M. Al-Bassam, G. Chandra, L. Song, G. L. Challis, Developmentally regulated volatiles geosmin and 2-methylisoborneol attract a soil arthropod to Streptomyces bacteria promoting spore dispersal. *Nature Microbiology*. **5,** 821-829 (2020).

92. M. C. Stensmyr, H. K. Dweck, A. Farhan, I. Ibba, A. Strutz, L. Mukunda, J. Linz, V. Grabe, K. Steck, S. Lavista-Llanos, A conserved dedicated olfactory circuit for detecting harmful microbes in Drosophila. *Cell*. **151,** 1345-1357 (2012).

93. A. Tran, A. Tang, C. T. O'loughlin, A. Balistreri, E. Chang, D. Coto Villa, J. Li, A. Varshney, V. Jimenez, J. Pyle, C. elegans avoids toxin-producing Streptomyces using a seven transmembrane domain chemosensory receptor. *Elife*. **6,** e23770 (2017).

94. L. Zaroubi, I. Ozugergin, K. Mastronardi, A. Imfeld, C. Law, Y. Gélinas, A. Piekny, B. L. Findlay, The ubiquitous soil terpene geosmin acts as a warning chemical. *Applied and Environmental Microbiology*. **88,** e00093-00022 (2022).

95. J. S. Metcalf, G. A. Codd, Cyanotoxins, in *Ecology of cyanobacteria II: their diversity in space and time*. 2012, Springer. p. 651-675.

96. X. He, D. E. Cane, Mechanism and Stereochemistry of the Germacradienol/Germacrene D Synthase of Streptomyces c oelicolor A3 (2). *Journal of the American Chemical Society*. **126,** 2678-2679 (2004).

97. J. Jiang, X. He, D. E. Cane, Geosmin biosynthesis. Streptomyces coelicolor germacradienol/germacrene D synthase converts farnesyl diphosphate to geosmin. *Journal of the American Chemical Society*. **128,** 8128-8129 (2006).

98. J. Jiang, C. P. Saint, D. E. Cane, P. T. Monis, Isolation and characterization of the gene associated with geosmin production in cyanobacteria. *Environmental science & technology*. **42,** 8027-8032 (2008).

99. D. E. Cane, X. He, S. Kobayashi, S. Ōmura, H. Ikeda, Geosmin biosynthesis in Streptomyces avermitilis. Molecular cloning, expression, and mechanistic study of the germacradienol/geosmin synthase. *The Journal of antibiotics*. **59,** 471-479 (2006).

100. G. G. Harris, P. M. Lombardi, T. A. Pemberton, T. Matsui, T. M. Weiss, K. E. Cole, M. KöKsal, F. V. Murphy Iv, L. S. Vedula, W. K. Chou, Structural Studies of Geosmin Synthase, a Bifunctional Sesquiterpene Synthase with αα Domain Architecture That Catalyzes a Unique Cyclization–Fragmentation Reaction Sequence. *Biochemistry*. **54,** 7142-7155 (2015).

101. M. KöKsal, W. K. Chou, D. E. Cane, D. W. Christianson, Structure of 2-methylisoborneol synthase from Streptomyces coelicolor and implications for the cyclization of a noncanonical C-methylated monoterpenoid substrate. *Biochemistry*. **51,** 3011-3020 (2012).

102. M. Chen, W. K. Chou, T. Toyomasu, D. E. Cane, D. W. Christianson, Structure and function of fusicoccadiene synthase, a hexameric bifunctional diterpene synthase. *ACS chemical biology*. **11,** 889-899 (2016).

103. C. Churro, A. P. Semedo-Aguiar, A. D. Silva, J. B. Pereira-Leal, R. B. Leite, A novel cyanobacterial geosmin producer, revising Geo A distribution and dispersion patterns in Bacteria. *Scientific Reports*. **10,** 8679 (2020).

104. L. Onorato, B. Stirmer, M. K. Showe, Isolation and characterization of bacteriophage T4 mutant preheads. *J Virol*. **27,** 409-426 (1978).

105. Q. Fang, W. C. Tang, A. Fokine, M. Mahalingam, Q. Shao, M. G. Rossmann, V. B. Rao, Structures of a large prolate virus capsid in unexpanded and expanded states generate insights into the icosahedral virus assembly. *Proc Natl Acad Sci U.S.A.* **119,** e2203272119 (2022).

106. Z. Chen, L. Sun, Z. Zhang, A. Fokine, V. Padilla-Sanchez, D. Hanein, W. Jiang, M. G. Rossmann, V. B. Rao, Cryo-EM structure of the bacteriophage T4 isometric head at 3.3-Å resolution and its relevance to the assembly of icosahedral viruses. *Proc Natl Acad Sci U S A*. **114,** E8184-E8193 (2017).

Chapter 2 A Two-Component Quasi-Icosahedral Protein Nanocompartment with Variable Shell Composition and Irregular Tiling

The work presented in this chapter was adapted from Dutcher, C. A., Andreas, M. P., & Giessen, T. W. (2024). A two-component quasi-icosahedral protein nanocompartment with variable shell composition and irregular tiling. bioRxiv. (1)

Author contributions: C.A.D. designed and constructed all plasmids used in this study and subsequently expressed and purified all protein samples. C.A.D. analyzed purified protein samples via negative stain TEM and DLS, performed protein crosslinking assays, and conducted thermal ramp stability experiments. M.P.A. prepared cryo-EM grids, collected and processed cryo-EM datasets, and built the protomer models for Sl-Enc1 and Sl-Enc2. T.W.G. aided with cryo-EM data processing and completed the 3D classification analysis of protomer interactions at the symmetry axes for mixed encapsulin shells.

2.1 Introduction

The HK97-fold is principally found in the major capsid proteins of bacteriophages of the ubiquitous viral order Caudovirales (2). It is named after the fold observed in the major capsid protein of Escherichia virus HK97, a double-stranded DNA bacteriophage natively infecting *Escherichia coli* (3). The HK97-fold is also found as a key structural component in other icosahedral protein assemblies including tailed archaeal viruses (4), the floor domains of animal-infecting viruses of the order Herpesviridae (5), and the shell proteins of encapsulins – self-

assembling enzyme-loaded protein nanocompartments – found in many bacteria and archaea (6-8).

The HK97-fold is characterized by three highly conserved structural features, the axial domain (A-domain), peripheral domain (P-domain), and the extended loop (E-loop) (3, 8). The structural plasticity and versatility of the HK97-fold is highlighted by the fact that its three core structural features – A-domain, P-domain, and E-loop – can be maintained in proteins with little sequence similarity while a range of diverse insertion domains have been found embedded at various positions within the HK97-fold (9, 10). In addition, icosahedral HK97 assemblies have been shown to exhibit dynamic pores able to open or close depending on the particular stimuli (11, 12).

Protein shells (encapsulins) and capsids (viruses) that utilize the HK97-fold exhibit an impressively large size distribution, ranging from 18 nm (triangulation number (T) of T=1, 60 protomers; empty *Myxococcus xanthus* encapsulin) to 180 nm (T=52, 3120 protomers; bacteriophage G) in diameter (13, 14). To assemble icosahedra greater than T=1, both pentamers and hexamers are needed to tile a closed shell or capsid (15, 16). Nearly all characterized HK97-fold protein assemblies use a single major capsid or shell protein to form icosahedral structures, relying on the built-in conformational flexibility of the HK97-fold, especially the E-loop, to assemble both oligomeric states – pentamers and hexamers – needed for assembling closed shells or capsids. There is a single virus – bacteriophage T4 – that utilizes two distinct HK97-fold proteins for separate structural roles, forming either homo-pentamers or homo-hexamers in the context of its assembled capsid (17, 18).

Encapsulins are the only non-viral structures known to utilize the HK97-fold (6-8). All so far studied encapsulins are composed of a single type of protomer that self-assembles into

defined shells with sizes ranging from 18 to 42 nm in diameter and triangulation numbers of T=1, T=3, or T=4 (8, 19, 20). Many prokaryotes utilize encapsulin shells to compartmentalize specialized metabolic processes (21, 22). This is accomplished through the selective encapsulation of specific enzymes, targeted to the encapsulin lumen via a C- or N-terminal targeting peptide or domain present in all native cargo proteins (23-26). So far, a number of encapsulin systems with different cargo enzymes have been studied and found to be involved in varied metabolic functions. Cargo enzymes include ferroxidases involved in iron storage (8, 19, 20, 27-29), peroxidases involved in oxidative stress resistance (25, 30-32), cysteine desulfurases involved in sulfur storage (33, 34), and terpene cyclases involved in terpenoid biosynthesis (35, 36). Enzyme encapsulation inside encapsulin shells can convey a number of distinct advantages, such as the formation of an optimized reaction environment, enzyme co-localization, intermediate sequestration, and the exclusion of unwanted proteins and molecules (37). Furthermore, the encapsulin shell may also serve to regulate cargo enzymes by modulating the flux of small molecule metabolites through its pores (6, 22, 38), whilst also stabilizing the oligomeric state of enzymatic cargo (32). Encapsulins are found in at least 35 prokaryotic phyla and have been classified into four distinct families based on operon structure and shell protein phylogeny (7, 34). Family 2, the most abundant class of encapsulins, is further subdivided into Family 2A and Family 2B based on the absence or presence, respectively, of a putative cyclic nucleotide-binding domain (CBD) inserted into the E-loop of the HK97-fold encapsulin shell protein (7). Interestingly, only present within Family 2B are encapsulin operons encoding two discrete encapsulin shell proteins, with a diversity of associated cargo types (6, 7). There is no precedent for a multi-component encapsulin shell formed from more than one HK97-fold shell protein.

In this chapter, we report the characterization of a Family 2B two-component encapsulin shell identified in *Streptomyces lydicus* (Sl). Through heterologous expression and purification studies, we show the differential assembly behavior of the two shell components – Sl-Enc1 and Sl-Enc2 – and demonstrate their ability to co-assemble into mixed shells with variable shell composition. We determined the structures of both *S. lydicus* shell proteins using cryo-electron microscopy (cryo-EM). Using 3D classification analyses, I confirm co-assembly into mixed shells and highlight the irregular tiling and random incorporation of the two shell proteins into two-component shells. This work expands the known assembly modes of HK97-fold proteins and lays the foundation for future studies aimed at elucidating the impact of two-component shells on encapsulin function.

2.2 Materials and Methods

2.2.1 Bioinformatic and phylogenetic analysis

Family 2B encapsulin amino acid sequences were accessed from the UniProtKB database on March 30, 2023, by initiating a search for sequences containing the two Pfam families PF00027 and PF19307, resulting in 2,892 initial sequences. To identify genome neighborhoods for all initial hits, the Enzyme Function Initiative-Genome Neighborhood Tool (EFI-GNT) was used (39, 40). Two-component encapsulin systems were identified by filtering for Family 2B encapsulin genes with start codons within 10 kb of each other. Operons containing encapsulin sequences with less than 400 amino acids were categorized as fragments and removed from the dataset, resulting in a final dataset containing 1,006 UniProtKB accessions in 503 unique two-component Family 2B operons. Encapsulin sequences were then downloaded in FASTA format and aligned using MAFFT (mafft.crbc.jp/alignment) (41) with default settings. The alignment

was forwarded to the Phylogenetic Tree function on the MAFFT server to generate a tree with the following settings: method-NJ with conserved sites (180 AAs), substitution model-JTT, heterogeneity among sites-ignore (alpha = infinity), bootstrap-1000. The phylogenetic tree was then visualized and annotated using iTOL V6.7.3 (42).

2.2.2 Molecular biology and cloning

The *S. lydicus* encapsulin shell genes were codon-optimized for *E. coli* using the Integrated DNA Technologies (IDT) codon optimization tool and ordered as gBlock gene fragments from IDT (Supplemental Table 2-1). pETDuet-1 vector was linearized by digestion with PacI and NdeI and gene fragments were inserted into multiple cloning site 2 (MCS2) of the linearized plasmid via Gibson assembly using the New England BioLabs (NEB) Gibson Assembly Master Mix and following the manufacturer's standard protocol. All gene modifications, including N-terminal His- and Strep-affinity tags, as well as cysteine modifications, were introduced to the *S. lydicus* encapsulin genes via primer extension with primers from IDT using an inverse PCR strategy. Outward-pointing PCR primer pairs were designed to amplify the entire template with the desired sequence insertions or mutations (Supplemental Table 2-2 and 2-3). Linear plasmid fragments were phosphorylated with T4 polynucleotide kinase and ligated with T4 DNA ligase following the manufacturer's protocol. The double cysteine mutant plasmid was constructed using Gibson Assembly to ligate two PCR-amplified fragments with the desired mutations. Invitrogen One Shot BL21(DE3) chemically competent *E. coli* cells were subsequently transformed with the respective assembled plasmids and validated via Sanger sequencing (Eurofins) (Supplemental Table 2-4).

2.2.3 Protein expression and purification

Protein expression for all constructs was carried out using Terrific Broth (TB) media supplemented with 100 μg/mL ampicillin. Cell cultures were inoculated from overnight starter cultures at a ratio of 1:100 and grown at 37˚C, 200 rpm until reaching an OD_{600} of 0.4-0.8. Then, protein expression was induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and cells grown further at 18˚C for 20-24 h. Cells were harvested by centrifugation at 8,000 rcf for 15 min at 4°C and cell pellets were either used immediately or stored at -80˚C.

Cell pellets were thawed and resuspended in 3 mL of Tris Buffered Saline (TBS) (20 mM Tris, 300 mM NaCl, pH 8.0) per 1 g of wet cell weight. Lysozyme (0.5 mg/mL), DNase (1 μg/mL), Benzonase nuclease (25 units/mL), and SIGMAFAST EDTA-free protease inhibitor cocktail (1 tablet per 100 mL) were added to the resuspension buffer. Cell suspensions were incubated at 4˚C while stirring for one hour. Cells were lysed by three consecutive passages through an EmulsiFlex-C3 homogenizer at 20,000 psi and subsequently centrifuged at 8,000 rcf for 15 min at 4˚C. The soluble fraction was then applied to a 5 mL packed Ni-NTA resin bed (GoldBio) and equilibrated with 10 column volumes of TBS. Resin was washed with a stepwise gradient in increments of 30 mM imidazole up to 90 mM. TBS buffer containing 250 mM imidazole was used to elute the remaining protein bound to the Ni-NTA resin. Protein samples were analyzed for purity using SDS-PAGE with 14% Novex Tris-Glycine Gels (Thermo-Fisher Scientific). Fractions containing pure protein were pooled and subsequently concentrated and buffer exchanged using 50 kDa MWCO Amicon Ultra-15 mL centrifugal units with TBS. The concentrated protein sample was either stored at 4˚C or immediately purified further via size exclusion chromatography (SEC). For SEC purification, the concentrated protein sample was centrifuged for 10 min at 10,000 rcf and then loaded onto an ÄKTA Pure FPLC system and

purified using a Superose 6 10/300 GL column pre-equilibrated with TBS. For cysteine mutants, proteins were purified in the same way, but the buffer contained 2-5 mM tris(2-carboxyethyl)phosphine (TCEP) throughout all purification steps.

2.2.4 Negative stain transmission electron microscopy

Purified encapsulin samples were imaged via negative stain transmission electron microscopy (TEM) with a 100 kV Morgagni transmission electron microscope at the University of Michigan Life Sciences Institute. 200 Mesh Gold Grids coated with formvar-carbon film (FCF- 200-Au, EMS) were glow-discharged using a PELCO easiGlow at 5 mA for 60 s to increase their hydrophilicity. To the glow-discharged grids, 3.5 μL of a 0.1 mg/mL protein sample in 20 mM Tris, 150 mM NaCl was applied and allowed to adsorb for 1 minute before excess liquid was blotted away with Whatman filter paper. The grid was first washed with distilled water and then with 0.75% (w/v) uranyl formate before being floated on an 11 μL drop of 0.75% (w/v) uranyl formate for one min. After removing excess stain via blotting, samples were immediately imaged or stored at room temperature for later use.

2.2.5 Native PAGE

Native PAGE samples were prepared by mixing purified protein samples with Invitrogen NativePAGE sample buffer (Thermo Fisher Scientific). Protein samples and a molecular weight marker were loaded onto Invotrogen NativePAGE 3 to 12% Bis-Tris gels (Thermo Fisher Scientific) in 1x Invitrogen NativePAGE running buffer. Samples were run at 150 V for 2 hours at 4˚C, and gels were stained using Sigma-Aldrich Ready Blue protein gel stain.

2.2.6 Dynamic light scattering (DLS) and thermal ramp experiments

DLS and thermal stability measurements were carried out using an Uncle instrument (Unchained Labs) at 25˚C in triplicate. Purified protein samples were adjusted to a concentration of 0.02 to 0.2 mg/mL, centrifuged at 20,000 rcf for 10 min at 4˚C and then immediately analyzed via DLS. Melting temperature (Tm) and aggregation temperature (Tagg) analysis was conducted over a 25˚C to 95˚C thermal ramp at 1°C per min, measuring static light scattering (SLS) at 266 nm.

2.2.7 Cryo-electron microscopy (cryo-EM)

Sample preparation: Purified Sl-Enc1 or Sl-Enc2 majority samples were concentrated to 2 mg/mL in 150 mM NaCl, 25 mM Tris pH 8.0. 3.5 µL of protein sample were applied to freshly glow discharged Quantifoil R1.2/1.3 Cu 200 mesh grids (EMS, Q225-CR1.3) and were frozen by plunging into liquid ethane using an FEI Vitrobot Mark IV (100% humidity, 22°C, blot force 20, blot time 4 s, wait time 0 s). The grids were immediately clipped and stored in liquid nitrogen until data collection.

Data collection: Cryo-EM movies for Sl-Enc1-His-Sl-Enc2 (Sl-Enc1 majority) were collected using a ThermoFisher Scientific Talos Arctica cryo-electron microscope operating at 200 kV equipped with a Gatan K2 direct electron detector. 1,658 movies were collected from a single grid using the Leginon (43) software package at a magnification of 45,000x, pixel size of 0.91 Å, defocus range of -0.8 µm to -1.8 µm, exposure time of 5 s, frame time of 200 ms, and a total dose of 45.5 e-/Å^2. Cryo-EM movies of Sl-Enc1-His-Sl-Enc2-His (Sl-Enc2 majority) were

collected using a ThermoFisher Scientific Krios G4i cryo-electron microscope operating at 300 kV equipped with a Gatan K3 direct electron detector with a BioQuantum energy filter operating at 20 eV slit width. 2,865 movies were collected from a single grid using the Leginon (43) software package at a magnification of 105,000x, pixel size of 0.84 Å, defocus range of -0.5 μm to -1.0 μm, exposure time of 3 s, frame time of 50 ms, and total dose of 54 e-/Å^2.

Image processing: CryoSPARC v4.2.1 was initially used to process the Sl-Enc1 majority dataset (44). 1,658 movies were imported, motion corrected using patch motion correction, and the CTF-fit was estimated using patch CTF estimation. Exposures with CTF fits worse than 5 Å were discarded from the dataset, resulting in 1,541 remaining movies. 222 particles were manually picked to create templates for template-based particle picking. Template picker was used to identify particles, and 77,425 particles were extracted with a box size of 440 pixels. Two rounds of 2D classification yielded 72,804 particles. Ab-initio reconstruction with 3 classes and I symmetry was then carried out. 72,760 particles were contained in the major class and were then used for heterogeneous refinement with 3 classes and I symmetry imposed. This yielded a major class with 69,590 particles, which were then subjected to homogeneous refinement against the heterogeneous refinement map with I symmetry imposed, per-particle defocus optimization, per-group CTF parameterization, and Ewald sphere correction enabled, resulting in a 2.58 Å resolution consensus map. Local resolution was calculated in cryoSPARC.

For all subsequent analyses of the Sl-Enc1 sample, cryoSPARC v4.4.0+231114 was used. To investigate the compositional heterogeneity and allowed protomer arrangements within the Sl-Enc1 majority sample, the 69,590 particles used to create the consensus refinement were initially symmetry expanded (I symmetry), yielding 4,175,400 particles, of which 3,675,400 particles were – due to computational limitations – used for further downstream processing. Four

separated masked 3D classification runs were carried out using 3,675,400 symmetry expanded particles with the following non-standard parameters: 16 classes, target resolution: 3.5, O-EM learning rate init: 0.75, O-EM learning rate halflife: 100, force hard classification, class similarity: 0.25. Four custom 3D classification masks were created for the four unique protomer interactions possible within a T=1 icosahedral shell using ChimeraX v1.6.1 and the volume tools job (dilation radius: 1, soft padding: 3) in cryoSPARC. All masks included only the HK97-domains, and not the CBD domains, of the two protomers in question. After manual inspection, select 3D classes were refined via masked C1 local refinements using the same masks as for the 3D classification jobs. The following non-standard parameters were used: use pose/shift gaussian prior, standard deviation of prior over rotation: 7, standard deviation of prior over shifts: 4, maximum alignment resolution: 0.05, re-center rotations each iteration, re-center shifts each iteration, number of extra final passes: 2, initial lowpass resolution: 6, force re-do GS split. To map symmetry-expanded particles of select 3D classes onto micrographs, the following protocol was used: first, a subset of 10,000 particles – created via the particle sets tool – from each 3D class of interest was re-centered using the volume alignments job in cryoSPARC. As the new center coordinates, the centers of the respective 3D classification masks were used which were measured in ChimeraX. The resulting re-centered particles were then re-extracted with a box size of 32 pixels. As these particles were exclusively used for visualization purposes, a small box size was sufficient. The re-extracted re-centered particles were then mapped onto micrographs and visualized via the inspect picks job. To create images of symmetry-expanded particles from multiple classes mapped onto the same micrograph, individually mapped micrographs were overlayed to create composite images.

CryoSPARC v4.2.1 was used to process the Sl-Enc2 majority dataset. 2,865 movies were imported, motion corrected using patch motion correction, and the CTF-fit was estimated using patch CTF estimation. Exposures with CTF fits worse than 5 Å were discarded from the dataset, resulting in 2,176 remaining movies. After manual picking and template creation, template picker was used to pick 134,284 particles which were subsequently extracted with a box size of 480 pixels. Two rounds of 2D classification yielded 116,239 particles. These particles were then used for heterogeneous refinement with 3 classes and I symmetry applied. 64,557 particles were contained in the major class and were then used for homogeneous refinement with I symmetry, per-particle defocus, and per-group CTF parameterization enabled. This yielded a 2.59 Å resolution consensus map. Local resolution was calculated in cryoSPARC.

Model building: Cryo-EM maps were initially opened in UCSF Chimera v 1.15 (45) and for the Sl-Enc2 majority dataset, an initial model of the *Streptomyces griseus* Family 2B encapsulin (PDB ID: 9BHU) was manually placed into the density and the fit further optimized using the Dock to Volume command. For the Sl-Enc1 majority dataset, an initial model of Sl-Enc2 was placed using the same procedure in USCF Chimera v 1.15. The resulting models were opened in Coot v 0.8.9.1 (46, 47) and manually refined against the maps using rigid body refinement, chain refinement, and iterative manual real space refinement along the length of the chain while also mutating chains to the respective protein (Sl-Enc1 or Sl-Enc2). The protomer models were then refined in Phenix v 1.20.1-4487 (48-50) using phenix.real_space_refine with default parameters. NCS was identified from the maps using phenix.map_symmetry, and NCS operators were applied using phenix.apply_ncs to make icosahedral models containing 60 copies of the respective protomer. These models were then refined using phenix.real_space_refine with NCS constraints, global minimization, and ADP refinements. The models were validated using

the comprehensive validation tool via phenix.validation_cryoem. The models were further inspected and manually refined until satisfactory geometry and density fit statistics were reached

2.2.8 Protein crosslinking assays

Purified encapsulin samples containing cysteine point mutations at 0.1 mg/mL (1 mL) were buffer exchanged using a 100 kDa MWCO Amicon Ultra-15 mL centrifugal unit into 20 mM Tris, 150 mM NaCl, 5mM EDTA at pH 7.3. A fresh 5mM stock of 1,11-bis(maleimido)triethylene glycol crosslinker (BM(PEG)3, Thermo-Fisher Scientific) was then prepared by dissolving the reagent in dimethylformamide (DMF). The crosslinker solution was added to the protein sample at a five-fold molar excess and the reaction mixture incubated at room temperature for 1 h in the dark. After incubation, the reaction was quenched by adding dithiothreitol (DTT) to a 10 mM final concentration, followed by incubation for 15 min at room temperature. Crosslinked protein samples were then analyzed via SDS-PAGE with 10% Novex Tris-Glycine gels (Thermo-Fisher Scientific).

2.2.9 Protein identification via mass spectrometry

Purified Sl-Enc1/His-Sl-Enc2 protein samples were analyzed via SDS-PAGE with a 14% Novex Tris-Glycine Gels (Thermo-Fisher Scientific). The SDS-PAGE gel was stained with Coomassie and the overexpressed bands at the correct molecular weights were excised and stored in water. The SDS-PAGE gel bands were transferred to the University of Michigan Biomedical Research Core Facilities (BRCF) Proteomics & Peptide Synthesis Core, where the protein was extracted from the SDS-PAGE gel and subjected to tryptic digest at 37˚C for 4 h,

quenched, and analyzed. The digested samples were then analyzed using a nano LC/MS/MS Waters HPLC system coupled to a Thermo-Fisher Scientific LTQ Orbitrap Velos Pro following standard MS/MS protocol. Data were searched using a local copy of Mascot (Matrix Science, UK).

2.2.10 Mass spectrometric analysis of crosslinked proteins

Crosslinked protein bands were excised from SDS-PAGE gels and transferred to the University of Michigan Department of Chemistry Mass Spectrometry facility for processing and analysis. Gel bands were washed with water and 80% acetonitrile to elute protein from the gel matrix. Protein was treated with dithiothreitol (DTT) and 2-iodoacetamide (IAA) to reduce the disulfide bonds and alkylate the protein. The sample was subsequently digested with trypsin for 4 h at 45˚C. Samples were analyzed on a Thermo-Fisher Scientific Orbitrap Fusion Lumos Tribrid mass spectrometer. Peptides were separated with a C18 desalting column (Thermo Acclaim PepMap, 75 μm × 25 cm) and reconstituted in 0.1% formic acid. A 60 min gradient from 4% to 76% acetonitrile with 0.1% formic acid, flow rate = 300 nL/min was used. HCD-MSMS was applied in data-dependent mode for a 3 second cycle time for all ions above 5e4 with 2+ to 7+ charge; applied 34% HCD energy. Proteome Discover 2.2 software was used to process the data and identify crosslinked peptides. The *S. lydicus* peptide sequences with affinity tag and cysteine modifications were used as sequence inputs.

2.3 Results

2.3.1 Two-component encapsulin operons are diverse and widespread in bacteria

To investigate the diversity and distribution of two-component encapsulin systems, we carried out a computational search to initially identify all Family 2B encapsulin shell genes encoded within 10 kb of each other in a given genome. Our search yielded 503 two-component operons representing 21% of all Family 2B systems (Figure 2-1A). Five distinct operon types could be identified based on the identities of co-encoded putative cargo proteins (6, 7). The five operon types encode terpene cyclases (TCs), prenyltransferases (PTs), TCs and PTs together (TC/PT), xylulose kinases (XKs), and cysteine desulfurases (CDs), in decreasing order of prevalence. In addition, a number of two-component operons do not encode any obvious cargo (Other). The two encapsulin shell genes are generally clustered together with conserved gene synteny. We define the shell protein encoded by the first (upstream) gene as Enc1 and the shell protein encoded by the second (downstream) gene as Enc2. Most cargos are encoded directly up- or downstream of Enc1-Enc2, except for in TC/PT operons where putative TC cargos can be located between the Enc1 and Enc2 genes while putative PT cargos are found directly downstream of Enc2.

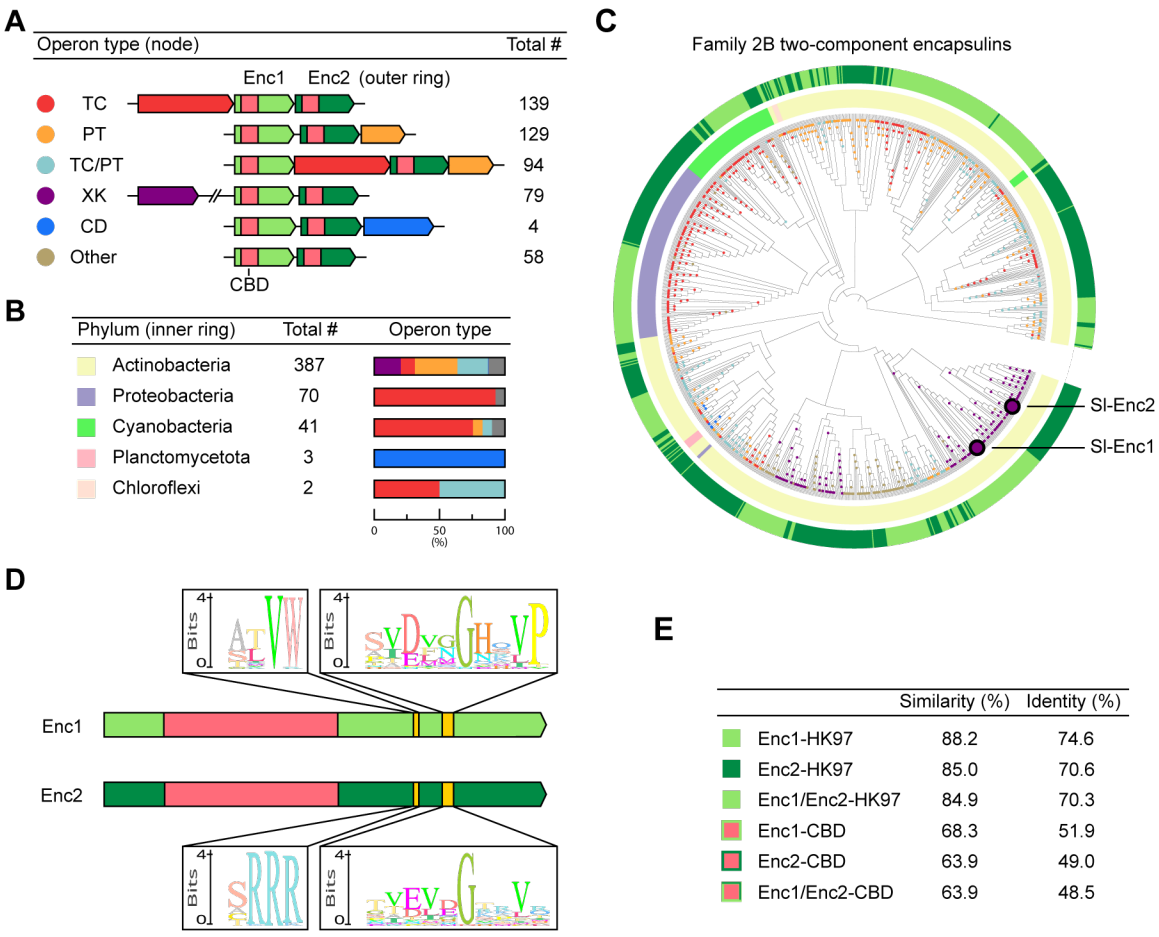


Figure 2-1: The HK97-fold and distribution and diversity of Family 2B two-component encapsulin systems. (**A**) Two-component Family 2B encapsulin operon types. TC: terpene cyclase, PT: polyprenyl transferase, XK: xylulose kinase, CD: cysteine desulfurase, CBD: cyclic nucleotide (cNMP)-binding domain. (**B**) Phylogenetic distribution of two-component encapsulin systems. (**C**) Phylogenetic tree of all identified shell proteins encoded in two-component Family 2B encapsulin operons. Outer ring: Enc1 vs Enc2. Inner ring: phylum. Nodes colored by cargo type. (**D**) Domain organization of the two *Streptomyces* lydicus (Sl) shell proteins highlighting two conserved sequence differences found in all Enc1 and Enc2 shell proteins (diagnostic motif 1 and 2). (**E**) Comparative sequence analysis of Enc1 and Enc2 shell proteins. HK97: HK97-domain.

Two-component operons are distributed across five bacterial phyla with more than 75% of systems found in the phylum Actinobacteria which also exhibits the largest diversity of operon types (Figure 2-1B). The other two phyla containing substantial numbers of two-component systems – Proteobacteria and Cyanobacteria – primarily encode TC-containing operons. Phylogenetic analysis of all encapsulin shell proteins identified in two-component operons highlights clustering within a phylum primarily based on cargo type (Figure 2-1C). Enc1s within

the same phylum sharing the same cargo type are more closely related to one another than to the Enc2s found in their respective operons. This observation may indicate a shared evolutionary trajectory of Enc1s, separate from Enc2s, and vice versa, prior to the formation of two-component operons. It may further suggest distinct and conserved roles for Enc1s and Enc2s in two- component systems.

Sequence analysis reveals two conserved diagnostic sequence differences – diagnostic motif 1 and 2 – between Enc1 and Enc2 shell proteins (Figure 2-1D). These short sequence stretches represent the only conserved substantial differences between Enc1 and Enc2. Using sequence alignments, the sequence identities of the HK97 domains of all Enc1s, all Enc2s, and all two-component shell proteins combined (Enc1s and Enc2s) were determined to be 74.6%, 70.6%, and 70.3%, further confirming the phylogenetic clustering observed above (Figure 2-1E). In addition, alignments focused only on the CBDs revealed a similar sequence identity trend – Enc1-CBDs: 51.9%, Enc2-CBDs: 49.0%, and Enc1/Enc2-CBDs: 48.5% – but with overall lower identities, highlighting that CBDs generally represent the most variable regions in Family 2B two-component shell proteins, including Sl-Enc1 and Sl-Enc2 (Figure 2-2).

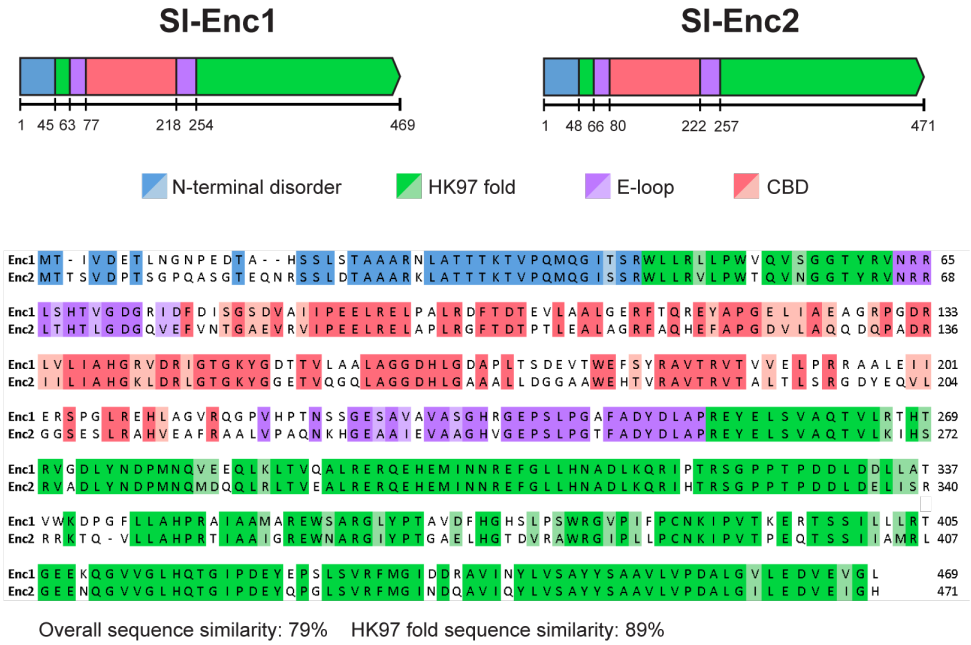


Figure 2-2: Sequence alignment of Sl-Enc1 and Sl-Enc2. A sequence alignment of Sl-Enc1 and Sl-Enc2 generated via Clustal Omega is shown. Sequences are colored by domain.

2.3.2 Heterologous expression of individual shell components highlights differential assembly behavior

Here, we focus on the two encapsulin shell proteins found in a putative XK two-component system from *S. lydicus* – Sl-Enc1 and Sl-Enc2. We first sought to determine whether Sl-Enc1 and Sl-Enc2 would individually assemble into encapsulin shells. Separate heterologous expression of Sl-Enc1 and Sl-Enc2 in *E. coli* BL21 (DE3), followed by affinity purification and size exclusion chromatography, yielded soluble purified protein (Figure 2-3A). Negative stain transmission electron microscopy (TEM) analysis of purified Sl-Enc1 and Sl-Enc2 highlighted that Sl-Enc1 did not assemble into regular encapsulin nanocompartments (Figure 2-3B). In contrast, Sl-Enc2 formed characteristic ca. 28nm shells, consistent with other T=1 Family 2B encapsulins (36). Native PAGE analysis further showed that Sl-Enc1 does not yield a single high

molecular weight band – as observed for Sl-Enc2. Rather, a broad smear is observed, confirming that Sl-Enc1 does not assemble into homogeneous protein shells, forming irregular aggregates instead (Figure 2-3C). This is surprising as Sl-Enc1 and Sl-Enc2 exhibit 79% overall sequence similarity and 89% sequence similarity of their shell-forming HK97-fold domains (Figure 2-2). The abovementioned diagnostic Enc1 and Enc2 sequence motifs, also found in Sl-Enc1 and Sl-Enc2, as well as the more variable CBDs may contribute to the observed differential assembly behavior of Sl-Enc1 and Sl-Enc2.

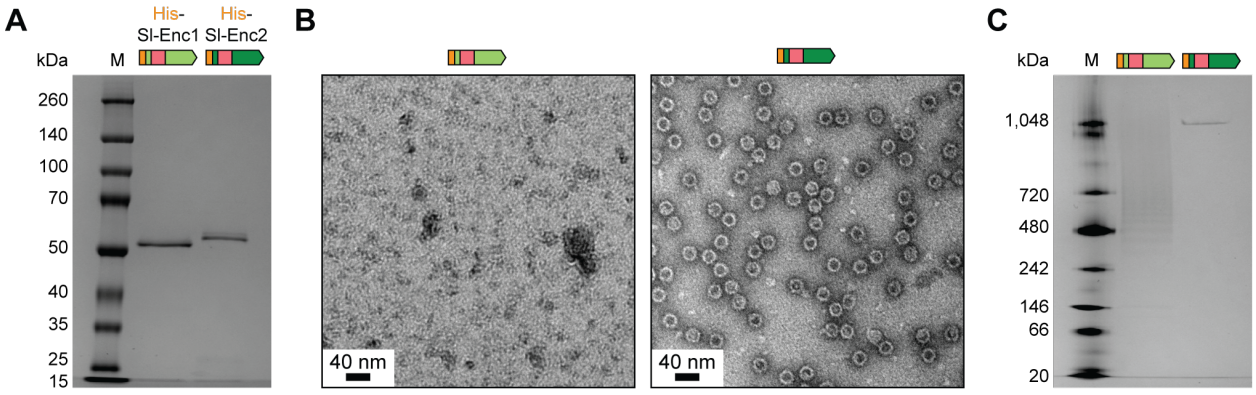


Figure 2-3: Separate analysis of Sl-Enc1 and Sl-Enc2. (**A**) SDS-PAGE analysis of individually expressed and purified Sl-Enc1 and Sl-Enc2. M: molecular weight marker. (**B**) Negative stain TEM micrographs of purified Sl-Enc1 and Sl-Enc2. (**C**) Native PAGE analysis of purified Sl-Enc1 and Sl-Enc2.

2.3.3 Sl-Enc1 and Sl-Enc2 co-assemble into compositionally variable two-component shells

To determine if Sl-Enc1 and Sl-Enc2 can co-assemble to form mixed encapsulin shells, despite the fact that Sl-Enc1 cannot form shells on its own, we pursued a differential His-tagging strategy. We constructed an expression plasmid encoding an Sl-Enc1-Sl-Enc2 two-gene operon with Sl-Enc1 carrying an N-terminal His-tag. As Sl-Enc1 cannot assemble into shells by itself, this tagging strategy would allow us to utilize affinity purification and size exclusion chromatography to confirm shell formation, followed by protein identification to investigate the

presence of both Sl-Enc1 and Sl-Enc2. Heterologous expression and purification of Sl-Enc1-Sl-Enc2 was carried out as described above. Clear encapsulin shells could be observed after purification via negative stain TEM (Figure 2-4A) and native PAGE (Figure 2-4B). The shells were of similar size and appearance compared to one-component Sl-Enc2 shells (Figure 2-3B). SDS-PAGE analysis showed the presence of two distinct bands (Figure 2-4C), identified as Sl-Enc1 (lower band) and Sl-Enc2 (upper band) via tryptic digest and LC-MS analysis. Repeat expressions and purifications of Sl-Enc1-Sl-Enc2 resulted in fairly consistent apparent ratios of Sl-Enc1 and Sl-Enc2 (ca. 60% Sl-Enc1 and 40% Sl-Enc2) (Figure 2-4C). Because only Sl-Enc1 is His-tagged, Sl-Enc1 must be incorporated into a mixed shell since it cannot form shells by itself and because one-component untagged Sl-Enc2 shells – which are likely formed as well – would have been removed during affinity purification. Notably, the incorporation of Sl-Enc1 did not alter the size or appearance of mixed encapsulin shells compared with homogeneous Sl-Enc2 shells. As Sl-Enc1 and Sl-Enc2 are highly similar in sequence, Sl-Enc1 appears to be able to replace Sl-Enc2 subunits in mixed shells. Following a similar differential His-tagging strategy as described above, a two-gene operon with only Sl-Enc2 carrying an N-terminal His-tag was constructed, expressed, and purified to determine if Sl-Enc1 could be included in shells containing primarily Sl-Enc2. Shell assembly was confirmed using TEM and native PAGE (Supplemental Figure 2-1, A and B). Although only a single major band was observed via SDS-PAGE analysis, LC-MS and tryptic digest of the excised gel band demonstrated that Sl-Enc1 was present (Supplemental Figure 2-1C). This suggests that Sl-Enc1 is a minor component of this mixed shell sample, although, given that the relative protein ratios are estimated at a bulk level, it does not exclude the possibility of some individual shells having high incorporation of Sl-Enc1.

Altogether, our differential His-tagging experiments establish that Sl-Enc1 incorporation into the encapsulin shell is variable and can range from 0% up to approximately 60% incorporation.

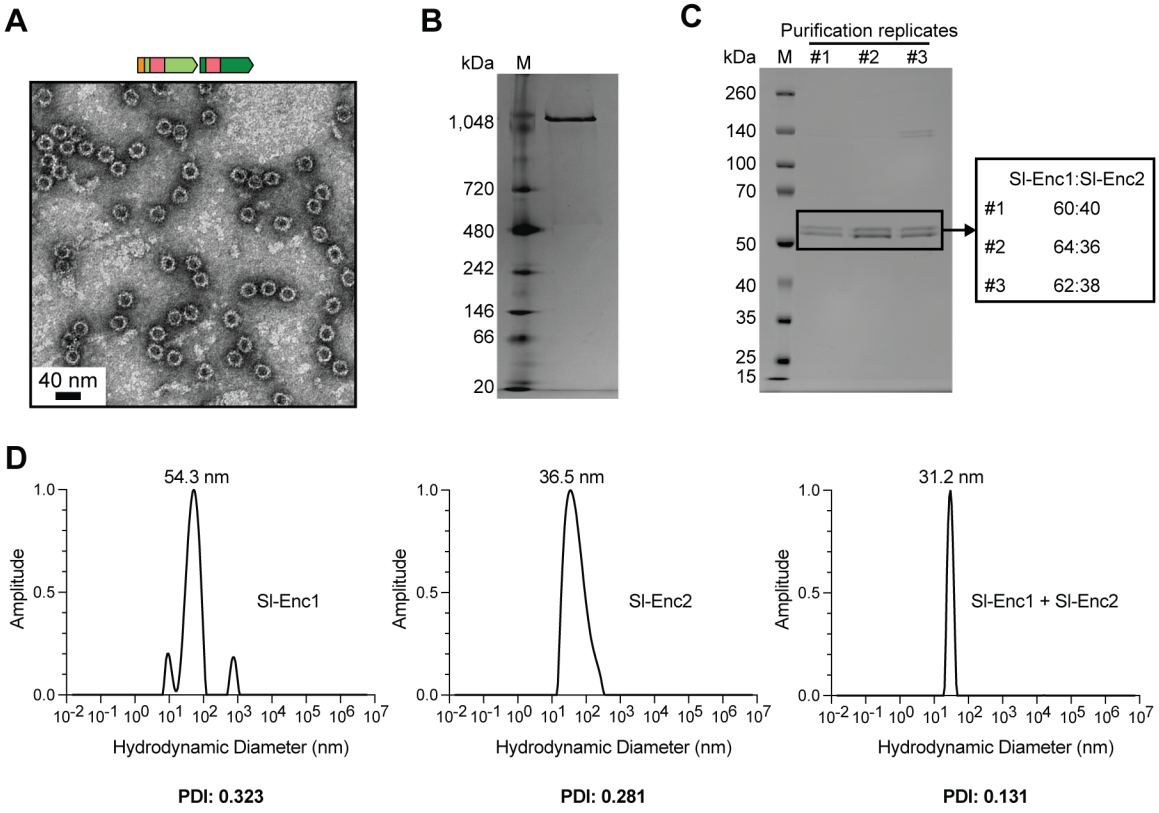


Figure 2-4: Analysis of mixed Sl-Enc1-Sl-Enc2 shells. (**A**) Negative stain TEM micrograph of purified Sl-Enc1 majority mixed shells. (**B**) Representative native PAGE analysis of an Sl-Enc1 majority mixed sample. (**C**) SDS-PAGE analysis of three replicate Sl-Enc1 majority mixed samples. The Sl-Enc1 to Sl-Enc2 ratio based on gel densitometry is shown on the right for individual purification replicates. (**D**) Dynamic light scattering (DLS) analysis of individually purified Sl-Enc1 (left) and Sl-Enc2 (middle) as well as of an Sl-Enc1 majority mixed sample (right). PDI: polydispersity index.

A potential preference for mixed shell formation is suggested by dynamic light scattering (DLS) analysis of purified samples with mixed shells exhibiting the narrowest size distribution and lowest polydispersity index (PDI) values (Figure 2-4D). The broad distribution and high PDI of Sl-Enc1 further confirms its inability to form homogeneous assemblies. Thermal ramps in combination with DLS and static light scattering (SLS) showed that mixed samples possess the highest melting temperature (T_m) of all analyzed samples, confirming that mixed shells are more stable than homogeneous Sl-Enc2 shells (Supplemental Figure 2-2).

2.3.4 Single particle cryo-EM analysis confirms mixed shell assembly and suggests irregular tiling of two-component shells

To gain a deeper understanding of the composition and structure of two-component shells, we performed single-particle cryo-EM analysis. In particular, we focused our efforts on two mixed samples with one of them exhibiting an excess of Sl-Enc1 (Figure 2-5A) and the other showing an excess of Sl-Enc2 (Figure 2-5B). Initially, we carried out consensus icosahedral (I) refinements and obtained symmetry-averaged cryo-EM maps at 2.58 Å (Sl-Enc1 majority) and 2.59 Å (Sl-Enc2 majority) (Supplemental Figure 2-3 and Supplemental Figure 2-4). We found that protomer excess resulted in consensus maps exclusively representing the majority component. That is to say, the Sl-Enc1 majority sample yielded a symmetry-averaged Sl-Enc1 shell map while the Sl-Enc2 majority sample resulted in a symmetry-averaged Sl-Enc2 shell density. As Sl-Enc1 cannot form shells on its own, the obtained symmetry-averaged Sl-Enc1 map has to be treated as artifactual. However, we were able to use this map to build an atomic model of the Sl-Enc1 protomer (Figure 2-6A). As mentioned above, Sl-Enc2 can assemble into complete shells by itself (Figure 2-3B). Therefore, we used the symmetry-averaged Sl-Enc2 map to build a model for a complete homogeneous Sl-Enc2 shell. The Sl-Enc2 shell is 28 nm in diameter, consists of 60 identical protomers, and exhibits T=1 icosahedral symmetry, similar to recently described Family 2B encapsulins (Fig. 4B) (35).

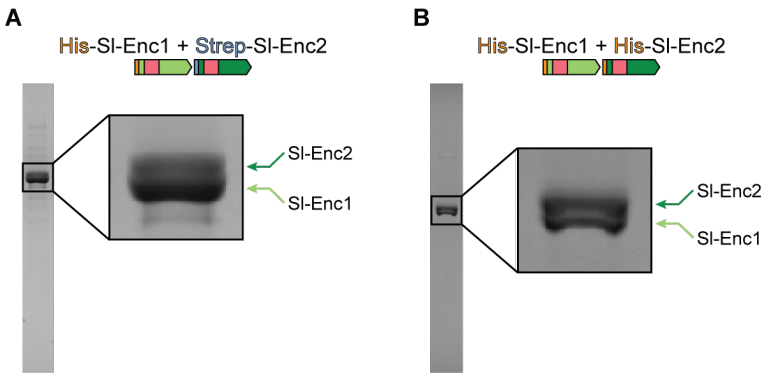


Figure 2-5: SDS-PAGE analysis of Sl-Enc1 majority and Sl-Enc2 majority samples subjected to cryo-EM analysis. (**A**) SDS-PAGE gel lane of protein sample used for cryo- EM data collection resulting in a symmetry-averaged Sl-Enc1 shell map. (**B**) SDS-PAGE gel lane of protein sample used for cryo-EM data collection resulting in a symmetry- averaged Sl-Enc2 shell map.

Both Sl-Enc1 and Sl-Enc2 protomers consist of two domains, a shell- forming canonical HK97-domain, conserved in all encapsulins, and a CBD domain inserted into the E-loop of the HK97-domain, only found in Family 2B encapsulins. The HK97-fold portions of both Sl-Enc1 and Sl-Enc2 exhibited high quality cryo-EM densities which allowed confident atomic model building. However, both CBDs exhibited weaker and lower resolutions densities, likely due to their inherent flexibility as highlighted by local resolution analysis (Supplemental Figure 2-3 and Supplemental Figure 2-4). To create complete Sl-Enc1 and Sl-Enc2 protomer models, we utilized AlphaFold predictions of the two CBDs, which fit the obtained lower resolution cryo-EM CBD maps very well and combined them with our HK97-fold domain models. While the overall structures of the Sl-Enc1 and Sl-Enc2 protomers are very similar, two major differences could be observed. First, Sl-Enc1 exhibits a clearly defined 23-residue N-arm, not present in the Sl-Enc2 protomer. Similar N-arms have been described in viral HK97-fold capsid proteins in addition to Family 2A encapsulins, but not Family 2B systems (9, 33, 34). In viruses, N-arms usually mediate inter-protomer interactions, can often adopt multiple dynamic conformations, and increase the overall stability of the viral capsid (3, 51-53). The second major difference between the Sl-Enc1 and Sl-Enc2 protomers is that their CBDs show different relative

orientations with respect to the HK97- domain with the Sl-Enc1 CBD being tilted by ca. 23° compared to the Sl-Enc2 CBD (Figure 2-6A and Figure 2-7). Overall, the homogeneous Sl-Enc2 shell is similar to recently structurally characterized Family 2B encapsulins with externally displayed CBDs being positioned above all two-fold pores (36). The two-fold pores are large and elongated with dimensions of 14 by 60 Å and partially occluded by the CBDs (Figure 2-6, C and D). The absence of any resolvable N-arms in the homogeneous Sl-Enc2 shell contributes to the large dimensions of these pores. Two additional types of pores are present in the Sl-Enc2 shell at the five- and three-fold axes of symmetry with diameters of 6 and 3 Å, respectively.

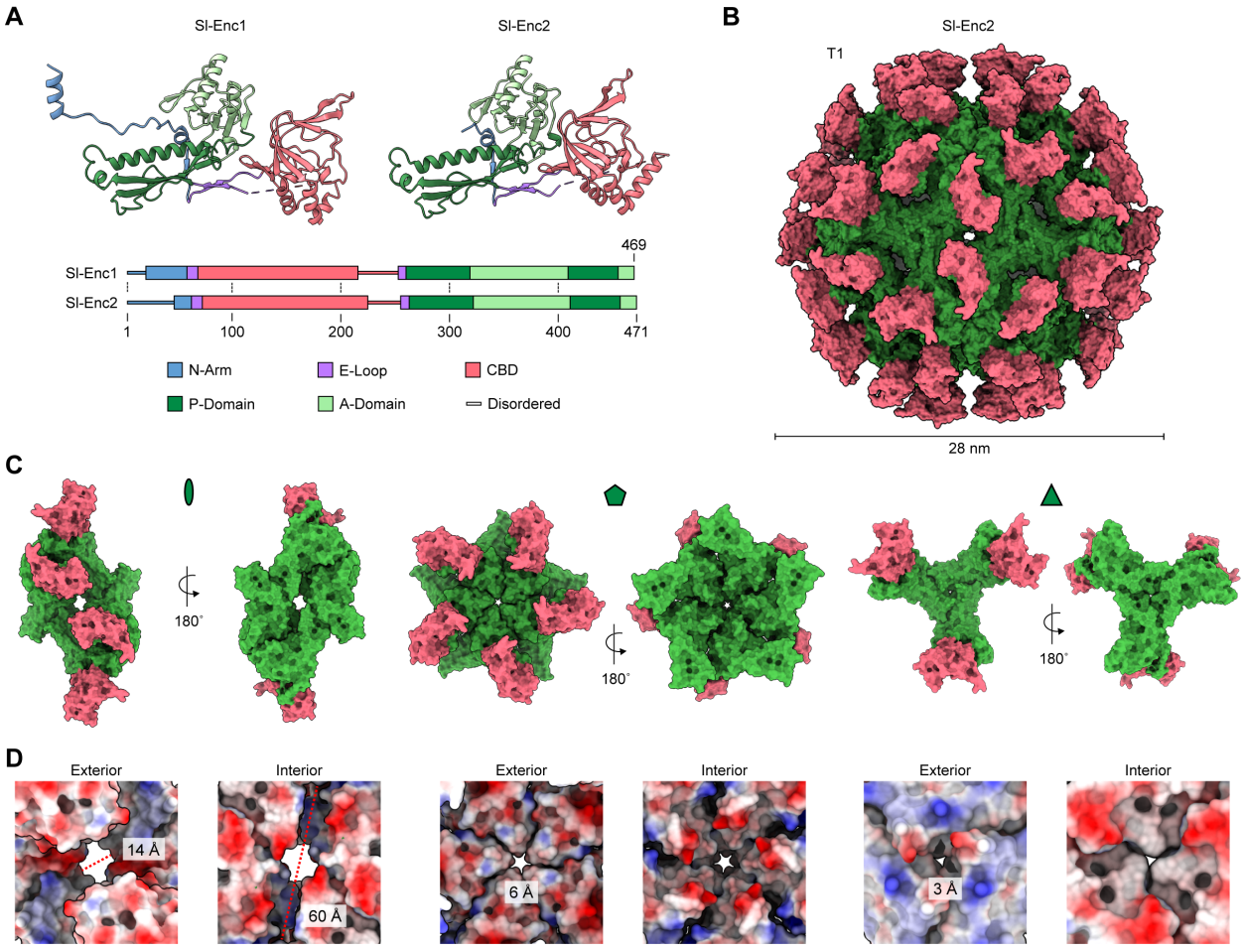


Figure 2-6: Structural analysis of mixed shells via single particle cryo-EM. (**A**) Protomer structures of Sl-Enc1 and Sl-Enc2. Structures colored by domain. (**B**) Homogeneous Sl-Enc2 shell with T=1 icosahedral symmetry.

HK97-domain colored green. CBD colored pink. (**C**) Exterior and interior views of the two- (left), five- (middle), and three-fold (right) interactions found in the homogeneous Sl-Enc2 shell. (**D**) Two- (left), five- (middle), and three-fold (right) pores of the homogeneous Sl-Enc2 shell shown as electrostatic surface. Both the exterior and interior are shown with pore diameters highlighted.

C1 refinements of mixed samples generally yielded low resolution maps of poor quality suggesting substantial compositional heterogeneity. To visualize both protomer types – Sl-Enc1 and Sl-Enc2 – in mixed shells, we carried out masked 3D classifications followed by masked local refinements on symmetry-expanded particles (I symmetry) (Supplemental Figure 2-5 to 2-8). We focused on the Sl-Enc1 majority sample where all shells can be assumed to be mixed as homogeneous Sl-Enc1 shells were not observed experimentally (Figure 2-3B). We created focus masks for the four unique interactions each protomer within a Family 2B shell must engage in (Figure 2-8A). It was our goal to identify all protomer interactions present in mixed two-component shells. We designed masks to completely encompass the HK97-domains of the two interacting protomers, omitting the CBDs due to their intrinsic conformational flexibility. We focused our analysis on two distinct interactions across the two-fold axis of symmetry (dubbed 2-fold (pore) and 2-fold (P-domain)), one across the three-fold axis, and one across the five-fold axis (Figure 2-8B). In principle, four distinct interactions are possible for all investigated interfaces: Sl-Enc1/Sl-Enc1, Sl-Enc2/Sl-Enc2, Sl-Enc1/Sl-Enc2, and Sl-Enc2/Sl-Enc1, with the two mixed protomer arrangements representing the same but mirrored interaction.

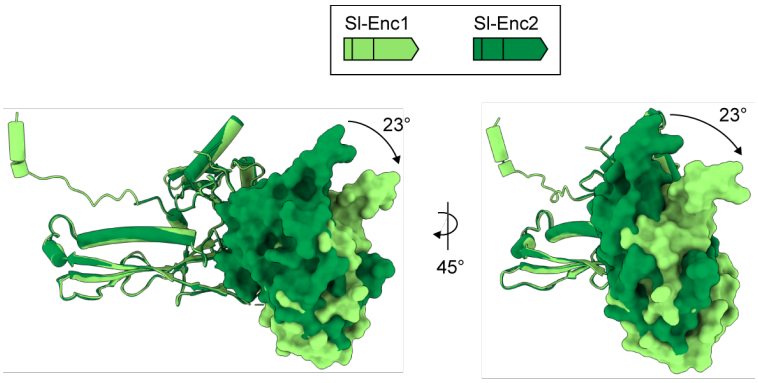


Figure 2-7: Comparison of CBD orientations within Sl-Enc1 and Sl-Enc2. Structural alignment of the Sl-Enc1 and Sl-Enc2 proteins highlighting the different CBD tilts relative to the HK97-domain. A relative tilt between the Sl-Enc2 and Sl-Enc1 CBDs of ca. 23° can be observed. CBDs shown as surface, HK97-domain as ribbon.

We were able to identify classes representing all four interactions for all masked 3D classifications (Figure 2-8C and Supplemental Figure 2-9 to 2-12). As the Sl-Enc1 and Sl-Enc2 protomers are highly similar in sequence and structure, we utilized one of the abovementioned diagnostic sequence motifs (diagnostic motif 1: Sl-Enc1: T_{337} VWK_{340} vs Sl-Enc2: $R_{340}RRK_{343}$) to differentiate between Sl-Enc1 and Sl-Enc2 (Figure 2-1D and Figure 2-8D). Our results suggest that mixed two-component shells exhibit irregular tiling of protomers with all possible permutations allowed. We further found that one of the investigated two-fold interactions – 2-fold (pore) – exhibits additional heterogeneity with respect to the presence or absence of resolved N-arms (Figure 2-9A). Specifically, two distinct symmetrical N-arm states could be observed in 3D classifications for Sl-Enc1/Sl-Enc1 interactions with N-arms either clearly resolved or completely absent. Symmetrical Sl-Enc2 or mixed interactions never exhibited resolved N-arms in our 3D classes. Thus, Sl-Enc1 protomers can exist in two different states, one with clearly resolved N-arms – as also observed in the symmetry-averaged consensus refinement – and one with completely unresolved N-arms. The discernable N-arm states result in substantial differences in the size and charge of the two-fold pores (Figure 2-9, A and B).

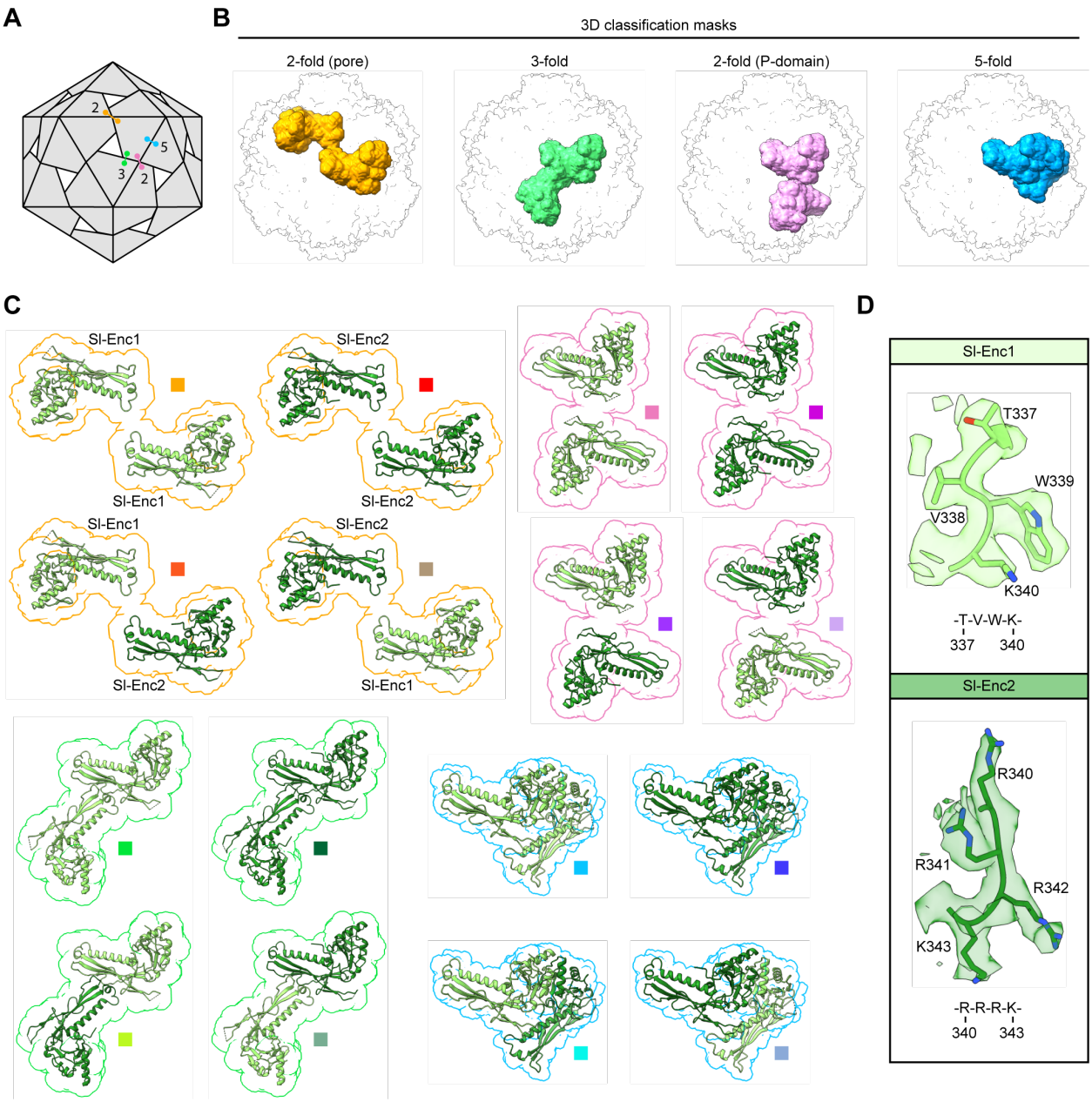


Figure 2-8: Protomer interaction analysis in mixed shells via 3D classification. (**A**) Schematic of a T=1 icosahedral protein shell highlighting the four unique interactions of a given protomer. Interactions are labelled by symmetry; orange: interaction across the 2-fold pore; green: interaction at the 3-fold pore; pink: 2-fold P-domain interaction; blue: 5-fold A-domain interaction. (**B**) 3D classification masks used to investigate all possible protomer interactions in mixed shells. (**C**) Observed protomer arrangements across the four unique interactions outlined in (A). Color scheme based on respective mask color. Sl-Enc1 shown in light green. Sl-Enc2 shown in dark green. (**D**) The diagnostic regions (diagnostic motif 1) within the Sl-Enc1 and Sl-Enc2 protomers used to unambiguously identify them. As an example, Sl-Enc1 from the Sl-Enc1-Sl-Enc1 2-fold P-domain interaction and Sl-Enc2 from the Sl-Enc2-Sl-Enc1 2-fold P- domain interaction is shown as cryo-EM density and atomic models with the respective sequence stretches shown below.

To show that all identified interactions and protomer arrangements are indeed present in most shells, we mapped a subset of 10,000 symmetry-expanded particles, contributing to the identified 3D classes, back onto the original micrographs (Figure 2-9, C to G). We used a subset of particles for visual clarity. This analysis suggests that all 16 possible positional permutations across the four investigated interactions are likely present in all shells (Figure 2-9H). Taken together, our analysis highlights that two-component shells consisting of Sl-Enc1 and Sl-Enc2 are pseudo-icosahedral and highly irregular with all possible protomer arrangements allowed. This compositional heterogeneity results in variable pore properties for the two-, three-, and five-fold pores. Pore variability is further exacerbated by Sl-Enc1 being able to exist with or without resolved N-arms.

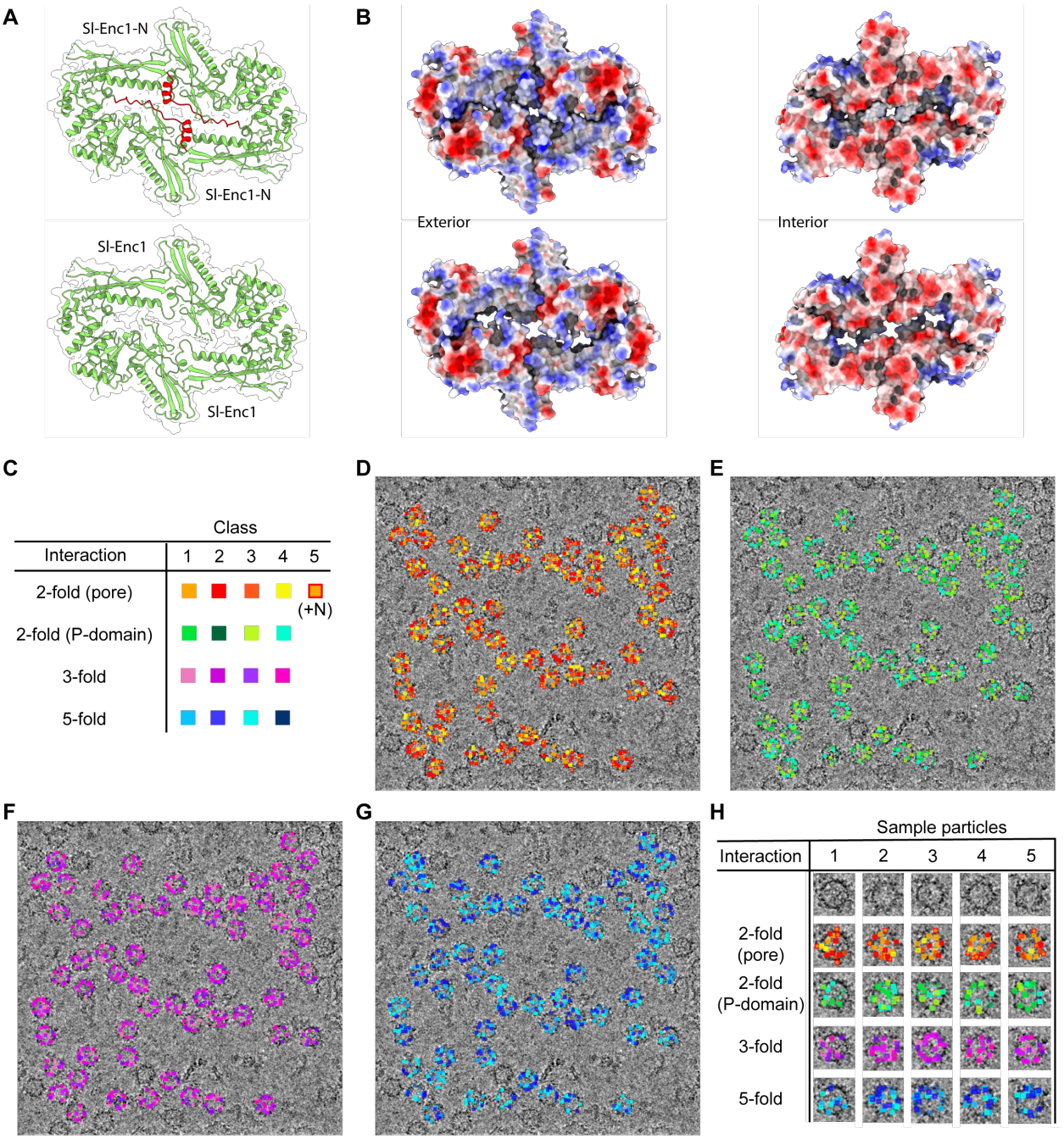


Figure 2-9: Two-fold pore variability and visualization of mixed shell heterogeneity. (**A**) Two-fold pore Sl-Enc1-Sl-Enc1 interactions observed via 3D classification analysis highlighting two resolved (top) or unresolved (bottom) N-arms (red). **(B)** Electrostatic surface representation of the two distinct two-fold pore Sl-Enc1-Sl-Enc1 arrangements. Top: two resolved N-arms; Bottom: unresolved N-arms. Left: exterior view. Right: interior view. (**C**) Color labelling scheme for the four unique interactions and protomer arrangements found across them used in subsequent panels. (**D**) Back-mapping of sub-particles contributing to the different classes identified for the 2-fold (pore) interaction via 3D classification onto a sample micrograph highlighting that all classes, representing unique protomer arrangements, can generally be found in each individual encapsulin particle. Each colored dot represents a single sub-particle sorted into one of the highlighted classes during 3D classification. (**E**) Back-mapping of sub-particles for the 2-fold (P-domain) interaction. (**F**) Back-mapping of sub-particles for the 3-fold interaction. (**G**) Back-mapping of sub-particles for the 5-fold interaction. (**H**) Five sample particles (1-5) from the sample

micrograph shown above are highlighted with all the distinct protomer arrangements shown using the color scheme of panel (A). All identified protomer arrangements across the four unique interactions can generally be found within each individual encapsulin particle. As a subset of 10,000 particles was used for this analysis, not all interactions that exist within each particle will be present.

2.3.5 Crosslinking analysis of two-component shells confirms heterogeneous two-fold interactions

To further investigate mixed shells, specifically with respect to the two-fold CBD interaction, analogous to the HK97-domain two-fold (P-domain) interaction discussed above, crosslinking studies followed by mass spectrometric analysis were carried out. Surface-exposed cysteines were introduced into the CBDs of either Sl-Enc1 (R195C) alone or both Sl-Enc1 and Sl-Enc2 (G198C) via site-directed mutagenesis (Figure 2-10A). The two kinds of mutant mixed shells were still able to properly assemble as confirmed by negative stain TEM of purified samples (Supplemental Figure 2-13, A to C). Site-selective crosslinking was carried out using a 17.8 Å cysteine-reactive double maleimide crosslinker (BM(PEG)3) (Figure 2-10A). As both Sl-Enc1 and Sl-Enc2 natively contain one buried cysteine residue, control crosslinking experiments with native mixed shells were carried out.

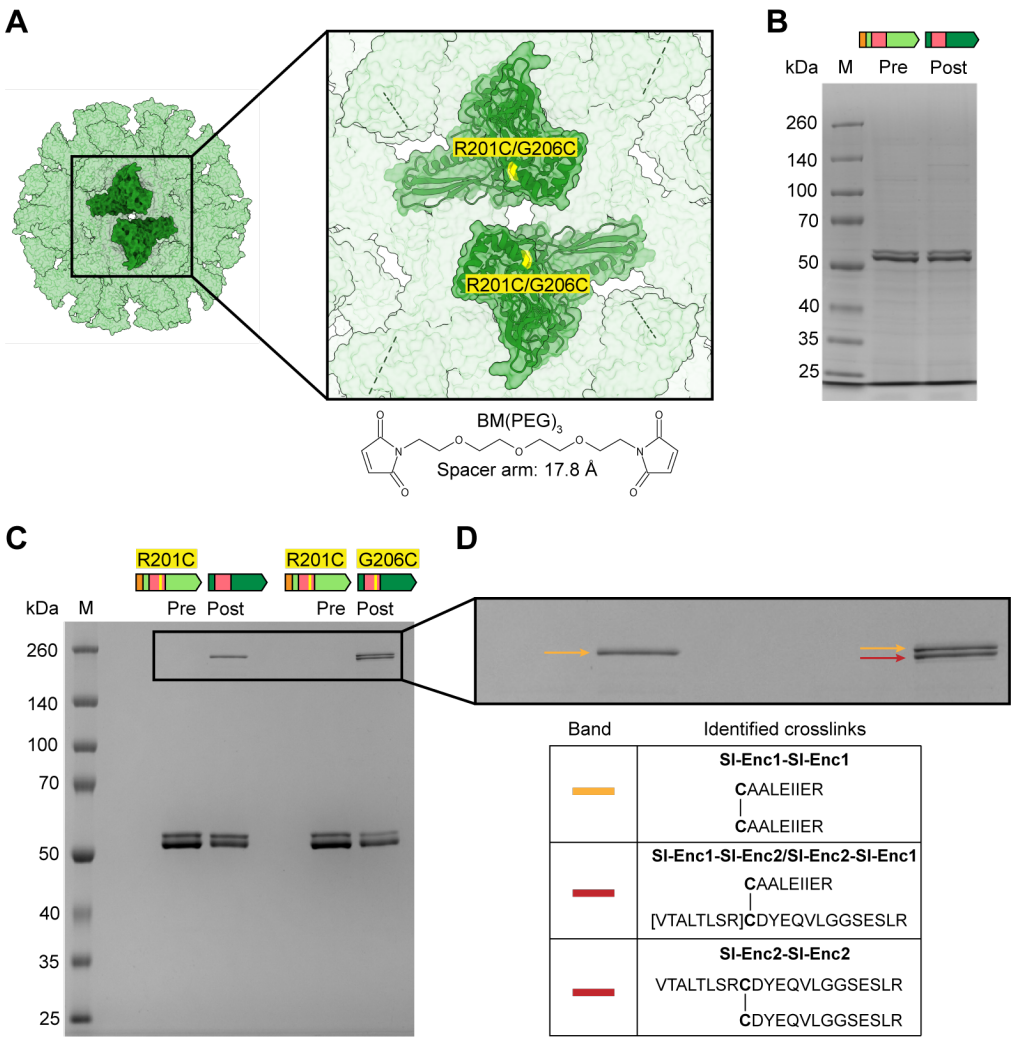


Figure 2-10: Site-specific crosslinking analysis of mixed shells. (**A**) Schematic of our crosslinking strategy based on CBD cysteine mutants (Sl-Enc1: R195C; Sl-Enc2: G198C). (**B**) SDS-PAGE analysis of a control crosslinking experiment using native mixed shells highlighting that no major higher molecular weight bands are present in the post crosslinking sample. (**C**) SDS-PAGE analysis of crosslinked mixed samples containing either one CBD cysteine (Sl-Enc1: R195C) or two CBD cysteines (Sl-Enc1: R195C and Sl-Enc2: G198C). The higher molecular weight bands observed in post crosslinking samples are outlined. (**D**) Mass spectrometric identification of crosslinks present in the higher molecular weight crosslinked bands. The identified crosslinks for each band are shown.

No crosslinking was observed in these control experiments with no higher molecular weight crosslinked bands visible on SDS-PAGE (Figure 2-10B). For mutant mixed shells containing only the Sl-Enc1 cysteine (C195), a single crosslinked band could be observed on SDS-PAGE while two bands could be clearly visualized when both Sl-Enc1 (C195) and Sl-Enc2 (C198) cysteines were present (Figure 2-10C). Subsequent tryptic digest and tandem mass spectrometric analysis of excised bands showed that the upper band, present in both samples,

contains Sl-Enc1-Sl-Enc1 crosslinked peptides and the lower band, only present in the double cysteine (Sl-Enc1: C195 and Sl-Enc2: C198) mixed shells, contains both Sl-Enc1-Sl-Enc2 and Sl-Enc2-Sl-Enc2 crosslinked peptides (Figure 2-10D, and Supplemental Figure 2-13D). These results confirm our 3D classification results of the two-fold (P-domain) interaction and highlight that all possible interactions across this two-fold interface are present in mixed shells.

2.4 Discussion

In this study, we report the structure and assembly principles of a two-component Family 2B encapsulin nanocompartment found in *Streptomyces lydicus*, highlighting its quasi-icosahedral geometry, variable shell composition, and irregular shell tiling. Our bioinformatic searches establish two-component encapsulin operons as widespread – 503 systems across five bacterial phyla, representing 21% of all Family 2B operons – and diverse – five distinct operon types with different cargo. Notably, the only operons to always encode two shell components are xylulose kinase (XK) systems. All other cargo types can be found in both two-component and single component operon arrangements (7). While this observation may be a result of the limited number and biased phylogenetic sampling of sequenced bacterial genomes, it could also indicate that some aspects of the two-component shell are essential for the proper functioning of specifically XK encapsulin systems. Our phylogenetic analysis further showed that Enc1s within a given phylum are more similar to each other than to Enc2s found within their respective operons (Figure 2-1E). This could potentially hint at separate evolutionary histories of Enc1s and Enc2s which may have existed as single component encapsulin systems in the past. The strong gene synteny for Enc1 and Enc2 genes, observed in current two-component systems, would further suggest distinct and conserved roles for each shell component, supported by our

experimental results demonstrating a preference for mixed shell formation (Supplemental Figure 2-1 and Supplemental Figure 2-2).

We confirmed the co-assembly of Sl-Enc1 and Sl-Enc2 into mixed shells using a differential His-tagging strategy (Figure 2-4) and further validated these findings with 3D classification analyses. We found that all possible protomer-protomer interactions are in principle allowed in mixed shells (Figure 2-8, and Figure 2-9, C to H). These findings demonstrate that both shell proteins can engage in all interactions across all interfaces which in turn implies that at the protomer-protomer interaction level, the sequence and structural differences between Sl-Enc1 and Sl-Enc2 do not restrict their interactions (Figure 2-8). CBDs positioned across from each other arranged around the two-fold pore can similarly tolerate all four CBD pairings based on crosslinking studies (Figure 2-10). These findings were surprising, as Sl-Enc1 is unable to form shells by itself (Figure 2-3, B and C) implying that there exists some intrinsic limitation in its self-assembly capacity. This fact, in combination with Sl-Enc1's propensity for aggregation rather than the formation of a single non-shell oligomeric state, may indicate that its inability to form homogeneous shells results from a cumulative structural effect. For example, slightly unfavorable angles between Sl-Enc1 protomer-protomer interfaces or the presence of too many defined N-arms – only observed for Sl-Enc1 – could potentially propagate if too many Sl-Enc1s are present in a given shell, resulting in an inability to form a closed compartment and a propensity towards aggregation. On the other hand, the observed preference for mixed shell formation may be partially attributed to the stabilizing effect of some, but not too many, defined Sl-Enc1 N-arms at the two-fold pore within mixed shells (Figure 2-9). As shell protein features of Family 2B encapsulins are highly conserved, it seems reasonable to

hypothesize that mixed shell formation represents a general feature of two-component encapsulin systems.

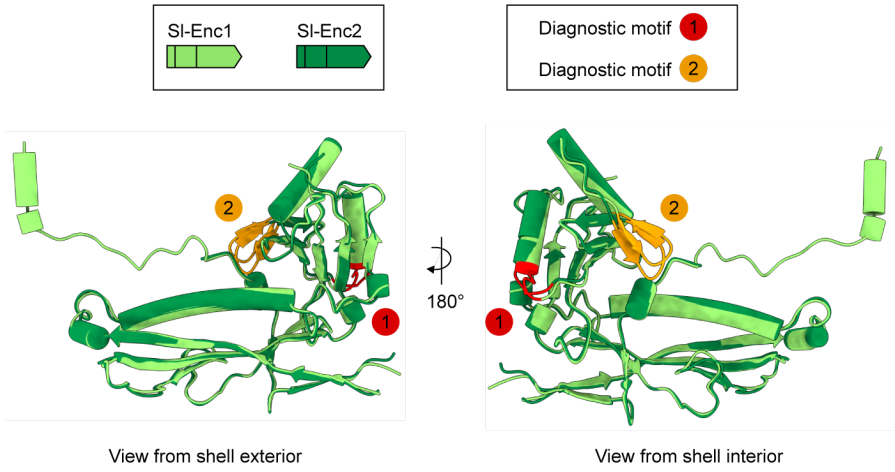


Figure 2-11: Location of the diagnostic motifs within the Sl-Enc1 and Sl-Enc2 protomers. The two conserved differences between Enc1 and Enc2 shell proteins are highlighted. CBDs not shown for visual clarity.

Using single particle cryo-EM, we found that the HK97-domain structures of Sl-Enc1 and Sl-Enc2 are nearly identical (Figure 2-6A), with only three regions exhibiting substantial differences. One such difference is the presence of a resolved N-arm for Sl-Enc1 that was notably absent for Sl-Enc2. Similar N-arms, resulting in a chainmail-like topology, have also been observed in Family 2A encapsulins (33, 34) and many HK97-fold viral capsid structures and are believed to contribute to the generally high stability of HK97-fold assemblies. The other two regions of structural deviation between Sl-Enc1 and Sl-Enc2 – also detectable at the sequence level – were dubbed diagnostic sequence motifs (Figure 2-1, D and E and Figure 2-11). One of these motifs – diagnostic motif 1 – is a stretch of four residues found in the A-domain located at the interface between two neighboring protomers within shell pentamers (Sl-Enc1: $T_{337}VWK_{340}$ vs Sl-Enc2: $R_{340}RRK_{343}$). Interestingly, these motifs are adjacent to the N-arms resolved for Sl-Enc1. The close proximity of this major difference between Sl-Enc1 and Sl-Enc2 to the N-arm could indicate that this diagnostic sequence motif may drive the preference for

mixed shells by improving overall shell stability through incorporation of structured Sl-Enc1 N-arms into Sl-Enc2 shells. The other primary difference between the HK97-domains of Sl-Enc1 and Sl-Enc2 – diagnostic motif 2 – is found on the luminal side of the A-domain (diagnostic motif 2: Sl-Enc1: A_{368}VDFHGHSLP$_{377}$ vs Sl-Enc2: G_{370}AELHGTDVR$_{379}$). Since this motif is not located at a protomer-protomer interface, it is unlikely to impact shell assembly preferences. However, the positioning of this loop in the shell interior makes it a candidate for facilitating cargo loading (Figure 2-11). With two distinct shell components and two distinct internal loops, two different cargo- loading mechanisms could facilitate the loading of multiple cargo proteins with approximately constant cargo stoichiometries determined by shell composition.

One of the defining features of Family 2B encapsulins is the presence of an insertion domain of so far unknown function within the E-loop of the HK97-fold. It has been previously shown that even though annotated as a cyclic nucleotide-binding domain (CBD) – specifically a cyclic adenosine monophosphate (cAMP)-binding domain – exhibiting the expected CBD protein fold (PF00027), this insertion domain does not bind any cyclic nucleotides (36). This can be rationalized by some of the generally conserved CBD binding site residues being divergent in Family 2B CBDs. It was suggested that some other unidentified small molecule ligands may bind to encapsulin CBDs instead (36). The CBDs in Sl-Enc1 and Sl-Enc2 are the most variable regions of these shell proteins as highlighted by their low sequence identities (Figure 2-1E). Further, the low local resolution observed via cryo-EM indicates a high degree of CBD conformational flexibility or mobility (Supplemental Figure 2-3 and Supplemental Figure 2-4). The binding sites of the two CBDs in Sl-Enc1 and Sl-Enc2 also deviate from canonical CBD binding sites. Although the specific roles of CBDs remain unclear, their positioning above the major two-fold pore likely has implications for overall encapsulin function. It would be expected

that the conformation and relative positioning of CBDs would greatly influence the level of occlusion at the two-fold pore, thereby regulating metabolite flux through the shell, potentially in a ligand-dependent fashion.

There are few known polyhedral protein compartments with shells or capsids that require multiple major, non-accessory or -dispensable, components and incorporate these components in a variable manner, as observed in the two-component encapsulin shell characterized in this work. One example are bacterial microcompartments (BMCs), a widespread class of bacterial protein compartments able to sequester multiple enzymes involved in carbon fixation or nutrient utilization (54). BMC shells are composed of three types of oligomers: hexamers (PF00936) that comprise most of the icosahedral faces (55), trimers (pseudo-hexamers) that influence the curvature and size of the compartment, also found within the icosahedral faces (56, 57), and pentamers (PF03319), necessary for forming closed shells, located at all shell vertices (58). Between three and eleven distinct shell components can be found across different predicted BMC operons (59). For one of the most well studied BMCs, the *Halothiobacillus neapolitanus* α-carboxysome, five distinct shell components are generally observed in natively isolated compartments (60). It has been suggested that this redundancy of shell components, especially for hexameric and trimeric oligomers, allows BMCs to adjust and fine-tune shell properties based on the dynamic needs of the host cell (57). For example, differences in pore charge and size likely influence the diffusion of small molecules across the BMC shell, and interactions with cargo proteins are usually mediated by specific shell proteins (61, 62). Since our studies of the *S. lydicus* two-component encapsulin were performed in the absence of cargo – no native cargo has yet been determined for XK Family 2B systems – it is possible that two-component encapsulins may benefit from the fine-tuning of shell properties for metabolite flux and having optimized

interaction surfaces for cargo or external binders. Some hexameric BMC shell proteins have also been shown to form both homo- and hetero-hexameric oligomers of variable composition when heterologously expressed (63, 64). While this heterogeneity more closely mirrors the random tiling observed for the *S. lydicus* two-component shell, the functional role of BMC hetero-hexamers remains to be determined, especially since preliminary evidence is mixed regarding potential stability benefits (64).

Besides the encapsulin described in this work, bacteriophage T4 (T4) is another known example of an HK97 assembly to preferentially utilize two separate protomers (65). Structural data has demonstrated that one T4 capsid protein (gp24) forms homo-pentamers that cap the T4 vertices, while a second T4 capsid protein (gp23) forms homo-hexamers that co-assemble with pentamers into the T4 prolate head with icosahedral T=13 ends closing the central cylindrical section (66, 67). All other HK97 capsids or shells with a triangulation number greater than T=1 rely on the intrinsic flexibility of the HK97-fold protomer – especially the E-loop – to form both pentamers and hexamers from a single protomer. The two T4 capsid proteins (gp23 and gp24) have only 17% amino acid sequence identity (17), much lower than the two S. lydicus encapsulin shell proteins (Figure 2-2), although their structures are highly similar. Several A-domain mutations have also been identified that allow gp23 to form pentamers and substitute for gp24 (67, 68). No evidence has been found to suggest the presence of hetero-pentamers or - hexamers in the T4 capsid, in contrast to the *S. lydicus* two-component encapsulin which is composed of seemingly random hetero-pentamers with variable composition of the two shell components Sl-Enc1 and Sl-Enc2.

There are a number of potential functions provided by a two-component encapsulin shell that are difficult to realize using a single shell system. For example, the encapsulation of multiple

cargos with distinct cargo encapsulation mechanisms can be easily envisioned in a two-component shell with distinct luminal shell binding sites for targeting peptides or cargo loading domains provided by two distinct shell components. Further, having more than one shell component may allow the fine-tuning of pore properties to optimize the metabolite flux of small molecules through the shell. While the function of CBDs found in Family 2B encapsulins remains to be determined, it has previously been hypothesized that small molecule ligand binding to externally displayed CBDs may modulate the properties of the shell or even the activity of encapsulated cargo in a dynamic fashion (36). Having two distinct CBDs present within a single shell may further expand this regulatory potential through the ability to bind two distinct ligands, thus integrating two separate regulatory signals.

In conclusion, we have demonstrated the existence of an HK97-fold protein compartment that preferentially uses two distinct shell components to form irregularly tiled encapsulin shells, expanding the known assembly modes of HK97-fold proteins. Our work lays the foundation for future studies aimed at elucidating the biological function of XK encapsulin systems and dissecting the functional logic of two-component encapsulins in general.

2.5 Supplementary Information

Supplemental Table 2-1: DNA sequences of gBlock gene fragments.

gBlock	DNA Sequence
Sl-Enc1	AAGTATAAGAAGGAGATATACAATGACAATCGTAGACGAGACGTTAAATGGGAACCCAGAGGAC ACCGCCCATTCGTCTTTGTCTACGGCTGCCGCCCGCAACTTGGCCACCACAACCAAAACTGTACCG CAAATGCAGGGCATCACATCCCGTTGGTTGCTTCGTTTGTTACCGTGGGTTCAGGTCTCAGGGGGC ACGTACCGCGTAAACCGCCGTTTATCACACACCGTGGGGGATGGACGTATCGACTTTGATATTTCT GGTAGTGACGTAGCGATTATCCCCGAAGAATTACGCGAACTTCCAGCGCTTCGCGACTTCACGGA TACTGAGGTCCTTGCTGCATTGGGGGAGCGTTTCACTCAGCGTGAGTATGCCCCAGGAGAGTTGA TCGCAGAGGCTGGGCGCCCAGGGGATCGCCTTGTTTTAATCGCGCATGGCCGCGTAGATCGCATT GGAACTGGAAAGTACGGCGACACCACCGTACTTGCAGCGCTTGCGGGGGGCGACCACCTGGGGG ACGCGCCGCTTACCTCTGACGAAGTTACATGGGAATTTTCGTATCGCGCGGTTACTCGCGTAACA GTAGTCGAGCTTCCGCGCCGCGCCGCCTTAGAAATCATCGAACGTTCCCCAGGTCTTCGTGAGCA TTTAGCCGGGGTACGTCAAGGGCCGGTTCACCCCACAAACAGTTCCGGCGAATCTGCTGTAGCCG

	TGGCTAGTGGTCACCGTGGCGAGCCGTCGCTGCCCGGAGCGTTCGCGGATTATGACTTAGCGCCC CGCGAGTACGAACTTAGCGTCGCTCAGACAGTATTGCGCACGCATACGCGCGTGGGGGATTTATA CAATGATCCTATGAACCAGGTGGAAGAACAATTAAAATTGACCGTCCAGGCACTGCGCGAACGC CAGGAACATGAGATGATCAACAATCGTGAGTTCGGATTGCTTCACAATGCAGACTTAAAGCAACG TATTCCTACGCGTAGCGGACCGCCTACTCCTGATGACTTGGATGATTTGCTGGCGACAGTTTGGAA GGACCCTGGCTTTCTTTTAGCGCATCCACGCGCCATCGCAGCAATGGCACGTGAATGGTCGGCTC GCGGACTGTATCCAACGGCGGTCGATTTTCATGGGCATTCTCTTCCATCCTGGCGTGGTGTTCCAA TCTTTCCGTGCAATAAAATTCCAGTCACTAAGGAACGCACAAGCTCCATTCTGCTTCTTCGCACCG GCGAGGAAAAGCAGGGCGTCGTCGGCCTTCACCAAACTGGGATTCCTGACGAGTATGAGCCAAG TCTGTCAGTTCGTTTCATGGGAATCGATGACCGCGCTGTAATTAACTACTTGGTTTCTGCGTACTA CAGTGCAGCCGTGCTTGTACCAGATGCATTGGGAGTCTTAGAGGATGTCGAAGTCGGGTTATGAA TTAACCTAGGCTGCTG
Sl-Enc2	AAGTATAAGAAGGAGATATACAATGACTACGTCCGTGGATCCAACTTCAGGTCCTCAGGCAAGTG GGACAGAGCAGAATCGCTCAAGCTTGGACACAGCCGCTGCTCGTAAACTGGCTACTACCACGAA AACGGTCCCGCAAATGCAGGGCATTTCATCACGCTGGTTGCTTCGTGTCTTACCGTGGACGCAGG TCAATGGGGGTACATACCGTGTTAATCGCCGCCTGACTCATACCCTGGGAGACGGTCAAGTGGAG TTCGTAAACACCGGCGCTGAAGTTCGTGTAATTCCGGAGGAGCTGCGTGAGTTGGCCCCGCTTCG TGGGTTTACCGATACTCCAACTCTGGAAGCCCTTGCGGGTCGCTTTGCACAACACGAGTTTGCTCC AGGCGATGTCTTAGCTCAGCAGGACCAACCGGCTGACCGCATCATCTTGATTGCTCACGGTAAGT TGGATCGTTTAGGCACTGGCAAGTATGGCGGAGAAACGGTGCAAGGGCAGTTGGCCGGCGGTGA CCACCTGGGAGCGGCCGCACTGTTAGACGGTGGTGCGGCCTGGGAGCATACTGTCCGCGCTGTTA CCCGCGTCACTGCACTGACTTTATCGCGCGGTGATTATGAACAAGTCCTGGGTGGTTCCGAGAGC TTACGTGCGCACGTCGAAGCTTTTCGTGCGGCCTTAGTACCGGCTCAGAATAAGCACGGCGAGGC TGCGATCGAAGTAGCTGCCGGTCATGTCGGGGAACCCTCCCTGCCAGGGACGTTTGCCGATTACG ATCTTGCTCCTCGTGAGTACGAGCTTTCAGTAGCCCAGACGGTTTTAAAGATTCACTCTCGTGTTG CCGACCTGTATAATGACCCCATGAACCAAATGGATCAGCAATTACGCTTAACTGTCGAGGCTTTG CGCGAACGCCAAGAACATGAGATGATCAACAACCGTGAATTTGGTTTACTGCATAATGCAGATTT AAAGCAGCGCATCCATACGCGCTCCGGCCCCCCAACACCTGACGACCTGGACGAGTTAATCTCTC GCCGCCGCAAAACGCAAGTGTTGTTGGCGCATCCACGCACTATTGCCGCCATTGGTCGTGAGTGG AACGCGCGTGGAATCTATCCGACGGGCGCCGAATTGCATGGAACAGATGTACGCGCTTGGCGTGG AATTCCTTTACTTCCTTGTAACAAAATTCCCGTAACCCCGGAACAAACAAGCAGTATCATTGCTAT GCGCTTGGGTGAGGAAAACCAGGGGGTAGTCGGATTGCACCAGACCGGCATTCCTGACGAGTAT CAACCAGGCTTGTCTGTTCGTTTTATGGGGATTAATGATCAGGCAGTCATCCAGTATCTTGTTAGC GCGTACTATTCAGCAGCAGTGTTGGTGCCTGATGCTTTGGGGATTCTGGAGGATGTTGAAATCGG GCATTGAATTAACCTAGGCTGCTGC
Sl-Enc1+Sl-Enc2	AAGTATAAGAAGGAGATATACAATGACAATCGTAGACGAGACGTTAAATGGGAACCCAGAGGAC ACCGCCCATTCGTCTTTGTCTACGGCTGCCGCCCGCAACTTGGCCACCACAACCAAAACTGTACCG CAAATGCAGGGCATCACATCCCGTTGGTTGCTTCGTTTGTTACCGTGGGTTCAGGTCTCAGGGGGC ACGTACCGCGTAAACCGCCGTTTATCACACACCGTGGGGGATGGACGTATCGACTTTGATATTTCT GGTAGTGACGTAGCGATTATCCCCGAAGAATTACGCGAACTTCCAGCGCTTCGCGACTTCACGGA TACTGAGGTCCTTGCTGCATTGGGGGAGCGTTTCACTCAGCGTGAGTATGCCCCAGGAGAGTTGA TCGCAGAGGCTGGGCGCCCAGGGGATCGCCTTGTTTTAATCGCGCATGGCCGCGTAGATCGCATT GGAACTGGAAAGTACGGCGACACCACCGTACTTGCAGCGCTTGCGGGGGGCGACCACCTGGGGG ACGCGCCGCTTACCTCTGACGAAGTTACATGGGAATTTTCGTATCGCGCGGTTACTCGCGTAACA GTAGTCGAGCTTCCGCGCCGCGCCGCCTTAGAAATCATCGAACGTTCCCCAGGTCTTCGTGAGCA TTTAGCCGGGGTACGTCAAGGGCCGGTTCACCCCACAAACAGTTCCGGCGAATCTGCTGTAGCCG TGGCTAGTGGTCACCGTGGCGAGCCGTCGCTGCCCGGAGCGTTCGCGGATTATGACTTAGCGCCC CGCGAGTACGAACTTAGCGTCGCTCAGACAGTATTGCGCACGCATACGCGCGTGGGGGATTTATA CAATGATCCTATGAACCAGGTGGAAGAACAATTAAAATTGACCGTCCAGGCACTGCGCGAACGC CAGGAACATGAGATGATCAACAATCGTGAGTTCGGATTGCTTCACAATGCAGACTTAAAGCAACG TATTCCTACGCGTAGCGGACCGCCTACTCCTGATGACTTGGATGATTTGCTGGCGACAGTTTGGAA GGACCCTGGCTTTCTTTTAGCGCATCCACGCGCCATCGCAGCAATGGCACGTGAATGGTCGGCTC GCGGACTGTATCCAACGGCGGTCGATTTTCATGGGCATTCTCTTCCATCCTGGCGTGGTGTTCCAA TCTTTCCGTGCAATAAAATTCCAGTCACTAAGGAACGCACAAGCTCCATTCTGCTTCTTCGCACCG GCGAGGAAAAGCAGGGCGTCGTCGGCCTTCACCAAACTGGGATTCCTGACGAGTATGAGCCAAG TCTGTCAGTTCGTTTCATGGGAATCGATGACCGCGCTGTAATTAACTACTTGGTTTCTGCGTACTA CAGTGCAGCCGTGCTTGTACCAGATGCATTGGGAGTCTTAGAGGATGTCGAAGTCGGGTTATGAT ATAAGAAGGAGATATACATGACTACGTCCGTGGATCCAACTTCAGGTCCTCAGGCAAGTGGGACA GAGCAGAATCGCTCAAGCTTGGACACAGCCGCTGCTCGTAAACTGGCTACTACCACGAAAACGGT CCCGCAAATGCAGGGCATTTCATCACGCTGGTTGCTTCGTGTCTTACCGTGGACGCAGGTCAATG GGGGTACATACCGTGTTAATCGCCGCCTGACTCATACCCTGGGAGACGGTCAAGTGGAGTTCGTA AACACCGGCGCTGAAGTTCGTGTAATTCCGGAGGAGCTGCGTGAGTTGGCCCCGCTTCGTGGGTT

	TACCGATACTCCAACTCTGGAAGCCCTTGCGGGTCGCTTTGCACAACACGAGTTTGCTCCAGGCG ATGTCTTAGCTCAGCAGGACCAACCGGCTGACCGCATCATCTTGATTGCTCACGGTAAGTTGGAT CGTTTAGGCACTGGCAAGTATGGCGGAGAAACGGTGCAAGGGCAGTTGGCCGGCGGTGACCACC TGGGAGCGGCCGCACTGTTAGACGGTGGTGCGGCCTGGGAGCATACTGTCCGCGCTGTTACCCGC GTCACTGCACTGACTTTATCGCGCGGTGATTATGAACAAGTCCTGGGTGGTTCCGAGAGCTTACGT GCGCACGTCGAAGCTTTTCGTGCGGCCTTAGTACCGGCTCAGAATAAGCACGGCGAGGCTGCGAT CGAAGTAGCTGCCGGTCATGTCGGGGAACCCTCCCTGCCAGGGACGTTTGCCGATTACGATCTTG CTCCTCGTGAGTACGAGCTTTCAGTAGCCCAGACGGTTTTAAAGATTCACTCTCGTGTTGCCGACC TGTATAATGACCCCATGAACCAAATGGATCAGCAATTACGCTTAACTGTCGAGGCTTTGCGCGAA CGCCAAGAACATGAGATGATCAACAACCGTGAATTTGGTTTACTGCATAATGCAGATTTAAAGCA GCGCATCCATACGCGCTCCGGCCCCCCAACACCTGACGACCTGGACGAGTTAATCTCTCGCCGCC GCAAAACGCAAGTGTTGTTGGCGCATCCACGCACTATTGCCGCCATTGGTCGTGAGTGGAACGCG CGTGGAATCTATCCGACGGGCGCCGAATTGCATGGAACAGATGTACGCGCTTGGCGTGGAATTCC TTTACTTCCTTGTAACAAAATTCCCGTAACCCCGGAACAAACAAGCAGTATCATTGCTATGCGCTT GGGTGAGGAAAACCAGGGGGTAGTCGGATTGCACCAGACCGGCATTCCTGACGAGTATCAACCA GGCTTGTCTGTTCGTTTTATGGGGATTAATGATCAGGCAGTCATCCAGTATCTTGTTAGCGCGTAC TATTCAGCAGCAGTGTTGGTGCCTGATGCTTTGGGGATTCTGGAGGATGTTGAAATCGGGCATTG ATTAACCTAGGCTGCTGC

Supplemental Table 2-2: DNA sequences of primers used to construct plasmids.

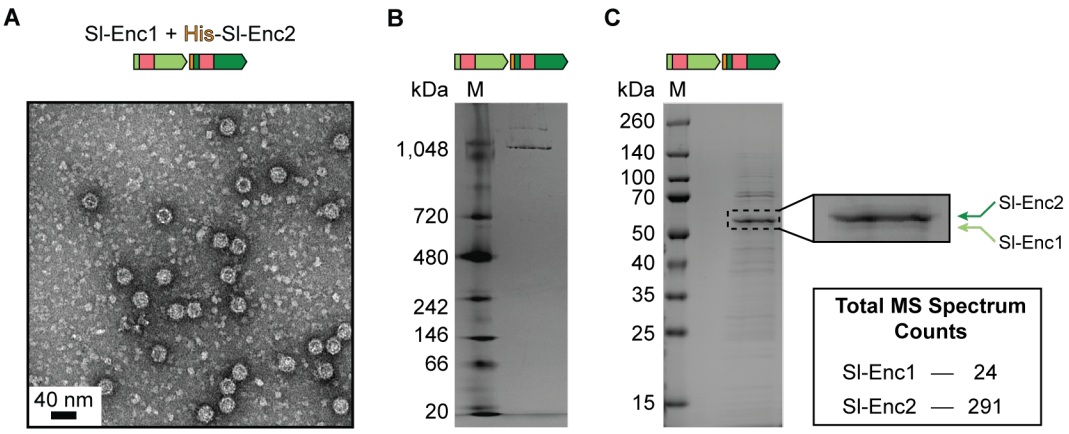
Primer	DNA Sequence
His-Sl-Enc1-iPCR-FW	ATGCACCACCACCATCACCATACAATCGTAGACGAGACGTTAAATGGG
pETDuet-1-iPCR-RV	TGTATATCTCCTTCTTATACTTAACTAATATACTAAGATGGGG
His-Sl-Enc2-iPCR-FW	ATGCACCACCACCATCACCATACTACGTCCGTGGATCCAACTTC
Double-Enc-His-Sl-Enc2-iPCR-FW	ACTACGTCCGTGGATCCAACTTC
Double-Enc-His-Sl-Enc2-iPCR-RV	ATGGTGATGGTGGTGGTGCATGTATATCTCCTTCTTATATCATAACCCG
Double-Enc-Strep-Sl-Enc2-iPCR-RV	CTTTTCGAACTGCGGGTGGCTCCACATGTATATCTCCTTCTTATATCATAACCC
His-Sl-Enc1-R201C-FW	TGTGCCGCCTTAGAAATCATCG
His-Sl-Enc1-R201C-iPCR-RV	GCGCGGAAGCTCGACTAC
His-Sl-Enc1-R201C-Gibson-Insert-RV	ACAGCGCGATAAAGTCAGTGC
His-Sl-Enc1-R201C-Gibson-Vector-FW	ACTGACTTTATCGCGCTGTGATTATGAACAAGTCCTGG
His-Sl-Enc1-R201C-Gibson-Vector-RV	GATTTCTAAGGCGGCACAGCGCGGAAGCTCGAC

Supplemental Table 2-3: Primer pairs used to construct the plasmids used in this study.

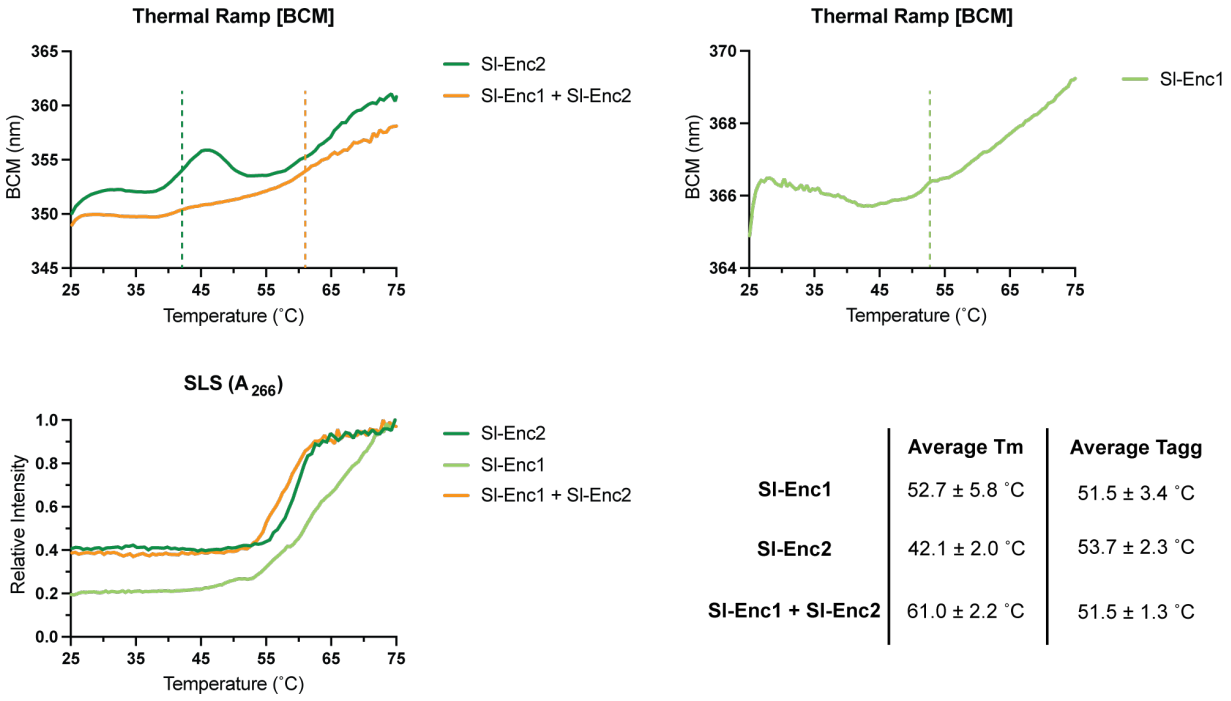
Plasmid	Cloning Method	Primers
His-Sl-Enc1	Inverse PCR	His-Sl-Enc1-iPCR-FW pETDuet-1-iPCR-RV
His-Sl-Enc2	Inverse PCR	His-Sl-Enc2-iPCR-FW pETDuet-1-iPCR-RV
Sl-Enc1 / His-Sl-Enc2	Inverse PCR	Double-Enc-His-Sl-Enc2-iPCR-FW Double-Enc-His-Sl-Enc2-iPCR-RV
His-Sl-Enc1 / Sl-Enc2	Inverse PCR	His-Sl-Enc1-iPCR-FW pETDuet-1-iPCR-RV
His-Sl-Enc1 / His-Sl-Enc2	Inverse PCR	His-Sl-Enc1-iPCR-FW pETDuet-1-iPCR-RV
His-Sl-Enc1 / Strep-Sl-Enc2	Inverse PCR	Double-Enc-His-Sl-Enc2-iPCR-FW Double-Enc-Strep-Sl-Enc2-iPCR-RV
His-Sl-Enc1(R201C) / Strep-Sl-Enc2	Inverse PCR	His-Sl-Enc1-R201C-FW His-Sl-Enc1-R201C-iPCR-RV
His-Sl-Enc1(R201C) / Strep-Sl-Enc2(G206C)	Gibson Assembly	<u>Insert</u> His-Sl-Enc1-R201C-FW His-Sl-Enc1-R201C-Gibson-Insert-RV
		<u>Vector</u> His-Sl-Enc1-R201C-Gibson-Vector-FW His-Sl-Enc1-R201C-Gibson-Vector-RV

Supplemental Table 2-4: Protein sequences used in this work

Protein	Protein Sequence
Sl-Enc1	MTIVDETLNGNPEDTAHSSLSTAAARNLATTTKTVPQMQGITSRWLLRLLPWVQVSGGTYRVNRRLSH TVGDGRIDFDISGSDVAIIPEELRELPALRDFTDTEVLAALGERFTQREYAPGELIAEAGRPGDRLVLIAH GRVDRIGTGKYGDTTVLAALAGGDHLGDAPLTSDEVTWEFSYRAVTRVTVVELPRRAALEIIERSPGLR EHLAGVRQGPVHPTNSSGESAVAVASGHRGEPSLPGAFADYDLAPREYELSVAQTVLRTHTRVGDLYN DPMNQVEEQLKLTVQALRERQEHEMINNREFGLLHNADLKQRIPTRSGPPTPDDLDDLLATVWKDPGF LLAHPRAIAAMAREWSARGLYPTAVDFHGHSLPSWRGVPIFPCNKIPVTKERTSSILLLRTGEEKQGVV GLHQTGIPDEYEPSLSVRFMGIDDRAVINYLVSAYYSAAVLVPDALGVLEDVEVGL*
His-Sl-Enc1	MHHHHHHTIVDETLNGNPEDTAHSSLSTAAARNLATTTKTVPQMQGITSRWLLRLLPWVQVSGGTYR VNRRLSHTVGDGRIDFDISGSDVAIIPEELRELPALRDFTDTEVLAALGERFTQREYAPGELIAEAGRPGD RLVLIAHGRVDRIGTGKYGDTTVLAALAGGDHLGDAPLTSDEVTWEFSYRAVTRVTVVELPRRAALEII ERSPGLREHLAGVRQGPVHPTNSSGESAVAVASGHRGEPSLPGAFADYDLAPREYELSVAQTVLRTHT RVGDLYNDPMNQVEEQLKLTVQALRERQEHEMINNREFGLLHNADLKQRIPTRSGPPTPDDLDDLLAT VWKDPGFLLAHPRAIAAMAREWSARGLYPTAVDFHGHSLPSWRGVPIFPCNKIPVTKERTSSILLLRTG EEKQGVVGLHQTGIPDEYEPSLSVRFMGIDDRAVINYLVSAYYSAAVLVPDALGVLEDVEVGL*
His-Sl-Enc1 (R201C)	MHHHHHHTIVDETLNGNPEDTAHSSLSTAAARNLATTTKTVPQMQGITSRWLLRLLPWVQVSGGTYR VNRRLSHTVGDGRIDFDISGSDVAIIPEELRELPALRDFTDTEVLAALGERFTQREYAPGELIAEAGRPGD RLVLIAHGRVDRIGTGKYGDTTVLAALAGGDHLGDAPLTSDEVTWEFSYRAVTRVTVVELPRCAALEII ERSPGLREHLAGVRQGPVHPTNSSGESAVAVASGHRGEPSLPGAFADYDLAPREYELSVAQTVLRTHT RVGDLYNDPMNQVEEQLKLTVQALRERQEHEMINNREFGLLHNADLKQRIPTRSGPPTPDDLDDLLAT VWKDPGFLLAHPRAIAAMAREWSARGLYPTAVDFHGHSLPSWRGVPIFPCNKIPVTKERTSSILLLRTG EEKQGVVGLHQTGIPDEYEPSLSVRFMGIDDRAVINYLVSAYYSAAVLVPDALGVLEDVEVGL*
Sl-Enc2	MTTSVDPTSGPQASGTEQNRSSLDTAAARKLATTTKTVPQMQGISSRWLLRVLPWTQVNGGTYRVNR RLTHTLGDGQVEFVNTGAEVRVIPEELRELAPLRGFTDTPTLEALAGRFAQHEFAPGDVLAQQDQPAD RIILIAHGKLDRLGTGKYGGETVQGQLAGGDHLGAAALLDGGAAWEHTVRAVTRVTALTLSRGDYEQ VLGGSESLRAHVEAFRAALVPAQNKHGEAAIEVAAGHVGEPSLPGTFADYDLAPREYELSVAQTVLKI HSRVADLYNDPMNQMDQQLRLTVEALRERQEHEMINNREFGLLHNADLKQRIHTRSGPPTPDDLDELI SRRRKTQVLLAHPRTIAAIGREWNARGIYPTGAELHGTDVRAWRGIPLLPCNKIPVTPEQTSSIIAMRLG EENQGVVGLHQTGIPDEYQPGLSVRFMGINDQAVIQYLVSAYYSAAVLVPDALGILEDVEIGH*
His-Sl-Enc2	MHHHHHHTTSVDPTSGPQASGTEQNRSSLDTAAARKLATTTKTVPQMQGISSRWLLRVLPWTQVNGG TYRVNRRLTHTLGDGQVEFVNTGAEVRVIPEELRELAPLRGFTDTPTLEALAGRFAQHEFAPGDVLAQ QDQPADRIILIAHGKLDRLGTGKYGGETVQGQLAGGDHLGAAALLDGGAAWEHTVRAVTRVTALTLS RGDYEQVLGGSESLRAHVEAFRAALVPAQNKHGEAAIEVAAGHVGEPSLPGTFADYDLAPREYELSVA QTVLKIHSRVADLYNDPMNQMDQQLRLTVEALRERQEHEMINNREFGLLHNADLKQRIHTRSGPPTPD DLDELISRRRKTQVLLAHPRTIAAIGREWNARGIYPTGAELHGTDVRAWRGIPLLPCNKIPVTPEQTSSII AMRLGEENQGVVGLHQTGIPDEYQPGLSVRFMGINDQAVIQYLVSAYYSAAVLVPDALGILEDVEIGH *
Strep-Sl-Enc2	MWSHPQFEKTTSVDPTSGPQASGTEQNRSSLDTAAARKLATTTKTVPQMQGISSRWLLRVLPWTQVN GGTYRVNRRLTHTLGDGQVEFVNTGAEVRVIPEELRELAPLRGFTDTPTLEALAGRFAQHEFAPGDVL AQQDQPADRIILIAHGKLDRLGTGKYGGETVQGQLAGGDHLGAAALLDGGAAWEHTVRAVTRVTALT LSRGDYEQVLGGSESLRAHVEAFRAALVPAQNKHGEAAIEVAAGHVGEPSLPGTFADYDLAPREYELS VAQTVLKIHSRVADLYNDPMNQMDQQLRLTVEALRERQEHEMINNREFGLLHNADLKQRIHTRSGPPT PDDLDELISRRRKTQVLLAHPRTIAAIGREWNARGIYPTGAELHGTDVRAWRGIPLLPCNKIPVTPEQTS SIIAMRLGEENQGVVGLHQTGIPDEYQPGLSVRFMGINDQAVIQYLVSAYYSAAVLVPDALGILEDVEI GH*
Strep-Sl-Enc2 (G206C)	MWSHPQFEKTTSVDPTSGPQASGTEQNRSSLDTAAARKLATTTKTVPQMQGISSRWLLRVLPWTQVN GGTYRVNRRLTHTLGDGQVEFVNTGAEVRVIPEELRELAPLRGFTDTPTLEALAGRFAQHEFAPGDVL AQQDQPADRIILIAHGKLDRLGTGKYGGETVQGQLAGGDHLGAAALLDGGAAWEHTVRAVTRVTALT LSRCDYEQVLGGSESLRAHVEAFRAALVPAQNKHGEAAIEVAAGHVGEPSLPGTFADYDLAPREYELS VAQTVLKIHSRVADLYNDPMNQMDQQLRLTVEALRERQEHEMINNREFGLLHNADLKQRIHTRSGPPT PDDLDELISRRRKTQVLLAHPRTIAAIGREWNARGIYPTGAELHGTDVRAWRGIPLLPCNKIPVTPEQTS SIIAMRLGEENQGVVGLHQTGIPDEYQPGLSVRFMGINDQAVIQYLVSAYYSAAVLVPDALGILEDVEI GH*

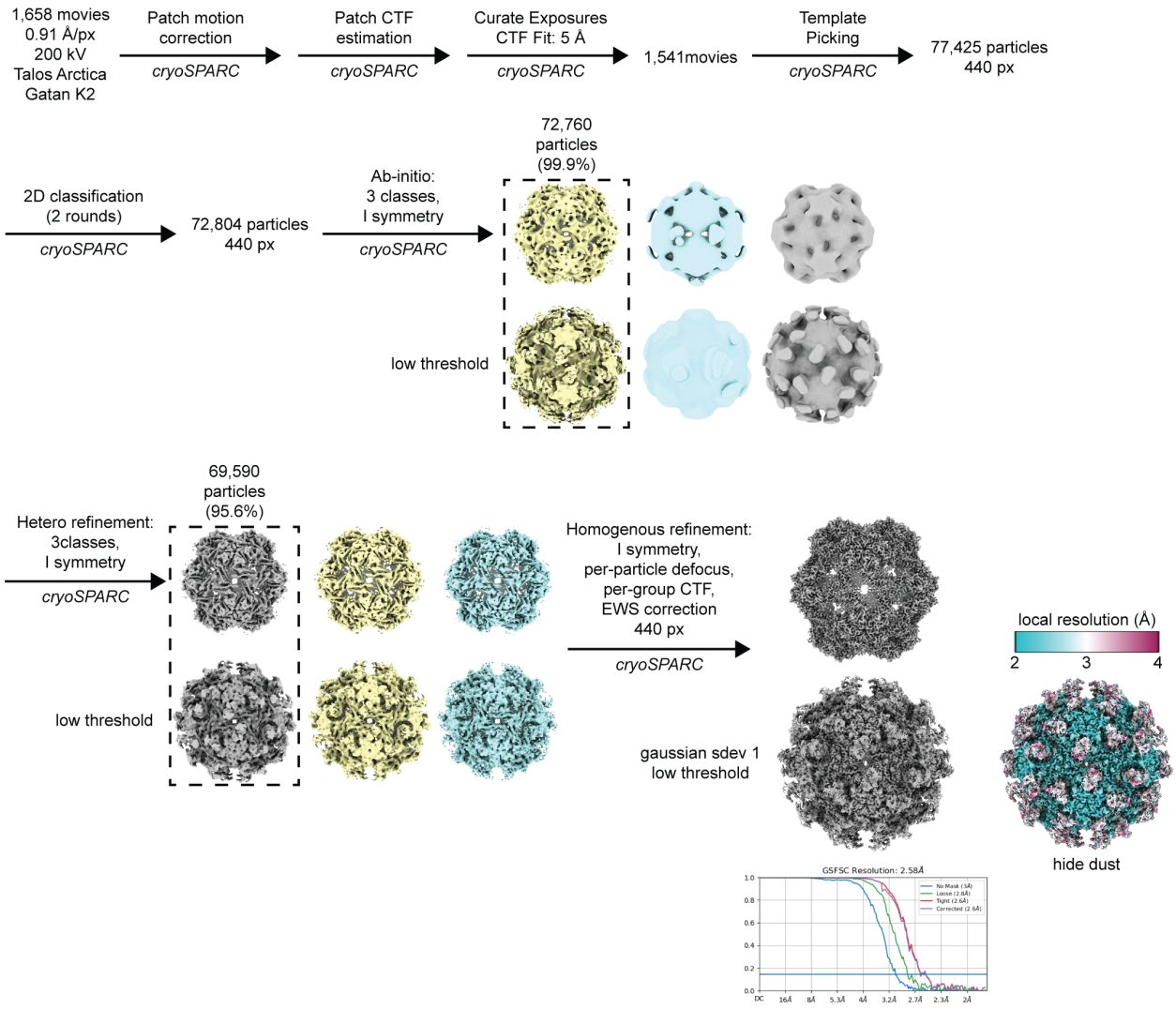


Supplemental Figure 2-1: Analysis of Sl-Enc2 majority shells. (**A**) Negative stain TEM micrograph of an Sl-Enc1 + His-Sl-Enc2 mixed shell sample purified via affinity purification and size exclusion chromatography. (**B**) Native PAGE analysis of an Sl-Enc1 majority mixed sample. M: molecular weight marker. (**C**) SDS-PAGE of purified Sl-Enc1 + His-Sl-Enc2. Outlined region on SDS-PAGE gel was excised and prepared for mass spectrometric analysis for protein identification. Total spectrum counts of Sl-Enc1 and Sl-Enc2 from mass spectrometry analysis are shown.

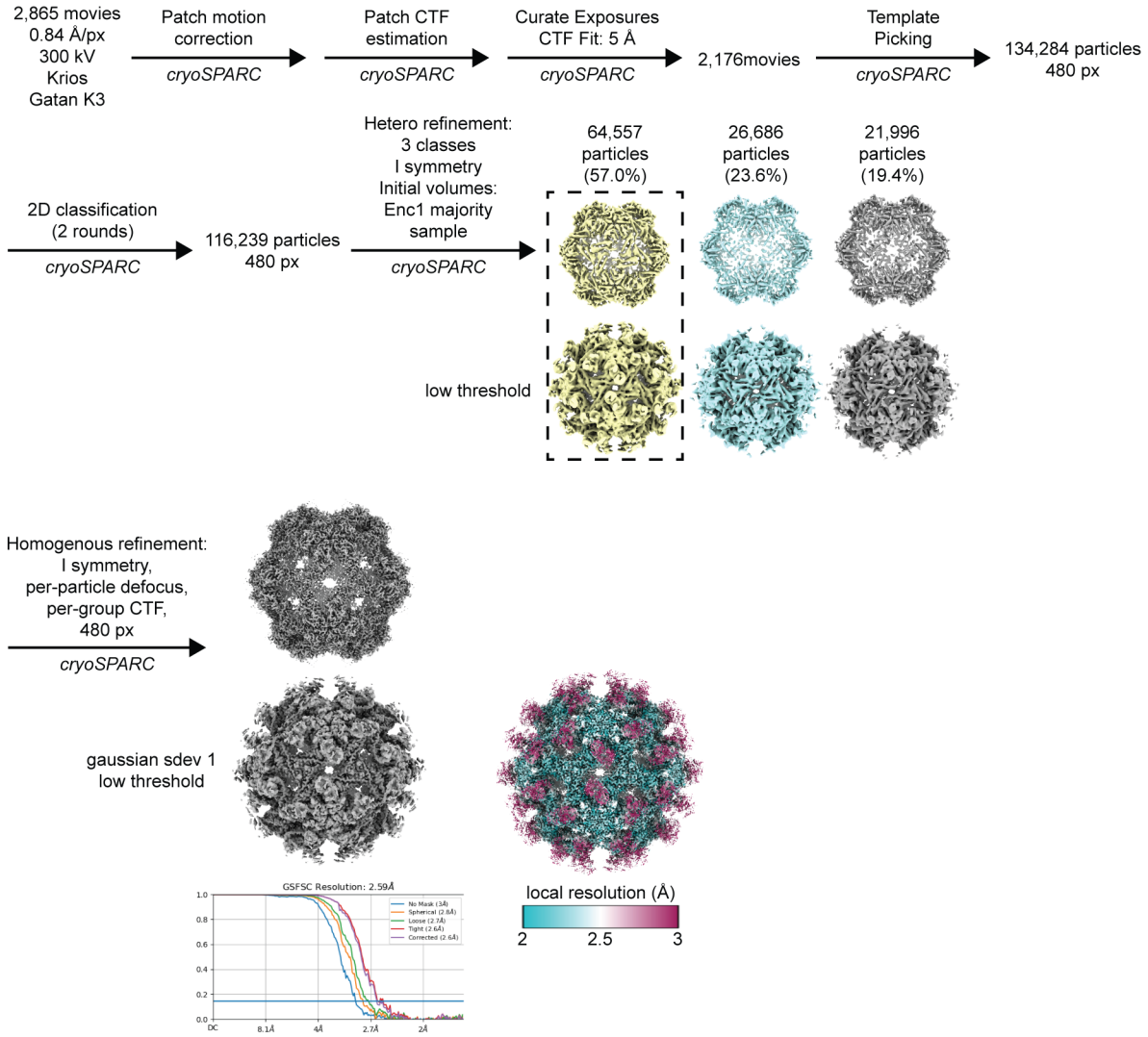


	Average Tm	Average Tagg
Sl-Enc1	52.7 ± 5.8 ˚C	51.5 ± 3.4 ˚C
Sl-Enc2	42.1 ± 2.0 ˚C	53.7 ± 2.3 ˚C
Sl-Enc1 + Sl-Enc2	61.0 ± 2.2 ˚C	51.5 ± 1.3 ˚C

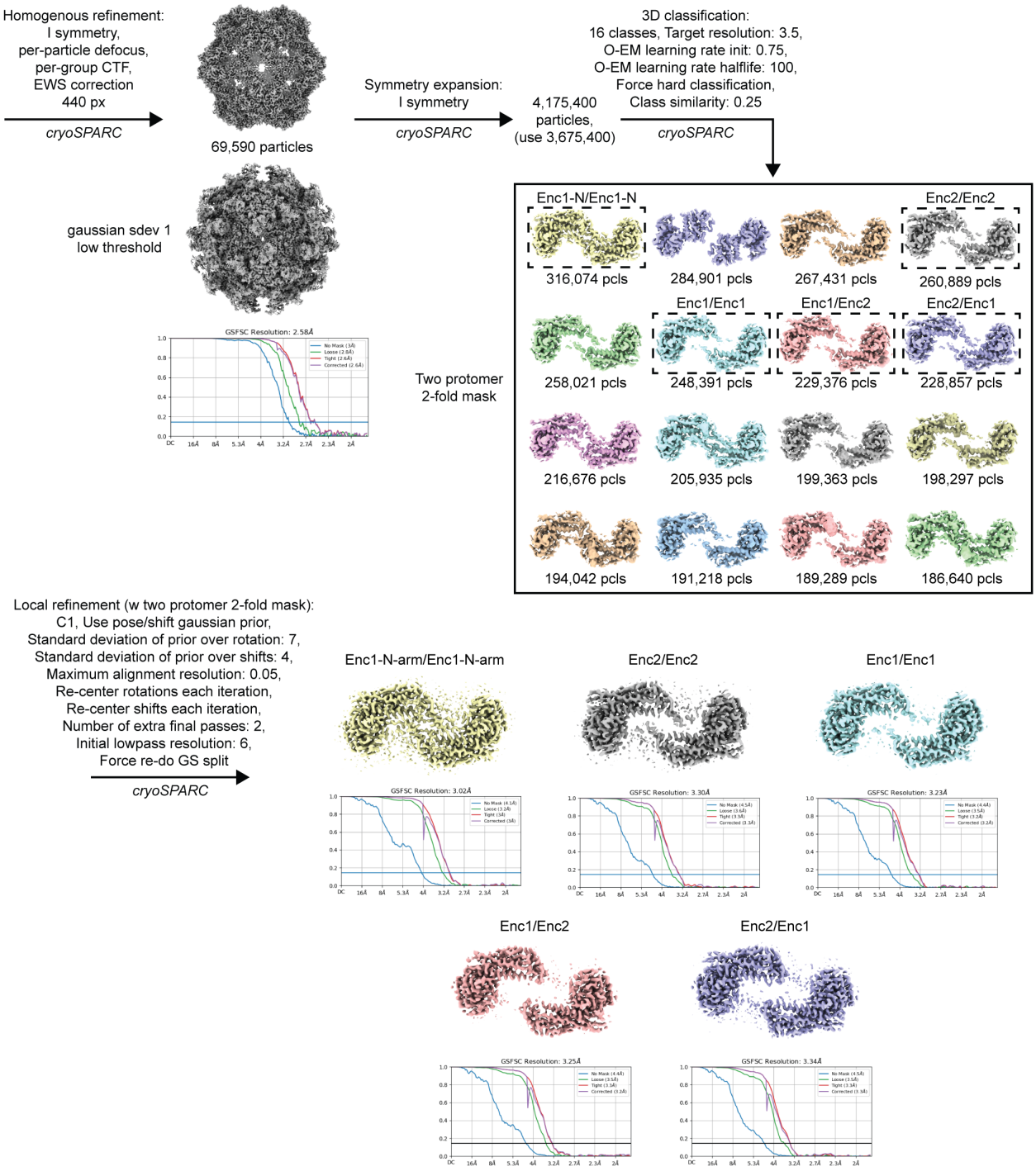
Supplemental Figure 2-2: Thermal ramp data for Sl-Enc1 and Sl-Enc2. (**A**) Thermal ramp data for Sl-Enc1, Sl-Enc2, Sl-Enc1 + Sl-Enc2. (B) SLS results for SL-Enc1, Sl-Enc1, and Sl-Enc1 + Sl-Enc2. (**C**) Average T_m and T_{agg} for Sl-Enc1, Sl-Enc2, and Sl-Enc1 + Sl-Enc2.



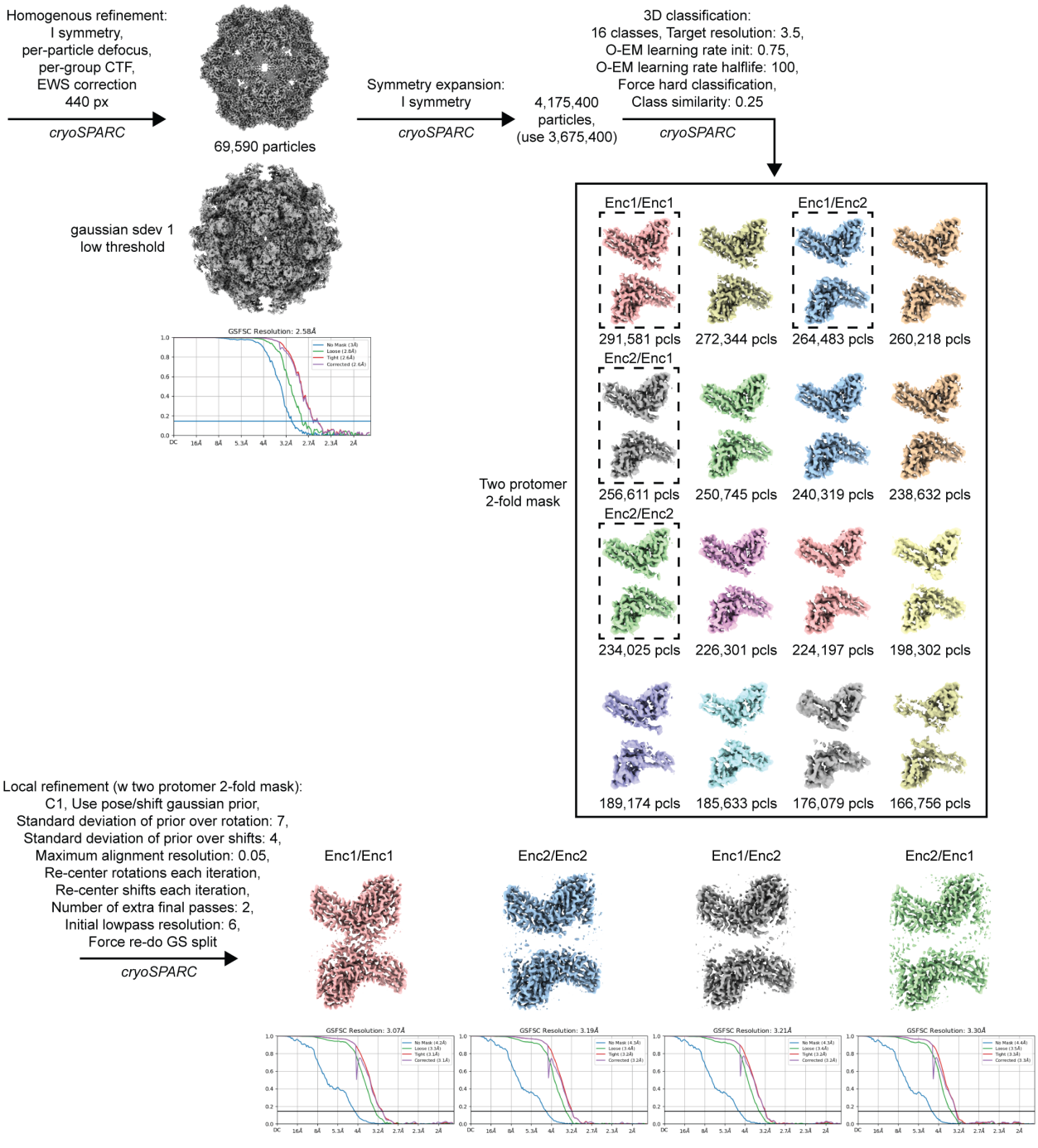
Supplemental Figure 2-3: Majority Sl-Enc1 cryo-EM data processing workflow. Cryo-EM workflow showing gold-standard Fourier shell correlation (FSC) curves for the final symmetry- averaged (I) consensus map as well as a local resolution analysis.



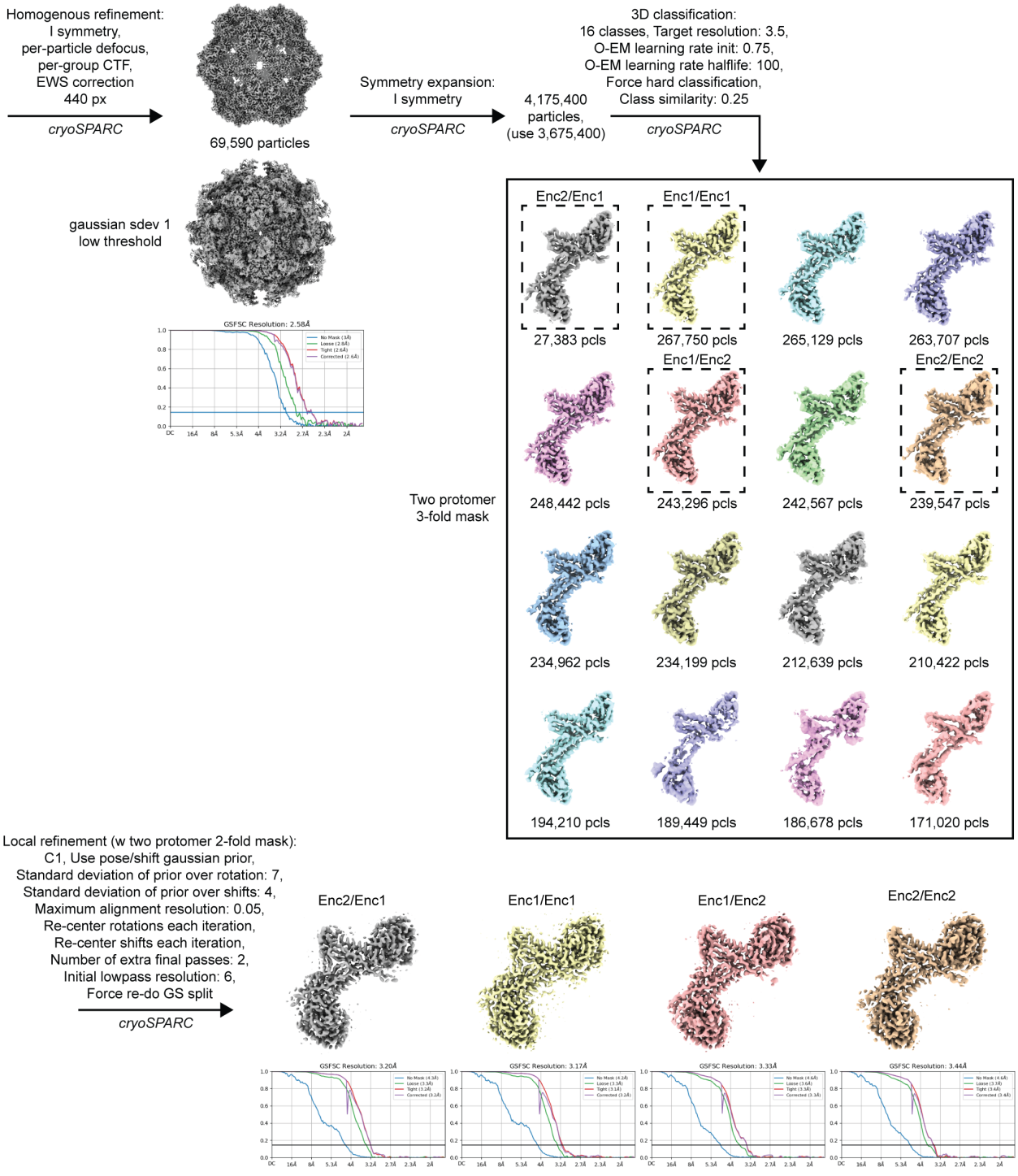
Supplemental Figure 2-4: Majority Sl-Enc2 cryo-EM data processing workflow. Cryo-EM workflow showing gold-standard Fourier shell correlation (FSC) curves for the final symmetry- averaged (I) consensus map as well as a local resolution analysis.



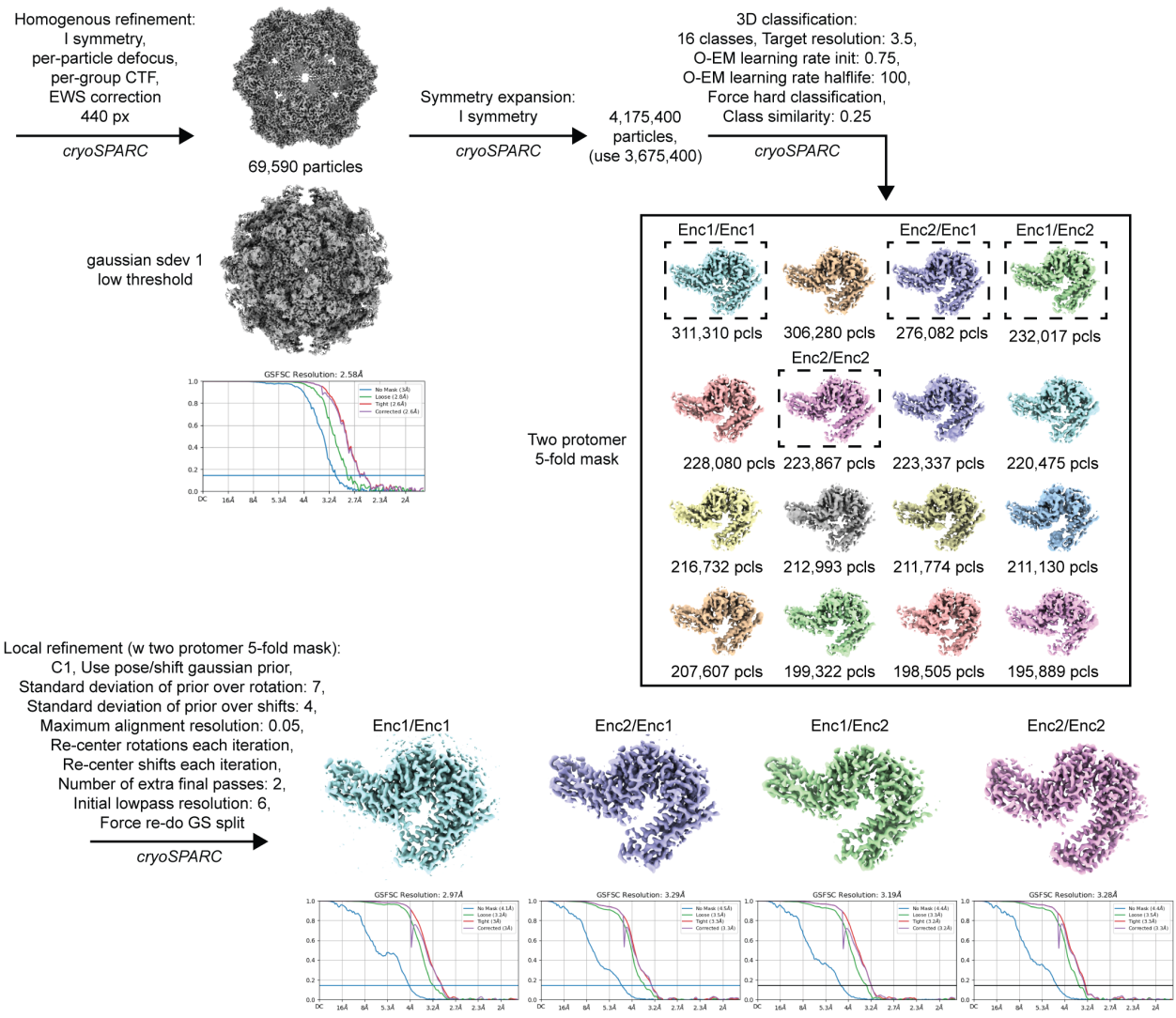
Supplemental Figure 2-5: Cryo-EM 3D classification analysis workflow for the 2-fold (pore) interaction.



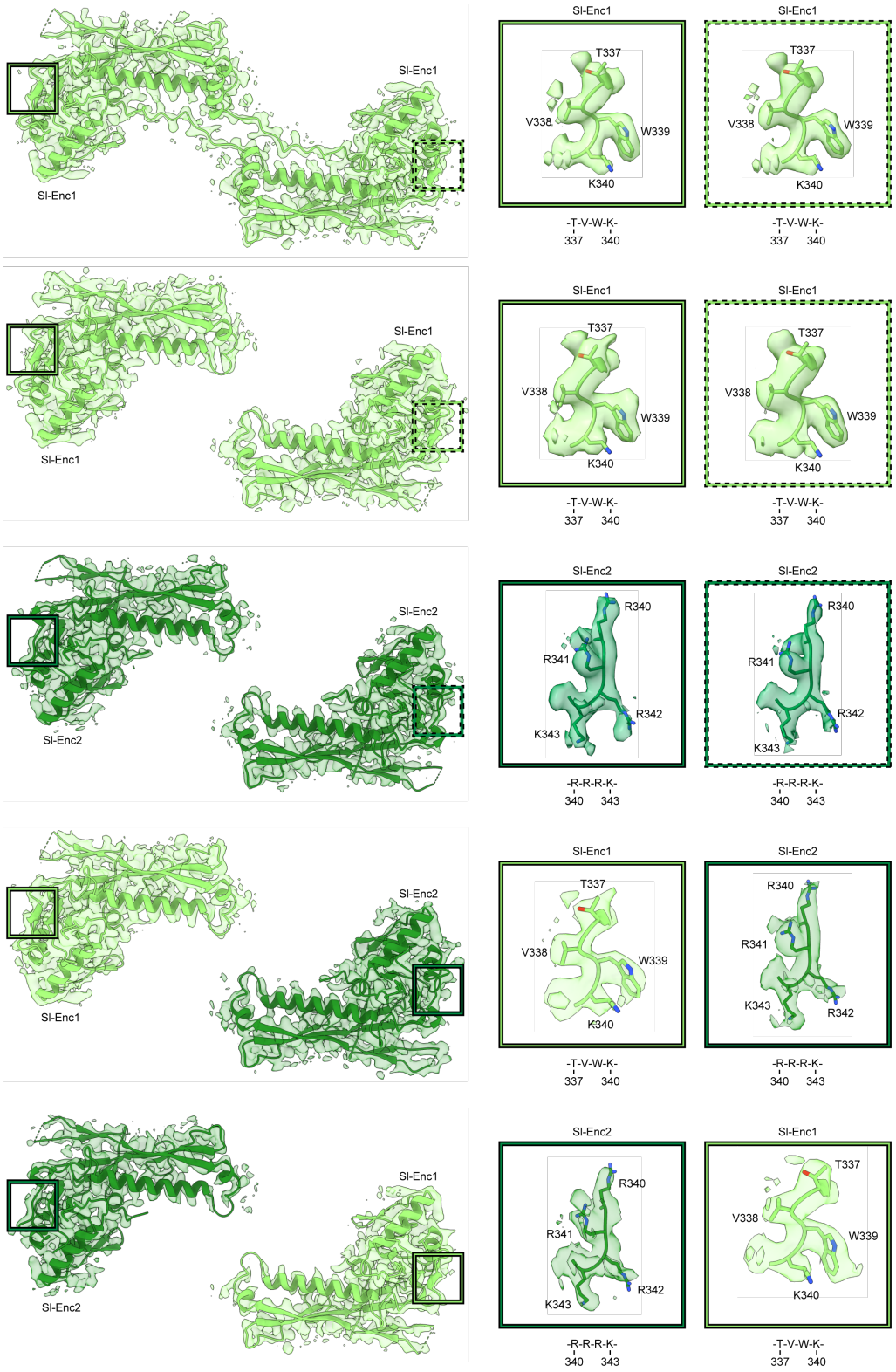
Supplemental Figure 2-6: Cryo-EM 3D classification analysis workflow for the 2-fold (P-domain) interaction.



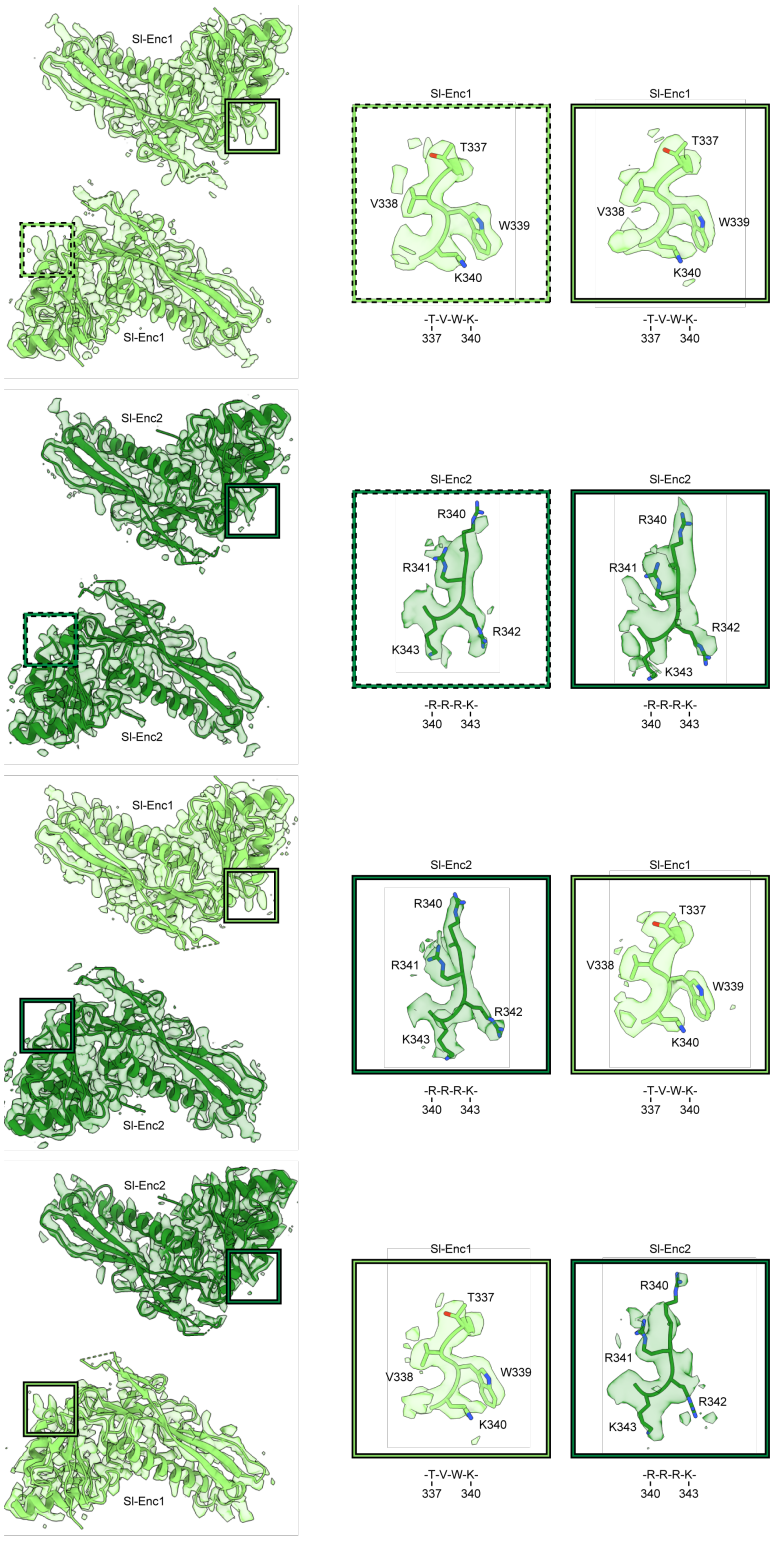
Supplemental Figure 2-7: Cryo-EM 3D classification analysis workflow for the 3-fold interaction.



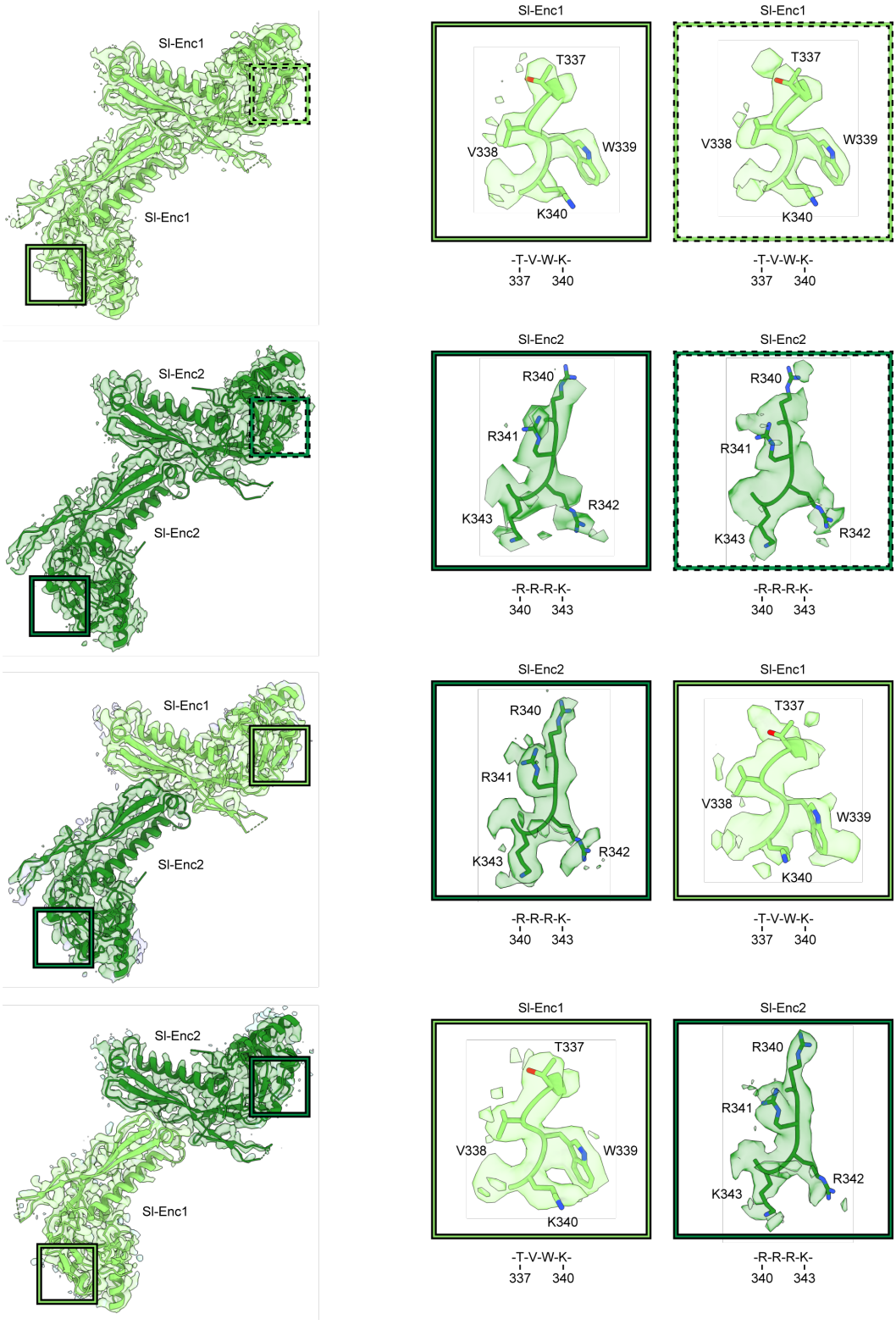
Supplemental Figure 2-8: Cryo-EM 3D classification analysis workflow for the 5-fold interaction.



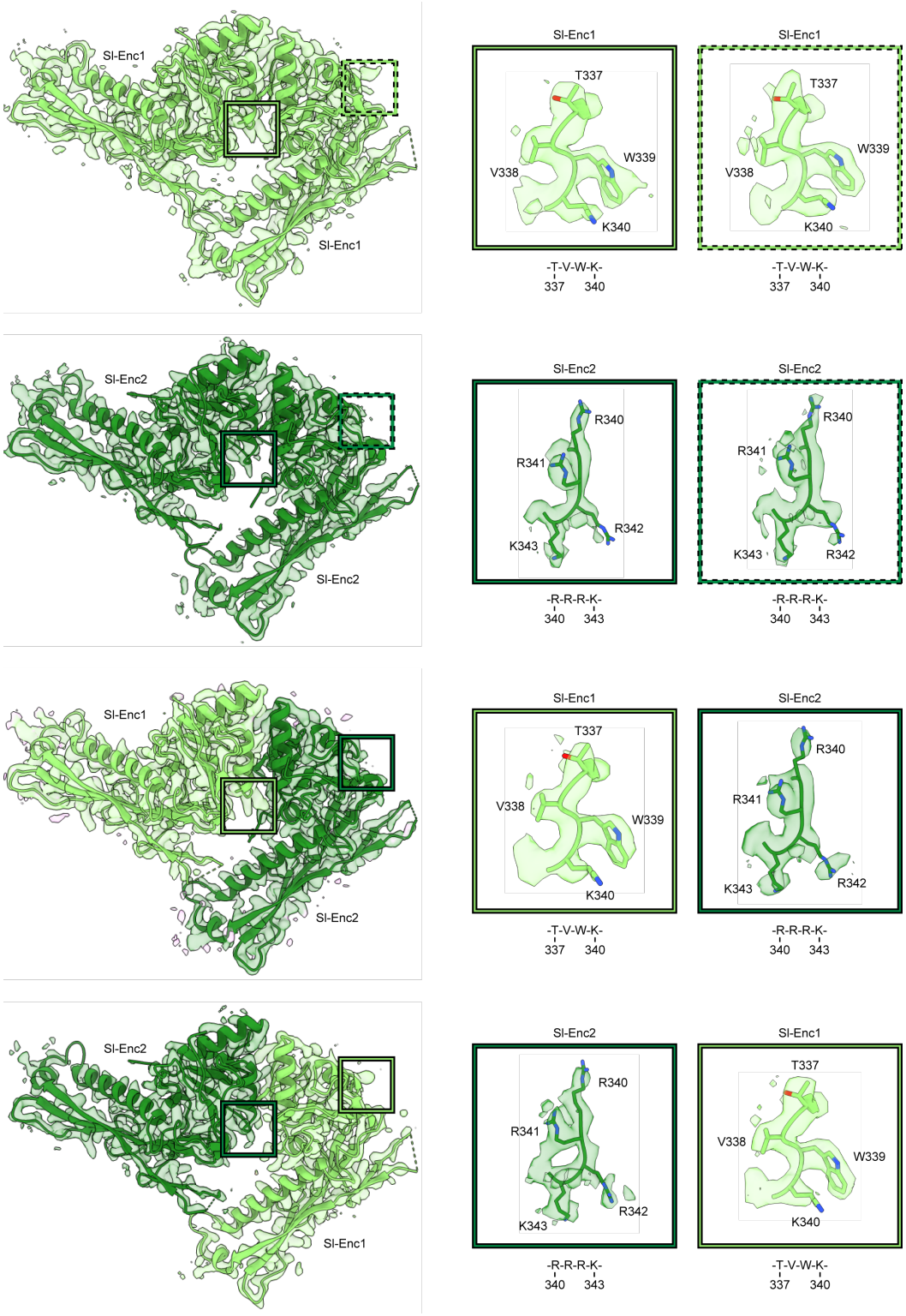
Supplemental Figure 2-9: Identified high quality classes for the 2-fold (pore) 3D classification analysis. Left: Cryo-EM densities and atomic models for the five identified classes representing all possible interaction permutations. The class shown on top represents the Sl-Enc1/Sl-Enc1 interaction with resolved N-arm. Right: Diagnostic regions located at the edge of the A-domain used to distinguish between Sl-Enc1 and Sl-Enc2.



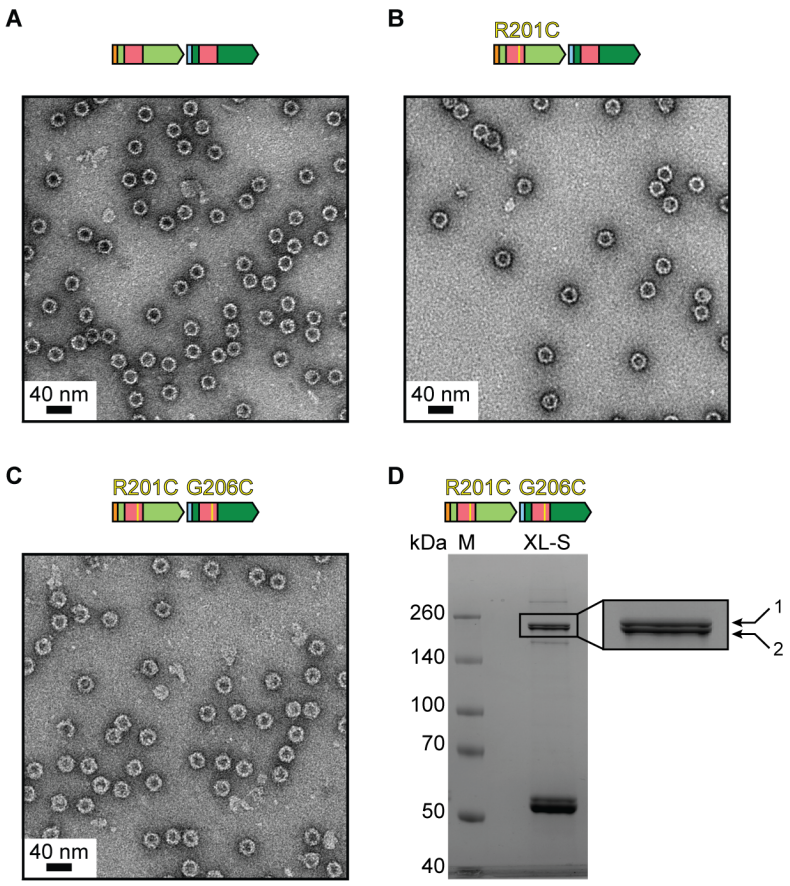
Supplemental Figure 2-10: Identified high quality classes for the 2-fold (P-domain) 3D classification analysis. Left: Cryo-EM densities and atomic models for the four identified classes representing all possible interaction permutations. Right: Diagnostic regions located at the edge of the A-domain used to distinguish between Sl-Enc1 and Sl-Enc2.



Supplemental Figure 2-11: Identified high quality classes for the 3-fold 3D classification analysis. Left: Cryo-EM densities and atomic models for the four identified classes representing all possible interaction permutations. Right: Diagnostic regions located at the edge of the A- domain used to distinguish between Sl-Enc1 and Sl-Enc2.



Supplemental Figure 2-12: Identified high quality classes for the 5-fold 3D classification analysis. Left: Cryo-EM densities and atomic models for the four identified classes representing all possible interaction permutations. Right: Diagnostic regions located at the edge of the A- domain used to distinguish between Sl-Enc1 and Sl-Enc2.



Supplemental Figure 2-13: Cysteine mutant shell assembly and crosslinking mass spectrometry. (**A**) Negative stain TEM micrograph of a mixed shell sample containing His-Sl-Enc1 and Strep-Sl-Enc2. (**B**) Negative stain TEM micrograph of a mixed shell sample containing mutant His-Sl-Enc1(R195C) and Strep-Sl-Enc2. (**C**) Negative stain TEM micrograph of a mixed shell sample containing mutant His-Sl-Enc1(R195C) and mutant Strep-Sl-Enc2 (G198C). (**D**) SDS-PAGE analysis of crosslinked double-cysteine mutant sample used for mass spectrometric analysis. M: molecular weight marker. XL-S: Crosslinked His-Sl-Enc1(R195C) + Strep-Sl-Enc2 (G198C) shells.

2.6 References

1. C. A. Dutcher, M. P. Andreas, T. W. Giessen, A two-component quasi-icosahedral protein nanocompartment with variable shell composition and irregular tiling. *bioRxiv* (2024).

2. Y. Liu, T. A. Demina, S. Roux, P. Aiewsakun, D. Kazlauskas, P. Simmonds, D. Prangishvili, H. M. Oksanen, M. Krupovic, Diversity, taxonomy, and evolution of archaeal viruses of the class Caudoviricetes. *PLoS Biol*. **19,** e3001442 (2021).

3. W. R. Wikoff, L. Liljas, R. L. Duda, H. Tsuruta, R. W. Hendrix, J. E. Johnson, Topologically linked protein rings in the bacteriophage HK97 capsid. *Science*. **289,** 2129-2133 (2000).

4. M. K. Pietilä, P. Laurinmäki, D. A. Russell, C.-C. Ko, D. Jacobs-Sera, R. W. Hendrix, D. H. Bamford, S. J. Butcher, M. K. Pietilä, P. Laurinmäki, D. A. Russell, C.-C. Ko, D. Jacobs-Sera, R. W. Hendrix, D. H. Bamford, S. J. Butcher, Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. *Proceedings of the National Academy of Sciences*. **110,** (2013-6-25).

5. M. L. Baker, W. Jiang, F. J. Rixon, W. Chiu, Common ancestry of herpesviruses and tailed DNA bacteriophages. *J Virol*. **79,** 14967-14970 (2005).

6. T. W. Giessen, Encapsulins. *Annu Rev Biochem*. **91,** 353-380 (2022).

7. M. P. Andreas, T. W. Giessen, Large-scale computational discovery and analysis of virus-derived microbial nanocompartments. *Nat Commun*. **12,** 4748 (2021).

8. M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology 2008 15:9*. **15,** (2008-08-31).

9. M. M. Suhanovsky, C. M. Teschke, Nature's favorite building block: Deciphering folding and capsid assembly of proteins with the HK97-fold. *Virology*. **479-480,** 487-497 (2015).

10. Z. H. Zhou, W. H. Hui, S. Shah, J. Jih, C. M. O'connor, M. B. Sherman, D. H. Kedes, S. Schein, Four levels of hierarchical organization, including noncovalent chainmail, brace the mature tumor herpesvirus capsid against pressurization. *Structure*. **22,** 1385-1398 (2014).

11. J. A. Jones, M. P. Andreas, T. W. Giessen, Exploring the extreme acid tolerance of a dynamic protein nanocage. *Biomacromolecules*. **24,** 1388-1399 (2023).

12. J. Ross, Z. Mciver, T. Lambert, C. Piergentili, J. E. Bird, K. J. Gallagher, F. L. Cruickshank, P. James, E. Zarazúa-Arvizu, L. E. Horsfall, K. J. Waldron, M. D. Wilson, C. L. Mackay, A. Baslé, D. J. Clarke, J. Marles-Wright, Pore dynamics and asymmetric cargo loading in an encapsulin nanocompartment. *Sci Adv*. **8,** eabj4461 (2022).

13. E. Eren, B. Wang, D. C. Winkler, N. R. Watts, A. C. Steven, P. T. Wingfield, Structural characterization of the Myxococcus xanthus encapsulin and ferritin-like cargo system gives insight into its iron storage mechanism. *Structure*. **30,** 551-563.e554 (2022).

14. J. Hua, A. Huet, C. A. Lopez, K. Toropova, W. H. Pope, R. L. Duda, R. W. Hendrix, J. F. Conway, Capsids and genomes of jumbo-sized bacteriophages reveal the evolutionary reach of the HK97 fold. *mBio*. **8,** (2017).

15. D. L. Caspar, A. Klug, Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol*. **27,** 1-24 (1962).

16. J. E. Johnson, J. A. Speir, Quasi-equivalent viruses: a paradigm for protein assemblies. *J Mol Biol*. **269,** 665-675 (1997).

17. A. Fokine, P. G. Leiman, M. M. Shneider, B. Ahvazi, K. M. Boeshans, A. C. Steven, L. W. Black, V. V. Mesyanzhinov, M. G. Rossmann, Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc Natl Acad Sci U S A*. **102,** 7163-7168 (2005).

18. M. L. Yap, M. G. Rossmann, Structure and function of bacteriophage T4. *Future Microbiol*. **9,** 1319-1327 (2014).

19. F. Akita, K. T. Chong, H. Tanaka, E. Yamashita, N. Miyazaki, Y. Nakaishi, M. Suzuki, K. Namba, Y. Ono, T. Tsukihara, A. Nakagawa, The crystal structure of a virus-like particle from the hyperthermophilic archaeon Pyrococcus furiosus provides insight into the evolution of viruses. *J Mol Biol*. **368,** 1469-1483 (2007).

20. T. W. Giessen, B. J. Orlando, A. A. Verdegaal, M. G. Chambers, J. Gardener, D. C. Bell, G. Birrane, M. Liao, P. A. Silver, Large protein organelles form a new iron sequestration system with high storage capacity. *Elife*. **8,** (2019).

21. J. A. Jones, T. W. Giessen, Advances in encapsulin nanocompartment biology and engineering. *Biotechnol Bioeng*. **118,** 491-505 (2021).

22. R. J. Nichols, C. Cassidy-Amstutz, T. Chaijarasphong, D. F. Savage, Encapsulins: molecular biology of the shell. *Crit Rev Biochem Mol Biol*. **52,** 583-594 (2017).

23. J. A. Jones, R. Benisch, T. W. Giessen, Encapsulin cargo loading: progress and potential. *J Mater Chem B*. **11,** 4377-4388 (2023).

24. W. J. Altenburg, N. Rollins, P. A. Silver, T. W. Giessen, Exploring targeting peptide-shell interactions in encapsulin nanocompartments. *Sci Rep*. **11,** 4951 (2021).

25. J. A. Jones, M. P. Andreas, T. W. Giessen, Structural basis for peroxidase encapsulation inside the encapsulin from the Gram-negative pathogen Klebsiella pneumoniae. *Nat Commun*. **15,** 2558 (2024).

26. S. Kwon, M. P. Andreas, T. W. Giessen, Structure and heterogeneity of a highly cargo-loaded encapsulin shell. *J Struct Biol*. **215,** 108022 (2023).

27. C. A. Mchugh, J. Fontana, D. Nemecek, N. Cheng, A. A. Aksyuk, J. B. Heymann, D. C. Winkler, A. S. Lam, J. S. Wall, A. C. Steven, E. Hoiczyk, A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *EMBO J*. **33,** 1896-1911 (2014).

28. D. He, S. Hughes, S. Vanden-Hehir, A. Georgiev, K. Altenbach, E. Tarrant, C. L. Mackay, K. J. Waldron, D. J. Clarke, J. Marles-Wright, Structural characterization of encapsulated ferritin provides insight into iron storage in bacterial nanocompartments. *Elife*. **5,** (2016).

29. D. He, C. Piergentili, J. Ross, E. Tarrant, L. R. Tuck, C. L. Mackay, Z. Mciver, K. J. Waldron, D. J. Clarke, J. Marles-Wright, Conservation of the structural and functional architecture of encapsulated ferritins in bacteria and archaea. *Biochem J*. **476,** 975-989 (2019).

30. K. A. Lien, K. Dinshaw, R. J. Nichols, C. Cassidy-Amstutz, M. Knight, R. Singh, L. D. Eltis, D. F. Savage, S. A. Stanley, A nanocompartment system contributes to defense against oxidative stress in. *Elife*. **10,** (2021).

31. R. Rahmanpour, T. D. Bugg, Assembly in vitro of Rhodococcus jostii RHA1 encapsulin and peroxidase DypB to form a nanocompartment. *FEBS J*. **280,** 2097-2104 (2013).

32. Y. Tang, A. Mu, Y. Zhang, S. Zhou, W. Wang, Y. Lai, X. Zhou, F. Liu, X. Yang, H. Gong, Cryo-EM structure of Mycobacterium smegmatis DyP-loaded encapsulin. *Proceedings of the National Academy of Sciences*. **118,** e2025658118 (2021).

33. R. Benisch, M. P. Andreas, T. W. Giessen, A widespread bacterial protein compartment sequesters and stores elemental sulfur. *Science Advances* (2024-02-02).

34. R. J. Nichols, B. Lafrance, N. R. Phillips, D. R. Radford, L. M. Oltrogge, L. E. Valentin-Alvarado, A. J. Bischoff, E. Nogales, D. F. Savage, Discovery and characterization of a novel family of prokaryotic nanocompartments involved in sulfur metabolism. *Elife*. **10,** (2021).

35. M. P. Andreas, T. W. Giessen, Heterologous expression and purification of encapsulins in *Streptomyces coelicolor*. *MethodsX*. **9,** 101787 (2022).

36. M. P. Andreas, T. W. Giessen, The biosynthesis of the odorant 2-methylisoborneol is compartmentalized inside a protein shell. *bioRxiv* 2024.2004. 2023.590730 (2024).

37. T. W. Giessen, P. A. Silver, Encapsulation as a strategy for the design of biological compartmentalization. *J Mol Biol*. **428,** 916-927 (2016).

38. L. S. R. Adamson, N. Tasneem, M. P. Andreas, W. Close, E. N. Jenner, T. N. Szyszka, R. Young, L. C. Cheah, A. Norman, H. I. Macdermott-Opeskin, M. L. O'mara, F. Sainsbury, T. W. Giessen, Y. H. Lau, Pore structure controls stability and molecular flux in engineered protein cages. *Sci Adv*. **8,** eabl7346 (2022).

39. N. Oberg, R. Zallot, J. A. Gerlt, EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *J Mol Biol*. **435,** 168018 (2023).

40. R. Zallot, N. Oberg, J. A. Gerlt, The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*. **58,** 4169-4182 (2019).

41. K. Katoh, J. Rozewicki, K. D. Yamada, MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. **20,** 1160-1166 (2019).

42. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. **49,** W293-W296 (2021).

43. C. Suloway, J. Pulokas, D. Fellmann, A. Cheng, F. Guerra, J. Quispe, S. Stagg, C. S. Potter, B. Carragher, Automated molecular microscopy: the new Leginon system. *J Struct Biol*. **151,** 41-60 (2005).

44. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods*. **14,** 290-296 (2017).

45. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*. **25,** 1605-1612 (2004).

46. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*. **60,** 2126-2132 (2004).

47. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*. **66,** 486-501 (2010).

48. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. Mccoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*. **66,** 213-221 (2010).

49. P. V. Afonine, B. P. Klaholz, N. W. Moriarty, B. K. Poon, O. V. Sobolev, T. C. Terwilliger, P. D. Adams, A. Urzhumtsev, New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr D Struct Biol*. **74,** 814-840 (2018).

50. D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. Mccoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, P. D. Adams, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol*. **75,** 861-877 (2019).

51. A. Huet, A. M. Makhov, J. B. Huffman, M. Vos, F. L. Homa, J. F. Conway, Extensive subunit contacts underpin herpesvirus capsid stability and interior-to-exterior allostery. *Nat Struct Mol Biol*. **23,** 531-539 (2016).

52. A. Huet, B. Oh, J. Maurer, R. L. Duda, J. F. Conway, A symmetry mismatch unraveled: How phage HK97 scaffold flexibly accommodates a 12-fold pore at a 5-fold viral capsid vertex. *Sci Adv*. **9,** eadg8868 (2023).

53. K. N. Parent, R. Khayat, L. H. Tu, M. M. Suhanovsky, J. R. Cortines, C. M. Teschke, J. E. Johnson, T. S. Baker, P22 coat protein structures reveal a novel mechanism for capsid maturation: stability without auxiliary proteins or chemical crosslinks. *Structure*. **18,** 390-401 (2010).

54. C. A. Kerfeld, C. Aussignargues, J. Zarzycki, F. Cai, M. Sutter, Bacterial microcompartments. *Nat Rev Microbiol*. **16,** 277-290 (2018).

55. C. A. Kerfeld, M. R. Sawaya, S. Tanaka, C. V. Nguyen, M. Phillips, M. Beeby, T. O. Yeates, Protein structures forming the shell of primitive bacterial organelles. *Science*. **309,** 936-938 (2005).

56. M. G. Klein, P. Zwart, S. C. Bagby, F. Cai, S. W. Chisholm, S. Heinhorst, G. C. Cannon, C. A. Kerfeld, Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. *J Mol Biol*. **392,** 319-333 (2009).

57. M. Sutter, B. Greber, C. Aussignargues, C. A. Kerfeld, Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell. *Science*. **356,** 1293-1297 (2017).

58. S. Tanaka, C. A. Kerfeld, M. R. Sawaya, F. Cai, S. Heinhorst, G. C. Cannon, T. O. Yeates, Atomic-level models of the bacterial carboxysome shell. *Science*. **319,** 1083-1086 (2008).

59. M. Sutter, M. R. Melnicki, F. Schulz, T. Woyke, C. A. Kerfeld, A catalog of the diversity and ubiquity of bacterial microcompartments. *Nat Commun*. **12,** 3809 (2021).

60. Y. Sun, V. M. Harman, J. R. Johnson, P. J. Brownridge, T. Chen, G. F. Dykes, Y. Lin, R. J. Beynon, L. N. Liu, Decoding the absolute stoichiometric composition and structural plasticity of α-carboxysomes. *mBio*. **13,** e0362921 (2022).

61. C. Fan, S. Cheng, S. Sinha, T. A. Bobik, Interactions between the termini of lumen enzymes and shell proteins mediate enzyme encapsulation into bacterial microcompartments. *Proc Natl Acad Sci U S A*. **109,** 14995-15000 (2012).

62. J. N. Kinney, A. Salmeen, F. Cai, C. A. Kerfeld, Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly. *J Biol Chem*. **287,** 17729-17736 (2012).

63. M. Sommer, M. Sutter, S. Gupta, H. Kirst, A. Turmo, S. Lechno-Yossef, R. L. Burton, C. Saechao, N. B. Sloan, X. Cheng, L. G. Chan, C. J. Petzold, M. Fuentes-Cabrera, C. Y.

Ralston, C. A. Kerfeld, Heterohexamers Formed by CcmK3 and CcmK4 Increase the Complexity of Beta Carboxysome Shells. *Plant Physiol*. **179,** 156-167 (2019).

64. L. F. Garcia-Alles, K. Root, L. Maveyraud, N. Aubry, E. Lesniewska, L. Mourey, R. Zenobi, G. Truan, Occurrence and stability of hetero-hexamer associations formed by β-carboxysome CcmK shell components. *PLoS One*. **14,** e0223877 (2019).

65. L. Onorato, B. Stirmer, M. K. Showe, Isolation and characterization of bacteriophage T4 mutant preheads. *J Virol*. **27,** 409-426 (1978).

66. Q. Fang, W. C. Tang, A. Fokine, M. Mahalingam, Q. Shao, M. G. Rossmann, V. B. Rao, Structures of a large prolate virus capsid in unexpanded and expanded states generate insights into the icosahedral virus assembly. *Proc Natl Acad Sci U.S.A.* **119,** e2203272119 (2022).

67. Z. Chen, L. Sun, Z. Zhang, A. Fokine, V. Padilla-Sanchez, D. Hanein, W. Jiang, M. G. Rossmann, V. B. Rao, Cryo-EM structure of the bacteriophage T4 isometric head at 3.3-Å resolution and its relevance to the assembly of icosahedral viruses. *Proc Natl Acad Sci U S A*. **114,** E8184-E8193 (2017).

68. K. Johnson, B. Condie, D. T. Mooney, A. H. Doermann, Mutations that eliminate the requirement for the vertex protein in bacteriophage T4 capsid assembly. *J Mol Biol*. **224,** 601-611 (1992).

Chapter 3 Encapsulation of Geosmin Synthase Within a Two-Component Protein Shell in *Myxococcus xanthus*

Author contributions: C.A.D. designed and constructed all plasmids used in this study and subsequently expressed and purified all protein samples. C.A.D. analyzed purified protein samples via negative stain TEM and DLS, conducted enzyme solubility experiments, and performed enzyme assays. M.P.A. prepared cryo-EM grids, collected and processed cryo-EM datasets, and built the protomer models for Mx-Enc1 and Mx-Enc2.

3.1 Introduction

The distinctive scent of wet earth, one of the most widely recognized aromas, originates from microbes found in the soil. More specifically, this distinct smell can be attributed to a few odoriferous compounds produced by these soil bacteria, including the terpenoid geosmin (1, 2). The name geosmin is a reflection of its primary association with its earthy scent, originating from the Greek words “ge” for “earth” and osme” for odor” (1). Geosmin is a very strong odorant, with a taste and odor threshold in the parts per trillion, much lower than many other odoriferous compounds associated with earthy scents (3). Geosmin is produced in many species of Actinobacteria, Myxobacteria, and Cyanobacteria, which primarily inhabit the soil and aquatic systems (4, 5). The biological function of geosmin is still unclear, though several published studies have documented the mixed effects of geosmin on the organismal behavior of ants, flies, worms, and even elephants (6-10). A notable study conducted using springtails demonstrated that

these arthropods are attracted to the scent of geosmin and 2-methylisoborneol, which is produced by the soil bacterium *Streptomyces coelicolor*. Consequently, springtails would opt to feed on geosmin-producing *S. coelicolor* colonies, and could then subsequently serve as a vector by which to distribute the bacterial spores (11). However, other organisms are deterred by the scent of geosmin, including humans who find geosmin to be a significant contributor to the musty flavors of contaminated drinking water, fish, and wine (12-15).

Geosmin is produced by the enzyme geosmin synthase (GeoA), which catalyzes the Mg^{2+}-dependent cyclization of farnesyl pyrophosphate (FPP) to create geosmin (16-19). Geosmin synthase is a bifunctional enzyme that catalyzes two distinct reaction steps. First, the N-terminal domain (NTD) cyclizes FPP, generating inorganic pyrophosphate (PP_i), germacradienol (major product), and germacrene D (minor product) (17, 20, 21). Then, the C-terminal domain (CTD) will convert germacradienol to geosmin via a proton-dependent cyclization reaction, also generating acetone (19). The structure of the full-length geosmin synthase has not yet been determined, though researchers have successfully crystallized and solved the structure of the NTD and found that it has an α-helical fold that is shared by many other Class I terpene cyclases (22). The same α-helical fold is also predicted for GeoA's C-terminus, though this construct, in addition to the full-length protein, could not be crystallized due to low expression and solubility (22).

Recently, analyses of microbial genomes have identified the presence of geosmin synthase genes in operons that include encapsulin nanocompartments (23). These operons have only been found in the three bacterial phyla known to contain organisms that produce geosmin, which are Actinobacteria, Myxobacteria, and Cyanobacteria. Encapsulins are self-assembling protein compartments that encapsulate enzymatic cargoes and are found in 35 prokaryotic phyla

(23-25). Encapsulins have been found to sequester a number of cargo enzymes for a variety of purposes, including ferroxidases and cysteine desulfurases for iron storage (25-30) and sulfur storage (31, 32), respectively, dye-decoloring peroxidases for oxidative stress resistance (33-36), and the terpene cyclase 2-methylisoborneol synthase (2-MIBS) for terpene biosynthesis (37, 38). It is intriguing to find *geoA* genes in encapsulin operons since many organisms are capable of producing geosmin without encapsulating their geosmin synthases in a protein shell (5, 16, 20, 39).

In this chapter, we report the initial characterization of a two-component geosmin synthase encapsulin from Family 2B identified in *Myxococcus xanthus*. Though geosmin production has previously been studied in *M. xanthus* (40), GeoA's encapsulation into a protein compartment has never been demonstrated. From the heterologous expression and purification of the two-component Mx operon, we demonstrate the differential assembly behavior of the two shell proteins and confirm their ability to form mixed-component shells in the absence of cargo. We further show the ability of single-protomer shells and mixed-protomer shells to load enzymatically active geosmin synthase cargoes, which exhibit similar initial rates of enzyme activity. Using cryo-EM, we have determined the structures of both shell proteins and found that cargo-loading of geosmin synthase is mediated at the three-fold pore. Further, we have identified two targeting peptide regions on the geosmin synthase and demonstrated their importance in cargo loading. In summary, this work highlights the novelty of a cargo-loading two-component encapsulin and presents preliminary evidence for encapsulation as a method of stabilizing the cargo enzyme.

3.2 Materials and Methods

3.2.1 Molecular biology and cloning

The genes used in this study were obtained from the genome of *Myxococcus xanthus* strain DK1622, and first codon optimized for *E. coli* expression using the Integrated DNA Technologies (IDT) codon optimization tool. The resulting gene fragments (Supplemental Table 3-1) were synthesized and inserted into the multiple cloning site (MCS) of the pET-21(+) plasmid by Twist Bioscience. Since the initial expression of His-Mx-Enc2 was very low, the Mx-Enc2 was codon optimized a second time using the Salis Lab CDS Calculator and ordered as a gBlock gene fragment from IDT (Supplemental Table 3-1). Primarily, plasmids were constructed via Gibson Assembly after performing PCR to amplify the pETDuet backbone and the gene insert of interest (Supplemental Table 3-2). Once the plasmid backbone was linearized, a Gibson Assembly reaction was performed using the New England Biolabs (NEB) Master Mix and the following manufacturer's instructions (Supplemental Table 3-3). For constructs containing two Mx-Enc genes, both were inserted into the MCS2 of the pETDuet plasmid with a short synthetic RBS in between. Constructs containing GeoA and one or both Enc genes were assembled such that GeoA was placed in MCS1 and the Enc gene(s) were placed in MCS2. For the His-Mx-Enc1_Mx-Enc2 construct, an inverse PCR strategy was used in which the Mx-Enc1_Mx-Enc2 plasmid was amplified with the His-tag added via primer extension. The resulting linear plasmid was phosphorylated with T4 polynucleotide kinase and ligated with T4 DNA ligase following the manufacturer's protocol. Electrocompetent *E. coli* BL21(DE3) cells were transformed with the assembled plasmids and grown on LB agar plates containing 100 μg/mL ampicillin that were incubated overnight at 37˚C. Gene sequences were validated by Sanger sequencing (Supplemental Table 3-4).

3.2.2 Protein expression and purification

Overnight starter cultures of Terrific Broth (TB) media supplemented with 100 μg/mL ampicillin were inoculated with *E. coli* BL21(DE3) cells containing expression plasmids and incubated while shaking overnight at 37˚C, 200 rpm. 2 mL of starter culture was used to inoculate flasks containing 100 mL TB media with 100 μg/mL ampicillin at a ratio of 1:100. Cultures were grown at 37˚C, 200 rpm until an OD600 of 0.4 – 0.8 was reached. Subsequently, cultures were induced with 0.15 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and cells were grown for 20-22 hours at 18˚C, 200 rpm. Cells were harvested by centrifugation at 8,000 rcf for 15 minutes at 4˚C and stored at -80˚C until use.

Frozen cell pellets were thawed on ice for 30 minutes and resuspended in 5 mL of Tris Buffered Saline (TBS) (20 mM Tris, 300 mM NaCl, pH 7.5) supplemented with 1 mM tris(2-carboxyethyl)phosphine (TCEP) per 1 g of wet cell weight. Lysozyme (0.5 mg/mL), DNase (1 μg/mL), Benzonase nuclease (25 units/mL), and SIGMAFAST EDTA-free protease inhibitor cocktail (1 tablet per 100 mL) were added to the resuspension buffer. Cell suspensions were incubated at 4˚C while gently rocking for 30 minutes. Cells were lysed by sonication for 4 minutes, pulsing for 8 seconds on and 10 seconds off, at 95 W with a Model 120 Sonic Dismembrator (Thermo-Fisher Scientific). Lysed cells were centrifuged at 8,000 rcf for 15 minutes at 4˚C to remove cell debris.

The soluble lysate was applied to a 2 mL packed Ni-NTA resin bed (GoldBio) and equilibrated with 10 column volumes of TBS. Subsequent Ni-NTA purification steps were conducted at room temperature. The Ni-NTA resin was then washed with TBS using a stepwise gradient in increments of 30 mM imidazole up to 90 mM to remove background proteins. TBS

buffer containing 250 mM imidazole was used to elute the His-tagged protein from the resin. Protein samples were then analyzed for purity with SDS-PAGE using 14% Novex Tris-Glycine Gels (Thermo-Fisher Scientific). Elution fractions containing purified protein were pooled and concentrated using a 50 kDa MWCO Amicon Ultra-15 mL centrifugal units. The concentrated protein sample was then either stored at 4˚C overnight or centrifuged for 10 minutes at 10,000 rcf to remove any precipitated protein in preparation for size exclusion chromatography (SEC). SEC was then conducted at 4˚C using a Superose 6 10/300 GL column pre-equilibrated with TBS on an ÄKTA Pure FPLC system.

3.2.3 GeoA Solubility Test

Frozen cell pellets containing GeoA with an N-terminal His-tag and TEV cleavage site were thawed on ice for 30 minutes. Once the pellet was sufficiently thawed, the pellet was distributed between 8 microcentrifuge tubes, each containing approximately 50 mg of wet cell pellet. Each 50 mg cell pellet was resuspended in 0.5 mL of a different buffer solution (Table 3-1). Small scoops of DNase and lysozyme were added to each tube, and cells were resuspended by briefly vortexing. Cells resuspended in buffer conditions 1-7 were incubated at 4˚C while gently rocking for 1-2 hours. The cell pellet resuspended in buffer condition 8 was incubated at room temperature while gently rocking before being centrifuged at 4˚C, 10,000 rcf for 10 minutes. The supernatant (soluble lysate) was immediately separated from the pellet (insoluble lysate) and stored at 4˚C. The other seven samples were lysed via sonication for 50 seconds, pulsing for 8 seconds on and 10 seconds off, at 95 W with a Model 120 Sonic Dismembrator. Following sonication, samples were also centrifuged at 4˚C, 10,000 rcf for 10 minutes, where the supernatant was immediately pipetted off of the pellet and into a separate tube. The insoluble

fractions for all samples were resuspended in 500 µL TBS 20/300, pH 7.5, 1 mM TCEP. All pellet, supernatant, and unlysed cell samples were analyzed via SDS-PAGE to determine the solubility of GeoA in different buffer conditions.

Table 3-1: Buffer conditions used for geosmin synthase solubility test.

Condition Number	Resuspension Buffer
1	TBS 20/300, pH 7.5, 1 mM TCEP
2	TBS 20/300, pH 7.0, 1 mM TCEP
3	TBS 20/300, pH 8.0, 1 mM TCEP
4	TBS 20/300, pH 7.5, 1 mM TCEP, 10% glycerol
5	TBS 20/300, pH 7.5, 1 mM TCEP, 20% glycerol
6	TBS 20/300, pH 7.5, 1 mM TCEP, 1% Triton X-100
7	TBS 20/300, pH 7.5, 1 mM TCEP, 1% CHAPS
8	B-PER Bacterial Protein Extraction Reagent (Thermo-Fisher Scientific)

3.2.4 Negative stain transmission electron microscopy

Formvar-enforced carbon grids (FCF-200-Au, EMS) were glow-discharged using a PELCO easiGlow at 5 mA for 60 seconds to increase their hydrophilicity. 3.5 µL of the purified Mx encapsulin samples at a concentration of 0.1 mg/mL in TBS were applied to the grids and allowed to adsorb for 1 minute. The excess liquid was blotted with Whatman filter paper. The grids were then briefly washed with 10µL distilled water, and the excess liquid was blotted. Grids were again briefly washed with 0.75% (w/v) uranyl formate and blotted again. Then, the grids were floated on a 10 µL drop of 0.75 (w/v) uranyl formate for 1 minute. After the excess stain was removed with blotting, samples were immediately imaged using a 100 kV Morgagni transmission electron microscope at the University of Michigan Life Sciences Institute.

3.2.5 Dynamic light scattering (DLS) and thermal ramp experiments

Mx encapsulin samples were analyzed with DLS and thermal ramp experiments using an Uncle instrument (Unchained Labs) at 25˚C in triplicate. Protein samples were centrifuged at

20,000 rcf for 10 minutes at 4˚C to remove protein aggregates. The protein samples were then adjusted to final concentrations of 0.05 and 0.1 mg/mL, and both concentrations were immediately analyzed via DLS. Melting temperature (Tm) and aggregation temperature (Tagg) analysis was conducted over a 25˚C to 95˚C thermal ramp at 1°C per min, measuring static light scattering (SLS) at 266 nm.

3.2.6 Protein identification via mass spectrometry

Purified Mx encapsulin protein samples were analyzed via SDS-PAGE using 14% Novex Tris-Glycine Gels (Thermo-Fisher Scientific). The SDS-PAGE gel was stained with Coomassie, and the protein bands at the correct molecular weights were excised and stored in water in a microcentrifuge tube. The SDS-PAGE gel bands were transferred to the University of Michigan Biomedical Research Core Facilities (BRCF) Proteomics & Peptide Synthesis Core, where the protein was extracted from the SDS-PAGE gel and subjected to tryptic digestion at 37˚C for 4 h, quenched, and analyzed. The digested samples were then analyzed using a nano LC/MS/MS Waters HPLC system coupled to a Thermo-Fisher Scientific LTQ Orbitrap Velos Pro following standard MS/MS protocol. The data were searched using a local copy of Mascot (Matrix Science, UK).

3.2.7 Cryo-electron Microscopy

Sample preparation: Purified samples of Mx-Enc1_GeoA were concentrated to 0.1 mg/mL in 150 mM NaCl, 25 mM Tris pH 8.0. 3.5 µL of protein was applied to freshly glow discharged (10 seconds, 5 mA) Quantifoil R1.2/1.3 Cu graphene oxide-coated 400 mesh grids

(EMS, Cat # GOQ400R1213Cu10) and were frozen by plunging into liquid ethane using an FEI Vitrobot Mark IV (100% humidity, 22 °C, blot force 0, blot time 4 seconds, wait time 30 seconds). The grids were immediately clipped and stored in liquid nitrogen until data collection. Purified samples of Mx-Enc1_Mx-Enc2_GeoA were concentrated to 3.3 mg/mL in 150 mM NaCl, 25 mM Tris pH 8.0. 3.5 µL of protein samples were applied to freshly glow discharged grids Quantifoil R1.2/1.3 Cu 200 mesh grids and were frozen by plunging into liquid ethane using an FEI Vitrobot Mark IV (100% humidity, 22 °C, blot force 20, blot time 4 seconds, wait time 0 seconds). The grids were immediately clipped and stored in liquid nitrogen until data collection.

Data collection: Cryo-EM movies were collected using an FEI Titan Krios G3 cryo-transmission electron microscope operating at 300 kV and equipped with a Gatan K3 Direct Detector with a Bioquantum Imaging Filter. SerialEM (41) was used to select targets and acquire movies. Settings for data collection were: defocus range -0.8 to -1.8 µm, dose of 50 $e^-/Å^2$, magnification of 105,000x, exposure time of 1.9 s. 4,585 movies were collected for the Mx-Enc1_Mx-Enc2_GeoA dataset, and 1,453 movies were collected for the Mx-Enc1_Mx-Enc2_GeoA dataset.

Data Processing: All data processing was performed using CryoSPARC v4.5.1 (42). For the Mx-Enc1_GeoA dataset, 4,585 movies were processed using CryoSPARC Live, which was used to perform patch motion correction and patch CTF estimation. Movies with CTF fits worse than 5 Å were rejected, resulting in 4,245 accepted exposures. Template picking was performed in CryoSPARC Live, and 139,267 particles were extracted with a box size of 480 pixels. Two rounds of 2D classification were performed on the particle set, resulting in 125,185 remaining particles. These particles were then used in ab-initio reconstruction with I symmetry imposed and

3 classes, with the main class containing 124,862 particles. Homogeneous refinement was then performed on this particle class with I symmetry imposed, defocus refinement enabled, global CTF refinement enabled, Ewald sphere correction (negative curvature) enabled, and minimize over per-particle scaling enabled, resulting in a 2.33 Å reconstruction. The particles were then polished using reference-based motion correction. A final homogenous reconstruction was then performed with I symmetry imposed, defocus refinement enabled, global CTF refinement enabled, Ewald sphere correction (negative curvature) enabled, and minimize over per-particle scaling enabled, resulting in a final 2.32 Å resolution reconstruction.

For the Mx-Enc1_Mx-Enc2_GeoA dataset, 1,453 movies were processed using CryoSPARC Live, which was used to perform patch motion correction and patch CTF estimation. Movies with CTF fist worse than 6 Å were rejected, resulting in 1,430 movies. Template picking was performed in CryoSPARC Live, and 198,597 particles were extracted using a box size of 480 pixels. Two rounds of 2D classification were performed on the particle set, resulting in 192,305 remaining particles. Initial volumes were created using ab-initio reconstruction with I symmetry imposed and three volume classes, with the main volume containing 192,290 particles. Homogeneous refinement was performed on these particles with I symmetry imposed, defocus refinement enabled, global CTF refinement enabled, Ewald sphere correction enabled (negative curvature), and minimize over per-particle scaling enabled resulting in a 2.44 Å reconstruction. The particles were then polished using reference-based motion correction. A final homogeneous reconstruction was then performed on the polished particles with I symmetry imposed, defocus refinement enabled, global CTF refinement enabled, spherical aberration fitting enabled, anisotropic magnification fitting enabled, tetrafoil fitting enabled,

Ewald sphere correction enabled (negative curvature), and minimize over per-particle scaling enabled, resulting in a 2.43 Å reconstruction.

Model building: AlphaFold 2 (43) models of protomers of MxEnc1 (Uniprot ID: Q1CYZ6) and MxEnc2 (Q1CYZ5) were manually placed in the respective icosahedral maps using the fit in map command in ChimeraX (44). Models were then iteratively refined using manual real space refinements in Coot (45, 46), followed by real space refinement using phenix.real_space_refine in Phenix v1.20.1-4487 (47) using default refinement parameters. Icosahedral symmetry operators were identified from the maps using phenix.map_symmetry and applied to the protomers to generate icosahedral shells using phenix.apply_ncs. BioMT operators were then identified using phenix.find_ncs and manually put in the headers of the PDB files.

3.2.8 Geosmin synthase activity assays

Geosmin synthase activity assays were performed using the EnzCheck pyrophosphate assay kit (Thermo-Fisher Scientific, Cat # E-6645). Assay components were added based on a final reaction volume of 100 µL, with the following EnzChek reagents added according to the manufacturer's instructions: 20X reaction buffer (1.0 M Tris- HCl, 20 mM MgCl, pH 7.5, 2 mM sodium azide), 2-amino-6-mercapto-7- methyl purine ribonucleoside (MESG), purine nucleoside phosphorylase (PNP), and inorganic pyrophosphatase (PPase). $MgCl_2$ was added to a final concentration of 10 mM for all reactions. This mixture was then divided into four separate tubes, with different enzyme samples added to each tube. Two samples of encapsulated GeoA were tested (Mx-Enc1_GeoA and Mx-Enc1_Mx-Enc2_GeoA). Since the two encapsulated GeoA samples were found to exhibit different cargo loading efficiencies, the overall protein concentration was standardized to the amount of total enzyme, which was determined from gel

densitometry of both samples. MilliQ water was then added to the mix, such that the final volume of the master mix was 95 µL per reaction. After the master mix was resuspended, 95 µL of the master mix containing the enzyme sample was aliquoted into 96 clear-bottom well plates. Assays were then initiated by the addition of 5 µL FPP ammonium salt (Cayman Chemical, Cat # 63250) or water into the sample wells. Each concentration of substrate was measured in triplicate. The final concentrations of FPP used were 0, 10, and 50 µM. After the addition of FPP, plates were immediately incubated in a BioTek Synergy HI microplate reader at 26˚C and the absorbance at 360 nm was measured every 22 seconds in sweep mode.

3.2.9 Gas chromatography mass spectrometry (GC-MS) analysis

Enzymatic reactions were prepared in sealed vials with a total reaction volume of 200 µL containing: 20 mM Tris, pH 7.5, 300 mM NaCl, 1 mM TCEP, 10 mM $MgCl_2$, 50 µM FPP, and 5 µM Mx-Enc1_Mx-Enc2_GeoA (total enzyme concentration estimated to be 1 µM based on gel densitometry). 100 µL of pentane was pipetted on top of the reaction and incubated for 24 hours at room temperature. After incubation, samples were centrifuged for 1 minute at 14,000 rcf and the aqueous layer was pipetted into a new tube. 200 µL pentane was added to the saved aqueous layer, vortexed to mix, and centrifuged again using the same conditions as before. This pentane layer was pooled with the previous pentane layer, and several small crystals of disodium sulfate were added to remove excess water. After a 30-minute incubation, the pentane layer containing organic volatiles was transferred to a new sample vial for analysis with GC-MS. A commercially available 1 mg/mL geosmin solution (Millipore Sigma, G5908) was used as a GC-MS standard. The sample was transferred to the University of Michigan Mass Spectrometry Core Facility for GC-MS analysis. 1 µL of each sample was injected into a Shimadzu QP-2010 GC-MS

instrument containing a DB-5 column, run in splitless mode with a column flow rate of 0.97 mL/min and a temperature ramp of 60°C to 280°C at 20°C/min. Data analysis was carried out in OpenChrom v 1.5.

3.3 Results

3.3.1 Heterologous production of encapsulin shell components demonstrates differential assembly behavior and ability to form mixed shells

This section focuses on the study of a two-component encapsulin operon from *M. xanthus* with a geosmin synthase cargo (Figure 3-1A). To determine if the gene products from the putative encapsulin operon form encapsulin shells, the genes were heterologously expressed in *Escherichia coli* and subsequently purified via nickel affinity purification and size exclusion chromatography (SEC). Purified protein samples were analyzed via negative stain transmission electron microscopy (TEM) to determine if encapsulin shells were formed. Mx-Enc1 was observed to form uniform encapsulin shells approximately 20 nm in diameter, consistent with the size of other T=1 encapsulins from Family 2B (Figure 3-1B) (38), whereas Mx-Enc2 was not observed to form any type of uniform assembly, including complete shells or shell facets (Figure 3-1B). Notably, the finding that Mx-Enc2 cannot assemble into uniform shells alone while Mx-Enc1 can is in direct contrast to the two-component encapsulin system observed in *S. lydicus*, where only Sl-Enc2 was observed to form encapsulin shells.

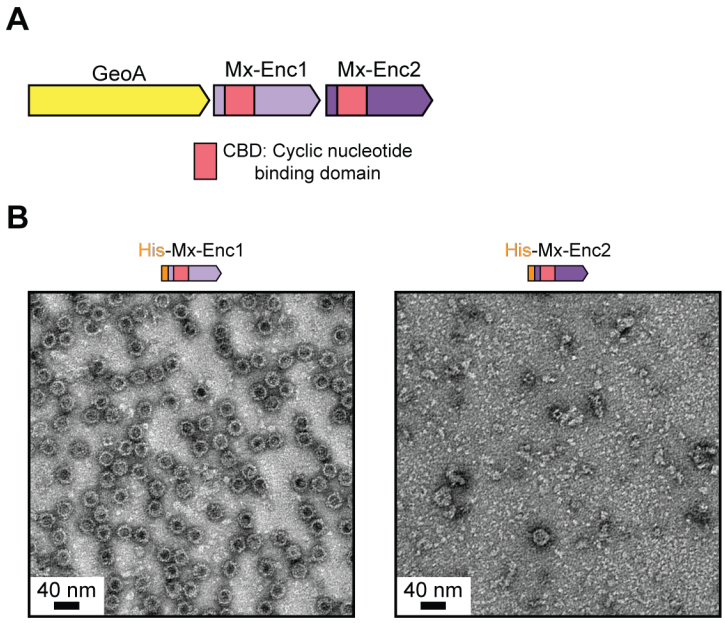


Figure 3-1: *Myxoccocus xanthus* Family 2B two-component encapsulin operon with a geosmin synthase (GeoA) cargo and heterologous production. (**A**) A two-component encapsulin operon containing a GeoA cargo identified in *M. xanthus*. (**B**) Negative stain TEM micrographs of individually expressed and purified Mx-Enc1 and Mx-Enc2.

A differential His-tagging strategy was then used to determine if Mx-Enc1 and Mx-Enc2 could form mixed shells. First, two expression plasmids were constructed containing Mx-Enc1 and Mx-Enc2, where an N-terminal His-tag was subsequently added to only one of the shell genes per plasmid. Heterologously expressed proteins were again purified via affinity chromatography, followed by size exclusion chromatography. Encapsulin shell formation was verified for both samples using negative stain TEM (Figure 3-2A). SDS-PAGE analysis demonstrated the high purity of the protein sample, and the SDS-PAGE bands corresponding to the size of Mx-Enc1 and Mx-Enc2 were excised and subjected to tryptic digestion and LC-MS analysis to determine the sample composition of the assembled shell (Figure 3-2B). Mx-Enc1 and Mx-Enc2 peptides were found in both samples, demonstrating that the two proteins can indeed co-assemble to form mixed shells. Unfortunately, the SDS-PAGE bands for Mx-Enc1 and Mx-Enc2 are indistinguishable due to their similar mass and charge, so the bulk ratios of the two proteins were unable to be determined. Dynamic light scattering (DLS) was used to determine

the polydispersity of shells for each sample, where diameters of 29.5 and 27.2 nm were observed, both similar to each other and consistent with the diameters observed in TEM micrographs (Figure 3-2C). The PDI for the mixed shell sample with His-tagged Mx-Enc2 is noticeably lower than that observed for the other dataset, suggesting that shells enriched with Mx-Enc2 may have improved stability. This finding was also observed for the *S. lydicus* two-component encapsulin, in which incorporation of Sl-Enc1 noticeably increased the monodispersity and thermal stability of the encapsulin shells when compared to homogenous Sl-Enc2 shells.

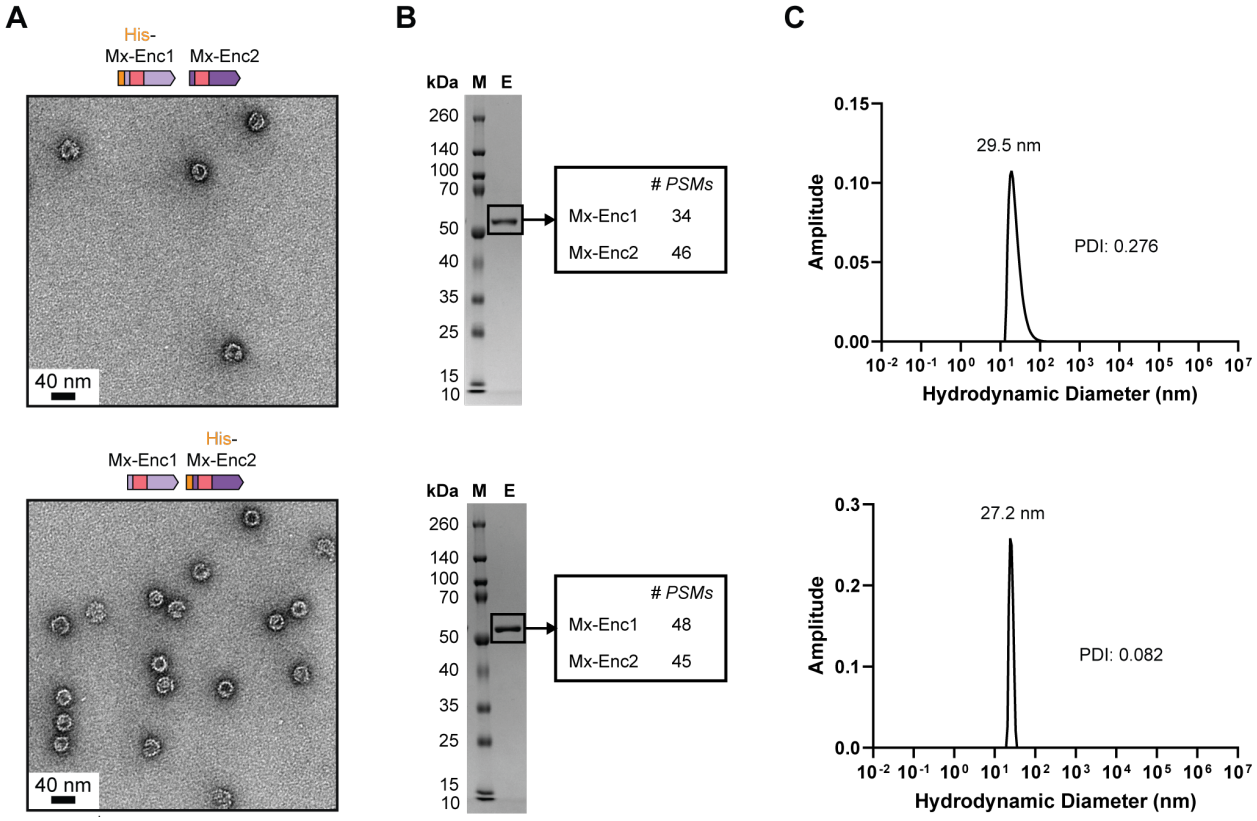


Figure 3-2: Analysis of mixed Mx-Enc1 and Mx-Enc2 shells. (**A**) Negative stain TEM micrograph of purified mixed shells with singly His-tagged Mx-Enc1 (top row) and a micrograph of mixed shells with singly His-tagged Mx-Enc2 (bottom row). (**B**) SDS-PAGE of purified mixed encapsulin shell samples. M: Molecular weight marker E: Encapsulin shell. The SDS-PAGE bands shown in the box were excised and subjected to tryptic digestion and subsequent LC-MS analysis. Peptide-spectrum matches (PSMs) from LC-MS for each shell component are listed. (**C**) Dynamic light scattering (DLS) analysis of His-Mx-Enc1 + Mx-Enc2 shells (top row) and Mx-Enc1 + His-Mx-Enc2 shells. PDI: Polydispersity index.

3.3.2 The M. xanthus two-component encapsulin's cargo is geosmin synthase

To determine if geosmin synthase is an encapsulin cargo, we constructed three plasmids: one co-expressing GeoA with both encapsulin shell genes, and two co-expressing GeoA with solely His-Mx-Enc1 or His-Mx-Enc2. Unfortunately, *E. coli* cells transformed with the GeoA_His-Mx-Enc2 plasmid exhibited very low expression of both proteins, and we were unable to obtain adequate amounts of purified protein to determine if there are interactions between Mx-Enc2 and GeoA. The remaining two constructs were heterologously expressed and then purified using nickel affinity chromatography. Since GeoA was not His-tagged, GeoA would only be present in the elution fraction if it interacted with the His-tagged encapsulin shell proteins. After the elution fractions were collected from the affinity purification step, the samples were further purified via SEC to isolate the assembled encapsulins. For GeoA co-expressed with His-Mx-Enc1 and His-Mx-Enc2, a major peak was observed at an elution volume of ~12 mL, characteristic for T=1 shells (Figure 3-3A). SDS-PAGE analysis of the major peak demonstrated bands corresponding to molecular weights of Mx-Enc1/Mx-Enc2 and GeoA (Figure 3-3B), and subsequent tryptic digest and LC-MS analysis identified both Mx-Enc proteins in the sample. Negative stain TEM micrographs confirmed the assembly of mixed encapsulin shells, with the density inside the shell indicating successful cargo-loading (Figure 3-3C). This result demonstrates that geosmin synthase is an encapsulin cargo, and mixed shells specifically are capable of loading geosmin synthase cargo.

The same purification protocol was used for the construct containing His-Mx-Enc1 and GeoA, where a peak at the characteristic T=1 elution volume for SEC was observed (Figure 3-3D). SDS-PAGE and TEM analysis demonstrate the encapsulation of GeoA within assembled, uniform encapsulin shells (Figure 3-3, E and F). Notably, the Mx-Enc1 shell demonstrated more

efficient cargo loading than the mixed shell sample, with GeoA accounting for roughly 25% of the total bulk protein sample as determined by gel densitometry, equating to approximately 15 monomers per shell. Comparatively, the mixed shell sample had slightly less efficient cargo loading, with only 20% of the total protein sample comprised of GeoA (12 monomers/shell). This result could indicate that Mx-Enc1 better facilitates interactions with the GeoA, allowing for greater cargo-loading capabilities. In both cases, however, cargo loading was nearly twice as high as that observed for the *S. griseus* Family 2B encapsulin with a 2-MIBS cargo (38).

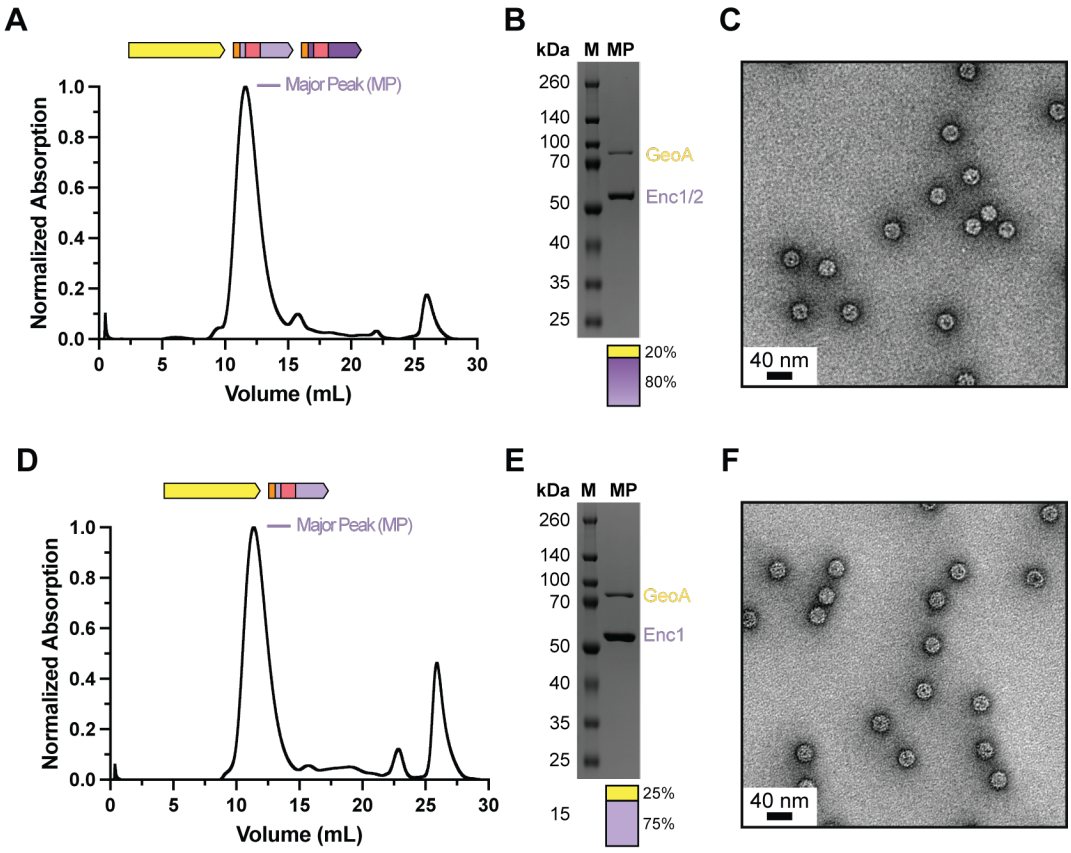


Figure 3-3: Cargo loading of geosmin synthase into Mx-Enc shells. (**A**) SEC of mixed Mx-Enc1/Mx-Enc2 shells with GeoA (Superose 6 column). The major peak (MP) is consistent with the size of T=1 encapsulin shells. (**B**) SDS-PAGE of SEC MP showing bands for GeoA and Mx-Enc1/Mx-Enc2. (**C**) Negative stain TEM micrograph of cargo-loaded mixed shells. (**D**) SEC of Mx-Enc1 shells with GeoA (Superose 6 column). (**E**) SDS-PAGE of SEC MP showing bands for GeoA and Mx-Enc1. (**F**) Negative stain TEM micrograph of cargo-loaded Enc1 shells.

3.3.3 Single particle cryo-EM analysis of Mx encapsulin shells

To determine the structures and cargo loading mechanism of the two-component Mx-Enc shell, single particle cryo-EM analysis was carried out using two separate Mx encapsulin samples: GeoA_Mx-Enc1 and GeoA_Mx-Enc1_Mx-Enc2. The icosahedral refinement of both samples yielded symmetry-averaged maps with resolutions of 2.32 Å (GeoA_Mx-Enc1) and 2.43 Å (GeoA_Mx-Enc1_Mx-Enc2). The map for the mixed shell sample resulted in a symmetry-averaged map for Mx-Enc2, from which the Mx-Enc2 atomic model could be built (Figure 3-4A). The map from the GeoA_Mx-Enc1 sample was used to build the Mx-Enc1 model (Figure 3-4A).

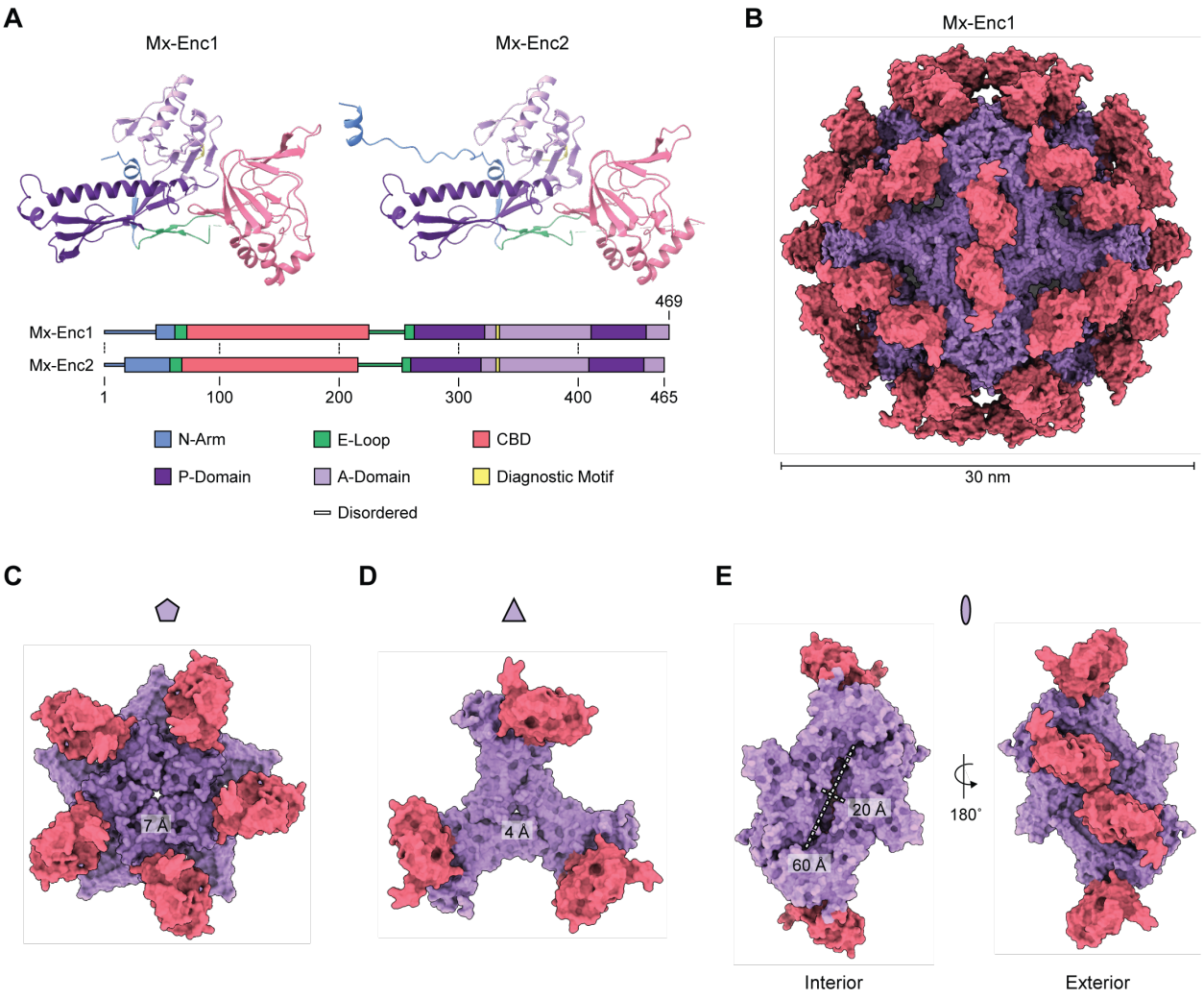


Figure 3-4: Structural analysis of Mx encapsulin shells via single particle cryo-EM. (**A**) Protomer structures of Sl-Enc1 and Sl-Enc2. Structures are colored by domain. (**B**) Homogeneous Mx-Enc1 shell with T=1 icosahedral symmetry. The HK97-fold is colored purple, and the CBD is colored pink. (**C**) Exterior view of the five-fold pore in the homogeneous Mx-Enc1 shell. (**D**) Exterior view of the three-fold pore in the homogenous Mx-Enc1 shell. (**E**) Exterior and interior views of the two-fold pore in the homogenous Mx-Enc1 shell. Pore diameters are highlighted.

The Mx-Enc protomer structures both contain the HK97-fold common to all encapsulins as well as a CBD characteristic of Family 2B encapsulins (Figure 3-4A). The protomer structures are highly similar to each other, with an extended N-terminal arm resolved only for Mx-Enc2. We investigated the structures to determine the location and sequence of diagnostic motif 1, which was identified in Chapter 2 as the most conserved sequence difference between all Enc1s and Enc2s. The diagnostic motif sequence for Mx-Enc1 is $A_{334}TVW_{337}$, and for Mx-Enc2 it is $S_{333}RRR_{336}$. Both Mx sequences are identical to the consensus sequence identified for Enc1s and

Enc2s, respectively. Similar to its location on the Sl-Enc proteins, this motif is located at the edge of the A-domain, where it may interact with neighboring protomers in the encapsulin shell (Figure 3-4A).

The Mx-Enc1 shell was found to be a T=1 icosahedral assembly, approximately 30 nm in diameter, consistent with previous analyses of the Mx encapsulin (Figure 3-4B) and matching the size of other Family 2B shells. The shells contain three primary pores, with the smallest measuring 7 Å in diameter at the five-fold symmetry axes and 6 Å at the three-fold symmetry axes (Figure 3-4, C and D). The largest pore is observed at the two-fold interface, with a long, interior cavity measuring 20 by 60 Å (Figure 3-4E). These large, two-fold pores are also observed for the other two Family 2B encapsulin shell structures (38). Altogether, the two-fold pores observed in all Family 2B shells are the largest reported to date. For both cryo-EM samples, additional density was observed on the interior of the encapsulin shell, corresponding to the encapsulated geosmin synthase. However, the resolution of the cargo was too low for model building, likely owing to the flexibility of GeoA within the shell.

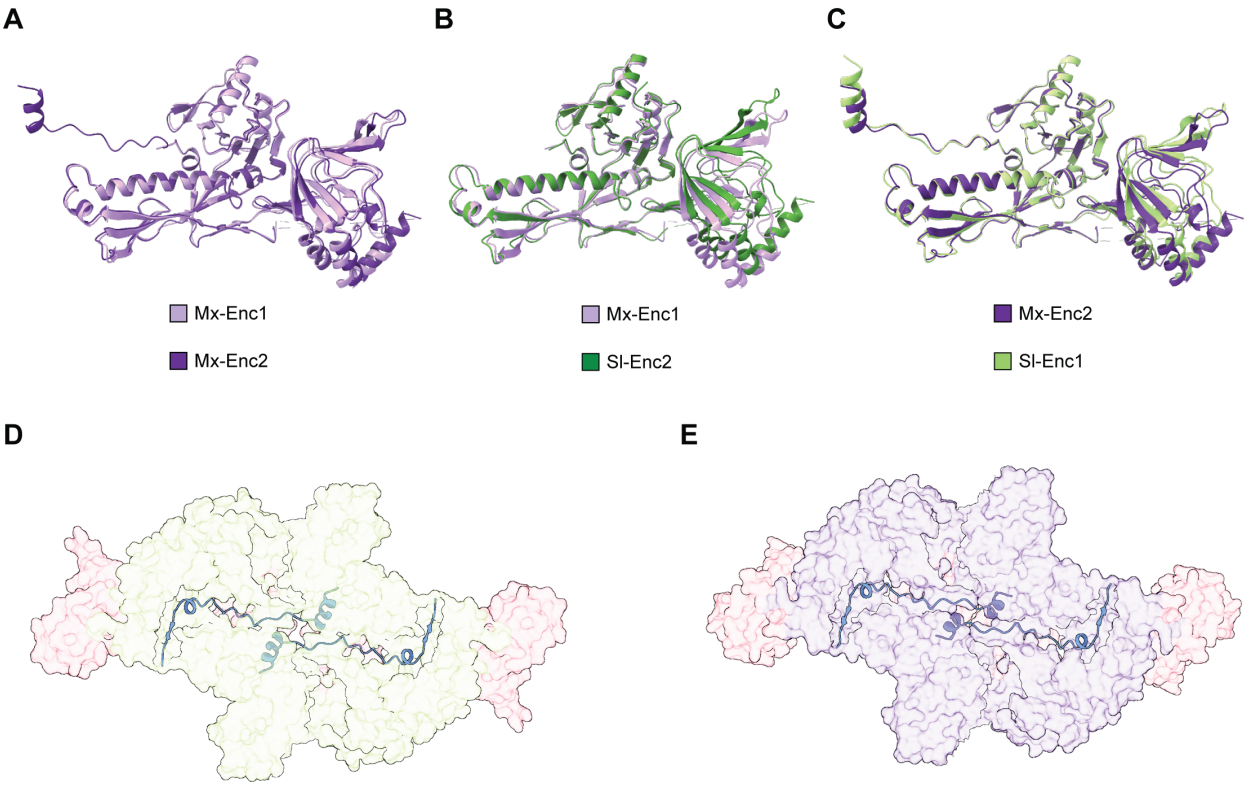


Figure 3-5: Structural comparison of two-component encapsulin protomers. Structural alignment of the following protomers (left to right): (**A**) Mx-Enc1 with Mx-Enc2; (**B**) Mx-Enc1 with Sl-Enc2; (**C**) Mx-Enc2 with Sl-Enc1. For the Sl-Enc1 and Sl-Enc2 protomers, the AlphaFold predictions for the CBDs were docked into the HK97-fold domain models. (**D**) Sl-Enc1/Sl-Enc1 interactions at the two-fold pore. N-arm is shown in blue. (**E**) Predicted Mx-Enc2/Mx-Enc2 interactions at the two-fold pore. N-arm is shown in blue.

The protomer structures of Mx-Enc1 and Mx-Enc2 are very similar, with the most notable difference being the presence of an extended N-arm resolved only in Mx-Enc2 (Figure 3-5A). The only other difference between the Mx-Enc1 and Mx-Enc2 structures is a slight variation in the orientation of the CBDs within the E-loops. Structural alignments of Mx-Enc1 and the self-assembling protomer from the *S. lydicus* encapsulin operon (Sl-Enc2) demonstrate that the structures of the two HK97-fold domains and the CBDs are nearly identical, with the only difference resulting from variations in CBD orientation (Figure 3-5B). Noticeably, the extended N-arm is absent in both protomer models, as they were not resolved from the cryo-EM data. Extended N-arms were also not resolved for the Family 2B encapsulin containing a 2-MIBS cargo, which uses only a single shell protein (38). Similarly, structural alignments of Mx-

Enc2 and the protomer incapable of self-assembly from the *S. lydicus* encapsulin operon (Sl-Enc1) show that the two proteins have nearly identical structures with the same N-arm orientation (Figure 3-5C). Within the context of the encapsulin shell, these N-arms are positioned at the major two-fold pore, significantly reducing its size (Figure 3-5, D and E). Extended N-arms have also been resolved for both Family 2A encapsulin structures and are similarly located at the two-fold pore (31, 32). Though the Mx-Enc2/Mx-Enc2 interaction at the two-fold pore is hypothetical, the Sl-Enc1/Sl-Enc1 interaction was observed via 3D classification analysis. Since the presence of N-arms significantly alters the size of the two-fold pore, they may serve to control metabolite flux through the shell, ultimately regulating the activity of the encapsulated enzyme.

3.3.4 Cargo loading is mediated at the three-fold axis of symmetry

Previous bioinformatic analyses of geosmin synthase encapsulin operons identified two highly conserved sequences in geosmin synthase that are predicted to be targeting peptides (23). The most conserved regions of these consensus sequences are identical, with a sequence of GPTGLGTSAAR. The enrichment of flexible glycine residues is also seen in targeting peptides in Family 1 cargoes (48). These two consensus sequences are found in regions of high predicted disorder in the middle of the protein and at the C-terminus (23). Two similar sequences were identified in the *geoA* gene of *M. xanthus*. According to the predicted AlphaFold structural model, these sequences are located in the flexible intergenic region that connects the two catalytic domains and at the C-terminus (Figure 3-6A).

To determine whether or not the bioinformatically identified motifs are responsible for cargo loading, four initial plasmids were constructed, each containing one mutated GeoA variant

alongside the His-Mx-Enc1 and His-Mx-Enc2 genes. The following GeoA mutants were constructed: (1) a deleted intergenic TP motif with no substitution; (2) substitution of the intergenic motif with an "ASSASSA" sequence; (3) C-terminal truncation removing the putative TP motif; and (4) a substituted intergenic motif with an "ASSASSA" sequence and a truncated C-terminus. However, when these four constructs were heterologously expressed, all cells died during the early exponential growth phase except for the GeoA mutant 4, with both putative TP sequences altered. Due to the inability to obtain a sufficiently large cell mass for protein purification with these three constructs, only the double TP mutant was used for further experimentation. The GeoA double mutant was co-expressed with His-Mx-Enc1 and His-Mx-Enc2 and subsequently purified with affinity chromatography followed by SEC. The SEC chromatogram for this sample showed the characteristic T=1 peak, indicating assembled encapsulin shells had formed (Figure 3-6B). SDS-PAGE analysis of this peak demonstrated co-elution of the mutant GeoA, though the amount of encapsulated GeoA had decreased by more than 50% compared to the non-mutant GeoA (Figure 3-6B and Figure 3-3B). This result indicates that the two identified TP sequences are important for cargo loading, as their deletion or mutation significantly decreases cargo-loading efficiency.

The two consensus TP motifs identified in geosmin synthases are nearly identical to a consensus motif identified in the targeting domain of another terpene cyclase encapsulin cargo, 2-MIBS (23). For the 2-MIBS encapsulin, this motif was similarly found to be nonessential for cargo loading, though its deletion significantly reduced the amount of encapsulated enzyme (38). For both terpene cyclase cargoes, it is likely that some level of cargo loading is still observed due to interactions between the shell and neighboring sequences located within regions of disorder.

The mode of binding is not thought to be sequence specific and is predicted to rely instead on interactions between primarily nonpolar and hydrophobic residues (38).

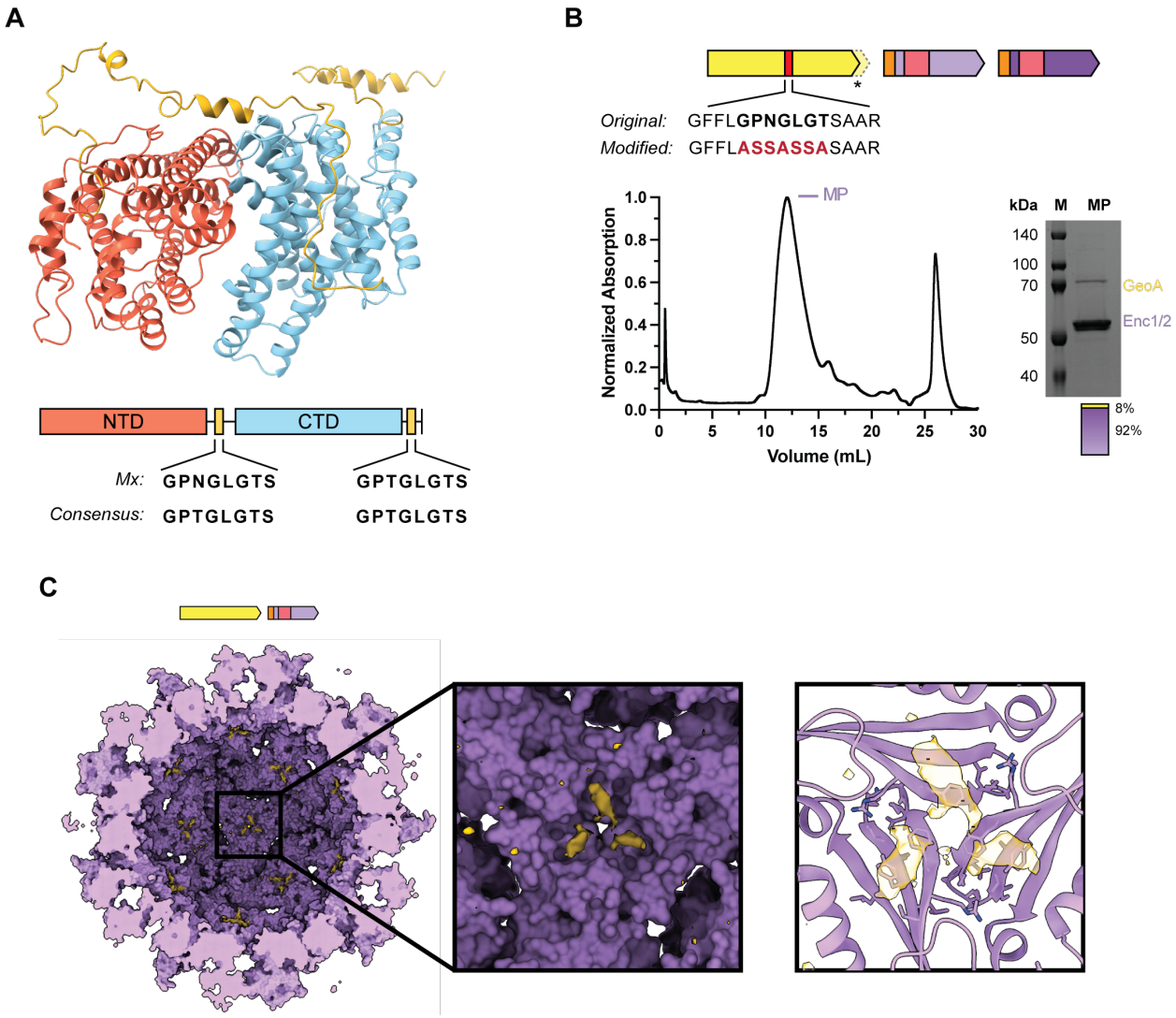


Figure 3-6: GeoA cargo loading into Mx shells. (**A**) AlphaFold predicted structure of *M. xanthus* GeoA with the N-terminal domain (NTD) in dark orange, the C-terminal domain (CTD) in blue, and the intergenic and C-terminal regions in yellow. The predicted targeting peptide sequences for *M. xanthus* GeoA are compared to the overall consensus sequences. (**B**) Operon structure showing the tested GeoA mutant with a truncated C-terminus and mutated TP in the intergenic region, co-expressed with two His-tagged shell proteins. Underneath the operon diagram is the corresponding SEC chromatogram, with an SDS-PAGE gel showing the protein composition of the SEC's major peak (MP). (**C**) Cut-away view of an Mx-Enc1 shell with extracted cryo-EM densities for the TP.

Analysis of the cryo-EM data also revealed extra density around the three-fold pores that could not be attributed to the shell proteins (Figure 3-6C). We believe this density corresponds to that of the targeting peptide, as similar non-shell density has also been observed at the three-fold

pore for all other structurally characterized Family 2 encapsulins (31, 32, 38). Unfortunately, the resolution was too low to determine the specific sequence interacting with the shell. However, given the results of the GeoA mutational analysis, it is probable that this density belongs to one or both of the identified targeting peptides.

3.3.5 Geosmin synthase activity is similar between encapsulins with different shell compositions

We next sought to determine the catalytic activity of GeoA. To accomplish this, we set out to generate purified stocks of non-encapsulated GeoA with the ultimate goal of comparing the enzymatic activity of free GeoA to that of encapsulated GeoA. An N-terminally His-tagged GeoA construct was generated, heterologously expressed, and subsequently, purification was attempted. After the first attempt at purification yielded no soluble protein, a solubility test was used to identify any potential conditions where GeoA exhibited improved solubility. *E. coli* pellets containing overexpressed GeoA were resuspended in eight lysis buffer conditions (Table 3-1) and lysed accordingly. After lysis, the soluble and insoluble lysate fractions were analyzed for protein content via SDS-PAGE (Figure 3-7). For all tested conditions, GeoA was found primarily in the cell pellet, indicating that its overexpression led to the formation of insoluble inclusion bodies. This finding is consistent with the heterologously expressed GeoA proteins from *S. coelicolor, S. avermitilis, and Nostoc punctiforme* (16, 20, 49). In all three cases, full-length GeoA was found to be insoluble and necessitated rescue from inclusion bodies or was co-expressed and purified with the chaperone protein combination GroES/GroEL. Overall, this result suggests that one function of encapsulation may be to stabilize the enzyme for improved activity.

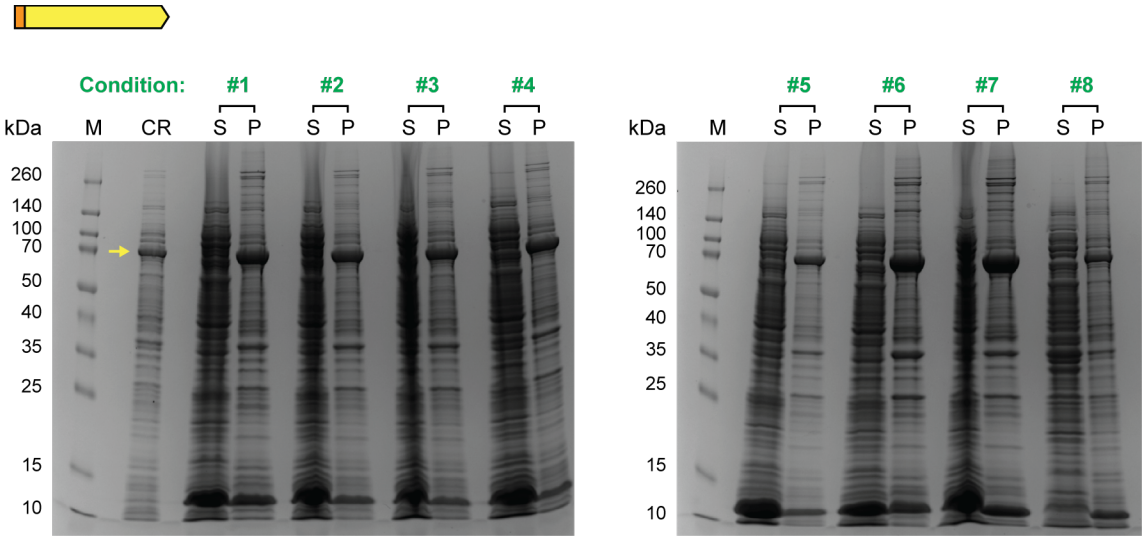


Figure 3-7: Geosmin synthase solubility test. SDS-PAGE was used to analyze the relative solubility of N-terminally His-tagged GeoA after cell lysis and centrifugation, showing the pre-lysis cell resuspension (CR), lysis supernatant (S), and lysis pellet (P) fractions. Lysis buffer conditions #1-8 used to resuspend the cell pellets are listed in Table 3-1. The yellow arrow points to the overexpressed GeoA band (86.9 kDa).

As we could not obtain soluble unencapsulated GeoA, we proceeded with the analysis of enzymatic activity using only encapsulated GeoA. We first wanted to confirm that the encapsulin cargo is a true geosmin synthase by verifying that it catalyzes the conversion of FPP to geosmin (Figure 3-8A). To achieve this, we incubated a purified sample of GeoA_Mx-Enc1_Mx-Enc2 with FPP and, subsequently, performed a pentane extraction to remove the organic, volatile compounds from the solution. The sample was then analyzed using GC-MS and compared to a geosmin standard. This comparison confirmed the production of geosmin (Figure 3-8B), thereby demonstrating the geosmin synthase activity of the cargo.

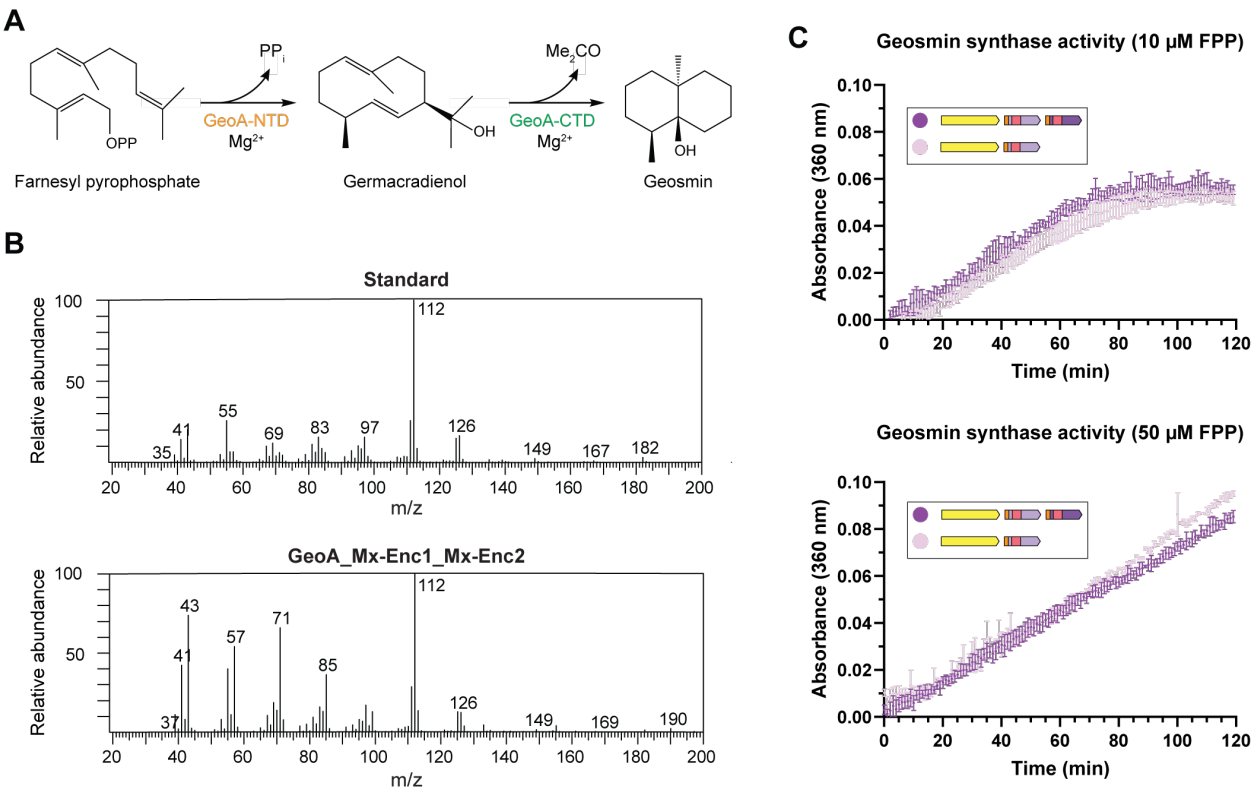


Figure 3-8: Geosmin synthase enzymatic activity. (**A**) Reaction scheme for the conversion of farnesyl pyrophosphate (FPP) to geosmin. The N-terminal domain (NTD) catalyzes the first step in the reaction, and the C-terminal domain (CTD) catalyzes the second step. (**B**) GC-MS analysis of GeoA enzymatic activity. Incubation of GeoA with FPP yielded geosmin, with its characteristic mass fragment at m/z +112 identified as the most abundant ion. The top chromatogram is from a geosmin synthase standard, and the bottom is from Mx-Enc1_Mx-Enc2_GeoA incubated with FPP. (**C**) Measurement of geosmin synthase activity for Mx-Enc1_GeoA and Mx-Enc1_Mx-Enc2_GeoA at two different substrate concentrations. Measurements were performed in triplicate, and error bars represent the standard deviation.

Finally, to determine if there is a difference in enzymatic activity between GeoA encapsulated in shells with different protein compositions, enzyme assays were conducted using two samples: Mx-Enc1_Mx-Enc2_GeoA and _Mx-Enc1_GeoA. For the mixed shell sample, we were unable to determine the bulk ratio of the two shell proteins since they do not form discrete bands on SDS-PAGE. We found that the initial rates of enzyme activity were comparable between the two shell compositions (Figure 3-8C) tested using two different concentrations of FPP. This result shows that, at least for the tested substrate concentrations, the shell composition did not alter the overall enzyme activity. This was surprising, as we had hypothesized that the

mixed encapsulin shell would have lower enzyme activity due to the presence of N-arms in the two-fold pore, limiting the enzyme's substrate access. However, it is worthwhile to note that this enzymatic assay is a bulk measurement and does not provide insight into the nuance or variability within a given sample. This is especially important for the mixed shell sample, where we have no way of knowing the level of Mx-Enc2 incorporation into each encapsulin and if these incorporation levels are consistent for all shells in the sample. Based on the observations made of the *S. lydicus* encapsulins, it stands to reason that there could be significant variability in the level of Mx-Enc2 and subsequent N-arm incorporation, which could serve to confound the results of this experiment.

3.4 Discussion

The work described in this chapter is the first demonstration of an encapsulin that has a geosmin synthase cargo. Further, this work reports the characterization of the first two-component encapsulin to have a confirmed and active enzymatic cargo, since the two-component encapsulin from *S. lydicus* was studied in the absence of cargo. We demonstrated the differential assembly behavior of Mx-Enc1 and Mx-Enc2, in which only Mx-Enc1 (Figure 3-1B) can form assembled encapsulin shells on its own. This finding was surprising because, for the *S. lydicus* two-component encapsulin, only Sl-Enc2 could form assembled shells. The difference in assembly behavior between the two sets of two-component encapsulin protomers indicates that the ability of a given protomer to self-assemble is not determined by whether it is an Enc1 or Enc2. This work has also shown that Mx-Enc1 and Mx-Enc2 can co-assemble to form mixed shells in the presence and absence of cargo (Figure 3-2 and Figure 3-3). However, because the two proteins could not be distinguished using SDS-PAGE, we were unable to determine the

overall ratio of the two protomers or assess whether these ratios changed with the addition of cargo. To identify differences in protomer interactions at the four unique symmetry interfaces between cargo-loaded and empty shells, a 3D classification analysis should be performed on the cryo-EM datasets collected from both samples. This analysis should provide a better understanding of mixed shell assembly and determine if the shell composition or tiling pattern changes to accommodate maximal cargo loading.

We have further shown that both Mx encapsulin protomers contain the HK97-fold and the CBD common to all Family 2B encapsulins (Figure 3-4). Mx-Enc2 was observed to have an extended N-arm that has previously only been identified on Sl-Enc1 and the two Family 2A encapsulins that encapsulate a cysteine desulfurase cargo (31, 32). In all four cases, including Mx-Enc2, the N-arm is found to thread through the major two-fold pore, almost completely closing it (Figure 3-4 and Figure 3-5). In Chapter 2, 3D classification analysis of the *S. lydicus* mixed shell demonstrated the presence of "open" two-fold pores, in which the N-arm was absent, and "closed" pores, where the N-arm was present. In all identified cases of the closed pore states, the N-arm was found to belong to Sl-Enc1, the protomer incapable of assembling encapsulin shells alone. This N-arm was not observed for Sl-Enc2. This analysis also demonstrated that some two-fold pores in the "open" state are comprised of Sl-Enc1 protomers, where the N-arm was vacant. We predict that after performing a 3D classification analysis at the axes of symmetry for mixed *M. xanthus* encapsulin shells, as discussed above, we would observe similar variability in the presence of closed and open two-fold pores in the mixed shells. We hypothesize that N-arms will only be observed in the two-fold pores of mixed shells containing Mx-Enc2.

Ultimately, regulating the size of the two-fold pore may have significant implications for cargo activity by altering the geosmin synthase's access to its substrate. It has been shown that

increasing the size of the encapsulin pore can result in increased enzymatic activity, likely resulting from increased substrate access due to a higher rate of diffusion through the shell (50-52). It stands to reason that the opposite is true as well, where decreasing the pore size or reducing the number of open pores would result in decreased enzymatic activity. However, preliminary data from enzyme activity assays demonstrates that geosmin synthase activity is comparable between encapsulin shells composed of both protomers and only Mx-Enc1 (Figure 3-8). One explanation could be that there is no significant difference in the openness of the shells, allowing for similar levels of substrate diffusion. It is also possible that the substrate concentrations used in these experiments were too high, which might have surpassed the limitations imposed by the reduced number of open two-fold pores in the mixed shell system. To better determine if there is a true difference in enzyme activity based on encapsulin shell composition, these assays should be repeated with lower substrate concentrations.

We have further shown evidence for a cargo loading mechanism in which two discrete targeting peptides are used to mediate cargo loading at the three-fold pore (Figure 3-6). When these two peptides were mutated and removed, lower levels of cargo loading were observed (Figure 3-6B). Unfortunately, the cryo-EM density was poor, so we were unable to determine the amino acid sequence of the targeting peptide. It is also unclear from the density which of the two putative targeting peptides interacts with the shell or if both are involved in facilitating cargo encapsulation. The simultaneous use of both cargo-loading peptides may also explain the lower resolution of the cryo-EM densities. We also observe consistently higher cargo loading for pure Mx-Enc1 shells compared to mixed shells, suggesting that the targeting peptides may have stronger or more specific interactions with the Mx-Enc1 protomer (Figure 3-3). In general, the results presented here are in agreement with the structural data for all other Family 2 encapsulins,

in which cargo-loading is mediated at the three-fold pore through the use of targeting sequences rich in proline, glycine, and alanine (23, 31, 32, 38).

The benefit of encapsulating geosmin synthase within a protein shell remains unclear. This work has presented evidence suggesting that one benefit of encapsulation may be to stabilize the enzyme since it is insoluble when expressed in the absence of an encapsulin shell (Figure 3-7). Although most GeoA homologs are reported to be largely insoluble when expressed heterologously, the organisms from which these homologs originate are still capable of producing geosmin. This suggests that there is soluble and active enzyme present within the cells (16, 20, 39). Furthermore, only 118 encapsulin operons with a putative GeoA cargo have been identified, though more than 1,000 predicted *geoA* genes have been identified in bacterial genomes (4, 23). Altogether, this shows that geosmin synthases do not necessarily need to be encapsulated to be soluble and active. Another potential advantage of GeoA encapsulation lies in the ability of the shell to prevent the escape of the reaction intermediate, germacradienol. Current evidence does not support the existence of a channel connecting GeoA's two catalytic domains (4, 19), and the co-localization of the CTD with its substrate may improve the overall rate of reaction (53).

In summary, this work has shown that two-component encapsulin nanocompartments can successfully load catalytically active cargo involved in terpenoid biosynthesis. Future experimentation will be essential in determining the primary benefit(s) of geosmin synthase encapsulation and specifically how a mixed shell may impact the overall encapsulin function.

3.5 Supplementary Information

Supplemental Table 3-1: DNA sequences of gene fragments.

gBlock	DNA Sequence
His-Mx-Enc1 + His-Mx-Enc2	AAATAATTTTGTTTAACTTTAAGTCTTCCATCACACTCGAAAGGCCCATCATGCACCACCACCATC ACCATAGTAACATCATTAAACCTGGATCTGATGCCGAAAAATCCCAGTTGTCGTTGGGGACCGCT GCCGCGCGCCAATTGGCGACCACCACAAAAAGTGTACCTCAGATGCAGGGCATCAGTTCACGTTG GTTATTGAAACTTCTGCCCTGGGTCCAGGTCTCAGGAGGAGTTTATCGTGTTAACCGCCGTCTTTC CTACGCCGTGGGGGATGGTCGCGTGACCTTCACCACGACAGGTGCCAAGGTGCAGGTCATTCCTC AGGAACTTTGCGAGCTGCCTTTATTACGCAGCTATGATGACGTAGAGGTGTTAACCGCACTGGCC AATCGCTTTGAACAAAAAACGTATAAGGCTGGGGACGTCATTACCGAGGTAGGTAAGGAGGCTG ATTGCATCGTTTTGATCGCGCATGGGAAGGTTAACAAGATCGGTGCTGGCAAGTACGGGGATGCC ACCGTCCTTGGCGTATTGGCCGACGGAGACCATTATTCGTATGAAGCTCTTCTTGAAAGCCAGGA TTATTGGAAGTTTACTGCTAAGGCTGCTACAGCAAGCACAGTTTTGGTTCTTCAACAATCAGATTT TGAGGCCGTTATGGCCCAGAGTCCCAGTCTGCATAAGCATGTCGAACAGTTCAAAGCACGTTCCA AGAAGAAACAGGATACGACGGGCCAAGCCGAGATTGAGTTAGCCGCGGGTCACACTGGCGAACC TGTCTTGCCGGGCACCTATGTTGACTATGAGACGTCACCTCGTGAATATGAACTTTCGGTAGCTCA AACTGTATTACAGATTCATACGCGTGTGGCCGACCTGTTTAATGAACCCATGAACCAAACCGAAC AGCAATTACGTCTTACCGTTGAAGCACTTAAAGAACGTAAAGAACACGAGCTGATTAACAATCGC GAGTTCGGCTTATTACACAACGCTGACTTAAAGCAACGTATCCATACGCGCCGCGGCCCTCCTAC GCCCGACGATATGGACGAGCTGCTGGCCACAGTCTGGAAGGAGCCATCATTCTTTCTTGCTCATC CCCGTGCCATCGCCGCTTTCGGGCAGGAATGCAGCCGCAAGGGTATCTACCCGACTAGTGTCGAC TTCAATGGGAACATGGTCCCAGCCTGGCGTGGTGTTCCCATCTTTCCCTGCAGCAAAATTCCGGTA AGTGACTCACGTACCACTTCTATTATGTTAATGCGTGCGGGTGAGAAAAATCAAGGCGTAATTGG ATTGCATCCCGGAACTATTCCAGACGAAATTGAGGCTGGTTTAAATGTGCGTTTTATGGGAATTA ACGAAAAAGCTATCATTAACTATTTGGTTACCAGTTACTTTTCAGCGGCAGTGTTGGTTCCCGATG CGCTGGGCATCTTAGAGTCAGTTGAAATTGGGCGCGGTGATTAGCCCTCCTCCCGCTCGGATGAC GTGAGGAAATCCACATGCATCATCACCACCATCATACGAAATTCGTTAAGACGGGTGATGACACA CCGCGCTTATCACTTGGAACTGCAGCGGCACGTCAGTTGGCTACGACGACGAAGACCGTCCCCCA GATGCAAGGTATTACACCCCGCTGGCTGCTGCGTATGTTGCCTTGGGTCGAAGTAACGGGTGGGA CCTACCGCGTAAATCGTCGCCTTTCTTACGCTGTAGCGGACGATCGCTTGGTATTCTCAAACATTG GAGCCAAAGTTCAAGTGATCCCTCAGGAATTGCTGAAACTGCCCTTACTTCGTGGGCTGTCAGGC GACGACGAGGTAATCGCAGCACTGGCATCTCAATTCACACAGGCAGAGTATAAAGCGGGTGACG TCATTGTAGAAGCCGGCGCCCCCGCTGAGCATGTATTCTTATTGGCTCACGGCAAAGCACGCAAG CTTCGCGCAGGCCCCTACGGTGCACCAGTAACCTTAGATGTTTTGGCAGACGGTGATCATTTCGGC GACCAAGCCGTCGTGGAGAGCGATGATGTATGGACCTTCACTGTCAAAGCCACTACGCCGTGCAC CGTATTATCACTGCCTCAGCAAGTCTTTGAGGGATTGATCGCACGTTCGGCAACATTGCGCGCTCA GGTAGAGCGTTACCGTGAACGCCTTAAGAAGCCGCAGGATAAGCTGGGGCAGGCTGCCATTCCAT TGGCCGCTGGGCATAGCGGGGAGCCAGACATCCAGGGAGGATACGTTGATTATGAATTAACGCC GCGTGAGTATGAGCTTTCGGTCGCCCAAACCGTTTTGCGCGTTCACACCCGTGTTTCGGACGTGTT TTCAGACCCAATGGACCAGACAGAACAGCAGTTACGTTTAACCATTGAGGCGCTTCGCGAACGTC AAGAACACGAACTGATCAACAATCGTGATTTTGGGCTTTTACACAATGCAGACCTGAAACAGCGT GTTCATACTCGTAATGGGCCGCCGACGCCAGATGATTTGGACGAGCTTTTATCCCGTCGTCGCAA GTCCCGTTTCTTCTTGGCACACCCGCGCACTATTGCTGCTTTTGGGCGCGAGTGCACGCGTCGTGG GATCTATCCAACACCGATCGAAGTGCAAGGGACTCCATGTATGGCCTGGCGCGGAGTTCCTCTGT TACCGTGTGATAAAATTCCTATTACGGAACAACGCACCTCCTCAATTCTTGTTCTGCGTACCGGGC AGGAAGATCAAGGGGTAATCGGTCTTCGCCGCACTGGTCTGCCAGACGAGGTTGAGCCCGGACTT TCCGTTCGCCGCATGGACGTGACCGATAAAGCGATCACAAATTACTTAGTGTCAACCTACTTTTCG GCAGCATTGTTAATCCCGGACGCACTGGGCGTTTTGGAAAATGTTGAGCTTGGTGGCTAG
His-Mx-Enc2: second codon optimization	AAGTATAAGAAGGAGATATACAATGCACCACCACCACCACCATACCAAATTTGTGAAAACTGGA GACGACACCCCGCGACTGTCGCTCGGCACGGCAGCCGCTCGTCAGTTGGCAACGACCACAAAAA CTGTTCCTCAGATGCAGGGAATTACACCTCGCTGGTTACTTCGCATGCTTCCTTGGGTTGAAGTTA CAGGCGGTACATACCGGGTAAACCGCCGTTTATCCTACGCCGTAGCAGATGATCGGCTGGTTTTC AGCAACATAGGGGCGAAAGTGCAGGTTATTCCCCAGGAATTACTGAAATTGCCGTTACTCCGTGG TCTCTCTGGTGATGATGAAGTCATTGCAGCCCTTGCTAGTCAGTTTACGCAGGCTGAATACAAAGC CGGTGACGTAATCGTAGAAGCTGGTGCACCTGCGGAACACGTGTTTCTGCTCGCACATGGTAAAG CACGTAAACTACGGGCTGGACCATATGGTGCTCCAGTGACACTCGATGTTCTGGCAGACGGCGAT CACTTTGGTGATCAGGCGGTAGTGGAATCCGATGACGTCTGGACGTTTACAGTTAAAGCTACGAC

	TCCCTGCACAGTTCTCTCTTTACCGCAGCAGGTTTTCGAAGGTTTGATTGCCCGTAGCGCGACTCT TAGAGCCCAGGTAGAAAGGTACCGGGAACGATTAAAAAAACCTCAGGATAAACTTGGCCAGGCC GCGATCCCGTTAGCAGCAGGTCACTCTGGAGAGCCTGATATCCAGGGTGGCTACGTAGACTATGA ATTGACGCCGCGTGAGTATGAATTGTCGGTTGCGCAGACAGTCTTACGTGTGCATACCCGGGTGA GTGATGTTTTCTCTGATCCTATGGATCAGACAGAACAACAGCTGCGGCTGACAATAGAAGCCTTA CGTGAACGCCAGGAGCATGAACTGATAAACAATAGGGATTTCGGTCTGCTCCATAACGCAGATTT AAAACAACGGGTTCATACCAGAAATGGTCCGCCCACACCGGATGATCTGGACGAACTGTTGAGTC GACGACGTAAAAGCCGTTTCTTCTTAGCACACCCCCGGACTATCGCTGCCTTTGGCAGAGAGTGC ACTCGTCGTGGTATCTATCCAACGCCCATTGAAGTACAGGGCACCCCTTGTATGGCATGGCGCGG TGTTCCTCTACTGCCCTGCGATAAAATCCCGATCACAGAACAGCGTACCAGTAGCATCTTGGTTCT GAGAACAGGCCAGGAAGATCAGGGTGTGATCGGATTACGTCGTACTGGGTTACCTGATGAAGTA GAACCCGGATTAAGTGTTCGAAGGATGGATGTGACAGATAAAGCAATCACGAACTACTTGGTCA GTACATACTTTAGCGCTGCGTTGTTAATCCCCGATGCATTAGGTGTATTAGAGAACGTTGAACTGG GCGGTTGATTAACCTAGGCTGCTGC
His-GGGS-TEV-GeoA	AAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGCACCACCACCATCACCATGGAGGAGG TTCCGAGAATCTTTATTTTCAGGGATCCACCGCGAAAAACAAGCAGCCTTTTGAGTTACCAGACTT CTACGTGCCGTGGCCGGCGCGTTTAAATCCCAATCTGGAAGGTGCTCGCGTACACTCAAAAGCCT GGGCCCGTGAACTGGGGATTATCGGCCGTCCAAAGGATGGATCAGCGCCCGAGATCTGGAGTGA AGCCAAATTTGATGCTATGGATTACGCCTTATTATGCGCTTACACTCACCCGGAAGCGCCCGGAC CGGAATTAGACTTGGTTACCGACTGGTACGTATGGGTTTTCTATTTTGATGACCACTTTCTGGAAT TATATAAACGTCCACAAGACCAAGTTGGTGCTAAGGCCTATTTGGATCGCCTGCCGCTGTTTATGC CAGTCGATCCCGCGGCAACCCCTCCGCCTCCTACTAATCCGGTAGAGGCTGGTCTTCTGGATTTGT GGAATCGTACCGTCCCCTCTCGTTCGATGGCTTGGCGTCGTCGCTTCTTTGAATCCACCAAACACT TGTTGGACGAATCATCATGGGAGCTGTCGAACATCTCTGACCGCCGCGTGTCGAATCCTATTGAA TACATCGAAATGCGTCGCAAAGTCGGTGGCGCCCCTTGGTCCGCCAACTTAGTTGAGCACGCTGT CTTTGCAGAGGTTCCTGATCGCGTCGCCGCTTCACGCCCGATGCGTGTTCTTAAAGACACTTTCTC AGACGCAGTACATCTTCGTAATGATCTTTTCTCGTACGAACGTGAAATTCTTGAGGAAGGAGAGT TGTCCAATGGGGTTTTAGTGATGGAGAAGTTTTTGAATATTTCGCCTCCGTCAGCCGCACATCTTG TAAATGAAGTTCTGACAAGTCGCCTTCAGCAGTTCGAGAATACCGTCTTAACTGAGTTGCCTAGTC TTTTTGTTGAGTTTGGGTTGAACCCCGTAGAACAAGCTCAAGTCCTGACCTATGTGCGCGGCCTGC AGGATTGGCAAAGCGGAGGCCACGAGTGGCACATGCGCAGCTCCCGTTACATGAATAAAGGTTC GGGTGGCGCAGGAGGGTTTTTTTTAGGTCCGAATGGATTGGGCACCTCTGCAGCGCGCCTGCCGC AATCGCCAACTGCCCTGGGTTTGACGCGTTTGAAGAACTTTTCACACGTCCCATACCAGCCTGTTG GTCCCGTTAAACTTCCTAAGTTCTATATGCCGTATAGTACAAAACCCAGTCCTCATCTTGACGCCG CTCGTCGTGATTCCAAGGCGTGGGCACGTCGTATGGGAATGTTAGATGTGCTGCCCGGAGTTCCA GGAGGGTATATCTGGGACGACCATAAATTCGATGTGGCTGACGTAGCATTGTGTGGGGCTTTGAT CCACCCGCATGCGACTGCCGCGCAACTTAACCTGTCGAGCTGTTGGTTAGTTTGGGGAACCTATG CTGATGACTACTTCCCTGCATTTTACGGACACACCAAAGATATGGCCGGTGCGAAGGTTTTTAATG CTCGCTTGGCCCTGTTCGTTCCAGAGGATGCCGGTGCAGTAGTTCCTCCCCCGACGAATCCCGTAG AACGCGGGTTAGCGGACCTTTGGGCGCGTACAACGGAAGGAGTTACGCCTGCAAGCCGCTCCTTA TTCCGTAAGGCAATTCTGGATATGACAGAGTCTTGGGTCTGGGAGTTGGCTAACCAAATCCAGAA CCGCATTCCAGACCCGATTGACTACGTTGAAATGCGTCGTCAAACTTTCGGGTCTGACTTGACAAT GTCACTGTCCCGTTTGGCACACGGAGACGCCTTACCGCCTGAAGTGTTTCATACACGTCCGATCCG CAGCTTGGAGAATTCCGCGGCAGATTACGCCTGTTTAATCAACGATGTGTTCTCGTACCAAAAGG AGATCGAATTTGAAGGAGAGTTGAATAACGGGGTTCTGGTGGTGCAGCGCTTCTTGGACTTGGAT CCGGCACGCGCCGTCAGCGTGGTGAACGATCTTATGACTGCGCGCATGCAGCAATTTGAGTATAT TATCGCGAATGAACTTGAGCCATTAGCGCGTAACTTTAATTTGGATGGGAAGGCACAAGATAAGT TGAAGCAATACGTGCAAAAATTACAGTGGTGGATGTCTGGGGTATTGATTTGGCATCAAACAGTA GATCGCTATAAGGAGTTCGAATTGCGTGCTTCCCGCAAATTAGCGCCCCGTCTGAGTTCGGGCCC CACCGGCCTTGGAACGTCAGCAGCCCGCATCACATCTTTATTTGCGAATCTGCGCTCTGGAGCCTA G

Supplemental Table 3-2: DNA sequences of PCR primers used to construct plasmids.

Primer	DNA Sequence
oMX01	ATTAACCTAGGCTGCTGCCAC
oMX02	CATTGTATATCTCCTTCTTATACTTAACTAATATA
oMX03	GTATAAGAAGGAGATATACAATGCACCACCACCATCACCATAG
oMX04	TGGCAGCAGCCTAGGTTAATCTAATCACCGCGCCCAATTTC
oMX05	GTGGCAGCAGCCTAGGTTAATCTAGCCACCAAGCTCAACAT
oMX06	TATAAGAAGGAGATATACATGCATCATCACCACCATC
oMX07	GTATATCTCCTTCTTATACTAATCACCGCGCCCAATTT
oMX08	CTTTAAGAAGGAGATATACCATGTCCACCGCGAAAAAC
oMX09	ATTATGCGGCCGCAAGCTCTAGGCTCCAGAGCGCAG
oMX10	AGCTTGCGGCCGCATAATG
oMX11	GGTATATCTCCTTCTTAAAGTTAAAC
oMX12	ATGACGAAATTCGTTAAGACGGGTGA
oMX13	GTATATCTCCTTCTTATACTAATCACCG
oMX14	GTATAAGAAGGAGATATACATGCACCACCACCACCACCATA

Supplemental Table 3-3: Primer pairs used to construct the plasmids in this study.

Plasmid Name	Cloning Method	Primers
His-GGGS-TEV-GeoA	N/a – plasmid constructed by TwistBio	N/a – plasmid constructed by TwistBio
His-Mx-Enc1	Gibson Assembly	*Insert:* oMX03 + oMX04 *Vector:* oMX01 + oMX02
His-Mx-Enc2	Gibson Assembly	*Vector:* oMX01 + oMX02
His-Mx-Enc1_Mx-Enc2	Inverse PCR	oMX12 + oMX13
Mx-Enc1_His-Mx_Enc2	Gibson Assembly	*Insert:* oMX14 + oMX05 *Vector:* oMX01 + oMX13
GeoA_His-Mx-Enc1_His-Mx-Enc2	Gibson Assembly	*Insert:* oMX08 + oMX09 *Vector:* oMX10 + oMX11
GeoA_His-Mx-Enc1	Gibson Assembly	*Insert:* oMX03 + oMX04 *Vector:* oMX01 + oMX02

Supplemental Table 3-4: Protein sequences used in this study. Underlined sequences represent expression tags that were kept or removed based on the plasmid design.

Name	Protein Sequence
GeoA (+/- His-tag & TEV site, underlined)	MHHHHHHGGGSENLYFQGSTAKNKQPFELPDFYVPWPARLNPNLEGARVHSKAWAREL GIIGRPKDGSAPEIWSEAKFDAMDYALLCAYTHPEAPGPELDLVTDWYVWVFYFDDHFLE LYKRPQDQVGAKAYLDRLPLFMPVDPAATPPPPTNPVEAGLLDLWNRTVPSRSMAWRRR FFESTKHLLDESSWELSNISDRRVSNPIEYIEMRRKVGGAPWSANLVEHAVFAEVPDRVAA SRPMRVLKDTFSDAVHLRNDLFSYEREILEEGELSNGVLVMEKFLNISPPSAAHLVNEVLT SRLQQFENTVLTELPSLFVEFGLNPVEQAQVLTYVRGLQDWQSGGHEWHMRSSRYMNKG SGGAGGFFLGPNGLGTSAARLPQSPTALGLTRLKNFSHVPYQPVGPVKLPKFYMPYSTKPS PHLDAARRDSKAWARRMGMLDVLPGVPGGYIWDDHKFDVADVALCGALIHPHATAAQL NLSSCWLVWGTYADDYFPAFYGHTKDMAGAKVFNARLALFVPEDAGAVVPPPTNPVER GLADLWARTTEGVTPASRSLFRKAILDMTESWVWELANQIQNRIPDPIDYVEMRRQTFGS DLTMSLSRLAHGDALPPEVFHTRPIRSLENSAADYACLINDVFSYQKEIEFEGELNNGVLV VQRFLDLDPARAVSVVNDLMTARMQQFEYIIANELEPLARNFNLDGKAQDKLKQYVQKL QWWMSGVLIWHQTVDRYKEFELRASRKLAPRLSSGPTGLGTSAARITSLFANLRSGA*
Mx-Enc1 (+/- His-tag, underlined)	MHHHHHHSNIIKPGSDAEKSQLSLGTAAARQLATTTKSVPQMQGISSRWLLKLLPWVQVS GGVYRVNRRLSYAVGDGRVTFTTTGAKVQVIPQELCELPLLRSYDDVEVLTALANRFEQK TYKAGDVITEVGKEADCIVLIAHGKVNKIGAGKYGDATVLGVLADGDHYSYEALLESQD YWKFTAKAATASTVLVLQQSDFEAVMAQSPSLHKHVEQFKARSKKKQDTTGQAEIELAA GHTGEPVLPGTYVDYETSPREYELSVAQTVLQIHTRVADLFNEPMNQTEQQLRLTVEALK ERKEHELINNREFGLLHNADLKQRIHTRRGPPTPDDMDELLATVWKEPSFFLAHPRAIAAF GQECSRKGIYPTSVDFNGNMVPAWRGVPIFPCSKIPVSDSRTTSIMLMRAGEKNQGVIGLH PGTIPDEIEAGLNVRFMGINEKAIINYLVTSYFSAAVLVPDALGILESVEIGRGD*
Mx-Enc2 (+/- His-tag, underlined)	MHHHHHHTKFVKTGDDTPRLSLGTAAARQLATTTKTVPQMQGITPRWLLRMLPWVEVT GGTYRVNRRLSYAVADDRLVFSNIGAKVQVIPQELLKLPLLRGLSGDDEVIAALASQFTQA EYKAGDVIVEAGAPAEHVFLLAHGKARKLRAGPYGAPVTLDVLADGDHFGDQAVVESD DVWTFTVKATTPCTVLSLPQQVFEGLIARSATLRAQVERYRERLKKPQDKLGQAAIPLAA GHSGEPDIQGGYVDYELTPREYELSVAQTVLRVHTRVSDVFSDPMDQTEQQLRLTIEALRE RQEHELINNRDFGLLHNADLKQRVHTRNGPPTPDDLDELLSRRRKSRFFLAHPRTIAAFGR ECTRRGIYPTPIEVQGTPCMAWRGVPLLPCDKIPITEQRTSSILVLRTGQEDQGVIGLRRTGL PDEVEPGLSVRRMDVTDKAITNYLVSTYFSAALLIPDALGVLENVELGG*

3.6 References

1. N. Gerber, H. Lechevalier, Geosmin, an earthy-smelling substance isolated from actinomycetes. *Applied microbiology*. **13,** 935-938 (1965).

2. R. G. Buttery, J. A. Garibaldi, Geosmin and methylisoborneol in garden soil. *Journal of agricultural and food chemistry*. **24,** 1246-1247 (1976).

3. N. N. Gerber, Volatile substances from actinomycetes: their role in the odor pollution of water. *CRC critical reviews in microbiology*. **7,** 191-214 (1979).

4. H. Xu, J. S. Dickschat, A Detailed View on Geosmin Biosynthesis. *ChemBioChem*. **24,** e202300101 (2023).

5. P. Garbeva, M. Avalos, D. Ulanova, G. P. Van Wezel, J. S. Dickschat, Volatile sensation: The chemical ecology of the earthy odorant geosmin. *Environmental Microbiology*. **25,** 1565-1574 (2023).

6. H. Huang, L. Ren, H. Li, A. Schmidt, J. Gershenzon, Y. Lu, D. Cheng, The nesting preference of an invasive ant is associated with the cues produced by actinobacteria in soil. *PLoS Pathogens*. **16,** e1008800 (2020).

7. M. C. Stensmyr, H. K. Dweck, A. Farhan, I. Ibba, A. Strutz, L. Mukunda, J. Linz, V. Grabe, K. Steck, S. Lavista-Llanos, A conserved dedicated olfactory circuit for detecting harmful microbes in Drosophila. *Cell*. **151,** 1345-1357 (2012).

8. L. Zaroubi, I. Ozugergin, K. Mastronardi, A. Imfeld, C. Law, Y. Gélinas, A. Piekny, B. L. Findlay, The ubiquitous soil terpene geosmin acts as a warning chemical. *Applied and Environmental Microbiology*. **88,** e00093-00022 (2022).

9. A. Tran, A. Tang, C. T. O'loughlin, A. Balistreri, E. Chang, D. Coto Villa, J. Li, A. Varshney, V. Jimenez, J. Pyle, C. elegans avoids toxin-producing Streptomyces using a seven transmembrane domain chemosensory receptor. *Elife*. **6,** e23770 (2017).

10. M. Wood, S. Chamaillé-Jammes, A. Hammerbacher, A. M. Shrader, African elephants can detect water from natural and artificial sources via olfactory cues. *Animal Cognition*. **25,** 53-61 (2022).

11. P. G. Becher, V. Verschut, M. J. Bibb, M. J. Bush, B. P. Molnár, E. Barane, M. M. Al-Bassam, G. Chandra, L. Song, G. L. Challis, Developmentally regulated volatiles geosmin and 2-methylisoborneol attract a soil arthropod to Streptomyces bacteria promoting spore dispersal. *Nature Microbiology*. **5,** 821-829 (2020).

12. C. S. Tucker, Off-flavor problems in aquaculture. *Reviews in Fisheries Science*. **8,** 45-88 (2000).

13. T. P. Heil, R. C. Lindsay, Volatile compounds in flavor-tainted fish from the upper Wisconsin River. *Journal of Environmental Science & Health Part B*. **23,** 489-512 (1988).

14. R. T. Lovell, I. Y. Lelana, C. E. Boyd, M. S. Armstrong, Geosmin and musty-muddy flavors in pond-raised channel catfish. *Transactions of the American Fisheries Society*. **115,** 485-489 (1986).

15. P. Darriet, M. Pons, S. Lamy, D. Dubourdieu, Identification and quantification of geosmin, an earthy odorant contaminating wines. *Journal of Agricultural and Food Chemistry*. **48,** 4835-4838 (2000).

16. D. E. Cane, R. M. Watt, Expression and mechanistic analysis of a germacradienol synthase from Streptomyces coelicolor implicated in geosmin biosynthesis. *Proceedings of the National Academy of Sciences*. **100,** 1547-1551 (2003).

17. X. He, D. E. Cane, Mechanism and Stereochemistry of the Germacradienol/Germacrene D Synthase of Streptomyces c oelicolor A3 (2). *Journal of the American Chemical Society*. **126,** 2678-2679 (2004).

18. J. Jiang, X. He, D. E. Cane, Geosmin biosynthesis. Streptomyces coelicolor germacradienol/germacrene D synthase converts farnesyl diphosphate to geosmin. *Journal of the American Chemical Society*. **128,** 8128-8129 (2006).

19. J. Jiang, X. He, D. E. Cane, Biosynthesis of the earthy odorant geosmin by a bifunctional Streptomyces coelicolor enzyme. *Nature Chemical Biology*. **3,** 711-715 (2007).

20. D. E. Cane, X. He, S. Kobayashi, S. Ōmura, H. Ikeda, Geosmin biosynthesis in Streptomyces avermitilis. Molecular cloning, expression, and mechanistic study of the germacradienol/geosmin synthase. *The Journal of antibiotics*. **59,** 471-479 (2006).

21. J. Jiang, C. P. Saint, D. E. Cane, P. T. Monis, Isolation and characterization of the gene associated with geosmin production in cyanobacteria. *Environmental science & technology*. **42,** 8027-8032 (2008).

22. G. G. Harris, P. M. Lombardi, T. A. Pemberton, T. Matsui, T. M. Weiss, K. E. Cole, M. KöKsal, F. V. Murphy Iv, L. S. Vedula, W. K. Chou, Structural Studies of Geosmin Synthase, a Bifunctional Sesquiterpene Synthase with αα Domain Architecture That Catalyzes a Unique Cyclization–Fragmentation Reaction Sequence. *Biochemistry*. **54,** 7142-7155 (2015).

23. M. P. Andreas, T. W. Giessen, Large-scale computational discovery and analysis of virus-derived microbial nanocompartments. *Nat Commun*. **12,** 4748 (2021).

24. T. W. Giessen, Encapsulins. *Annu Rev Biochem*. **91,** 353-380 (2022).

25. M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology 2008 15:9*. **15,** (2008-08-31).

26. T. W. Giessen, B. J. Orlando, A. A. Verdegaal, M. G. Chambers, J. Gardener, D. C. Bell, G. Birrane, M. Liao, P. A. Silver, Large protein organelles form a new iron sequestration system with high storage capacity. *Elife*. **8,** (2019).

27. F. Akita, K. T. Chong, H. Tanaka, E. Yamashita, N. Miyazaki, Y. Nakaishi, M. Suzuki, K. Namba, Y. Ono, T. Tsukihara, A. Nakagawa, The crystal structure of a virus-like particle from the hyperthermophilic archaeon Pyrococcus furiosus provides insight into the evolution of viruses. *J Mol Biol*. **368,** 1469-1483 (2007).

28. D. He, S. Hughes, S. Vanden-Hehir, A. Georgiev, K. Altenbach, E. Tarrant, C. L. Mackay, K. J. Waldron, D. J. Clarke, J. Marles-Wright, Structural characterization of

encapsulated ferritin provides insight into iron storage in bacterial nanocompartments. *Elife*. **5,** (2016).

29. D. He, C. Piergentili, J. Ross, E. Tarrant, L. R. Tuck, C. L. Mackay, Z. Mciver, K. J. Waldron, D. J. Clarke, J. Marles-Wright, Conservation of the structural and functional architecture of encapsulated ferritins in bacteria and archaea. *Biochem J*. **476,** 975-989 (2019).

30. C. A. Mchugh, J. Fontana, D. Nemecek, N. Cheng, A. A. Aksyuk, J. B. Heymann, D. C. Winkler, A. S. Lam, J. S. Wall, A. C. Steven, E. Hoiczyk, A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *EMBO J*. **33,** 1896-1911 (2014).

31. R. J. Nichols, B. Lafrance, N. R. Phillips, D. R. Radford, L. M. Oltrogge, L. E. Valentin-Alvarado, A. J. Bischoff, E. Nogales, D. F. Savage, Discovery and characterization of a novel family of prokaryotic nanocompartments involved in sulfur metabolism. *Elife*. **10,** (2021).

32. R. Benisch, M. P. Andreas, T. W. Giessen, A widespread bacterial protein compartment sequesters and stores elemental sulfur. *Science Advances* (2024-02-02).

33. J. A. Jones, M. P. Andreas, T. W. Giessen, Structural basis for peroxidase encapsulation inside the encapsulin from the Gram-negative pathogen Klebsiella pneumoniae. *Nat Commun*. **15,** 2558 (2024).

34. K. A. Lien, K. Dinshaw, R. J. Nichols, C. Cassidy-Amstutz, M. Knight, R. Singh, L. D. Eltis, D. F. Savage, S. A. Stanley, A nanocompartment system contributes to defense against oxidative stress in. *Elife*. **10,** (2021).

35. R. Rahmanpour, T. D. Bugg, Assembly in vitro of Rhodococcus jostii RHA1 encapsulin and peroxidase DypB to form a nanocompartment. *FEBS J*. **280,** 2097-2104 (2013).

36. Y. Tang, A. Mu, Y. Zhang, S. Zhou, W. Wang, Y. Lai, X. Zhou, F. Liu, X. Yang, H. Gong, Cryo-EM structure of Mycobacterium smegmatis DyP-loaded encapsulin. *Proceedings of the National Academy of Sciences*. **118,** e2025658118 (2021).

37. M. P. Andreas, T. W. Giessen, Heterologous expression and purification of encapsulins in *Streptomyces coelicolor*. *MethodsX*. **9,** 101787 (2022).

38. M. P. Andreas, T. W. Giessen, The biosynthesis of the odorant 2-methylisoborneol is compartmentalized inside a protein shell. *bioRxiv* 2024.2004. 2023.590730 (2024).

39. B. Gust, G. L. Challis, K. Fowler, T. Kieser, K. F. Chater, PCR-targeted Streptomyces gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proceedings of the National Academy of Sciences*. **100,** 1541-1546 (2003).

40. J. S. Dickschat, H. B. Bode, T. Mahmud, R. Müller, S. Schulz, A novel type of geosmin biosynthesis in myxobacteria. *The Journal of organic chemistry*. **70,** 5174-5182 (2005).

41. D. N. Mastronarde, SerialEM: a program for automated tilt series acquisition on Tecnai microscopes using prediction of specimen position. *Microscopy and Microanalysis*. **9,** 1182-1183 (2003).

42. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods*. **14,** 290-296 (2017).

43. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, Highly accurate protein structure prediction with AlphaFold. *nature*. **596,** 583-589 (2021).

44. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*. **30,** 70-82 (2021).

45. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*. **60,** 2126-2132 (2004).

46. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*. **66,** 486-501 (2010).

47. D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. Mccoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, P. D. Adams, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol*. **75,** 861-877 (2019).

48. J. A. Jones, R. Benisch, T. W. Giessen, Encapsulin cargo loading: progress and potential. *J Mater Chem B*. **11,** 4377-4388 (2023).

49. Z. Wang, Y. Xu, J. Shao, J. Wang, R. Li, Genes associated with 2-methylisoborneol biosynthesis in cyanobacteria: isolation, characterization, and expression in response to light. *PLoS one*. **6,** e18665 (2011).

50. E. M. Williams, S. M. Jung, J. L. Coffman, S. Lutz, Pore engineering for enhanced mass transport in encapsulin nanocompartments. *ACS Synthetic Biology*. **7,** 2514-2517 (2018).

51. M. C. Jenkins, S. Lutz, Encapsulin nanocontainers as versatile scaffolds for the development of artificial metabolons. *ACS Synthetic Biology*. **10,** 857-869 (2021).

52. S. Kwon, M. P. Andreas, T. W. Giessen, Pore engineering as a general strategy to improve protein-based enzyme nanoreactor performance. *bioRxiv* (2024).

53. C. M. Jakobson, D. Tullman-Ercek, M. F. Slininger, N. M. Mangan, A systems-level model reveals that 1, 2-Propanediol utilization microcompartments enhance pathway flux through intermediate sequestration. *PLoS computational biology*. **13,** e1005525 (2017).

Chapter 4 Conclusions and Future Directions

4.1 Concluding remarks

This thesis reports the first known encapsulins to form mixed shells with a variable shell composition. I have completed the structural and biochemical characterization of the two-component encapsulin systems from *Streptomyces lydicus* and *Myxococcus xanthus.* For both encapsulin systems, I demonstrated the differential assembly behavior of the two shell components and their ability to form mixed shells in the absence of cargo. Specifically, for the *M. xanthus* encapsulin, I further showed that mixed shells are capable of loading an enzymatically active terpene cyclase cargo. The structures of four encapsulin shell protomers and two encapsulin shells were determined using cryo-EM, demonstrating the presence of the HK97-fold and putative cyclic nucleotide-binding domain on all protomers. Lastly, 3D-classification analysis of the *S. lydicus* cryo-EM data revealed that the two-component mixed encapsulin shell uses an irregular tiling pattern with a variable shell composition.

The two-component encapsulins described in this work are the first reported protein assemblies that use the HK97-fold to preferentially utilize two shell components in an irregular tiling pattern. Only one other assembly built with this protein fold is known to utilize two separate protomers, which is bacteriophage T4 (T4) (1). T4 uses one capsid protein (gp24) to form homo-pentamers and another capsid protein (gp23) to form homo-hexamers. These homo-oligomers then co-assemble to form the T4 icosahedral shell with a triangulation number of T=13 (2, 3). Otherwise, all other structures built using the HK97-fold with a triangulation

number greater than T=1, including all encapsulins, utilize the intrinsic flexibility of the HK97-fold to form both types of oligomers. In contrast to the T4 capsid, the two-component shell protomers do not seem to have distinct and separate structural roles akin to those of the gp23 and gp24 protomers.

For both two-component encapsulins investigated, only one shell protomer in each operon (Sl-Enc2 and Mx-Enc1) was found to be capable of assembling into homogenous shells on its own. The difference in assembly behavior observed for the two encapsulin systems characterized in this work shows that the ability to self-assemble into shells is not a universal feature of all Enc1s or Enc2s, which are named based on their order relative to each other within the Enc operon. Separate alignments of all Enc1 and Enc2 sequences identified two diagnostic motifs that represent the most substantial and conserved differences between all Enc1s and Enc2s. One such motif, termed Diagnostic Motif 1, is a strongly conserved stretch of four residues that is located at the interface between neighboring A-domains within the same pentamer. For Diagnostic Motif 1, the Enc1 consensus sequence is ATVW, and the Enc2 consensus sequence is SRRR. In both *M. xanthus* and *S. lydicus*, the Enc proteins' diagnostic motif sequences are identical to these consensus sequences. Based on the varied assembly behavior observed between the two different encapsulin systems, it is clear that this motif is not solely responsible for the ability of the protomers to assemble shells. However, due to its location at a shell interface, this motif may contribute to the preferential formation of mixed shells.

The ability of a given two-component encapsulin protomer to self-assemble may be related to the rigidity of their N-arms. For the two encapsulin systems discussed in this work, the protomers unable to self-assemble shells individually (Sl-Enc1 and Mx-Enc2) were found to

have resolved (rigid), extended N-arms similar to those found in both Family 2A encapsulin shell structures and some viruses containing the HK97-fold (4-7). In the 3D classification analysis of the *S. lydicus* two-component shell dataset, two different states of the two-fold pore were observed for Sl-Enc1. One state showed the absence of density corresponding to N-arms in the two-fold pore, with the overall pore in the "open" conformation. The other state contained the Sl-Enc1 N-arms threaded through the pore, nearly completely obstructing it. For two-fold symmetry axes composed of Sl-Enc2 protomers, no N-arms were resolved in any conformation. Aside from the two-component encapsulins described in this work, only one other structure of a Family 2B encapsulin has been determined, which is from a single-component operon (8). This protomer structure lacks rigid N-arms, leaving the two-fold pores open. In all three cases, defined N-arms are absent from Family 2B encapsulin shells that can self-assemble using only a single shell protein. Overall, this suggests that the presence of defined N-arms might be restricting the ability of Sl-Enc1 and Mx-Enc2 protomers to assemble into uniform encapsulin shells. One possible explanation is that the unfavorable angles between protomer-protomer interfaces could propagate through the structure, preventing the formation of a closed, icosahedral compartment. However, this dissertation provides evidence that mixed shell assemblies are preferred and suggests that the N-arms may be important for stabilizing the two-component shell structure, thus promoting mixed shell formation. Additionally, the N-arms may be functionally important for the regulation of the cargo enzymes by controlling the pore size and therefore altering metabolite flux through the shell.

Due to the variability observed in mixed shell assembly, it is difficult to pinpoint how exactly the encapsulin shells benefit from the incorporation of a second shell protein. Every other characterized encapsulin has been shown to require only a single shell protein, many of which

are stable without utilizing an extended N-arm (9-11). Some of these encapsulins have even been shown to have dynamic pores that do not require the use of an additional shell protein (12, 13). Comparatively, bacterial microcompartments (BMCs), a separate class of protein-based organelles, utilize between three and eleven distinct shell components (14). The redundancy of shell proteins is thought to be beneficial by allowing the BMC to fine-tune its shell properties based on the dynamic needs of the host cell, specifically by mediating interactions with particular cargo proteins and altering pore properties to exert greater control over metabolite flux through the shell (15-17). Though the use of a second shell protein in the encapsulin may provide overall structural benefits, it is also very likely that its incorporation is important for the regulation or biological function of the overall encapsulin. Ultimately, further experimentation is needed to investigate the potential advantages conferred by a two-component shell on its cargo.

4.2 Future work

4.2.1 Future experiments with the M. xanthus two-component encapsulin

Additional experimentation is needed to complete the characterization of the *M. xanthus* two-component encapsulin as described in Chapter 3. To better understand the assembly principles guiding mixed shell formation, 3D classification analysis of the existing cryo-EM datasets is necessary. This technique was previously used to analyze the *S. lydicus* mixed shell datasets in Chapter 2, where we observed all possible interactions between two pairs of shell protomers at every symmetrical interface, demonstrating a random tiling pattern. Ideally, this analysis would also allow us to determine if the same observation is true for the cargo-loaded *M. xanthus* two-component encapsulin or if cargo loading alters the shell composition and assembly

preferences. 3D classification analysis may also enable an improvement in the resolution of the targeting peptide densities observed at the three-fold pore. Higher-resolution data would hopefully allow us to identify specific interactions between residues on the shell interior and the GeoA targeting peptide.

In this study, only preliminary investigations into the enzymatic activity of GeoA were conducted. The results from this initial experiment demonstrated no significant activity difference between GeoA samples with different shell protomer compositions. However, only two non-zero substrate concentrations were tested. Though we are still unsure of the overall benefit provided by a two-component shell, one hypothesis is that the cellular advantage lies in the mixed shell's ability to confer additional regulation on the encapsulated enzyme, potentially through the two-fold pores. To further test the possibility that differences in shell composition, and therefore, two-fold pore accessibility, alter the encapsulated enzyme's activity, a complete kinetic analysis should be completed with a broader range of substrate concentrations to determine the kinetic parameters. If no significant differences in enzyme activity are observed at other substrate concentrations, then the advantage of encapsulating GeoA in a two-component shell might lie elsewhere, or it may not be specifically advantageous.

The last set of experiments that I propose is aimed at better understanding the role of the CBD. This domain shares its overall fold with many other proteins known to bind cAMP or cGMP, including gated ion channels, transcriptional regulators, and protein kinases (18-20). The *E. coli* Catabolite Activator Protein (CAP) is one of the most well-studied examples of this protein fold and contains both a cAMP-binding domain and a DNA-binding domain (21). Upon cAMP binding to CAP, the cAMP binding domain undergoes a conformational shift that is propagated through the rest of the protein complex, ultimately resulting in the repositioning of

the DNA-binding domain, enabling it to bind DNA (21). In this example, cAMP binding to CAP serves to allosterically regulate the activity of a separate domain. Based on this, we hypothesize that the binding of a small molecule to the CBDs displayed on the shell's exterior may induce a similar conformational change that propagates through the rest of the shell structure and serves to regulate cargo enzyme function. Since the CBDs in all Family 2B encapsulin structures are positioned directly above the two-fold pore, we specifically predict that a conformational change induced by small molecule binding will influence the size or accessibility of this major pore, thus regulating the cargo enzyme activity by controlling substrate access. It was recently shown that cAMP does not bind to the CBD of the Family 2B encapsulin from *S. griseus* (8), though another small molecule may.

To test this hypothesis, we plan to repeat the previously established enzyme assays in the presence of other small molecules and see if there is a difference in enzymatic activity resulting from encapsulin structural changes. Specifically, we plan to test small molecules found in the methylerythritol phosphate (MEP) isoprenoid biosynthesis pathway, which most bacteria use to generate the isoprenoid precursors DMAPP and IPP (22, 23).

4.2.2 Proposed studies of other two-component encapsulins

Though this work has provided broad insights into the structures and assembly patterns of two-component encapsulins, many new questions have been posed over the course of this investigation. Above all, the biggest question remains: what is the benefit of a two-component shell? Further complicating the answer to this question is the significant diversity of two-component encapsulin operons (24). Five distinct operon types have been identified based on cargo identities and co-occurrence, including four unique enzymatic cargoes. Furthermore, these

two-component encapsulin operons are found in five bacterial phyla. This significant phylogenetic diversity indicates that there may not be a straightforward answer to this overarching question and that the benefit of a two-component shell may vary significantly depending on the bacterial phyla in which it is found and/or its enzymatic cargo. To address this fundamental question, comprehensive studies on the structure, assembly, and biological function of other two-component encapsulins are necessary.

It is still unknown why some protomers in two-component operons cannot assemble shells, despite having strong structural similarity to other HK97-fold proteins. Structural studies of new two-component encapsulin systems may allow for the elucidation of specific sequences or structural features that are specific to protomers that are unable to assemble on their own. The structure determination and 3D classification analysis of more mixed shell encapsulins may also highlight different assembly patterns and incorporation ratios of the two proteins, if they form mixed shells at all. For example, one operon type is characterized by the presence of two putative cargoes (a terpene cyclase and a polyprenyl transferase) alongside the two encapsulin shell components. It is possible that, in such operons, two homogenous types of encapsulins are formed, with each cargo interacting exclusively with one encapsulin protomer. The investigation of new two-component encapsulins with different cargoes would allow us to determine if there are structural features unique to encapsulins that use a particular cargo. Experiments aimed at elucidating the cargo-loading mechanism for other two-component encapsulin cargoes would provide additional insight into the potential benefits of a mixed shell. One hypothesis is that some cargo proteins may selectively interact with only one of the two shell protomers, thus promoting overall encapsulin stability or improved enzyme activity.

In addition to conducting structural analyses of other two-component encapsulins, it is also essential to conduct physiological studies *in vivo* to understand the cellular benefits and uses of mixed encapsulin shells. In general, very few published works on encapsulins have studied them in the context of their native organism (4, 5, 25-27). Using techniques such as RT-qPCR, the physiological conditions in which encapsulin components are upregulated can be determined (4, 5). To specifically determine if both shell proteins are essential for the proper functioning of two-component encapsulins, a series of knockout mutants can be generated in which individual operon components are removed from the native organism. Then, the cells can be monitored under various growth conditions to determine if there are significant differences in growth rate or cell morphology between the knockout strains.

However, one particular challenge of some proposed experiments is the lack of tools currently available to conduct single-particle analysis of encapsulin samples, resulting from their small particle size and subtle differences in mass and charge. Though bulk measurement methods can be utilized to approximate the relative levels of the two shell proteins, they do not provide information regarding the level of protomer incorporation into each encapsulin or if the protomer composition is consistent for all shells in the sample. Fortunately, advancements in native mass spectrometry instrumentation and methodology may soon enable its use for single-particle analysis of mixed shell encapsulins, specifically to determine the shell protomer stoichiometries and cargo loading ratios (28, 29).

4.3 References

1. L. Onorato, B. Stirmer, M. K. Showe, Isolation and characterization of bacteriophage T4 mutant preheads. *J Virol*. **27,** 409-426 (1978).

2. Q. Fang, W. C. Tang, A. Fokine, M. Mahalingam, Q. Shao, M. G. Rossmann, V. B. Rao, Structures of a large prolate virus capsid in unexpanded and expanded states generate insights into the icosahedral virus assembly. *Proc Natl Acad Sci U.S.A.* **119,** e2203272119 (2022).

3. Z. Chen, L. Sun, Z. Zhang, A. Fokine, V. Padilla-Sanchez, D. Hanein, W. Jiang, M. G. Rossmann, V. B. Rao, Cryo-EM structure of the bacteriophage T4 isometric head at 3.3-Å resolution and its relevance to the assembly of icosahedral viruses. *Proc Natl Acad Sci U S A*. **114,** E8184-E8193 (2017).

4. R. Benisch, M. P. Andreas, T. W. Giessen, A widespread bacterial protein compartment sequesters and stores elemental sulfur. *Science Advances* (2024-02-02).

5. R. J. Nichols, B. Lafrance, N. R. Phillips, D. R. Radford, L. M. Oltrogge, L. E. Valentin-Alvarado, A. J. Bischoff, E. Nogales, D. F. Savage, Discovery and characterization of a novel family of prokaryotic nanocompartments involved in sulfur metabolism. *Elife*. **10,** (2021).

6. M. M. Suhanovsky, C. M. Teschke, Nature's favorite building block: Deciphering folding and capsid assembly of proteins with the HK97-fold. *Virology*. **479-480,** 487-497 (2015).

7. W. R. Wikoff, L. Liljas, R. L. Duda, H. Tsuruta, R. W. Hendrix, J. E. Johnson, Topologically linked protein rings in the bacteriophage HK97 capsid. *Science*. **289,** 2129-2133 (2000).

8. M. P. Andreas, T. W. Giessen, The biosynthesis of the odorant 2-methylisoborneol is compartmentalized inside a protein shell. *bioRxiv* 2024.2004. 2023.590730 (2024).

9. M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, N. Ban, Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology 2008 15:9*. **15,** (2008-08-31).

10. J. A. Jones, M. P. Andreas, T. W. Giessen, Structural basis for peroxidase encapsulation inside the encapsulin from the Gram-negative pathogen Klebsiella pneumoniae. *Nat Commun*. **15,** 2558 (2024).

11. T. W. Giessen, B. J. Orlando, A. A. Verdegaal, M. G. Chambers, J. Gardener, D. C. Bell, G. Birrane, M. Liao, P. A. Silver, Large protein organelles form a new iron sequestration system with high storage capacity. *Elife*. **8,** (2019).

12. J. A. Jones, M. P. Andreas, T. W. Giessen, Exploring the extreme acid tolerance of a dynamic protein nanocage. *Biomacromolecules*. **24,** 1388-1399 (2023).

13. J. Ross, Z. Mciver, T. Lambert, C. Piergentili, J. E. Bird, K. J. Gallagher, F. L. Cruickshank, P. James, E. Zarazúa-Arvizu, L. E. Horsfall, K. J. Waldron, M. D. Wilson, C. L. Mackay, A. Baslé, D. J. Clarke, J. Marles-Wright, Pore dynamics and asymmetric cargo loading in an encapsulin nanocompartment. *Sci Adv*. **8,** eabj4461 (2022).

14. M. Sutter, M. R. Melnicki, F. Schulz, T. Woyke, C. A. Kerfeld, A catalog of the diversity and ubiquity of bacterial microcompartments. *Nat Commun*. **12,** 3809 (2021).

15. M. Sutter, B. Greber, C. Aussignargues, C. A. Kerfeld, Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell. *Science*. **356,** 1293-1297 (2017).

16. C. Fan, S. Cheng, S. Sinha, T. A. Bobik, Interactions between the termini of lumen enzymes and shell proteins mediate enzyme encapsulation into bacterial microcompartments. *Proc Natl Acad Sci U S A*. **109,** 14995-15000 (2012).

17. J. N. Kinney, A. Salmeen, F. Cai, C. A. Kerfeld, Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly. *J Biol Chem*. **287,** 17729-17736 (2012).

18. M. Gomelsky, M. Y. Galperin, Bacterial second messengers, cGMP and c-di-GMP, in a quest for regulatory dominance. *The EMBO journal*. **32,** 2421-2423 (2013).

19. H. M. Berman, L. F. Ten Eyck, D. S. Goodsell, N. M. Haste, A. Kornev, S. S. Taylor, The cAMP binding domain: an ancient signaling module. *Proceedings of the National Academy of Sciences*. **102,** 45-50 (2005).

20. N. Kannan, J. Wu, G. S. Anand, S. Yooseph, A. F. Neuwald, J. C. Venter, S. S. Taylor, Evolution of allostery in the cyclic nucleotide binding module. *Genome biology*. **8,** 1-13 (2007).

21. N. Popovych, S.-R. Tzeng, M. Tonelli, R. H. Ebright, C. G. Kalodimos, Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proceedings of the National Academy of Sciences*. **106,** 6927-6932 (2009).

22. M. Rohmer, From molecular fossils of bacterial hopanoids to the formation of isoprene units: discovery and elucidation of the methylerythritol phosphate pathway. *Lipids*. **43,** 1095-1107 (2008).

23. B. Marshall, K. Amritkar, M. Wolfe, B. Kaçar, R. Landick, Evolutionary flexibility and rigidity in the bacterial methylerythritol phosphate (MEP) pathway. *Frontiers in Microbiology*. **14,** 1286626 (2023).

24. M. P. Andreas, T. W. Giessen, Large-scale computational discovery and analysis of virus-derived microbial nanocompartments. *Nat Commun*. **12,** 4748 (2021).

25. M. P. Andreas, T. W. Giessen, Heterologous expression and purification of encapsulins in *Streptomyces coelicolor*. *MethodsX*. **9,** 101787 (2022).

26. D. Kim, J. Choi, S. Lee, H. Hyun, K. Lee, K. Cho, Mutants defective in the production of encapsulin show a tan-phase-locked phenotype in Myxococcus xanthus. *journal of microbiology*. **57,** 795-802 (2019).

27. K. A. Lien, K. Dinshaw, R. J. Nichols, C. Cassidy-Amstutz, M. Knight, R. Singh, L. D. Eltis, D. F. Savage, S. A. Stanley, A nanocompartment system contributes to defense against oxidative stress in. *Elife*. **10,** (2021).

28. D. J. Reid, S. Thibert, M. Zhou, Dissecting the structural heterogeneity of proteins by native mass spectrometry. *Protein Science*. **32,** e4612 (2023).

29. V. Yin, P. W. Devine, J. C. Saunders, A. Barendregt, F. Cusdin, A. Ristani, A. Hines, S. Shepherd, M. Dembek, C. L. Dobson, Stochastic assembly of biomacromolecular complexes: impact and implications on charge interpretation in native mass spectrometry. *Chemical Science*. **14,** 9316-9327 (2023).