THE BIOPHYSICAL CHARACTERIZATION OF THE A-CARBOXYSOME AND A MINIMAL CARBOXYSOME

KRISTI LEE TURTON
Bachelor of Science, **University of Lethbridge, 2019**

A thesis submitted
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

BIOCHEMISTRY

Department of Chemistry and Biochemistry
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

THE BIOPHYSICAL CHARACTERIZATION OF THE A-CARBOXYSOME AND A MINIMAL CARBOXYSOME

KRISTI LEE TURTON

Date of Defense: July 6, 2022

Dr. H.-J. Wieden	Professor	Ph.D.
Dr. T. Patel	Associate Professor	Ph.D.
Thesis Co-Supervisors		
Dr. U. Kothe		
Thesis Examination Committee Member	Professor	Ph.D.
Dr. E. Schultz		
Thesis Examination Committee Member	Professor	Ph.D.
Dr. J.-D. Hamel	Assistant Professor	Ph.D.
Chair, Thesis Examination Committee		

DEDICATION

This thesis is dedicated to my family and the many friends that have been with me during my academic career. The major reason I was able to persevere through this journey is because of the people in my life. I am unable to truly express the gratitude and love I have for every one of you.

ABSTRACT

Cells use organelles, micelles, and proteinaceous structures to compartmentalize cellular processes. Compartmentalization has the benefit of providing metabolic control, flux, and protective storage of biomaterials. Carboxysomes are proteinaceous compartments that enclose proteins and metabolites involved in the metabolism of carbon dioxide. Due to the nature and biophysical properties of the carboxysome, use in biotechnology applications is being pursued.

Using biophysical characterization methods such as size exclusion chromatography-Multi-Angle Light Scattering (SEC-MALS), Transmission Electron Microscopy (TEM), and Analytical Ultracentrifugation (AUC), structural information of wild type carboxysomes expressed in *E. coli* and carboxysomes engineered to have a minimal set of proteins, known as the minimal carboxysome are obtained. AUC and TEM provided detail with respect to size and mass distributions of individual particles. Data in this thesis indicate the need for further optimization of existing carboxysome purification strategies.

PREFACE

The work described in this thesis was conducted solely by me with the following exceptions. Mass spectrometry experiments and basic data processing was conducted by Fan Mo and Ying Lao at the University of Lethbridge and University of Manitoba, respectively. Transmission Electron Microscopy imaging was conducted by Dr. Zev Ripstein and Timothy Vos at the University of Manitoba. Analytical Ultracentrifugation experiments were performed by myself with the assistance of Amy Henrickson; basic analysis and data processing was conducted by Amy Henrickson and Dr. Borries Demeler at the University of Lethbridge. The methods section on the analytic ultracentrifugation experiments was written and provided by Amy Henrickson.

Chapter 1 contains figures and content from a review targeted for submission to ACS Synthetic Biology, written by myself and Dr. H.-J. Wieden titled, “Engineering the Carboxysome: A Roadmap Towards Novel Applications.”

ACKNOWLEDGMENTS

There is a lot of people to thank for helping me on this journey. Firstly, thank you to my supervisor Dr. Hans-Joachim Wieden. I appreciate how much I have learned and have loved every opportunity I have received while being in your lab group. Thank you to my co-supervisor Dr. Trushar Patel and my committee members Dr. Ute Kothe and Dr. Elizabeth Schultz for all the guidance and support.

I would also like to thank Corteva Agrisciences and Rory Degenhardt for the collaboration in the deepYellow challenge, Dr. Hasan Uludağ and the RJH Biosciences team for my Industry Internship, and finally to Dr. Laura Keffer-Wilkes for being the greatest RNA Innovation program coordinator one could ask for (although you have supported me in infinitely more ways since my start at the UofL). Thank you to Dr. Soumya Deo who worked on this project previously. Thank you to Fan Mo, Tyler Mrozowich, and Amy Henricksen for help with experiments. Thanks to Luc Roberts, Taylor Sheahan, Dora Capatos, and Jessica Semmelrock for reference protein samples and/or supporting experimental data. Thank you to previous and present UofL iGEM members: Luke Saville, Marcel Ezra Michailides, Catrione Lee, Chris Isaac, and Sydnee Calhoun- Wiki Freeze and trips to Original Joes would have never been the same without you all. Let me also not forget my favourite Neuroscientist Emily Hagens: I will save the teasing for another time. Thank you to the Wieden, Kothe, Patel, and Demeler lab members past and present: My lab family and human Wikipedia's. Thank you to Nicole Perl, Anthony Devasahayam, Simmone D'Souza, Davinder Kaur, and Dr. Harland Brandon for being such good friends and being there for me when I need it most.

Thank you to my family: Your constant stream of support, strength, and laughter are one of the things that made the hard lab days and the weekend work sessions easier. I love you.

And finally, thank you to Cheyanne Leckie and Fabian Rohden: My confidants, unofficial therapists, comedians, “side-kicks”, part-time personal chefs, and best friends. I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DTT: 1,4-Dithiothreitol
2-OG: 2-oxo-glutarate
2-PG: 2-phosphoglycolate
PGP: 2-phosphoglycolate phosphatase
3-PGA: 3-phosphoglycerate
Ala: Alanine
AUC: Analytical Ultracentrifugation
AaLS: Lumazine Synthase from *Aquifex Aeolicus*
BMC: Bacterial Microcompartment
Bp: Base pairs
CBB: Calvin-Benson-Bassham
CA: Carbonic Anhydrase
CB: Carboxysome
α-CB: α-Carboxysome
β-CB: β-Carboxysome
CAT: Catalase
CDS: Coding Sequence
CV: Column Volume
Cryo-EM: Cryogenic-electron microscopy
GLYK: D-glycerate 3-kinase
DLS: Dynamic Light Scattering
dT: double terminator
EDTA: Ethylenediaminetetraacetic acid
EP: Encapsulation Peptide
FDH: Formate dehydrogenase
Glu: Glutamine
GDC: Glycine decarboxylase
AGT1: Glyoxylate aminotransferase
GCD: Glycolate dehydrogenase
GOX: Glycolate oxidase
GGT: Glyoxylate aminotransferase
GCL: Glyoxylate carboligase
GXO: Glyoxylate oxidase
GdnHCl: Guanidinium chloride
R_H: Hydrodynamic radius
HPR1/HPR2: Hydroxypyruvate reductase
IAA: Iodoacetamide
IPTG: Isopropyl ß-D-1-thiogalactopyranoside
kDa: Kilodalton
LB: Luria Broth
MAL: malate
MALS: Multi-Angle Light Scattering

mRBS: medium Ribosomal Binding Sight
CH2-THF: Methylenetetrahydrofolate
mCB: Minimal Carboxysome
MDH: malate dehydrogenase
NADP-ME: NADP-malic enzyme
OAA: Oxaloacetic acid
OD: Optical Density
ODC: Oxalate decarboxylase
PA: Polyacrylamide
PEP: Phosphoenolpyruvate
PEPC: Phosphoenolpyruvate Carboxylase
PMSF: phenylmethylsulfonyl fluoride
PNC: Protein Nanocompartment
PPDK: Pyruvate phosphate dikinase
Pyr: Pyruvate
RI: Refractive Index
rpm: Revolutions per minute
RBS: Ribosomal Binding Site
RuBP: Ribulose-1;5-bisphosphate
RuBisCO: Ribulose-1;5-bisphosphate carboxylase oxygenase
SHM: Serine transhydromethyltransferase
SEC-MALS: Size exclusion chromatography- Multi-angle Light Scattering
SEC: Size exclusion chromatography
S: Svedberg unit
GFP (+36): Supercharged GFP
TSR: Tartronate semialdehyde reductase
THF: Tetrahydrofolate
TEM: Transmission Electron Microscopy
V: Volts
V_0: Void Volume

CHAPTER 1: INTRODUCTION

1.1 Introduction

Bacterial microcompartments (BMCs) are naturally occurring, large, macromolecular structures that form a "shell" which can assemble into well-defined compartments capable of encapsulating other biomolecules and metabolites. BMCs naturally encapsulate toxic molecules that are part of cellular pathways, typically associated with producing energy (Kerfeld *et al.*, 2010). The α-carboxysome (α-CB) is a BMC commonly found in cyanobacteria or proteobacteria that maintains high concentrations of CO_2 in order to increase the efficiency of the encapsulated protein RuBisCO (ribulose-1,5-bisphosphate carboxylase/oxygenase) which uses CO_2 to convert ribulose-1,5-bisphosphate into 3-phosphoglycerate (Rae *et al.*, 2013).

The α-CB from the cyanobacteria *Halothiobacillus neapolitanus* (the most studied α-CB to date) consists of eleven individual proteins (including eight shell proteins) forming an icosahedral structure (Bonacci *et al.*, 2012; Rae *et al.*, 2013). The α-carboxysome's ability to adapt to different volumes of cargo and its unique semi-permeable surface, enables the storage and delivery of small molecules.

The use of proteinaceous particles, such as the α-CB in new applications, expands technology beyond the ever-popular liposome or polymer-based compartmentalization systems. Currently, complex α-CB particles, besides having unique biophysical characteristics and functions, also have limitations; we do not have a full understanding on the mechanisms of function and formation of the resulting shells. What is known, is that the variability in assembly and encapsulation of cargo can cause differences of shell protein stoichiometry during assembly, suggesting that the resulting α-CB particles are heterogeneous in size and structure (Schmid *et*

al., 2006; Y. Sun *et al.*, 2019). One way to stream-line the studies regarding α-CB heterogeneity is to use readily accessible biophysical methods to accurately assess structure and size. Biophysical characterization is a benefit for both, verifying wild type α-CBs for subsequent analysis, or for accessing engineered BMCs (Figure 1.1). Furthermore, biophysical approaches also complement methods that are less labour intensive, are faster, and are able to provide additional information such as molecular weight and size.

In this thesis I utilize several biophysical methods, Analytical Ultracentrifugation (AUC), Transmission Electron microscopy (TEM), and Multi-Angle Light Scattering (MALS) for characterizing purified α-CBs and minimal carboxysomes (mCBs). The goal of this thesis is to 1) determine which biophysical methods can be used to study α-CB proteins *in vitro* and 2) to better analyze the heterogeneous α-CB particles, that as of this point in time have no reported characterization in the literature, when expressed in *E. coli*. To this end, the wild type CB from *H. neapolitanus* was analyzed by comparing it against an mCB that contains only two of the original eight shell proteins. The mCB has not been fully characterized *in vitro* or when expressed in *E. coli.*

A full characterization of the mCBs is the first step towards their potential applications in biotechnology. In agriculture for example, carboxysomes could be inserted into the genome of C3 plants host plant to improve their capacity for carbon fixation. Common genetic engineering practices to introduce new genes into plant genomes can be challenging (Yin *et al.*, 2017) and therefore, using mCBs reduces cost and time, as well as reducing the metabolic burden on the host plant. In manufacturing, using mCBs to compartmentalize novel proteins and metabolites in bacteria such as *E. coli* can improve production of therapeutics or other valuable products. The structure, shape, size, and function of a mCB could be fine-tuned towards each specific

application. However, this requires standardized biophysical methods to characterize mCBs and CBs in general.

The work presented will not only provide details regarding different α-CB or mCB structures (the sample heterogeneity) to be used for future applications but will also support the use of biophysical methods to improve studies of the α-CB or similar proteinaceous particles. Additional work with a monomeric proteinaceous particle, lumazine synthase from *Aquifex aeolicus*, was performed with the goal of biophysical characterization. Due to limited progress, this project is explained in detail in Appendix II.

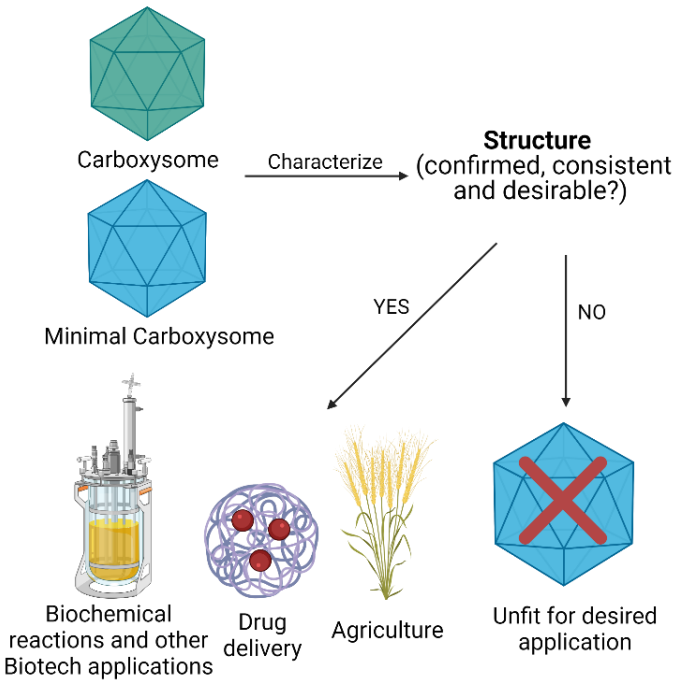


Figure 1.1 Overview of thesis goals. The α-CB and minimal carboxysome, once purified from the cell, were characterized using biophysical methods. Through the results of these methods, the structure and other features of these BMCs were determined, allowing assessment of their use for specific applications. Created with BioRender.com

CHAPTER 2: ENGINEERING THE CARBOXYSOME: A ROADMAP

Select content and figures from this section were prepared for submission to the ACS Synthetic Biology Journal

2.1 Bacterial Microcompartments

Bacterial microcompartments (BMC) are proteinaceous organelles that naturally compartmentalize proteins and/or metabolites that are a part of specific metabolic processes. BMCs are grouped based on the type of protein that they compartmentalize as a part of the metabolic process, both anabolic or catabolic (Kirst *et al.*, 2019). To date, 68 different types of BMCs have been described, most being identified by metagenomic analysis (Sutter *et al.*, 2021). Among these types of BMCs, the α-CB is one of the most studied BMC to date and is therefore of particular interest for biophysical analysis.

Most BMC proteins are expressed on operons with many BMCs having additional select proteins encoded on satellite loci (Sutter *et al.*, 2021). As BMCs are essential for the respective metabolic processes, they are highly conserved and are expressed as operons to allow for tight expression control (A. Liu *et al.*, 2009). BMC shell proteins also exhibit high structural conservation, which allows for classification via protein family domains (pfam domains), corresponding to their type; hexameric, trimeric, and pentameric (Figure 2.1) (Gonzalez-Esquer *et al.*, 2016; Kerfeld *et al.*, 2016; Kirst *et al.*, 2019; Lee *et al.*, 2019).

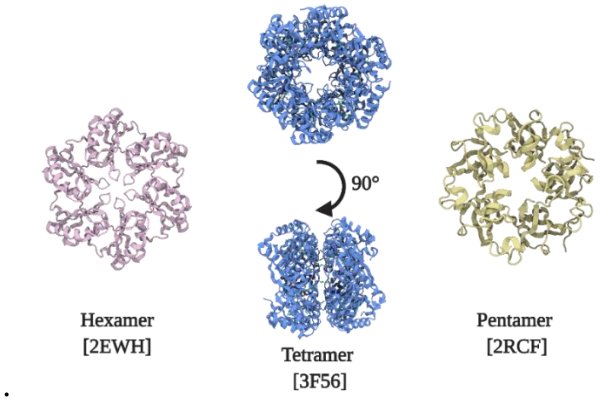


Figure 2.1 Crystal structures representing the three types of shell proteins that can constitute a BMC. Representative BMC proteins are Csos1A from *Halothiobacillus neapolitanus* (2EWH, PDB accession code), Csos1D from *Prochlorococcus Marinus* MED4 (3F56) and OrfA, also known as Csos4A, from *Halothiobacillus neapolitanus* (2RCF). Trimeric shell proteins are rotated 90˚ to show the full structure. Created with BioRender.com

2.2 Carboxysomes

The carboxysome (CB) is currently the only identified catabolic BMC (Kirst *et al.*, 2019). Found in phototrophic bacteria and cyanobacteria species, the CB encapsulates the protein Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) which is an enzyme associated with the Calvin-Benson-Bassham (CBB) cycle. The CBB cycle, also present in C3 and C4 plants, allows for the fixation of carbon in plants (Figure 2.2) (Espie *et al.*, 2011; Kerfeld *et al.*, 2015; Mallmann *et al.*, 2014; Orf *et al.*, 2016). RuBisCO is an inefficient enzyme where binding carbon dioxide is outcompeted by oxygen binding at increasing oxygen concentrations and higher temperatures (Espie *et al.*, 2011). Through compartmentalization via the carboxysome, high CO_2 levels can be maintained to increase RuBisCO activity by reducing the probability of an oxygen side reaction.

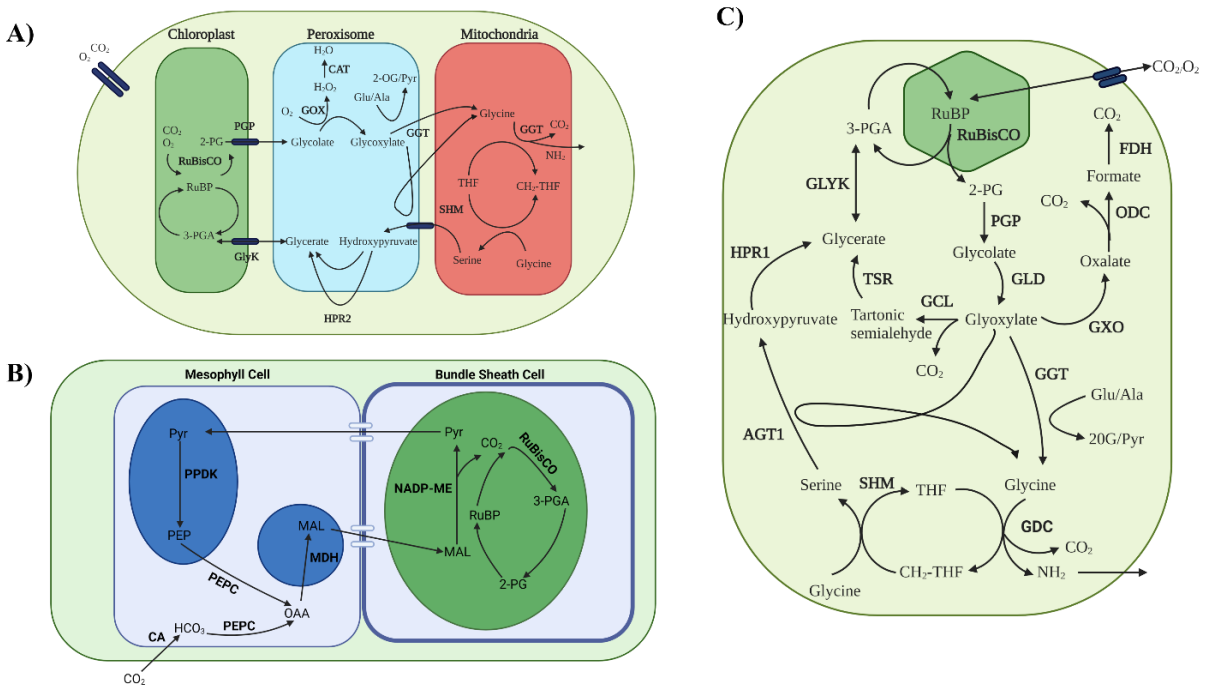


Figure 2.2 The Calvin cycle in phototrophic organisms. A) The Calvin cycle in C3 and **B)** C4 plants **C)** The Calvin cycle in Cyanobacterial species. Within C3 plants, RuBisCO is found in an organelle called chloroplast. In Cyanobacteria, RuBisCO is encapsulated in the carboxysome which is located in the cytosol. Associated CBB enzymes are in bold. Metabolites and by-products are not bolded (adapted from (Mallmann *et al.*, 2014; Orf *et al.*, 2016). The associated protein and metabolites described in the figure are as follows: 2-OG (2-oxo-glutarate), 2-PG (2-phosphoglycolate), 3-PGA (3-phosphoglycerate), AGT1 (glyoxylate aminotransferase), Ala (alanine), CAT (catalase), CH2-THF (methylenetetrahydrofolate), FDH (formate dehydrogenase), GCL (glyoxylate carboligase), GDC (glycine decarboxylase), GGT (glyoxylate aminotransferase), GLD (glycolate dehydrogenase), Glu (glutamine), GLYK (D-glycerate 3-kinase), GOX (glycolate oxidase), GXO (glyoxylate oxidase), HPR1/HPR2 (hydroxypyruvate reductase), MAL (malate), MDH (malate dehydrogenase), NADP-ME (NADP-malic enzyme), OAA (oxaloacetic acid), ODC (oxalate decarboxylase), PEP (Phosphoenolpyruvate), PEPC (Phosphoenolpyruvate carboxylase), PGP (2-phosphoglycolate phosphatase), PPDK (pyruvate phosphate dikinase), Pyr (pyruvate), RuBisCO (ribulose-1;5-bisphosphate carboxylase oxygenase), RuBP (ribulose-1;5-bisphosphate), SHM (serine transhydromethyltransferase), THF (tetrahydrofolate), TSR (tartronate semialdehyde reductase). Created with BioRender.com.

Carbon fixation utilizing the CB proceeds as follows: 1) CO_2 and HCO_3^- enter the cell and both diffuse through the carboxysome shell, 2) HCO_3^- is converted to CO_2 by Carbonic anhydrase. 3) CO_2 is bound by RuBisCO along with Ribulose-1,5-bisphosphate (RuBP) within the CB to produce two 3-phosphoglycerate molecules. 4) 3-phosphoglycerate (3-PGA) leaves the

carboxysome and moves either through the rest of the carbon fixation/photorespiration cycle or is funneled into other metabolic processes in the cell such as the glycolysis pathway (Figure 2.3)(Espie *et al.*, 2011; Kerfeld *et al.*, 2016; So *et al.*, 2004).

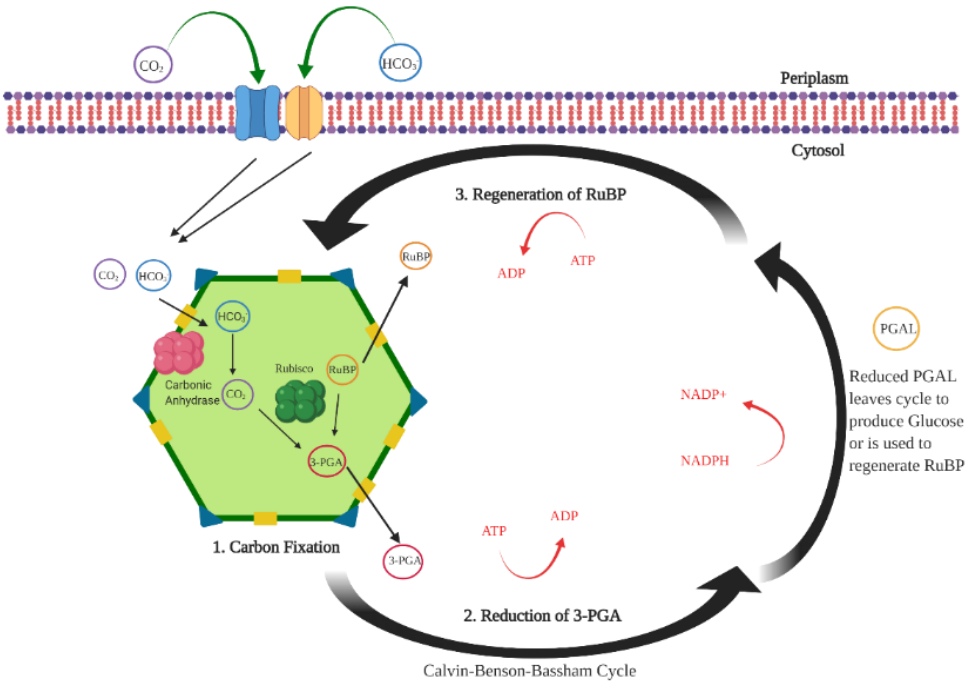


Figure 2.3 Assembly and function of carboxysomes. Created with BioRender.com.

CBs are unique to phototrophic bacterial species as plants do not have their RuBisCO proteins compartmentalized but rather localized within the chloroplast organelle (Figure 2.3). Consequently, C3 plants typically have low RuBisCO activity and therefore, less efficient carbon fixation. C3 plants have a fixation rate of 2-5 mol CO_2 fixed per mole RuBisCO active site^{-1} per sec^{-1} (V_{C02}) while Cyanobacteria have been shown to exhibit a rate of 12-14 V_{CO2} (Whitney *et al.*, 2011). Due to this, many research groups are in the process of implementing CBs within plant chloroplasts in the hopes to improve carbon fixation rates (Lin *et al.*, 2014; Long *et al.*, 2018; Occhialini *et al.*, 2016).

2.2.1 α-Carboxysomes vs β-Carboxysomes

There are two types of carboxysomes, alpha and beta (α-CB and β-CB), that exhibit similar conservation in shell protein types and icosahedral structures (Figure 2.4) (Kerfeld *et al.*, 2016). The α-CB is the smaller particle of the two, with an average diameter ranging from 100-200 nm and a shell thickness of 4 nm (Kerfeld *et al.*, 2005). The β-CB, in contrast, ranges from 200-400 nm in diameter depending on the species (Sutter *et al.*, 2019). The respective diameter ranges have been determined using electron microscopy. The size variability has been hypothesized to be due to a number of factors such as environmental conditions (light and CO_2 concentrations), the particular assembly mechanisms, variability in stoichiometry of shell proteins during assembly, or the cargo loading differences of RuBisCO (Y. Sun *et al.*, 2019). The RuBisCO cargo (type 1A) within the α-CB forms lattice structures when packaged (Metskas, 2022) while the β-CB with RuBisCO (type 1B) has a stacked packaging conformation (Kerfeld *et al.*, 2015; Schmid *et al.*, 2006). Carbonic Anhydrase (CsosCA, formally known as Csos3 in α-CB's) is the other cargo of both CB types and has been shown to be imbedded in the shell structure, rather than within the lumen (Baker *et al.*, 2000). The shell structure of both CB types can be slightly variable in stoichiometry and structure, allowing for a flexible structure in response to different amounts of RuBisCO and metabolites compartmentalized within the shell (Y. Sun *et al.*, 2019).

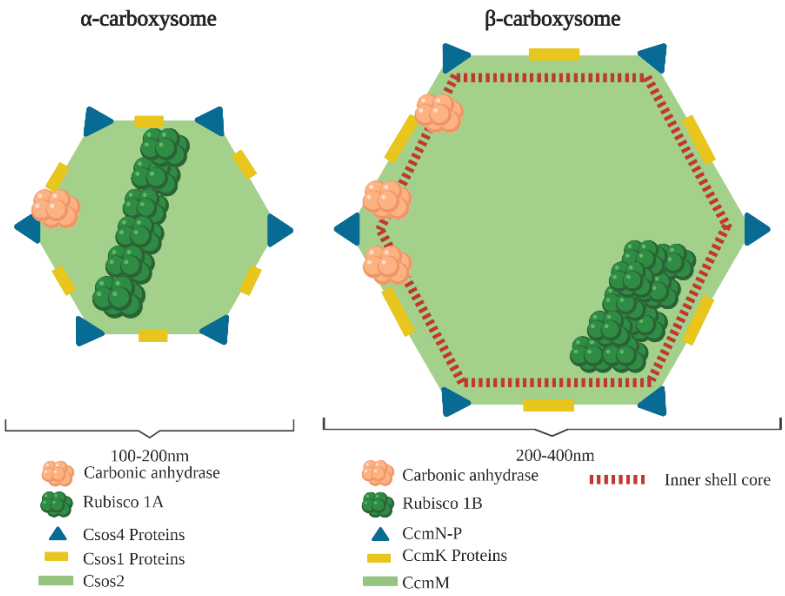


Figure 2.4 The structure and composition of the α- and β-carboxysome. The size ranges, shell types and protein cargos for each type of CB is highlighted, assuming the presence of all shell protein types. The differences in RuBisCO packing of each CB is also shown. Created with BioRender.com.

The α-CBs, like all BMCs, is organized within an operon with the exception of the shell protein Csos1D, which is expressed from a downstream or upstream satellite loci (Roberts *et al.*, 2012). Figure 2.5 compares several representative α-CB operons to highlights that the order of genes within the main operon and positioning of the Csos1D satellite loci can differ between species. In contrast, the genetic organization of β-carboxysomes (not shown) is more specific: each species contains a main locus with one of the CcmK1/2 genes, which is followed by CcmL, CcmM, CcmN and for the majority, CcmO. The last gene, CcmP, is always observed in a satellite locus (Sommer *et al.*, 2017). As my thesis focuses on the α-CB from *H. neapolitanus*, further detail will be provided below on the specific shell proteins and cargos present in this CB type.

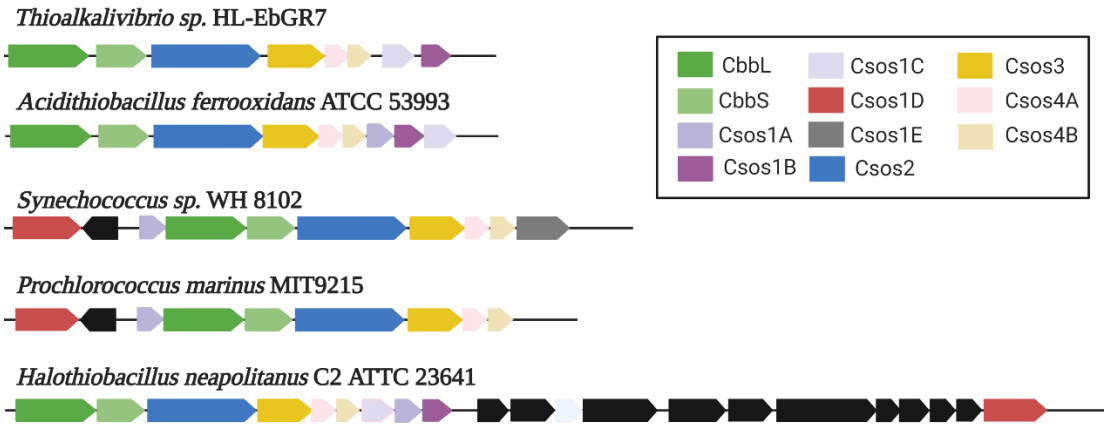


Figure 2.5 Organization of α-CB operons in select bacterial and cyanobacterial species. The respective shell proteins and Rubisco are represented as colored coding sequences. Unlabelled genes (black) are found upstream/downstream of the operon but are not a part of the operon. The position of coding sequences is based on data from the Microbes Online Database. Created with BioRender.com.

2.2.2 α-carboxysome Shell Proteins

α-carboxysomes consist of hexametric, pentameric, or trimeric subunits formed by the shell proteins. Shell proteins in α-CB species consist of Csos2 (which exists as two isoforms), various paralogs of Csos1 and Csos4, and in some cases, Csos1D. Paralogs of Csos1 consist of Csos1A, B, C, D, and E and paralogs of Csos4, 4A and 4B. Their structure and function are summarized in Table 2.1. For the α-CB found in the species *H. neapolitanus* specifically, the described shell proteins, apart from Csos1E, constitute the complete shell structure. The exact stoichiometry of each shell protein within the α-CB remains elusive; only the average stoichiometry of the β-CB's has been determined so far (Y. Sun *et al.*, 2019). Each shell protein element starts as a monomeric protein subunit. These subunits then interact with each other to form higher level shell structures (Figure 2.6). Each shell component, once it has formed their hexameric,

pentameric, and trimeric structures, interact with other shell proteins to form higher order structures and eventually forms the entire shell structure.

Table 2.1 Types of shell proteins that can make up the α-CB. The Shell structure, function and essential role of each shell protein identified in α-CBs is described.

Shell Protein Name	Structure	Function	Necessary for CB formation?	References
Csos1A	hexamer; varies from other Csos1 proteins at C-terminus; pfam00936	Potentially facilitates diffusion of metabolites or proteins into the lumen	Yes	(Fridlyand *et al.*, 1996; Jablonsky *et al.*, 2011; Kerfeld, 2017; Rolland *et al.*, 2016; So *et al.*, 2004)
Csos1B	Hexamer; Crystal structure has not been determined; Pfam00936	Hypothesized to interact with Carbonic Anhydrase	No	(Frey *et al.*, 2016; Kerfeld, 2017; Q. Sun *et al.*, 2019)
Csos1C	two layered Hexamer; Pfam0936	Weakly binds to the large RuBisCO subunit	No	(Frey *et al.*, 2016; Kerfeld *et al.*, 2005; Lopez-Sagaseta *et al.*, 2016)
Csos1D	trimer structure that forms dimers of trimers with one open and one closed pore: two Pfam0936 domains	Hypothesized to facilitate metabolite or protein transport into the lumen	No	(Frey *et al.*, 2016; Shih *et al.*, 2016; Tabita *et al.*, 2008)
Csos1E	Similar C-terminus of Csos1D; Structure not determined; Pfam00936	Found in species that live in low light environments; Function unknown	Unknown*	(Cai *et al.*, 2009; Frey *et al.*, 2016; Roberts *et al.*, 2012)
Csos2	Trimer structure; Has 3 domains and 2 isoforms; Pfam11288	Acts as a protein scaffold. Binds RuBisCO; initiates CB formation	Yes	(Frey *et al.*, 2016; Sommer *et al.*, 2017; Sutter *et al.*, 2017)
Csos4A/4B	Pentamer structures; Pfam03319	Makes up vertices of the CB; hypothesized to interact with Carbonic Anhydrase	No	(Axen *et al.*, 2014; Frey *et al.*, 2016; Q. Sun *et al.*, 2019)

** Structure-function studies of Csos1E have not yet been conducted and the implications for α-CB assembly of Csos1E containing species have not been determined.*

The resulting higher order structures composed of these shell proteins form pores that can have different electrostatic properties and consequently are hypothesized to control permeability and passage of metabolites (Cai *et al.*, 2015). The hexameric pores are suggested to allow permeability of small molecules such as CO_2 or HCO_3^- while the pore formed from the trimer or pentamers have been suggested to provide passage of proteins and metabolites (Klein *et al.*, 2009). In both cases, the permeability of each shell type has only been modelled through computational simulations (Faulkner *et al.*, 2020). Despite how little is known about the

functionality of the pores within each shell structure, it is apparent that their control of metabolic transport in and out of the cell is an important component of the α-CB structure, and therefore of large interest in novel applications. Despite allowing the passage of small proteins and metabolites, the shell proteins prevent the leakage of RuBisCO and CO_2, which therefore retains the high RuBisCO activity and high concentrations of CO_2. Compartmentalization of CO_2 prevents the cytotoxic effects caused by the high concentrations required for increasing the carboxylation activity of RuBisCO (Kerfeld *et al.*, 2015; Kinney *et al.*, 2011).

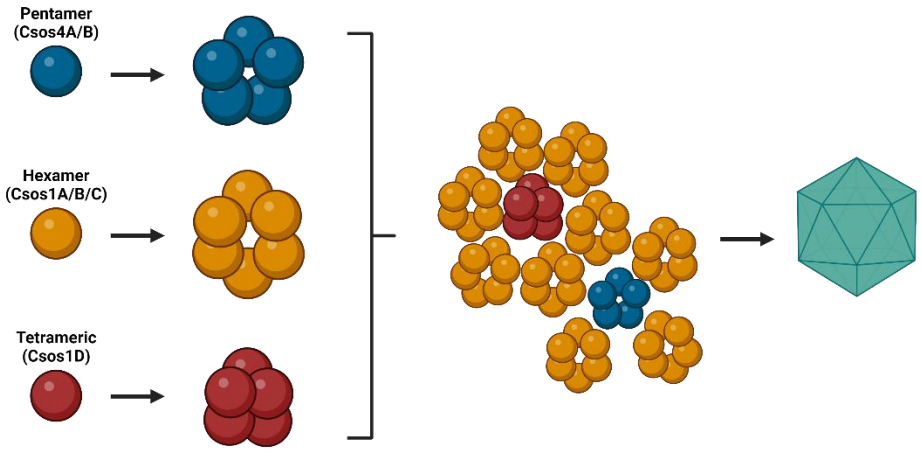


Figure 2.6 The basic assembly mechanism of the α-CB. As the stepwise assembly of the α-CB is not fully determined, the basic assembly is shown here, with monomeric shell proteins forming the higher order hexameric, pentameric, etc. structures before interacting with each other to form the overall CB shell. Created with BioRender.com

2.2.3 Assembly and Encapsulation Mechanisms of the α-Carboxysome

The detailed mechanism of assembly for the α-CB has yet to be elucidated, although several theories have been developed (Oltrogge *et al.*, 2020). α-CBs typically do not form empty shells (unlike the β-CB) suggesting that the interaction of cargo is necessary for α-CB formation (Dai *et al.*, 2018; Menon *et al.*, 2008) and efficient cargo encapsulation. A recent publication

suggests that assembly initiation proceeds via the interaction between the RuBisCO cargo and shell protein Csos2 (Oltrogge *et al.*, 2020) with Carbonic Anhydrase being incorporated immediately after (Zang *et al.*, 2021). The encapsulation peptide (EP) (a sequence found in BMC cargo proteins that recognizes and binds to the lumen of the shell structure) within the RuBisCO peptide sequence is bound to the N-terminal region of the Csos2 shell protein (Oltrogge *et al.*, 2020). The other shell proteins are hypothesized to be subsequently incorporated with the RuBisCO-Csos2 initiation complex. However, the detailed steps through which their assembly occurs are still not known, although several protein-protein interactions between shell proteins and cargo are known (Y. Liu *et al.*, 2018; Zang *et al.*, 2021). Once assembled, the α-CB typically takes on an icosahedral, flexible structure.

2.3 Developments in Synthetic Biology: The Minimal Carboxysome

Although many groups have engineered both types of CBs, developing chimeric CBs (Cai *et al.*, 2015), changing cargo (Hagen *et al.*, 2018), and using shell proteins for other applications (Huang *et al.*, 2019), the development of a CB with a minimal set of proteins has been of recent interest. For future applications, the CB can be minimized genetically (thus being called a minimal carboxysome, mCB) while maintaining desired functionalities. For example, Long *et al.*, 2018 used the genes Csos1A and Csos2 to create a minimal gene set to introduce α-CBs into the chloroplasts of tobacco plants (Long *et al.*, 2018). The pores formed by the two shell proteins provided proper metabolic flux in and out of the shell, while the interactions between the hexamer and tetramer allowed for the formation of mCBs. The structures of these mCBs differ from the native CBs due to the lack of pentameric proteins, shell proteins that are responsible for forming the vertices of the icosahedral CB structure. Without pentamers, the carboxysome will either form an icosahedral structure with structural gaps or elongated structures due to the lack of

vertices formation (Long *et al.*, 2018). Tan *et al.*, 2021 recently produced a library of mCBs, developed by selecting a variety of combinations of hexameric and pentameric shell proteins, while using the same shell protein, Csos1D, and cargo. Subsequently they characterized the structures and the ability for the mCB to protect encapsulated cargo from different extreme environments which included temperature and chemical denaturation (Tan *et al.*, 2021). In this thesis however, only a single hexameric protein (Csos1A) and Shell protein (Csos2) was used (described further in Chapter 5).

2.4 Methods for the Biophysical Characterization of Carboxysomes

This thesis uses several methods to characterize the structure of the α-CB and mCB. Although this is not an exhaustive set of methods, the key methods used in this thesis can provide details regarding the size, molecular weight, relative abundance of certain sized particles, and in some cases overall shape of the protein shell. The following section describes the methods used in technical detail, as well as the data that can be obtained with them.

2.4.1 Transmission Electron Microscopy

Transmission electron microscopy (TEM) can be considered one of the more classical methods for accessing the structure of CBs *in vitro* and *in vivo*. The first visualization was achieved in the 1980's (Burghardt *et al.*, 2006). TEM uses electron beams (Figure 2.7), where the diffraction of the electrons, once they hit the biomolecule or a cell that is fixed on a grid (for example a carbon coated grid), allows for high resolution contrast images (Bonacci *et al.*, 2012). Contrast imaging is facilitated by the negative staining of a sample. The use of stains like uranyl acetate or uranyl formate to stain the background grid provides increased resolution of the

biomolecule, particles, or organisms within a sample. TEM specifically is the first method to determine the structure and average size of the α-CB.

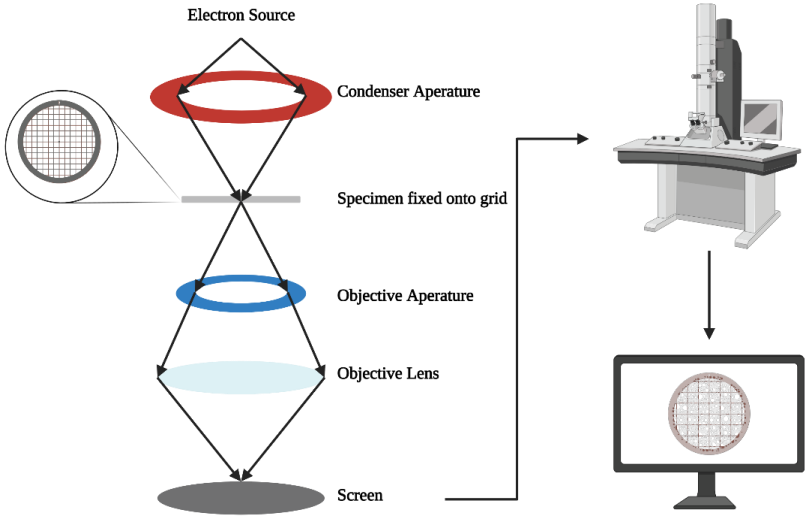


Figure 2.7 A basic schematic of Transmission Electron Microscopy. Adapted from (Shang *et al.*, 2016). Created with BioRender.com.

2.4.2 Light Scattering Methods

Light scattering occurs when an incoming beam of light is deflecting off a particle, resulting in several scattered light beams with different trajectories. Measuring this light scattering can be used to determine the size (diameter or radius) and molecular weight of biomolecules in solution (Andersson *et al.*, 2003). For spherical particles, a mathematical model known as Lorenz-Mie scattering can be used to accurately describe the size and molecular weight of the particle. Experimental methods that use light scattering for such analyses allow for the derivation on both the particle size and shape, as well as the hydration sphere around the particle in solution. There are many different technical approaches to using light scattering for biophysical characterization, but for the purposes of this thesis, only Dynamic Light Scattering (DLS) and Multi-Angle Light Scattering (MALS) will be described (Figure 2.8). Both methods

use a single beam of light but vary in the number of detectors. DLS uses a single detector at a fixed angle, evaluating the fluctuation of light intensity at the given angle over time. MALS simultaneously uses multiple detectors at different angles. DLS provides the average value for the light scattering of a particular sample and consequently one value for the size and molecular weight of particles. Therefore, it is useful for the analysis of homogeneous samples (Kaasalainen *et al.*, 2017).

Instead of giving average values, MALS provides the accurate identification of sizes, shape, and molecular weights of particles within a sample as MALS can be coupled to chromatography-based separation techniques such as a size exclusion chromatograph (SEC)(used in this study). This sequential combination, SEC-MALS allows the separation of particles within a sample by size and then use MALS to obtain specific size, shape and mass values as biomolecules separate on the SEC column. Therefore, SEC-MALS is a promising method for discerning different CB species within a sample.

MALS (and DLS) typically provides a larger value for the size of a particle as compared to other methods such as TEM, as it includes the hydration sphere around the particle (Chen *et al.*, 2012). The particle diameter is inferred from the radius of gyration, which is the root mean square distance of the particle's center of mass. The mathematical concepts for the Lorenz-Mie mathematical model are discussed in section 2.4.2.2.

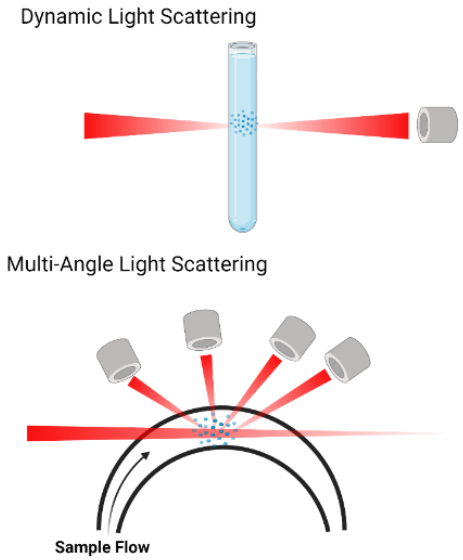


Figure 2.8 Dynamic Light scattering vs. Multi-Angle Light Scattering. Created with BioRender.com.

2.4.2.1 The Refractive Index

To assess the quality of light scattering data, one can first refer to the refractive index (RI) of the biomolecule. The refractive index is a relative number that describes the ability of light to travel through a material and is defined as the ratio of the speed of light in a vacuum over the phase velocity of light in solution, and depends on the wavelength of light used (Han *et al.*, 2020). For perspective, the refractive index of water is 1.33. Although a refractive index can have a value below 0, negative values are typically only seen with “metamaterials” (synthetic materials). The refractive index values determined by the SEC-MALS system are expected to vary as the solution resulting from the chromatography step will contain different amounts of particles/proteins, whereas Dynamic Light Scattering will generate a single value.

2.4.2.2 Calculating the Hydrodynamic Radius and Molecular Weight of Carboxysomes

I light scattering of Lorenz-Mei was analyzed using the ASTRA software (Wyatt Technology) and used to determine the Molecular weight and diameter of the spherical particle using the Rayleigh-Gans Approximation equation given below (Bohren, 1998):

$$I=I_0\left(\frac{1+\mathrm{Cos}^2\theta}{2R^2}\right)\left(\frac{2\pi}{\lambda}\right)^4\left(\frac{n^2-1}{2^2+2}\right)^2\left(\frac{d}{2}\right)^6 \quad (1)$$

The variable I refers to the light intensity, I_0 refers to the buffer solution. R is the distance between the particle and the light detector. Θ refers to the light scattering angle, λ the wavelength of light used (in this case 280 nm to detect protein), n being the refractive index determined by refractometer device, and d, the diameter of the particle.

Equation 1 includes the diameter of the particle but not the shape of the respective particle. Alternatively, the stokes radius equation (Khokhlov, 2000) below can be used to determine biophysical characteristics, and where the determined radius also accounts for the hydration sphere that surrounds the molecule in solution. This equation is especially important as it is applied to spherical or "sphere-like" particles. The radius is therefore more accurately called the hydrodynamic radius. This equation also factors in the effects of the solvent by accounting for the diffusion rate of the molecule as it travels through the solvent.

$$R_H=a=\frac{K_BT}{6\pi\eta D} \quad (2)$$

Where:

$$D=\frac{K_BT}{f} \quad (3)$$

The hydrodynamic radius (R_H) or radius (a) of the particle is dependent on the Boltzmann constant, temperature (T in kelvin), the sample viscosity (η) and the diffusion coefficient, D. As the diffusion coefficient is dependent on the frictional coefficient (f), it can be defined by the relative shape of the particle.

For the consideration of spherical particles such as the Carboxysome, the hydrodynamic radius of the particle depends on the viscosity of the solution which can be determined using a viscometer. To determine the molecular weight, Rayleigh based equations are used (Kratochvl, 1987).

$$\frac{KC}{R\theta} = \frac{1}{MwP(\phi)} + 2A_2C \qquad (4)$$

Where:

$$P(\phi) = \frac{1}{1 + \frac{16}{3\lambda_0^2}\pi^2 R_H^2 Sin^2\frac{\theta}{2}} \qquad (5)$$

And:

$$K = \frac{2\pi^2 n^2}{\lambda_0^2 N_A}\left(\frac{d_n}{d_c}\right)^2 \qquad (6)$$

The molecular weight determined by equation 4, contains the optical constant (K), sample concentration (C), second virial coefficient (A_2) and scattering equation ($P(\phi)$) which relates back to the hydrodynamic radius.

The $\frac{d_n}{d_c}$ values indicate the change of refractive index (d_n) relative to the change of solute concentration (d_c) in solution. Proteins have the same relative $\frac{d_n}{d_c}$ value of 0.184 (shown in

equation 6 and defines the optical constant), so we can assume that the α-CB has the same value. This results in the following equation:

$$K=0.034\frac{2\pi^2 n^2}{\lambda_0^2 N_A} \quad (7)$$

The optical coefficient is therefore based on the refractive index (n), the initial wavelength (λ_0) and Avogadro's number (N_A).

2.4.3 Analytical Ultracentrifugation

Analytical Ultracentrifugation (AUC) is an advanced method that can determine the biophysical characteristics of a particle by coupling centrifugation with measurements in real time (e.g., absorbance, fluorescence, and light scattering). AUC can accurately determine the size and molecular weight of a given particle but also allows to characterize the heterogeneity of samples, spatial conformation of species, and in some cases the kinetics of interconversion between two species in solution (Ralston, 1993). Analysis by AUC is done by monitoring the sedimentation of particles in solution as a centrifugal force is applied to the sample. The sedimentation coefficient that is determined through this method is dependent on the velocity of the particle in response to the gravitational acceleration (Dam *et al.*, 2004).

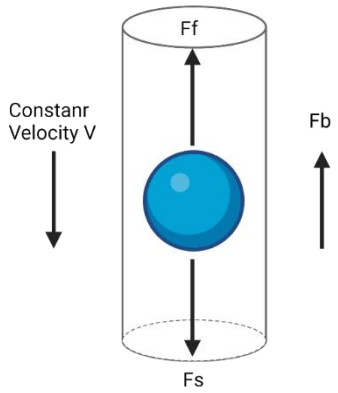


Figure 2.9 Forces experienced by a particle during Ultracentrifugation. Particles during ultracentrifugation are exposed to gravitational force (F_s), Buoyant force (F_b) and frictional force (F_f). Adapted from (Ralston, 1993). Created with BioRender.com.

The sedimentation and the relative forces that are applied to the particles in solution (Figure 2.9) can be used to obtain the molecular weights, stoichiometry of complexes, and binding kinetics. The sedimentation coefficient of a particle can be obtained using the following equation (Cole *et al.*, 2008):

$$s = \frac{u}{\omega^2 r} = \frac{M(1-\upsilon\rho)}{N_A f} \quad (8)$$

Where ω is the angular velocity, r is the axis of rotation, N_A is Avogadro's number, u is velocity of the particle, and M is the molecular weight of the particle. The molecular weight is dependent on the buoyant force of the particle in solution in which the effective molar weight (mass of fluid displaced by the particle) m_0, is used to derive the mass and molecular weight of the particle.

$$F_b = -m_0 \omega^2 r \quad (9)$$

$$m_0 = m\upsilon\rho = \frac{M}{N_A}\upsilon\rho \quad (10)$$

The effective molecular weight is also dependent on the partial specific volume (υ), the density of the solvent (ρ), and the mass of a single particle in solution (m). The molecular weight and other properties can also be determined via Frictional force (F_f) and gravitational force (F_s):

$$F_f = -fV \quad (11)$$

$$Fs = m\omega^2 = \frac{M}{N_A}\omega r^2 \quad (12)$$

As the velocity of the particle will depend on its mass and its ability to move through solution (influenced by the particles shape), the resulting sedimentation coefficients can be used to determine biophysical properties through subsequent mathematical modelling (Cole *et al.*, 2008). Modelling includes the fitting of the obtained data using 2-D spectral analysis (the modelling done in this thesis, further explained in section 3.6), for the identification and characterization of individual particles in a sample that may be diverse in size and shape (Brookes *et al.*, 2010). This mathematical analysis is supported by Monte Carlo simulation modelling (which is a probability model that aids the analysis of data with the presence of absorbance signal noise) that is provided by the Ultrascan software. Other similar tools and methods are available (Aziz *et al.*, 2007), A basic schematic of how samples are prepared for AUC can be seen in Figure 2.10.

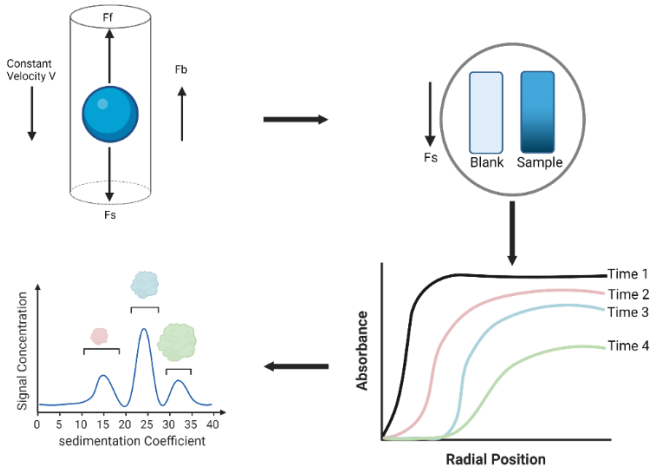


Figure 2.10 Typical Analytic Ultracentrifuge analysis set-up. The analyte containing samples are placed in a cell which can contain either one or two samples, including a reference (blank) sample. The cell is then introduced into a specialized rotor which subsequently is centrifuged in an Optima AUC Centrifuge. The AUC conducts a series of consecutive reads (e.g. fluorescence or absorbance) over time monitoring the sample particles move towards the bottom of the cell following the applied gravitational force. The resulting data reads, corresponding to change in radial position of the particle, can be used to extrapolate properties such as sedimentation coefficients. Created with BioRender.com.

CHAPTER 3: MATERIALS AND METHODS USED TO STUDY THE α-CARBOXYSOME AND MINIMAL CARBOXYSOME

3.1 Cloning and Construct Design

The following Chapter describes the methods used in all the work described in the thesis apart from Appendix II. All reagents were obtained from New England Biolabs, Fischer Scientific, Sigma Aldrich, Promega, and Biobasic unless otherwise stated.

The pHnCBS1D plasmid (Figure 3.1) containing the α-CB operon and Csos1D gene (naturally expressed on a satellite locus) from *H. neapolitanus* (Klein *et al.*, 2009) was obtained from Addgene (#52065). The plasmid construction has been previously published (Bonacci *et al.*, 2012). The Csos1D gene and a natural cyanobacterial ribosomal binding site (RBS) was cloned at the end of the operon to allow for expression control by T7. In this thesis the α-CB operon and Csos1D were expressed using the T7 promoter and IPTG induced expression. The corresponding α-CB associated protein sequences are summarized in Appendix I.

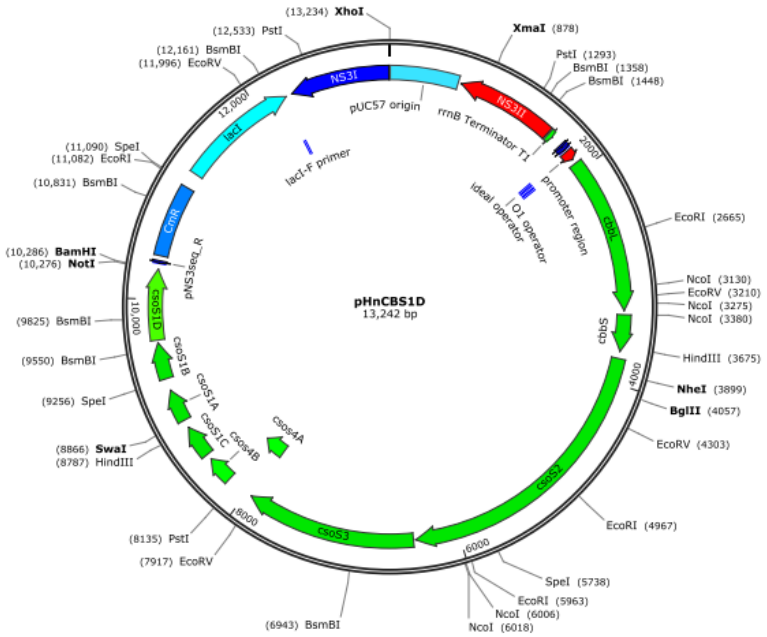


Figure 3.1 The pHnCBS1D plasmid (Addgene #52065) with the α-CB operon and Csos1D. (Bonacci *et al.*, 2012)**.** Coding sequences of the α-CB genes are shown in green. The promoter and operator are upstream of the CbbL coding sequence and the relevant restriction sites are indicated. Image was obtained from addgene.com.

The minimal carboxysome construct, (pET28a(+)Csos1A2RuBisCOls) was synthesized and cloned by BioBasic Inc. The DNA insert consists of Csos2, RuBisCO large subunit, RuBisCO small subunit, and Csos1A genes in the respective order, with synthetic regulatory regions for each gene. This insert was introduced into pET28a (+) using XhoI and XbaI restriction sites (Figure 3.2 and Figure A1.1). The Csos1A gene is under control of the T7 promoter while the other three genes are under the control of the constitutive promoter BBa_J23105 (Smolke, 2009). For other transcriptional regulation, different double terminator (dT) sequences were used. Translational regulation for all genes includes different medium Ribosomal binding sites (mRBS). The use of different mRBS and dT sequences helped to increase the success of gene synthesis and cloning. Csos2 is controlled by mRBS BBa_B0034 and dT BBa_B0015, Csos1A by mRBS BBa_B0031 and dT BBa_B1002, Large RuBisCO

subunit mRBS BBa_B0030 and dT BBa_B1006, and Small RuBisCO subunit by mRBS BBa_B0032 and dT BBa_J61048. The coding sequences were codon optimized to *E. coli* K12 using IDT's (Integrated DNA Technologies Inc) Codon Optimization Tool. At the 5' and 3' end of each coding region within the construct specific restriction enzyme cut sites were integrated, such that genes can be removed, replaced, or allow for the fusion/insertion of other genes such as fluorescent reporters. Cut site pairs were chosen based on buffer compatibility (Thermo Fisher Scientific) to allow for efficient double digestions if required. The chosen restriction enzyme cut sites, if identified in the construct (using NEBcutter V2.0, New England Biolabs), were removed from the coding regions and replaced with a different codon pair as per the *E. coli* K12 codon usage table (Maloy, 1996). The cut sites were also confirmed to be absent from the regulatory regions of the construct.

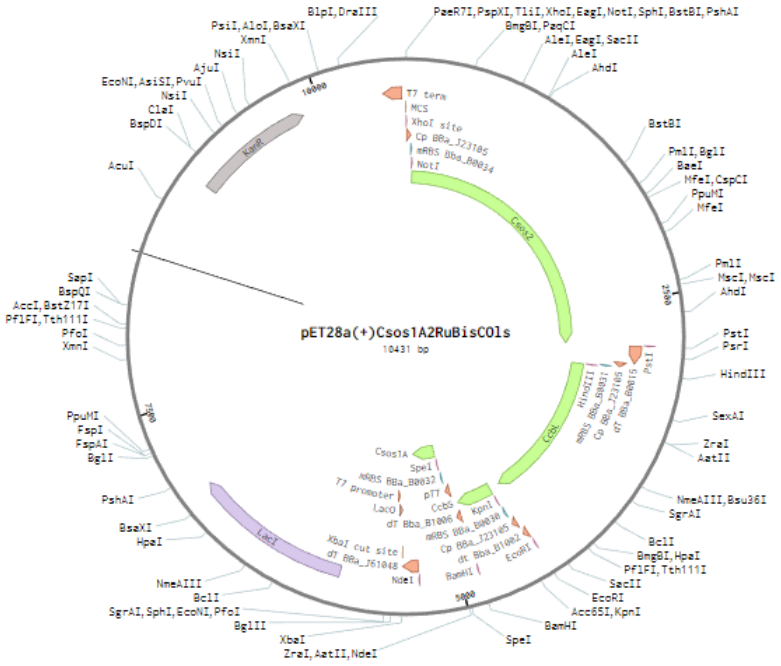


Figure 3.2 Plasmid map of the minimal carboxysome construct pET28a(+) Csos1A2RuBisCOls. The promoters (red), ribosomal binding sites (blue), coding sequences (green), inserted cut sites (pink), and terminators (red) are annotated on the plasmid map. Any sequence from the iGEM registry is denoted by their BBa number. The plasmid map was created in Benchling.

3.2 Protein Expression in *E. coli*

The α-Carboxysomes were expressed according to published protocols with minor modifications (Bonacci *et al.*, 2012). Briefly, the pHnCBS1D and pET28(+)Csos1A2RuBisCOls plasmids were transformed into DH5α and BL21 (DE3) cells using commercially purchased chemically competent cells (New England Biolabs). Cells were transformed by adding 1 µL of plasmid (50-100 ng/µL) into 10 µL of competent cells. After 30 minutes incubation on ice, cells were heat shocked at 42 ˚C for 20 seconds and placed back on ice for 1 minute. The transformed cells were then diluted in 950 µL of LB media in a 1.5 mL microcentrifuge tube and placed at 37 ˚C and shaken at 220 rpm for 1.5 hours. ~200 µL of the cells in LB was then spread plated onto prewarmed Luria Broth (LB) Agar plates supplemented with 0.025mg/mL final concentration of Chloramphenicol.

For protein expression, BL21 (DE3) colonies containing the plasmid were streak plated from a glycerol stock, picked, and then grown overnight in 50 mL of LB in 150-250 mL Erlenmeyer flasks supplemented with 0.025 mg/mL final concentration of Chloramphenicol. The next day, 500 mL of LB in 2 L Erlenmeyer flasks were inoculated with the overnight culture to a final $OD_{600\ nm}$ of 0.1 and supplemented with 25 mg/mL of Chloramphenicol. Cultures were grown at 37˚C at 220 rpm to $OD_{600\ nm}$ of ~0.4-0.6 in a New Brunswick Innova incubator (Eppendorf). Expression was induced with IPTG (Isopropyl ß-D-1-thiogalactopyranoside) at a final concentration of 0.5 mM, and cells were grown at 20˚C for 20 hours at 180 rpm. Cultures were harvested by centrifugation at 4,000 g for 15-20 minutes using the JLA 8.1000 rotor and Avanti JXN-26 centrifuge. The resulting cell pellets were stored at -20˚C. Average yields ranged from 6-9 g of cells per litre of culture. Overexpression was confirmed by SDS-PAGE.

3.3 Protein Purification of the α-Carboxysome

The α-CB purification protocol is based on previous work (So *et al.*, 2004) with minor modifications. Briefly, the *E. coli* cell pellets were suspended on ice in ~50 mL of TEMB buffer (5mM Tris-HCl pH 8.0, 1mM EDTA, 10mM $MgCl_2$, 20mM $NaHCO_3$) for ~7 g of dry cell mass. Cells were opened by sonication using a 1/5" sonication tip and a Branson Sonifier 450 (Branson Ultrasonic corp.). Sonication was done (on ice) 30 times (30 seconds) at 50% duty and an output of 6. The following steps were performed at 4˚C unless otherwise stated. Cell debris and unopened cells were separated from the lysate by centrifugation at 12,000 x g for 20 minutes using a JA25.50 (Beckman Coulter) rotor with an Avanti JXN-26 centrifuge. The resulting supernatant was then centrifuged for an additional 20 minutes at 40,000 x g to collect the carboxysome particles. The resulting pellet was resuspended in 20 mL of TEMB buffer with 1 mg/mL of lysozyme followed by incubation for 30 minutes at room temperature. The cell suspension was centrifuged at 40,000 x g for 30 minutes with the JA25.50 rotor (Avanti JXN-26) to obtain an enriched carboxysome pellet. The pellet was then resuspended in 3-5 mL of TEMB buffer, using a glass rod and gently pipetting. The volume was chosen to ensure a high concentration sample. The obtained suspension was then cleared by centrifugation at 3,000 x g for 1 minute using a BIOShield 4 x 250mL Swinging-Bucket rotor (Sorvall Legend RT centrifuge). The supernatant (1-2 mL between two tubes) was loaded onto a 10-50% weight/volume linear (34 mL) sucrose gradient (Gradient Master 108, Biocomp Instruments). The Carboxysomes were separated by centrifugation at 26,000 rpm (SW28 rotor, Beckman Coulter) for 18 minutes in an OPTIMA XPN-100 Centrifuge and subsequently fractionated at room temperature using the ÄKTA prime system equipped with a fractionation pump and absorbance of 280 nm was recorded. Fractions containing assembled α-CBs were diluted in

TEMB buffer and the α-CBs were collected by centrifugation for 82 minutes in a JA25.50 Rotor at 40,000 x g (Avanti JXN-26). The resulting final pellet was resuspended in TEMB containing 50% glycerol, flash frozen and stored at -80˚C. Before subsequent analysis after thawing, α-CBs containing samples were shaken overnight at 4˚C and aggregates were subsequently removed by centrifugation at 13,000 x g for 15-30 minutes (accuSpin Micro 17, Fisher Scientific).

3.4 Transmission Electron Microscopy

Purified α-carboxysomes were diluted in TEMB buffer to a final concentration of ~1 mg/mL. Carbon grids (Electron Microscopy Sciences) were made hydrophilic by exposing to a glow discharge for 15 seconds using a plasma cleaner (Sigma). 5 μL of the diluted α-CBs were added on top of the carbon grid and incubated at room temperature for 2-5 minutes. Excess liquid was removed using whatman filter paper and the grid was washed three times with water. Samples were stained with 3% Uranyl acetate (Electron Microscope Sciences) for 5-10 seconds. Excess stain was removed, and the grid was left to dry for 5-10 minutes at room temperature. The stained grids were imaged using a FEI Talos F200X S/TEM microscope equipped with a Ceta camera (Thermo Fisher Scientific).

3.5 Size Exclusion Chromatography- Multi-Angle Light Scattering (SEC-MALS)

The SEC-MALS system used was an ÄKTA Pure purification system coupled to the DAWN MALS unit with Refractive Index Optilab detector (Wyatt Technologies) (defined as DAWN Optilab in this thesis). Purified α-CBs were separated on a 10/300 S400 Sepharose column (Cytiva). The column was equilibrated with 3 column volumes (CV) of distilled water and 3 CV of TEMB buffer, both filtered through 0.1 μm cellulose nitrate membrane filters (Cytiva). A-CBs at 96 and 206 picomoles of total protein within a volume of 500 μL were

injected onto the column and the samples were resolved at a flow rate of 0.4 mL/min using the ÄKTA Pure system (room temperature). As the α-CBs eluted from the column (at 40 mL total or ~1.7 CV), they passed through the DAWN Optilab MALS system. No fractions were collected unless further analysis of fractions was performed. Fractions were collected using the F9-C fraction collector (Cytiva) attached to the ÄKTA Pure system at 4˚C.

Analysis of the collected scattering data was performed using the ASTRA software (Wyatt Technologies). Light scattering was interpreted using the sphere model that follows the Lorenz-Mie Theory. For an improved data fit and to remove inaccurate scattering data, data from light scattering detectors 1-5 was excluded from the data set (due to excessive signal noise from the detector) (Some *et al.*, 2019). Subsequently the molecular weight and hydrodynamic radius were determined.

3.6 Analytical Ultracentrifugation (AUC)

Purified α-CBs were loaded into an AN-60 Ti Rotor (Beckman Coulter) with 450 μL of sample per cell. AUC experiments were performed at the Canadian Center for Hydrodynamics at the University of Lethbridge. Multiwavelength sedimentation velocity AUC experiments were performed in an Optima AUC centrifuge (Beckaman Coulter), using a 2-channel epon-charcoal centerpiece, with a 1.2 cm pathlength fitted with Quartz windows. Samples were analyzed at 14,500 rpm (AN-60-Ti) at 4°C, and sedimentation was monitored using the UV absorbance optics of the instrument in intensity mode, measuring from 240 – 290 nm every 2 nm. A-CBs were prepared in TEMB buffer; the density and viscosity were determined to be 1.001410 g/cm^2 and 1.00485 cP, respectively, using UltraScan (Demeler *et al.*, 2016). The partial specific volume was set to the default value of 0.7200 mL/g.

All data were analyzed using UltraScan, version 6348. Initially, each multiwavelength dataset was analyzed with two-dimensional spectrum analysis, to fit the boundary and remove time and radially

invariant noise (Brookes *et al.*, 2010). Iteratively refined 2-Dimensional Spectrum Analysis models for each wavelength were used to simulate the entire MW-AUC experiment on a common time grid. From a spectral scan collected from 200-600 nm on a Genesys 10s benchtop spectrophotometer (Thermo Fisher Scientific), it was noted that protein, DNA, and Mie scattering contributed to the signal. Therefore, the carboxysomes MW-AUC data was deconvoluted into protein, DNA, and lipid nanoparticles (to account for the Mie scattering signal) spectral profiles. To do so, a dilution series of each species was collected from 210-310 nm. The dilution series for each was then globally fitted to a sum of gaussians using the spectrum fitter program in UltraScan, and scaled to 1 OD at A280, A260, and A230, for protein, DNA, and lipids, respectively. The absorbance spectral data was fitted against the following controls: BSA for the protein signal, a 4,048 base pair (bp) plasmid for DNA, and empty lipid nanoparticles for Mie scattering profiles. The profiles were then used to deconvolute the multi-wavelength data at each time and radial position (Henrickson *et al.*, 2021; Horne *et al.*, 2020; J. Zhang *et al.*, 2017).

3.7 DNase I Treatment of Purified Carboxysomes

Purified carboxysomes were treated with DNase I (Sigma) to determine the identity of the present nucleic acids, either RNA or DNA. The identification is done by developing agarose gels of DNAse I treated samples. Seeing no bands in the sample lane indicates that only DNA was present, whereas visible bands indicate the presence of RNA. Further analysis using a RNase A digestion was found to be unnecessary as this would not cleave DNA as indicated by the previous experiment. A-CBs were treated with 1 μL DNase (1 u/μL), and plasmid DNA (pet28a(+) with a 513 bp insert) was used as a control. The respective 20 μL reactions were incubated at 37˚C overnight, treated with 10 μL of 8 M Urea, heated at 100˚C, and then separated on an 0.5%

agarose gel for 16 V for 25 hours at 4˚C. The gel was stained with Ethidium bromide and imaged under UV light (302 nm).

3.8 Mass Spectrometry Analysis

Purified carboxysomes were incubated overnight at 4˚C under continuous shaking at ~100 rpm (Orbit Shaker, Lab-Line) to help dissolve carboxysomes. The next day, the purified α-CBs were centrifuged at 13, 000 x g for 15 minutes in a microcentrifuge (accuSpin Micro 17, Fisher Scientific) to remove particulates originating from incomplete lysis of the *E. coli* cell membrane. The resulting 500 µL containing ~ 50 µg of total protein was concentrated using an Amicron centrifugal filters (Cytiva). The resulting α-CB samples were then treated with 1,4-Dithiothreitol (DTT) (10 mM final concentration) for 30 minutes at 57˚C. After cooling to room temperature, the proteins were then alkylated with 50 mM Iodoacetamide (IAA) for 45 minutes in the dark. Excess DTT was added, and the solution was subjected to trypsinolysis (overnight at 37˚C). Following overnight digestion, the α-CB sample was analysed on a Orbitrap Mass Spectrometer (Thermo Fisher Scientific). The collected data was analyzed using the proteome Discoverer software V2.2 (Thermo Fisher Scientific). Gene Ontology analysis was performed using PANTHER classification system V17.0 (Thomas *et al.*, 2003).

CHAPTER 4: The BIOPHYSICAL CHARACTERIZATION OF THE α-CARBOXYSOME

4.1 Introduction

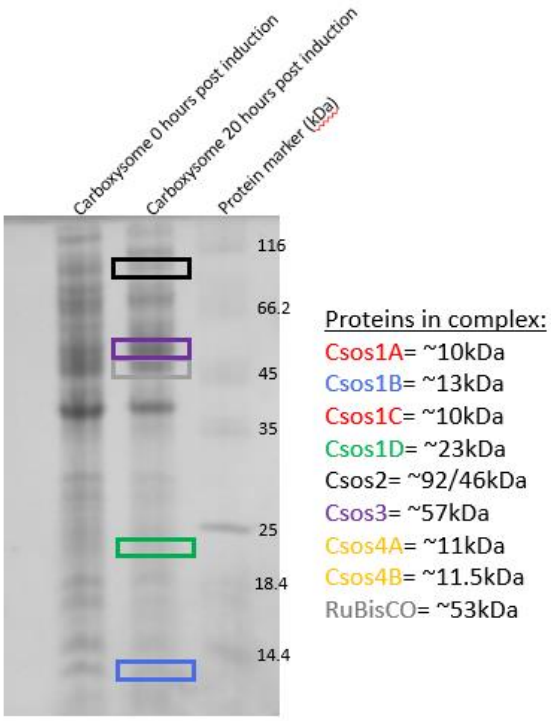
The following chapter reports the preliminary biophysical characterisation of purified α-CBs from *H. neapolitanus* using TEM, SEC-MALS, and AUC. To the best of our knowledge, purified and recombinantly expressed α-CBs have only been characterised using DLS and not SEC-MALS or AUC (Dou *et al.*, 2008; Li *et al.*, 2020; Long *et al.*, 2018). DLS limits the ability to assess the impacts of purification parameters, cargo loading, and rational engineering of the constituent proteins on, for example, assembly and shape of α-CBs. Despite the limited access to AUC by industry and researchers, it is a highly advantageous method to characterize purified α-CBs and engineered BMC structures. AUC is sensitive and can provide higher resolution with respect to the distribution of α-CB species present in a sample/ α-CBs preparation than DLS. This is in particularly true for very large biomolecular assemblies such as the α-CBs (averaging at 250 MDa), because the use of S400 sepharose as a part of the SEC-MALS limits the upper end of the achievable resolution. Furthermore, AUC can provide a novel way to analyze the structural details, as well as the assembly and encapsulation mechanisms of the α-CB. Beyond this study, establishing the use of AUC for other BMCs from different species can be highly beneficial for determining variability in structure, assembly, or encapsulation between BMC types and species.

DLS is constrained to reporting only the average molecular weight and hydrodynamic radius of a particle, limiting the ability to determine heterogeneity. However, SEC-MALS, despite its limitations in resolving large particles, is the best available chromatography-light scattering method to analyze molecular species present *in vitro*, such as multiple α-CBs that are drastically different in size or structure (in particular, smaller assembly intermediates).

4.2 Results

4.2.1 Purification of the α-Carboxysome Using Sucrose Gradient Ultracentrifugation

For biophysical analysis, the α-CBs were expressed and purified from *E. coli*. The wild type α-CB operon was expressed from the pHnCBS1D plasmid with expression being controlled by an inducible T7 promoter. The confirmed overexpression of the α-CB with each shell protein indicated by color can be found in Figure 4.1.



4.1 Overexpression of α-CBs in *E. coli*. Shell and cargo proteins were identified by size and are indicated by the color-coded table and boxes. Leaky expression from the T7 promotor results in α-CB bands present in the pre-induction lane. 10 µL of 1.0 OD 600 nm of cell lysate was analyzed on a 12% SDS-PAGE (180 V for 45 minutes) and stained with Coomassie G-250.

Purification of the α-CB from the cell lysate is achieved, primarily due to its unique size (250 MDa), through the use of a single sucrose gradient ultracentrifugation purification step (So *et al.*, 2004), removing the need for purification tags or disassembly prior to purification.

Sucrose gradient ultracentrifugation is a commonly used purification method for large particles, which provides high purity and enables separation of partially assembled α-CBs (Bonacci *et al.*, 2012). Subsequent ion exchange chromatography can be used as a secondary purification step, as many of the shell proteins have high positive charges which are presented on the outer shell structure (Sutter *et al.*, 2019). This method has been successfully applied only to smaller (~40 nm) synthetic β-CBs (Sutter *et al.*, 2017).

After cell lysis, α-CB enriched samples were split into two aliquots and loaded onto a 10-50% sucrose gradient. After centrifugation, both tubes were fractionated using an ÄKTA Prime system; the resulting chromatogram for each gradient is shown in Figure 4.2. An SDS-PAGE showing α-CBs during the purification process is shown in Figure S4.1. After purification was complete, the α-CB proteins were analyzed on a 12% SDS-PAGE and 15% Tris-Tricine PAGE (Figure 4.3).

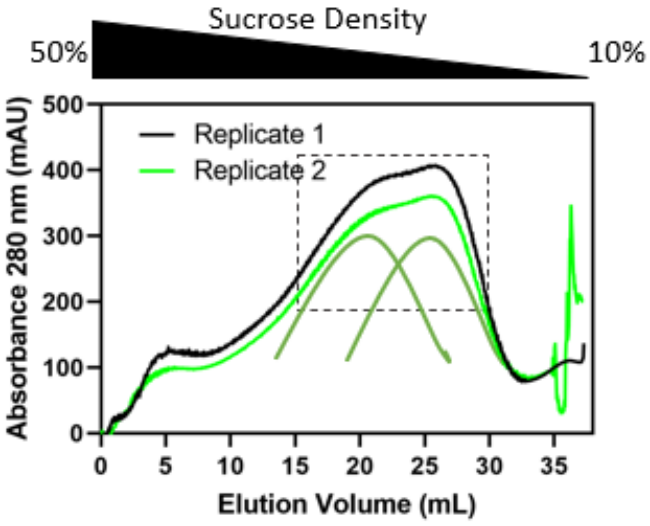


Figure 4.2 Absorbance profile of sucrose gradient of α-CBs purified from *E. coli*. Replicates 1 and 2 (black and green) are technical replicates of the same α-CB lysate sample. Each sucrose gradient tube was loaded with half the α-CB preparation and were separated on a 10-50% linear gradient. The grey box indicates the fractions that were pooled for further experimental use. The green curves indicate the hypothesized gaussian distributions of proteins as they are being fractionated.

As shown in Figure 4.1, both technical replicates show a similar absorbance profile with only slightly varying amount of protein. Figure S4.1 (a Polyacrylamide gel that summarizes the purification process) indicates a high amount of protein within the sample that was loaded onto the two sucrose gradients, corresponding to the high mAU 280 nm values in the chromatogram. However, after fractionation, white debris was observed which was later identified as likely being cell or membrane debris carried over from incomplete cell opening. This was determined through discussions with the David Savage lab as they are the first to attempt the purification (Bonacci *et al.*, 2012) and are more versed with the troubleshooting aspects of α-CB purification from *E. coli*. Due to incomplete lysis, these particulates sedimented with the a-CB and were therefore retained in the α-CB preparation. The use of lysozyme, as an alternative to the more stringent reagents proposed by the original protocol (So *et al.*, 2004), likely caused this issue. To mitigate any contamination, damage to equipment, and to maintain high quality samples for analysis, prior to any biophysical experiment the α-CBs were shaken overnight at 4˚C and then centrifuged at 13,000 x g for 15 minutes or more. Shaking encourages the α-CBs to dissolve and the centrifugation resulted in the removal of the particulates. Therefore, the samples were quantified before use and the respective yield is stated with every experiment as opposed to after purification was completed. As the α-CB samples were stored at -80˚C and the α-CBs are resistant to proteolysis (Li *et al.*, 2020), it was expected that no damage occurred due to the temporary storage with the cell debris.

Analysis of the fractionations corresponding to volume 12-30 mL, shown in Figure 4.1, indicate overlapping peaks. The peaks can be deconvoluted using two gaussian distributions, as indicated by the green curves in figure 4.2. I hypothesized that these two peaks indicated that the

α-CBs preparation contained two α-CB sub-populations. Both peaks were pooled and used for subsequent analysis. As a future alternative experiment, the peaks should be analyzed separately to determine the biophysical attributes of each subpopulation. Typically, proteins detected at a lower sucrose density corresponding to lower molecular mass would indicate a population of α-CBs that are not fully assembled.

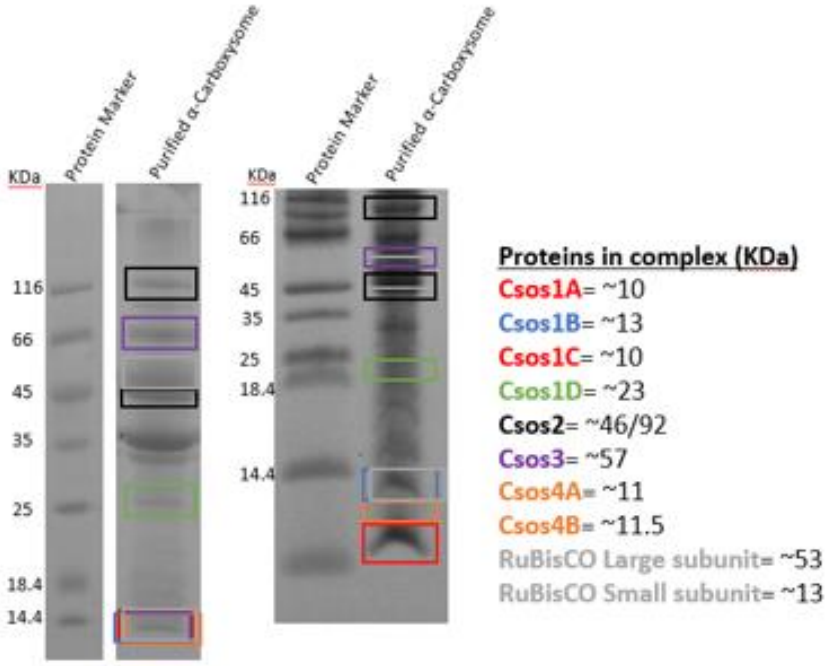


Figure 4.3 Polyacrylamide Gel analysis of the purification of α-CBs. (Left) A 12% SDS-PAGE of purified α-CBs. 10 µL of protein were loaded and separated at 180 V for ~45 minutes. **(Right)** A 15% Tris-Tricine gel of the same purified α-CB sample (10 µL 200 V for 1.5 hours). The gel was stained with Coomassie G-250. α-CB proteins were identified by size; each associated protein is indicated by a colored boxes that corresponding to the color-coded table on the right of the Figure. The α-CBs samples were used for subsequent mass spectroscopy analysis to determine the identification of other proteins within the sample. In gel digestion was not used.

Analysis using Polyacrylamide gels (Figure 4.3) allowed for the identification of some α-CB proteins by size. As there are multiple unidentified bands in the sample, mass spectrometry was performed to confirm the presence of α-CB proteins and to identify *E. coli* protein contaminants.

4.2.2 Mass Spectrometry Analysis of Purified α-Carboxysomes

With the purification of α-CBs and identification of α-CB proteins by size, a second validation step was needed to ensure that all α-CB components are present and are of high quality. Secondly, as Figure 4.3 shows multiple other bands, further analysis to identify protein contaminants was required. The obtained protein data is summarized in supplementary tables S4.1 and S4.2 respectively. All the α-CB shell proteins and cargo were detected. However, a large number of *E. coli* protein contaminants were also identified. As the mass spectrometry was performed using a volume of the total purification sample, all peptides present after purification could be determined. Using the Proteome Discoverer software (Thermo Fisher Scientific), peptides not corresponding to α-CB proteins were compared to the proteins of the *E. coli* K12 proteome. The specific identification of bands corresponding to unknown proteins in Figure 4.3 was not conducted, as only a total protein sample was analyzed and not band specific in-gel trypsin digestions. Therefore, no identification of the prominent band at 35 kDa, was possible. However, the mass spectrometry data in Table S4.2 suggests that the 35 kDa protein may be a membrane protein (accession number E2QJG1). The membrane protein has a molecular weight of 37.2 kDa and has a high peptide-spectrum match value (PSM) of 40 suggesting this protein is very prominent within the sample. Mass spectrometry was done after biophysical analysis, due to issues in access to an orbitrap mass spectrometer. Therefore, the biophysical analysis was performed on samples containing the above-described contaminants.

Gene ontology analysis of the *E. coli* proteins (Figure 4.4) indicates a range of contaminants from several different protein classes such as translation (which includes ribosomal proteins) and bacterial cytoskeletal proteins. Therefore, the data reported here may also contain the biophysical characterization of other *E. coli* biomolecules, proteins, and protein complexes.

All analysis hereafter is performed on a " α-CB sample" that may report size, shape, and mass distributions of both α-CBs and *E. coli* proteins. However, this data represents a critical analysis of the currents state of α-CB purification techniques and can form the basis for rational improvements to the purity.

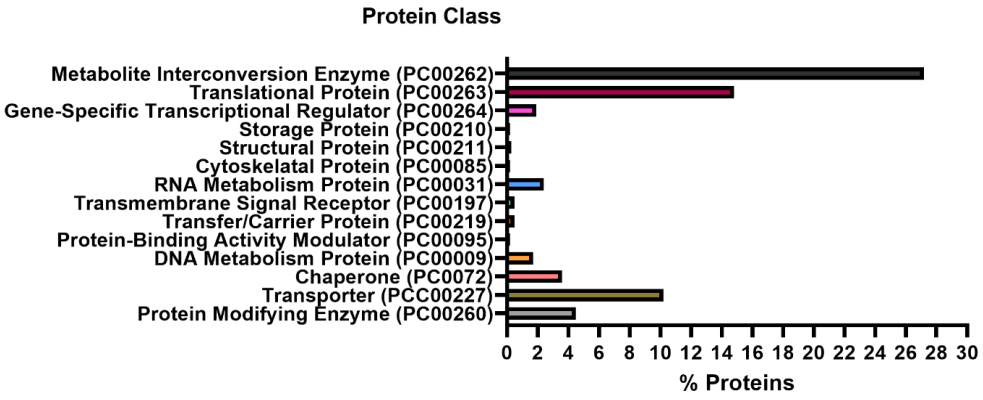


Figure 4.4 Gene ontology analysis of *E. coli* proteins found in α-CB samples. *E. coli* proteins were identified in the raw mass spectrometry file using the Proteome Discoverer software (Thermo Fisher Scientific). The proteins with more than one peptide hit were analyzed using the PANTHER classification system. The percentage of each protein contaminant in its protein class is represented in the bar graph (n=373).

4.2.3 TEM Analysis

To visually inspect the morphology and size distributions of the α-CB samples purified using the described approach**,** TEM was performed. Micrographs of negative stained α-CB samples were analysed to 1) confirm that the α-CB samples are properly assembled, 2) assess if any *E. coli* protein complexes (e.g., from the translation machinery) can be identified, and 3) provide insight into the size distribution of the particles present in the purified α-CB sample. A representative TEM image from the imaging experiment is shown in Figure 4.5.

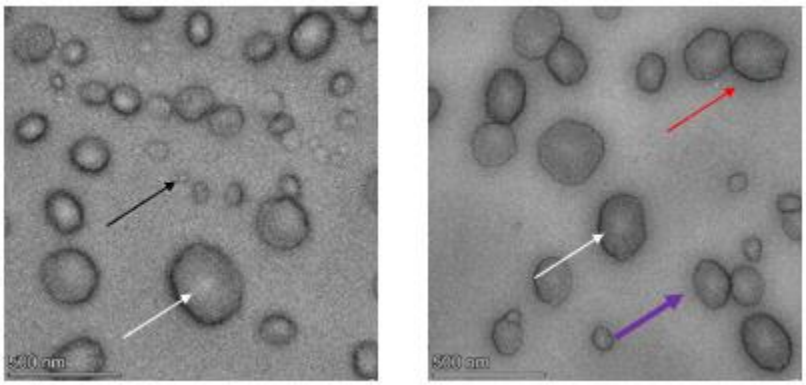


Figure 4.5 Transmission Electron Microscopy images of the α-CB sample. Samples were stained on carbon coated grids using uranyl acetate and imaged using a FEI Talos F2000X S/TEM microscope. The left image is at 360,000 X magnification (1554 pixel size). The right image is at 36,000 X magnification (1602 pixel size). The black arrow indicates RuBisCO octamers or contaminants, the red arrow indicates the expected α-CB shape. The thick purple arrow indicates an elongated particle, and white arrows indicate areas of particle collapse. n=52.

The very small particles in the obtained micrographs (Figure 4.5, black arrow) likely indicate RuBisCO octamers, suggesting that free RuBisCO is present due to the purification or the result of α-CB collapse during staining (Bonacci *et al.*, 2012). However, it cannot be excluded that these smaller particles may be contaminating *E. coli* proteins consistent with the Mass Spectrometry analysis (Figure S4.1 and S4.2). Constituents with the presence of free RuBisCO, the obtained micrographs suggest that a fraction of the particles are collapsed. Large particles can collapse on themselves, typically when they are dehydrated during the negative staining process. Such a collapse is common when imaging of BMCs or other large spherical particles (Kennedy *et al.*, 2020). All other structures observed in the micrographs are elongated or have shapes deviating from the expected icosahedral structures expected for α-CBs. The latter might also be due to the negative staining procedure.

The distribution of particle sizes reported in Figure 4.6 and Table S4.3 reveals a large range of particle diameters and has been previously observed during expression in the natural host

(Schmid *et al.*, 2006). As the loading of RuBisCO and the stoichiometry of shell proteins are amendable, the size and shape of the α-CB shells can be variable (Y. Sun *et al.*, 2019). This is likely due to the encapsulation and assembly mechanisms of the α-CB which are not completely understood. However, as distribution analysis has never been performed in recombinant expression, the extent of size variability is unknown. The average diameter for the particles seen in the TEM images is 122 ± 61 nm (n=52) consistent with the distribution reported by Bonacci et al., 2012, 136 ± 30 nm, respectively. Therefore, TEM indicated that there are α-CB-like particles present in the sample and the population has an average diameter comparable to previously published data. The size distributions shown in Figure 4.5 is also reminiscent the two sub-populations seen in the sucrose gradient ultracentrifugation experiment in Figure 4.1. The first population has a size of 0-120 nm and the second at 121-280 nm. The two groups of particle size (and corresponding mass) elute at a specific range of sucrose density, resulting in the two respective absorbance maxima in Figure 4.1.

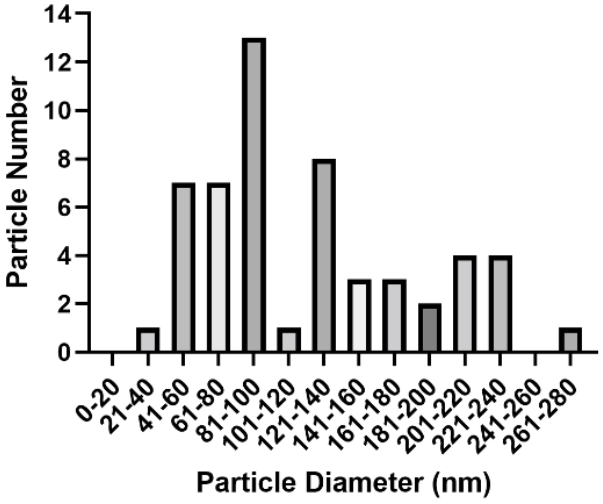


Figure 4.6 Size distribution of α-CB particles (n=52). Particles sizes were measured using the images in Figure 4.4 and then binned in 20 nanometer bins.

4.2.4 Biophysical Analysis of α-CBs using SEC-MALS

SEC-MALS experiments were performed 1) to identify and characterize assembled α-CBs and 2) to potentially identify and characterize other species within the sample. α-CB samples were injected onto a 10/300 (10 X 300 mm bed dimensions with a 24 mL bed volume) Size Exclusion Chromatography column with Sephacryl S400 resin (Cytiva) coupled to a ÄKTA Pure purification system connected to the DAWN Optilab apparatus for detecting light scattering of the particles within the sample. The largest limitation was identified when performing the experiment was the resolving power of the used chromatography material. The S400 sephacryl resin used as a part of the size exclusion chromatography step limits the method to resolving particles with a molecular weight between $2x10^4$ – $8x10^6$ Da. The α-CBs are much larger (with a reported average size of 250 MDa) than the exclusion limit of the column and therefore would likely elute in the void volume of the column. Figure 4.7 shows the chromatograms of the four replicates as they move through the purification system and enter the MALS detection system. The column volume (V_c) and void volume (V_o) of the column are given in the figure. The chromatograms confirm that the particles in the α-CB sample eluting from the column (particles of all sizes) are within the void volume (<8 ml of elution volume). The other peaks and their identities are further discussed in section 4.2.4.1. Despite limitations in resolution, particles could still be further analyzed to determine the average molecular mass and size.

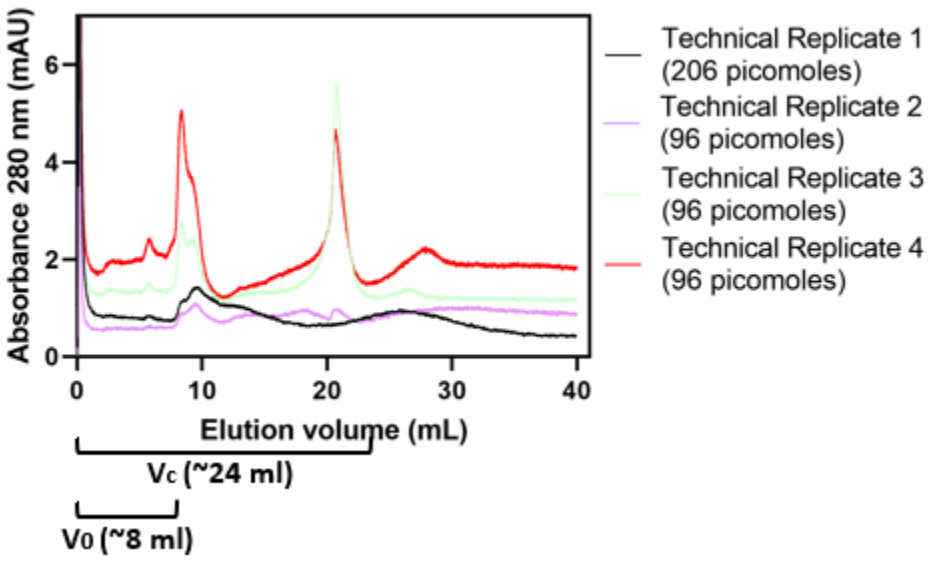


Figure 4.7 Chromatograms of α-CBs during MALS analysis. The absorbance values were collected by the UV detector on the ÄKTA Purification system coupled to the DAWN Optilab system. The void volume (V_0) and column volume (V_C) are indicated.

The proteins identified in the void volume were then analyzed using MALS with a summary of the data shown in Figure 4.8 with the light scattering, molecular weight, 280 nm absorbance, and RI shown. Replicate 2, although shown in Figure 4.7, contained extremely noisy data that maxed out the refractometer and therefore could not be used for further analysis (as indicated by the horizontal line in the spectral data in the chromatogram for this replicate in Figure 4.8). To assess the quality of light scattering data, it is important to first analyze the refractive index (RI) and its alignment to the light scattering data. The RI is used to determine that the molecular weight and hydrodynamic radius of the protein particles and as such, must be aligned to the light scattering values that it is derived from. Negative RI values within the data sets would indicate dissolved air which should have been removed from the buffer prior to the assay or from an absence of protein being eluted (Harvey, 2005). Replicates 3 and 4 indicate sufficient alignment of the light scattering and refractive index data while the first replicate is not

well aligned. The light scattering does show horizontal values perhaps suggesting that the concentration of the sample was exceeding the light scattering detection limit. This would also explain that in most of the light scattering data, the mass of the protein could not be determined.

Using the Astra software provided with the DAWN Optilab system and focusing on the proteins eluted in the void volume, which are expected to have α-CB-like particles, the molecular weight and the hydrodynamic radius of the peaks could be determined. It is unknown as to why, despite using technical replicates, that the elution profiles for each replicate is different. It can be hypothesized that the heterogeneity of the proteinaceous particles' sizes and mass within the samples may cause these inconsistencies. The RI has a scale of 0.0-1.0 and therefore, in order to compare the light scattering and 280 nm absorbance, these two values are also normalized to the 0.0-1.0 scale.

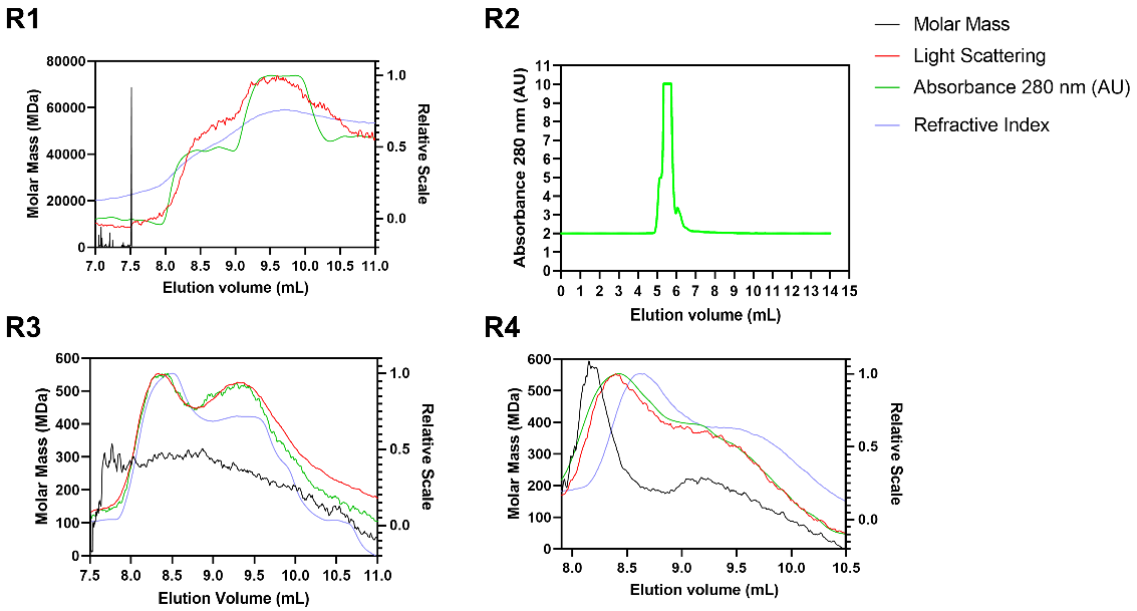


Figure 4.8 Light scattering, 280 nm absorbance, RI, and molecular weight of α-CB samples. Samples (206 picomoles for technical replicate 1 (R1) and ~96 picomoles for technical replicates 2,3 and 4 (R2,3,4)) were loaded on a S400 Sepharose 10/300 column coupled to the DAWN Optilab MALS system. The figures show a snapshot of the ~8-10 mL point of the chromatogram that is expected to contain fully α-CB-like particles.

The SEC-MALS data in Figure 4.8, indicates two spectral maxima at 280 nm. There is a trend of the molecular weight differences at the 8.5 and 9.5 mL peaks, averaging 255.6 MDa and 194.5 MDa respectively when analyzing replicates 3 and 4 (Table S4.5 and graphically shown in Figure 4.8). Molecular weight data could not be determined for replicate 1. These values are comparable to the average α-CB molecular weight of 250 MDa. The hydrodynamic radius is consistent between the two peaks, with a value around 122-123 nm, as indicated in Table S4.6. This is the same average diameter seen with the previously shown TEM experiment (Figure 4.5) that showed an average of 122 nm. Although there are two peaks in absorbance observed which may indicate different α-CB particles being eluted off the SEC column together, the lack of resolution from the column does not confirm this hypothesis. Overall, working with the heterogeneous α-CB samples in the SEC-MALS will provide only average size and mass and provides insufficient information for full characterization.

4.2.4.1 Identification of Low Molecular Weight Particles in SEC-MALS Experiments

The chromatograms from the SEC-MALS experiments (Figure 4.7) indicated smaller protein particles within the sample. When monitoring absorbance at 280 nm, indicating proteins containing aromatic amino acids, the elution profile shows maxima at ~21, 27 and 35 mL in the case of the first replicate (Figure 4.7), suggesting that there were other proteins (either associated with the α-CB or not). As there were indications of very small particles in the TEM data, we wanted to identify the protein particles and determine if they are small α-CBs or *E. coli* protein contaminants. To determine what these species of protein might be, several methods were used.

The first strategy used to identify that particle based on their size utilizing light scattering data. However, as shown in Tables S4.5 and S4.6, most absorbance maxima beyond the 10 mL void volume had unreliable light scattering data using the Lorenz-Mie model. Therefore, the

molecular weight and the hydrodynamic radius could not be confirmed as these proteins are much smaller in relation to the assumed assembled α-CBs indicated in the 280 nm absorbance profile in Figure 4.7; their concentration was likely not high enough to have sufficient light scattering. Size references were analyzed using SEC-MALS to determine if any of these peaks could be RuBisCO octamers or to determine a relative size of the particle (Figure 4.9).

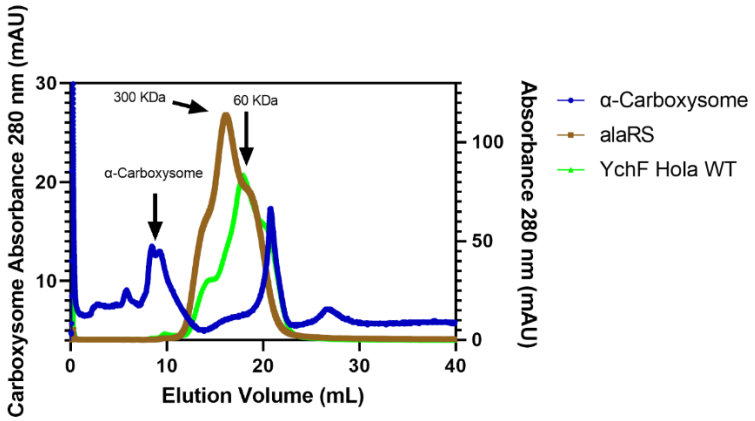


Figure 4.9 Size comparisons of α-CB sample to *E. coli* proteins using size exclusion chromatography. AlaRS and YchF (65,500 picomoles and 40,000 picomoles respectively) were injected on a S400 Sepharose 10/300 column, and the absorbance at 280nm was recorded using the ÄKTA Pure system coupled to the DAWN Optilab. The void volume (V_0) and column volume (V_C) are indicated. The technical replicate 3 from the α-CB SEC-MALS experiments was used for sample comparison.

The reference samples are alanine tRNA synthetase (AlaRS, 97.4 kDa) in its homotetrameric form (resulting in a size of ~ 300 kDa) and ribosomal binding ATPase protein YchF (60.5 kDa). Based on comparative sizes, the proteins eluting at ~21 mL and ~25 mL, as reflected by the respective absorbance peaks at 280nm, are close or past the end of the column volume (>10 kDa). As the RuBisCO octamers are expected to have a mass of ~560 kDa, the peaks at ~21 mL and ~35 mL are more likely to be smaller proteins such as Csos1A or Csos1C shell proteins with a molecular mass of ~10 kDa and could therefore be the proteins at ~21 mL. However, these peaks may also be *E. coli* contaminants within the sample. Attempts to manually

collect the fractions corresponding to the absorbance maxima at ~21 mL and ~35 mL from the SEC-MALS via the waste collection tube (using an α-CB sample with higher concentration than that used in the light scattering analysis described above) failed.

An additional experiment was performed to determine if the peaks are concentration dependent and/or if they are the result of the used equipment, such as the ÄKTA purification system (Figure 4.10). α-CB samples at different concentrations were analyzed using the SEC-MALS apparatus (20˚C) and on an ÄKTA Pure system (4˚C).

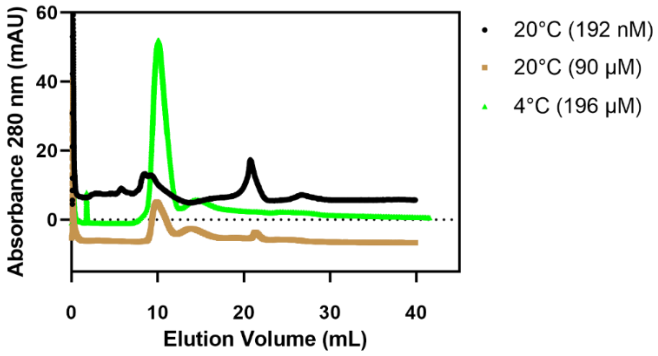


Figure 4.10 α-CB samples analyzed by different purification systems. α-CB samples at different concentrations were injected on the same S400 Sepharose 10/300 column and eluted at 0.4 mL/min. The column was then coupled to a ÄKTA Pure at 4˚C or to a ÄKTA Pure attached to the MALS DAWN Optilab apparatus at room temperature.

When comparing the three chromatograms with the 192 nM, 90 µM, and 196 µM α-CB samples, the data in Figure 4.10 suggests that the absorbance peak at ~20 mL is concentration dependent. As the protein concentration increases, the spectral maxima at ~21 mL get smaller. Despite the change of temperature for the 196 µM sample, both 20˚C samples show a potential concentration dependence. This is an unusual observation as concentration dependence would suggest that all spectral maxima would increase with the increase of protein concentration. This either suggests some type of human-error during the experiment or hints at a concentration-

dependent equilibrium favouring the formation of the complex or particle species eluting at 21 mL. Unfortunately, no replicate is available to confirm this observation. However, because this data set is incomplete (three concentrations of protein sample analyzed at each temperature, 4˚C and 20˚C), it is difficult to predict if temperature dependent behaviour may also be occurring with the size exclusion chromatography experiments. Although the molecular weight of the proteins eluting at ~21 and ~35 mL have a relative size of 10 kDa or smaller, as shown in Figure 4.9, the identity and their exact molecular weight is unknown. However, due to their size, they are confirmed to not be an assembled α-CB or RuBisCO octamer.

4.2.5 Analytical Ultracentrifugation Indicates α-CB Heterogeneity

AUC was used to further analyze the heterogenicity of the α-CB samples. The α-CB samples in TEMB buffer and the TEMB buffer alone were first tested on a spectrometer (Figure S4.4) to determine 1) if EDTA in the TEMB buffer is causing a signal that could interfere with the spectral analysis during the AUC experiment, 2) if there are signs of high amounts of aggregation, and 3) if there is enough protein content in the α-CB sample with respect to the AUC analysis. EDTA absorbs strongly at 230 nm and can absorb up to 260 nm and therefore may affect the absorption values of nucleic acids which absorb at 240 nm and 260 nm (Shen, 2019). Figure S4.2, left panel shows that the TEMB buffer alone does not absorb above 240 nm suggesting that the EDTA is not causing absorbance signals which can be misinterpreted as biomaterial. The sample themselves show no significant absorption above 300 nm, indicating that there is no aggregation of proteins.

The AUC experiments were all performed at 4˚C to prevent potential protein aggregation. The first experiment only recorded absorbance at 280 nm while the second experiment used multi-wavelength analysis recording absorbance between 210-310 nm. The two technical

replicates from the multi-wavelength analysis show the same sedimentation coefficient (s) profile (Figure 4.11 Panel A). The majority of particles identified fall between 100 and 500s, with some larger particles above 500s. Table S4.4 shows each identified particle with the corresponding size and molecular mass. The purified α-CBs sample therefore contains a multitude of proteinaceous species with a high diversity of sizes. The average α-CB molecular mass reported in the literature (250 MDa) corresponds to a sedimentation coefficient of 500-600s. However, Figure 4.11 panel B, suggests a sedimentation profile that is significantly different than the α-CB sample in panel A, with more species above 500s albeit at a lower concentration as indicated by the Y-axis.

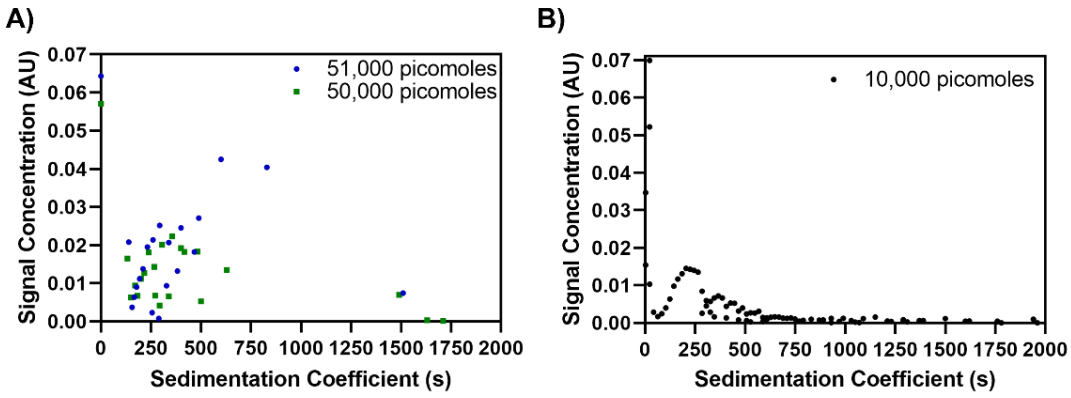


Figure 4.11 Sedimentation coefficient analysis of α-CBs samples. α-CBs (450μL) were analyzed using the AUC with **A)** 10,000 picomoles of protein for a single wavelength (280 nm) scan. **B)** Multi-wavelength analysis was conducted using 51,000 picomoles (replicate 1) and 50,000 picomoles (replicate 2) The concentration of a particle at a specific sedimentation coefficient is reported as its relative signal at 280 nm.

We moved on to analyzing the distribution of the masses of identified particles using the sedimentation coefficient from the AUC data (Figure 4.12). With 10, 000 picomoles of protein, ~20% have a mass larger than 300 MDa and 44% with a mass less than 100 MDa. At ~50,000 picomoles only 9% of particles have a mass less than 300 MDa and ~70% of the particles have a mass less than 100 MDa. The mass of the identified particles corresponding to their sedimentation coefficients is shown in Table S4.4. Although the interpretation of the data with

respect to the contribution of α-CBs is putative due to the above identified *E. coli* protein contamination, using AUC still allows to identify the biophysical attributes of individual particles within the sample.

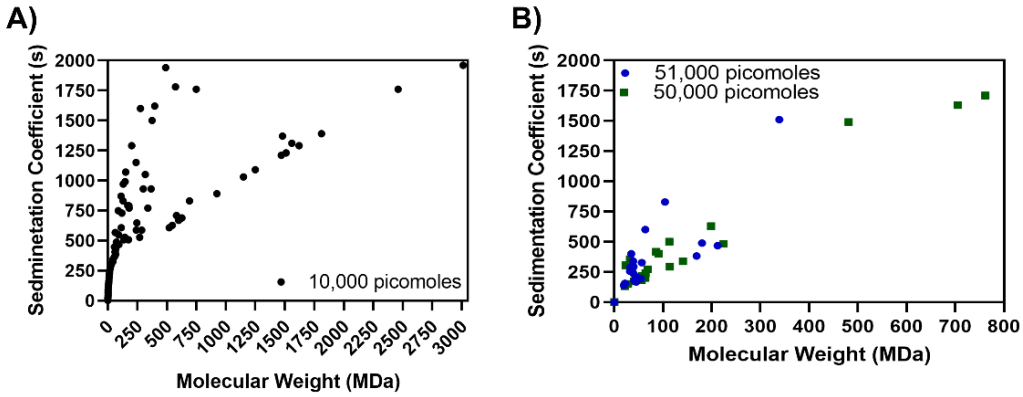


Figure 4.12 Molecular weight of different proteinaceous species in AUC sedimentation velocity experiments. A) 10,000 picomoles of protein **B)** 51,000 and 50,000 picomoles of protein. The X-axis indicates the determined molecular weight as the particles that were sedimented. The Y-axis indicates the corresponding sedimentation coefficients.

To further analyze the particle shapes in the α-CB samples, the frictional coefficients were determined using the AUC data. Analyzing the frictional coefficient can provide a relative idea of the particle structure. Frictional coefficients report how elongated a particle is, as it is centrifuged at high speeds. The value of coefficient indicates if the biomolecules shape is rod-like, spherical, etc. (Edwards *et al.*, 2020). Perfectly spherical particles have a frictional coefficient around 1 and particles with elongated or flexible structures have a coefficient of 2 or higher (Urban *et al.*, 2016). As shown in Figure 4.13, there is a distribution of frictional coefficient values between 1-6 with a value of 1 indicating globular or sphere-like structures as expected for the α-CB shell (Cannon *et al.*, 1983). Any frictional values higher than 2 may be further elongated shapes of α-CBs or caused by *E. coli* protein contaminants. With AUC, we are unable to specifically identify what each identified particle is and can only describe the sample as a whole. At 10,000

picomoles, there are 15 particles that have a frictional ratio higher than 4 compared to the ~50,000 picomole sample (panels A and B in Figure 4.13) which except for two particles, are 3 or lower. It is expected that most particles would have a frictional coefficient of 1-2. Therefore, the data suggests that there are a significant number of particles that are not within the expected frictional coefficient values of 1-2. This could be an effect of the centrifugation speeds used in the AUC experiment or are a feature of the *E.coli* contaminants.

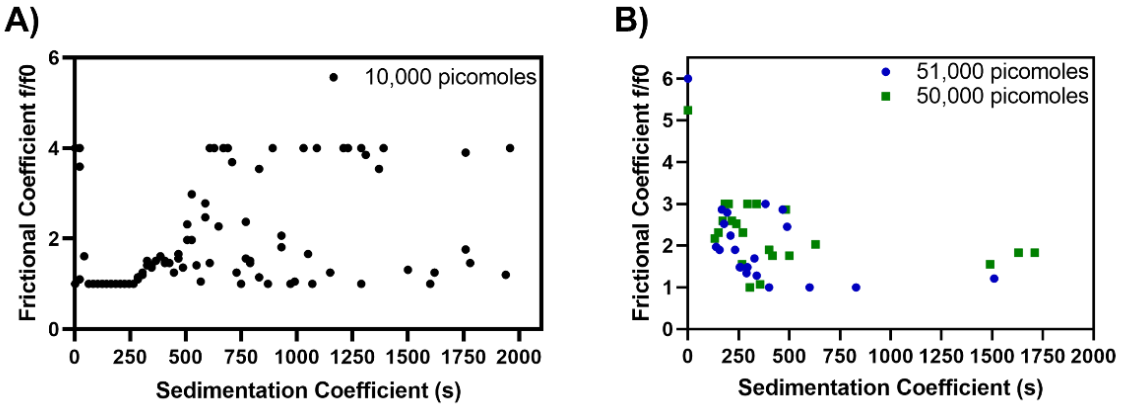


Figure 4.13 Frictional coefficient distributions of α-CB samples. A-CBs were studied at either **A)** 4˚C or **B)** 20˚C. The frictional coefficient on the X-axis describes the sedimentation coefficient of a particle while under centrifugal force, while the Y-axis indicates the frictional coefficient of the particle.

4.2.5.1 AUC Indicates Nucleic Acids in α-CB samples

During our initial AUC experiments using α-CB samples (Table S4.4), we observed an absorbance peak between ~240 and 260 nm, indicating the presence of nucleic acids. Using AUC multiwavelength analysis, we further analyzed the nucleic acids in the α-CB samples between 210-310 nm and subsequently at 240 and 280 nm. If there are interactions between the nucleic acids and the α-CB proteins they should sediment together and thus both be present at the same sedimentation coefficient. The comparison between wavelengths and sedimentation coefficient (Figure 4.14), indeed shows a potential correlation between the sedimentation of the protein and

nucleic acid, such as at ~23s and from 100-500s for both replicates. However, after this point, the nucleic acids identified in the sedimentation coefficient profile are not correlated to the protein set of data. The data alone suggests that the nucleic acids and α-CB particles are co-sedimenting (indicating similar size) but does not indicate any interactions between the two.

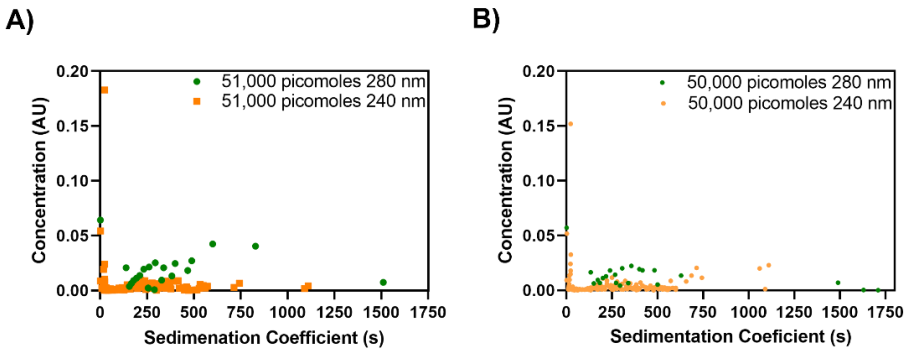


Figure 4.14 The presence of nucleic acids in α-CB samples. Multiwavelength analysis was performed on two technical replicates detecting tyrosine's in proteins (280 nm) and nucleic acids (240 nm) for **A)** sample at 51,000 picomoles of protein and **B)** 50,000 picomoles of protein. The X-axis indicates the sedimentation coefficients, and the Y-axis indicates the relative signal concentration which is a measure of the relative quantity of each type of particle sedimented to the bottom of the cell.

Molecular weight analysis (Figure 4.15 Panel A) indicates that the nucleic acids have sizes between 100-300 MDa. As some of the proteinaceous particles are of similar sizes (Table S4.4), it is possible that nucleic acids at this same size would purify with protein in solution without any interaction between them. The observed nucleic acid species between the two replicates vary in terms of mass, suggesting that they may not be highly stable within the buffer environment or during centrifugation, and that the nucleic acid themselves are unstable. Interestingly, we see that the nucleic acids also have multiple structures as suggested by the distribution of frictional coefficients (Figure 4.15 Panel B). Nucleic acids have a frictional coefficient of 4 or higher (Katsura *et al.*, 2000), but the nucleic acids in the purified α-CBs have species that are almost globular in structure, which have a frictional coefficient of 1-2. The high frictional coefficient

values may indicate that the nucleic acids are highly dense and coiled, for example, like genomic bacterial DNA. The observed nucleic acids may therefore be intact globular, and fragmented *E. coli* chromosomal DNA. Additional peaks at a coefficient of 4 or higher may be sheared nucleic acids resulting from purification and/or centrifugation (Katsura *et al.*, 2000).

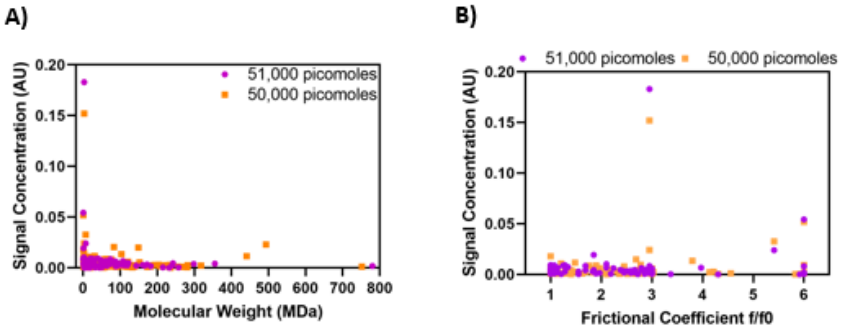


Figure 4.15 Molecular weight and frictional coefficient of nucleic acids in α-CB samples. Data was derived from a multi-wavelength analysis where nucleic acids were detected at 240 nm**.** The Y-axis indicates the relative signal concentration which is a measure of the relative quantity of each type of particle sedimented to the bottom of the cell. **A)** molecular weight based on sedimentation velocity experiments. **B)** Frictional coefficients based on sedimentation velocity experiments.

To further identify what type of nucleic acids (RNA or DNA) were in the purified α-CB, DNase I digest assays were performed. The assay involves the incubation of the α-CB sample with DNaseI followed by analysis of the reaction products on an agarose gel stained with ethidium bromide (Figure 4.16). If no degradation of the nucleic acid in the sample is observed, suggesting RNA as the nucleic acid, a RNase A treatment would be added to the assay to confirm the presence of RNA. If both treatments do not result in degradation of the nucleic acid, this would suggest that respective nucleic acid is not accessible to the enzymes and likely located inside of the respective particle. However, if DNaseI digestion of the sample is successful, it will indicate that the nucleic acid observed is DNA and that no DNA is encapsulated within, for

example the assembled α-CBs present, or is not being protected by any of the *E. coli* protein contaminants.

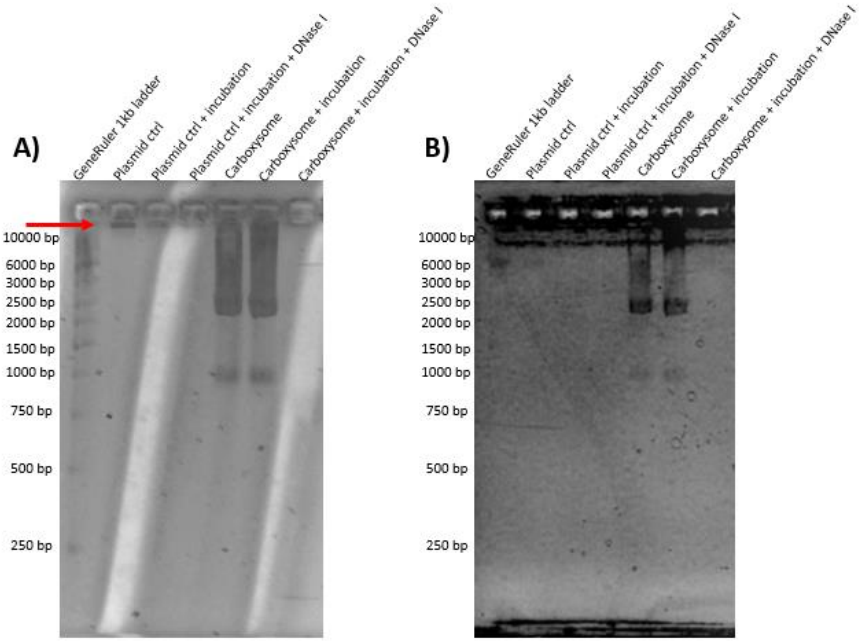


Figure 4.16 Treatment of α-CB samples with DNase I to determine the presence of DNA. α-CB samples were treated with DNase I and incubated at 37˚C overnight to ensure complete reaction. **A)** The agarose gel imaged with a Edvotek U.V transilluminator. **B)** The same gel imaged with a Axygen Gel Documentation System. Plasmid DNA was used as reaction controls and are indicated by the red arrow. α-CBs (15 µL) were analyzed on a 0.5% agarose gel at 16 V for 25 hours. The gel was stained with ethidium bromide.

Two imagers were used for gel analysis. The Edvotek transilluminator (Panel B) was used to provide enough resolution for the DNA ladder and the plasmid control, while the Axygen imager was used to provide better resolution of the individual bands of the DNA in the α-CB samples and to avoid the background lines caused by the transilluminator (Panel A). Incubation of the α-CB samples at 37˚C overnight in the absence of DNase I did not cause changes to the DNA within the protein sample, indicating the absence of DNase or RNase activity in the sample (e.g. from the *E. coli* contaminants). To ensure complete DNA digestion has occurred, DNaseI treatment was done overnight. Attempts at 2-4 hour showed incomplete digestion (data not

shown). Using overnight treatment with DNase I resulted in complete degradation of the nucleic acids in the α-CB sample (no DNA was observed in lane 7 of the agarose gel), therefore suggesting that DNA is the only nucleic acid present.

Many attempts were made to identify the sizes of the nucleic acids, but many of the large fragments were difficult to resolve or did not enter the gel, suggesting very large DNA. The use of a low voltage over a longer period of time (16 V for 25 hours) in combinations with the use of urea and 100˚C heat to unfold the α-CB protein, as well as other proteins prior to loading, improved the migration of the DNA and allowed for the best resolution of some of the smaller fragments in the gel.

Figure 4.16 indicates that there are large amounts of DNA fragments at ~900 base pairs and ~2,300 base pairs. The control plasmid is expected to be 7012 base pairs and therefore migrates a little higher than expected, indicating that it is either nicked or linear. The length of the larger fragments could not be defined and are only ~10,000 base pairs or larger as the resolution of these fragments as well as the ladder is poor. Because the plasmid expressing the α-CB operon is 13,218 bp, these DNA fragments could be plasmid contaminants from the α-CB purification. Comparing the DNase I assay to the AUC data, we can confirm that the nucleic acid observed is DNA and that there are multiple sizes of DNA fragments that correspond to the different nucleic acid species identified during the AUC experiment. The assay did not confirm that copurifying DNA is encapsulated into the α-CB or that there is a formation of a DNA-α-CB complexes.

4.3 Discussion

The α-CB is a highly complex and dynamic structure which has evolved to improve the carbon fixation cycle in phototrophic bacteria and cyanobacteria (Turmo *et al.*, 2017). As BMCs

such as the α-CBs are composed of multiple shell proteins and have complex biogenesis mechanisms, finding easy and accessible ways to characterize these BMCs will be beneficial for assessing *in vitro* samples for mechanistic studies or to characterize engineered variants. We tested TEM, SEC-MALS, and AUC biophysical methods to characterize the α-CB from *H. neapolitanus* when expressed and purified from *E. coli*. Use of other methods such as electron tomography (Schmid *et al.*, 2006) or atomic force microscopy (Sutter *et al.*, 2016) are limited by their cost, accessibility, and the need for extensive data analysis.

TEM allows for the analysis of large proteinaceous particles as it provides both a visual confirmation of large particles within the sample and allows a baseline assessment regarding the size distribution of the particles, with an average of 122 ± 61 nm. The TEM analysis also confirmed the observation that two sub-populations are present, first observed during the sucrose gradient ultracentrifugation during α-CB purification. Interestingly, the TEM analysis showed elongated structures, as indicated in Figure 4.5, or other structures that deviated from the typically icosahedral structures that α-CBs adopt. The first reason deviated structures may be seen is that the operon is expressed in *E. coli* instead of *H. neapolitanus*. Transcription in *E. coli* of the wild type α-CB operon from *H. neapolitanus* could result in changes in the number of shell and cargo proteins being produced due to cryptic regulatory elements. Consistent with this, the use of the native operon in the recombinant host has been hypothesized to cause changes to the protein expression and therefore affects the assembly of α-CB (Bonacci *et al.*, 2012). This may also relate to the deviated structures observed in the frictional coefficient analysis of α-CBs using AUC data (Figure 4.14). The second reason may be due to collapsing of particles with negative staining. The white areas in the middle of the particles indicated by the white arrows in Figure 4.4 suggest particle collapse. Particle collapse is a common occurrence with negative staining of

large particles such as α-CB's and other BMCs due to the dehydration of the sample during staining (Kennedy *et al.*, 2020). Unfortunately, particle collapse will affect the apparent diameter of the particle and the accuracy of the size distribution observed in the TEM experiments. The last reason for the observation of particles that are not consistent with the expected shape of the α-CB is that these observed particles are actually *E. coli* protein contaminants, or the result of interactions of these contaminants with the assembling α-CB. However, specific *E. coli* protein complexes have not been identified.

Despite the quality issues and particle collapse, the visual inspection of these diverse sizes can help access the sensitivity of the other biophysical methods such as AUC; if large size distribution of particles is seen with AUC, it's a complementary observation even if the data from the TEM (diameter) cannot be directly correlated to the AUC data.

The use of SEC-MALS is a highly accessible and high-quality method for determining the size and molecular weight of biomolecules. As the α-CB has an average mass of 250 MDa, the separation of protein species this large using the S400 Sepharose resin failed, as the resolution limit of this resin is 2000-8x10^6 Da. It was hypothesized that the SEC column would be able to resolve the smaller particles identified in the TEM experiment while the larger particles would be eluted in the void volume. However, the separation of all particles despite differences of size and mass was not observed. SEC-MALS, however, was still able to provide an average molecular weight (~225 MDa) and size (~122 ± 61 nm) of the particles that is comparable to the mass and size of α-CBs in the literature. However, these values are putative due to presence of *E. coli* protein contamination within the sample which also limits assessing the success of the different biophysical characterisation methods.

AUC provided a significant amount of detail regarding the protein species present in the α-CB sample, even though AUC may also be analyzing *E. coli* protein contaminants resulting from the standard purification strategy and which were identified by mass spectrometry. To improve the purification of the α-CB, an improvement in the cell opening (to ensure full lysis) coupled with a second purification step such as an ion exchange chromatography might improve the purity of the α-CB and remove a significant amount of the *E. coli* protein contaminants. The second limitation to this data is the differences in data that may be based on total protein concentrations (Figure 4.11). The differences in size distribution correlating to changes in the protein concentrations suggests that this experimental condition influences α-CB 3D-structure or assembly. However, without additional experimental replicates, and compounded by the issue of protein contaminants, the data remains inconclusive and limits the conclusions that we can draw from the data set. The second limitation is the presence of DNA. We cannot determine at this point, if the DNA is affecting the sedimentation of the α-CB particles and therefore affecting the collected data. Moving forward, improvement to the purity of the purification is required to ensure that any proteinaceous particles observed during the biophysical analysis can be attributed to be fully assembled α-CBs or its assembly intermediates. If successful at this stage, further optimization of TEM to decrease particle collapse, and optimization of AUC to determine if the observed concentration dependence is a consideration.

Based on the experiments performed using all three biophysical methods, TEM and AUC can indicate sample heterogeneity as well as detailed size-mass distributions and 3-D structural details of α-CBs. Through future work, while addressing the sample quality through improving the protein purification method and other experimental caveats (increasing the number of technical replicates and the removal of impurities such as DNA and *E. coli* proteins), the

biophysical data collected can be used to infer α-CB shapes, sizes and heterogenicity when expressed and purified from *E. coli*. Implications of this preliminary biophysical characterization will aid the use of BMCs in other applications. For example, in health applications, having consistent particle sizes is of critical importance when BMC are used as drug delivery systems. The work presented in this thesis therefore provides a basic framework for studying heterogeneous α-CBs using the two methods AUC and TEM.

4.3.1 Lessons Learned

Based on the detailed analysis using biophysical approaches, there are several ways to improve the data quality and experiments described. Improvement of purification by using a different sucrose gradient centrifugation approach (or multiple sucrose gradient purification steps), as well as ensuring complete cell lysis, can improve the separation of any misassembled α-CBs, avoid contamination with host proteins, and the formation of α-CB species that are not of interest (being large or small). The introduction of DNase I digestion during purification will help determine the influence of nucleic acids in purified α-CBs. As there is a possibility that the nucleic acids might affect the biophysical data of the proteinaceous species in the α-CB sample, as well as the assembly of the α-CB, the removal of nucleic acids would provide improved biophysical data.

If optimization of the purification method provides improved AUC data, the use of a fluorescent protein fused to the RuBisCO protein can help analyze the differences in RuBisCO cargo loading between different particles in the sample. The use of AUC with a fluorescence optics and detector can track the sedimentation of α-CBs with RuBisCO. The fluorescence can also be quantified, providing information on how much cargo is encapsulated. The use of fluorescence when fused to RuBisCO will indicate the variability of cargo loading and how that

may influence the heterogeneity of the produced α-CBs. Future experiments at lower centrifugal force may also help determine more detail about the shape and size of the assembly products present in the purified α-CB as they move through solution. Beyond the improvements for AUC experiments, improving the imaging of α-CBs will also be a benefit. Negative stain TEM imaging is limited to providing information on particle size and does not provide information regarding the detail in molecular structure. Moving towards single particle reconstruction-based imaging methods using Cryo-electron microscopy (Cryo-EM) can provide higher detail of the α-CBs 3-dimensional structure. Cryo-EM of a minimal α-CBs has recently been reported, including preliminary work towards obtaining a three-dimension structure (Evans, 2022).

4.4 Supplementary Figures

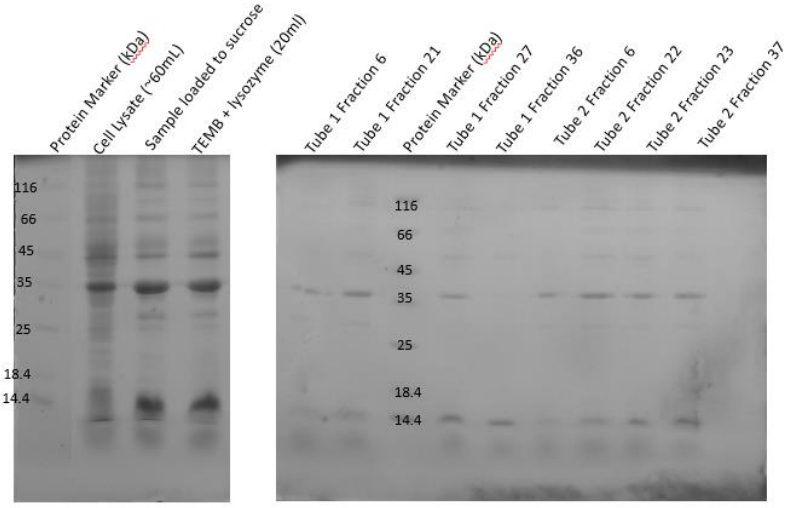


Figure S4.1 α-CBs purified from *E. coli*. The two gels presented show **Left)** the cell opening steps of the *E. coli* pellets containing overexpressed α-CBs. **Right)** Samples from the fractionation of α-CBs separated using sucrose gradient centrifugation (see Figure 4.1). Fraction numbers correspond to the volume at the sample was located in the gradient. Samples were resolved on a 12% SDS-PAGE at 180V for 45 minutes and stained with Coomassie G-250.

Table S4.1 α-Carboxysome proteins identified in mass spectrometry experiments. Purified α-CBs were injected onto a Orbitrap liquid column mass spectrometer. The raw data was analyzed using the Proteome Discoverer (Thermo Fisher Scientific) by comparing the raw data to the expected α-CB protein sequences. The Accession number is the Uniprot ID of the protein, the coverage % indicates the amount of total Protein sequence identified by mass spectrometry, # peptides being the number of identified peptides by mass spectrometry, # PSMs indicating the number of peptide spectra matching the identified protein, # unique peptides corresponding to the peptide sequences unique to the peptide group and Score Sequest HT being the overall score indicating the quality of protein identification based on the mass spectrometry data.

Accession	Description (# of amino acids; molecular weight kDa)	Coverage [%]	# Peptides	# PSMs	# Unique Peptides	Score Sequest HT
P45689	Major carboxysome shell protein CsoS1A (99;10)	76	9	506	1	1354
P45690	Carboxysome shell protein CsoS1B (110;11.3)	54	6	373	1	915
P45688	Carboxysome shell protein CsoS1C (98;9.9)	76	9	502	1	1342
D0KZ73	Carboxysome shell protein CsoS1D (213;23.4)	58	10	11	10	25.
O85041	Carboxysome assembly protein CsoS2B (869;91.9)	74	56	326	56	883
O85042	Carboxysome shell carbonic anhydrase (514;57.3)	60	23	45	23	113
O85043	Carboxysome shell vertex protein CsoS4A (83;8.9)	31	2	4	2	4
O85044	Carboxysome shell vertex protein CsoS4B (81;8.8)	19	2	4	2	6
O85040	Ribulose bisphosphate carboxylase large chain (473;52.6)	67	33	498	33	1121
P45686	Ribulose bisphosphate carboxylase small subunit (110;12.8)	93	8	50	8	113

Table S4.2 Contaminant *E. coli* proteins identified in mass spectrometry experiments. α-CB samples were injected onto a Orbitrap liquid column mass spectrometer. The raw data was analyzed using the Proteome Discoverer (Thermo Fisher Scientific) by comparing the raw data to the *E. coli* K12 proteome. The Accession number is the Uniprot ID of the protein, the coverage % indicates the amount of total Protein sequence identified by mass spectrometry, # peptides being the number of identified peptides by mass spectrometry, # PSMs indicating the number of peptide spectra matching the identified protein, # unique peptides corresponding to the peptide sequences unique to the peptide group and Score Sequest HT being the overall score indicating the quality of protein identification based on the mass spectrometry data.

Accession	Description of *E. coli* proteins (# of amino acids; molecular weight kDa)	Coverage [%]	# Peptides	# PSMs	# Unique Peptides	Score Sequest HT
C3TRK2	Chaperone protein DnaK (638;69.1)	73	47	126	47	47
C3SL97	ATP synthase subunit alpha (513;55.2)	64	30	65	29	30
C3TIN7	Succinate dehydrogenase (588;64.4)	62	25	58	25	25

Q548M1	60 kDa chaperonin (548; 57.3)	60	25	57	25	25
C3SLA7	ATP synthase subunit beta (460; 50.3)	65	23	51	23	23
E2QJQ6	Bifunctional protein putA (1320;143.7)	30	30	41	30	30
E2QJG1	Membrane protein (346; 37.2)	41	12	40	12	12
C3TGB2	30S ribosomal protein S1 (557;61.1)	52	25	37	25	25
E2QPE4	NadH-quinone oxidoreductase subunit G (910;100.4)	40	25	39	25	25
E2QJ13	DNA-directed RNA polymerase subunit beta (1342;150.6)	32	33	37	33	33
E2QJ06	Elongation factor Tu (394;43.3)	64	20	35	20	20
E2QFJ4	Elongation factor Tu (394;43.3)	64	20	35	20	20
C3SIA2	DNA-directed RNA polymerase subunit beta' (1407; 155.1)	31	34	38	34	34
C3SSK2	ATP-dependent zinc metalloprotease FtsH (644;70.7)	43	22	33	22	22
E2QEJ9	Polyribonucleotide nucleotidyltransferase (711;77.1)	38	21	32	21	21
E2QGA5	Lactose operon repressor (360;38.6)	53	14	26	14	14
E2QPE7	NADH-quinone oxidoreductase subunit C/D (596;68.2)	47	20	25	20	20
C3TIN2	Succinate dehydrogenase iron-sulfur subunit (238;26.8)	64	13	24	13	13
C3TPJ2	Outer membrane protein assembly factor BamA	27	20	26	20	20
C3SR02	50S ribosomal protein L5 (179;20.3)	58	12	22	12	12
C3SQX2	30S ribosomal protein S3 (233;25.97)	49	10	23	10	10
E2QPE5	NADH dehydrogenase (445;49.3)	40	15	21	15	15
E2QKG8	Ribonuclease E (1061;118.3)	23	19	24	19	19
C3SQV7	50S ribosomal protein L2 (273;29.8)	40	8	19	8	8
C3SX72	Enolase (432;45.6)	45	13	19	13	13
C3SKQ2	Transcription termination factor Rho (419;46.63)	45	17	23	17	17
C3SLR2	Small heat shock protein IbpA (137;15.77)	61	6	17	6	6
C3SIC2	50S ribosomal protein L1 (234;24.8)	49	12	18	12	12
E2QHM5	Small heat shock protein IbpB (142; 16.1)	61	8	21	8	8
C3TQA2	Dihydrolipoyl dehydrogenase (474;50.7)	43	14	20	14	14
C3SR27	30S ribosomal protein S5 (167;17.6)	62	8	18	8	8
E2QFQ1	33 kDa chaperonin (294;32.7)	44	7	13	7	7
E2QPV2	Outer membrane protein assembly factor BamC (344;36.8)	59	12	15	12	12
C3TPN2	30S ribosomal protein S2 (241;26.7)	46	9	17	9	9
E2QGA4	Beta-galactosidase (1024;116.3)	19	18	19	18	18
Q7BGE6	10 kDa chaperonin (97;116.3)	65	6	17	6	6
E2QGW3	L-lactate dehydrogenase (396;42.7)	35	11	17	11	11
C3SR67	DNA-directed RNA polymerase subunit alpha (329;36.5)	55	12	16	12	12
C3SR62	30S ribosomal protein S4 (206;23.5)	37	8	16	8	8

C3SR52	30S ribosomal protein S13 (118;13.1)	62	8	16	8	8
C3SR17	50S ribosomal protein L6 (177;18.9)	59	10	17	10	10
C3SFP2	50S ribosomal protein L9 (149;15.6)	51	7	15	7	7
C3SLA2	ATP synthase gamma chain (287;31.6)	53	13	15	13	13
B9VUA5	Acriflavine resistance protein A (397;42.1)	39	11	13	11	11
E2QPE6	NADH dehydrogenase (166;18.6)	69	8	15	8	8
Q3HSD9	Peptidylprolyl isomerase (623;68.1)	24	10	14	10	10
E2QF45	Acetyltransferase component of pyruvate dehydrogenase complex (630;66.1)	26	15	16	15	15
C3SIB2	50S ribosomal protein L7/L12 (121;12.3)	50	6	19	6	6
C3SQS7	Elongation factor G (704;77.5)	27	12	13	12	12
E2QI99	2-oxoglutarate dehydrogenase E1 component (933;105)	18	12	13	12	12
Q5UES8	Tryptophanase (476;53.3)	36	14	16	14	14
C3SR37	50S ribosomal protein L15 (144;14.96)	57	9	14	9	9
C3SQU7	50S ribosomal protein L4 (201;22.1)	45	7	12	7	7
E2QGI0	Trigger factor (432;48.2)	25	10	13	10	10
C3TCN2	DNA-binding protein (137;15.5)	61	8	12	8	8
E2QHS4	ATP synthase subunit b (156;17.2)	47	8	13	8	8
E2QI94	Citrate synthase (427;47.98)	23	8	12	8	8
E2QIV8	ATP-dependent protease ATPase subunit HslU (443;49.6)	33	13	15	13	13
C3TIH2	Protein TolB (430;45.9)	36	10	12	10	10
E2QF90	Protease do (474;49.3)	31	10	13	10	10
C3TIL2	Succinyl-CoA ligase [ADP-forming] subunit beta (388;41.4)	39	11	13	11	11
E2QGF2	Protein translocase subunit SecD (615;66.6)	19	11	14	10	11
E2QPX8	Outer membrane protein assembly factor BamB (392;41.9)	34	9	10	9	9
E2QNW3	D-tagatose-1,6-bisphosphate aldolase subunit GatZ (420;46.95)	21	5	10	5	5
E2QK11	Alpha-galactosidase (451;50.6)	19	7	11	7	7
E2QK93	Fumarate reductase (602;65.95)	30	12	13	12	12
C3SQW7	50S ribosomal protein L22 (110;12.2)	52	5	10	5	5
C3SR72	50S ribosomal protein L17 (127;14.4)	46	7	13	7	7
C3SFQ7	30S ribosomal protein S6 (131;15.2)	63	7	12	7	7
C3SQZ2	50S ribosomal protein L14 (123;13.5)	50	5	13	5	5
C3T6W2	Glyceraldehyde-3-phosphate dehydrogenase (331;35.51)	38	9	10	9	9
C3TLA7	Chaperone protein HtpG (624;71.38)	26	13	13	13	13
C3TIK7	Succinyl-CoA ligase [ADP-forming] subunit alpha (289;29.8)	51	11	11	11	11
E2QF44	Pyruvate dehydrogenase E1 component (887;99.6)	19	12	12	12	12
E2QF48	Aconitate hydratase 2 (865;93.4)	15	11	13	11	11

C3SSN7	Transcription elongation factor NusA (495;54.8)	27	11	12	11	11
C3SQU2	50S ribosomal protein L3 (209;22.23)	43	8	9	8	8
C3SRX7	50S ribosomal protein L13 (142;16.0)	54	6	11	6	6
C3SQN2	Peptidyl-prolyl cis-trans isomerase (196;20.8)	41	4	9	4	4
E2QDT7	UPF0301 protein YqgE (187;20.8)	37	6	11	6	6
Q548B5	21 kDa hemolysin (191;20.0)	48	7	9	7	7
C3SQS2	30S ribosomal protein S7 (156;17.6)	57	7	9	7	7
E2QMR3	Phosphoenolpyruvate synthase (792;87.37)	17	11	11	11	11
C3SIB7	50S ribosomal protein L10 (165;17.7)	55	7	9	7	7
E2QQ54	Outer membrane protein assembly factor BamD (245;27.8)	33	6	10	6	6
E2QDR0	Phosphoglycerate kinase (387;41.1)	39	10	10	10	10
C3SQW2	30S ribosomal protein S19 (92;10.4)	49	5	8	5	5
C3SZ18	ATP-dependent RNA helicase SrmB (444;49.9)	27	8	9	8	8
E2QFG6	50S ribosomal protein L24 (104;11.3)	39	6	9	6	6
E2QGV5	PTS system mannitol-specific EIICBA component (637;67.9)	21	9	9	9	9
E2QJB4	Formate acetyltransferase (760;85.3)	14	9	10	9	9
C3TJK2	D-alanyl-D-alanine carboxypeptidase (403;44.4)	30	8	10	8	8
E2QGX6	ADP-L-glycero-D-manno-heptose-6-epimerase (310;34.9)	33	9	10	9	9
E2QDQ9	Fructose-bisphosphate aldolase (359;39.2)	39	7	7	7	7
E2QES6	MreB, subunit of longitudinal peptidoglycan synthesis/chromosome segregation-directing complex (347;36.9)	24	6	9	6	6
C3TM82	Membrane protein YajC (110;11.9)	28	5	9	5	5
C3TRH7	30S ribosomal protein S20 (87;9.7)	43	4	10	4	4
C3SVY0	Lysine--tRNA ligase (505;57.6)	22	9	9	8	9
E2QNW2	Galactitol-specific phosphotransferase enzyme IIA (150;16.9)	44	4	8	4	4
C3SYM7	Protein GrpE (197;21.8)	44	5	7	5	5
C3TGI7	Serine--tRNA ligase (430;48.4)	26	9	9	9	9
C3SR12	30S ribosomal protein S8 (130;14.1)	41	5	9	5	5
E2QH56	AcrB RND-type permease, subunit of AcrAB-TolC multidrug efflux transport system (1049;113.5)	11	8	8	8	8
C3T982	Protein YdgH (314;33.9)	35	9	9	9	9
E2QFX6	DcrB protein (185;19.8)	43	7	8	7	7
E2QFC4	Membrane protein (134;14.2)	60	5	7	5	5
C3TQK2	Cell division protein FtsZ (383;40.3)	26	8	8	8	8
E2QIA0	Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex (405;44.0)	22	7	8	7	7

E2QPI0	Semialdehyde dehydrogenase (337;36.4)	27	8	9	8	8
E2QLN3	DNA topoisomerase 1 (865;97.3)	14	8	8	8	8
E2QIY6	Membrane protein (577;66.6)	16	9	9	9	9
E2QHL4	tRNA-2-methylthio-N(6)-dimethylallyladenosine synthase (474;53.6)	23	8	8	8	8
E2QEQ9	Conserved protein (132;19.95)	39	4	6	4	4
C3TJX7	AhpC component, subunit of alkylhydroperoxide reductase (187;20.7)	45	7	8	7	7
E2QN35	RNA chaperone ProQ (232;25.9)	38	6	7	6	6
C3SI42	DNA-binding protein (90;9.5)	58	4	6	4	4
C3UV71	Predicted rhodanese-related sulfurtransferase (143;15.6)	50	5	7	5	5
E2QJC5	Lipid A export ATP-binding/permease protein MsbA (582;64.5)	19	7	7	7	7
E2QFC2	Proline--tRNA ligase (572;63.6)	20	7	7	7	7
E2QEJ7	ATP-dependent RNA helicase DeaD (645;72.5)	16	7	7	7	7
C3T2E7	NADH-quinone oxidoreductase subunit I (180;20.5)	39	7	7	7	7
C3T2B7	NADH-quinone oxidoreductase subunit B (220;25.0)	33	6	7	6	6
C3TAU7	Predicted lipoprotein (222;24.4)	34	6	7	6	6
E2QMH7	Protein ydgA (502;54.6)	19	7	8	7	7
E2QMH5	Fumarase A (548;60.3)	18	7	7	7	7
C3SL92	ATP synthase subunit delta (177;19.3)	51	4	5	4	4
C3T8I7	Glutaredoxin (115;12.9)	46	4	6	4	4
C3SQX7	50S ribosomal protein L16 (136;15.3)	32	3	6	3	3
E2QDS3	Transketolase (663;72.1)	13	8	8	8	8
C3TDS2	Penicillin-binding protein activator LpoB (213;22.5)	40	6	7	6	6
C3SRY2	30S ribosomal protein S9 (130;14.8)	36	5	8	5	5
C3T8H7	Superoxide dismutase (193;21.3)	25	3	4	3	3
C3SQT7	30S ribosomal protein S10 (103;11.7)	41	4	6	4	4
E2QGH6	Cytochrome o ubiquinol oxidase subunit II (315;34.9)	17	3	5	3	3
E2QJQ7	Proline:sodium symporter PutP (502;54.3)	3	2	5	2	2
E2QHX6	Uroporphyrinogen-III C-methyltransferase (393;42.9)	19	7	7	7	7
C3TLW2	Cytochrome O ubiquinol oxidase subunit I (663;74.3)	5	2	5	2	2
E2QPB7	Lipopolysaccharide core heptose(II)-phosphate phosphatase (200;22.4)	34	5	6	5	5
E2QDP1	Glycine dehydrogenase (decarboxylating) (957;104.3)	9	6	6	6	6
C3SYP2	30S ribosomal protein S16 (82;9.2)	57	6	6	6	6
E2QKA8	Protein hflK (419;45.5)	15	6	7	6	6
E2QI58	PTS N-acetyl glucosamine transporter subunits IIABC (648;68.3)	10	6	7	6	6

E2QEI3	Penicillin-binding protein activator LpoA (678;72.9)	12	7	7	7	7
C3SRV3	Malate dehydrogenase (312;32.3)	27	6	6	6	6
C3TLB7	Nucleoid-associated protein YbaB (109;12.0)	51	4	6	4	4
C3T992	NAD(P) transhydrogenase subunit beta (462;48.7)	17	4	4	4	4
E2QK06	UPF0141 membrane protein yjdB (547;61.7)	11	5	6	5	5
C3SPR7	Glycogen synthase (477;52.8)	18	7	7	7	7
E2QGR4	Inner membrane lipoprotein yiad (219;22.2)	16	3	6	3	3
E2QF93	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase (274;29.9)	21	5	6	5	5
E2QHV7	ATP-dependent RNA helicase RhlB (421;47.1)	20	6	6	6	6
E2QNV9	Galactitol-1-phosphate 5-dehydrogenase (346;37.4)	12	4	6	4	4
C3SIC7	50S ribosomal protein L11 (142;14.9)	28	4	6	4	4
C3SME2	50S ribosomal protein L28 (78;9.0)	41	5	7	5	5
E2QMG6	NAD(P) transhydrogenase subunit alpha (510;54.6)	13	6	6	6	6
E2QPY2	Cytoskeleton protein RodZ (336;36.2)	18	4	5	4	4
C3TRJ7	Chaperone protein DnaJ (376;41.0)	19	6	6	6	6
E2QJE0	Outer membrane protein F (362;39.3)	20	6	6	6	6
E2QFN8	Protein damX (428;46.0)	21	5	5	5	5
E2QE97	Transaldolase (317;35.2)	25	6	6	6	6
C3SM87	DNA-directed RNA polymerase subunit omega (91;10.2)	70	4	4	4	4
C3SIV2	50S ribosomal protein L31 (70;7.9)	53	4	5	4	4
C3TPM7	Elongation factor Ts (283;30.4)	22	6	6	6	6
E2QPB9	Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase (322;36.2)	20	5	5	5	5
E2QQ51	Chaperone clpB (823;91.7)	11	6	6	6	6
E2QMU7	Osmotically-inducible lipoprotein E (112;12.0)	36	3	5	3	3
E2QFC5	Lipoprotein (271;29.4)	29	5	5	5	5
E2QJT3	Glucans biosynthesis glucosyltransferase H (847;96.9)	8	5	6	5	5
C3SSB7	ABC transporter ATP-binding protein (269;29.1)	26	4	5	4	4
E2QQJ4	Lipoprotein (379;40.1)	17	5	5	5	5
C3TJK7	UPF0250 protein YbeD (87;9.8)	55	3	5	3	3
E2QPN9	Glutamate--tRNA ligase (471;53.8)	14	5	5	5	5
E2QPZ6	Cysteine desulfurase IscS (404;45.1)	18	5	5	5	5
E2QH68	Adenylate kinase (214;23.5)	25	6	6	6	6
E2QGS0	Glycine--tRNA ligase beta subunit (689;76.7)	11	6	6	6	6
E2QLR3	Phage shock protein (222; 25.5)	24	4	4	4	4

E2QGX4	L-threonine 3-dehydrogenase (341;37.3)	18	5	5	5	5
C3SSQ7	30S ribosomal protein S15 (89;10.3)	43	3	4	3	3
E2QEB5	RNA polymerase sigma factor RpoD (613;70.2)	9	5	5	5	5
C3SLK2	Membrane protein insertase YidC (548;61.5)	11	4	5	4	4
Q6KCW9	ABC transporter permease (251;28.0)	24	4	4	4	4
C3T6A2	Cold shock protein (69;7.4)	78	4	5	4	4
C3TBD7	Uncharacterized protein (316;35.7)	21	5	5	5	5
C3TBH7	Probable thiol peroxidase (168;17.8)	39	4	5	4	4
E2QG35	Aminoacyl-histidine dipeptidase (485;52.9)	12	5	6	5	5
E2QP51	50S ribosomal protein L25 (94;10.7)	39	4	5	4	4
Q93R41	Isocitrate dehydrogenase [NADP] (416;45.7)	16	5	5	5	5
C3SYQ7	50S ribosomal protein L19 (115;13.1)	41	4	4	4	4
E2QN19	PTS mannose transporter subunit IIAB (323;35.1)	19	4	4	4	4
C3T887	Major outer membrane lipoprotein (78;8.3)	42	3	4	3	3
E2QK92	Succinate dehydrogenase iron-sulfur subunit (244;27.1)	20	4	5	4	4
E2QK76	Aspartate ammonia-lyase (478;52.3)	14	5	5	5	5
E2QQK7	Sulfite reductase [NADPH] hemoprotein beta-component (570;63.9)	11	5	5	5	5
E2QL02	Peptidyl-prolyl cis-trans isomerase (206;22.2)	24	5	5	5	5
C3SVB2	Biosynthetic arginine decarboxylase (658;73.8)	7	3	4	3	3
E2QJ37	Isocitrate lyase (434;47.5)	16	5	5	5	5
C3SYU7	Putative yhbH sigma 54 modulator (113;12.8)	41	3	4	3	3
Q0P6M2	RNA polymerase sigma factor (191;21.7)	28	3	4	3	3
C3TLT2	ATP-dependent Clp protease proteolytic subunit (207;23.2)	29	3	4	3	3
C3SIY7	Regulator of ribonuclease activity A (161;17.4)	20	4	5	4	4
C3T6W7	Peptide methionine sulfoxide reductase MsrB (137;15.4)	47	3	3	3	3
C3T7Q7	50S ribosomal protein L20 (118;13.5)	23	3	5	3	3
E2QFQ3	Phosphoenolpyruvate carboxykinase [ATP] (540;59.6)	12	5	5	5	5
E2QGR6	Uncharacterized protein yiaF (276;30.1)	18	4	4	4	4
E2QHJ3	Rare lipoprotein A (362;37.5)	17	4	4	4	4
E2QP97	DNA gyrase subunit A (875;96.9)	8	5	5	5	5
E2QPH0	Histidine-binding periplasmic protein (260;28.5)	22	4	4	4	4
C3SS87	Lipopolysaccharide ABC transporter, ATP-binding protein LptB (241;26.8)	23	4	4	4	4
E2QLX3	Aldehyde dehydrogenase A (479;52.3)	11	5	5	5	5
E2QL31	PTS system trehalose(Maltose)-specific transporter subunits IIBC (472;50.9)	9	4	5	4	4

C3SSG2	50S ribosomal protein L21 (103;11.6)	36	4	5	4	4
E2QFX3	Lead, cadmium, zinc and mercury-transporting ATPase (732;76.8)	9	5	5	5	5
E2QK21	Lysine--tRNA ligase (505;57.8)	10	4	4	3	4
C3TGX2	ATP-binding component of 3rd arginine transport system (242;27.0)	23	3	3	3	3
E2QQQ9	Membrane-bound lytic murein transglycosylase A (365;40.4)	14	3	3	3	3
E2QF26	Protein translocase subunit SecA (901;101.9)	7	5	5	5	5
C3SFP7	30S ribosomal protein S18 (75;8.98)	48	4	4	4	4
C3SLB2	ATP synthase epsilon chain (139;15.1)	43	3	3	3	3
C3SY12	Protein RecA (353;37.95)	14	4	4	4	4
C3TH97	D-alanyl-D-alanine carboxypeptidase penicillin-binding protein 6 (400;43.6)	15	3	3	3	3
C3TPI7	Chaperone protein skp (161;17.7)	27	4	4	4	4
C3TR42	Chaperone SurA (428;47.3)	11	4	4	4	4
Q14F07	Thioredoxin 1 (144;16.0)	19	2	3	2	2
E2QFC7	Methionine import ATP-binding protein MetN (343;37.8)	18	4	4	4	4
E2QGW7	Protein-export protein SecB (155;17.3)	38	3	4	3	3
C3STZ7	30S ribosomal protein S21 (71;8.5)	32	4	4	4	4
E2QK81	Conserved protein (117;11.95)	31	2	3	2	2
E2QPT7	NadP-dependent malic enzyme (759;82.4)	8	4	4	4	4
E2QK03	Glycine/betaine ABC transporter (500;54.8)	11	4	4	4	4
E2QPY4	Nucleoside diphosphate kinase (143;15.4)	38	3	3	3	3
E2QK13	Fumarate hydratase (548;60.1)	10	4	4	4	4
E2QIA5	Cytochrome bd-I terminal oxidase subunit I (522;58.2)	8	4	4	4	4
C3TKM2	Peptidyl-prolyl cis-trans isomerase (164;18.2)	27	4	4	4	4
C3TLS7	ATP-dependent Clp protease ATP-binding subunit ClpX (424;46.3)	11	4	4	4	4
E2QIB4	18K peptidoglycan-associated outer membrane lipoprotein (173;18.8)	21	3	4	3	3
E2QNZ9	Periplasmic beta-glucosidase (765;83.5)	7	4	4	4	4
C3SJN7	GTP-binding protein TypA (607;67.3)	9	4	4	4	4
E2QMS8	Translation initiation factor IF-3 (144;16.6)	30	3	3	3	3
C3SR07	30S ribosomal protein S14 (101;11.6)	38	4	4	4	4
E2QKJ3	NADH dehydrogenase (434;47.3)	12	4	4	4	4
C3SG27	Modulator of FtsH protease HflC (334;37.6)	19	4	4	4	4
C3T122	Crr, subunit of enzyme II [glc], trehalose PTS permease, EIIBCMalX and N-acetylmuramic acid PTS permease (169;18.2)	31	4	4	4	4
C3SKH2	A late step of protoheme IX synthesis (398;45.2)	12	4	4	4	4
C3TJ62	PhoH-like protein (359;40.6)	14	4	4	4	4

E2QI31	Sec-independent protein translocase protein TatA (89;9.7)	30	3	4	3	3
C3SFH7	Membrane protein (447;49.7)	9	3	3	3	3
E2QMS9	Threonine--tRNA ligase (642;74.0)	9	4	4	4	4
E2QHJ7	Ribosomal silencing factor RsfS (105;11.6)	33	3	4	3	3
C3TID7	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (250;28.54)	20	4	4	4	4
E2QQE5	Alanine--tRNA ligase (876;95.9)	5	4	4	4	4
E2QNW0	Galactitol permease IIC component (451;48.3)	12	3	3	3	3
Q14F23	Integration host factor subunit alpha (99;11.3)	41	4	4	4	4
E2QJA2	Thioredoxin reductase (321;34.6)	15	3	3	3	3
C3SSG7	50S ribosomal protein L27 (85;9.1)	33	3	4	3	3
E2QJA0	ATP-binding/permease protein cydC (573;62.9)	6	3	3	3	3
E2QFK2	Peptidyl-prolyl cis-trans isomerase (270;28.9)	14	3	3	3	3
C3SR57	30S ribosomal protein S11 (129;13.8)	22	2	3	2	2
E2QG48	DUF1440 domain-containing membrane protein (204;23.0)	13	2	3	2	2
C3SQY2	50S ribosomal protein L29 (63;7.3)	46	2	2	2	2
Q2EVI2	Ribose-phosphate pyrophosphokinase (315;34.2)	16	4	4	4	4
E2QJE1	Asparagine--tRNA ligase (466;52.5)	10	3	3	3	3
E2QNY2	Methionine--tRNA ligase (677;76.2)	6	4	4	4	4
Q2LD69	Inhibitor of g-type lysozyme (133;14.8)	32	3	3	3	3
E2QE62	UPF0441 protein YgiB (223;23.5)	20	3	3	3	3
E2QE63	Uncharacterized protein (386;44.9)	11	3	3	3	3
E2QGH8	Predicted lipoprotein (192;20.9)	18	3	3	3	3
C3TGK7	Leucine-responsive regulatory protein (164;18.9)	19	3	3	3	3
C3SZ57	Signal peptidase I (324;35.9)	11	3	3	3	3
E2QJB6	UPF0142 protein ycaO (586;65.6)	5	3	3	3	3
E2QLT1	UPF0283 membrane protein YcjF (353;39.4)	12	3	3	3	3
C3SQV2	50S ribosomal protein L23 (100;11.2)	19	2	3	2	2
C3T1Y2	Acetyl-coenzyme A carboxylase carboxyl transferase subunit beta (304;33.3)	14	3	3	3	3
E2QNQ4	6-phosphogluconate dehydrogenase, decarboxylating (468;51.45)	9	3	3	3	3
E2QQB8	Peptidoglycan-binding protein (149;16.1)	25	3	3	3	3
E2QJ08	Transcription termination/antitermination protein NusG (181;20.5)	20	2	2	2	2
E2QI28	Ubiquinone/menaquinone biosynthesis C-methyltransferase UbiE (251;28.1)	16	3	3	3	3
E2QIS2	Protein FdhE (309;34.7)	14	3	3	3	3
C3T582	Ferritin (165;19.4)	19	3	3	3	3
E2QER0	Protease degQ (455;47.2)	7	2	2	2	2

E2QKW5	D-amino acid dehydrogenase (432;47.6)	10	3	3	3	3
E2QKH6	Malonyl CoA-acyl carrier protein transacylase (309;32.4)	15	2	2	2	2
C3SR32	50S ribosomal protein L30 (59;6.5)	58	3	3	3	3
C3SV92	S-adenosylmethionine synthase (384;41.9)	12	3	3	3	3
C3SQR7	30S ribosomal protein S12 (124;13.7)	19	2	3	2	2
E2QJI7	Escherichia coli IMT2125 genomic chromosome, IMT2125 (57;6.6)	54	3	3	3	3
C3STK2	Conserved protein (101;11.0)	26	2	2	2	2
E2QEK3	Translation initiation factor IF-2 (890;97.2)	4	3	3	3	3
E2QLV5	Universal stress protein F (144;15.9)	24	3	3	3	3
E2QEY5	LPS-assembly protein LptD (784;89.6)	4	2	2	2	2
E2QLK5	Oligopeptide transport ATP-binding protein oppD (337;37.1)	18	3	3	3	3
C3T2B2	NADH-quinone oxidoreductase subunit A (147;16.5)	18	2	3	2	2
E2QLK2	Peptide ABC transporter substrate-binding protein (543;61.0)	4	2	3	2	2
C3SS42	Aerobic respiration control sensor protein ArcB (778;87.9)	4	3	3	3	3
C3TLR7	DNA-binding protein HU-1 (90;9.22)	34	3	3	3	3
C3T137	Cysteine synthase (323;34.5)	13	3	3	3	3
C3SHL2	LexA repressor (202;22.3)	15	3	3	3	3
E2QKY9	Outer-membrane lipoprotein LolB (207;23.6)	14	2	3	2	2
C3TDX2	3-oxoacyl-[acyl-carrier-protein] synthase 2 (413;43.0)	12	2	2	2	2
C3SQY7	30S ribosomal protein S17 (84;9.7)	31	2	2	2	2
E2QF18	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase (355;37.8)	11	3	3	3	3
C3TM72	Protein-export membrane protein SecF (323;37.8)	10	2	2	2	2
E2QEQ2	N-acetylneuraminate lyase (297;32.6)	12	3	3	3	3
E2QN34	Tail-specific protease (682;76.7)	4	3	3	3	3
E2QHR7	Glutamine--fructose-6-phosphate aminotransferase [isomerizing] (609;66.8)	6	2	2	2	2
E2QPC0	Bifunctional polymyxin resistance protein ArnA (660;74.4)	6	3	3	3	3
E2QIV3	Glycerol kinase (502;56.2)	7	3	3	3	3
C3SSP7	Ribosome-binding factor A (133;15.1)	21	3	3	3	3
E2QIW0	Cell division protein ftsN (319:35.8)	9	3	3	3	3
E2QHT6	D-ribose pyranase (139;15.2)	23	2	2	2	2
C3TJ02	Flavodoxin (176:19.73)	20	2	2	2	2
E2QPX9	Membrane protein (206;22.1)	16	2	2	2	2
C3SZX2	Histidine--tRNA ligase (424;47.0)	7	2	2	2	2
C3SIZ2	Cell division protein ZapB (81;9.6)	44	2	2	2	2

C3TJ07	Ferric uptake regulation protein (148;16.8)	17	2	2	2	2
E2QLN1	Peptidase (349;39.4)	10	2	2	2	2
C3T0M7	Detox protein (111;11.4)	23	2	2	2	2
E2QME7	UPF0482 protein YnfB (113;13.0)	18	2	2	2	2
E2QIY9	N-acetyl-gamma-glutamyl-phosphate reductase (334;35.9)	10	2	2	2	2
C3THM2	Glutamine ABC transporter substrate-bindnig protein (248;27.2)	11	2	2	2	2
E2QIN1	Uncharacterized lipoprotein ybjP (171;18.9)	16	2	2	2	2
C3TPL7	Ribosome-recycling factor (185;20.6)	16	2	2	2	2
Q14F22	Integration host factor subunit beta (94;10.6)	28	2	2	2	2
E2QI74	Negative modulator of initiation of replication (181;20.3)	14	2	2	2	2
E2QPZ5	NifU-like protein (128;13.8)	20	2	2	2	2
C3SEX7	Inner membrane protein yjgP (366;40.3)	9	2	2	2	2
C3T127	Phosphoenolpyruvate-protein phosphotransferase (575;63.5)	4	2	2	2	2
C3TRF7	Peptidyl-prolyl cis-trans isomerase (149;16.1)	19	2	2	2	2
C3TJM2	Lipoyl synthase (321;36.0)	10	2	2	2	2
C3SJ37	Triosephosphate isomerase (255;27.0)	8	2	2	2	2
C3SGH2	Anaerobic C4-dicarboxylate transporter (433;45.7)	3	2	2	2	2
E2QQ41	Uncharacterized tRNA/rRNA methyltransferase yfiF (345;37.8)	13	2	2	2	2
C3SG17	Adenylosuccinate synthetase (432;47.3)	6	2	2	2	2
E2QF54	Glucose dehydrogenase (796;86.7)	3	2	2	2	2
C3TM37	N utilization substance protein B homolog (139;15.7)	15	2	2	2	2
C3SGD7	Elongation factor P (188;20.6)	14	2	2	2	2
E2QG79	Oxygen-dependent choline dehydrogenase (556;61.9)	5	2	2	2	2
C3T3Y7	GTP cyclohydrolase 1 (222;24.8)	11	2	2	2	2
E2QIB0	Colicin uptake protein TolQ (230;25.6)	11	2	2	2	2
Q2LD72	Site-determining protein (270;29.6)	10	2	2	2	2
E2QIS5	Formate dehydrogenase-O, major subunit (1016;112.5)	2	2	2	2	2
E2QEY6	DnaJ-like protein DjlA (271;30.6)	7	2	2	2	2
E2QPR5	RpoE-regulated lipoprotein (191;20.8)	14	2	2	2	2
E2QGB8	Delta-aminolevulinic acid dehydratase (324;35.6)	6	2	2	2	2
E2QIH6	Glutamine transport ATP-binding protein glnQ (240;26.7)	10	2	2	2	2
C3TGN2	Translation initiation factor IF-1 (72;8.2)	39	2	2	2	2
E2QKK9	Spermidine/putrescine import ATP-binding protein PotA (378;43.0)	5	2	2	2	2
E2QQ11	Serine hydroxymethyltransferase (417;45.3)	4	2	2	2	2
C3SZQ2	Iron-binding protein IscA (107;11.5)	21	2	2	2	2

E2QJB8	Phosphoserine aminotransferase (362;39.7)	8	2	2	2	2
C3TG87	UPF0434 protein YcaR (60;6.9)	35	2	2	2	2
E2QKK1	Lipoprotein-releasing system ATP-binding protein LolD (233;25.4)	10	2	2	2	2
E2QGK0	Oligopeptidase A (680;77.1)	4	2	2	2	2
E2QFR3	Fe/S biogenesis protein NfuA (191;20.9)	16	2	2	2	2
C3SKB2	Magnesium and cobalt transport protein CorA (316;36.6)	11	2	2	2	2
E2QMT8	L-cystine transporter tcyP (463;48.7)	6	2	2	2	2
E2QJ40	B12-dependent methionine synthase (1227;136.0)	2	2	2	2	2
C3TGJ7	Outer-membrane lipoprotein carrier protein (204;22.6)	14	2	2	2	2
E2QEV4	Isoleucine--tRNA ligase (938;104.3)	2	2	2	2	2
C3T7R7	Phenylalanine--tRNA ligase alpha subunit (327;36.8)	6	2	2	2	2
E2QQ65	Ribosome maturation factor RimM (182;20.6)	18	2	2	2	2
E2QJD9	Aspartate aminotransferase (396;43.6)	5	2	2	2	2
E2QLK6	Oligopeptide transport ATP-binding protein oppF (334;37.2)	5	2	2	2	2
E2QJ70	Glycerol-3-phosphate acyltransferase (827;93.6)	2	2	2	2	2
C3SRL2	AccC, subunit of biotin carboxylase and acetyl-CoA carboxylase (449;49.3)	5	2	2	2	2
E2QL03	Amino acid permease (470;51.6)	6	2	2	2	2
E2QLP5	DeoR faimly transcriptional regulator (249;27.6)	7	2	2	2	2
E2QJA1	ATP-binding/permease protein cydD (588;65.1)	3	2	2	2	2
E2QEY8	RNA polymerase-associated protein RapA (968;109.7)	2	2	2	2	2
C3SPD2	Cell division protein FtsX (352;38.5)	5	2	2	2	2
E2QPF8	Phosphate acetyltransferase (714;77.2)	3	2	2	2	2
E2QJF5	Paraquat-inducible protein B (546;60.4)	4	2	2	2	2
E2QPI4	3-oxoacyl-ACP synthase (406;42.6)	5	2	2	2	2
E2QN80	Arginine--tRNA ligase (577;64.6)	4	2	2	2	2

Table S4.3 Diameter of particles and average diameter of particle population. Diameters were determined by counting pixels using the paint software. The entire population from both images were averaged together to obtain the average particle diameter and standard deviation. For elongated structures, the width was measured.

Left Image Diameter (nm)	**Right Image Diameter (nm)**
191	218
90	125
91	135
112	173
73	189
63	276
134	208
91	231
58	202
147	87
92	74
133	83
57	173
228	138
50	61
150	93
89	160
234	173
126	215
126	
228	
71	
84	
99	
42	
55	
65	
126	
55	
97	
84	
79	
31	
81	
42	
Average Diameter: 122 ± 61; n=52	

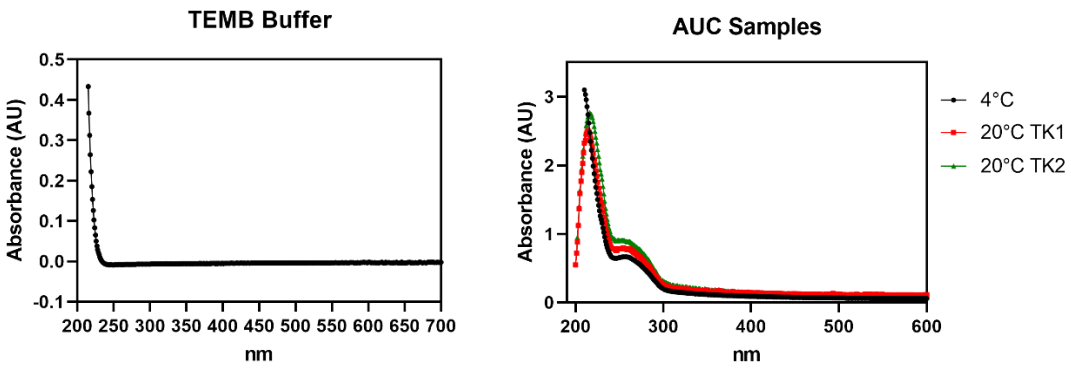


Figure S4.2 Spectral analysis of α-CBs and buffers used in AUC experiments. Left) Absorbance analysis of TEMB buffer. **Right)** Absorbance analysis of the three technical replicates of purified α-CB samples used for AUC experiments. Amount of protein within the sample was determined by SDS-PAGE gel analysis using imageJ and Coomassie stain. Analysis was used to determine sample quality and if doing multi-scan analysis and analysis at 240 nm and 280 nm on the AUC was valid.

Table S4.4 Sedimentation coefficient, molecular mass, and percent abundance of species of proteinaceous particles identified in AUC experiments. Each data from each scan from Figure 4.5 is recorded. Peaks close to average α-CB literature molecular weight (± $5x10^7$ Da) are bolded. Peaks at around the size of the *E. coli* 70S ribosome is italicized.

4˚C			20˚C replicate 1			20˚C replicate 2		
Sedimentation Coefficient (s)	Molecular weight (Da)	Abundance of species (%)	Sedimentation Coefficient (s)	Molecular weight (Da)	Abundance of species (%)	Sedimentation Coefficient (s)	Molecular weight (Da)	Abundance of species (%)
3	2.26E+04	3.74	1.00	6.38E+04	16.01	1	5.21E+04	19.41
3	1.81E+05	8.44	138.89	1.96E+07	5.18	133.33	2.14E+07	5.61
23.172	5.60E+05	16.97	155.56	2.20E+07	0.92	150	2.80E+07	2.14
23.172	3.30E+06	2.50	166.67	4.53E+07	1.58	172.22	4.08E+07	3.21
23.172	3.88E+06	12.66	177.78	4.11E+07	2.25	183.33	5.60E+07	2.30
43.343	*2.53E+06*	0.70	194.44	5.50E+07	2.79	200	6.38E+07	3.80
63.515	2.20E+06	0.41	211.11	4.47E+07	3.42	216.67	5.76E+07	4.31
83.687	3.33E+06	0.60	233.33	4.04E+07	4.85	238.89	6.41E+07	6.16
103.86	4.60E+06	0.96	255.56	3.20E+07	0.57	266.67	3.66E+07	4.85
124.03	6.00E+06	1.55	261.11	3.31E+07	5.33	272.22	6.85E+07	2.31
144.2	7.52E+06	2.37	288.89	3.33E+07	0.20	294.44	1.14E+08	1.42
164.37	9.15E+06	2.84	294.44	3.96E+07	6.27	305.56	2.32E+07	6.85
184.55	1.09E+07	3.17	327.78	5.66E+07	2.34	338.89	1.41E+08	2.26
204.72	1.27E+07	3.53	338.89	3.91E+07	5.16	355.56	3.22E+07	7.57
224.89	1.47E+07	3.47	383.33	1.69E+08	3.28	400	9.08E+07	6.51

245.06	1.67E+07	3.39	400	3.48E+07	6.09	416.67	8.62E+07	6.19
265.23	1.88E+07	3.28	466.67	**2.12E+08**	4.53	483.33	**2.24E+08**	6.22
285.4	2.42E+07	0.63	488.89	1.80E+08	6.75	500	1.13E+08	1.78
285.4	2.59E+07	2.05	600	6.38E+07	10.58	628.57	**1.99E+08**	4.57
305.58	3.06E+07	1.10	828.57	1.04E+08	10.05	1485.7	4.81E+08	2.37
305.58	3.26E+07	1.44	1514.3	3.39E+08	1.86	1628.6	7.05E+08	0.11
325.75	4.26E+07	1.38				1714.3	7.62E+08	0.05
325.75	4.73E+07	0.69						
345.92	4.41E+07	0.39						
345.92	4.66E+07	1.62						
366.09	5.64E+07	1.74						
386.26	6.74E+07	1.62						
406.43	6.26E+07	0.33						
406.43	6.59E+07	1.07						
426.61	6.74E+07	1.29						
446.78	5.76E+07	1.28						
466.95	8.54E+07	0.19						
466.95	9.38E+07	0.76						
487.12	7.37E+07	0.96						
507.29	1.37E+08	0.16						
507.29	1.76E+08	0.58						
527.47	1.45E+08	0.65						
527.47	**2.71E+08**	0.06						
547.64	9.29E+07	0.64						
567.81	6.33E+07	0.75						
587.98	**2.41E+08**	0.35						
587.98	**2.87E+08**	0.08						
608.15	1.15E+08	0.23						
608.15	5.21E+08	0.33						
628.32	5.47E+08	0.36						
648.49	**2.46E+08**	0.40						
668.67	6.01E+08	0.37						
688.84	6.28E+08	0.31						
709.01	5.82E+08	0.29						
729.18	1.20E+08	0.33						
749.35	8.91E+07	0.25						
769.52	1.81E+08	0.10						
769.52	3.39E+08	0.12						
789.7	1.70E+08	0.16						
789.7	1.79E+08	0.12						
830.04	1.29E+08	0.24						

830.04	6.93E+08	0.22						
870.38	1.12E+08	0.19						
890.56	9.24E+08	0.20						
930.9	**3.01E+08**	0.29						
930.9	3.67E+08	0.02						
971.24	1.31E+08	0.09						
991.41	1.46E+08	0.30						
1031.8	1.15E+09	0.15						
1051.9	3.17E+08	0.05						
1072.1	1.52E+08	0.02						
1092.3	1.25E+09	0.28						
1152.8	**2.39E+08**	0.38						
1213.3	1.47E+09	0.12						
1233.5	1.51E+09	0.09						
1294	**2.02E+08**	0.11						
1294	1.62E+09	0.20						
1314.2	1.56E+09	0.06						
1374.7	1.48E+09	0.16						
1394.9	1.81E+09	0.15						
1495.7	3.75E+08	0.28						
1596.6	**2.77E+08**	0.11						
1616.7	3.97E+08	0.12						
1757.9	7.49E+08	0.06						
1757.9	2.46E+09	0.10						
1778.1	5.73E+08	0.00						
1939.5	4.90E+08	0.23						
1959.7	3.01E+09	0.01						

Table S4.5 Summary of molecular weight values of four α-CB sample technical replicates from SEC-MALS. Values obtained using the ASTRA software for the DAWN Optilab MALS equipment (Wyatt Technologies).

Molecular weight (MDa)				
Peak (mL)	Replicate 1	Replicate 3	Replicate 4	Average + standard deviation of standard value
~8.5	VU	300.1 ± 7.3	211.0 ± 36.8	255.6 ± 22.1
~9.5	VU	245.0 ± 26.5	144.0 ± 31.0	194.5 ± 28.8
~21	VU	VU	VU	VU
~27	VU	VU	VU	VU
~35	VU	NA	NA	VU

NA= Not applicable; VU= Value unavailable (unable to determine value from data)

Table S4.6 Summary of hydrodynamic radius values of four α-CB sample technical replicates from SEC-MALS. Values obtained using the ASTRA software for the DAWN Optilab MALS equipment (Wyatt Technologies).

Hydrodynamic Radius (nm) + Standard deviation				
Peak (mL)	Replicate 1	Replicate 3	Replicate 4	Average + standard deviation of standard value
~8.5	NA	126.8 ± 19.1	117.5 ± 25.6	122.2 ± 31.9
~9.5	143.6 ± 11.9	116.5 ± 16.5	111.0 ± 23.7	123.7 ± 17.4
~21	VU	188.3 ± 31.5	122.0 ± 23.6	155.2 ± 27.6
27	Na	VU	VU	VU
35	VU	NA	NA	VU

NA= Not applicable; VU= Value unavailable (unable to determine value from data)

CHAPTER 5: TOWARDS THE BIOPHYSICAL CHARACTERIZATION OF A MINIMAL CARBOXYSOME

5.1 Introduction

α-carboxysomes have valuable properties such as resistance to pH changes and protease degradation, control of metabolite and protein permeability into the shell, large size, and resistant to disassembly due to cycles of freeze thaw (Klein *et al.*, 2009; Tan *et al.*, 2021); these properties can support industrial applications such as use of the α-carboxysome as a biomolecular nanoreactor. However, the complexity of the α-carboxysome, with often more than eight shell proteins (Kinney *et al.*, 2011), means that engineering for specific uses can be challenging.

However, minimal carboxysomes (mCBs) are a viable option, as the number of shell proteins can be reduced to only two (Long *et al.*, 2018). mCBs are carboxysomes that are engineered to be made up of a minimal gene set, usually containing only essential shell proteins and cargo to form fully assembled particles. Therefore, implementation is much easier and without the loss of structural integrity or functionality. mCBs also provide a simpler experimental model (Ochoa *et al.*, 2021). Biophysical characterization of mCBs can be a beneficial including quick methods to screen different engineered structures.

Work with mCBs has been recently described in two publications (Long *et al.*, 2018; Tan *et al.*, 2021). Long *et al.*, 2018 first described a mCB as an approach towards introducing the α-CB into plants to improve carbon fixation efficiency. The authors used a bottom-up approach for building α-CBs in tobacco plants, discovering that with only two shell proteins (Csos1A and only one of Csos2 isoforms), an icosahedral compartment could form within the tobacco leaf tissue and improve RuBisCO activity. Tan *et al.*, 2022 expanded on the creation of mCBs and provided

a library of mCBs which were characterized with respect to their shell structure, demonstrating the resiliency of mCB to drastic changes in physical and chemical environments.

There are indications that the removal of shell proteins, including pentameric proteins, cause elongated structures (Cai *et al.*, 2009; Long *et al.*, 2018). Therefore, being able to analyze structural differences between different CBs is important for our understanding of CB assembly. (Figure 5.1).

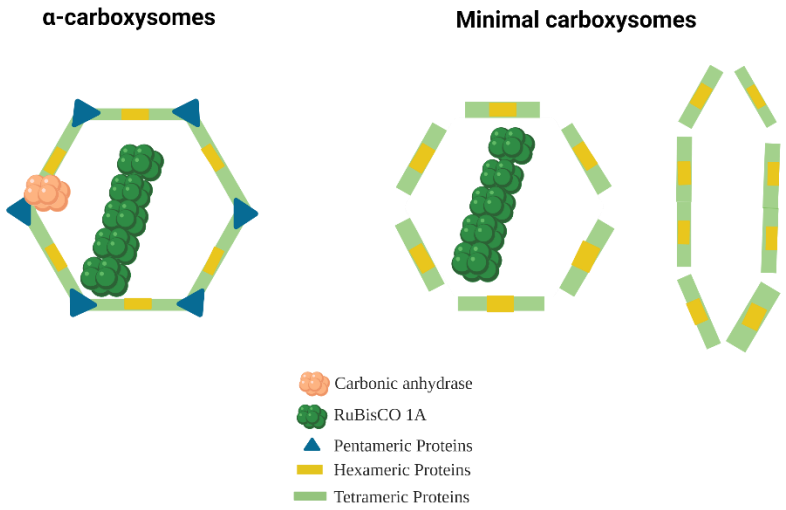


Figure 5.1 Hypothesized structural difference between purified α-CBs and mCBs. Purified α-CBs are typically icosahedral in shape. The mCBs used in this thesis have been previously observed to have both an icosahedral and elongated shape when expressed in Tobacco leaves (Long *et al.*, 2018)**.** Created with Biorender.com.

The objectives of the current mCB study were the same as for the α-CB - to biophysically investigate its structure and obtain insights into its assembly or encapsulation mechanisms. The α-CB in this case can be used as reference to detect any alternative structures, such as the elongated shapes indicated by Long *et al.*, 2018. The construct design was adapted from Tan *et*

al., 2021 because they had used *E. coli*-based promoters and ribosomal binding sites to provide better control of expression in the recombinant host. The reported mCB is specifically composed of the shell protein Csos1A, Csos2 (with the possibility of both isoforms) and the small and large RuBisCO subunits which make up the cargo. The design was adapted in two ways: 1) The specific constitutive promoters, ribosomal binding sites, and terminator regions were chosen based on gene synthesis requirements and for expression in *E. coli*. For each gene, inclusion of an inducible or constitutive promoter was based on the designs by Tan *et al.,* 2019. 2) To improve the modularity of the construct, additional restriction enzyme cut sites were added to the 5' and 3' ends of the coding sequence of each CB gene.

Based on my work on the α-CB, TEM and AUC will be used to characterize the shell shape and heterogenicity of the mCB. As the average particle diameter described for this particular mCB is ~100 nm (Long *et al.*, 2018) and the particle molecular weight is not known, it is likely that the technical limitations of the SEC-MALS analysis discussed in Chapter 4, will also apply to mCBs. Therefore, structural analysis using TEM and AUC will be prioritized.

5.2 Current work and Future Directions

The work with the mCB was limited partially due to delayed (by several months) production of the construct by the gene synthesis provider (Biobasic Inc.). For such a large insert (5,445 bp), the sequence was synthesized in three different segments. The construct (Figures 3.2 showing the vector map and Appendix I Figure A1.2 summarizing the entire sequence) contains multiple promoters, ribosomal binding sites, and double terminators, all of which are prone to forming secondary structures that can cause issues during DNA synthesis, sequencing, and other auxiliary methods such as site directed mutagenesis. Due to these characteristics, there was unexpected difficulties with producing one of the three segments. Furthermore, large sequences

such as this are difficult to ligate into plasmids. As pET28a has 5,371 base pairs and is therefore of a similar size as the sequence to be inserted, issues in proper and efficient ligation can arise. Furthermore, the resulting large construct of 10,816 base pairs is also difficult to work with in terms of sequencing and obtaining sufficient plasmid yields. However, all the cloning steps and design considerations for ensuring gene synthesis is within the scope of commercial gene synthesis companies and feasibility of the project was confirmed with the provider prior to issuing the synthesis order.

Although it would have been easier to synthesize the mCB construct with regulatory elements from the original operon (fewer), the introduction of regulatory elements that have been characterized specifically for use in *E. coli* will provide better control of expression and therefore the assembly of the mCB. Previous work suggests that the operon-based expression in *E. coli* results in differences in transcript copy number compared to the natural expression system (Cyanobacteria), which can affect the carboxysome structure (Bonacci *et al.*, 2012). However, refactoring the entire α-CB operon from *H. neapolitanus* using regulatory elements from *E. coli* is incredibly labour intensive and costly (Temme *et al.*, 2012). Although such a refactored operon would likely provide better expression control for the mCB due to its smaller set of genes, its construction is beyond the scope of this thesis. The mCB analysis will provide additional data that can aid in determining the different types of structures, sizes, masses, and abundance of assembly products. Due to the constraints in commercial gene synthesis beyond my own control, the following section will only describe the future directions of this project.

Once the mCB is obtained, the project will start with the production of the mCBs. Expression and purification of the pETa(+)Csos1A2RuBisCOls construct can be done following the same methods as the α-CBs, highlighted in Chapter 3. As there are only four proteins that

constitute the mCB, and all of them are of different molecular weights (in which Csos1A is ~10 kDa, Csos2 is ~46 or 92 kDa depending on the isoform, The RuBisCO small subunit is ~13 kDa, and RuBisCO large subunit is 63 kDa), identification their expression by size using SDS-PAGE will be easier than for the α-CB (as the α-CB has multiple analogs of the same size). Mass spectrometry will be used to properly assess the proteins (both quality and size), because differences could drastically change the downstream assembly and purification of the respective mCB. mCB purification will still be attempted using a linear sucrose gradient. As described in the previous chapter, optimizing the purification protocol, including the DNase I digestion identified, will be the first step to enable biophysical characterization of the mCB.

The next steps would be the biophysical characterization of the mCB. As with the α-CB, AUC experiments can provide a multitude of data surrounding the mCB heterogenicity. Because the mCB lacks pentameric shell proteins and other paralogs and analogs, the structures likely will differ from the α-CB structure. Therefore, we will compare the mass and the shape (using the frictional coefficient data), to determine if major differences in mCB composition is affecting the assembly and three-dimensional structure of the shell.

Beyond multiwavelength analysis as described in chapter 4, further steps can be taken to study the shape, assembly, and encapsulation mechanisms of the mCB. The construct design of the mCB was done specifically to facilitate experiments. The main addition in the mCB construct are restriction enzyme recognition sites between each coding sequence. Additional, specific cut sites will enable three possible outcomes, 1) the fusion of fluorescent proteins or other markers to the respective coding sequence, 2) the removal of shell protein or cargo coding sequences, and 3) the addition of shell protein genes. These restriction sites allow for further engineering of the mCB facilitating the construction of increasingly more complex 3-dimensional carboxysome

structures or further studies using fluorescent proteins. Fluorescence is of interest as it can be used to determine changes in the loading of RuBisCO into the mCB and that could be correlated to certain carboxysome sizes. For example, a spherical particle may load more or less RuBisCO octamers as compared to an elongated mCB of the same volume. For shell proteins, fluorescence can indicate the changes in stoichiometry (Y. Sun *et al.*, 2019) as particles are assembled. The ability to add additional shell proteins could also advance the development of CB chimeras in which shell proteins from the β-CB are used to replace their homologs in the α-CB (Cai *et al.*, 2015). Applying this method to mCBs can allow for the engineering of new carboxysomes with specific properties.

The final main step of the future directions is to perform Cryo-EM based structure determination of the mCB in tandem with the α-CB. Current reports using Cryo-EM to study CBs are very limited (Evans, 2022), and other imaging techniques that have been used for α-CBs only include electron tomography (Schmid *et al.*, 2006). Cryo-EM can be used to determine the protein-protein interactions within a complex and provide a high-resolution crystal structure of a mCB and α-CB. A high-resolution crystal structure can provide more structural (atomic) information regarding how the shell protein units, such as hexamers and pentamers, interact with each other to form the different possible structures.

CHAPTER 6: CONCLUDING REMARKS

In this thesis, I have described the first preliminary work towards the biophysical characterization of the α-Carboxysome using Analytical Ultracentrifugation, size-exclusion chromatography-Multi-Angle Light Scattering, in conjunction with Transmission Electron Microscopy. Although heterogenicity of α-CBs *in vitro* and *in vivo* has been observed previously (Bonacci *et al.*, 2012; Long *et al.*, 2018; Y. Sun *et al.*, 2019), there has been little effort to fully describe sample populations. As knowledge of the α-CB advances towards engineering for novel applications, such as implementation in plant cells to improve crop yields, it is increasingly important to have detailed biophysical understanding to support the applications.

My work, although preliminary, shows that AUC and TEM are complementary methods that can indicate the molecular weight, shape, and size distribution of the α-CB samples. TEM allows for the determination of particle size and shape through visual analysis and through biophysical modelling, AUC is able to determine particle molecular weight and shape. Both methods can be used for the determination of biophysical attributes that can be compared to each other to provide robust data sets of heterogeneous α-CB samples. Finally, through my AUC experiments I was able to observe that nucleic acids remain in the sample after purification.

It should be noted that my data currently has several caveats. First, the purification of the α-CB, with issues in full cell lysis, contaminants, and separation indicates that further optimization is required. The additional discovery of nucleic acids in the purified α-CB also indicates that additional steps to the original protocol (So *et al.*, 2004) are required. The second caveat is that although the AUC experiments provided valuable data, additional replicate experiments at specific temperatures and different total protein amounts to confirm the effect of concentration and temperature on the samples are required.

Despite the limitations to this study, we have achieved our main goal, to determine which biophysical method are accurate enough to provide data to support the engineering and characterization of highly complex α-CB samples. Biophysical analysis can provide support for future α-CB research which is trending towards use in improving metabolic processes for manufacturing purposes. For example, the use of α-CB in enhancing carbon fixation in plants (by enhancing RuBisCO activities) (Hanson *et al.*, 2016; Long *et al.*, 2018) and the development of carboxysomes or other bacterial microcompartments for use in manufacturing, such as providing enhanced food supplementation for food stock (Kirst *et al.*, 2022).

Appendix I: CARBOXYSOME NUCLEIC ACID AND PROTEIN SEQUENCES

Table A1.1 Carboxysome peptide sequences.

Carboxysome Protein	Amino Acid Sequence
RuBisCO Large Subunit	MAVKKYSAGVKEYRQTYWMPEYTPLDSDILACFKITPQPGVDREEAAAAVAAESSTGTWT TVWTDLLTDMDYYKGRAYRIEDVPGDDAAFYAFIAYPIDLFEEGSVVNVFTSLVGNVFGF KAVRGLRLEDVRFPLAYVKTCGGPPHGIQVERDKMNKYGRPLLGCTIKPKLGLSAKNYGR AVYECLRGGLDFTKDDENINSQPFMRWRDRFLFVQDATETAEAQTGERKGHYLNVTAPTP EEMYKRAEFAKEIGAPIIMHDYITGGFTANTGLAKWCQDNGVLLHIHRAMHAVIDRNPNH GIHFRVLTKILRLSGGDHLHTGTVVGKLEGDRASTLGWIDLLRESFIPEDRSRGIFFDQD WGSMPGVFAVASGGIHVWHMPALVNIFGDDSVLQFGGGTLGHPWGNAAGAAANRVALEAC VEARNQGRDIEKEGKEILTAAAQHSPELKIAMETWKEIKFEFDTVDKLDTQNR
RuBisCO Small Subunit	MAEMQDYKQSLKYETFSYLPPMNAERIRAQIKYAIAQGWSPGIEHVEVKNSMNQYWYMWK LPFFGEQNVDNVLAEIEACRSAYPTHQVKLVAYDNYAQSLGLAFVVYRGN
Csos1A	MADVTGIALGMIETRGLVPAIEAADAMTKAAEVRLVGRQFVGGGYVTVLVRGETGAVN AAVRAGADACERVGDGLVAAHIIARVHSEVENILPKAPQA
Csos1B	MATTHGIALGMIETRGLVPAIEAADAMTKAAEVRLVGRSFVGGGYVTVMVRGETGAVN AAVRAGADACERVGDGLVAAHIIARVHSEVEIILPETPEDSDSAWCIANLNS
Csos1C	MAAVTGIALGMIETRGLVPAIEAADAMTKAAEVRLVGRQFVGGGYVTVLVRGETGAVN AAVRAGADACERVGDGLVAAHIIARVHSEVENILPKAPEA
Csos1D	MNNIDLRVYSFIDSLQPQLASYLATSSQGFLPVPGDACLWIEVAPGMAVHRLSDIALKAT NVRLGEQVVERAFGSMEIHYRNQSDVLASGEAVLREINHAQEDRLPCRIAWKEIIRAITP DHATLINRQLRKGSMLLPGKSMFILETEPAGYIVQAANEAEKAAHVTLIDVRAFGNFGRL TMMGSEAETEEAMRAAEATIASINARARRAEGF
Csos2	MPSQSGMNPADLSGLSGKELARARRAALSKQGKAAVSNKTASVNRSTKQAASSINTNQ VRSSVNEVPTDYQMADQLCSTIDHADFGTESNRVRDLCRQRREALSTIGKKAAKTTGKPS GRVRPQQSVVHNDAMIENAGDTNQSSSTSLNNELSEICSIADDMPERFGSQAKTVRDICR ARRQALSERGTRAVPPKPQSQGGPGRNGYQIDGYLDTALHGRDAAKRHREMLCQYGRG TAPSCKPTGRVKNSVQSGNAAPKKVETGHTLSGGSVTGTQVDRKSHVTGNEPGTCRAVT GTEYVGTEQFTSFCNTSPKPNATKVNVTTTARGRPVSGTEVSRTEKVTGNESGVCRNVTG TEYMSNEAHFSLCGTAAKPSQADKVMFGATARTHQVVSGSDEFRPSSVTGNESGAKRTIT

	GSQYADEGLARLTINGAPAKVARTHTFAGSDVTGTEIGRSTRVTGDESGSCRSISGTEYLSN EQFQSFCDTKPQRSPFKVGQDRTNKGQSVTGNLVDRSELVTGNEPGSCSRVTGSQYGQSK ICGGGVGKVRSMRTLRGTSVSGQQLDHAPKMSGDERGGCMPVTGNEYYGREHFEPFCTS TPEPEAQSTEQSLTCEGQIISGTSVDASDLVTGNEIGEQQLISGDAYVGAQQTGCLPTSPRFN QTGNVQSMGFKNTNQPEQNFAPGEVMPTDFSIQTPARSAQNRITGNDIAPSGRITGPGML ATGLITGTPEFRHAARELVGSPQPMAMAMANRNKAAQAPVVQPEVVATQEKPELVCAPRS DQMDRVSGEGKERCHITGDDWSVNKHITGTAGQWASGRNPSMRGNARVVETSAFANRN VPKPEKPGSKITGSSGNDTQGSLITYSGGARG
Csos3 (Carbonic Anhydrase)	MNTRNTRSKQRAPFGVSSSVKPRLDLIEQAPNPAYDRHPACITLPERTCRHPLTDLEANE QLGRCEDSVKNRFDRVIPFLQVVAGIPLGLDYVTRVQELAQSSLGHTLPEELLKDNWISG HNLKGIFGYATAKALTAATEQFSRKIMSEKDDSASAIGFFLDCGFHAVDISPCADGRLKG LLPYILRLPLTAFTYRKAYAGSMFDIEDDLAQWEKNELRRYREGVPNTADQPTRYLKIAV YHFSTSDPTHSGCAAHGSNDRAALEAALTQLMKFREAVENAHCCGASIDILLIGVDTDTD AIRVHIPDSKGFLNPYRYVDNTVTYAQTLHLAPDEARVIIHEAILNANRSDGWAKGNGVA SEGMRRFIGQLLINNLSQIDYVVNRHGGRYPPNDIGHAERYISVGDGFDEVQIRNLAYYA HLDTVEENAIDVDVGIKIFTKLNLSRGLPIPIAIHYRYDPNVPGSRERTVVKARRIYNAI KERFSSLDEQNLLQFRLSVQAQDIGSPIEEVASA
Csos4A	MKIMQVEKTLVSTNRIADMGHKPLLVVWEKPGAPRQVAVDAIGCIPGDWVLCVGSSAA REAAGSKSYPSDLTIIGIIDQWNGE
Csos4B	MEVMRVRSDLIATRRIPGLKNISLRVMEDATGKVSVACDPIGVPEGCWVFTISGSAARFG VGDFEILTDLTIGGIIDHWVT

CTAGTCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAA
AATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGC
GTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGA
TACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTA
GGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCC
CCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGT
AAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGA
GGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTA
GAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAG
TTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTG
CAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTC
TACGGGGTCTGACGCTCAGTCTAGTTCAGCCAGCTCGTCGTGATGTCGAAACCCAAG
CCACCCTCAGAGGTGAAGGCCGCTTCGAGCACATCGGGAAGCGTGTCGACGACGGT
GCTCGGTGCCTCGGGTAAGAGCCAGATAGATGCGGTGACGTTCACCGTCGCGATCGT
GGCGGGAACCACGGTCACTGCGTCCGTTACAACTCGCACGTCGTCGGCCAGCACGAC
TGACTCGACGGCTTCGATTAGCCCGGGAGAGGCAAGGCCATCAGGCTCGGTAGACA
GAATGCTGATTAACACCTCGCCGGGAGCAGGGCTGCTCACAGCCGCGTCCTTCACCC
GTGGGTCAGCCGTCAGCGCTTGATAGCGGTACCAGGCCGCACCGCCTGCGGTCGAGC
TGCCCTTGATCCGCTCGATGGTGCGATCGCGAAGCTCGCCATCCGGCTCATTCGGCT

GCCGCACCACGCCGTAGAACGCGGCAAGGTTGTCGAGGTCGGGGCCGTTTGAGTAG
CGAAGCAGTGTCGCAAGAAGTGCGTCGTTGATCCGCTGACGCAGGATCAGCTCGCG
GGCTGCGCAGACCTCCAGCAGCTTGATGACCGGGTCGGATTCGAGGATGGCGGTGTA
GCTGGCGTCACGCGATCGCAGGTCGTCGATCAAGTCCTGCAGGATCAGTTCAAAGTC
CAGCGCTTCGATGATGGTGGGCGCGGGAATCGTAGCAAAGTCAAGAACGGTCATGA
GACGACTAAGCCCTCCAGCGTGATACGCCTGCCCTCGGGGATGTAGTAGCCGATCAG
GTTCAGCTCGACTTGACCAGCTGCGCTGGCTGAGACGATGCGAACCTTCTCCAGCTT
CAGCCGTGGCTCCCAGCGATCGAGCGCTTCAGCTGTGGCCGCCACCAGGTCAACGAT
GAGGGACTGGTTGATCGGTCTAGTCATCAAATAAAACGAAAGGCTCAGTCGAAAGA
CTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAAT
CCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGG
ACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGAAGGCCATCCTGACGGATG
GCCTTTTTGCGTTTCTACTCTAGTTTCGAATTGTGAGCGCTCACAATTTCGAAACCCC
AGGCTTGACACTT*tatgcttccggctcgtataatgtgtggaattgtgagcggataacaatttcagctaggtgcgcgaatccccc*
*atcctttcaggaggaactc*atggcagttaaaaagtatagtgctggtgtaaaagaataccggcagacctattggatgccggaatacacaccgt
tggattccgacatccttgcatgcttcaaaatcaccccacaaccgggtgttgatcgcgaagaagccgcagccgcggttgcagcagaatcttca
accggcacatggaccaccgtgtggaccgatttgctgaccgacatggactactacaaaggccgtgcctatcgcattgaagacgtacccggtg
acgatgcggcattctatgcctttatcgcctacccaatcgacctgttcgaagaagggtcagttgttaacgtgtttacctcactggttggtaacgtgt
tcggcttcaaagcggtacgcggcctgcgtctggaagatgttcgcttcccactcgcctacgttaaaacctgtggcggcccaccgcacggtatt
caagtcgaacgcgacaagatgaacaaatatggtcgcccactgttgggttgcaccatcaagccaaaacttggtttgtctgcgaaaaactacgg
ccgtgccgtatacgagtgcctccgtggcggcctcgacttcactaaagatgatgaaaacatcaattctcagccgttcatgcgctggcgcgatcg
cttcttgttcgtacaagacgcgaccgaaactgctgaagcccaaaccggcgaacgcaaaggccattacctcaacgtaacggcgccaactcct
gaagaaatgtacaagcgcgcagaattcgccaaagaaattggcgcgccaatcattatgcacgactacatcaccggtggcttcacggccaaca
ctggcttggccaagtggtgtcaagacaacggcgtactgctgcacatccaccgtgcgatgcatgcggttatcgaccgtaacccgaaccacgg
tattcacttccgtgttctgaccaagattctgcgtttgtcgggtggcgatcacctgcacaccggtaccgttgtcggcaaactggaaggcgaccgt
gcctctactctgggctggattgatttgctccgcgaatcgtttatccctgaagatcgctcgcgcggtatcttcttcgatcaagactggggttcaatg
ccaggcgtattcgctgtggcctctggtggtattcacgtatggcacatgcctgcgctggttaacatctttggtgacgactctgtcctccaattcggt
ggcggtacgctgggtcatccatggggcaacgctgccggtgctgctgccaaccgtgttgctctggaagcctgcgtagaagcgcgtaaccaa
ggccgcgatatcgaaaaagaaggcaaagaaattctgactgctgctgcacagcacagcccagaactgaagattgccatggaaacttggaaa
gagatcaaattcgaatttgacactgtcgacaaactcgacactcaaaatcgttgatccctcgtaccacacaacatactaaggtgagtaaccatgg
ctgaaatgcaggattacaagcaaagcctcaaatatgagactttctcttatcttccacccatgaacgcggaacgcatccgcgctcaaatcaagta
cgcaattgctcaaggctggagccccggcattgagcacgtagaagtgaaaaactccatgaaccaatattggtacatgtggaaacttcccttctt
cggcgaacaaaatgtcgacaacgtgttggctgaaattgaagcgtgtcgtagtgcgtatccaacacaccaggtcaaactggtggcttatgaca
actatgcgcaaagcttaggtctggccttcgtggtctaccgcggcaactaagtcaagtctgtcattgcgctgtgcgtttctctacgcacagcactt
attcaagatggggtcaaacatgccttcacagtcaggaatgaatcctgccgacctgagcggactctctggcaaggaactggcacgcgcacgc
cgcgctgcactatccaagcaagggaaagcagctgtttctaataaaacggctagcgtaaaccgtagcactaaacaggcggcatcttcgatca
atacaaatcaggtgcgctcttctgtaaatgaagtgcccactgattaccaaatggcggatcaattgtgctctacgattgatcatgctgactttggta
ccgaaagcaatcgcgttagagatctctgccgccaacgcagaggaggcactatcaactatcggtaaaaaagcggctaaaaccaccggcaagc
cgtcgggtcgcgttcgaccacagcaatcagtggttcacaacgacgcaatgatcgaaaatgccggtgatactaaccaatcatcgtccacttcat
taaataatgaactttccgaaatctgctccatagcagacgacatgccggagcgttttggttcacaagccaaaaccgtccgtgatatctgccgtgc
acgccgtcaagcgctctctgagcgtggaactcgcgccgtgccgccaaagccgcaatctcaaggtggtccaggacgcaatggctatcaaatt
gatggatacctagataccgcacttcatggccgcgatgccgccaagcgccaccgtgaaatgctctgtcaatacggccgcggcacagcacctt
cctgcaagccaacaggccgtgtcaaaaattctgtacagtcgggcaacgcagcgccaaaaaaggttgaaaccggtcacaccctatccggcg
gatctgttacgggcacgcaagtggatcgtaaatctcatgtgactggaaacgagccgggcacttgccgagcagtcacgggcaccgagtacg
taggtactgagcaattcacctctttttgcaataccagccccaagccaaatgcgacgaaggtcaatgtgaccacaacggctcgtggtcgccctg
ttagcggtacggaagtttcacggaccgagaaggtaactggcaacgaatccggtgtctgccgtaacgttaccggcaccgaatacatgagtaat
gaagctcacttttctctatgtggaacagccgcaaagccttcacaagcggataaagtcatgttcggcgccacagcacgaacgcatcaagtggt

cagtggcagtgatgaattcaggccctcttctgttacgggtaacgaatcgggtgcaaaacgcacaattaccggctcgcagtacgcagacgaa
ggtcttgcgcgactcacgatcaacggagcacctgcaaaagtagccagaacccacaccttttgcgggctctgacgttaccggcacggaaatc
ggtcgctctactcgcgtaactggtgatgaaagcggttcgtgtcgttcaatctcaggcaccgagtatctcagtaacgagcaattccaatccttttg
tgacacaaaacctcaacgcagcccgttcaaggttggccaagatcgcacgaacaagggtcagtctgtgactggtaacttggttgatcgttccg
aactggttacaggtaacgaaccaggttcatgctctcgggttacaggctctcagtatggccaaagcaaaatctgcggtggtggcgtgggaaaa
gtgcgctcaatgcgcacccttcgcggcacctcagtatctggccaacagctagatcatgccccaaagatgtccggtgacgagcgcggcggg
tgcatgcccgtcaccggtaatgagtactacggtcgtgaacatttcgaaccgttttgtacgagcaccccagagcccgaagctcaatcaactgaa
caatcattgacctgtgaaggacaaattattagcggcacttcagttgacgccagtgatttggtcacaggaaatgaaatcggtgaacagcaactc
atcagcggtgacgcctatgttggcgcgcagcagacaggttgccttcccactagtccacgcttcaaccaaactggcaatgttcagtcaatgggt
tttaagaacaccaatcagccagaacaaaactttgcaccaggtgaagtaatgcctactgactttagtattcaaaccccagctcgctcggctcaga
atcgcattacaggtaacgacattgcgccctcaggtcgcattacaggccctggtatgctggcaaccggcttgattacaggaaccccgaattc
aggcacgctgcgcgcgagttggttggttctccacaacccatggcaatggccatggccaaccgtaataaggctgctcaagcacctgttgtgca
gccagaagtggttgcaactcaggaaaagcctgagttggtatgtgcaccaagaagcgatcaaatggatcgtgtgagtggcgaaggcaaaga
acgttgccacatcactggcgatgactggtcagttaacaagcacatcaccggtacagccggtcaatgggcgagtggtcgcaacccttccatg
cgcggtaatgcgcgtgtggtcgaaaccagcgcgtttgccaatcgcaatgtgccaaaacctgaaaagccgggctccaagatcacgggcagt
agtggtaatgacacccaaggtagtctgatcacttactccggcggcgcgcgcggttgattaagtaaagtgtaacgatcatgaacacccgtaac
acacgaagcaagcaacgcgcaccgtgtggtgttagctcatcagtcaaacctcggcttgacttgattgagcaagcaccaaaccctgtctatga
ccgccatcctgcttgtataacgttgcctgagcgtacctgccggcacccgttaactgaccttgaagccaacgaacaactgggtcgttgcgagg
atagcgtcaagaaccggtttgatcgcgttatccctttcttgcaagttgttgctggcattcctcttggtctggattatgttacccgtgttcaagaatta
gcccagtcgtcgcttggacatacgctgcccgaagaattactcaaagataattggatcagtggacacaatttaaaaggcatttttggctacgcga
ctgctaaagcactaacagcagccacggaacaattcagtcgtaaaataatgtctgagaaggacgattccgcatcggccattggcttcttcctgg
attgcgggttccacgcagtggacataagcccttgtgccgatggtcgtctcaaagggttactgccttacatattgcgtttgccccttacggcattc
acctatcgtaaagcctacgcaggctcgatgttcgatattgaagatgatctggcacagtgggagaaaaatgaactccgccgttatcgtgaaggt
gttccaaatacagcggatcagccaacacgatacctgaaaattgctgtgtatcatttcagcacctctgacccgacacactctggctgcgcggca
cacggcagtaatgatcgtgcagcactggaagcggctttaacccagctgatgaaattcagagaagcggttgaaaatgcccattgctgcggcg
caagtatcgatattttactgattggcgttgatacggatacggatgccattcgcgttcatattccggatagcaaaggttttttgaatccgtatcgttat
gttgacaacacagtaacttatgcgcaaacactacatctggcgccggatgaggctcgtgtgattattcacgaagcaattctcaacgcaaaccgc
agcgatggttgggctaaaggaaatggagtagccagcgaggggatgcgtcgttttattggtcagcttttgatcaacaacctctcgcaaatcgatt
acgtagtaaatcgtcatggtggtcgctatccacccaatgatattggtcatgctgagcgatatatcagtgttggtgatggttttgatgaagttcaaat
ccggaatttagcctactacgcgcatttggatacggttgaagaaaatgcgattgatgtggatgtgggaatcacaattttcaccaaacttaatttga
gtcgaggtttaccgattccgattgccatccactatcgctatgaccccaatgttccaggctccagagaaagaaccgtggtcaaagcaagacgg
atatacaacgccattaaagagcggttctcatccttggatgagcagaatctattgcagtttcgtttgagcgttcaggcgcaggatatcggaagcc
cgattgaagaggttgcatccgcatgaaatcatgcaagttgagaaaacgttggtttcaacaaaccgtattgctgatatgggtcacaaaccacta
ttagtggtatgggagaaaccgggcgcgcccaggcaggtcgccgtggatgcgattggctgcataccgggcgattgggtttgtgcgttgggt
catcggcagcacgagaggctgcaggaagcaagtcttatccctctgatttgacgattatcgggattattgatcagtggaatggtgag**taatgg**ga
agtaatgcgcgttcgttccgacctaatcgcaacacgcaggattcccggtcttaaaaatatctctttgcgtgtcatggaggatgctacgggtaag
gtcagtgtcgcttgcgatcccattggcgttcccgagggatgttgggtctttacgattagcggctctgccgctcggtttggcgtgggtgattttga
gattctcacggatttgacgattggtggcatcatcgatcactgggtaacttgaccatcgctagatgagttgattttgaatgagtctttattgaggag
agaagaaatggcagcagtaacaggtattgcactgggtatgattgaaacacgtggtctggttccagcgattgaagctgccgatgccatgacca
aggccgccgaagtacgtttggttggccgtcaatttgttggtggtggttacgtgaccgttttggtccgtggtgaaaccggtgccgtcaacgcag
cagttcgtgcgggcgctgatgcctgcgaacgagtcggcgatggtctggtcgcggcgcatatcattgcccgtgtccattccgaagtcgaaaa
catcctgccgaaagcccctgaagcttaaggattgggaaagacgaaccggcgcaggcttgttccggttctttgcataaagtgacagcttagga
gtttatttaaatggctgatgtaactggtattgctctgggtatgatcgaaacacgtggcttagttcctgcgattgaagcagcggacgccatgacta
aagcggctgaagtgcgtttggtcggtcgtcaatttgttggtggcggttacgtcaccgtattggttcggggcgaaacaggcgctgttaacgccg
ctgttcgtgctggcgccgatgcttgcgaacgtgttggtgatggcttggttgctgcgcacatcattgcgcgtgtccactcagaagtagaaaacat
tctgcctaaggcgccacaagcctaagtcagatattcctaagacggctcacttcggcgacacccgtcgccgaaaccgtctaaccaaatttctg
ggctcgttctgaacctgctcaacactagttttagaggatctgttatggcaacgactcacggtattgccctgggcatgattgaaacacgaggatt

ggttcctgccattgaagccgcagatgccatgaccaaagcggcggaagtccgtctggtcggacgatcatttgttggcggcggttacgtgacc
gtaatggttcgtggtgagacaggtgcagtaaatgctgccgttcgtgcgggtgctgacgcctgtgaacgtgttggcgatggcctggttgctgc
gcacatcattgcgcgcgttcattctgaagttgagatcatcctacccgagacgcccgaagactcagattccgcgtggtgtatcgcaaatctgaat
agctaatgtctagtagggaagatgcgcatgaacaacattgatttgcgcgtctattcgtttatcgattcgcttcaaccgcagcttgcatcctatcttg
cgacatcatcgcaaggctttcttcccgttccgggtgatgcctgcttgtggattgaagtcgcgccgggcatggccgtccatcgcctcagtgatat
tgcgctaaaggccacgaacgttcgtctcggcgaacaggtagtcgagcgtgcttttggctcgatggaaattcattaccgaaaccaaagcgacg
ttctcgcatccggtgaggccgttttaagagaaatcaatcacgcgcaagaagatcgtctgccttgtcgcatcgcatggaaagagatcattcgag
cgattacccccgatcatgccaccttgatcaatcgccagttgcgtaaaggctctatgctgttgccgggcaaaagcatgttcatccttgaaacaga
accggcaggttatattgttcaggctgccaacgaggccgagaaagcagctcatgtgactctgatcgatgtacgtgcttttggtaactttggtcgc
ctgaccatgatgggcagcgaagcggaaaccgaagaagccatgcgggcggctgaggccacaatcgcaagcatcaacgcgcgtgcgcgt
cgcgctgaagggttctaaGCGGCCGCGAGGTTCCAACTTTCACCATAATGAAATAAGATCACTA
CCGGGCGTATTTTTTGAGTTATCGAGATTTTCAGGAGCTAAGGAAGCTAAAATGGAG
AAAAAAATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTAAAGAACAT
TTTGAGGCATTTCAGTCAGTTGCTCAATGTACCTATAACCAGACCGTTCAGCTGGATA
TTACGGCCTTTTTAAAGACCGTAAAGAAAAATAAGCACAAGTTTTATCCGGCCTTTA
TTCACATTCTTGCCCGCCTGATGAATGCTCATCCGGAGTTCCGTATGGCAATGAAAG
ACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAGC
AAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTC
TACACATATATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTA
AAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCACCAG
TTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACTATGGGC
AAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCAT
GCCGTCTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGC
GATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTGAATT
CCTACTAGTCGAAGCGGCATGCATTTACGTTGACACCATCGAATGGTGCAAAACCTT
TCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAATGTGA
AACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTT
CCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAA
GCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGG
CAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTC
GCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGT
GTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTC
TCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCA
TTGCTGTGGAAGCTGCCTGCACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCA
GACACCCATCAACAGTATTATTTTCTCCCATGAAGACGGTACGCGACTGGGCGTGGA
GCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTC
TGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATT
CAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAAC
CATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCA
GATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGA
TATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATATCCCGCCGTT
AACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCT
GCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGT
GAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGG
CCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAT
CCCACAGCCGCCAGTTCCGCTGGCGGCATTTTAACTTTCTTTAATGAATCTAGTGACA
AGCCGGGGCAGACGTGAGCCGTAGTCCCGTCGCCAGACGCGGGTGCCCACGGGCGT

CGTCAGGATGTCCGTAATTGACTGCCGGAGGTGGTCAATGCCCTTCAGCTCCTTGCC
ACTGTCACGGCTCATGCCTCGGGTCATTAGTCGCCCGCTCCGGTATCTTCACTGGCTT
CGATGATTGCCGCCCCGCAGCTGCAGAGGTCACCGATCCGAGCAGTCGGCCTCTGGT
TGGTAAAGACCGTGCGACTGCCGGTGATGATCGTGTTCAAGCCATGCAGGGGGCAG
GCGTGAAGGTCATCCTTTCGGGCCACGGGGCGGCTGTTGACGAAGGTGTCGTCGCTC
CCGGTGATGATGATCCCGCCGTGATCGGTCACGTCGTTTAGTCGAGCGATGCCTGGC
GTCGTAGTCACGGGTTTAGGTCAATACGACTTGCGGTCACTGTAACGTTGCCCTCGG
CGGTCACGTTAACGTCGCCTTGGGCTTCGACTTGCGCCTCCTGCACAAGGATCACAA
TCCGTCCTTGGGCTGCGGTGAGGTCGATCTTGTACTCATGCGCTTCGCGGTCGTACTG
GATGATTGAGTCATCCTCGAACTGCGTCTTTTGGATCGTTTCTTTGTCCTCGATCTGG
GGGTAGTCAGTCGAGAACGCGCCGGGCATCGCGAAGCCCTGACTGATCTCGCCGGA
GGGGGCCATCACGACGACGGCCTCACCGACCTCGGGCGCCCACCAGAACCGATCCTT
GCCCGCTCGCTGCGTGAGCCACGGAATCCAGTCAGTGAGGAGCAGCGCCTCGCCGCT
CTCCTCGTCTTCCTCGATCGCGACACGGATCAGCCCCTTGGGATAGTCAGCCTCGGCT
ACCCTGCCTACGCGGAGCAAGTTGCCGTGACGCCGACTGTCTCGAGTAT

Figure A1.1 pHnCBS1D plasmid DNA sequence (13,218 bps). Lowercase bases are part of the α-carboxysome operon, and the uppercase represent the other regions of the plasmid. The genes for RuBisCO large subunit (dark green), RuBisCO small subunit (light green), Csos1A (pink), Csos1B (purple), Csos1C (grey), Csos1D (red), Csos2 (dark blue), Csos3 (yellow), Csos4A (dark red) and Csos4B (gold) are highlighted. The overlapping stop and start codons for Csos4A and Csos4B respectively are bolded. The T7 promoter and RBS region is italicized. Other regions not highlighted are regions of unknown function within the operon. Sequence was retrieved from addgene.com

TAATCCGGATATAGTTCCTCCTTTCAGCAAAAAACCCCTCAAGACCCGTTTAGAGGC
CCCAAGGGGTTATGCTAGTTATTGCTCAGCGGTGGCAGCAGCCAACTCAGCTTCCTT
TCGGGCTTTGTTAGCAGCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAGtttacggctag
ctcagtcctaggtactatgctagcaaagaggagaaaggcggccgcatgccttcacaaagcgggatgaatccagcggacttgtcaggtttaag
cgggaaagaactggcccgtgcccgccgcgctgcactttcgaaacagggaaaagctgcggtatctaacaaaacagcctcagtcaaccgctc
aaccaagcaggctgcttcctccattaacacaaatcaggttcgctcgtccgtcaacgaggtgccaactgattaccagatggcggatcaactgtg
ctcaacaatcgaccacgcagactttggtacagagtctaaccgtgtccgtgacttgtgtcgtcagcgccgtgaggcattatcaacgattgggaa
gaaggcggcaaaaacgactgggaaaccgagcggacgtgtacgtccccagcaatctgttgtgcacaatgatgctatgattgagaatgcaggt
gataccaaccagtctagtagtacgtcgttaaacaacgaacttagtgaaatctgctccattgcagacgacatgccggagcgcttcggtagccaa
gctaagacagtgcgtgatatctgccgcgcacgccgccaggcgctgagtgaacgcgggactcgtgctgttccacctaaaccccagtcccaa
ggaggcccgggggcgcaatggttaccaaattgacggctacttagataccgctttacacggtcgtgacgccgcaaagcgccatcgcgagatg
ctgtgtcagtacggccgcggcaccgcgccaagctgtaaacctactggacgcgtgaaaaacagcgtccagtcagggaacgcagcaccgaa
aaaagtggaaaccggccatacactgtcgggcgggtctgttacgggcacccaggtggaccgcaagtcgcatgttactggaaatgagcccgg
gacttgtcgcgcggttaccggtactgaatacgtaggcacagaacagtttacctctttttgtaatacttccccgaaacccaacgctactaaagtca
acgtgacgactacggcacgcggagcccagtgtctggcacggaggtttctcgcacggagaaggttaccgggaatgagtccggggtttgcc
gtaacgtaaccggtacagaatacatgagtaatgaggcgcatttctccctttgtggtactgccgctaaacccagtcaagcggataaggtaatgtt
tggggcgacagcacgtacccaccaagtagtaagcggatctgatgagttccgtccgtctagcgtcactggaaatgaatccggtgccaagcgt
accattacaggttcccaatatgccgatgaggggttagcacgcttgactattaacggagccccagccaaggtcgcacgcacgcatactttcgc

aggttccgatgtaaccgggaccgaaattggacgtagcacccgcgttaccggtgatgaatctgggtcttgtcgctctatttcggggacagaata
tctttcgaatgagcagttccaatctttctgcgacaccaagccccagcgttcacctttcaaagtgggccaggaccgtacgaataaagggcaatc
agtcaccggaaatcttgtggaccgtagtgagttagttacaggcaatgagcctggctcgtgttcacgtgttacaggaagccagtacggccaatc
aaagatttgtggtggcggcgtgggaaaggtacgttcaatgcgtactttacgtggcacttcggtatctgggcagcaattggaccatgcacccaa
gatgagtggtgacgaacgcggaggttgtatgccagtcaccggaaatgaatactacgggcgtgagcatttcgagccgttttgtacatcaacgc
cggagccagaagctcaatctacggaacagagcttgacatgtgaaggccaaattattagtgggacctccgttgacgcgtccgacttagtaact
gggaatgaaattggagagcaacaattgattagtggagacgcctatgtcggggcccaacaaaccggatgtttgcccacctcgcctcgtttcaa
tcaaacgggtaacgtccaaagtatgggcttcaagaacaccaaccaaccggaacagaatttcgccccaggtgaagtgatgccaacggatttc
agtattcaaacgcccgcgcgcagcgcacaaaaccgtatcacaggaaacgacattgccccctcgggccgcatcactggtccgggcatgttg
gctacgggcttgattactggaacaccagagtttcgtcacgctgcacgtgagttagtagggtcgccacagccaatggccatggctatggccaa
tcgcaataaagctgctcaagcgccagtggtccaaccagaggtagtggcaacccaggagaagccagaattagtttgcgcgccacgctccga
ccaaatggatcgtgtaagcggagaagggaaagagcgttgtcatattacaggtgacgattggtctgtcaataagcatatcactggaacggcag
ggcagtgggcatccgggcgtaatccctcgatgcgcggtaatgcgcgtgtcgtggaaacttctgctttcgcgaatcgcaacgtgcctaagcct
gagaagccgggttcaaaaattacgggaagctcagggaacgacactcagggttccttaattacttactctgggggggcccgtggataactgc
agccaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctctactagagtcaca
ctggctcaccttcgggtgggcctttctgcgtttatatttacggctagctcagtcctaggtactatgctagctcacacaggaaaccaagcttatgg
ctgtcaagaaatattccgctggggttaaagaatatcgccagacctactggatgcccgaatacacccccccttgacagtgacattttagcatgtttc
aaaatcacgccgcagccaggagtggatcgtgaagaggctgctgcggctgtggctgcggagagttcgacggggacttggacgacggtgtg
gactgacttattgaccgatatggattactacaaagggcgtgcttaccgtatcgaggacgtacctggtgatgacgcggcgttttacgcatttattg
cctatcccattgatttatttgaggaggggagcgtggtaaatgtattcacttccttggtaggaaatgtgtttggcttcaaggccgtacgcggtttgc
gcttggaagacgtccgttttccattggcgtacgtaaaaacctgcggtggtcctccacacgggattcaagttgagcgtgataagatgaataagt
acggacgcccgctgttgggctgtacaattaaaccgaaacttggtctgtccgccaaaaactacggtcgtgcagtctatgagtgtctgcgtggag
gacttgattttacgaaggatgacgagaacatcaattcacaaccatttatgcgttggcgcgaccgtttcttatttgtacaagatgcgacagaaact
gccgaggcccagaccggagagcgcaagggtcattacctgaatgtgaccgctcccacccctgaggaaatgtacaaacgcgcggagttcgc
aaaggagattggggcaccaattattatgcacgactatatcaccggtgggtttactgctaatactggccttgcaaaatggtgtcaagataatggt
gttctgcttcatattcaccgcgctatgcacgcagtaattgaccgcaatcccaaccatgggattcactttcgcgtcctgacgaagatcctgcgttt
atcgggtggtgatcacttgcataccggcactgttgtaggcaaattagagggagatcgtgcgagcacccttgggtggatcgacctgttgcgcg
aatcatttattccagaggatcgctcgcgcggtatcttcttcgatcaggattgggggtcgatgcctggtgtttttgcggtcgcctccggagggatt
cacgtctggcacatgcccgctttggttaacatctttggggatgactctgtccttcaatttggtggcggcacactggggcatccatgggggaac
gcagccggtgccgctgcaaatcgcgttgcactggaggcgtgtgtagaggcgcgcaatcaaggccgtgatatcgaaaaggaaggcaagg
agattttaactgccgcggctcaacacagtccggagttaaaaatcgcgatggaaacctggaaagagattaaatttgagttcgatactgtggata
aacttgacacgcagaatcgctaagaattccgcaaaaaaccccgcttcggcggggttttttcgctttacggctagctcagtcctaggtactatgc
tagcattaaagaggagaaaggtaccatggcggaaatgcaggactacaagcagtcgcttaaatatgagacattctcttacttgccaccaatgaa
tgccgaacgcatccgcgctcaaatcaagtacgccattgctcagggttggtcgcccggtattgaacatgtcgaagtgaagaactctatgaacc
agtattggtatatgtggaagttgccatttttcggagaacaaaatgttgataatgttttggcggagattgaagcatgtcgcagcgcttatcccacac
atcaagtcaaattagtcgcatacgacaattatgcccagagtttagggttggctttcgttgtttaccgtggtaactaaggatccaaaaaaaaaccc
cgcccctgacagggcgggggttttttttaatacgactcactatagggaatacaagctacttgttcttttgcatcacacaggaaagactagtatgg
ccgacgtcaccgggattgcacttggcatgatcgagactcgtggactggtgcctgctattgaagcggcggatgccatgacgaaggctgccga
agttcgtcttgtaggtcgtcaattcgttggcggcggctatgttaccgtgttagttcgtggggaaacaggtgctgtgaatgccgcagtccgtgct
ggcgcagatgcctgcgagcgtgttggtgatggtttagtagccgcgcacatcattgcgcgtgtgcatagtgaagtcgaaaacatcttacctaaa
gcgccacaggcgtaacatatgccggcttatcggtcagtttcacctgatttacgtaaaaacccgcttcggcgggtttttgcttttggaggggcag
aaagatgaatgactgtccacgacgctatacccaaaagaaaTCTAGAGGGGAATTGTTATCCGCTCACAATTC
CCCTATAGTGAGTCGTATTAATTTCGCGGGATCGAGATCTCGATCCTCTACGCCGGA

CGCATCGTGGCCGGCATCACCGGCGCCACAGGTGCGGTTGCTGGCGCCTATATCGCC
GACATCACCGATGGGGAAGATCGGGCTCGCCACTTCGGGCTCATGAGCGCTTGTTTC
GGCGTGGGTATGGTGGCAGGCCCCGTGGCCGGGGGACTGTTGGGCGCCATCTCCTTG
CATGCACCATTCCTTGCGGCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGC
TTCCTAATGCAGGAGTCGCATAAGGGAGAGCGTCGAGATCCCGGACACCATCGAAT
GGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGG
GTGGTGAATGTGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCT
TATCAGACCGTTTCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGG
GAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACA
ACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCT
GCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGC
CAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGG
TGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATG
ACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTTCCGGCGTTATTTCTTGA
TGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAAGACGGTACGCG
ACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGG
CCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACT
CGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGG
TTTTCAACAAACCATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGT
TGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCG
CGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTA
TATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGT
GGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCC
CGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCC
CCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAG
CGGGCAGTGAGCGCAACGCAATTAATGTAAGTTAGCTCACTCATTAGGCACCGGGAT
CTCGACCGATGCCCTTGAGAGCCTTCAACCCAGTCAGCTCCTTCCGGTGGGCGCGGG
GCATGACTATCGTCGCCGCACTTATGACTGTCTTCTTTATCATGCAACTCGTAGGACA
GGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAGCGCGAC
GATGATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCCTTC
GTCACTGGTCCCGCCACCAAACGTTTCGGCGAGAAGCAGGCCATTATCGCCGGCATG
GCGGCCCCACGGGTGCGCATGATCGTGCTCCTGTCGTTGAGGACCCGGCTAGGCTGG
CGGGGTTGCCTTACTGGTTAGCAGAATGAATCACCGATACGCGAGCGAACGTGAAG
CGACTGCTGCTGCAAAACGTCTGCGACCTGAGCAACAACATGAATGGTCTTCGGTTT
CCGTGTTTCGTAAAGTCTGGAAACGCGGAAGTCAGCGCCCTGCACCATTATGTTCCG
GATCTGCATCGCAGGATGCTGCTGGCTACCCTGTGGAACACCTACATCTGTATTAAC
GAAGCGCTGGCATTGACCCTGAGTGATTTTTCTCTGGTCCCGCCGCATCCATACCGCC
AGTTGTTTACCCTCACAACGTTCCAGTAACCGGGCATGTTCATCATCAGTAACCCGTA
TCGTGAGCATCCTCTCTCGTTTCATCGGTATCATTACCCCCATGAACAGAAATCCCCC
TTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCT
TTATCAGAAGCCAGACATTAACGCTTCTGGAGAAACTCAACGAGCTGGACGCGGAT
GAACAGGCAGACATCTGTGAATCGCTTCACGACCACGCTGATGAGCTTTACCGCAGC
TGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAG

ACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGC
GTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTAGCGATA
GCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCA
CCATATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAG
GCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGA
GCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAA
CGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAG
GCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAAT
CGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTT
TCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATAC
CTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGT
ATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCG
TTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAA
GACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGG
TATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGA
AGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTT
GGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCA
AGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTA
CGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAACA
ATAAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAA
CGGGAAACGTCTTGCTCTAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATAT
GGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAATCTATCGATTG
TATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCC
AATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTT
CCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGA
TCCCCGGGAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATA
TTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTG
TCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAA
CGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACA
AGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCAGTCGTCACTCA
TGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATT
GATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAAC
TGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTG
ATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAAGA
ATTAATTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGG
TTCCGCGCACATTTCCCCGAAAAGTGCCACCTGAAATTGTAAACGTTAATATTTTGTT
AAAATTCGCGTTAAATTTTTGTTAAATCAGCTCATTTTTTAACCAATAGGCCGAAATC
GGCAAAATCCCTTATAAATCAAAAGAATAGACCGAGATAGGGTTGAGTGTTGTTCCA
GTTTGGAACAAGAGTCCACTATTAAAGAACGTGGACTCCAACGTCAAAGGGCGAAA
AACCGTCTATCAGGGCGATGGCCCACTACGTGAACCATCACCCTAATCAAGTTTTTT
GGGGTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTA
GAGCTTGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAAAGCGAA

AGGAGCGGGCGCTAGGGCGCTGGCAAGTGTAGCGGTCACGCTGCGCGTAACCACCA
CACCCGCCGCGCTTAATGCGCCGCTACAGGGCGCGTCCCATTCGCCA

Figure A1.2 pET28a(+)Csos1A2RuBisCOls plasmid DNA sequence (10,431 bps). Lowercase bases are the α-carboxysome genes with regulatory regions and the uppercase bases are the other regions of the plasmid. Constitutive promoters are shown in pink, inducible promoter T7 in grey, terminators in blue, RBS in red, and added restriction sites are underlined. Genes for Csos2 (teal), RuBisCO large subunit (green), RuBisCO small subunit (purple) and Csos1A (yellow) are also highlighted.

Appendix II: TOWARDS THE BIOPHYSICAL CHARACTERIZATION OF LUMAZINE SYNTHASE VARIANT AaLS-13

A2.1 Introduction

Lumazine synthase from thermophilic archaeal species *Aquifex aeolicus* (AaLS) is a highly thermostable Protein Nanocompartment (PNC) with a melting temperature of 120˚C (Ladenstein *et al.*, 2013). AaLS naturally encapsulates and stores the enzyme riboflavin synthetase to facilitate the production of riboflavin (X. Zhang *et al.*, 2006) as depicted in Figure A2.1. In contrast to the carboxysome and any other PNCs or BMCs, the AaLS shell takes an active part within the metabolic pathway to produce riboflavin. AaLS phosphorylates 3,4-dihydroxy-2-butanone-4-phosphate yielding 6,7-dimethyl-8-ribityl-lumazine which is then capable of binding to riboflavin synthetase to produce riboflavin.

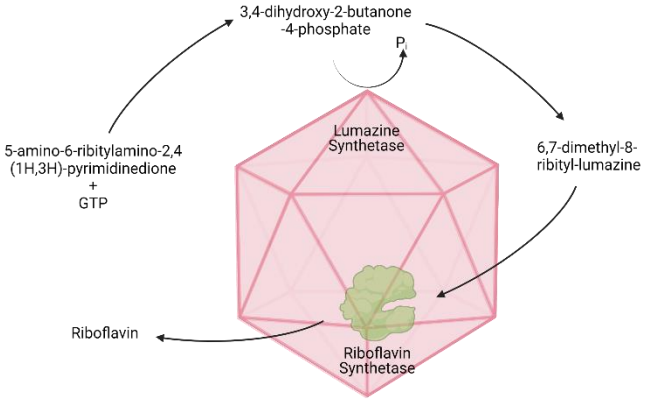


Figure A2.1 Riboflavin biosynthesis in *Aquifex aeolicus* involving the Lumazine Synthase shell and its cargo riboflavin synthetase. The AaLS shell is shown in pink and riboflavin synthetase in green. GTP stands for Guanosine-5'-triphosphate and Pi is an inorganic phosphate.

AaLS has an icosahedral shell structure naturally formed by 60 monomeric units with the binding site for riboflavin synthetase (for the function of encapsulating the protein) within the interior of the shell (Azuma, Edwardson, *et al.*, 2018). Wild type AaLS has an interior volume of 270 nm^3 (Azuma *et al.*, 2017; Ladenstein *et al.*, 2013). The shell itself is stable at neutral pH but results in larger shell structures at different pH, slightly increasing this initial volume (X. Zhang *et al.*, 2006). AaLS shells also have the benefit of being able to be resistant to proteases, extreme pH, temperature, etc. Like many other PNCs, AaLS can encapsulate protein cargo and prevent other biomolecules from entering the inside of the icosahedral shell due to its semipermeable pores (Azuma, Bader, *et al.*, 2018). However, the exact mechanism of metabolic transport through these pores remains unknown.

Several studies have modified the lumen of the shell to facilitate the encapsulation of positively charged cargo or assembly without cargo using high ionic buffers (Figure A2.2A). A negatively charged variant, AaLS-neg, contains sequence modifications in which Arg83, Thr86, Thr120 and Gln123 were changed to glutamic acid residues that cause the interior of the PNC to have an overall negative charge (Ladenstein *et al.*, 2013). The variant AaLS-13, an evolutionarily optimized variant of AaLS-neg, is able to expand into a 360-subunit icosahedral structure (increasing the internal volume to 15 600 nm^3 compared to the original 270 nm^3) causing larger gaps within the shell wall (Azuma, Bader, *et al.*, 2018; Ladenstein *et al.*, 2013).

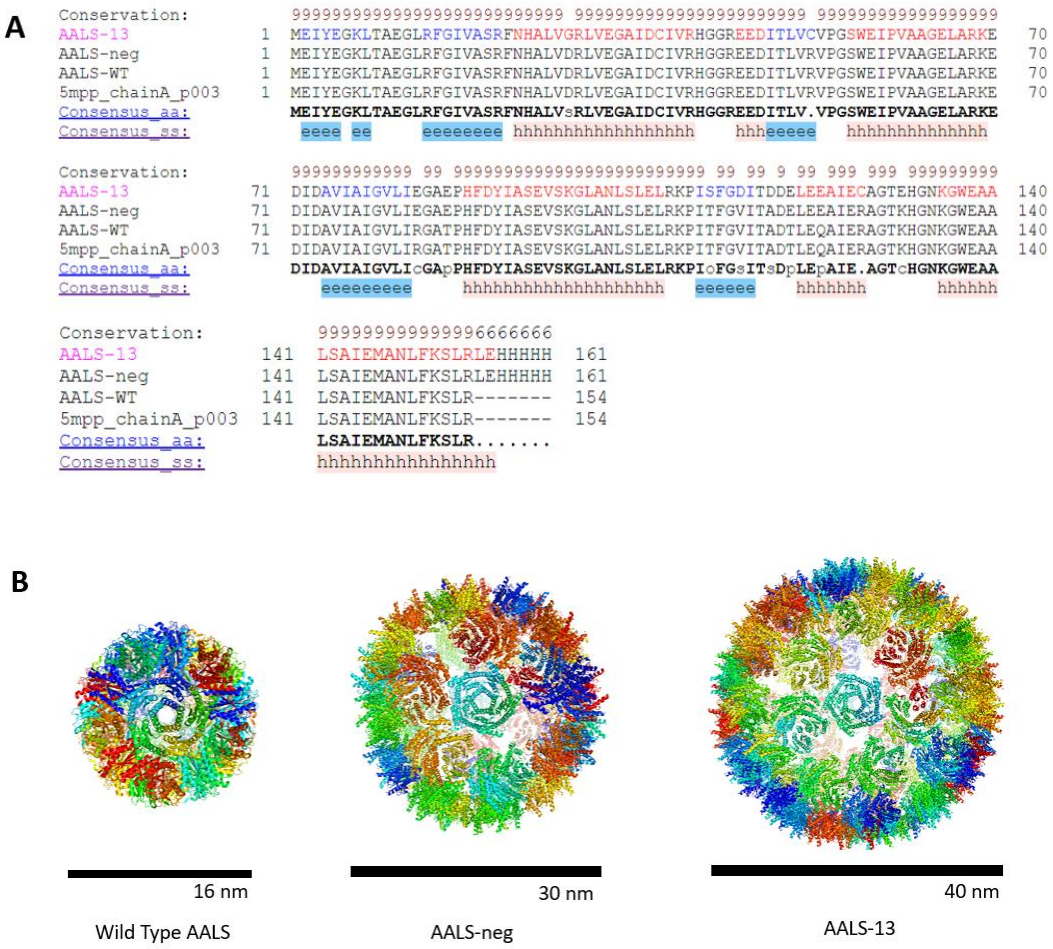


Figure A2.2 Sequence and structure differences between AaLS-WT, AaLS-neg, and AaLS-13. A) Sequence and secondary structure alignment of three AaLS proteins: Wild type (pdb:5MPP), AaLS-neg (pdb:5MQ3), and AaLS-13 (pdb:5MQ7) using the PROMALS3D online tool. The conservation score is given at the top of the alignment where mutations are represented by gaps. The 5mpp_chainA_p003 is the Wild type variant automatically found by the tool. The consensus secondary structure is shown at the bottom of the alignment where h represents alpha helices and e represents β-sheets. **B)** The structures and sizes of AaLS-WT, AaLS-neg, and AaLS-13 as determined by electron microscopy.

These gaps have been determined to be ~4 nm in diameter as compared to the ~1 nm diameter of the Wild type (Sasaki *et al.*, 2017). The crystal structure of AaLS-WT, AaLS-13, and AaLS-neg variants are compared in Figure A2.2 panel B. Although these larger gaps allow for uncontrolled entry of small molecules into the shell, the high affinity binding of the positively charged substrate and the highly negatively charged interior prevents the entry of negatively

charged molecules and the escape of the cargo (Azuma, Edwardson, *et al.*, 2018). These gaps also allow influx of larger, positively charged metabolites that can interact with the cargo without the need for cargo release. In the Wild type AaLS the influx occurs through the shell pores and is therefore more restrictive.

A2.1.2 Assembly and Encapsulation Mechanisms

The assembly of AaLS is facilitated by interactions with the protein cargo riboflavin synthetase (Azuma *et al.*, 2017). The C-terminal domain of riboflavin synthetase can be used as an encapsulation peptide for other proteins if required. For example, the C-terminal domain has been used to facilitate encapsulation of GFP into the AaLS shell. The AaLS-13 and AaLS-neg variants can assemble spontaneously with the presence of positively charged cargo or in high salt conditions (Azuma & Hilvert, 2018). Highly positively charged proteins such as supercharged green fluorescent protein (with a net charge of +36) can induce the assembly of the shell (Azuma & Hilvert, 2018; Azuma *et al.*, 2016).

Despite the variability in AaLS assembly, Wild type or variants, the disassembly of already formed shells has only been achieved under denaturing conditions such as treatment with 3 M guanidinium Chloride (GdnHCl). This chaotropic induced disassembly has been found to be reversible with insignificant protein losses, once native conditions are restored (Azuma *et al.*, 2017).

A2.1.3 Advancements in Lumazine Synthase Applications

Beyond its natural function as a part of riboflavin synthesis, AaLS has been engineered to create higher order structures that contain shell surface modifications and/or modifications for drug delivery applications. The AaLS shell can be easily chemically modified or engineered for

cell targeting, peptide display, and imaging (Min *et al.*, 2014; Song *et al.*, 2015). Additionally, AaLS has been used in vaccine development where antigens are displayed on the shell to display and allow for recognition by T-cells (Azuma, Edwardson, *et al.*, 2018; Ra *et al.*, 2014). Despite the multi-faceted research on AaLS engineering, only two approaches have been pursued to encapsulate different cargos with AaLS so far: The encapsulation of specific mRNA (Terasaka *et al.*, 2018) and encapsulation of selenocysteine (Wang *et al.*, 2018).

A2.1.4 Objectives

The original objective was identical to the α-CB: biophysically characterize AaLS-13 using the methods SEC-MALS, AUC, and TEM to advance the use of AaLS-13 as a drug delivery system. The progress in this regard was limited due to improper assembly of the AaLS-13 shell. The following sections will describe what has been done to improve assembly and give future directions to attain the proper assembly of AaLS-13. Although the focus in this chapter is AaLS-13, some sections will include AaLS-WT which was intended to be used as a control.

A2.2 Experimental Methods

The following section describes the methods used in the work described in Appendix II. All reagents were obtained from New England Biolabs, Fischer Scientific, Sigma Aldrich, Promega, or Biobasic unless specifically stated.

A2.2.1 Construct Design

The DNA sequences were designed by first optimizing the coding sequences for AaLS-WT and AaLS-13 for translation in *E. coli* and adding a hexa-histidine tag at the C-terminus. EcoRI or HindIII restriction sites within the Coding sequence (CDS) were manually removed and the

restriction sites were then added to the respective 5’ and 3’ ends for cloning into the final expression plasmid.

The DNA sequences were then synthesized and cloned into pUC57-Kan by Biobasic Inc. The constructs were transformed into DH5α *E. coli* cells and cell cultures were produced for plasmid extraction using a commercial kit (Biobasic) The pUC57-Kan constructs containing the AaLS genes were restricted using EcoRI (5’) and HindIII (3’) to excise the respective gene sequences and then purified by gel extraction (BioBasic). The obtained fragments were then ligated into the expression vector pET28a(+)-Kan via overnight T4 ligation at room temperature. Integration of the inserts were verified by sanger sequencing and double restriction tests using EcoRI and HindIII.

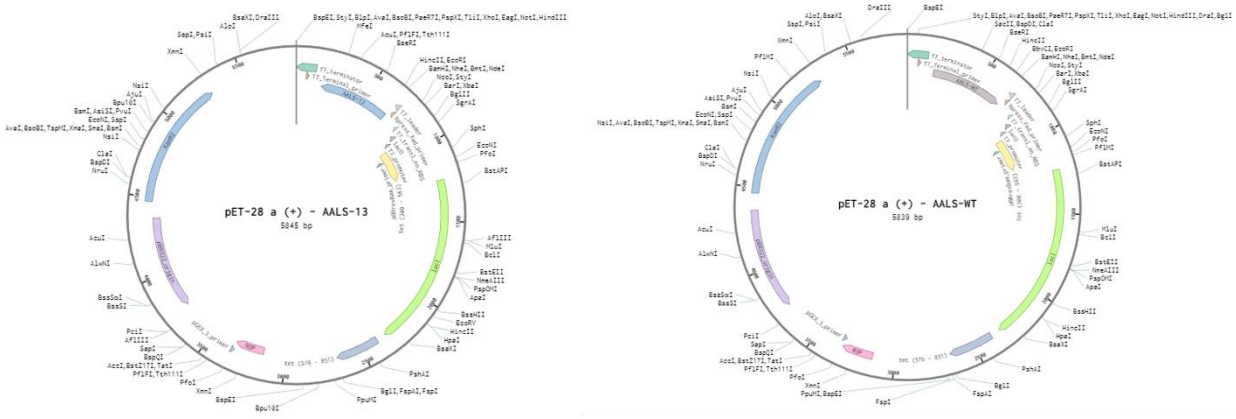


Figure A2.3 Plasmid maps of AaLS-WT and AaLS-13 pET28a (+) expression plasmids. The promoters, ribosomal binding sites, coding sequences, and terminators are annotated on the plasmid map. The plasmid map was created in Benchling.

A2.2.2 Overexpression of AaLS-WT and AaLS-13

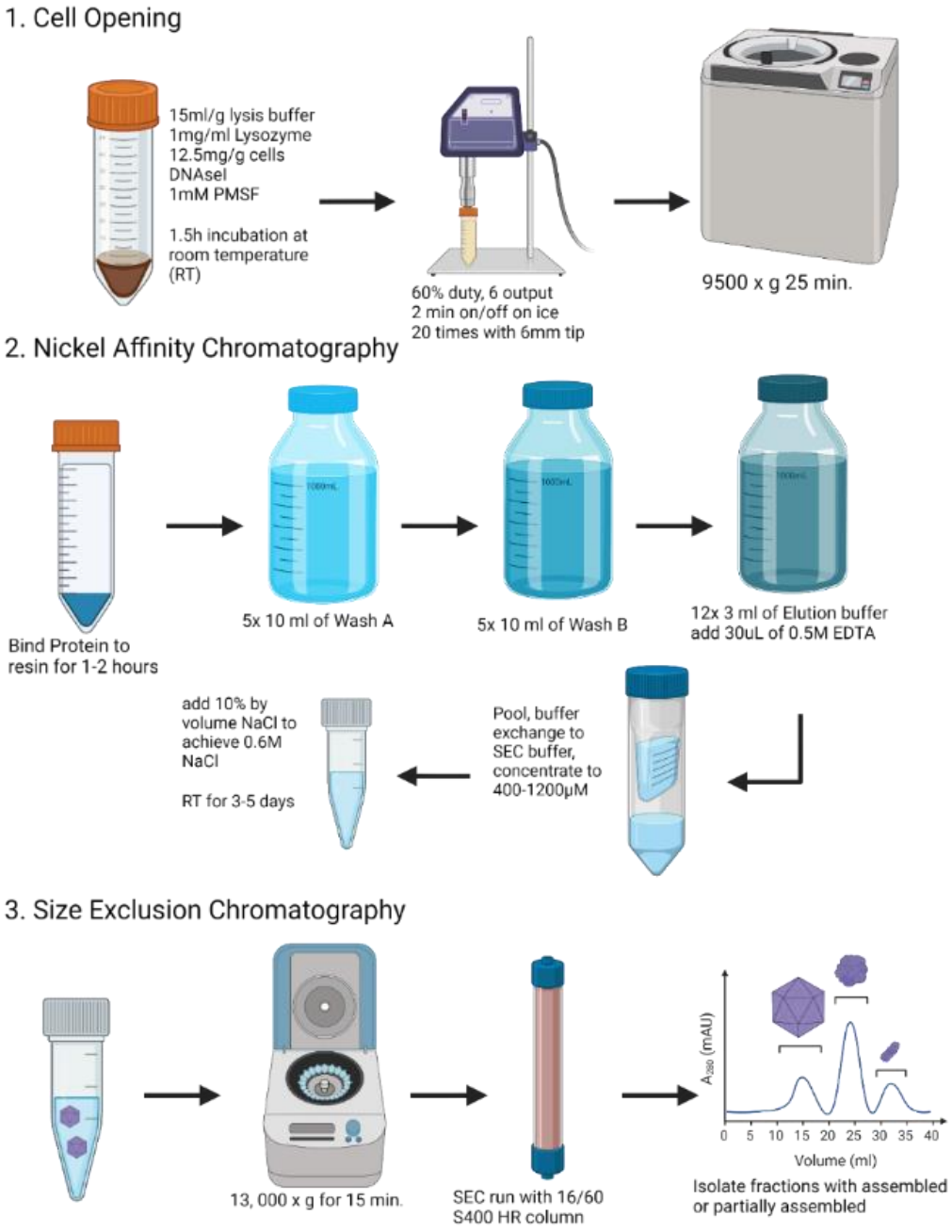
AaLS-WT and AaLS-13 pET28a(+) expression constructs were transformed into BL21 (DE3) cells for protein overexpression. 1 µL of 100 µM DNA was pipetted into 10-20 µL of

BL21 (DE3) cells and left on ice for 30 minutes. The cells were subsequently heat shocked at ~42˚C for 20 seconds before being placed on ice for 2 minutes. 950 µL of LB medium was added to the cells and shaken at ~220 rpm at 37˚C for 1 hour. 200-400 µL of the transformation mixture was spread onto a LB plate supplemented with 0.05 mg/mL final concentration of kanamycin. For overexpression, cells were spread on a LB plate supplemented with 0.05 mg/mL final concentration of kanamycin and incubated at 37˚C overnight. A single colony was picked and used to inoculate 50 mL LB media supplemented with 50 mg/mL of kanamycin final concentration followed by incubation over night at 37˚C with shaking at 180 rpm in a New Brunswick Innova incubator (Eppendorf). The overnight culture was used to inoculate a 500 mL culture at a final OD_{600} of 0.1. Cultures were grown at 37˚C to an OD_{600} of 0.4-.07 at which the samples were induced with 0.1 µM IPTG and left to grow overnight at 180 rpm at 25˚C for 18 hours or 20˚C at 20 hours. Cells were harvested by centrifugation at 5,000 x g for 15-20 minutes (JLA 8.1000 rotor with Avanti JXN-26 centrifuge) at 4˚C. The resulting cell pellets were stored at -20˚C until further use.

A2.2.3 Nickel Affinity Purification of AaLS-WT and AaLS-13

The following purification steps were all performed at room temperature unless noted otherwise (Figure A2.4). Cells were resuspended in 15 mL/g of lysis buffer (100 mM phosphate buffer pH 8.0, 300 mM NaCl, 10 mM imidazole, 1 mM EDTA, 0.05% Tween-20). 1 mg/mL of lysozyme, DNase, 10 mM phenylmethylsulfonyl fluoride (PMSF), and 12.5 mg/g cells of Sodium deoxycholate were added and the culture was incubated for 1.5 hours. The cell suspension was then sonicated 20 times, on ice with settings at 50% duty, 6 output, with 2 minutes intervals using a 1/5” sonication tip and a Branson Sonifier 450 (Branson Ultrasonic corp.). The cell debris in the lysate was removed by centrifugation at 9,500 x g for 30 minutes using a JA25.50 Beckman

Coulter rotor (Avanti JXN-26). The resulting supernatant was bound to 5 mL nickel sepharose resin slurry (GE Healthcare) for 1 hour. The resin was washed with 50 mL wash buffer A (50 mM phosphate buffer pH 8.0, 800 mM NaCl, 20 mM imidazole) followed by 50 mL wash buffer B (50 mM phosphate buffer pH 8.0, 800 mM NaCl, 80 mM imidazole). Proteins were eluted from the column using 36 mL of elution buffer (50 mM phosphate buffer pH 8.0, 800 mM NaCl, 500 mM imidazole), and 5 mM EDTA was added immediately after each elution was collected. After purification, the elution samples were buffer exchanged into the SEC buffer (50 mM phosphate buffer pH 8.0, 200 mM NaCl, 5 mM EDTA) and concentrated using an ultracentrifugation membrane with a 3 kDa molecular weight cut off (Cytiva).



Lysis buffer= 100mM phosphate buffer pH 8.0, 300mM NaCl, 10mM imidazole, 1mM EDTA, 0.05%, Tween-20
Wash A= 50mM phosphate buffer pH 8.0, 800mM NaCl, 20mM imidazole
Wash B= 50mM phosphate buffer pH 8.0, 800mM NaCl, 80mM imidazole
Elution buffer= 50mM phosphate buffer pH 8.0, 800mM NaCl, 500mM imidazole
SEC Buffer= 50mM phosphate buffer pH 8.0, 200mM NaCl, 5mM EDTA

Figure A2.4 A schematic of purifying and assembling AaLS proteins. Created with Biorender.com.

A2.2.4 The Assembly of AaLS-WT and AaL-13

AaLS-13 nickel affinity purified samples in SEC buffer were initiated for assembly by the addition of 10% v/v of 5 M NaCl (0.6 M NaCl final concentration) and left for 3-5 days at room temperature. As a control for non-assembly, samples were left at 0.2 M. Aggregates formed during assembly were removed by centrifugation at 13, 000 x g for 15 minutes (accuSpin Micro 17, Fisher Scientific) and the resulting assembled AaLS were subsequently separated on a S400 HR 16/60 size exclusion column at 1 mL/min in SEC buffer using a ÄKTA Prime (Cytiva) purification system at room temperature.

A2.2.5 Western Blot

Protein samples were analyzed on a 12% Polyacrylamide (PA) gel for ~45 minutes at 180 V. Using a mini-transblot cell from Biorad, samples on the PA gel were transferred (100 mAmp) onto a 45 μm nitrocellulose membrane (Cytiva) for 1 hour at 4˚C in transfer buffer (48 mM Tris, 39 mM Glycine, 20% Methanol, pH 9.2). The PA gel was stained thereafter with Coomassie G-250 and the nitrocellulose membrane was stained with 0.5% Ponceau S to confirm transfer. The membrane was de-stained and incubated in blocking buffer (20 mM Tris, 150 mM NaCl, pH 7.5, 3% skim milk) for 1 hour and washed with TBS buffer (20 mM Tris, 150 mM NaCl pH 7.5) twice for 5 minutes. Samples were incubated in 1:5,000 anti-His antibody (Monoclonal Anti-6X His-tag antibody produced in mouse, Sigma Aldrich) and diluted in TBS with 5% skim milk overnight. The membrane was then washed with TBS+TT (TBS with 0.05% Triton X 100 and 0.2% Tween 20) for 1 minute, twice with TBS for 10 minutes, and subsequently incubated with a 1:10,000 dilution of secondary antibody (anti-mouse igG antibody, HRP conjugate, Sigma Aldrich) for 1 hour, washed once with TBS+TT for 10 minutes and the proteins were detected with luminol using an Amersham 1600 imager (Cytiva) with a 10-20 minute autoexposure time.

A2.2.6 Purification of Supercharged GFP

Purification was performed at room temperature unless noted otherwise. Cells were resuspended in 15 mL/g of lysis buffer (20 mM phosphate buffer pH 7.4, 2 M NaCl). ~1 mg/mL of lysozyme, a few crystals of DNase I, 10 mM PMSF, and 12.5 mg/g cells of deoxycholate were added to the suspended cells and was incubated for 1.5 hours. The cell lysate was then sonicated on ice at 50% duty, 6 output, 2 minutes on/off for 20 times using a 1/5" sonication tip and a Branson Sonifier 450 (Branson Ultrasonic corp.). Cell debris were removed from the lysate by centrifugation at 9,500 x g for 30 minutes (JA25.50 Beckman coulter rotor and Avanti KXN-26 centrifuge) and the resulting supernatant was bound to the nickel Sepharose resin (3 mL of slurry) for 1 hour. The resin was washed with 100 mL lysis buffer and 40 mL wash buffer B (20 mM phosphate buffer pH 7.4, 2 M NaCl, 20 mM imidazole). Samples were eluted with 36 mL elution buffer (20 mM phosphate buffer pH 7.4, 2 M NaCl, 250 mM imidazole) and 5 mM final concentration of EDTA was added immediately after elution. After purification, the elution samples were buffer exchanged into the lysis buffer and concentrated using an ultracentrifugation membrane with a 3 kDa molecular weight cut off (Cytiva). The GFP (+36) samples (2 samples at 500 μL each) were injected onto a Superdex 75 10/300 column (Cytiva) and eluted at 4˚C using an ÄKTA Pure system (Cytiva) at a flow rate of 0.4 mL/min. Fractions containing GFP(+36) were pooled, buffer exchanged into a storage buffer (20 mM phosphate buffer pH 7.5, 1 M NaCl, 50% glycerol), and concentrated using an ultrafiltration membrane with a 3 kDa molecular weight cut off (Cytiva). Samples were then flash frozen and stored at -80˚C until used for further experiments.

A2.2.7 Determination of peptide secondary structure using I-Tasser

The AaLS-13 peptide sequence with both the C-terminal His-tag and N-terminal tag was used as an input on the I-Tasser online server (Zheng *et al.*, 2021). To constrain the structure predictions, the AaLS-WT crystal structure from the pdb database (5MPP) was used to align the peptide sequence. The resulting structures were used for analysis. The superimposing of the predicted AaLS-13 structure with the two tags to the hexameric AaLS-WT structure was performed using Pymol.

A2.3 Results

A2.3.1 AaLS proteins can be Successfully Overexpressed in *E. coli*

Our initial work was to confirm successful overexpression of AaLS-WT and AaLS-13 using SDS-PAGE. Western blots were used to confirm the identity of the respective AaLS in the SDS-PAGE via detection of the His-tag fused. Proteins were overexpressed in BL21 (DE3) cells and induced by IPTG. In ~30% of the conducted overexpressions of AaLS-13, bands at ~24 kDa and ~21 kDa were observed. Only a single band at ~24 kDa band was seen for AaLS-WT (Figure A2.5).

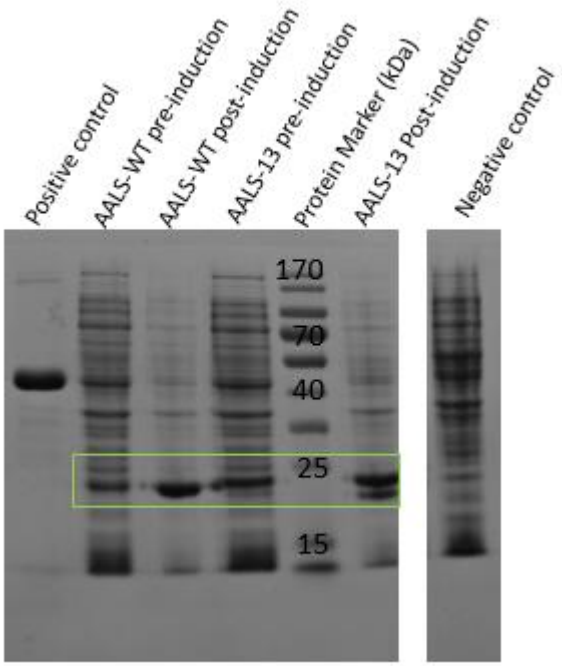


Figure A2.5 Overexpression of AaLS-WT and AaLS-13. 1.0 OD_{600} of cells were resuspended in 8 M of Urea. 5 µL of the lysate was analyzed on a 12% SDS-PAGE. The PAGE was developed at 180 V for 45 minutes. The gel was stained with Coomassie G-250. Purified EF-Tu protein was used as a positive control and BL21 (DE3) cell lysate was used as a negative control. Bands corresponding to the overexpressed AaLS proteins are highlighted. The image is cropped to remove unrelated samples from the gel image.

Overexpression is observed at the final time point (post-induction) for both AaLS-WT and AaLS-13 when compared to the pre-induction and post-induction samples. There is also an indication of leaky expression, indicated by the signal in the negative control (Briand *et al.*, 2016). At this point, there are no other bands that suggest oligomerization when the proteins are overexpressed. In the overexpression shown in Figure A2.5 two bands are detected for AaLS-13. A western blot confirmed that both the 24 and the 21 kDa band have a His-tag and therefore are AaLS-13 proteins with the C-terminal His-tag (Figure A2.6).

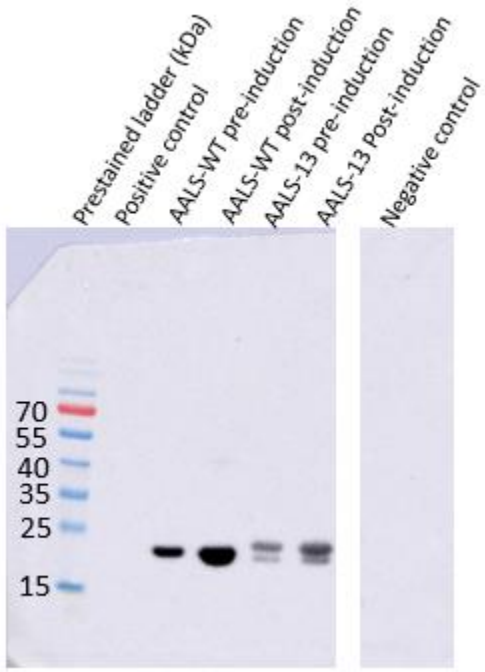


Figure A2.6 A western blot of overexpressed AaLS proteins. T_0 represents preinduction and T_2 indicating post induction. 5 µL of 1.0 OD 600 nm cell lysates were analyzed on a 12% PA gel at 180 V for 45 minutes. Samples were then transferred onto a 45 µm nitrocellulose membrane at 100 mAMP for 60 minutes. The membrane after the addition of a primary and secondary antibody was stained with luminol and imaged on an Amersham 1600 imager. The membrane was manually exposed for 10 minutes. Purified EF-TuHis protein was used as a positive control and BL21 (DE3) cell lysate was used as a negative control. The image is cropped to remove unrelated samples from the gel image.

No bands could be observed in the empty BL21 (DE3) cells indicating that *E. coli* proteins are not contributing to the band intensity within the AaLS samples. The positive control is not visible as time during manual exposure was not long enough. A repeat experiment (Figure SA2.1) shows a different positive control with the same results as Figure A2.6. The repeated western blot also only shows the one 24 kDa band in lanes 5 and 6 despite using the same sample as Figure A2.6. The 21 kDa band, despite being a AaLS-13 protein (which is likely a truncated AaLS-13 protein that has a His-tag) is therefore less stable over freeze-thaw cycles. There is no indication that different oligomeric states are detected by western blot. From this point, efforts

were towards purifying the AaLS-13 as it was our main focus for biophysical characterization and therefore optimization in purification and assembly.

A western blot was performed on a AaLS-13 nickel affinity purification to confirm if the elution's contained higher order AaLS-13 structures. The purification optimizations are discussed in the next section. The western blot contained several purification samples including the supernatant before binding to the resin, elution samples that contain the hypothesized oligomerization bands, and the overall pooled sample (Figure A2.7). In the supernatant and faintly seen in the elution sample, there are multiple bands suggesting AaLS-13 proteins with His-tags at ~20kDa, 50kDa, 70kDa, and 100kDa. These additional bands suggests that AaLS-13 starts to assemble after cell lysis. However, the pooled sample does not show these multiple states suggesting that the oligomerization is either affected by the concentrating of the sample, the use of the SEC buffer, or the protein has aggregated and was subsequently removed in one of the centrifugation steps.

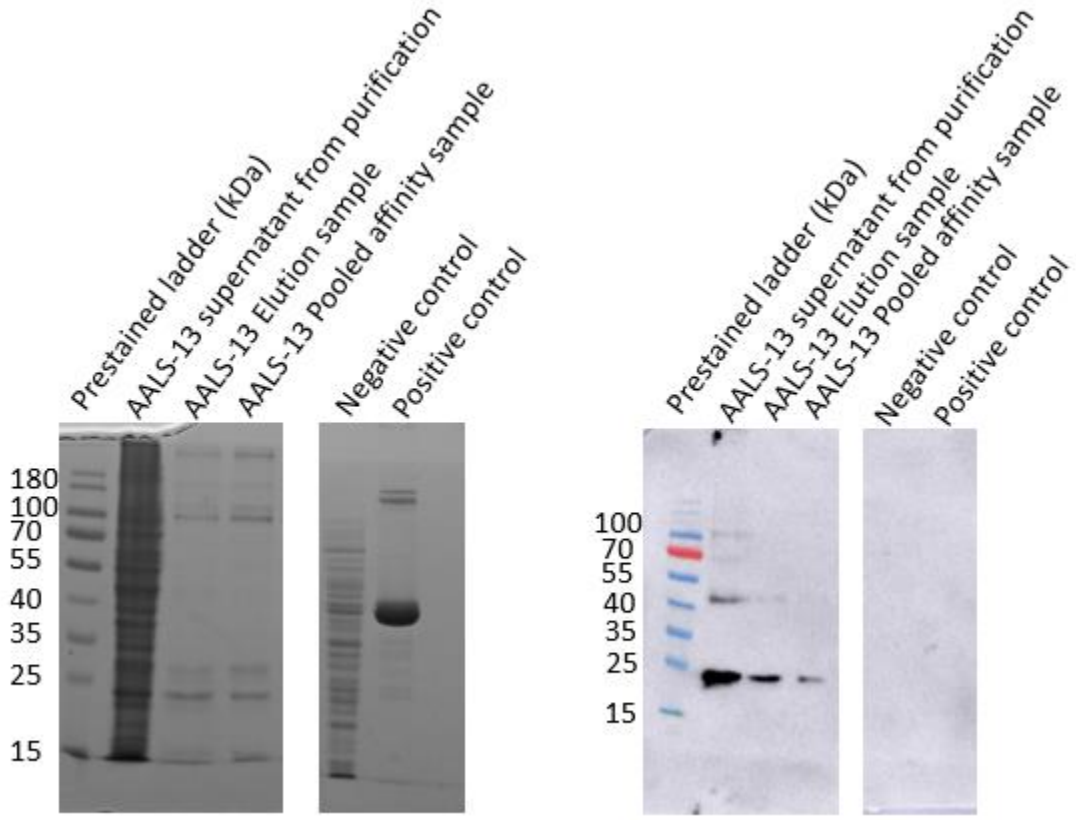


Figure A2.7 Western blot analysis of purified AaLS-13 proteins. The 12% PA gel on the left was stained with Coomassie G-250 and the 45 μm nitrocellulose membrane on the right was stained with luminol. AaLS-13 samples from the purification were analyzed on a 12% PA gel at 180 V for 45 minutes. Transfer onto a nitrocellulose membrane was performed (100 mAmp for 60 minutes). Purified EF-TuHis protein was used as a positive control and BL21 (DE3) cell lysate was used as a negative control. The image is cropped to remove unrelated samples from the gel image.

A2.3.2 Toward Optimizing AaLS-13 Protein Purification using Nickel Affinity Chromatography and Size Exclusion Chromatography

Initial attempts to purify AaLS-13 proteins used the buffers described in section 2.2.3, except for the lysis buffer that did not initially contain 1 mM EDTA and 0.05% Tween-20. These additions were made in further experiments to improve protein stability and AaLS-13 assembly (Azuma & Hilvert, 2018). Nickel affinity purification was successful in binding the AaLS-13

protein to the nickel NTA resin and subsequent elution, but with large quantities of contaminants (Figure A2.8).

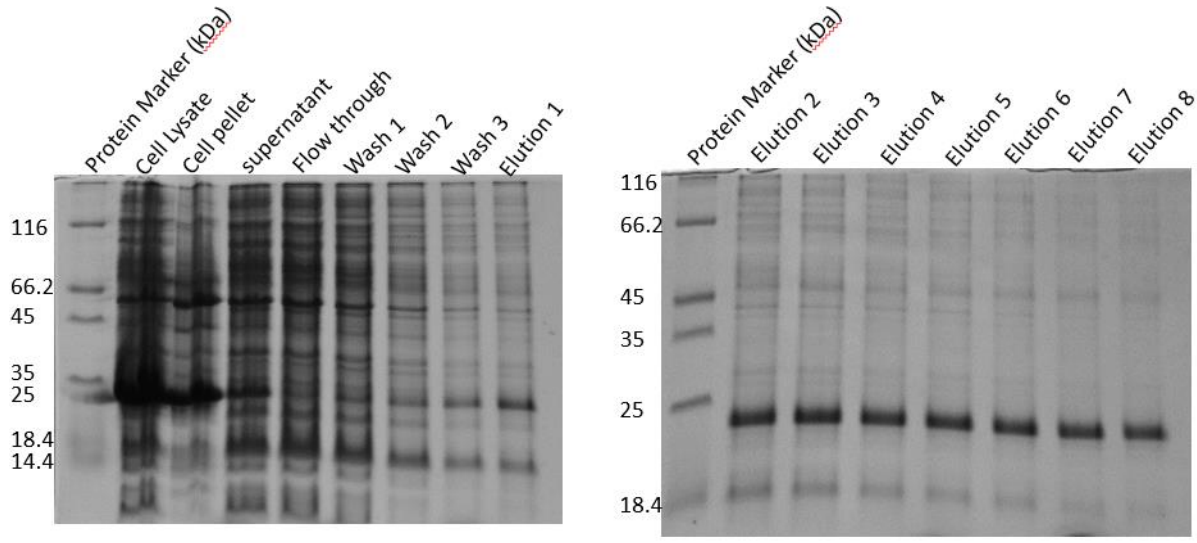


Figure A2.8 AaLS-13 proteins purified using affinity chromatography. 10 μL of AaLS-13 samples from the purification was analyzed on a 15% Tris-Tricine PAGE (200 V for 1.5 hours). The gel was stained with Coomassie G-250. The cell opening (left gel Lanes 2-4), binding and washing to the resin (left gel Lanes 5-8) and elution from the resin (left gel lane 9 and right gel lanes 2-8) are noted.

As samples were concentrated for the AaLS-13 assembly, aggregation was observed. Previous reports indicate that aggregation is a common observation in AaLS-13 purification and was determined to not result in significant losses (Azuma & Hilvert, 2018). AaLS-13 assembly was assessed by separating the sample on a 16/60 S400 Sepharose column (Cytiva). An example of AaLS-13 proteins in a SEC experiment can be seen in Figure A2.9. Assembled particles are expected to elute at ~30 mL while misassembled or unassembled AaLS-13 proteins are expected to elute afterwards. However, after induction of assembly through increased salt concentration and incubation for 3-5 days at room temperature, no complete assembly was observed. The aggregates are hypothesized to be all the AaLS-13 proteins that were assembled. Assembly using the same buffer conditions was attempted several times. Attempts to resolubilize the aggregates

using sonication (Azuma & Hilvert, 2018) and protein unfolding using urea were conducted but did not improve the assembly of the AaLS-13 (data not shown).

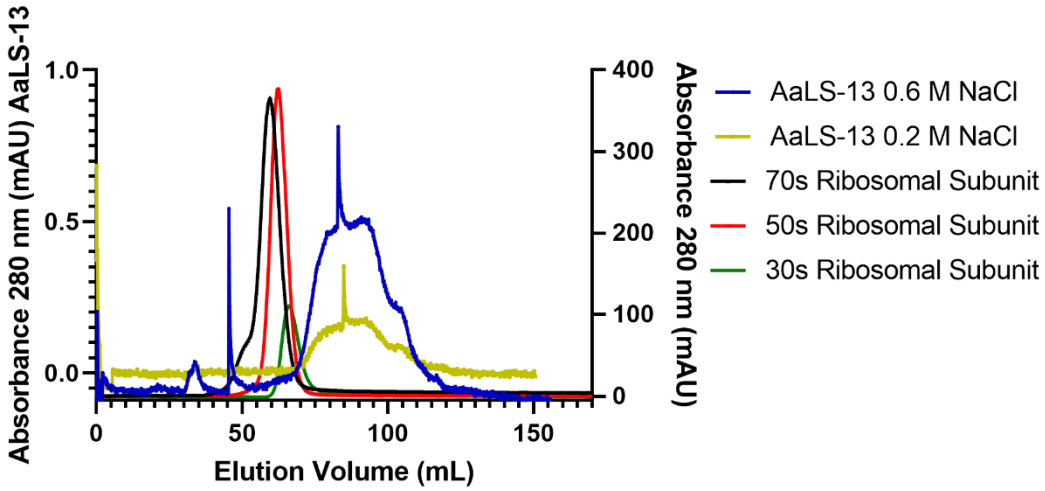


Figure A2.9 SEC of AaLS-13 at different salt concentrations. AALS-13 proteins were injected onto a 16/60 S400 Sepharose HR column on a ÄKTA Prime (Cytiva) system at room temperature to assess assembly when treated with high or low salt conditions. 28,640 picomoles of AaLS-13 protein in 0.6 M NaCl and 32,820 picomoles of AaLS-13 protein in 0.2 M NaCl were injected onto the column. Flow rate was 1 ml/min. The sharp peaks in some of the elution profiles (~50 mL and ~90 mL) are artifacts caused by the pausing of the purification system. The ribosomal proteins were from previous experiments and provided by a fellow lab member to be used as a size standard.

The lysis buffer used for nickel affinity purification was modified by the addition of EDTA and Tween-20, to prevent aggregation and hydrophobic interactions with cellular components, respectively. The original published protocol suggested these amendments if issues in assembly are observed (Azuma & Hilvert, 2018). However, aggregation and low assembly still persisted after sample concentration and with the addition of 5 M NaCl (0.6 M NaCl final concentration) (Figure A2.10). The majority of AaLS-13 protein eluted after 80 mL of elution volume, suggesting that only misassembled AaLS-13 proteins are present, as this corresponds to a

molecular weight of less than 1 MDa, as compared to the 30s ribosomal subunit reference sample (Culver *et al.*, 2008). The expected size of AaLS-13 is ~3 MDa (Sasaki *et al.*, 2017).

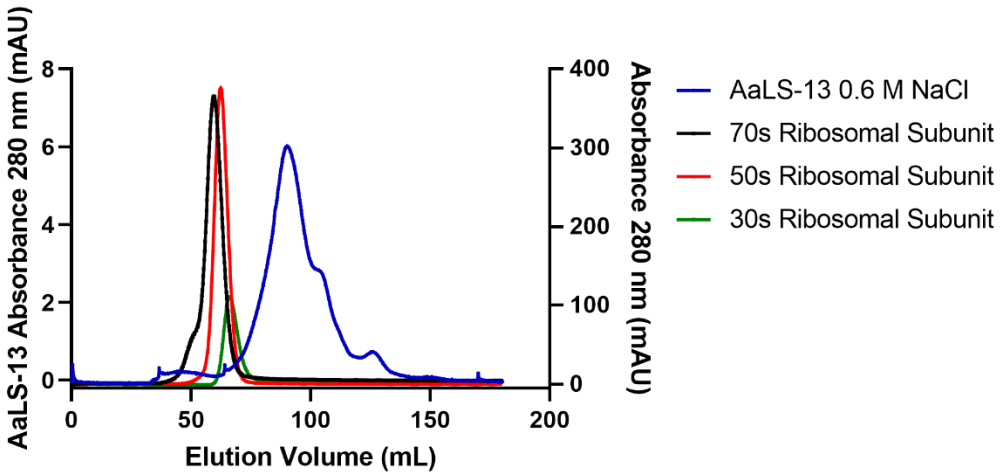


Figure A2.10 SEC Chromatogram of AALS-13 purified via nickel affinity. AaLS-13 samples were injected on a 16/60 S400 Sepharose HR column on a ÄKTA Prime (Cytiva) system at room temperature. 546,890 picomoles of AaLS-13 protein in 0.6 M salt conditions was injected onto the column. Flow rate was 1 ml/min. The ribosomal proteins were from previous experiments and provided by a fellow lab member to be used as a size standard.

A2.4 Discussion

We assume that the failed assembly is due to two reasons: First, the lack of cargo that would otherwise aid in the assembly of the structure. Aside of AaLS-13 assembly as an empty shell (Azuma & Hilvert, 2018), other methods have been described to facilitate stable assembly of AaLS-13. The use of a highly charged proteins can be used to facilitate assembly (Azuma, Bader, *et al.*, 2018). The second possible reason for the lack of assembly is a mistake in the cloning of the AaLS protein constructs, which was discovered after the previously described purification and assembly attempts. The pET28a expression vector provides an N-terminal tag (MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSEF), that contains a N-terminal His-tag, a thrombin cleavage site and a T7 tag. The AaLS-13 coding sequence that was designed and

inserted into the vector already contained a C-terminal His-tag. To the best of our knowledge, only C-terminal His-tags have been used for AaLS-13 purification so far. The large tag and its positioning at the opposite terminus could be contributing to problems during assembly.

To address the first hypothesis, we propose a new purification method using cargo. The purified cargo (which will be supercharged GFP) could be added to the AaLS-13 shell at elution from the nickel affinity column, to allow for assembly *in vitro*. Adding a cargo therefore can be used as an alternative method, allowing for the continuation of biophysical characterization using SEC-MALS and AUC in the future.

As assembly can start already during nickel affinity purification, there is a chance of encapsulating contaminants in this step. The unfolding of the AaLS-13 protein while bound to the nickel column using buffer containing GdnHCl will also decrease contaminants that may be in the purified proteins during elution due to unspecific interactions with the shell protein; Contaminants will unfold and will not be retained on the column during the wash. The addition of GdnHCl will also prevent the formation or partial assembled AaLS-13 compartments prior to the introduction of cargo.

In conclusion, the plan was to produce and optimize a protocol in which the AaLS-13 protein is unfolded on the column and refolded (Azuma, Edwardson, *et al.*, 2018) prior to the elution into the cargo sample. The cargo chosen for initial assembly attempts is supercharged GFP (+36), which has been used previously to facilitate assembly and for encapsulation studies of AaLS-13 (Azuma, Bader, *et al.*, 2018). The protein itself has been engineered to contain more charged residues (Thompson *et al.*, 2012) which will interact with the negatively charged lumen of AaLS-13. The purification of GFP (+36) has been successfully performed and the purification

of AaLS-13 using the folding and unfolding method will be done in future experiments (Figure A2.11).

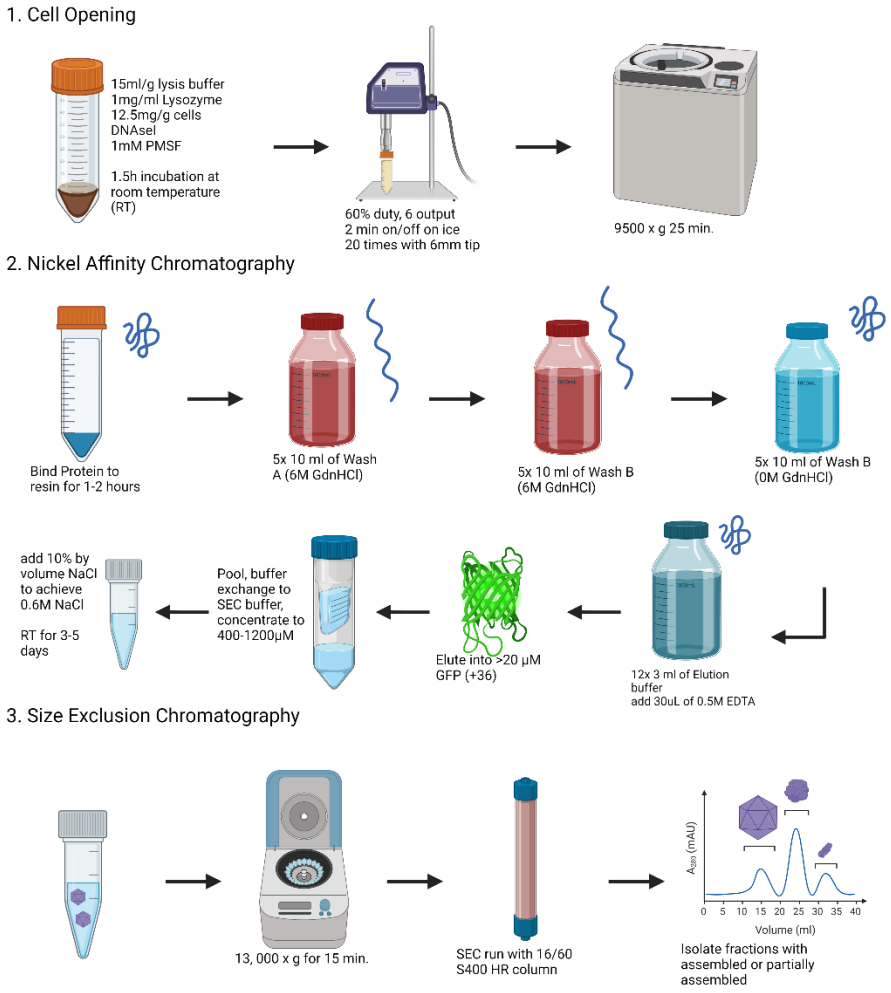


Figure A2.11 Schematic of the unfolding nickel affinity purification and assembly protocol using GFP (+36) of AaLS-13. Created with Biorender.com.

GFP (+36) containing a N-terminal hexa-histidine tag was obtained from Addgene (#62937) and transformed into BL21 (DE3) cells for overexpression and purification. The first step of purification was nickel affinity purification using the batch purification method (Figure A2.12). Initial elution's contain large amounts of contaminants suggesting that efforts towards

the removal of contaminants and proteins that can bind weakly to the nickel Sepharose resin should be performed. Elution's 6-12 (a total volume of 28 mL) were pooled used for SEC purification as they showed the least contaminants. The GFP (+36) protein (29.5 kDa) was buffer exchanged and concentrated to 1 mL using an ultracentrifuge filter with a 10 kDa molecular weight cut off (Cytiva) for SEC purification.

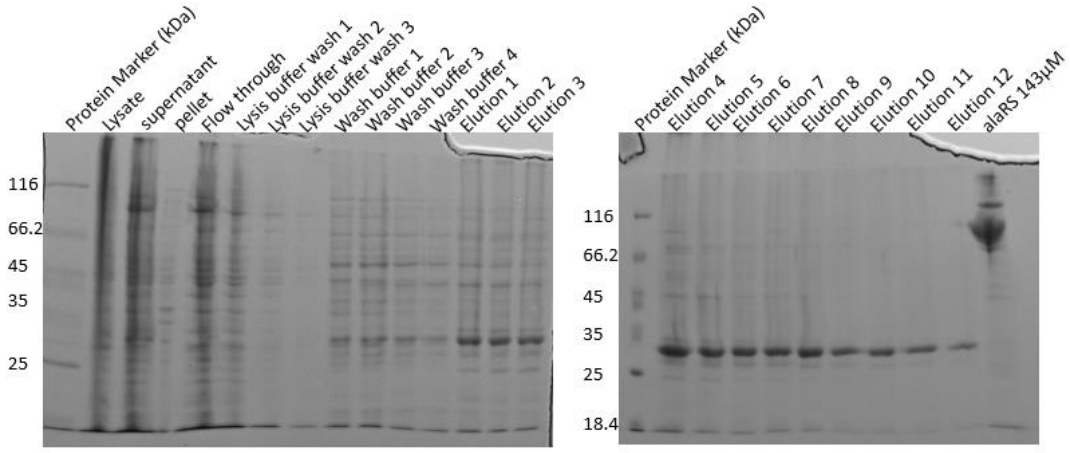


Figure A2.12 Nickel affinity purification of GFP (+36). 10 μL of GFP (+36) protein was analyzed on a 12% SDS-PAGE and run at 180 V for 45 minutes. The gel was stained with Coomassie G-250.

After concentrating, the GFP (+36) protein preparation was split into two 500 μL aliquots which were separated on a Superdex 75 10/300 GL increase column (Cytiva). The chromatograms of the samples can be found in Figure A2.13. Both experiments have consistent elution profiles with the maximum peak containing GFP (+36) at ~13 mL.

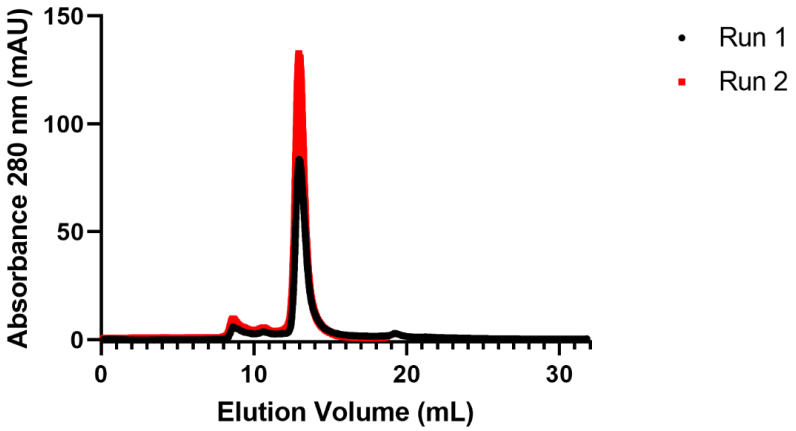


Figure A2.13 Chromatogram of GFP (+36) SEC. Samples were injected as 2 x 500µL replicates on a Superdex 75 10/300 GL increase column on a ÄKTA Pure system (Cytiva). Flow rate was set at 0.4 mL/min and protein was collected in 0.5 mL fractions.

The peak fractions (13-14 mL elution volume) were analysed on an 12% SDS-PAGE (Figure A2.14). The samples after pooling were found to be of high purity (~98%) based on Coomassie staining (Figure AS2.2), concentrated to 3 mL at 68 µM, and stored for future use.

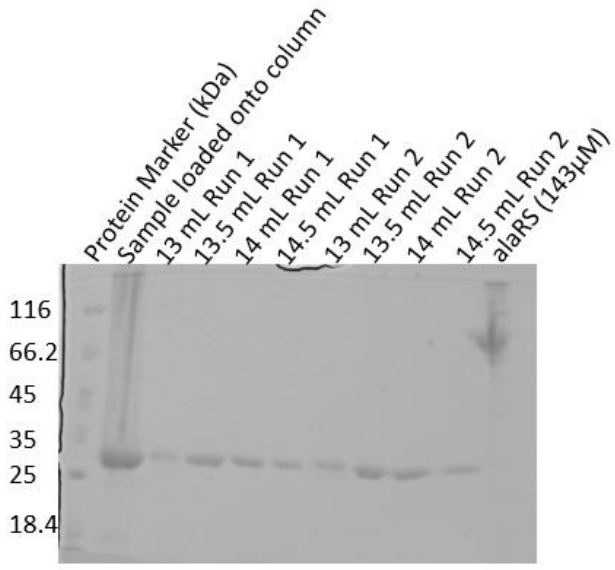


Figure A2.14 Analysis of GFP (+36) proteins separated on a Superdex 75 10/300 GL SEC column. 10 µL of GFP (+36) protein were analyzed on a 12% SDS-PAGE developed at 180 V for 50 minutes. The gel was stained with Coomassie G-250.

To address the second hypothesis, that the N-terminal tag may be affecting the assembly of the LS particle, the structure with the tag (Figure A2.15) was modelled using I-TASSER (Zheng *et al.*, 2021). Structural modelling was done to determine if the N-terminal tag is structured and if this structure may be in a region that could affecting assembly. The LS monomer protein with the long N-terminal tag and C-terminal His-tag was submitted to the I-TASSER server (Zheng *et al.*, 2021), to predict the structure of the tag. The crystal structure of the AaLS-13 monomer with the C-terminal His-tag (PDB: 5MQ7: Chain A) was used as a template for the model. The AaLS-13 crystal structure in the intended design and the 5 predicted structural models are shown in Figure A2.15.

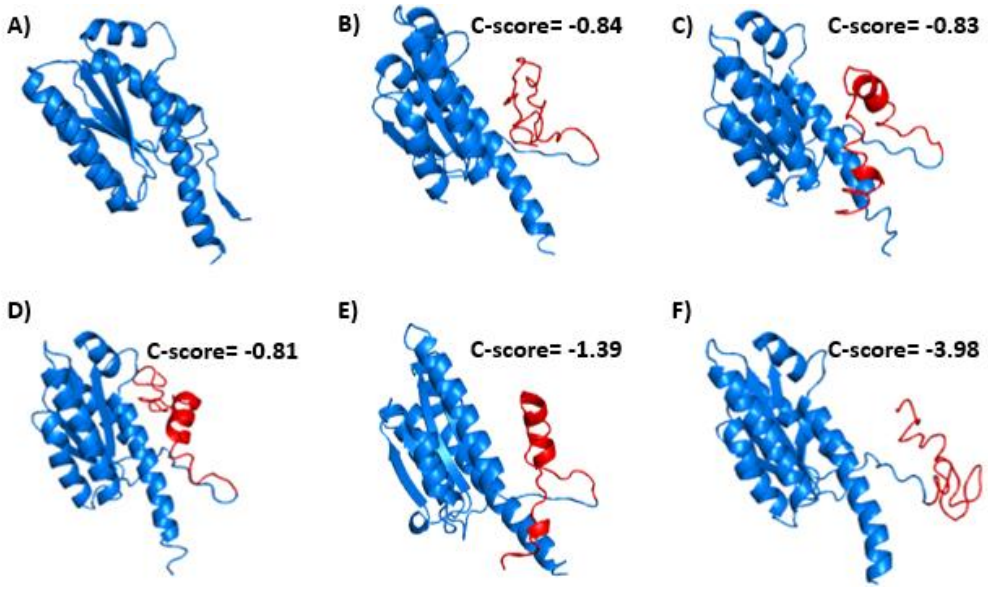


Figure A2.15. Predicted structures of AaLS-13 with the N-terminal Tag. The AaLS-13 amino acid sequence with the N- and C-terminal tag was modelled using the I-TASSER server with the use of the AaLS-13 crystal structure with the C-terminal His-tag (PDB: 5MQ7) as a modelling constraint. **A)** AaLS-13 with C-terminal His-tag, **B)-F)** I-TASSER models 1 to 5 with the N-terminal tag in red. The confidence scores (C-score) for each model are shown.

All 5 models are evaluated based on the C-score which is the confidence score that indicates the quality of the model's protein secondary structure prediction. The C-score also takes the base

structure template (the AaLS-WT crystal structure on the pdb database) into consideration where models 1, 2 and 3 are considered the most accurate. The I-TASSER server can also predict a TM-score for assessing structural similarity between structures. The TM score of the first model is 0.61 ± 0.14 which suggests accurate modelling of the structure to the template. The server typically only runs a TM score with the best C-score model. Based on the C-score and TM-score, we decided to use model one for further analysis which was superimposing the structure onto the pentameric unit of AaLS-13 (Figure A2.16). The pentamer is the first order structure of the AaLS-13 shell; pentamers interact with each other to form the final structure.

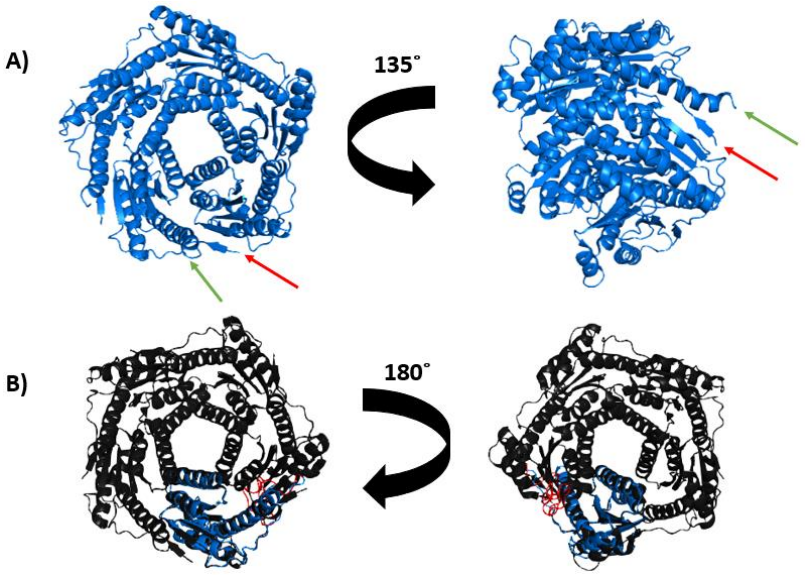


Figure A2.16 AaLS-13 pentamers compared to the AaLS-13 monomer with N-terminal His-tag. A) AaLS-13 pentamer with concave face forward. The pentamer is then rotated 135˚ to show the location of the C-terminus and N-terminus in the structure where both are positioned on the side that faces the shell lumen (highlighted by the red and green arrows respectively) **B)** The AaLS-13 pentamer (grey) superimposed with the AaLS-13 monomer (blue) with the N-terminal tag (red). The concave face is forward on the left side of the figure. The structure is rotated 180˚ to show the position of the tag on the lumen side of the pentamer.

Based on superimposing the AaLS-13 protein with the N-terminal tag onto the pentameric subunit, it is evident that the tag points towards the outer region of the pentamer as well as the

concave side of the subunit that would otherwise be a part of the lumen of AaLS-13. From this, it likely that the N-terminal tag affects the assembly. First, as the tag has no secondary structure and is near the outer edge of the pentamer, it is highly dynamic and can cause interference with other pentameric subunits during assembly or affect any other protein-protein interactions that would facilitate assembly. This may explain why higher order structures were observed in western blot experiments (Figure A2.7). Such interference would also be consistent with the observed partial, but not full assembly of the AaLS particles in our SEC experiments (Figure A2.9). Secondly, as the N-terminal tag faces inside the shell, it may also be disrupting the assembly of the pentameric subunits to facilitate assembly. As the tag does have positively charged amino acids it will also affect the overall charge of the lumen which could counteract the shell assembly through increasing the salt concentration. Based on this assumption, the removal of the N-terminal tag is of the largest priority to obtain assembly and afterwards the intended biophysical characterization without cargo.

A2.5 Future Directions

Lumazine synthase is a simple yet extremely stable protein cage that tolerates protein engineering and encapsulation of different cargo's, making it an interesting target for a drug delivery system or vaccination tool. However, homogeneity, structure, and stability are the largest factors in building a drug delivery system (Pircalabioru *et al.*, 2020). Therefore, it is important to have a quick and easy way of analyzing particles during pre-clinical and clinical development. Using methods such as SEC-MALS and AUC can be used as reliable experimental methods for analyzing protein cages and determining their viability for such applications.

Despite the many publications describing assembled AaLS proteins (Azuma, Bader, *et al.*, 2018; Azuma *et al.*, 2017; Azuma *et al.*, 2016; Terasaka *et al.*, 2018; Worsdorfer *et al.*, 2011), all

attempts to assemble and purify AaLS-13 failed. Amendments to the purifications buffers and sonication methods did not improve the assembly. The presence of an extended N-terminal tag seems to be the major reason for this. Removing the tag from the coding sequence can be achieved by inverse PCR. If assembly issues still arise, the next approach is to use supercharged cargo and protein unfolding of AaLS-13 during nickel affinity purification. Avoiding high amounts of contaminants and providing stability through cargo is hypothesized to provide large concentrations of assembled protein cages that can be used for biophysical characterization using AUC and SEC-MALS.

A2.6 Supplemental Figures

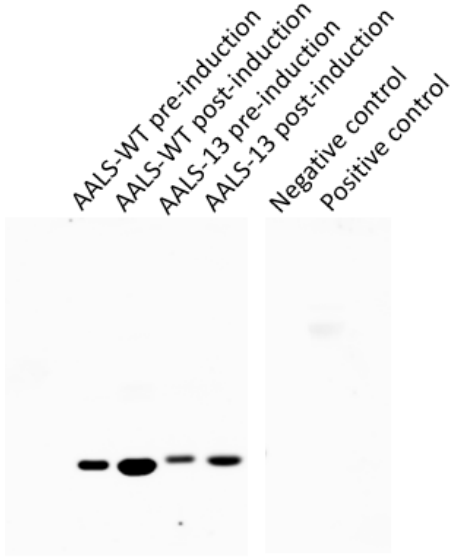


Figure AS2.1 Western blot of AaLS over-expressions. T_0 represents preinduction and T_2 indicating post induction (18 hours). 10 μL of 1.0 $OD_{600\ nm}$ cell lysates were analyzed on a 12% SDS-PAGE at 180 V for 45 minutes. Samples were then transferred onto a 45 μm nitrocellulose membrane at 100 mAMP for 60 minutes. Samples were stained with luminol and imaged using chemiluminescence on an Amersham 1600. Samples were manually exposed for 15 minutes. BL21 (De3) cells were used as a negative control and purified alaRS protein (143 μM) was used as a positive control. The image is cropped to remove unrelated samples from the gel image.

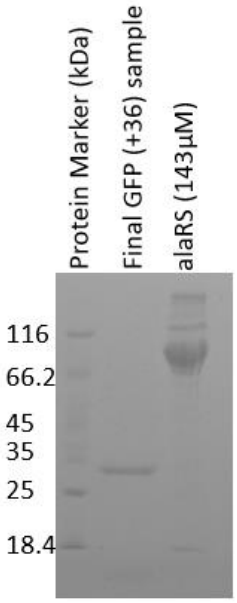


Figure AS2.2 The final GFP (+36) purified sample in storage buffer. The buffer contained 20 mM phosphate buffer with 1M NaCl and 20% glycerol)**.** 10 μL of purified GFP(+36) were analyzed on a 12% SDS-PAGE gel (180 V for 50 minutes). Stained with Coomassie G-250.

REFERENCES

Andersson, M., Wittgren, B., & Wahlund, K. G. (2003). Accuracy in multiangle light scattering measurements for molar mass and radius estimations. Model calculations and experiments. *Anal Chem, 75*(16), 4279-4291. doi:10.1021/ac030128+

Axen, S. D., Erbilgin, O., & Kerfeld, C. A. (2014). A taxonomy of bacterial microcompartment loci constructed by a novel scoring method. *PLoS Comput Biol, 10*(10), e1003898. doi:10.1371/journal.pcbi.1003898

Aziz, Z., Daugherty, M., de la Torre, J., Demeler, B., Douady, C., et al. (2007). *Analytical ultracentrifugation: techniques and methods*. Cambridge: Royal Society of Chemistry.

Azuma, Y., Bader, D. L. V., & Hilvert, D. (2018). Substrate Sorting by a Supercharged Nanoreactor. *J Am Chem Soc, 140*(3), 860-863. doi:10.1021/jacs.7b11210

Azuma, Y., Edwardson, T. G. W., & Hilvert, D. (2018). Tailoring lumazine synthase assemblies for bionanotechnology. *Chem Soc Rev, 47*(10), 3543-3557. doi:10.1039/c8cs00154e

Azuma, Y., & Hilvert, D. (2018). Enzyme Encapsulation in an Engineered Lumazine Synthase Protein Cage. *Methods Mol Biol, 1798*, 39-55. doi:10.1007/978-1-4939-7893-9_4

Azuma, Y., Zschoche, R., & Hilvert, D. (2017). The C-terminal peptide of *Aquifex aeolicus* riboflavin synthase directs encapsulation of native and foreign guests by a cage-forming lumazine synthase. *J Biol Chem, 292*(25), 10321-10327. doi:10.1074/jbc.C117.790311

Azuma, Y., Zschoche, R., Tinzl, M., & Hilvert, D. (2016). Quantitative Packaging of Active Enzymes into a Protein Cage. *Angew Chem Int Ed Engl, 55*(4), 1531-1534. doi:10.1002/anie.201508414

Baker, S. H., Williams, D. S., Aldrich, H. C., Gambrell, A. C., & Shively, J. M. (2000). Identification and localization of the carboxysome peptide Csos3 and its corresponding gene in *Thiobacillus neapolitanus*. *Arch Microbiol, 173*(4), 278-283. doi:10.1007/s002030000141

Bohren, C. a. H., D. (1998). *Absorption and Scattering of Light by Small Particles*. New York: John Wiley & Sons.

Bonacci, W., Teng, P. K., Afonso, B., Niederholtmeyer, H., Grob, P., et al. (2012). Modularity of a carbon-fixing protein organelle. *Proc Natl Acad Sci U S A, 109*(2), 478-483. doi:10.1073/pnas.1108557109

Briand, L., Marcion, G., Kriznik, A., Heydel, J. M., Artur, Y., et al. (2016). A self-inducible heterologous protein expression system in *Escherichia coli. Sci Rep, 6*, 33037. doi:10.1038/srep33037

Brookes, E., Cao, W., & Demeler, B. (2010). A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. *Eur Biophys J, 39*(3), 405-414. doi:10.1007/s00249-009-0413-5

Burghardt, R. C., & Droleskey, R. (2006). Transmission electron microscopy. *Curr Protoc Microbiol, Chapter 2*, Unit 2B 1. doi:10.1002/9780471729259.mc02b01s03

Cai, F., Menon, B. B., Cannon, G. C., Curry, K. J., Shively, J. M., et al. (2009). The pentameric vertex proteins are necessary for the icosahedral carboxysome shell to function as a CO2 leakage barrier. *PLoS One, 4*(10), e7521. doi:10.1371/journal.pone.0007521

Cai, F., Sutter, M., Bernstein, S. L., Kinney, J. N., & Kerfeld, C. A. (2015). Engineering Bacterial Microcompartment Shells: Chimeric Shell Proteins and Chimeric Carboxysome Shells. *Acs Synth Biol, 4*(4), 444-453. doi:10.1021/sb500226j

Cannon, G. C., & Shively, J. M. (1983). Characterization of a homogenous preparation of carboxysomes from Thiobacillus neapolitanus. *Arch Microbiol, 134*(1), 52-59. doi:10.1007/BF00429407

Chen, Z. H., Kim, C., Zeng, X. B., Hwang, S. H., Jang, J., et al. (2012). Characterizing size and porosity of hollow nanoparticles: SAXS, SANS, TEM, DLS, and adsorption isotherms compared. *Langmuir, 28*(43), 15350-15361. doi:10.1021/la302236u

Cole, J. L., Lary, J. W., T, P. M., & Laue, T. M. (2008). Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. *Methods Cell Biol, 84*, 143-179. doi:10.1016/S0091-679X(07)84006-4

Culver, G. M., & Kirthi, N. (2008). Assembly of the 30S Ribosomal Subunit. *EcoSal Plus, 3*(1). doi:10.1128/ecosalplus.2.5.3

Dai, W., Chen, M., Myers, C., Ludtke, S. J., Pettitt, B. M., et al. (2018). Visualizing Individual RuBisCO and Its Assembly into Carboxysomes in Marine Cyanobacteria by Cryo-Electron Tomography. *J Mol Biol, 430*(21), 4156-4167. doi:10.1016/j.jmb.2018.08.013

Dam, J., & Schuck, P. (2004). Calculating Sedimentation Coefficient Distributions by Direct Modeling of Sedimentation Velocity Concentration Profiles. In *Methods in Enzymology* (Vol. 384, pp. 185-212): Academic Press.

Demeler, B., & Gorbet, G. E. (2016). Analytical Ultracentrifugation Data Analysis with UltraScan-III. In S. Uchiyama, F. Arisaka, W. F. Stafford, & T. Laue (Eds.), *Analytical Ultracentrifugation: Instrumentation, Software, and Applications* (pp. 119-143). Tokyo: Springer Japan.

Dou, Z., Heinhorst, S., Williams, E. B., Murin, C. D., Shively, J. M., et al. (2008). CO2 fixation kinetics of *Halothiobacillus neapolitanus* mutant carboxysomes lacking carbonic anhydrase suggest the shell acts as a diffusional barrier for CO2. *J Biol Chem, 283*(16), 10377-10384. doi:10.1074/jbc.M709285200

Edwards, G. B., Muthurajan, U. M., Bowerman, S., & Luger, K. (2020). Analytical Ultracentrifugation (AUC): An Overview of the Application of Fluorescence and Absorbance AUC to the Study of Biological Macromolecules. *Curr Protoc Mol Biol, 133*(1), e131. doi:10.1002/cpmb.131

Espie, G. S., & Kimber, M. S. (2011). Carboxysomes: cyanobacterial RubisCO comes in small packages. *Photosynth Res, 109*(1-3), 7-20. doi:10.1007/s11120-011-9656-y

Evans, S., Al-Hazeem, M., Mann, D., Smetacek, N., Beavil, A., Sun, Y., et al. (2022). Single-particle cryo-EM analysis of the shell architecture and internal organization of an intact α-carboxysome. *Preprint*. doi:10.1101/2022.02.18.481072

Faulkner, M., Szabo, I., Weetman, S. L., Sicard, F., Huber, R. G., et al. (2020). Molecular simulations unravel the molecular principles that mediate selective permeability of carboxysome shell protein. *Sci Rep, 10*(1), 17501. doi:10.1038/s41598-020-74536-5

Frey, R., Mantri, S., Rocca, M., & Hilvert, D. (2016). Bottom-up Construction of a Primordial Carboxysome Mimic. *J Am Chem Soc, 138*(32), 10072-10075. doi:10.1021/jacs.6b04744

Fridlyand, L., Kaplan, A., & Reinhold, L. (1996). Quantitative evaluation of the role of a putative CO2-scavenging entity in the cyanobacterial CO2-concentrating mechanism. *Biosystems, 37*(3), 229-238. doi:10.1016/0303-2647(95)01561-2

Gonzalez-Esquer, C. R., Newnham, S. E., & Kerfeld, C. A. (2016). Bacterial microcompartments as metabolic modules for plant synthetic biology. *Plant J, 87*(1), 66-75. doi:10.1111/tpj.13166

Hagen, A., Sutter, M., Sloan, N., & Kerfeld, C. A. (2018). Programmed loading and rapid purification of engineered bacterial microcompartment shells. *Nat Commun, 9*(1), 2881. doi:10.1038/s41467-018-05162-z

Han, Y., Li, D., Li, D., Chen, W., Mu, S., et al. (2020). Impact of refractive index increment on the determination of molecular weight of hyaluronic acid by muti-angle laser light-scattering technique. *Sci Rep, 10*(1), 1858. doi:10.1038/s41598-020-58992-7

Hanson, M. R., Lin, M. T., Carmo-Silva, A. E., & Parry, M. A. (2016). Towards engineering carboxysomes into C3 plants. *Plant J, 87*(1), 38-50. doi:10.1111/tpj.13139

Harvey, A., Kaplan, S., & Burnett, J. (2005). Effect of dissolved air on the Density and Refractive Index of water. *Int J Thermophys, 26*(5), 1495-1514. doi:10.1007/s10765-005-8099-0

Henrickson, A., Kulkarni, J. A., Zaifman, J., Gorbet, G. E., Cullis, P. R., et al. (2021). Density Matching Multi-wavelength Analytical Ultracentrifugation to Measure Drug Loading of Lipid Nanoparticle Formulations. *ACS Nano, 15*(3), 5068-5076. doi:10.1021/acsnano.0c10069

Horne, C. R., Henrickson, A., Demeler, B., & Dobson, R. C. J. (2020). Multi-wavelength analytical ultracentrifugation as a tool to characterise protein-DNA interactions in solution. *Eur Biophys J, 49*(8), 819-827. doi:10.1007/s00249-020-01481-6

Huang, J., Ferlez, B. H., Young, E. J., Kerfeld, C. A., Kramer, D. M., et al. (2019). Functionalization of Bacterial Microcompartment Shell Proteins With Covalently Attached Heme. *Front Bioeng Biotechnol, 7*, 432. doi:10.3389/fbioe.2019.00432

Jablonsky, J., Bauwe, H., & Wolkenhauer, O. (2011). Modeling the Calvin-Benson cycle. *BMC Syst Biol, 5*, 185. doi:10.1186/1752-0509-5-185

Kaasalainen, M., Aseyev, V., von Haartman, E., Karaman, D. S., Makila, E., et al. (2017). Size, Stability, and Porosity of Mesoporous Nanoparticles Characterized with Light Scattering. *Nanoscale Res Lett, 12*(1), 74. doi:10.1186/s11671-017-1853-y

Katsura, S., Yamaguchi, A., Hirano, K., Matsuzawa, Y., & Mizuno, A. (2000). Manipulation of globular DNA molecules for sizing and separation. *Electrophoresis, 21*(1), 171-175. doi:10.1002/(SICI)1522-2683(20000101)21:1<171::AID-ELPS171>3.0.CO;2-U

Kennedy, N. W., Hershewe, J. M., Nichols, T. M., Roth, E. W., Wilke, C. D., et al. (2020). Apparent size and morphology of bacterial microcompartments varies with technique. *PLoS One, 15*(3), e0226395. doi:10.1371/journal.pone.0226395

Kerfeld, C. A. (2017). A bioarchitectonic approach to the modular engineering of metabolism. *Philos Trans R Soc Lond B Biol Sci, 372*(1730). doi:10.1098/rstb.2016.0387

Kerfeld, C. A., & Erbilgin, O. (2015). Bacterial microcompartments and the modular construction of microbial metabolism. *Trends Microbiol, 23*(1), 22-34. doi:10.1016/j.tim.2014.10.003

Kerfeld, C. A., Heinhorst, S., & Cannon, G. C. (2010). Bacterial microcompartments. *Annu Rev Microbiol, 64*, 391-408. doi:10.1146/annurev.micro.112408.134211

Kerfeld, C. A., & Melnicki, M. R. (2016). Assembly, function and evolution of cyanobacterial carboxysomes. *Curr Opin Plant Biol, 31*, 66-75. doi:10.1016/j.pbi.2016.03.009

Kerfeld, C. A., Sawaya, M. R., Tanaka, S., Nguyen, C. V., Phillips, M., et al. (2005). Protein structures forming the shell of primitive bacterial organelles. *Science, 309*(5736), 936-938. doi:10.1126/science.1113397

Khokhlov, A. a. G., E. . (2000). *Lectures on Physical Chemistry of Polymers*. Moscow: Mir.

Kinney, J. N., Axen, S. D., & Kerfeld, C. A. (2011). Comparative analysis of carboxysome shell proteins. *Photosynth Res, 109*(1-3), 21-32. doi:10.1007/s11120-011-9624-6

Kirst, H., Ferlez, B. H., Lindner, S. N., Cotton, C. A. R., Bar-Even, A., et al. (2022). Toward a glycyl radical enzyme containing synthetic bacterial microcompartment to produce pyruvate from formate and acetate. *Proc Natl Acad Sci U S A, 119*(8). doi:10.1073/pnas.2116871119

Kirst, H., & Kerfeld, C. A. (2019). Bacterial microcompartments: catalysis-enhancing metabolic modules for next generation metabolic and biomedical engineering. *BMC Biol, 17*(1), 79. doi:10.1186/s12915-019-0691-z

Klein, M. G., Zwart, P., Bagby, S. C., Cai, F., Chisholm, S. W., et al. (2009). Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. *J Mol Biol, 392*(2), 319-333. doi:10.1016/j.jmb.2009.03.056

Kratochvl, P. (1987). *Classical Light Scattering from Polymer Solutions (Polymer Science Library, 5)*. Amsterdam: Elsevier Science Ltd. .

Ladenstein, R., Fischer, M., & Bacher, A. (2013). The lumazine synthase/riboflavin synthase complex: shapes and functions of a highly variable enzyme system. *FEBS J, 280*(11), 2537-2563. doi:10.1111/febs.12255

Lee, M. J., Palmer, D. J., & Warren, M. J. (2019). Biotechnological Advances in Bacterial Microcompartment Technology. *Trends Biotechnol, 37*(3), 325-336. doi:10.1016/j.tibtech.2018.08.006

Li, T., Jiang, Q., Huang, J., Aitchison, C. M., Huang, F., et al. (2020). Reprogramming bacterial protein organelles as a nanoreactor for hydrogen production. *Nat Commun, 11*(1), 5448. doi:10.1038/s41467-020-19280-0

Lin, M. T., Occhialini, A., Andralojc, P. J., Devonshire, J., Hines, K. M., et al. (2014). beta-Carboxysomal proteins assemble into highly organized structures in *Nicotiana* chloroplasts. *Plant J, 79*(1), 1-12. doi:10.1111/tpj.12536

Liu, A., & Fletcher, D. (2009). Biology under construction: in vitro reconstitution of cellular function. *Nat Rev Mol Cell Biol, 10*(9), 644-650. doi:10.1038/nrm2746

Liu, Y., He, X., Lim, W., Mueller, J., Lawrie, J., et al. (2018). Deciphering molecular details in the assembly of alpha-type carboxysome. *Sci Rep, 8*(1), 15062. doi:10.1038/s41598-018-33074-x

Long, B. M., Hee, W. Y., Sharwood, R. E., Rae, B. D., Kaines, S., et al. (2018). Carboxysome encapsulation of the CO2-fixing enzyme Rubisco in tobacco chloroplasts. *Nat Commun, 9*(1), 3570. doi:10.1038/s41467-018-06044-0

Lopez-Sagaseta, J., Malito, E., Rappuoli, R., & Bottomley, M. J. (2016). Self-assembling protein nanoparticles in the design of vaccines. *Comput Struct Biotechnol J, 14*, 58-68. doi:10.1016/j.csbj.2015.11.001

Mallmann, J., Heckmann, D., Brautigam, A., Lercher, M. J., Weber, A. P., et al. (2014). The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. *Elife, 3*, e02478. doi:10.7554/eLife.02478

Maloy, S. R., Stewart, V.J. & Taylor, R. K. (1996). *Genetic Analysis of Pathogenic Bacteria : a Laboratory Manual* Plainview, NewYork: Cold Spring Harbor Laboratory Press.

Menon, B. B., Dou, Z., Heinhorst, S., Shively, J. M., & Cannon, G. C. (2008). *Halothiobacillus neapolitanus* carboxysomes sequester heterologous and chimeric RubisCO species. *PLoS One, 3*(10), e3570. doi:10.1371/journal.pone.0003570

Metskas, L., Ortega, D., Oltrogge, L., Blikstad, C., Laughlin, T., Savage, D., & Jensen, G. (2022). Rubisco forms a lattice inside alpha-carboxysomes. *bioRxiv*. doi:10.1101/2022.01.24.477598

Min, J., Kim, S., Lee, J., & Kang, S. (2014). Lumazine synthase protein cage nanoparticles as modular delivery platforms for targeted drug delivery. *RSC Advances, 4*(89), 48596-48600. doi:10.1039/C4RA10187A

Occhialini, A., Lin, M. T., Andralojc, P. J., Hanson, M. R., & Parry, M. A. (2016). Transgenic tobacco plants with improved cyanobacterial Rubisco expression but no extra assembly factors grow at near wild-type rates if provided with elevated CO2. *Plant J, 85*(1), 148-160. doi:10.1111/tpj.13098

Ochoa, J. M., & Yeates, T. O. (2021). Recent structural insights into bacterial microcompartment shells. *Curr Opin Microbiol, 62*, 51-60. doi:10.1016/j.mib.2021.04.007

Oltrogge, L. M., Chaijarasphong, T., Chen, A. W., Bolin, E. R., Marqusee, S., et al. (2020). Multivalent interactions between CsoS2 and Rubisco mediate alpha-carboxysome formation. *Nat Struct Mol Biol, 27*(3), 281-287. doi:10.1038/s41594-020-0387-7

Orf, I., Timm, S., Bauwe, H., Fernie, A. R., Hagemann, M., et al. (2016). Can cyanobacteria serve as a model of plant photorespiration? - a comparative meta-analysis of metabolite profiles. *J Exp Bot, 67*(10), 2941-2952. doi:10.1093/jxb/erw068

Pircalabioru, G. G., & Chifiriuc, M. C. (2020). Nanoparticulate drug-delivery systems for fighting microbial biofilms: from bench to bedside. *Future Microbiol, 15*, 679-698. doi:10.2217/fmb-2019-0251

Ra, J. S., Shin, H. H., Kang, S., & Do, Y. (2014). Lumazine synthase protein cage nanoparticles as antigen delivery nanoplatforms for dendritic cell-based vaccine development. *Clin Exp Vaccine Res, 3*(2), 227-234. doi:10.7774/cevr.2014.3.2.227

Rae, B. D., Long, B. M., Badger, M. R., & Price, G. D. (2013). Functions, compositions, and evolution of the two types of carboxysomes: polyhedral microcompartments that facilitate CO2 fixation in cyanobacteria and some proteobacteria. *Microbiol Mol Biol Rev, 77*(3), 357-379. doi:10.1128/MMBR.00061-12

Ralston, G. B. (1993). *Introduction to analytical ultracentrifugation* (Vol. 1). California: Beckman.

Roberts, E. W., Cai, F., Kerfeld, C. A., Cannon, G. C., & Heinhorst, S. (2012). Isolation and characterization of the *Prochlorococcus* carboxysome reveal the presence of the novel shell protein CsoS1D. *J Bacteriol, 194*(4), 787-795. doi:10.1128/JB.06444-11

Rolland, V., Badger, M. R., & Price, G. D. (2016). Redirecting the Cyanobacterial Bicarbonate Transporters BicA and SbtA to the Chloroplast Envelope: Soluble and Membrane Cargos Need Different Chloroplast Targeting Signals in Plants. *Front Plant Sci, 7*, 185. doi:10.3389/fpls.2016.00185

Sasaki, E., Bohringer, D., van de Waterbeemd, M., Leibundgut, M., Zschoche, R., et al. (2017). Structure and assembly of scalable porous protein cages. *Nat Commun, 8*, 14663. doi:10.1038/ncomms14663

Schmid, M. F., Paredes, A. M., Khant, H. A., Soyer, F., Aldrich, H. C., et al. (2006). Structure of *Halothiobacillus neapolitanus* carboxysomes by cryo-electron tomography. *J Mol Biol, 364*(3), 526-535. doi:10.1016/j.jmb.2006.09.024

Shang, T.-T., Liu, X.-Y., & Gu, L. (2016). Interface of transition metal oxides at the atomic scale. *Sci China Phy Mech, 59*(9), 697001. doi:10.1007/s11433-016-0122-x

Shen, C.-H. (2019). Chapter 7 - Detection and Analysis of Nucleic Acids. In C.-H. Shen (Ed.), *Diagnostic Molecular Biology* (pp. 167-185): Cambridge, Academic Press.

Shih, P. M., Occhialini, A., Cameron, J. C., Andralojc, P. J., Parry, M. A., et al. (2016). Biochemical characterization of predicted Precambrian RuBisCO. *Nat Commun, 7*, 10382. doi:10.1038/ncomms10382

Smolke, C. D. (2009). Building outside of the box: iGEM and the BioBricks Foundation. *Nat Biotechnol, 27*(12), 1099-1102. doi:10.1038/nbt1209-1099

So, A. K., Espie, G. S., Williams, E. B., Shively, J. M., Heinhorst, S., et al. (2004). A novel evolutionary lineage of carbonic anhydrase (epsilon class) is a component of the carboxysome shell. *J Bacteriol, 186*(3), 623-630. doi:10.1128/jb.186.3.623-630.2004

Some, D., Amartely, H., Tsadok, A., & Lebendiker, M. (2019). Characterization of Proteins by Size-Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS). *J Vis Exp* (148). doi:10.3791/59615

Sommer, M., Cai, F., Melnicki, M., & Kerfeld, C. A. (2017). beta-Carboxysome bioinformatics: identification and evolution of new bacterial microcompartment protein gene classes and core locus constraints. *J Exp Bot, 68*(14), 3841-3855. doi:10.1093/jxb/erx115

Song, Y., Kang, Y. J., Jung, H., Kim, H., Kang, S., et al. (2015). Lumazine Synthase Protein Nanoparticle-Gd(III)-DOTA Conjugate as a T1 contrast agent for high-field MRI. *Sci Rep, 5*, 15656. doi:10.1038/srep15656

Sun, Q., Tsai, S. L., & Chen, W. (2019). Artificial scaffolds for enhanced biocatalysis. *Methods Enzymol, 617*, 363-383. doi:10.1016/bs.mie.2018.12.007

Sun, Y., Wollman, A. J. M., Huang, F., Leake, M. C., & Liu, L. N. (2019). Single-Organelle Quantification Reveals Stoichiometric and Structural Variability of Carboxysomes Dependent on the Environment. *Plant Cell, 31*(7), 1648-1664. doi:10.1105/tpc.18.00787

Sutter, M., Faulkner, M., Aussignargues, C., Paasch, B. C., Barrett, S., et al. (2016). Visualization of Bacterial Microcompartment Facet Assembly Using High-Speed Atomic Force Microscopy. *Nano Lett, 16*(3), 1590-1595. doi:10.1021/acs.nanolett.5b04259

Sutter, M., Greber, B., Aussignargues, C., & Kerfeld, C. A. (2017). Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell. *Science, 356*(6344), 1293-1297. doi:10.1126/science.aan3289

Sutter, M., Laughlin, T. G., Sloan, N. B., Serwas, D., Davies, K. M., et al. (2019). Structure of a Synthetic beta-Carboxysome Shell. *Plant Physiol, 181*(3), 1050-1058. doi:10.1104/pp.19.00885

Sutter, M., Melnicki, M. R., Schulz, F., Woyke, T., & Kerfeld, C. A. (2021). A catalog of the diversity and ubiquity of bacterial microcompartments. *Nat Commun, 12*(1), 3809. doi:10.1038/s41467-021-24126-4

Tabita, F. R., Satagopan, S., Hanson, T. E., Kreel, N. E., & Scott, S. S. (2008). Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *J Exp Bot, 59*(7), 1515-1524. doi:10.1093/jxb/erm361

Tan, Y. Q., Ali, S., Xue, B., Teo, W. Z., Ling, L. H., et al. (2021). Structure of a Minimal alpha-Carboxysome-Derived Shell and Its Utility in Enzyme Stabilization. *Biomacromolecules*. doi:10.1021/acs.biomac.1c00533

Temme, K., Zhao, D., & Voigt, C. A. (2012). Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc Natl Acad Sci U S A, 109*(18), 7085-7090. doi:10.1073/pnas.1120788109

Terasaka, N., Azuma, Y., & Hilvert, D. (2018). Laboratory evolution of virus-like nucleocapsids from nonviral protein cages. *Proc Natl Acad Sci U S A, 115*(21), 5432-5437. doi:10.1073/pnas.1800527115

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res, 13*(9), 2129-2141. doi:10.1101/gr.772403

Thompson, D. B., Cronican, J. J., & Liu, D. R. (2012). Engineering and identifying supercharged proteins for macromolecule delivery into mammalian cells. *Methods Enzymol, 503*, 293-319. doi:10.1016/B978-0-12-396962-0.00012-4

Turmo, A., Gonzalez-Esquer, C. R., & Kerfeld, C. A. (2017). Carboxysomes: metabolic modules for CO2 fixation. *FEMS Microbiol Lett, 364*(18). doi:10.1093/femsle/fnx176

Urban, M. J., Holder, I. T., Schmid, M., Fernandez Espin, V., Garcia de la Torre, J., et al. (2016). Shape Analysis of DNA-Au Hybrid Particles by Analytical Ultracentrifugation. *ACS Nano, 10*(8), 7418-7427. doi:10.1021/acsnano.6b01377

Wang, S., Al-Soodani, A. T., Thomas, G. C., Buck-Koehntop, B. A., & Woycechowsky, K. J. (2018). A Protein-Capsid-Based System for Cell Delivery of Selenocysteine. *Bioconjug Chem, 29*(7), 2332-2342. doi:10.1021/acs.bioconjchem.8b00302

Whitney, S. M., Houtz, R. L., & Alonso, H. (2011). Advancing our understanding and capacity to engineer nature's CO2-sequestering enzyme, Rubisco. *Plant Physiol, 155*(1), 27-35. doi:10.1104/pp.110.164814

Worsdorfer, B., Woycechowsky, K. J., & Hilvert, D. (2011). Directed evolution of a protein container. *Science, 331*(6017), 589-592. doi:10.1126/science.1199081

Yin, K., Gao, C., & Qiu, J.-L. (2017). Progress and prospects in plant genome editing. *Nat Plants, 3*(8), 17107. doi:10.1038/nplants.2017.107

Zang, K., Wang, H., Hartl, F. U., & Hayer-Hartl, M. (2021). Scaffolding protein CcmM directs multiprotein phase separation in beta-carboxysome biogenesis. *Nat Struct Mol Biol, 28*(11), 909-922. doi:10.1038/s41594-021-00676-5

Zhang, J., Pearson, J. Z., Gorbet, G. E., Colfen, H., Germann, M. W., et al. (2017). Spectral and Hydrodynamic Analysis of West Nile Virus RNA-Protein Interactions by Multiwavelength Sedimentation Velocity in the Analytical Ultracentrifuge. *Anal Chem, 89*(1), 862-870. doi:10.1021/acs.analchem.6b03926

Zhang, X., Konarev, P. V., Petoukhov, M. V., Svergun, D. I., Xing, L., et al. (2006). Multiple assembly states of lumazine synthase: a model relating catalytic function and molecular assembly. *J Mol Biol, 362*(4), 753-770. doi:10.1016/j.jmb.2006.07.037

Zheng, W., Zhang, C., Li, Y., Pearce, R., Bell, E. W., et al. (2021). Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Rep Methods, 1*(3). doi:10.1016/j.crmeth.2021.100014