Integrated Experimental and Bioinformatics Approach to Investigate Metabolic Channelling

by

Kris Hon

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Graduate Department of Biochemistry
University of Toronto

© Copyright by Kris Hon 2019
Abstract

INTEGRATED EXPERIMENTAL AND BIOINFORMATICS APPROACH TO INVESTIGATE METABOLIC CHANNELLING

Doctor of Philosophy

2019

Kris Hon

Graduate Department of Biochemistry

University of Toronto

Metabolic channelling refers to the phenomenon where metabolic intermediates are directed between enzymatic active sites without release into the bulk cytosol. While disparate computational and experimental studies have initiated characterization of metabolic channelling, there is a lack of comprehensive studies to fully evaluate the extent of channelling in nature. In addition, experimental techniques to engineer metabolic channelling have yielded inconsistent results which highlight the need to better understand channelling. In this thesis, I leverage a combined bioinformatics and experimental approach in order to understand the extent of channelling in nature and explore the use of a compartment system to engineer channelling. A list of enzyme pairs predicted to benefit from metabolic channelling was generated by leveraging several publicly available data sources, and compared with protein interaction and gene fusion
datasets to validate predictions. The minimal overlap for channelling predictions led to additional exploration to better understand what factors are important in driving channelling. Based on phylogenetic analysis of gene fusions and visualizations of the prediction list in the context of the global metabolic network, fusions were typically found in non-essential metabolic pathways while enzymes in the prediction list were enriched at convergence points of multiple metabolic pathways. In conjunction with the bioinformatics approach, I explored the use of the bacteriophage HK97 capsid as a compartment system for the introduction of channelling between enzyme pairs of interest. I was able to target enzymes of interest into an assembled capsid within E. coli through fusion with a C-terminal targeting sequence. Encapsulated enzymes were found to be active, albeit at a reduced activity level in comparison to similar quantities of unencapsulated enzyme. Multiple enzymes could also successfully targeted into the capsid, with different enzymes being encapsulated at varying efficiencies. Charge appeared to be one of the factors that affected targeting levels, as altering the charge of the capsid interior was found to affect both targeting efficiency and enzymatic activity. An additional application identified for the HK97 capsid was to protect encapsulated enzymes, as a targeted glycoside hydrolase was successfully protected from precipitation at elevated temperatures in comparison to its unencapsulated form.
Acknowledgments

During the course of my graduate experience, I’ve been very fortunate to receive an enormous amount of support from a wide range of individuals which has enabled my growth both personally and professionally. All my interactions and learning have driven my development from an undergraduate student to an industry professional who now advises executives from leading Canadian and global entities. First and foremost, I want to thank my supervisor, John Parkinson. Your constant guidance and support was a key driver for my development, and your willingness to allow me to pursue opportunities outside of academia has led me to where I am today. I also want to thank my committee members, Alan Davidson and Shana Kelley, for their insightful discussions and recommendations during committee meetings and throughout my project, as well as Karen Maxwell, for providing lab space and resources to execute the experimental studies described within this thesis.

An important part of any graduate experience is the group of co-workers who are always available for stimulating scientific discussion and support across all aspects of life, many of whom have become lifelong friends. I want to thank Diane Bona for educating me on a number of microbiology and biochemical techniques and helping troubleshoot issues during my experiments. Her mentorship provided a solid foundation which allowed me to independently drive the research for this thesis. I also want to thank Senjuti Saha, Nichole Cumby, Yvonne Tsao, Smriti Kala, Lisa Pell, Kelly Reimer and Mostafa Fatehi, my day-to-day interactions with this group made the graduate experience easier and much more enjoyable. Following your continued success beyond the lab has been a pleasure. Not to be excluded are the many members of the Davidson/Maxwell and Parkinson labs for all of their helpful suggestions, advice, and assistance when preparing for meetings and talks throughout my time in the department.

Given the wide range of enzyme systems investigated in this thesis, I would not have been able to complete a number of experiments without the help of collaborators across various departments and disciplines. To this end, I would like to thank Cody Sarch for supplying arabinofuranosidase constructs and his assistance in performing the arabinofuranosidase activity assays, Olivia Molenda for supplying dehalogenase constructs and her assistance in performing the dehalogenase activity assays, Veronica Yim for supplying laccase constructs, and Dr. Robert
Flick for performing laccase activity assays. In addition, I would like to thank Dr. Xuejian Xiong for her assistance in conducting the statistical analyses highlighted throughout Chapter 2.

Last, but certainly not least, I would like to thank all my family and friends for their unwavering support throughout my graduate experience. They have consistently kept me motivated, despite the hurdles created by taking on a full-time position in industry during the thesis writing period. I would like to thank my friends Jeff Hong and Ted Livingston, who have constantly pushed the borders in their professional careers and their incredible success has served as my continued source of inspiration. I would also like to thank my partner, Sunny, for all of her support and encouragement during the defense process. Finally, I want to thank my parents, Kam and Rebecca, who have always supported my education and career path decisions throughout the manner years.
Table of Contents

List of Tables ............................................................................................................................................... ix
List of Figures ............................................................................................................................................... x
List of Abbreviations ................................................................................................................................... xi
Chapter 1: Introduction ................................................................................................................................. 1

1.1 Biological compartmentalization .................................................................................................................... 2
  1.1.1 Eukaryotic compartments .......................................................................................................................... 2
  1.1.2 Prokaryotic compartments .......................................................................................................................... 3

1.2 Genomic surveys reveal bacterial compartmentalization is a widespread phenomenon ...................... 5

1.3 Biological benefits of compartmentalization ................................................................................................. 6
  1.3.1 Increasing the kinetic efficiency of enzymes ............................................................................................. 7
  1.3.2 Isolating toxic/inhibitory intermediates from the cytosol ........................................................................ 9
  1.3.3 Stabilizing insoluble enzymes and protecting sensitive enzymes from harmful effects ...................... 10

1.4 Micro-compartments offer an effective route to drive channelling .............................................................. 12
  1.4.1 Engineered de novo lipid compartments .................................................................................................. 12
  1.4.2 Strategies for protein-based compartments ............................................................................................... 13
  1.4.3 Engineering de novo protein compartments ............................................................................................. 14
  1.4.4 Adopting existing protein compartment-like structures for use as a compartment system .......... 16
  1.4.5 Converting specialized micro-compartments into flexible encapsulation systems .............................. 17

1.5 Viruses as enzyme compartments .................................................................................................................. 19
  1.5.1 Previous work in the field ......................................................................................................................... 20
  1.5.2 Limitations of current protein compartment systems ............................................................................. 23
  1.5.3 HK97 capsid as a potential compartment ................................................................................................. 24
  1.5.4 Improvements and potential applications ............................................................................................... 26

1.6 Applications of engineered protein compartments for metabolic channelling ......................................... 26
1.6.1 Previous work on characterizing metabolic channelling ................................................................. 26
  1.6.1.1 Computational characterizations of metabolic channelling ......................................................... 28
  1.6.1.2 Experimental studies to introduce metabolic channelling into enzyme systems ........ 30

1.6.2 Limitations, challenges and opportunities in channelling .......................................................... 32

1.7 Project goals ........................................................................................................................................ 34

2.1 Introduction ........................................................................................................................................... 37

2.2 Methods ................................................................................................................................................. 38
  2.2.1 Sources of data ................................................................................................................................. 38
  2.2.2 Identification of enzyme pairs with the potential to benefit from channelling ......................... 39
  2.2.3 Metabolic network reconstruction ............................................................................................... 40
  2.2.4 Hyperscape Network Representations .......................................................................................... 40

2.3 Results ................................................................................................................................................... 41
  2.3.1 Surveys of thermodynamic, toxic and inhibitory datasets predict 9,161 unique reaction pairs that represent candidates for metabolic channelling ................................................................. 41
  2.3.2 Channelling predictors are supported by gene fusion evidence and multifunctional enzymes 48
  2.3.3 Examining channelling predictions in the context of the global metabolic network ............ 54
  2.3.4 Visualizing biochemical pathway complexity ............................................................................. 57

2.4 Discussion ......................................................................................................................................... 63
  2.4.1 Extent of metabolic channelling in nature ............................................................................... 65
  2.4.2 Role of channelling in regulating metabolic flux .................................................................. 68

Chapter 3: Exploring the Potential of the HK97 Capsid as a Compartment System ......................... 72

3.1 Introduction ........................................................................................................................................... 72

3.2 Materials and Methods ....................................................................................................................... 74
  3.2.1 Engineering the HK97 Prohead and targeted enzyme constructs ............................................ 74
  3.2.2 Protein expression ....................................................................................................................... 76
  3.2.3 Capsid purification from induced samples ............................................................................. 76
3.2.4 Purification of His-tagged enzymes

3.2.5 SDS-PAGE and Western blot

3.2.6 Gel densitometry and average protein copy number per capsid

3.2.7 Enzyme activity assays

3.2.7 Transmission Electron Microscopy to visualize capsids

3.3 Results

3.3.1 Assembled HK97 capsids can successfully encapsulate target proteins

3.3.2 Enzymes display activity when targeted inside of HK97 capsids

3.3.3 Insoluble proteins can be isolated within purified capsids

3.3.4 Altering capsid charge interior can affect activity and targeting levels of targeted enzymes

3.3.5 Encapsulating multiple enzymes into HK97 Proheads

3.3.6 Capsids can stabilize enzymes at higher temperatures

3.4 Discussion

Chapter 4 – Future Directions

4.1 Summary

4.2 Multiple capsid alterations to optimize multi-enzyme system encapsulation and activity for metabolic engineering

4.3 Isolation of insoluble proteins purified within HK97 capsids for biochemical studies

4.4 Identifying additional reaction pairings that can benefit from metabolic channelling for subsequent experimental testing and industrial application

4.5 Implications of Thesis Findings

5.0 References

Appendix A – Supplemental Figures

Appendix B – List of Supplemental Tables
List of Tables

Table 1. A summary of characterized compartments and their respective functions in prokaryotic organisms

Table 2. A summary of experiments adapting proteins and protein compartments as protein containers for biotechnological applications

Table 3. The ten highest scoring predicted channelling pairs based on free energy values derived from the Mavrovouniotis group-contribution method

Table 4. The ten highest scoring predicted channelling pairs based on free energy values derived from experimentally determined free energy values of compounds

Table 5. Select properties of enzymes in multi-enzyme targeting experiments

Table 6. Predictions and targeting results of various enzyme pair combinations associated with multi-enzyme targeting experiments
List of Figures

Figure 1. The potential roles of compartmentalization in metabolic pathways.................................................. 8

Figure 2. Protein fusions can generate compartment-like structures of desired size and shape. .............. 15

Figure 3. Selective targeting of proteins inside the CCMV capsid using an engineered coiled-coil targeting system. ......................................................................................................................................................... 21

Figure 4. The assembly pathway of the HK97 capsid. .................................................................................. 25

Figure 5. Localization of target proteins to a synthetic scaffold to increase kinetic efficiency for a reaction cascade .......................................................................................................................................................... 32

Figure 6. Application of the Mavrovouniotis group-contribution method to estimate Gibbs Free Energy value.................................................................................................................................................................................................. 43

Figure 7. Comparisons of enzyme pair channelling predictions with natural channelling examples potentially found in gene fusions and protein complexes. .......................................................... 51

Figure 8. Heatmap of gene fusion distributions across various taxonomic groups..................................... 52

Figure 9. Network representation of enzyme-substrate relationships. ....................................................... 56

Figure 10. Representations of channelling predictions and gene fusions in the pentose phosphate pathway illustrating the complexities in a section of the global metabolic network........................................... 61

Figure 11. Controlling metabolic flux through suppressed enzyme expression in a pathway. ...................... 67

Figure 12. Two potential methods of encapsulating target proteins into HK97 capsids............................ 81

Figure 13. GFP is successfully encapsulated into HK97 Prohead I capsids. .................................................. 83

Figure 14. Alkaline phosphatase is encapsulated and active inside of HK97 capsids................................. 86

Figure 15. Decreased PhoA activity in capsids is likely due to substrate diffusion limitation. .................... 88

Figure 16. Targeting of laccases into HK97 capsids. ...................................................................................... 90

Figure 17. Capsid mutations affect alkaline phosphatase activity but not targeting levels. ....................... 92

Figure 18. Targeting of β-gal and GalD into HK97 capsids......................................................................... 94

Figure 19. Targeting of multiple enzymes into HK97 capsids. ................................................................. 96

Figure 19. The activity of NHA GH62 samples with different temperature treatments............................ 98

Figure 20. The interaction of the maturation protease with the capsid shell of an HK97 capsid. ........... 101
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>°C</td>
<td>degrees Celsius</td>
</tr>
<tr>
<td>Å</td>
<td>angstroms</td>
</tr>
<tr>
<td>ΔG</td>
<td>Gibbs free energy change</td>
</tr>
<tr>
<td>β-gal</td>
<td>beta-galactosidase</td>
</tr>
<tr>
<td>AChE</td>
<td>acetylcholinesterase</td>
</tr>
<tr>
<td>AdhD</td>
<td>alcohol dehydrogenase</td>
</tr>
<tr>
<td>AtoB</td>
<td>acetoacetyl-CoA thiolase</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine Tri-phosphate</td>
</tr>
<tr>
<td>BLAST</td>
<td>basic local alignment search tool</td>
</tr>
<tr>
<td>BMC</td>
<td>bacterial micro-compartment</td>
</tr>
<tr>
<td>BRENDA</td>
<td>Braunschweig Enzyme Database</td>
</tr>
<tr>
<td>CCMV</td>
<td>cowpea chlorotic mottle virus</td>
</tr>
<tr>
<td>CO₂</td>
<td>carbon dioxide</td>
</tr>
<tr>
<td>CoA</td>
<td>co-enzyme A</td>
</tr>
<tr>
<td>DMSO</td>
<td>dimethyl sulfoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>E. coli</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>EC</td>
<td>enzyme classification</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>EGFP</td>
<td>enhanced green fluorescent protein</td>
</tr>
<tr>
<td>Etu</td>
<td>ethanol utilization compartment</td>
</tr>
<tr>
<td>Eut</td>
<td>ethanolamine-utilization compartment</td>
</tr>
<tr>
<td>GalA</td>
<td>alpha-galactosidase</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>GalD</td>
<td>galactose dehydrogenase</td>
</tr>
<tr>
<td>GalK</td>
<td>galactokinase</td>
</tr>
<tr>
<td>GBD</td>
<td>GTPase protein binding domain</td>
</tr>
<tr>
<td>GDP</td>
<td>guanosine diphosphate</td>
</tr>
<tr>
<td>GFP</td>
<td>green fluorescent protein</td>
</tr>
<tr>
<td>GH62</td>
<td>glycoside hydrolase family 62</td>
</tr>
<tr>
<td>GluK</td>
<td>glucokinase</td>
</tr>
<tr>
<td>gp5</td>
<td>HK97 major capsid protein</td>
</tr>
<tr>
<td>GTP</td>
<td>guanosine triphosphate</td>
</tr>
<tr>
<td>HIV</td>
<td>human immunodeficiency virus</td>
</tr>
<tr>
<td>HK-97</td>
<td>bacteriophage Hong-Kong 97</td>
</tr>
<tr>
<td>HMGR</td>
<td>hydroxyl-methylglutaryl-CoA reductase</td>
</tr>
<tr>
<td>HMGS</td>
<td>hydroxyl-methylglutaryl-CoA synthase</td>
</tr>
<tr>
<td>HRP</td>
<td>horseradish peroxidase</td>
</tr>
<tr>
<td>k&lt;sub&gt;cat&lt;/sub&gt;</td>
<td>turnover number</td>
</tr>
<tr>
<td>KCl</td>
<td>potassium chloride</td>
</tr>
<tr>
<td>kDa</td>
<td>kilodaltons</td>
</tr>
<tr>
<td>KDPGal</td>
<td>2-keto-3-deoxy-6-phosphogalactonate</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>K&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Michaelis-Menten constant</td>
</tr>
<tr>
<td>LB</td>
<td>lysogeny broth</td>
</tr>
<tr>
<td>LC50</td>
<td>lethal concentration required to kill 50% of a population</td>
</tr>
<tr>
<td>LPS</td>
<td>lipopolysaccharide</td>
</tr>
<tr>
<td>MGC</td>
<td>Mavrovouniotis group-contribution</td>
</tr>
<tr>
<td>MgCl&lt;sub&gt;2&lt;/sub&gt;</td>
<td>magnesium chloride</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MHT</td>
<td>methyl halide transferase</td>
</tr>
<tr>
<td>MOPS</td>
<td>3-morpholinopropane-1-sulfonic acid</td>
</tr>
<tr>
<td>MTAN</td>
<td>methylthioadenosine nucleosidase</td>
</tr>
<tr>
<td>MTR</td>
<td>5-methylthioribose</td>
</tr>
<tr>
<td>NaCl</td>
<td>sodium chloride</td>
</tr>
<tr>
<td>NADPH</td>
<td>reduced nicotinamide adenine dinucleotide phosphate</td>
</tr>
<tr>
<td>NaOH</td>
<td>sodium hydroxide</td>
</tr>
<tr>
<td>NHA</td>
<td>arabinofuranosidase from <em>Nectria haematococca</em></td>
</tr>
<tr>
<td>Ni-NTA</td>
<td>nickel - nitrilotriacetic acid</td>
</tr>
<tr>
<td>NIST</td>
<td>National Institute for Standards and Technology</td>
</tr>
<tr>
<td>nm</td>
<td>nanometers</td>
</tr>
<tr>
<td>OAS</td>
<td>O-Acetyl-L-serine</td>
</tr>
<tr>
<td>OD&lt;sub&gt;600&lt;/sub&gt;</td>
<td>optical density measured at 600 nm</td>
</tr>
<tr>
<td>PalB</td>
<td><em>Pseudozyma antarica</em> lipase B</td>
</tr>
<tr>
<td>Pdu</td>
<td>propanediol-utilization compartment</td>
</tr>
<tr>
<td>PEG</td>
<td>polyethylene glycol</td>
</tr>
<tr>
<td>Pfam</td>
<td>protein family</td>
</tr>
<tr>
<td>pH</td>
<td>power of hydrogen</td>
</tr>
<tr>
<td>phoA</td>
<td>alkaline phosphatase A</td>
</tr>
<tr>
<td>pI</td>
<td>isoelectric point</td>
</tr>
<tr>
<td>PPI</td>
<td>protein-protein interaction</td>
</tr>
<tr>
<td>PRPP</td>
<td>phosphoribosyl pyrophosphate</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>rpm</td>
<td>rotations per minute</td>
</tr>
<tr>
<td>RuBisCO</td>
<td>ribulose-1,5-bisphosphate carboxylase/oxygenase</td>
</tr>
</tbody>
</table>
**S. cerevisiae**  \( \text{Saccharomyces cerevisiae} \)

**SAM**  S-adenosyl methionine

**SDS-PAGE**  sodium dodecyl sulfate polyacrylamide gel electrophoresis

**SH3**  SRC Homology 3 Domain

**TEM**  transmission electron microscopy

**THF**  tetrahydrofolate

**Tris**  tris(hydroxymethyl)aminomethane

**Tris-HCl**  Tris-hydrochloride

**μg**  microgram

**μL**  microlitre

**V_{max}**  maximum reaction velocity

**w/v**  weight/volume

**WT**  wild type

**Xgal**  5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
Chapter 1: Introduction

Metabolic channelling is the phenomenon where metabolic intermediates are passed between the active sites of biochemically sequential enzymes without release into the bulk cytosol [1]. Channelling fulfills various biological roles, including increasing reaction efficiency through enzyme systems, and sequestering toxic or inhibitory enzymes that can cause damage to host organisms [1]. In nature, more elegant examples of directed channelling include the generation of gene fusions [2] and the formation of protein complexes [1]. Another alternative to introduce channelling involves compartmentalization, where enzymes and substrates are encapsulated in isolated environments [1]. There has been strong interest in studying compartments to understand how they can be utilized in engineered enzyme systems. Studies have previously focused on eukaryotic compartments such as the mitochondrion, which localizes enzymes from the citric acid cycle to drive the production of ATP [3], and chloroplasts which play a critical role in carbon fixation to build complex macromolecules for eukaryotic organisms [4]. Recent studies in Clostridium difficile and Salmonella enterica have demonstrated how compartmentalization of pathway segments in prokaryotic organisms allows the utilization of unique compounds as carbon sources or storing important metabolites [5, 6]. The compartments that house the relevant enzymes in a pathway are typically named after the metabolite that it processes (i.e., propanediol-utilization, ethanolamine-utilization) [5, 6]. The ability to reconstitute such compartments in non-native hosts [7-9] has garnered heavy interest for manufacturing compounds of interest through metabolic channelling or stabilizing enzymes for bioremediation purposes. However, there is a lack of understanding of how these compartments can be manipulated to fine-tune enzyme systems to behave in pre-defined manners, as well as what enzymes demonstrate a benefit from encapsulation. This lack of understanding will limit the current usefulness of compartments for industrial applications. In addition, no extensive survey of metabolic channelling has been performed, and the factors that determine whether enzymes can benefit from channelling are not completely understood. In this thesis, I characterize the use of the bacteriophage HK97 capsid as a compartment system for implementation in industrial applications as it is a well-characterized compartment-like structure that is structurally robust [10, 11]. Furthermore, I perform a survey of the extent of channelling
in nature to understand its prevalence and predict enzyme pairings that could benefit from channelling.

1.1 Biological compartmentalization

Biological compartmentalization is the method by which biological molecules are localized within enclosed structures [12]. The physical separation of biological reactions provides the vital spatial organization necessary to sustain life and allow metabolism to function smoothly [12]. Since the coining of the term, “organula”, by Karl August Möbius in the 1880s when describing the cellular structures of unicellular organisms [13], extensive studies of various biological compartments have revealed the wide variety of roles they play in living organisms. Some of these roles include increasing efficiency of biochemical processes by providing an optimal environment [14] and protecting encapsulated contents from inhibitors or proteases [14]. With increased interest in the roles of eukaryotic compartments [15], and serendipitous discoveries of compartments in prokaryotic organisms (to be discussed in following sections), it is evident that compartmentalization is ubiquitous throughout nature.

1.1.1 Eukaryotic compartments

The discovery and characterization of organelles across diverse eukaryotic organisms revealed common roles that compartments play across organisms. One identified role involves the maintenance of a localized environment inside the container to enhance the function of its encapsulated components. For example, the lysosome maintains a pH 4.5 – 5.0 environment within the pH 7.2 cytosol in order to optimize the activity of its enclosed acid phosphatases [16]. Another role is to provide an enclosed area to protect its contents from components in other compartments. The endoplasmic reticulum protects polypeptide chains from cytosolic degradation by proteases and allows for the correct folding of proteins in cisternae sacs before subsequent transport to the Golgi apparatus [17]. The nucleus isolates genomic DNA from nucleases and other DNA altering agents (e.g., reactive oxygen species, free radicals) that may be present in the cytosol [18, 19]. A third role identified includes localizing enzymes and metabolic intermediates to increase the efficiency of biochemical reactions. In the mitochondrial cristae, enzymes in the electron transport chain are localized in the membrane to efficiently
channel electrons between various carriers for ATP synthesis. This phenomenon is also seen in plant chloroplasts; light-dependent reactions which generate ATP and NADPH occur at the thylakoid sacs, while the carbon fixation reactions occur in the chloroplast lumen [20]. Compartmentalization plays a crucial role in many biochemical processes in eukaryotic organisms, with lipid compartments representing a key differentiator between eukaryotes and prokaryotes. However, the concept of compartmentalization is not unique to eukaryotes. The next section highlights how prokaryotes have evolved unique solutions to achieve compartmentalization.

1.1.2 Prokaryotic compartments

Widespread compartmentalization was commonly thought to be absent in prokaryotes. However, recent investigations have not only revealed that presence of compartments across a number of prokaryotic organisms, but also demonstrated that the compartments aid in biochemical processes unique to prokaryotes. Polygonal inclusion bodies were first reported in 1956 after their discovery in electron micrograph images of the cyanobacterium Phormidium uncinatum carboxysome [21]. The polyhedral bodies were first purified in 1973 from Halothiobacillus neapolitanus and found to contain the enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) within a proteinaceous shell later known as the carboxysome [22]. Carboxysomes were initially thought to act as either storage bodies [23] or micro-compartments/simple organelles that aided in the fixation of CO2 [24]. Subsequent studies revealed that RuBisCO was active in intact, purified carboxysomes and had similar kinetic and inhibitor-binding properties to cytoplasmic RuBisCO [25]. RuBisCO production and the number of carboxysomes were increased in CO2-limiting conditions, with most RuBisCO found to be localized inside of carboxysomes [25]. Additional evidence showed that carboxysome mutations preventing proper assembly, and ectopic expression of carbonic anhydrase in the cyanobacterial cytosol, both prevented CO2 fixation from occurring [25]. Taken together, these findings show that carboxysomes function as carbon-fixation organelles by isolating RuBisCO away from inhibitors and competing enzymes to enhance carbon fixation efficiency.

Similar bacterial micro-compartments (BMCs) have been identified in other organisms that allow the metabolism of compounds that generate toxic intermediates (Figure 1). The
propanediol-utilization (Pdu) compartment, first found in *Salmonella enterica* in 1994 [26], was the first example of a heterotrophic organism containing a micro-compartment shell protein gene. It processes the metabolite 1,2-propanediol, which is a major product formed in the anaerobic degradation of rhamnose and fucose moieties in plant cell walls, bacterial exopolysaccharides and glycoconjugates from intestinal epithelial cells [5]. A toxic intermediate, propionaldehyde, is formed during the conversion of 1,2 propanediol to the product propionyl-CoA, which diffuses out of the micro-compartment for use in the citric acid cycle. Deletion of genes required for Pdu compartment formation leads to increased propionaldehyde levels, followed by *S. enterica* growth arrest and increased DNA mutagenesis rate [5]. These findings revealed that compartments are used to sequester toxic intermediates and protect the cell from their damaging effects. The ethanolamine utilization compartment, identified in *Clostridium difficile*, was found to perform a similar function by sequestering the toxic and volatile acetaldehyde during the conversion of ethanolamine to acetyl-CoA [27]. Finally, an ethanol utilization (Etu) compartment was recently discovered in *Clostridium kluyveri* that allowed bacterial growth using ethanol as a carbon source [28]. Structural studies revealed a unique pore structure which only enabled ethanol entry into the compartment [28]. Pores refer to openings on compartment shells that permit diffusion of substrates into and out of the compartment [29]. Different compartments have been shown to have customized pore structures to allow diffusion of specific compounds into and out of the compartment lumen to optimize the activity of encapsulated enzymes [30]. While new micro-compartments are continually being discovered in prokaryotes (Table 1), there are likely many compartments yet to be discovered that can reveal novel roles and contribute to the understanding of how prokaryotes utilize compartmentalization.
Table 1. A summary of characterized compartments and their respective functions in prokaryotic organisms.

<table>
<thead>
<tr>
<th>Compartment</th>
<th>Function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carboxysome</td>
<td>Increases carbon fixation efficiency by localizing RuBisCO</td>
<td>Codd, G.A., 1998 [23]</td>
</tr>
<tr>
<td>Propanediol utilization (Pdu)</td>
<td>Converts 1,2 propanediol to acetyl CoA while isolating the toxic intermediate propionaldehyde</td>
<td>Bobik et al., 1999 [31]</td>
</tr>
<tr>
<td>Ethanolamine utilization</td>
<td>Converts ethanolamine to acetyl-phosphate while isolating the volatile intermediate acetaldehyde</td>
<td>Pitts et al., 2012 [27]</td>
</tr>
<tr>
<td>Ethanol utilization</td>
<td>Converts ethanol to acetyl-CoA</td>
<td>Heldt et al., 2009 [28]</td>
</tr>
<tr>
<td>Glycyl-radical enzyme</td>
<td>Involved in the fermentation of fucose and rhamnose to propionaldehyde, which undergoes further processing similar to Pdu compartments</td>
<td>Axen et al., 2014 [32]</td>
</tr>
<tr>
<td>Bacterial micro-compartment of unknown function</td>
<td>Associated with amidohydrolases and deaminases that indicate its potential involvement in the metabolism of nitrogenous compounds</td>
<td>Axen et al., 2014 [32]</td>
</tr>
</tbody>
</table>

1.2 Genomic surveys reveal bacterial compartmentalization is a widespread phenomenon

With the discovery of bacterial compartments, several groups have performed computational studies in order to identify additional compartments and their associated enzymes. These studies also serve to identify the biological principles behind encapsulation for utilization in biotechnological applications. An early study searching for carboxysome shell gene homologs in various bacteria found seven functionally distinct families of bacterial micro-compartment loci distributed over 40 different genera of bacteria [33]. Initial functional characterization of computationally predicted micro-compartments arose from genomic analysis of *Clostridium phytofermentans* and bioinformatics analysis of *Planctomycetes* and *Verrucomicrobia* phyla genomes [34] [35]. In both cases, micro-compartments were identified that metabolize fucose or rhamnose to propionate and propanol, generating a toxic compound (lactaldehyde) in the process. These compartments differ from each other by either degrading fucose or rhamnose with
an aldolase [34] or utilizing a glycyl radical enzyme to degrade a propanediol intermediate generated from separate pathways metabolizing fucose or rhamnose [35].

Subsequent computational surveys evaluated the co-occurrence of protein functional groups in BMC loci across 113 sequenced bacterial genomes [36]. Although experimentally determined loci were regularly identified, loci with novel gene structures were not detected. A more recent study utilizing a similar gene-centric approach identified 358 bacterial species containing BMCs and grouped loci based on protein domain families (Pfams) [37]. However, this study failed to predict new signature enzymes or define novel BMC locus types. To find novel components, a new algorithm was developed which incorporated previously characterized features of compartments and introduced flexible criteria to enable the identification of compartments with a subset of features thought to be required for a fully functional compartment [32]. Implementation of the algorithm not only identified previously characterized compartments such as the Pdu and Eut compartments, but also found new compartment types including the glycyl-radical enzyme compartment, hybrids of compartments with multiple functions and a large class of putative compartments with a potential role in processing nitrogen-containing compounds [32]. Subsequent experimental work verified the existence of the glycyl-radical enzyme compartments [38, 39], with the identification of iron-sulfur binding sites on the compartment shell believed to be involved in generating glycyl radicals in the compartment [39]. Glycyl-radical enzymes in these compartments form complexes with an accompanying activating enzyme to facilitate degradation of complex compounds such as choline, propanediol and fuculose phosphate [39, 40].

1.3 Biological benefits of compartmentalization

As highlighted in the previous section, biological compartments serve a number of roles in living organisms. Given their functional flexibility, compartments represent an attractive option for metabolic engineers to construct a system that elicit the desired effect. Although these compartments are designed for very specific functions, the general principles as to how these roles are achieved can theoretically be applied to synthetic biology or metabolic engineering projects. The biological principles behind these roles and how they can be applied are discussed in the following sub-sections.
1.3.1 Increasing the kinetic efficiency of enzymes

Encapsulation of enzymes in compartments can increase the kinetic efficiency through reducing free diffusion and ensuring that concentrations drive direction of reaction (Figure 1). Confinement within an enclosed compartment results in an increase in local concentration of enzymes and substrates within a confined compartment, helping drive the aforementioned outcomes. Given the volume of a compartment would be significantly lower than the total volume of the cytosol, the concentration of a single enzyme and a single substrate is much higher inside of a compartment than it would be in the cytosol. This effective increase in concentration shifts the reaction equilibrium towards the formation of the product, assuming that products do not themselves accumulate within in the compartment. The second reason as to why compartmentalization can increase kinetic efficiency, primarily in multi-enzyme systems, is the isolation of enzymes and substrates away from competing enzymes. In the cytosol, metabolic intermediates can potentially be diverted away from the desired reaction path by competing enzymes in the vicinity. As one progresses along a biochemical pathway, the potential pool of metabolic intermediates within the pathway decreases depending on the number of enzymes in alternate pathways that utilize the intermediates within the pathway of interest. By confining enzymes within a restricted space, the intermediate is more likely to be processed by the co-localized enzyme. This can be critical for metabolic engineering applications where such co-localization can promote the desired progression of reactions, leading to an overall increase kinetic efficiency for the same quantity of substrate.

This method is a very attractive option in metabolic engineering as it can effectively increase the yield of a product given a certain amount of time or starting material. In the industrial and pharmaceutical industries, finding an enzyme or enzyme system that can produce a compound of interest yields significant cost savings over conventional means such as chemical synthesis and purification from natural sources. Adding an extra step of compartmentalization can further enhance the efficiency of the process and provide greater cost savings. The production of biofuels utilizing enzymes from varying organisms that convert compounds to the desired biofuel is a significant area of interest within this category. For example, *E. coli* was successfully engineered to synthesize 3-methyl-1-butanol by adding genes from *Bacillus subtilis, Lactococcus lactis* and *Saccharyomyces cerevisiae* [41]. Deletion of competing
pathways and removal of feedback inhibition loops resulted in increased production from 0.056 g/L to 0.12 g/L of 3-methyl-1-butanol [41]. Additional rounds of random mutagenesis further increased production nearly 40-fold to 4.4 g/L [42].

![Diagram of metabolic pathways](image)

**Figure 1. The potential roles of compartmentalization in metabolic pathways.**

In a hypothetical pathway, metabolite 1 is converted to metabolite 2 (a toxic/inhibitory compound) by enzyme A. Metabolite 2 can be converted to metabolite 3 by enzyme B or metabolite 4 by enzyme C. Compartmentalization can increase the production of metabolite 3 by co-localizing both reactions to increase local concentrations of substrate and enzyme (top row) or by preventing access by enzyme C to metabolite 2 (middle row). In addition, compartmentalization can reduce the levels of metabolite 2 as it can be quickly converted by enzyme B inside of a compartment. This acts to protect the host from the toxic effects of metabolite 2 by preventing accumulation in the cell.

Although effective, the process of generating multiple mutations in competing pathways can be challenging. Deleting genes also carries the risk of eliciting un-intended effects that can
affect host fitness and decrease its survivability. Encapsulation offers an attractive alternative by co-localizing the necessary enzymes required to produce the correct metabolic intermediates into a compartment and directing synthesis along the correct pathway. This can limit diversion of metabolic intermediates to competing pathways and increase synthetic efficiency by reducing the amount of initial substrate needed to produce a specified amount of the desired compound.

1.3.2 Isolating toxic/inhibitory intermediates from the cytosol

In nature, there are many reactions that produce compounds that can be toxic or inhibit important processes in either native host or heterologous host systems. One approach for engineering strains to produce valuable but toxic metabolites is through increasing host tolerance to the metabolite. This is currently achieved through mutagenesis. For example, directed evolution of an *E. coli* strain was performed to maximize tolerance of isobutanol in an effort to increase production [43]. Identification of the mutations in the resulting strain and replication of the mutations in the parent strain successfully increased the strain’s tolerance to isobutanol but indirectly decreased the strains tolerance to hexane and chloramphenicol [43]. Compartmentalization offers an attractive alternative solution to overcome unexpected side effects, such as decreased tolerance to other metabolites. Instead of mutagenizing the strain in the earlier example, compartmentalizing the enzymes important in isobutanol synthesis could act to localize some of the isobutanol and prevent its intercalation in the membrane lipid bilayer, and subsequently alleviate the inhibition of growth [44]. This concept can be applied to other systems that produce valuable but toxic metabolites. By compartmentalizing the enzyme or enzymes that generate toxic or inhibitory molecules with a biochemically sequential enzyme that converts the intermediate to a less harmful compound, potential effects of the harmful molecule can be alleviated as it is quickly converted into a non-harmful form. This technique can be beneficial for engineering various enzyme systems where heterologously expressed enzymes generate compounds that are toxic to the host organism. It can also increase the range of enzymes that can be included into model expression systems such as *E. coli* and *S. cerevisiae* for industrial or pharmaceutical purposes, providing cost savings advantages as highlighted in an earlier section.
1.3.3 Stabilizing insoluble enzymes and protecting sensitive enzymes from harmful effects

In the cytosol, enzymes are constantly exposed to cellular compounds and components that can decrease their kinetic efficiency or reduce their half-lives in the cell through degradation. Compartmentalization can help protect enzymes from such outcomes as compartments are structurally stable and can act as barriers against harmful compounds or degradative enzymes. The exclusion of unwanted entities will allow encapsulated enzymes to operate efficiently for an extended period of time, as compared to its unencapsulated form. The ability to protect encapsulated enzymes could be utilized in several applications. One application of compartments is to stabilize heterologously expressed enzymes within the compartment lumen by preventing aggregation and degradation of target enzymes. This can enable downstream characterization of the structure and function of encapsulated enzymes which were previously not possible for the target enzyme. Previous approaches to stabilize enzymes for further biochemical study include immobilizing single and multi-enzyme systems on media such as agarose, cellulose and nucleic acids. These approaches increased enzyme stability and activity as compared to the freely diffusing forms of the enzymes [45].

Compartmentalization can offer an easier alternative to stabilizing enzymes than the above approaches, with an additional benefit of being easily purified in an intact manner. This attribute can be useful for isolating insoluble enzymes in an active form inside of the compartment. Recently, a study utilizing the capsid from bacteriophage P22 was used to encapsulate a partially soluble α-galactosidase [46]. Capsids were successfully isolated with the enzyme inside of the capsid in an active form, showing that enzymes can be targeted away from insoluble aggregates into protein containers [46]. After isolation of the compartments, the encapsulated enzyme could be recovered for subsequent biochemical analyses by incubation in buffers that disrupt interactions of compartment structural components, freeing the encapsulated enzyme without denaturing the enzymes themselves. Chaperones are naturally occurring examples of proteins that act in this manner. They form compartments and help repair misfolded proteins by providing an environment to unfold and refold in the proper configuration. In the well-documented GroEL/GroES chaperone system present within *E. coli*, polypeptide substrates bind to hydrophobic regions within the GroEL chamber before subsequent binding of GroES and
ATP [47]. ATP is then hydrolyzed to help promote correct folding of the substrate before the binding of new ATP releases GroES and the folded product polypeptide [47]. To my knowledge, there have been no studies that target insoluble and inactive proteins into protein containers and successfully recover activity.

Beyond immobilization and naturally occurring chaperones, enzymes have been shown to be successfully stabilized in other systems. Silica gels were designed with pores to localize various proteins (e.g., lysozyme, α-lactalbumin, metmyoglobin, alkaline phosphatase) within the pores [48, 49]. At increased temperatures that would normally denature the aforementioned proteins, analyses using circular dichroism spectroscopy revealed that all proteins were more stable when encapsulated in the pores as compared to its free solution form. The temperature required to unfold α-lactalbumin was increased by approximately 25°C to 30°C upon silica gel encapsulation. Proteins unfold in a number of stages with varying levels of activation energy that contribute to overall stability [50]. It was hypothesized that confinement in a tight space acted to increase the activation energy during the unfolding process, leading to its enhanced stability. Similar stability results were also observed with varying salt concentrations and pH. In this scenario, the confined spaces within the silica gel pores prevented the localization of charged particles in the vicinity of the encapsulated enzyme. By shielding these enzymes, the effects of salt and pH were effectively neutralized as compared to the enzymes in free solution. Additional studies have been performed to see whether similar effects can be achieved using protein containers. In one study, *Escherichia coli* hydrogenase I, which catalyzes the oxidase of molecular hydrogen to form hydrogen gas, was encapsulated inside P22 capsids. Encapsulation resulted in a small increase in relative hydrogenase activity upon incubation at higher temperature, showing that protein compartments may offer an alternative strategy to stabilize enzymes [51].

In summary, compartmentalization has been shown to provide a number of benefits in biochemical systems including increasing kinetic efficiency, isolating toxic/inhibitory intermediates and stabilizing/protecting enzymes. In the next section I describe various approaches by which compartmentalization may be achieved to introduce channelling between enzymes of interest.
1.4 Micro-compartments offer an effective route to drive channelling

Lipid-based and protein-based structures that resemble containers have been common targets for studies, given their potential for a variety of applications, including compound storage. Subsequent sub-sections below highlight studies to date that have characterized potential compartment systems and generated insights that can inform the design of future compartment structures. With increased focus on discovering new compartments and manipulating previously characterized compartments for new applications, the knowledge gained from the highlights can be used to express the necessary components in heterologous organisms and encapsulate specific target enzymes to facilitate channelling.

1.4.1 Engineered *de novo* lipid compartments

Several studies have emulated the use of lipid-based compartments in eukaryotes by generating lipids and oil emulsions to form *in vitro* compartments. One example are liposomes, which have been investigated for use as compartments in various applications. They are generated by the hydration of lipid films in a controlled manner and supplemented with the addition of polymers to provide added stability. In one study, *Drosophila melanogaster* acetylcholinesterase (AChE) was encapsulated into vesicles and shown to be encapsulated by measuring AChE activity before encapsulation and after encapsulation following the addition of pronase [52]. Pronase inactivates free and externally bound AChE but does not pass through the liposomal bilayer to inactive encapsulated AChE [52]. Using the initial system, adjustments to the lipid composition, the buffer ionic strength and the addition of a His-tag were performed to determine important factors in enzyme encapsulation in liposomes [52]. Negatively charged lipids were found to decrease AChE encapsulation due to repulsion with the negative charge of AChE [52]. The addition of a His-tag to AChE enhanced encapsulation within liposomes conjugated to nickel-NTA moieties [52]. Increased salt concentrations negatively affected AChE encapsulation efficiency as charges of the Ni-NTA moiety and AChE were masked, preventing the enzyme-liposome interactions necessary for encapsulation [52].

In a separate study, cell-sized lipid vesicles up to 5 μM in diameter were synthesized and loaded with T7 promoter-containing plasmids, *E. coli* extracts and T7 RNA polymerase to allow for simultaneous transcription and translation of GFP inside each vesicle [53]. Fluorescence
microscopy confirmed the production of GFP inside of the vesicle, with expression efficiency determined to be significantly higher within the vesicle than in the solution outside [53]. The vesicle membrane was also found to protect the encapsulated GFP, as addition of proteinase K reduced fluorescence intensity outside of vesicles but did not have an effect on fluorescence inside of lipid compartments [53].

Interestingly, natural membrane-bound compartments have also been co-opted as compartments. For example, methyl halide transferase (MHT) was localized within yeast vacuoles using a 16 residue N-terminal tag found on carboxypeptidase Y, which is naturally targeted into vacuoles [54]. Due to the increased concentration of S-adenosyl methionine (SAM), a MHT substrate, when localized inside vacuoles as compared to other locations in the yeast cell, it was hypothesized MHT activity would be increased upon encapsulation [54]. Methyl iodide production was increased upon MHT encapsulation in vacuoles by nearly 15-fold, with the effect abolished in vacuole-deficient strains of yeast [54]. Given the importance of methyl halides in the production of agricultural and industrial precursors, the study presented a unique option in co-opting organelles to aid in the production of chemicals that overcome current energy intensive while increasing the range of initial substrates from which methyl halides can be created [54]. Although lipid-based compartments have stably encapsulated enzymes of interest, the difficulty in generating lipid-based compartments and the lack of demonstrated success in in vivo prokaryotic systems are major drawbacks in adopting lipid-based compartments for metabolic channelling purposes. As a result, interest has shifted from lipid-based compartments to protein-based systems that will be described in the following sections.

1.4.2 Strategies for protein-based compartments

A review of existing literature has identified three classes of strategies to create customized protein compartments. The first strategy involves adapting proteins to interact in a specific manner to create cage-like structures. Existing residues and domains important for interactions are either mutated or removed and replaced by new residues / interactions motifs in specific regions of proteins. This allows the proteins to interact in specific geometries that promote the formation of protein cages, which are cage-like structures of varying geometries that can be assembled using one of more proteins. The second strategy is utilizing existing proteins
that naturally self-associate to form cages and adapting them to house heterologously-expressed enzymes. These entities differ from naturally existing micro-compartments as proteins are not naturally targeted inside the self-assembling cages and natural mechanisms are not available to target proteins into these protein cages. As a result, new *in vitro* or *in vivo* targeting mechanisms will need to be developed to specifically encapsulate desired target enzymes. The third, and more recent, strategy is to adopt micro-compartments for use as general compartments. These three strategies will be discussed in more detail in the subsequent sections.

### 1.4.3 Engineering *de novo* protein compartments

A unique yet challenging strategy to engineer compartmentalization is by designing peptides as large building blocks that are capable of self-assembling into the desired size and shape. The process of computer-driven structure design is complex; nonetheless, in one recent study, the authors showed that it was possible to adopt such an approach to design a protein-based structure of specified dimensions [55]. In their study, they exploited the propensity for low-energy protein-protein interfaces to drive interactions by developing a customized computational program that creates compartments with a desired size, geometry and subunits count [55]. Based on designs generated by the program, a 24-subunit, 13-nm diameter complex with octahedral symmetry and a 12-subunit, 11-nm diameter complex with tetrahedral symmetry were successfully engineered as determined by X-ray crystallography verification [55]. Similarly, a fusion protein engineered with the trimeric *E. coli* 2-keto-3-deoxy-6-phosphogalactonate (KDPGal) aldolase, a four residue helical linker and the dimeric N-terminal domain of the *E. coli* FkpA protein successfully acted as building blocks for a cage structure [56]. Upon assembly, the 24-subunit cage had a central cavity with 10 nm pores (Figure 2), a 13 nm inner diameter and 22.5 nm outer diameter [56]. Interestingly, other unintended structures involving 12 or 18 subunits were also formed during assembly. The presence of alternate structures highlights a key challenge in designing oligomer fusions to create of protein cages as exact geometries cannot be precisely controlled and undesired assemblies can arise [56]. Nonetheless, the 24-subunit cage represented the largest assembly created from engineered protein subunits at the time of publication and provided a basis to form a more homogeneous sample of defined subunit number.
Another study created amphiphilic building blocks as a way of creating organelle-like complexes \textit{in vivo}. This design concept leverages the principle where hydrophobic surfaces interact in a manner that isolates hydrophobic surfaces from aqueous solution. By combining a hydrophobic motif and hydrophilic motif in one peptide, multiple copies of the peptide can assemble into a higher order structure driven by the interaction of hydrophobic motifs. Twenty copies of a VPGE\textsubscript{F}G motif, found to have elastin-like properties that allow it to rearrange in a variety of geometries, acted as the hydrophilic domain for future sub-functionalization and solubilisation of the lipid compartment in solutions [57]. A similar motif (VPGE\textsubscript{F}G) with a substitution of the polar residue, glutamine, for a strongly hydrophobic residue, phenylalanine, was utilized as the hydrophobic domain of the building block [57]. A fusion of GFP to the building block resulted in many fluorescent higher-order structures being formed, which signified that the building block was able to assemble into compartment-like structures [57]. Only a very small variation in the ratio of hydrophilic to hydrophobic domains resulted in successful compartment formation [57]. The unnatural amino acid par-azido-L-phenylalanine, which can be readily functionalized with dyes and other compounds for subsequent imaging applications, could be successfully incorporated in the amphiphilic protein without affecting its assembly [57].

![Figure 2](image_url)

\textbf{Figure 2. Protein fusions can generate compartment-like structures of desired size and shape.}

(a) The trimeric KDPGal aldolase protein (green) and the dimeric N-terminal domain of FkpA (orange) were fused together with a short linker (blue) to create a building block for the assembled cage. (b) The assembled cage consists of 24 building blocks. The N-terminal domain of FkpA forms the “edge” of the cage with 2-fold symmetry, while the KDPGal aldolase forms the “vertices” of the cage with 3-fold symmetry. This figure is reproduced with permission from Lai \textit{et al.} [56]
1.4.4 Adopting existing protein compartment-like structures for use as a compartment system

While the design and synthesis of novel compartments can be challenging, adapting / co-opting existing protein compartment-like structures that do not naturally compartmentalize proteins offers an additional route to selectively encapsulate target enzymes into compartments. One such example involves ferritin, a protein that naturally forms a tightly packed cage with small pores and functions as an iron storage complex. Ferritin variants were engineered using an approach termed metal-templated interface redesign, whereby specific structures would be assembled upon selective copper chelation between the subunit interfaces [58]. The addition of metal-chelating groups and the re-engineering of complementary interactions resulted in the controlled formation of ferretin cages only in the presence of copper ions [58]. The interior of the capsid was functionalized with cysteine residues resulting in the successful localization of fluorophores in the cage interior. The fluorophores could only interact with cysteine residues only if the ferretin cage was not initially assembled [58]. The ability to control both the assembly and loading of specific targets into the ferretin cage broaden its application for biotechnology and nanotechnology.

Another example of a protein that self-assembles into a cage is lumazine synthase. Lumazine synthase from *Aquicus aeolicus* naturally assemble into icosahedral protein shells containing either 60 or 180 enzyme copies and catalyzes the penultimate step in riboflavin biosynthesis [59]. To convert lumazine synthase into a compartment for industrial applications, rounds of mutagenesis were conducted by selecting for the altered compartment’s ability to protect *E. coli* from poly-arginine-tagged HIV protease; an enzyme which is toxic to *E. coli*. The best performing mutant compartment encapsulating an average of 7 HIV protease dimers per compartment [59]. Characterization of the mutations revealed targeted residues associated with the lumen of the compartment, as well as the subunit interface, were responsible for the increase encapsulation capacity of lumazine synthase [59]. A subsequent study using GFP mutated to contain 26 additional positive residues yielded approximately 100 GFP molecules per compartment [60]. The ability to modify a compartment to maximize loading is a valuable design feature for future applications such as creating reaction vessels and encapsulating multi-
enzyme systems, where increasing loading capacity can increase the production of metabolites from a single capsid.

### 1.4.5 Converting specialized micro-compartments into flexible encapsulation systems

Native prokaryotic micro-compartments have two significant properties which were thought to limit their potential to be used in channelling applications in comparison to general protein containers. The first characteristic is the selectivity of pores on micro-compartment shells. This selectivity may limit the rate in which substrates of interest can diffuse into the compartment shell, which in turn limits the reaction rate of the target enzymes to be encapsulated within the shell. The second characteristic is the selectivity of encapsulation within micro-compartments since targeting peptides are typically required to ensure encapsulation of enzymes within a micro-compartment [8, 9]. Insufficient characterization of a micro-compartment’s target peptide can reduce potential targeting levels and effectiveness for channelling applications.

Further research into micro-compartment and nano-compartment targeting peptides has generated insights into the key characteristics of targeting sequences which enable selective targeting into compartments. Attachment of an 18-residue peptide to the N-terminus of GFP resulted in its encapsulation inside Pdu micro-compartments [8]. Previous studies demonstrated that heterologous expression of 5 different ethanolamine utilization (Eut) compartment shell proteins resulted in the reconstitution of the compartment in a heterologous host [9]. Fusing a 19 residue peptide to the N-terminus of GFP and expressing the construct in a Eut shell protein *E. coli* background led to successful encapsulation of GFP within Eut compartments [9]. To test whether heterologously expressed proteins could be active upon encapsulation, β-galactosidase was targeted into Eut compartments and probed with X-gal to assay for activity. Formation of the X-gal cleavage product was localized as determined by differential interference contrast microscopy after an overnight induction of β-galactosidase expression and incubation of X-gal [9]. This resulted demonstrated that X-gal could diffuse into the compartment shell despite the presence of selective pores and that β-galactosidase was still active upon encapsulation. Further analysis of other signal sequences that can target proteins into the Pdu micro-compartment revealed that a common hydrophobic motif drives targeting [61]. While the Pdu and Eut
compartments rely on an N-terminal targeting peptide for enzyme encapsulation, a C-terminal peptide was shown to be responsible for targeting native dye-decolouring peroxidase into the encapsulin nano-compartment from *Thermotoga maritima* [7]. More recently, a 17 residue peptide was found to successfully target *Synechococcus elongatus* carboxysomal proteins inside a carboxysome assembled in *Nicotiana benthamiana* [62]. A summary of currently characterized targeting mechanisms is provided in Table 2.
Table 2. A summary of experiments adapting proteins and protein compartments as protein containers for biotechnological applications

<table>
<thead>
<tr>
<th>Name of system</th>
<th>Protein(s) targeted</th>
<th>Method of Targeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pdu micro-compartment</td>
<td>• GFP&lt;br&gt;• Glutathione S-transferase&lt;br&gt;• Maltose-binding protein</td>
<td>N-terminal peptide, <em>in vivo</em> [8, 9]</td>
</tr>
<tr>
<td>Eut micro-compartment</td>
<td>• GFP&lt;br&gt;• β-galactosidase</td>
<td>N-terminal peptide, <em>in vivo</em> [9, 63]</td>
</tr>
<tr>
<td>CCMV capsid</td>
<td>• Horseradish Peroxidase</td>
<td><em>In vitro</em> capsid re-assembly [64]</td>
</tr>
<tr>
<td></td>
<td>• GFP and <em>Pseudozyma antartica</em> Lipase B co-encapsulation</td>
<td>Engineered coiled-coil interaction, <em>in vitro</em> [65]</td>
</tr>
<tr>
<td>Lumazine synthase</td>
<td>• GFP&lt;br&gt;• HIV protease</td>
<td>Poly-arginine tag to a modified compartment with increased negative charge, <em>in vivo</em> [59, 66]</td>
</tr>
<tr>
<td>P22 capsid</td>
<td>• GFP and mCherry&lt;br&gt;• Alcohol dehydrogenase D&lt;br&gt;• α-galactosidase&lt;br&gt;• CelB, glucokinase and galactokinase fusion</td>
<td>Fusion to naturally encapsulated scaffold domain, <em>in vivo</em> [46, 67, 68]</td>
</tr>
</tbody>
</table>

1.5 Viruses as enzyme compartments

Initial characterization of micro-compartment revealed striking similarities between the structure of compartment shell monomers and viral capsid monomers [7]. Given that micro-compartments can be modified for industrial applications, it is logical to expect that viral capsids could be amenable to similar modifications. Initial explorations involving natural protein cage systems such as ferretin, heat shock proteins and virus capsids for industrial applications allowed for the mineralization of inorganic compounds within the cage interior [69-71]. Drawing inspiration from these previous studies, interest has risen for the use of capsids from plant viruses and bacteriophages as compartments for enzymes. Capsids from both systems are assembled through building blocks of one or more types of shell proteins. Plant virus capsids are assembled from capsomeres arranged in an icosahedral (e.g., brome mosaic virus, cowpea chlorotic mosaic virus) or tubular (e.g., tobacco mosaic virus) structure [72]. Similarly, bacteriophage capsids are assembled from hexamers and / or pentamers of the capsid protein into icosahedral structures, with the aid of scaffolding or stabilization proteins to ensuring proper assembly [73]. The scaffolding protein can be found as a separate peptide (e.g., T7, P22) or fused to the capsid protein (e.g., HK97). Following proper assembly of the capsid, the scaffolding protein either
diffuses out of the capsid shell or is degraded prior to the loading of the viral genome into the capsid.

Structural studies between phage capsids and the nano-compartment encapsulin revealed that the protein monomers and the overall architectures were very similar, lending credence to the idea that plant and phage capsids could be altered to become compartments [7]. However, the lack of understanding of how enzymes were targeted into phage capsids as well as a lack of techniques to engineering targeting into these structures delayed the adoption of capsids as potential compartments.

1.5.1 Previous work in the field

The first instance of utilizing viruses as an enzyme container was the localization of horseradish peroxidase (HRP) inside of the cowpea chlorotic mottle virus (CCMV) capsid [64]. The CCMV capsid is approximately 28 nm in diameter, and has a unique property where incubation at pH 7.5 dissociates the capsid, while incubation at pH 5 promotes reassembly of the capsid [64]. HRP was successfully encapsulated in CCMV capsids an initial *in vitro* incubation of HRP and dissociated capsid dimers at pH 7.5, followed by a shift to pH 5 to promote reassembly of the CCMV capsid with HRP localized inside [64]. HRP was estimated to be encapsulated at a ratio of approximately 1 copy of HRP per CCMV capsid [64]. Activity of the encapsulated HRP was confirmed through confocal microscopy by detecting the conversion of dihydrorhodamine 6G into the fluorescent substrate rhodamine 6G in purified CCMV samples [64]. A subsequent study aimed to target multiple copies of the same enzyme into CCMV capsids by introducing a coiled-coil targeting system (Figure 3). Using this technique, the altered CCMV capsid could encapsulate up to an average of 15 molecules of GFP per capsid [74].
To understand the effect of encapsulating varying amounts of enzyme on enzymatic activity, lipase B from *Pseudozyma antarctica* was encapsulated into CCMV capsids in varying ratios. Encapsulated enzymes were found to have a significantly higher initial reaction velocity as compared to the same concentration of encapsulated enzyme [65]. However, increasing the amount of encapsulated enzyme from 1.3 copies of PalB per capsid to 4 copies resulted in a decreased initial velocity, dropping from a 5-fold to only a 2-fold increase of initial velocity as compared to unencapsulated enzyme [65]. These results were further corroborated by experiments exploring PalB activity in capsids with or without GFP co-encapsulated. Despite a similar level of PalB encapsulation, co-encapsulation with GFP was found to decrease the initial velocity of PalB [65].

Another system that has been explored for use as a compartment was the capsid from bacteriophage MS2. Alkaline phosphatase (phoA) was conjugated to DNA and encapsulated *in vitro* by combining the enzyme with MS2 capsid dimers and osmolyte to promote correct reassembly of the capsid [75]. Targeting levels were estimated to be approximately 8 copies of phoA per MS2 capsid [75]. In this system, the activity of the encapsulated phoA was approximately 75% of the activity the unencapsulated phoA, which contradicts the findings about encapsulated enzymes previously mentioned for the CCMV system [75]. While no explanation was offered for this discrepancy, a possible explanation may involve the amount of protein loaded into the compartment system. As highlighted in the CCMV – PalB study, an increased amount of enzyme loaded into the CCMV capsid led to decreased activity of PalB [65]. When comparing the amount of encapsulated PalB in the CCMV capsid and phoA in the MS2 capsid, there is a greater amount of encapsulated protein in the MS2 capsid study. With an
increased amount of encapsulated protein, substrates may take longer to diffuse into enzyme active sites, contributing to lower reaction rates.

The final capsid system to highlight involves the capsid of bacteriophage P22. The P22 capsid assembles through the aid of a scaffold protein that allows the capsid proteins to orient in the appropriate cage structure. After initial assembly, the scaffold protein diffuses out of the capsid interior before subsequent maturation steps and genome loading yields the mature P22 capsid head. A method was developed to target proteins of interest inside of P22 capsids by fusing proteins to a truncated version of the scaffold protein [76]. Approximately 100 copies of GFP and mCherry fused to the scaffold were packaged per assembled capsid, and were found to be in closer proximity than the fusion in free solution as determined by fluorescence resonance energy transfer [67]. Subsequent targeting with alcohol dehydrogenase (AdhD) yielded almost 250 copies per capsid [77]; however, the activity of the encapsulated enzyme was approximately 70 fold lower as compared to the unencapsulated form [77]. Neither expanding the internal volume nor forming large pores on the capsid pores by using a whiffleball mutant capsid resulted in increased activity [77]. The Whiffleball capsid is a mutant structure where the capsid protein pentamers are not present in the assembled capsid. This unique phenomenon was found for the HK97 capsid upon mutation of residue E219 to a lysine [10] and for the P22 capsid upon extended heating at 75°C [78].

The P22 capsid was also used to prevent Pyrococcus furiosus α-galactosidase (GalA) from forming insoluble aggregates in E. coli. While the P22 capsid reliably encapsulated GalA, only a combination of expanding the capsid internal volume and forming whiffleball capsids resulted in GalA turnover measurements that aligned to previously reported levels [46]. This finding indicates that both internal volume and pore sizes are important to allow for sufficient metabolite flow within the capsid that result in activity levels of the unencapsulated enzyme. While the enzyme was successfully localized, attempts to free the enzyme from the capsid interior were unsuccessful, as various treatments to isolate the encapsulated enzyme resulted in protein precipitation. A final reported application of P22 capsids was to elicit a reaction cascade through multiple enzymes. The enzymes CelB (a β-galactosidase), GalK (galactokinase) and GluK (glucokinase) were fused together with spacer peptides and attached to the P22 scaffold for co-encapsulation [68]. In this study, the authors sought to test the hypothesis that coupling sequential enzymes within a compartment system can increase production of a downstream
metabolite by localizing the sequential enzymes within P22 capsids. Upon encapsulation of the enzymes, no demonstrable increase in downstream metabolite production was observed (as compared to the unencapsulated enzymes). To help understand why this system did not behave as predicted, a mathematical model was developed to understand why no activity enhancement was observed and to better predict scenarios which can lead to increased reaction speeds. The model suggested that altering turnover and Michaelis-Menten constant (Km) of enzymes in a specific direction would result in a successful cascade reaction [68]. Modification of the enzymes resulted in kinetic efficiencies that were similar to the values predicted by the model [68]. The order of fused enzymes was found to affect both the assembly of the viral capsid as well as the activity [68]. Taken together, these findings suggest that encapsulation of enzymes within a compartment does not necessarily result in enhanced rates of reaction (leading to an increase in the production of downstream metabolites), which has implications for future designs of multi-enzyme systems.

1.5.2 Limitations of current protein compartment systems

Although increasing numbers of compartment systems are being discovered and modified for downstream applications, there are several drawbacks associated with current systems. The CCMV and MS2 systems, which rely on opening the capsid and releasing the viral genome before reassembly with cargo protein, can only target enzymes in vitro. As a result, in vivo applications cannot be explored with this system unless capsids are functionalized to allow for uptake into cells and remain intact upon uptake. Another drawback of existing systems (i.e., CCMV capsid, MS2 capsid and the encapsulin nano-compartment) is the limited encapsulation capacity of their capsids. The above compartment systems have inner diameters of approximately 18 nm, and were reported to hold 10 or fewer copies of target proteins ranging from 25 to 50 kDa in size [7, 65, 75]. The limited encapsulation capacity of certain systems can limit the effectiveness of multi-enzyme systems as limited enzyme encapsulation may result in no noticeable kinetic benefit. A third drawback, found primarily with micro-compartments and certain nano-compartments, is the selectivity of pores on the compartment shell. Studies of the Pdu, Eut and Etu compartments showed several selectivity mechanisms that control which substrates can enter. While X-gal was able to enter Eut compartment [9], the rate of diffusion may have been limited by the selectivity of pores as X-gal diffusion was allowed to occur
overnight. The encapsulin nano-compartment was also shown to have 3 pores with varying electrostatic potentials that could affect substrate diffusion across the compartment shell [7]. Selectivity can greatly decrease the flexibility of a system by limiting the number of enzymes can be encapsulated to those in which substrates can diffuse into the compartment lumen.

A final drawback to current systems is the difficulty in targeting a range of enzymes into capsids for high-throughput studies, particularly when investigating potential channelling between enzyme pairs. Given the combinations of enzymes being co-targeted in this type of study, investigating channelling activity would be both time-consuming and labour intensive. This also applies for studies which involve encapsulating many putative insoluble enzymes for future biochemical characterization. In the P22 system, enzymes are targeted through fusions to the scaffold protein with linkers connecting multiple enzymes together before a subsequent thrombin cleavage step. The process of making fusions to target a large collection of multiple enzymes carries a greater risk as one element may not function and could lead to disruption of the entire system. In addition, a study targeting multiple enzymes into P22 capsids found that varying the order of the fusion had an effect on the activity of the enzymes [68]. This creates even more complications as multiple versions of the same fusion will need to be made to find the ideal arrangement to maximize kinetic activity. With respect to micro compartments, not all genes found in micro-compartment loci have been fully characterized, meaning there could be unknown ancillary proteins that are essential for full micro-compartment functionality. If expression of many genes is required for full micro-compartment functionality, heterologous expression and targeting of enzymes may be difficult to achieve and could yield suboptimal activity for multi-enzyme systems.

1.5.3 HK97 capsid as a potential compartment

In attempts to alleviate some of the limitations outlined above, viral capsids have been explored for their potential as compartments in applications such as enzyme cascades, rescuing insoluble enzymes, and increasing kinetic efficiency. Amongst these, the HK97 capsid offers promise as a compartment system as it is a well-characterized system that is easy to reconstitute and manipulate in vivo. The HK97 capsid is composed of 415 copies of the major capsid protein, which makes up the capsid shell, and 12 copies of the portal protein which assemble into a ring
structure [11]. The portal assembly is located on one vertex of the capsid and has a 30Å pore through which the genome is packaged into the capsid [11]. The major capsid and portal proteins, along with ~60 copies of the HK97 maturation protease, co-assemble into a structural intermediate approximately 45 nm in diameter known as Prohead I (Figure 4a). The protease is targeted inside the assembling capsid by a 63 residue C-terminal targeting sequence found on the maturation protease [79]. The protease removes the N-terminus from each of the assembled major capsid proteins and proteolyzes itself, resulting in the creation of the second structural intermediate Prohead II [11]. The proteolyzed material diffuses out of the capsid through pores approximately 10 Å in diameter before subsequent DNA packaging and maturation steps create the final mature capsid that is 60 nm in diameter [80].

(a)

(b)

Figure 4. The assembly pathway of the HK97 capsid.

(a) 12 copies of the portal protein, approximately 60 copies of the maturation protease and 415 copies of the major capsid protein co-assemble into a structural intermediate known as Prohead I. The maturation protease is encapsulated inside of the capsid shell and the portal complex is found at the vertex of the capsid. The protease is thought to be localized in the assembling capsid through the aid of a C-terminal, 63 residue targeting sequence. Upon assembly, the maturation protease activates and cleaves the delta domain of the major capsid protein as well as itself. The proteolyzed material diffuses out of pores on the capsid shell to form the Prohead II structural intermediate. Subsequent DNA loading and maturation steps yield the fully mature capsid.

(b) Expressing the genes encoding the portal, protease and major capsid proteins result in the assembly of HK97 Prohead II capsids (yellow arrows) which can be purified from E. coli cells and visualized using TEM.
1.5.4 Improvements and potential applications

There are several advantages of using the HK97 capsid as a protective compartment for enzymes or as a nanoreactor containing multi-component enzymes system. Firstly, the HK97 capsid is the most well characterized bacteriophage capsid; its assembly has been thoroughly characterized and crystal structures of its structural intermediates and final mature form have been solved. This allows for systematic alterations of capsid components to identify changes that can enhance the ability of the HK97 capsid to act as a compartment. Secondly, the C-terminal targeting sequence found on the maturation protease represents a mechanism that can be readily exploited to allow targeting of heterologously-expressed enzymes inside of the HK7 capsid [79]. Thirdly, capsid assembly and enzyme targeting occur in vivo, which may allow in vivo applications such as introducing engineered pathways into E. coli, using HK97 capsids as a vessel. Finally, the presence of pores along the shell and in the portal assembly provides locations with which substrates can diffuse across the compartment shell.

1.6 Applications of engineered protein compartments for metabolic channelling

One major area where compartments can be utilized is to introduce metabolic channelling into biological systems of interest. The ability to control metabolic flow also plays important roles in regulation in vivo and represents an enormous opportunity for synthetic biologists and metabolic engineers to further improve the efficiency of systems of interest. In the following sub-sections, I highlight computational studies attempting to characterize channelling in nature and experimental attempts to introduce channelling into enzyme systems of interest to give context to how my subsequent chapters fill in current gaps in the field of channelling.

1.6.1 Previous work on characterizing metabolic channelling

In the field of metabolic channelling, there are two major areas that are still left unanswered. The first area involves understanding the extent of channelling in nature. Specific examples of channelling have been identified in various organisms, such as the hydrophobic tunnel channelling of a hydrophobic molecule between active sites in tryptophan synthase [81]
and complexes like the fatty acid synthesis complex that directly shuttle substrates to different active sites for sequential processing [82]. However, factors underlying channelling have not been fully explored and systematic studies to identify reactions that may be undergoing channelling in nature are lacking. Investigating this question may lead to the identification of new mechanisms underlying channelling, clarity in which situations channelling will elicit desired effects such as increased kinetic efficiency and reduced cellular toxicity, and finally, new sets of enzyme systems that can benefit from channelling that may also have industrial or pharmaceutical importance. These findings can then be further confirmed experimentally through techniques such as transient time analysis, isotope dilution and enrichment, and challenges from competing side reactions or inhibitors [83]. For example, in one study, the transient time for bifunctional enzyme thymidylate synthetase - dihydrofolate reductase (TS–DHFR) was compared to the transient time of the un-fused, mono-functional enzymes [84]. The transient time of a sequential reaction is an observable lag in a reaction time course and is the time required to reach the steady-state flux for the coupled enzyme’s intermediate compound. The bifunctional enzyme produces a τ close to zero, while the same reaction with the two un-fused enzymes results in a 25 s lag before reaching steady state. In a separate experiment, an isotopically labelled substrate and excuse unlabelled substrate was incubated with bifunctional TS-DHFR or monofunctional TS and DHFR to determine whether channelling occurs in the bifunctional enzyme [85]. The ratio of labelled to unlabelled product provides an indication of whether channelling occurs (i.e., a significantly higher percentage of labelled to unlabelled product will indicate the presence of channelling). 55% of labeled product was measured when conducting the experiment with monofunctional enzymes, while almost all product was labeled when the substrates were incubated with the bifunctional enzyme. These findings, along with several other observations, lead to the hypothesis that channelling occurs between TS and DHFR).

The second major area explores the possibility of engineering channelling into enzymatic pathways of interest. While an exploration of channelling in nature can provide clues to what enzymes can benefit from channelling, it is important to validate the current understanding experimentally and determine whether it is feasible to utilize channelling for metabolic engineering. If enzyme combinations that are predicted to benefit from channelling can successfully be engineered to undergo channelling, it will re-enforce the current understanding of
channelling and highlights the incredible potential for creating efficient design of enzymatic systems. However, if attempts to engineer channelling result in lower than expected or no channelling effects, it will indicate the presence of non-documented biochemical factors that influence channelling. This scenario would highlight the need to better understand what factors are important for effecting channelling prior to future attempts at engineering channelling into enzyme systems. One method of identifying additional factors involves a combinatorial approach of computational and experimental studies. The computational approach can be used to determine the extent of channelling in nature and inform predictions as to what enzymes could benefit from it. The experimental approach will be used to test predicted combinations and recursively refine the computational approach to increase the likelihood of correctly predicting enzyme combinations that will benefit from channelling.

1.6.1.1 Computational characterizations of metabolic channelling

Surveys of the prevalence of metabolic channelling in nature have been limited in number and in scope, utilizing only one of the many mechanisms that can elicit channelling as the criteria for surveying. Huthmacher et al. explored protein-protein interactions reported in *Escherichia coli* and *Saccharomyces cerevisiae* databases to identify potentially channeled reactions. Interactions were compared to organism-specific metabolic networks generated by linking reactions that share at least one non-currency metabolite (highly abundant reactants or cofactors used in many reactions such as H₂O, ATP, NADH etc.) with the thinking that any interacting enzymes that are linked in the metabolic networks could be undergoing metabolic channelling [86]. Candidates for metabolic channelling were determined by looking at linked reactions in the metabolic networks that had previous evidence for protein interactions. With this criterion, 1759 reaction pairs from *E. coli* and 2508 reaction pairs from *S. cerevisiae* were postulated to be involved in metabolic channelling. While this study served as an initial step in identifying enzymes that may benefit from channelling, the criterion used in the study is not an accurate predictor of channelling and simplifies what is required for channelling to occur. For example, non-sequential reactions may interact and recruit intermediate enzyme(s) to generate a channeled reaction across a number of enzymes. Non-interacting enzymes may also be co-localized within restricted spaces (e.g., micro-compartments, organelles, phase separation) and undergo channelling without ever interacting. Therefore, the criteria that two enzymes physically
interacting being required for channelling is not a necessary nor sufficient condition to predict channelling between enzymes.

Exploration of the Gibbs free energy changes of reactions catalyzed by enzymes in this study (i.e., difference in the Gibbs free energy between all substrates and all products) led to the hypothesis that the channelling pairs highlighted in the study may be involved in regulation. The high negative Gibbs free energy change of one of the enzymes in each pair, and the presence of those enzymes at metabolic branching points (where multiple reaction paths can be taken by an intermediate) are hallmarks of regulatory enzymes in metabolic networks. This is due to the high negative Gibbs free energy reaction occurring after a branch point that can act to pull metabolic flux towards the high negative Gibbs free energy reaction [87]. This phenomenon is caused by the depletion of the metabolic intermediate that represents the substrate of the high negative Gibbs free energy reaction. To maintain the desired equilibrium, pre-cursor metabolites will be converted into the substrate for that reaction, creating a “pulling” metabolic flow effect. Through examining metabolic reactions in the context of large scale protein interaction networks, ~200 pairs of reactions from yeast and E. coli were predicted to channel using the combined network and protein-interaction approach [87]. A large proportion of predicted channelling pairs have not been experimentally characterized, and further attempts to identify attributes of the enzyme pairs that would indicate an increased likelihood of channelling (i.e., path lengths, interaction probabilities in randomized networks) did not reveal any attributes that are strong indicators of potential channelling between enzymes. A later study modeled the effects of compartmentalizing enzyme pairs in a small segment of the glycolysis pathway to investigate the effects of co-localization on kinetic efficiency [88]. Localizing reactions with a low equilibrium constant followed by a sequential reaction with a high equilibrium constant resulted in the greatest increases in catalytic efficiency [88]. The absence of localizing a subsequent enzyme in the pathway resulted in accumulation of metabolic intermediate which acted to decrease metabolic flux through the pathway. The findings that enzyme co-localization can decrease the accumulation of intermediate substrates, depending on the pathway relationships of the co-localized enzymes and their kinetic parameters, suggested that channeling may be useful to overcome thermodynamic barriers caused by unfavourable reactions limiting metabolic flow within a biochemical pathway. In combination with the previous study where reactions with high negative Gibbs free energy are found at branch points, the findings suggest that metabolic flow
can be tightly controlled around branch points – namely through the presence of an energetically unfavourable reaction before the metabolic branch point and an energetically favourable sequential reaction occurring at the same branch point. This has implications in targeted metabolic engineering and reducing pathway cross-talk to elicit a desired outcome.

### 1.6.1.2 Experimental studies to introduce metabolic channelling into enzyme systems

Prior to the discovery that natural compartments are effective for introducing metabolic channelling in organisms, one of the earliest methods used to introduce channelling was by creating enzyme fusions. By co-localizing enzymes, the transit time of metabolites between active sites of the fused enzymes would decrease in comparison to the non-fused enzymes, leading to faster production of the desired product. Several pairs have been the focus of gene fusions as an attempt to increase the combined catalytic efficiency through the enzymes (i.e., to increase the production of a downstream metabolite). For example, fusing β-galactosidase and galactose dehydrogenase doubled the yield of NADH, which is a co-product along with D-galactono-1,4-lactone and helps tracks the progression of the coupled reaction [89]. In addition, a decreased transient time (i.e., lag time to reach steady-state flux) was also observed for the fused enzyme as compared to the unfused enzymes, which supports the hypothesis for channelling between the two enzymes [89]. While no statistical significance for the change in activity was indicated in the initial study, the combination of an increase in NADH production and reduced transit time was viewed as a channelling effect. Fusion of malate dehydrogenase and citrate synthase yielded differing results. One group argued that oxaloacetate was being channelled between the active sites of the fusion protein based on a reduction of transient time and protection of the channelled intermediate from competing reactions [90]. Another study argued that the intermediate travels similarly to freely diffusing enzymes and that the fusion protein altered the kinetics of the two enzymes, rather than restricting movement of the metabolite towards the sequential active site [91]. A third study analyzing this fusion attributed channelling to a difference in electrostatic potential between the two active sites, rather than restricted diffusion between the enzymes [92]. However, during the exploration of this claim, no clear evidence of any significant electrostatic potential difference was identified that would facilitate channelling the intermediate between the active sites [91]. Based on the information presented
across these studies, particularly the transient time decrease and competitor reaction protection findings, it appears likely that oxaloacetate may be directed towards citrate synthase. However, the loss of transient time decrease and competitor reaction protection through increased ionic strength (proposed to mask a charged surface directing oxaloacetate diffusion) does not directly prove channelling is introduced through a charged surface. Other factors, such as denaturation of the enzyme or negative affects on the enzymatic active sites, are not explicitly ruled out in this scenario.

The use of gene fusions for channelling was improved by creating fusions of protein interaction domains to form scaffolds that can localize multiple enzymes [93]. Unlike gene fusions, the idea involves protein scaffolds composed of discrete units that can bind to a specific target enzyme to create an ensemble of enzymes. This collection of enzymes can then work in unison to convert an initial substrate to a final product through a defined set of biochemical reactions. By fusing combinations of the protein-protein interaction domains SH3 (PFAM PF00018), PDZ (PFAM PF00595) and GBD (PFAM CL0202) on a scaffold and their corresponding ligands to a target protein, the optimal stoichiometries of the localized enzymes can be obtained. Using this method, yeast hydroxyl-methylglutaryl-CoA synthase (HMGS), yeast hydroxyl-methylglutaryl-CoA reductase (HMGR) and acetoacetyl-CoA thiolase (AtoB) from *E. coli* were successfully localized to a synthetic scaffold [93] (Figure 5a). Utilization of the scaffold increased the production of mevalonate, a precursor to the anti-malarial compound artemisinin, by nearly 70-fold as compared to the non-scaffolded enzymes [93] (Figure 5b). The increase is attributed to the alleviation of a metabolic flux bottleneck created by HMGR, which catalyzes the 3rd of 3 sequential reactions in the cascade. HMGR catalyzes a thermodynamically unfavourable reaction which leads to a build-up of toxic intermediates earlier in the pathway such as hydroxymethylglutaryl-CoA, which is generated by the first enzyme in the reaction cascade, AtoB [93]. The modularity of this system as well as the flexibility to create larger scaffolds with many enzymes combined in a specified stoichiometry make it an attract system to engineering reaction chains to produce compounds of pharmaceutical or industrial interest.
Figure 5. Localization of target proteins to a synthetic scaffold to increase kinetic efficiency for a reaction cascade.

(a) Interaction domains SH3, PDZ and GBD were fused together with short linkers to form a scaffold. Target enzymes yeast hydroxyl-methylglutaryl-CoA synthase (HMGS), yeast hydroxyl-methylglutaryl-CoA reductase (HMGR) and acetoacetyl-CoA thiolase (AtoB) are conjugated to interaction domain ligands (orange square, circle, triangle) to localize the target enzymes to the synthetic scaffold. (b) Ratios of the different target domains are altered to find the optimal stoichiometry for kinetic activity through the three-enzyme cascade. A ratio of 1 GBD domain, 2 SH3 domains and 2 PDZ domains resulted in an increase in kinetic activity of over 70-fold as compared to the non-scaffolded enzyme mix. This figure is reproduced with permission from Dueber et al. [93].

1.6.2 Limitations, challenges and opportunities in channelling

Despite advances made by previous computational and experimental studies on metabolic channelling, there were key limitations to the studies that need to be considered. The first limitation with previous computational studies of metabolic channelling is that different channelling mechanisms were explored independently, with no considerations of other potential mechanisms of channelling in each study. Looking at multiple mechanisms of channelling (i.e., overlapping gene fusion/protein complex data with enzymes that may be undergoing channelling toxic/inhibitory intermediates) can reveal more complex and larger-scale regulatory patterns. Examining one channelling mechanism in isolation will not reveal such mechanisms. To date, no comprehensive, large-scale study of metabolic channelling in nature has been performed. This represents a large opportunity to not only understand the potential of channelling for engineering purposes but to identify new mechanisms that can lead to channelling. A further challenge associated with current attempts to introduce channelling into enzymes is that intermediate substrates are not restricted from diffusing into the bulk cytosol. While the use of gene fusions and scaffolds bring enzymes in close proximity, this only serves to increase the likelihood that
metabolites will diffuse to the biochemically sequential enzyme, rather than restrict diffusion of metabolites (a requirement for metabolic channelling). Fusions and scaffolds do not create a significant barrier restricting metabolite diffusion to the cytosol, which is a key feature associated with metabolic channelling. This creates an opportunity for using compartments to generate channelling between enzymes as it not only co-localizes enzymes in a confined space, but it also restricts diffusion of metabolic intermediates into the cytosol.

Secondly, there are some inherent challenges in performing a large-scale and comprehensive survey of channelling, such as the quality and size of datasets available for analysis. Gene fusion and protein complex data is limited by what is available from sequenced organisms and protein-protein interaction data, both of which only represent a small fraction of what is known in nature. Any potential enzyme combinations identified or predicted to be undergoing channelling are likely represent a very small fraction of the full extent of channelling in nature. In addition, there is a limited understanding of what biochemical factors are indicative of good channelling candidates, which will be described in further detail in Chapter 2. One of the hypothesized factors is based on the thermodynamic values of reactions. Given the limited experimental information available regarding the thermodynamics of all biochemical reactions in nature, predicting enzymes that can benefit from channelling based on predicted thermodynamic values may lead to potential false positives and negatives.

Thirdly, there is a lack of appropriate visualization tools to highlight any novel channelling complexities identified in a potential large-scale study. Metabolic networks can be visualized using tools such as Cytoscape, where a network representation can be created with compounds as nodes and enzymes catalyzing the conversion of one compound to another as edges [94]. The challenge with visualizing channelling within a metabolic network is that channelling occurs over multiple reactions, with participating reactions also containing additional features such as physical linkages, thermodynamic directionality, etc. Overlaying multiple layers of features within a traditional metabolic network representation will result in difficulties interpreting the visualization and generating insights. Hypergraphs are a potential solution in integrating multiple features into an easy-to-interpret network diagram. They are based on a principle where multiple nodes can be joined by non-linear edges called hyperedges [95]. Several studies have demonstrated the value in using hypergraphs to better depict the complexity inherent in data [96, 97]. Although hypergraphs are not currently supported by a
number of network tools, including Cytoscape, a recently-developed software (Hyperscape) has been developed to integrate hypergraphs into metabolic network visualizations [95]. Hyperscape may represent a good solution for displaying channelling and other network complexities within metabolic networks.

Finally, previous attempts to engineer metabolic channelling have yielded inconsistent gains across the highlighted systems (whether through gene fusions or scaffolding). Changes range from a 2-fold increase in kinetic efficiency using gene fusions [89], to a 70-fold increase in kinetic efficiency using a protein scaffold [93]. The fusion and scaffold systems described previously do not fit under the definition of channelling, as metabolites are not restricted from diffusing into the cytosol. While compartmentalization offers the opportunity to solve the issue of substrate diffusion to the cytosol, there are key challenges to consider. The first challenge is determining which enzymes can yield kinetic benefit within a compartment. In one study, the enzymes CelB (which catalyzes the hydrolysis of lactose to galactose and glucose), galactokinase (which phosphorylates galactose using ATP to generate galactose-1-phosphate) and glucokinase (which phosphorylates glucose using ADP to generate glucose-6-phosphate) were successfully localized in P22 capsids. While all three enzymes were active inside the compartment, no kinetic benefit was observed. A better understanding the various factors (i.e., enzyme properties and kinetics, compartmentalization efficiencies) that lead to a channelling benefit is necessary prior to the widespread application of compartments for channelling. A second major challenge is the ability to control the encapsulation efficiency of enzymes into compartments. Most targeting studies currently aim to maximize encapsulation efficiency into compartments or target average amounts per capsid through in vitro reconstitution experiments. Given the importance of stoichiometry in maximizing kinetic efficiencies as seen with the artemisinin pathway protein scaffold study [93], the ability to control the relative abundance of multiple enzymes will be crucial in optimizing multi-enzyme systems encapsulated in compartments.

1.7 Project goals

The overarching goal of this project was to reveal the potential for coupling metabolic reactions that alleviate thermodynamic and physiological constraints that limit biochemical
systems (i.e., decrease reaction efficiency), with a view to understanding how coupling of reactions may be applied for metabolic engineering purposes. To achieve this goal, two complementary approaches were undertaken to address two key questions in the field. The first question involves understanding what enzymes can benefit from channelling. Previous computational studies on metabolic channelling have been limited and only explored the potential presence of channelling using simplified criteria (e.g., protein interactions) across select organisms (i.e., *E. coli, S. cerevisiae*). There is an opportunity to conduct a more comprehensive analysis and identify key insights which can be applied to metabolic engineering. The second question relates to whether compartments can be utilized to constrain reactions and result in potential benefits of channelling such as enhancing reaction efficiency or avoiding the accumulation of toxic intermediates. While methods such as gene fusions and protein scaffolds have increased the yield of metabolites of interest, their success has been inconsistent across different enzyme systems. In addition, the limitations of existing compartment systems, such as small compartment size, pore selectivity and the inability to encapsulate enzymes *in vivo*, limit their flexibility for various metabolic engineering applications.

To address the first question, I undertook a systematic survey of channelling to identify candidates for metabolic channelling based on currently understood mechanisms of channelling. Candidate enzymes that may benefit from metabolic channelling were identified utilizing three criteria outlined by a previous graduate student, Chris Sanford. The first criterion was to pair a thermodynamically unfavourable reaction with a biochemically sequential, thermodynamically favourable reaction of greater magnitude. As outlined earlier, introducing channelling in a system with an energetically unfavourable first reaction and an energetically favourable second reaction of greater magnitude can eliminate the kinetic bottleneck generated by the first reaction, allowing for increased kinetic flux across both reactions. The second criterion was to pair reactions that have a common toxic intermediate, while the third criterion was to pair reactions that have an inhibitory compound as a common intermediate. The later two criteria allow for the toxic/inhibitory compound to be isolated from the cytosol, limiting potential damage to the host organism. After identifying candidates, the list of predictions was evaluated against potentially channeled reactions in gene fusions and protein complexes to examine if there was any correlation between channelling and the presence of gene fusions and protein complexes. If a
correlation exists, this might support the hypothesis that gene fusions or protein complexes are viable mechanisms to introduce widespread channelling in nature.

To address the second question, I explored the possibility of exploiting the HK97 capsid as a compartment to introduce channelling into candidates identified in the previous goal. The initial step for this goal involved targeting proteins and enzymes into the capsid and exploring methods to maximize targeting. This was followed by encapsulating an enzyme and investigating the effect encapsulation had on enzymatic activity. After demonstrating the ability to target enzymes into capsids, I then described two systems of industrial importance for encapsulation. The first focused on enzymes that are unstable under endogenous conditions. While previous work has been done on insoluble enzymes with partial activity, there have been no reports on the use of compartmentalization to rescue insoluble and inactive enzymes. HK97 capsids may provide an opportunity to sequester insoluble enzymes from aggregation in cells and provide an isolated environment that may allow enzymes to fold in the proper confirmation. The second system focused on enhancing activity of enzymes in non-native conditions typically found in industrial processes. Previous work has shown that an enzyme’s activity can be extended to previously non-permissive pH and salt concentrations when encapsulated inside silica pores [49]. Targeting enzymes inside HK97 capsids may have similar stabilizing effects that could be utilized in conditions where the HK97 capsid remains intact to impart the effect on encapsulated enzymes.
Chapter 2: A systematic analysis of metabolic channelling from a thermodynamic and evolutionary perspective

2.1 Introduction

Genes and their products do not operate in isolation but rather form parts of integrated biochemical pathways. There is increasing evidence that for many pathways, their components exhibit varying degrees of spatial organization arising from effects of macromolecular crowding, sub-cellular compartmentalization or the formation of discrete complexes [98-100]. During evolution, many organisms have exploited spatial organization of enzymes as a method of promoting the channelling of metabolic intermediates to either enhance the efficient production of key metabolites or reduce exposure to toxic or otherwise inhibitory metabolites. While examples of this sort of metabolic channelling exist in nature (Sections 1.1, 1.2, 1.6.1.2), the absence of systematic surveys makes it unclear which reactions are most likely to benefit from co-localization.

Metabolic channelling has many potentially applications, including the field of bioremediation; however, further investigation is necessary to fully understand how to introduce channelling between enzymes of interest. In this chapter, I build on the work undertaken by Chris Sanford [101] by exploiting the wealth of existing genomic data in order to identify which enzymes are most likely to be co-localized and benefit from metabolic channelling mechanisms. Chris utilized available genomic, proteomic, biochemical, thermodynamic and pathway datasets to perform a systematic survey to address these questions and build on previous findings in the field. Based on Chris’ original methodology, predicted reactions that are likely to benefit from co-localization either from a thermodynamic perspective or through minimizing exposure to otherwise toxic or inhibitory intermediates were identified. Next, gene fusion and protein-protein interaction datasets were leveraged to systematically explore the incidence of metabolic channelling across the three domains of life, as well as gain a global overview within the context of the global metabolic network. Using these approaches, a total of 9,161 pairs of reactions that could benefit from metabolic channelling were identified. Specific examples of channelling identified from the global overview were discussed, along with how novel instances of gene
fusion events may have arisen as a consequence of overcoming thermodynamic and/or inhibitory constraints.

The final aspect of this chapter pertains to visualization of metabolism. Our predictions highlight the large number of reactions that have the potential to benefit from being physically connected to improve efficiency and protect the host organism; however, it is difficult to illustrate these relationships through traditional methods for visualizing metabolic relationships. Traditionally, network visualizations and KEGG-based diagrams have been used to visualize the relationship between enzymes and metabolites in metabolic pathways. However, current visualization methods lack the ability to communicate the complexities in metabolic pathways in a clear, concise manner. While network representations can connect enzymes with substrate or product metabolites across all pathways, visualizing many enzymes and metabolites across a number of different pathways can lead to confusion. Incorporating additional layers of information (e.g., regulation and reaction efficiency) can only be performed using existing edges or nodes by adjusting thickness and colour, with addition of multiple layers making such diagrams increasing difficult to interpret. KEGG-based representations are easier to follow; however, they cannot show how reactions from various pathways are linked together in an effective manner. Despite the additional flexibility associated with adding design features to KEGG-based representations to display multiple layers of information, key relationships are missed between the chosen representation and other areas of metabolism. Given these drawbacks, new methods for visualization are needed to better display the complexities of metabolic systems. Overall, the goal of this chapter is to advance the understanding of when a potential benefit from metabolic channelling can be realized, and identify enzyme pairs that are most likely to demonstrate a potential benefit when experimentally coupled.

2.2 Methods

2.2.1 Sources of data

Data on multifunctional enzymes were obtained from SwissProt and Prolinks [102, 103]. For Prolinks, a database of putative gene fusions, only predictions of high-confidence were included (P < 0.05). Thermodynamic data was obtained from two sources: the Thermodynamics
of Enzyme-catalyzed Reactions Database (TECRDB) hosted by the National Institute for Standards and Technology (NIST), which contains free energy values for ~400 reactions [104]; and a theoretically derived dataset containing free energy values for 874 reactions from the iJR904 E. coli metabolism model [105], that were calculated using the Mavrovouniotis group-contribution method [106]. Data on compound toxicity was obtained from ChemIDPlus which contains information on over 380,000 compounds [107]. 254 toxic compounds were identified using a cutoff of lethal concentration required to kill 50% of a population (LC50) \( \leq 100 \) ppm. An enzyme inhibition list of 343 inhibitors was obtained from the Braunschweig Enzyme Database (BRENDA - Version 0702), with inhibitors defined as any compound previously reported to inhibit any enzyme that also acted as substrates for two or more enzymes [108].

Protein interaction datasets were obtained from studies on the human soluble protein complex study, the yeast protein complex study, and the E. coli protein complex study [109-111]. Clustering of phylogenetic profiles generated from SwissProt data was performed using Cluster 3.0 [112].

### 2.2.2 Identification of enzyme pairs with the potential to benefit from channelling

To identify potential enzyme pairs that may be involved in channelling from the above data sources, one of three criteria had to be met. The first criteria is that the first enzyme in a pair must catalyze an energetically unfavourable reaction (\( \Delta G > 0 \)) and be paired with a second enzyme that catalyzes a biochemically sequential, energetically favourable reaction (\( \Delta G < 0 \)) with a free energy magnitude greater than the first reaction. This represents an enzyme pair that may undergo channelling in order to alleviate any thermodynamic bottleneck arising from the energetically unfavourable reaction. The second criteria requires the first enzyme in an enzyme pair to generate a toxic compound as a product, as determined from ChemIDPlus, with the second enzyme using the toxic compound as a substrate and converting it to a non-toxic compound. This represents an enzyme pair that may undergo channelling to reduce the level of the toxic compound such that it does not accumulate to lethal levels. The final criteria involves the generation of an inhibitory compound (as determined from BRENDA) as a product of the first enzyme in an enzyme pair, with a second enzyme converting the inhibitory compound to a non-inhibitory compound. This represents an enzyme pair that may undergo channelling to
reduce the level of the inhibitory compound below a threshold where significant inhibitory effects are observed.

2.2.3 Metabolic network reconstruction

Data on the connectivity of metabolic pathways was obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) ligand database [113]. Release 42.0 of this database contains 13,779 compounds are connected by 4,224 enzymes. To create a meaningful list of chemical intermediates, 36 compounds representing so called ‘currency metabolites’ were removed [114]. Using in house scripts written in Perl, enzyme pairs were generated by searching for enzymes that had the same non-currency metabolite in common. This resulted in a final list of 78,680 pairs of consecutive enzymes connected by 4,480 different metabolites. This list was used to construct a bipartite network, in which both enzymes and compounds are represented by nodes and edges between these nodes represent reactions involving the enzyme and compound. Network visualization was performed using Cytoscape version 2.8.1 [94].

The visualization of the pentose phosphate pathway to demonstrate the presence of reactions predicted to benefit from channelling was created with Adobe Illustrator CS4 using the KEGG depiction as a template. Gene fusions in the pentose phosphate pathway identified from the Rosetta Stone and UniProt data sets were highlighted in the depiction, while reaction pairs predicted to benefit from channelling are highlighted with red chevrons that start from the initial substrate, and travel through the enzymes catalyzing the reactions that create the final product.

2.2.4 Hyperscape Network Representations

The protocol outlined in Cromar et al. was followed to create Hyperscape network representations [95]. Briefly, linkages between compounds and enzymes were listed to generate the underlying metabolic network. Additional linkage lists describing channelling pairs that exist in the network (i.e., initial substrate to first enzyme, first enzyme to intermediate substrate, intermediate substrate to second enzyme, and second enzyme to final product) were then generated in order to create hyperedges overlaid on top of the metabolic network. The information was entered into the web application (http://www.compsysbio.org/hyperscape) to generate a network visualization that integrates hyperedges with the underlying network.
2.3 Results

2.3.1 Surveys of thermodynamic, toxic and inhibitory datasets predict 9,161 unique reaction pairs that represent candidates for metabolic channelling

To examine the extent to which channelling has been exploited in nature as well as the potential for channelling to result in beneficial effects, I utilized a methodology developed by a previous lab member, Chris Sanford, which investigates reaction thermodynamics, compound toxicities and compound inhibition potentials to identify enzyme pairs that could benefit from metabolic channelling as described in the Section 2.2. In these surveys, the aim was to both predict instances of channelling that may be exploited for industrial or pharmaceutical purposes, as well as reveal the extent to which different organisms have adopted channelling for beneficial purposes. Among the 78,680 pairs of enzymes (defined as unique enzyme classification (EC) numbers) that share a common substrate as defined by KEGG (i.e., the product of the first enzyme is a substrate of the second enzyme), identifying pairs that might benefit from metabolic channelling by either alleviating thermodynamic, toxic or inhibitory constraints would generate insights as the whether such pairings offer evolutionary advantages that warrant maintenance and optimization over time by nature. Focusing on the former, the laws of thermodynamics dictates that reactions that operate with low equilibrium coefficients are prone to inhibition by the accumulation of their product [88]. This was demonstrated when altering kinetic values of 4 enzymes in the glycolysis pathway (phosphoglyceratekinase; phosphoglycerate mutase; enolase and pyruvate kinase) and calculating the levels of the end metabolite (pyruvate) produced in the simulation with or without co-localization [88]. Results from the simulation indicated that the increase in catalytic efficiency, due to co-localization, is greatest for coupled reactions with an overall low equilibrium constant [88].

To explain why this phenomenon occurs, one must reference the Gibbs free energy equation below:

$$\Delta G = \Delta G^\circ + RT \ln \left( \frac{[B]}{[A]} \right)$$

$\Delta G$ represents the Gibbs free energy of the reaction, $\Delta G^\circ$ represents the standard Gibbs free energy of formation for compounds in the reaction, $R$ is the gas constant, $T$ represents the temperature, $[A]$ represents the concentration of substrates in a reaction, and $[B]$ represents the
concentration of products in a reaction. An energetically unfavourable reaction with a low equilibrium constant would typically be associated with a positive standard Gibbs free energy of formation. If an energetically favourable reaction (with a negative standard Gibbs free energy of formation) that catalyzes a biochemically sequential downstream reaction is co-localized with the aforementioned unfavourable reaction, it can quickly convert the product of the 1st reaction to another compound. This would act to minimize the concentration of the product [B] in the unfavourable reaction and give rise to a situation where the concentration of substrate is much greater than the concentration of product (i.e., [A] >> [B]). In this situation, the expression $RT \ln ([B]/[A])$ may become a large negative value that overcomes the positive standard Gibbs free energy of formation and leads to an overall negative Gibbs free energy of reaction.

To identify reactions that may undergo the above phenomenon, Chris Sanford developed a simplified methodology leveraging Gibbs free energy values of reactions [101]. Unfavourable reactions will have $K_{eq}$ values less than 1, corresponding to a $\Delta G$ value that aligns with a description of the 1st reaction in the aforementioned scenario. Favourable reactions have $K_{eq}$ values greater than 1, resulting in negative $\Delta G$ values that aligns with a description of the 2nd reaction in the aforementioned scenario. By pairing a low $K_{eq}$ reaction (which forms a thermodynamic reaction barrier in a pathway) with higher $K_{eq}$ reactions, reaction pairings to alleviate thermodynamically barriers can be identified using $\Delta G$ values. Given the Gibbs free energy of reaction equation, the magnitude of the $\Delta G$ value correlates with $K_{eq}$ and facilitates a comparison as to whether a reaction is more energetically favourable than another reaction. Reaction pairings where a positive $\Delta G$ is associated with the first reaction (i.e., is energetically unfavourable), and a second reaction that has a negative $\Delta G$ with a greater absolute magnitude than the $\Delta G$ of the first reaction would satisfy the aforementioned criteria. Given this equation, the magnitude of the $\Delta G$ value directly correlates with $K_{eq}$ and facilitates a comparison as to whether a reaction is more energetically favourable than another reaction. While it’s important to acknowledge that this methodology does not factor in rates of the reverse reactions and how biochemical systems do not operate at equilibrium (equilibrium only occurs in conditions where the concentrations of substrate and product remain unchanged over time), it is still useful as an initial exploratory step to better understand metabolic channelling.

To predict thermodynamically favourable pairs where channelling has the capability to improve reaction efficiency, Gibbs free energy values ($\Delta G$) were obtained from two independent
sources, an experimentally derived dataset (subsequently referred to as NIST) [104] and a predicted dataset obtained by the Mavrovouniotis group-contribution method (subsequently referred to as MGC) [105]. The MGC method is based on the principle where different chemical groups have certain free energies, and the free energy of any compound equalling the sum of the free energies of all chemical groups in the compound (Figure 6).

<table>
<thead>
<tr>
<th>Group label</th>
<th>Chemical Group</th>
<th>Gibbs Free Energy Contribution (kJ/mol)</th>
<th>Group Count</th>
<th>Total Gibbs Free Energy Contribution (kJ/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>N/A</td>
<td>+86.1</td>
<td>1</td>
<td>+86.1</td>
</tr>
<tr>
<td>1</td>
<td>-COO$^{3-}$</td>
<td>+42.2</td>
<td>2</td>
<td>+84.4</td>
</tr>
<tr>
<td>2</td>
<td>-CH=</td>
<td>-389.3</td>
<td>1</td>
<td>-389.3</td>
</tr>
<tr>
<td>3</td>
<td>-CH$_3$</td>
<td>-60.3</td>
<td>1</td>
<td>-60.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>-279.1 kJ/mol</strong></td>
</tr>
</tbody>
</table>

**Figure 6. Application of the Mavrovouniotis group-contribution method to estimate Gibbs Free Energy value.**

Crotonate is divided into 4 separate groups. The sum of the free energy values of each group and with a residual free energy value (due to the molecule being part of an open-chain) closely matches the experimentally-determined free energy value of crotonate [106].

957 unique pairs of reactions were identified after integrating free energy values from the MGC method with the NIST dataset in which: 1) a product from the first reaction is a substrate for the second reaction; (2) the first reaction is thermodynamically unfavourable ($\Delta G > 0$); and, (3) the second reaction is thermodynamically favourable such that the net $\Delta G$ of both reactions is greater than zero (i.e., the reaction pair overall is thermodynamically favourable and therefore a putative candidate for channelling) (Appendix B, Supplemental Table 1). To see if these pairs were dominated by a limited range of substrates, a random sampling test was performed where 100 instances of 957 unique pairs were selected from the KEGG database and the number of unique intermediate compounds was calculated from the selected unique pairs. A statistically significant difference between the number of intermediate compounds from the random sampling and the predicted thermodynamic-alleviating list would indicate that the enzyme pair predictions were not selected by chance. Enzyme pairs in the prediction list were found to share significantly
fewer intermediate compounds than by chance (182 vs 228 expected, standard deviation = 9.56, p-value < 1 E \(-5\), independent T-test), with many of these compounds involved in multiple reaction pairings. For example, chorismate is involved in 15 reaction pairs predicted to be thermodynamically favourable. Chorismate represents a critical branch point metabolite, which is a region in the metabolic network where multiple reactions converge at a single metabolite. It links indole and tryptophan metabolism, as well as the ubiquinone, folate and alkaloid biosynthesis pathways, and its use is tightly regulated to ensure the proper balance of amino acids and secondary metabolite production is maintained [115].

To prioritize pairs that are likely of most interest for downstream industrial applications, pairs were ranked on the basis of the product of the free energy values. This scoring method enriches for the biggest differences in free energy values, which is predicted to yield the greatest increase in kinetic efficiency (upon alleviating thermodynamic reaction barriers). The top 10 unique metabolites that give the largest difference in free energy as determined from each thermodynamic dataset (MGC and NIST) are shown in Tables 3 and 4 respectively. 12 of the 19 (63%) distinct intermediate compounds were found to each be involved in ten or more reaction pairs (compared to 20% of all KEGG metabolites). These findings are consistent with previous studies suggesting that reaction pairs with large differences in free energy are typically found at points the metabolic pathway where multiple reactions converge, and were postulated to regulate metabolic flux [116]. For example, glyoxylate is a ubiquitous compound associated with purine, caffeine and amino acid biosynthesis as well as the glyoxylate cycle in plants and microbes. Glyoxylate is predicted to be a candidate for channeling in 50 reactions in both the NIST and MGC prediction dataset. A second compound to highlight is phosphoenolpyruvic acid, which is involved in the critical glycolysis and gluconeogenesis pathways [117]. It is one of the select intermediates amongst the highest scoring predicted potential channelling pairs derived from both the NIST and MGC datasets and is associated with 36 and 23 predicted channelling pairs across the NIST and MGC datasets respectively.

Given that these reactions are found at locations in the metabolic network where multiple reactions converge to the same intermediate metabolite, one can hypothesize that flow can be regulated by localizing biochemically subsequent enzymes to direct flow along a specific path. Large differences in free energy between reactions can prevent flow along multiple reactions paths, until a regulatory mechanism co-localizes enzymes catalyzing such reactions to promote
flow under specific situation(s). The potential for these metabolites to occur at these convergence points will be explored in a later section. In summary, 957 unique pairs of enzymes were identified to have the potential for alleviating thermodynamic bottlenecks in metabolic pathways. Experimentally coupling enzyme pairs in this list, particularly pairs which occur at metabolic branch points, may be useful when engineering pathways in hosts which express enzymes catalyzing competing reactions (to limit the diversion of intermediate metabolites to competing pathways).

Table 3. The ten highest scoring predicted channelling pairs based on free energy values derived from the Mavrovouniotis group-contribution method.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Channeled Metabolite</th>
<th>Top Score</th>
<th>MGC and NIST reaction dataset</th>
<th>KEGG ID</th>
<th>KEGG reaction dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number of reactions</td>
<td>Percentage of reactions</td>
<td>Number of reactions</td>
<td>Percentage of reactions</td>
</tr>
<tr>
<td>1</td>
<td>Glyoxylate</td>
<td>$2.2 \times 10^{10}$</td>
<td>50</td>
<td>5.2</td>
<td>C00048</td>
</tr>
<tr>
<td>2</td>
<td>Glycerone phosphate</td>
<td>$8.4 \times 10^9$</td>
<td>40</td>
<td>4.2</td>
<td>C00111</td>
</tr>
<tr>
<td>3</td>
<td>Adenosine</td>
<td>$7.8 \times 10^9$</td>
<td>16</td>
<td>1.7</td>
<td>C00212</td>
</tr>
<tr>
<td>4</td>
<td>Hydroxypyruvate</td>
<td>$7.2 \times 10^9$</td>
<td>3</td>
<td>0.3</td>
<td>C00168</td>
</tr>
<tr>
<td>5</td>
<td>(s)-Lactaldehyde</td>
<td>$5.4 \times 10^9$</td>
<td>10</td>
<td>1.0</td>
<td>C00424</td>
</tr>
<tr>
<td>6</td>
<td>Acetaldehyde</td>
<td>$5.3 \times 10^9$</td>
<td>21</td>
<td>2.2</td>
<td>C00084</td>
</tr>
<tr>
<td>7</td>
<td>(2R)-2-Hydroxy-3- (phosphonooxy)-propanal</td>
<td>$5.2 \times 10^9$</td>
<td>20</td>
<td>2.1</td>
<td>C00118</td>
</tr>
<tr>
<td>8</td>
<td>5-Phospho-alpha-D-ribose 1-diphosphate</td>
<td>$4.2 \times 10^9$</td>
<td>35</td>
<td>2.4</td>
<td>C00119</td>
</tr>
<tr>
<td>9</td>
<td>Phosphoenol-Pyruvate</td>
<td>$4.1 \times 10^9$</td>
<td>23</td>
<td>3.8</td>
<td>C00074</td>
</tr>
<tr>
<td>10</td>
<td>D-Fructose 6-phosphate</td>
<td>$4.1 \times 10^9$</td>
<td>58</td>
<td>6.1</td>
<td>C00085</td>
</tr>
</tbody>
</table>
Table 4. The ten highest scoring predicted channelling pairs based on free energy values derived from experimentally determined free energy values of compounds.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Channeled Intermediate</th>
<th>Highest Score in Henry</th>
<th>Number of Reactions in Henry AND NIST</th>
<th>% of Reactions in Henry and NIST</th>
<th>KEGG ID</th>
<th>Number of Reactions in KEGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chorismate</td>
<td>$3.60 \times 10^{10}$</td>
<td>15</td>
<td>1.57%</td>
<td>c00251</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>Ubiquinone-8</td>
<td>$3.12 \times 10^{10}$</td>
<td>49</td>
<td>5.12%</td>
<td>c00399</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>2-C-Methyl-D-erythritol 2,4-cyclodiphosphate</td>
<td>$1.53 \times 10^{10}$</td>
<td>1</td>
<td>0.10%</td>
<td>c11453</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>L-Glutamine</td>
<td>$1.35 \times 10^{10}$</td>
<td>28</td>
<td>2.93%</td>
<td>c00064</td>
<td>41</td>
</tr>
<tr>
<td>5</td>
<td>all-trans-Octaprenyl diphosphate</td>
<td>$6.61 \times 10^{9}$</td>
<td>1</td>
<td>0.10%</td>
<td>c04146</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>(S)-Dihydroorotate</td>
<td>$5.08 \times 10^{9}$</td>
<td>2</td>
<td>0.21%</td>
<td>c00337</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>Phosphoenolpyruvate</td>
<td>$4.42 \times 10^{9}$</td>
<td>36</td>
<td>3.76%</td>
<td>c00074</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>sn-Glycerol 3-phosphate</td>
<td>$4.10 \times 10^{9}$</td>
<td>5</td>
<td>0.52%</td>
<td>c00093</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>(S)-Lactate</td>
<td>$4.10 \times 10^{9}$</td>
<td>2</td>
<td>0.21%</td>
<td>c00186</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>(S)-Malate</td>
<td>$4.10 \times 10^{9}$</td>
<td>2</td>
<td>0.21%</td>
<td>c00149</td>
<td>14</td>
</tr>
</tbody>
</table>

Next, a list of enzyme pairs was generated based on its potential to reduce cellular exposure of toxic compounds upon introduction of channelling. Exploring reactions that match this criteria has will be beneficial for industrial applications where the generation of toxic compounds from engineered pathways limit the production of a desired metabolite. If too much toxic metabolite is produced, the cell will die and effectively halt the production of product of interest. For example, early attempts to engineer a pathway to produce precursor metabolites for the anti-malarial drug, artemisinin, were not successful due to the build-up of toxic isoprenoid precursors [118]. ChemIDplus, a chemical database hosted by the National Library of Medicine (US), which contains toxicological data on 380,000 compounds as a source of toxic compounds, was leveraged in order to perform this analysis. Due to the lack of standardized intracellular toxicity information, analyses were performed based on the minimum reported lethal concentration in any organism. Using a LC50 cutoff of 100ppm and custom scripts evaluating the ChemIDplus data set (developed by Chris Sanford), 254 potentially toxic compounds that represent the intermediates of 4,858 pairs of reactions were identified (Appendix B, Supplemental Table 2). Amongst these are known examples of toxic intermediates that have previously been the target of experiments. For example, O-Acetyl-L-serine (OAS), a toxic intermediate highlighted in the channelling predictions, was predicted to be potentially channeled by 23 enzyme pairs. Previous studies have shown that increased levels of OAS are toxic to cells
and that over-expression of an exporter of OAS overcomes the toxicity [119]. The most prevalent toxic intermediates are formate and formaldehyde, which are intermediates in 655 and 404 pairs of reactions respectively (~22% of all reaction pairs). Both metabolites are generated during the metabolism of methanol and have previously been shown to illicit cytotoxic effects in rats and rabbits [120, 121].

Finally, a list of enzyme pairs was generated based on their potential benefit to limit cellular exposure of inhibitory compounds upon the introduction of channelling. Restricting the diffusion of inhibitory intermediates through channelling will reduce the likelihood of engineered systems disrupting other metabolic pathways in host organisms, which will maximize the output of the engineered system. Leveraging the BRENDA database as a source of inhibitory compounds and custom scripts developed by Chris Sanford to evaluate the database, a list of 343 compounds that represent the intermediate of 3,570 pairs of reactions was generated (Appendix B, Supplemental Table 3). Amongst these are known examples of inhibitory intermediates that have previously been the target of experiments. One such inhibitory intermediate identified in the dataset is 6-phospho-D-gluconate, an intermediate from the pentose phosphate pathway. The pentose phosphate pathway is vital for producing the sugars required to synthesize DNA and RNA: ribose and deoxyribose. In this pathway, 6-phospho-D-gluconate is converted to a number of metabolic precursors for ribulose-5-phosphate and glyceraldehyde-3-phosphate. At higher concentrations, it inhibits several enzymes, including glucose-6-phosphate isomerase, which catalyzes one of the earliest reactions in glycolysis pathway [122]. Among the sets of toxic and inhibitory compounds, there were several identified that have previously been associated with channelling, including ammonia, carbon monoxide and 5, 10-Methylenetetrahydrofolate [123-125]. Given the importance of these compounds in many reactions in metabolism, intermediates associated with channelled reactions may play regulatory roles within the overall metabolic network.

Overall 9,161 unique reaction pairs were predicted to potentially benefit from channelling through the removal of thermodynamic and / or toxic / inhibitory constraints, several of which have already been the subject of experimental investigations. This pool of reactions represents the total potential for engineering metabolic channelling into biochemical systems based on the three criteria used to identify channelling examples in this study. A natural follow-up question relates to the proportion of enzyme pairs in the prediction list that are already naturally
occurring. If only a small proportion of examples are found naturally, engineering metabolic channelling into systems of interest may represent a reasonable method to improve performance of the metabolic system. It is also important to note that attempts to survey natural examples of channelling will be limited, as relatively few proteins and genomes have been sequenced out of the available universe of proteins / genomes. Some mechanisms of effecting metabolic channelling in natural enzyme systems include gene fusions [81] and protein complexes [82], which are characterized by looking at decreased transient time (to reach steady state) and increased production of radiolabeled product in the presence of excess unlabeled metabolic intermediates (as described in Section 1.6.1). In the next section, I describe the systematic exploration of gene fusion and protein-interaction data, and their overlap with the channelling-benefitting prediction list, to understand its potential for use in metabolic engineering.

2.3.2 Channelling predictors are supported by gene fusion evidence and multifunctional enzymes

Aside from co-localizing enzymes through compartmentalization [88], metabolic channelling can also occur in protein complexes or multifunctional enzymes generated through gene fusion events. To date, several large scale protein-protein interaction (PPI) datasets have been generated that detail protein complexes associated with a range of organisms including E. coli, yeast and human [109-111]. A previous survey of PPIs derived from a variety of sources found an enrichment of direct physical interactions associated with pairs of enzymes that share metabolites compared to enzymes separated by several reaction steps [86]. Similar to an analysis performed by Chris Sanford, I compared the prediction list of 9,161 enzyme pair channelling candidates with the PPI datasets to identify the overlaps between the channelling candidates and PPI datasets and whether these overlaps were statistically significant (Figure 7a). Due to concerns over the relative quality of different PPI datasets, analyses were restricted to 3 high quality protein interaction datasets (see Methods). The E. coli interaction dataset only contained 6 proteins pairs annotated with Enzyme Commission (EC) numbers catalyzing biochemically sequential reactions. One EC pairing (2.7.4.6 and 2.7.7.7) overlapped with the inhibitory compound channelling prediction list, while no overlaps were observed with the positive scoring or toxic compound channelling predictions. Of the 27 human interacting protein pairs annotated with EC numbers catalyzing biochemically sequential reactions, four pairs overlapped with the
predicted channelling list described earlier. Two of the four pairs were found in the toxic compound channelling predictions, while the other two pairs overlapped with the inhibitory compound channelling predictions. Interestingly, no overlap was detected with the positive scoring predictions list. The *S. cerevisiae* interaction dataset contained 28 interacting pairs that catalyze biochemically sequential reactions. Only one of the 28 pairs, AICAR transformylase (EC:2.1.2.3) and adenylosuccinate lyase (EC:4.3.2.2), overlapped with the predicted channelling list.

To understand whether channelling may occur between enzymes within protein complexes that do not directly interact, the shortest pathlengths of predicted channelling pairs was compared to the shortest pathlengths of interacting proteins in the context of each organism’s interaction network. If channelling pairs were biologically more likely to occur between non-interacting proteins, the analysis would demonstrate that predicted channelling pairs have a shorter pathlength that is statistically-significant in comparison to the shortest pathlength of interacting proteins. Unlike previous studies [86, 126], no significant enrichment for shorter path lengths associated with consecutive enzymes as compared to non-consecutive enzymes was found (Figure 7a). When analyzing the shortest pathlengths of non-interacting channelling pairs, there was also no enrichment for shorter pathlengths when compared to the interacting protein pairs for the 3 interaction datasets. If enrichment for shorter pathlengths was identified, the reduction in pathlength increases the likelihood that channeling may occur through interaction with intermediaries. For example, a reduction of shortest pathlength from 3 to 2 would indicate that only a single intermediary may be necessary to bring two candidate enzymes in close spatial proximity. In all, only a limited number of protein complexes were found to overlap with predicted channelling reactions.

An alternative mechanism for whereby channelling can be introduced involves gene fusions, where two or more enzyme activities are combined within the same polypeptide chain. From Chris Sanford’s analysis, 5,448 enzymes are functionally annotated with multiple EC numbers (of the 118,000 enzymes listed in the SwissProt database [102]), implying that they catalyze multiple reactions. He identified subset of 3,963 enzymes catalyzed sequential biochemical reactions covering 144 unique EC pairings (Appendix B, Supplemental Table 4). The Prolinks database is an alternate resource which provides data on 4.6 million gene fusion events from 170 genomes, predicted using the Rosetta Stone approach where two of more
proteins encoded for separately in a genome also appear as fusions, either in related or distant organisms [103, 127]. Gene fusions identified with this approach are referred to as the Rosetta Stone dataset in this thesis. To investigate how many of the predicted channelling pairs overlap with gene fusions involving sequential reactions, I compared the list of predicted channelling, toxic intermediate and inhibitory intermediate enzyme pairs to both the SwissProt and Rosetta Stone datasets was performed (Figure 7b). The thermodynamic alleviating and inhibitory intermediate channelling predictions displayed significant overlap to the gene fusion datasets relative to the entire set of KEGG reactions (35 of 957 pairs and 37 of 3553 pairs respectively, at p-values of 1.37 E-21 and 2.69 E-9 respectively). When comparing the list of 35 pairs and 37 pairs, only 3 enzyme pairs overlapped across the 2 subsets of enzyme pairs (purine-nucleoside phosphorylase [EC:2.4.2.1] and thymidine phosphorylase [EC:2.4.2.4]; N-acylmannosamine kinase [EC:2.7.1.60] and N-acetylneuraminic acid lyase [EC:4.1.3.3]; dihydroneopterin aldolase [EC:4.1.2.25] and 2-amino-4-hydroxy-6-hydroxymethylidihydropteridine diphosphokinase [EC:2.7.6.3]). These enzyme pairs are found in the pyrimidine metabolism, amino/nucleotide sugar metabolism and folate biosynthesis respectively. While the pairs are not closely linked in the metabolic network, they metabolize central reactions and may represent essential reactions in their respective pathways (as depicted by KEGG representations). The toxic intermediate channelling predictions did not display any significant overlap, with only 13 pairs overlapping with the gene fusion datasets. These results suggest that gene fusions may be useful for channelling inhibitory intermediates and alleviating thermodynamic constraints, as well as introduce channelling into such reactions in bioengineering applications.
Figure 7. Comparisons of enzyme pair channelling predictions with natural channelling examples potentially found in gene fusions and protein complexes.

(a) Shortest pathlength comparisons for E. coli, yeast and human protein-protein interaction networks. Interaction networks of all proteins within the protein-protein interaction network (“All proteins”), all enzymes within the protein-protein interaction network (“All enzymes”), all consecutive biochemical reactions, and all protein-protein interactions containing enzymes pairs predicted to alleviate thermodynamic constraints (“All thermodynamic channelling reactions”) were subject to shortest pathlength analyses. Z-tests were used to calculate the significance in differences of average shortest pathlengths between networks of the same organism (●●● indicates a statistically significance difference in average shortest pathlength as compared to the average shortest pathlength for all proteins).

(b) Overlaps of gene fusions (SwissProt and Rosetta Stone) with enzyme pairs predicted to benefit from channelling by alleviating thermodynamic constraints and sequestering toxic or inhibitory intermediates. The overlap between each predicted dataset and the union of the gene fusion datasets was tested for significance using a hypergeometric test. The overlaps for the thermodynamic and inhibitor datasets were found to be significant, while the overlap of the toxic dataset was not significant.
Figure 8. Heatmap of gene fusion distributions across various taxonomic groups.

Gene fusions that catalyze biochemically consecutive reactions, represented by pairs of EC numbers, were identified from the SwissProt database. Phyla where the specified gene fusion was identified are marked by yellow in the heat map. Black denotes the lack of the specified fusion in the corresponding phylum. Taxa are organised according to kingdom, with gene fusions clustered based on their similarity of fusion profiles (i.e., the presence of specific gene fusions in specific phyla) to each other using Cluster 3.0. Enzymes of notable clusters in the heatmap are highlighted in adjacent boxes and linked to the corresponding cluster by a dotted line. The heatmap was visualized using Java TreeView [128].
In addition to exploring potential overlap with channelling predictions, the gene fusion datasets provide an opportunity to explore the phylogenetic distribution of metabolic channelling. Potential insights that can be gained from this analysis include the adaptation of organisms through the evolution of certain metabolic pathways, and whether certain adaptations arise independently in unrelated phyla. These insights provide an understanding of how channelling is implemented across different organisms (i.e., select reactions in specific phyla or widespread channelling of one reaction) to increase the success rate of engineering channelling.

112 gene fusions identified in the SwissProt database were clustered based on their presence across 33 phyla. Fusions were classified into 10 major clusters according to the profile of their presence across phyla, with enzymes (and their corresponding EC number) present within select clusters highlighted in adjacent boxes (Figure 8). A subset of clusters contain fusions where EC pairings are associated with only 1 or 2 phyla. For example, Cluster 2 is largely restricted to proteobacteria, and contains fusions with six EC numbers associated with phenylalanine, tyrosine and tryptophan biosynthesis. Since all six EC numbers involve reactions occurring downstream of chorismate (a precursor to the synthesis of phenylalanine, tyrosine, and tryptophan), one can hypothesize that a specific synthesis of a specific amino acid can be promoted through the expression of select fusions that control the flux of chorismate and downstream intermediates towards the specified amino acid. The ability to control metabolic flux in this manner likely signifies tight expression regulation of the fusions in Cluster 2 to ensure an appropriate balance of amino acids levels is maintained.

Fusions in Cluster 6 is largely represented by Streptophytes (green plants) and includes several enzymes that perform sugar interconversions linked to the synthesis of components of the plant cell wall. For example, the fusion of fucokinase (EC:2.7.1.52), which interconverts between 6-deoxy-galactose and fucose-1-phosphate, with fucose-1-phosphate guanylyltransferase (EC:2.7.7.30), which interconverts between fucose-1-phosphate and GDP-fucose, may help enhance synthesis of xyloglucan, an important cell wall precursor [129]. Other clusters contain EC pairings found across a wide range of phyla. For example, Cluster 4 contains an EC pairing in the folate synthesis pathway found in all but 4 phyla analyzed. Given the intermediate generated (5, 10-Methenyl-THF) is utilized as a substrate by a number of other enzymes, the fusion of these enzymes may prioritize the production of folate, a key precursor for nucleic acid synthesis. Overall, the phylogenetic distribution of gene fusions indicates that
channelling between enzyme pairs is more commonly found in specific phyla and that these fusions are generally involved in non-essential pathways. Such fusions may represent exploratory adaptations with the goal of developing competitive advantages in their respective environments rather than necessary adaptations to fundamental processes.

2.3.3 Examining channelling predictions in the context of the global metabolic network

During the gene fusion analysis, several instances of metabolic channelling were identified that may have a role in regulating metabolic flux through pathway redirection. Given that regulation of flux may happen at branch points in a metabolic pathway, a question that arises pertains to how frequently each class of channelling candidates occurs at metabolic branch points. Understanding the prevalence of potential channelling reactions at metabolic branch points can highlight the feasibility of using channelling for pathway redirection. To evaluate this question in his initial analysis, Chris Sanford defined branch points were defined based on 2 approaches. The first approach is based on KEGG pathway memberships, which define branch points based on manually annotated pathways from model organisms (e.g., human, yeast, \textit{E. coli}). Although a large number of branch points can be identified in this manner, physiologically relevant pathway connections in other species may be missing, leading to the oversight of potential branch points. The second approach complements the first approach by leveraging a global metabolic network and identifying intermediate metabolites which are associated with multiple reactions in the network. The list of branch points generated through both approaches were used to evaluate whether the level of overlap between the enzyme pairs in the prediction list and enzyme pairs at branch points was significant. Evaluating this overlap relies on the assumption that the number of canonical KEGG-defined biochemical pathways associated with a metabolite is correlated to the likelihood that that the metabolite is involved with an important branch point. Enrichment of channelled intermediates at branch points would indicate that metabolic channelling is a mechanism to regulate / direct metabolic flux across the metabolic network.

To understand whether enzyme pairs in the prediction list are enriched at branch points, Chris determined the number of KEGG defined pathways associated with an intermediate
metabolite in the prediction list (Figure 9a). Leveraging the prediction list described earlier, I found that intermediate metabolites associated with free energy predictions demonstrated the largest enrichment for metabolites involved in three or more KEGG pathways with a ~2.5-fold increase compared with a control group of intermediates of all consecutive enzyme pairs (p < 10^{-10}, hypergeometric test). Enzyme pair predictions based on limiting other criteria also showed significant increases for metabolites involved in three or more KEGG pathways (toxic: p < 10^{-5}, inhibitory: p < 10^{-5}, hypergeometric test). Given the significance in overlap between the SwissProt and Rosetta stone gene fusion datasets and the predictions list, the enrichment at branch points for the presence of intermediate metabolites in gene fusion enzyme pairs was explored (Figure 9a). The enrichment of intermediate metabolites in three or more KEGG pathways was found to be significant for the Rosetta stone dataset (p < 10^{-10}, hypergeometric test). Enrichment was found for fusions in the SwissProt dataset, albeit at a non-significant level. This finding indicates that gene fusions can be found, but not preferred, at branch points.

Another analysis which Chris utilized to evaluate the enrichment of enzyme pairs in the prediction list at branch points involves the prevalence of branch point metabolites as channelled intermediates in the enzyme pair prediction list. Figure 9b shows the normalized frequency of metabolites as a function of their connectivities for each class of channelling predictions described earlier in this thesis. As with the KEGG pathway memberships, metabolites associated with alleviating thermodynamic barriers showed the greatest correlation to branching factor with a 5-fold enrichment for the most highly connected metabolites (21+ reactions, p < 10^{-10} hypergeometric test). Metabolites channeled in enzyme pair predictions isolating toxic and inhibitory intermediates were also significantly enriched for highly connected compounds. Together with the KEGG pathway analysis, these analyses demonstrate a clear relationship between branch point metabolites and the various classes of channelling-benefitting predictions spanning such metabolites. Intermediates in channelling predictions that alleviate thermodynamic constraints had the strongest fold enrichments for both their prevalence in KEGG pathways and their prevalence at metabolic network ranch points. This suggests that branch points may represent attractive locations for the evolution or engineering of regulatory mechanisms in the metabolic network (e.g., through allosteric interactions, feedback mechanisms, or regulation of expression), as the presence of thermodynamic barriers around
certain intermediate metabolites may be leveraged to control metabolic flow through specific reactions as needed (e.g., starvation of required metabolites).

Figure 9. Network representation of enzyme-substrate relationships.

To understand any potential enrichment of enzyme pairs predicted to benefit from channelling at metabolic branch points, the fold enrichment of different enzyme pair datasets as compared to the set of all consecutive enzymes was evaluated on 2 bases.

(a) Reaction pair membership of channeled intermediates and (b) pathway membership of channeled intermediates were evaluated for channelling predictions and gene fusion datasets and compared against a list of consecutive reactions from the KEGG database. “Consecutive” refers to the complete set of biochemically consecutive enzyme pairs, whereas the remaining data sets contain enzyme pair
predictions: “Positive” pairs were predicted by their Gibbs free energy values, “Rosetta” pairs are gene fusions predicted using the Rosetta stone method and found in the Prolinks database, “SwissProt” pairs are multifunctional enzymes, “Toxic” pairs are enzymes acting on a common toxic metabolite and “Inhibitory” pairs are enzymes which both act on a known enzymatic inhibitor. Statistical significance based on a hypergeometric test is indicated with dots: p-value < 10^{-2}(•), < 10^{-5}(••), < 10^{-10}(•••).

(c) A segment of the network is displayed with grey squares representing enzymes, diamond nodes represent repeated compounds in the network and circle nodes present non-repeated compounds. The colour of the compounds represents the biological process in which they are involved in. Blue edges represent reactions found in gene fusions, while red edges indicate reactions predicted to overcome thermodynamic barriers when channeled. Purple edges are reactions that have been found in gene fusions and are part of the predicted thermodynamic channelling pair dataset.

(d) Select channelling clusters in the fatty acid synthesis pathway and the shikimate synthesis pathway within the global metabolic network are highlighted which contain multiple gene fusion and/or channelling prediction reactions.

2.3.4 Visualizing biochemical pathway complexity

While these results clearly demonstrate a relationship between channelling potential and metabolic branch points, as noted earlier, KEGG pathways represent artificial constructs that do not capture the full potential for metabolites to be involved in multiple reactions. Recognizing this limitation, previous studies have constructed metabolic networks on the basis of shared substrates, enabling the discovery of potentially new pathways [130-132]. To complement the previous analysis, Chris Sanford developed a network approach was adopted to derive a more accurate representation of metabolite branch points. From KEGG reaction data, he constructed a bi-partite graph that contains 6,739 nodes (2,896 enzymes and 3,843 compounds) linked by 6,055 edges in 1,076 connected components and visualized the network using Cytoscape [94] (Figure 9c). The average shortest pathlength of the network is 29.7 and the average number of neighbours is 1.8 across the entire network. This representation reveals the breadth to which enzymes and substrates / products participate in reactions across the metabolic network. This view enables quick identification of enzyme hubs involved with the largest number of reactions, offsetting the weakness of KEGG pathway visualizations that are limited to enzymes only in a specific metabolic pathway. Although this view the direct measurement of the number of potential routes available through a metabolite, it may also introduce a number of false positives since the depicted reactions may only be present in a limited set of organisms. A compound that is processed differently in different organisms may have multiple associated reactions, even though it may only be associated with a subset of these reactions in any one organism. In addition to a single large connected component consisting of 1,305 nodes (615 enzymes and 690
metabolites) connected by 1,437 edges, the network contains an additional 1,075 subnetworks ranging in size from 2 to 130 nodes. This representation splits fundamental biochemical processes necessary for life from other secondary processes and provides an understanding regarding the prevalence of gene fusions and potential channelling-benefitting reactions in the two groups.

With Chris’ bi-partite network representation as the foundation, I integrated the updated predicted channelling list into the representation to identify whether certain pathways or regions of the metabolic network are associated with a high concentration of potential channelling enzyme pairs. Further exploration of the visualization highlighted how channelling predictions are not evenly distributed across pathways, but rather appear to form a backbone of ‘channelling routes’. Among the more notable examples are two such pathways involving at least five intermediate metabolites. The longest identified involved 7 biochemically sequential enzymes that channel 6 distinct intermediates in the fatty acid biosynthesis pathway – a finding that is not surprising given the results of the earlier phylogenetic analysis (Figure 9d). In the first superpathway, all reactions are found in the fatty acid synthase protein complex that is responsible for extending the length of fatty acid side chains to a desired level [133]. The second superpathway involves enzymes in the phenylalanine, tyrosine and tryptophan biosynthesis pathways (Figure 9e). The initial gene fusion involving dehydroquinate synthase (EC:4.2.3.4) and dehydroquinate dehydratase (EC:4.2.1.10) may represent a mechanism to overcome natural metabolic flow based on favourable thermodynamics through the enzyme pair of dehydroquinate synthase and quinate dehydrogenase (EC: 1.1.1.24). This energetically favourable route would act to direct metabolites away from the production of the three amino acids. Subsequent gene fusions channel intermediates towards chorismate production, where it can be converted into a number of different metabolites. The gene fusion event between anthranilate synthase (EC: 4.1.3.27) and 4-amino-4-deoxychorismate synthase (EC:2.6.1.85) acts to direct metabolites away from amino acid synthesis towards folate biosynthesis. The enzyme pair, chorismate mutase (EC:5.4.99.5) and 4-amino-4-deoxychorismate synthase, predicted to alleviate a thermodynamic barrier, will act to channel chorismate towards the synthesis of tryptophan. Finally, another thermodynamic barrier alleviating reaction pair, isochorismatase (EC:3.3.2.1) and isochorismate synthase (EC:5.4.4.2, formerly 5.4.99.6), will lead to the channelling of chorismate towards the production of compounds in the KEGG pathway which assist in ferric iron acquisition. Beyond
the long-connected routes, there are many additional smaller routes which are involved in a
diverse range of biochemical processes, including purine metabolism, biotin metabolism, folate
biosynthesis, porphyrin and chlorophyll synthesis and amino sugar synthesis. These smaller
routes suggest highways of metabolic flux may exist throughout the metabolic network to route
the conversion of substrates to a desired end product through a complex series of consecutive
reactions.

While most of the smaller channelling routes described above are associated with gene
fusion channelling predictions, other identified routes were also found to be associated with
enzyme pairs predicted to alleviate thermodynamic barriers and involve three or four metabolic
intermediates. Given that these predictions are based on the occurrence of intermediates with
higher free energies, one should not expect to identify such chains of consecutive and
overlapping free energy predictions. The first reaction in a reaction pair is an energetically
unfavourable reaction followed by a second reaction that is an energetically favourable reaction
with a higher absolute free energy, it is impossible to have a third reaction in a channelling chain
as the requirement for the 2nd reaction in the chain of 3 to be an energetically unfavourable
reaction. Since the 2nd reaction in this scenario will always be an energetically favourable
reaction, free energy chains will be limited to a consecutive chain of 2 reactions at a maximum.

To understand why channelling routes with consecutive gene fusions / channelling-
benefitting enzyme pairs may occur, I examined the pentose phosphate pathway in more detail.
This pathway had an increased prevalence of potential channelling-benefitting reactions through
gene fusions and thermodynamic barrier-alleviating enzyme pairs, including a chain of four
intermediate metabolites linked through reactions involving thermodynamic barrier-alleviating
enzyme pairs. To understand why this and other channelling routes may exist, I leveraged a
KEGG pathway depiction as a template for mapping gene fusions and enzymes from the
predictions list to see whether nearby reactions are influencing the need for channelling routes
(Figure 10a). This depiction illustrates the complexity inherent in attempting to visualize these
relationships. A majority of free energy predictions involve fructose-6-phosphate and
glyceraldehyde-3-phosphate, which represent highly connected branch points within the
pathway. This is not surprising given the number of pathways these metabolites also appear in,
including carbon fixation, glycolysis, amino acid synthesis and synthesis of secondary
metabolites. The number of enzymes pairs from the predictions list around the same area of the
metabolic network poses a challenge for representing both the directionality and the number of channelling predictions in an easy to interpret manner.

Despite this challenge, this depiction demonstrates the interconnectivity and complexities that may exist within a pathway, and between pathways, that a simplistic mapping onto a KEGG defined pathway will fail to represent. There are various examples in which characterized enzyme fusions and enzyme pairs predicted to alleviate thermodynamic barriers when channelled overlap. These examples provide a potential rational for the existence of the fusion. For example, coupling the thermodynamically favoured channeled reaction of ribose-phosphate pyrophosphokinase (EC:2.7.6.1) and transketolase (EC:2.2.1.1) with the gene fusion of transketolase and transaldolase (EC:2.2.1.2) can lead to the generation of fructose-6-phosphate, which is found at a major metabolic branch point. Without this fusion, the initial substrate D-ribulose-5-phosphate can be converted into a number of different metabolites across the pentose phosphate pathway.

Further evaluation of the pentose phosphate pathway depiction reveals how gene fusions can overcome a thermodynamically favourable reaction and convert a substrate into a different product. One example involves the enzymes phosphoexoisomerase (EC: 5.3.1.8) and glucose-6-phosphate isomerase (EC:5.3.1.9) which does not a predicted energetically favourable reaction route. In a large number of paired reactions in Figure 10a, glucose-6-phosphate isomerase would undergo a thermodynamically favourable reaction when coupled with a number of other enzymes. In the absence of channelling, this would pose a major thermodynamic barrier to channel intermediates towards phosphoexoisomerase. However, gene fusions identified in the SwissProt database with phosphoexoisomerase (EC: 5.3.1.8) and glucose-6-phosphate isomerase (EC:5.3.1.9) and three fusions identified using the Rosetta stone method [phosphoexoisomerase (EC: 5.3.1.8) and phosphomannose mutase (EC: 5.4.2.8); phosphomannose mutase (EC: 5.4.2.8) and mannose-1-phosphate guanylyltransferase (EC: 2.7.7.22); mannose-1-phosphate guanylyltransferase (EC: 2.7.7.22) and GDP-mannose 4,6-dehydratase (EC: 4.2.1.47) ] may allow for channelling of fructose-6-phosphate into the amino sugar and nucleotide sugar metabolism pathway instead of competing pathways such as the pentose phosphate pathway and glycolysis. Channelling of metabolites in such a manner may be useful in high growth scenarios for the production of glycoconjugates, which are important components of fungal and bacterial cell walls [134].
Figure 10. Representations of channelling predictions and gene fusions in the pentose phosphate pathway illustrating the complexities in a section of the global metabolic network.

(a) An initial KEGG-based representation was created to demonstrate the complexities found in the pentose phosphate pathway. Gene fusions are highlighted in blue, while enzyme pairs predicted to benefit from channelling are displayed by red chevrons. Blackened regions in the chevrons represent the starting, channelled, and final metabolite in the enzyme pair. Maroon-coloured chevrons represent channelling pairs highlighted in the results section. Yellow-boxed metabolites represent metabolites repeated in the diagram.

(b) By applying a Hyperscape visualization [95], complex relationships between pathways are made clearer. The Hyperscape representation employs different coloured arcs in sets of four to illustrate channelling pairs, while gene fusions are highlighted in blue hyperedges. Pink circles represent enzymes while gold squares represent metabolites in the network. Red hyperedges represent channelling pairs.
highlighted in the results section. The original KEGG representation is displayed as an inset beside the Hyperscape representation.

In an attempt to overcome visualization limitations of this KEGG pathway-based representation, use of the Hyperscape tool, recently developed by members of the Parkinson lab, was explored as a solution (Figure 10b). Hyperscape is a network visualization tool that utilizes hypergraphs, which have been shown to more effectively represent biological relationships within network representations [95]. Similar to the KEGG pathway-based representation, enzymes and metabolites are depicted as differently shaped nodes. A key difference is the use of different edge types to display multiple levels of information and enable quicker and simpler consumption of information. Thin, straight edges are used to depict connections between enzymes and metabolites, much like KEGG pathway-based representations. By not layering complex information on this edge (as seen in other visualization methods), features of the naturally occurring pathway, such as reaction directionality, can be clearly seen. The second type of edge is found in sets of 4 coloured, directional arcs that highlight reactions that are predicted to benefit from metabolic channelling. Channelling predictions are represented using a series of 4 arrowed arcs that go from the initial substrate to the 1st enzyme, the 1st enzyme to the intermediate substrate, the intermediate substrate to the 2nd enzyme and finally, the 2nd enzyme to the product. This allows for delineation of channelling routes where certain reactions (i.e., reaction pairs where β-D-fructose-6-phosphate or PRPP are the intermediate metabolite) are involved in a number of potential channelling pairs. The final type of edge is a hyperedge, which is designed to connect to any number of nodes and can highlight key features in the network. In the redesigned visualization of the pentose phosphate pathway, enzymes found in gene fusions and select enzyme pairs predicted to benefit from channelling are grouped using a different coloured hyperedges.

Together, the features from this new visualization provides improvements over previous visualizations (bipartite networks, KEGG pathway representation with chevrons) by seamlessly integrating multiple layers of information to facilitate the generation of key insights in areas within a metabolic network. This representation can be expanded to other areas of the network to rationalize the presence of key regulatory mechanisms or genomic landmarks (i.e., gene fusions).
2.4 Discussion

Metabolic channelling has the potential to provide several benefits, including: 1) alleviation of thermodynamic constraints; 2) minimizing exposure to toxic or inhibitory intermediates; 3) preventing loss of volatile substrates; and 4) directing substrates through specific pathways to produce products of interest [135, 136]. Channelling has been exploited by a wide range of organisms to improve the function of metabolic processes (e.g., the citric acid cycle and photosynthesis). In addition to its importance in various organisms, channelling can also be an intriguing solution for solving metabolic engineering problems [137-139]. However, while several studies have focused on specific pairs of reactions or pathway segments that are predicted to benefit from channelling [2, 140], few studies have attempted to capture the full extent to which nature has exploited channelling, or identify those pathways that might be successfully targeted in metabolic engineering applications. Here, a systematic set of analysis were performed to understand the potential for channelling, as well as the mechanisms by which channelling has been effected in nature.

Based on three separate criteria, 9,161 unique pairs of enzymes (distinct EC numbers) were predicted to offer benefits through channelling. Amongst these are 957 pairs predicted to alleviate thermodynamic constraints (thermodynamic pairs); 4,858 pairs that limit exposure to toxic intermediates (toxic pairs); and 3,570 that limit exposure to inhibitory intermediates (inhibitory pairs). The prediction list of reaction pairs relieving thermodynamic constraints was generated using experimentally defined free energy values and free energy values calculated through the ‘group contribution method’, where a compound is broken up into different chemical groups that are each assigned free energy values. The free energy value of a compound is then determined as the sum of the free energies of all chemical groups in the compound. Together, both datasets provided free energy values for 2,083 of 9,750 reactions (based on the KEGG database). Consequently, the 957 reaction pairs identified here are expected to be only a subset of all pairs predicted to benefit from channelling. Further, it should be appreciated that these datasets are also limited by the physiological conditions (e.g., pH and temperature) under which the free energy values were obtained. Differences in pH and temperature affect the free energy differences in reactions, meaning reactions that previously did not match the channelling criteria may undergo channelling at non-physiological combinations of pH and temperature.
Nonetheless, these subsets provide useful frameworks that may be exploited for metabolic engineering applications. For example, pathway segments being engineered to produce a compound of interest is limited due to a thermodynamic bottleneck that slows production or the production of a toxic or inhibitory intermediate that kills hosts cells and limit production yield. If the pathway components that are responsible for the thermodynamic bottleneck or toxic / inhibitory compound production are present on the prediction list, introducing channelling between such components can be a potential solution for alleviating pathway roadblocks. In turn, this will allow for production of metabolites that would not be previously possible. The presented predicted channelling set represents a large increase in predicted channelling pairs, as a previous study on metabolic channelling identified only 4,200 pairs of enzymes that may be undergoing channelling (~4,200) [126]. This larger list was created despite having fewer metabolites being termed as currency metabolites as compared to the previous study, which would reduce the number of identifiable enzyme pairs predicted to benefit from channeling as the pool of metabolites that can be channeled is reduced.

Upon generation of the prediction list, analyses performed to identify the number of enzyme pairings which may already be channeled in nature and understand the mechanisms by which channelling is achieved. The overlap would provide insight as to the potential of leveraging channelling when engineering systems. If there is already a significant overlap of currently channeled enzyme systems and the prediction list, it would indicate that organisms in nature have likely sampled a large proportion of possible channelling combinations and that the potential to engineer novel channelling systems will be limited. Conversely, a lack of overlap may indicate that nature has not sampled a lot of potential channelling combinations, which highlights a large potential to create new channelling combinations through metabolic engineering. Two potential mechanisms for channelling were explored: gene fusions and protein complexes [141, 142]. Since the proteins encoded by the fusions are linked in close proximity, there is a stronger likelihood that metabolites could be channelled for subsequent processing if the fusion catalyzes biochemically sequential reactions. Comparison between datasets revealed a significant enrichment for thermodynamic and inhibitory pairs within the gene fusion datasets. No such enrichment was observed for toxic pairs. A survey of existing studies to identify examples where gene fusions have been used to successfully introduce metabolic channelling yielded a limited number of studies. Such investigations revealed only modest improvements in
reaction kinetics [89, 91] and may be insufficient to effectively channel intermediate metabolites. This may explain the lack of overlap between the prediction list that channel toxic intermediates and gene fusions, as improvements from gene fusions may not be enough to lower levels of toxic intermediates to non-fatal levels. Comparisons between the prediction list and protein interaction datasets (representing protein complexes) did not reveal any significant overlaps, indicating that protein complexes are not commonly used to alleviate thermodynamic barriers, or isolate toxic / inhibitory intermediates. Further discussion of this analysis is found in the Section 2.4.1.

A third mechanism, encapsulation within micro-compartments is beginning to receive considerable attention. As highlighted in a number of studies, compartmentalization may offer a more efficient route to channel intermediates [6, 31, 143-145]. To date, such compartments have been associated with two major metabolic processes: carbon fixation (carboxysomes) [23]; and aldehyde oxidation (e.g., 1,2-propanediol, ethanol and ethanolamine utilization compartments) [6, 31, 143]. Only one enzyme pair from the enzymes previously shown to be associated micro-compartments overlapped with the predicted beneficial channelling pairs, alcohol dehydrogenase (EC: 1.1.1.1) and aldehyde dehydrogenase (EC: 1.2.1.5). Furthermore, the ethanolamine utilization compartment is thought to restrict the loss of the otherwise volatile aldehyde intermediate [6] – a channelling consideration, that due to lack of appropriate data, was not explored in this study. Nevertheless, micro-compartments represent a potential system for effecting channelling between enzyme pairs due to their ability to restrict metabolite flow out of the compartment and house multiple enzymes in close vicinity – features which are critical for introducing channelling between enzymes.

2.4.1 Extent of metabolic channelling in nature

From a taxonomic perspective, gene fusions involving reaction pairs were largely restricted to individual phyla representing taxon specific innovations. For example, within Streptophyta, several fusions were identified which involved enzymes responsible for the synthesis of plant cell wall precursors. However, a number of more conserved gene fusions were also identified, typically associated with core metabolic processes such as fatty acid synthesis. Previous studies have shown that taxon-specific metabolic processes tend to result in the generation of novel pathways that do not interfere with core, conserved pathways [130].
Consequently, innovations with enzymes peripheral to core metabolic processes are less prone to causing detrimental host fitness than enzymes involved in core, conserved pathways due to higher concentration of essential genes in core metabolic processes [146].

During these analyses, several fusions involving two reactions indirectly linked through a third, intermediate reaction were identified. For example, GTP diphosphokinase (EC: 2.7.6.5) is indirectly linked to guanosine-3’, 5’-bis(diphosphate) 3’- diphosphatase (EC:3.1.7.2) through exopolyphosphatase (EC: 3.6.1.11). Similarly, histidinol dehydrogenase (EC:1.1.1.23) is indirectly linked to histidinol-phosphate transaminase (EC:2.6.1.9) through 3.1.3.15. Such instances represent a potential mechanism through which metabolic flux may be controlled through expression of the enzyme catalyzing the intermediate reaction, particularly where substrate intermediates may be channeled into one of several pathways (Figure 11). The relatively small number of gene fusions involving channelling reactions identified here may be due to a relatively poor sampling of genomes. The caveat of leveraging gene fusions to introduce robust channelling is the limited benefit in reaction speed, as demonstrated by the fusion of β-galactosidase and galactose dehydrogenase which yielded only a 2-fold increase in reaction speed [89]. With the emergence of metagenomics, one may expect to determine if such paucity is due to sampling biases or the inefficient nature of gene fusion as a route for channelling (to be discussed in section 4.4).

![Diagram showing metabolic pathways and gene fusions](image1.png)
Figure 11. Controlling metabolic flux through suppressed enzyme expression in a pathway.

In this diagram, metabolite 1 is converted to metabolite 4 through a series of reactions catalyzed by enzymes A, B and C. Metabolite 2 can also be converted to other metabolites by competing reactions, such as metabolite 5 through enzyme D. By suppressing the expression of enzyme B, metabolite 3 is not generated in the pathway. As a result, metabolic flux from metabolite 2 to metabolite 4 is halted, and metabolite 2 may be converted into other metabolites by enzyme D in a competing pathway. Upon expression and co-localization / channelling of enzyme B with enzyme A, metabolite 3 is generated, enabling the production of metabolite 4 through the channelling of metabolites 2 and 3 (through enzymes B and C).

Beyond gene fusions, an alternative method by which enzymes may be brought together in close proximity to elicit channelling behaviours is through the formation of protein complexes. A classic example involves the fatty acid synthase complex, which employs a prosthetic group involving an acyl carrier protein that transports fatty acid intermediates to various active sites within the complex, enabling fatty acid chain elongation [133]. Previous studies attempting to predict metabolic channelling benefitting enzyme pairs through protein-protein interactions identified 1,759 and 2,508 pairs of physically interacting enzymes involved in consecutive reactions, for *E. coli* and yeast respectively [86, 147]. From analyses described in this thesis, 61 pairs of consecutive reactions were identified which directly interact across the three interaction datasets studied (*E. coli*, yeast, and human). This more limited set, in comparison to the aforementioned previous studies, is likely due to a larger exclusion of pairs of reactions linked through currency metabolites. For example, one study included reactions linked by ubiquitous metabolites such as NAD, pyruvate, ammonium and ATP, which was explicitly excluded in analyses in this thesis due to its prevalence in a large number of reactions and its subsequent impact in linking non-relevant reactions [86]. Of the 61 consecutive reactions, only six reaction pairs overlapped with the channelling-benefitting predictions list – three were from the inhibitor channelling dataset, two were from the toxic intermediate channelling dataset and one pairing predicted to alleviate thermodynamic constraints. The lack of overall between the predictions list and the PPI datasets may be interpreted several ways. Many of the initial predictions may simply be incorrect, resulting in a large number of false negative hits. PPI datasets are also generally poor at determining transient interactions. There have been several examples where transient complexes called metabolons were discovered to undergo channelling [139]. The transient complexes may reveal more potential channelling examples which cannot be captured by
traditional PPI experiments and analyses. Finally, many enzymes are located in biological membranes. PPI datasets typically lack the inclusion of membrane proteins as common methods have difficulty defining interactions in membranes [148, 149].

In addition to directly interacting, proteins that interact indirectly (i.e., through an accessory protein) offer an additional route for channelling. To explore this further, the shortest path lengths was calculated between the predicted channelling enzyme pairs, as well as enzymes in the protein interaction datasets. Analyses did not reveal any significant enrichment in shortened path lengths for either directly interacting or indirectly interacting enzyme pairs. This finding is in contrast to the previous studies which suggest that there was a significant enrichment in shortest path lengths in both an *E. coli* and yeast dataset [86]. However, as noted above, the previous study included many reactions that channel a currency metabolite, whereas the analyses described in this thesis deliberately did not involve currency metabolites. While the enzyme pair prediction methodology described in this thesis led to the identification of some non-physiologically relevant reaction pairs, removing currency metabolites significantly reduced the identification of additional non-physiologically relevant reaction pairs. Furthermore, to reduce the incidence rate of false positive assignments, analyses were performed on more limited sets of high quality protein complex data derived from high throughput proteomics studies as opposed to large scale interaction datasets derived from a variety of different sources and techniques [86].

### 2.4.2 Role of channelling in regulating metabolic flux

One of the potential benefits of channelling explored included the ability to regulate metabolic flux across so called metabolic *branch points*, where a metabolite may be directed into one of several downstream pathways. Through analyses of KEGG defined pathways and metabolic network connections, the predictions list was found to be enriched for reaction pairs spanning branch points. Previous attempts to channel metabolites in a specific manner across branch points have focused on the use of gene knockouts or over-expression [150-152]. For example, computational methods such as OptKnock and RobustKnock, have been used to predict reaction knockouts that optimize metabolite production by precluding flux in competing pathways [153, 154]. However, modelling errors can yield situations where a desired chemical
is not produced at all. An alternative approach to optimizing flux involves the overexpression of proteins immediately downstream of the branch point metabolite [155, 156]. For example, overexpression of chorismate mutase, which is downstream of the metabolic branch intermediate chorismate, favoured the production of metabolites for synthesizing tyrosine and phenylalanine [155]. One drawback to using knockouts is that resulting strains can demonstrate sub-optimal growth rates [157]. Furthermore, gene knockouts cannot be performed on essential genes. Channelling may offer an additional route to regulate branch point redirection. By taking into account thermodynamic constraints, one can predict where fusions across branch points are likely to be most effective. Analyses described in this thesis highlighted multiple reactions with low Gibbs free energy changes upstream of branch points, and high Gibbs free energy changes downstream, indicating the potential for regulation at this region in the metabolic network.

This principle has been used as the basis for designing biological sensors for specific metabolites. A study coupled glucose oxidase, which oxidizes glucose and generates hydrogen peroxide during the process, with hexokinase, an enzyme that converts glucose to glucose-6-phosphate using ATP [158]. Given hexokinase’s stronger affinity for glucose in the presence of ATP, a solution containing hexokinase and glucose oxidase can successfully sequester glucose to glucose oxidase, depending on the concentration of ATP present [158]. If the concentration of glucose rises above the concentration of ATP in solution, glucose will be processed by glucose oxidase to generate hydrogen peroxide and a detectable change in current [158]. By leveraging glucose’s affinity for the different enzymes (dependent on the concentration of ATP), a response curve for glucose oxidase was engineered over a much smaller concentration range to resemble a biological switch that can act as a biological logic gates in systems of interest [158]. This would allow an engineered system to be sensitive to minor changes in glucose or ATP concentration and enable a controlled response under specific conditions (i.e., certain glucose / ATP concentrations). However, the implementation of such fusions may not be straightforward. A recent study attempted to direct a branch point metabolite (carbamoyl phosphate) towards the production of pyrimidine through the fusion of enzymes CarB, which generates carbamoyl phosphate, and PyrB. However, this fusion resulted in no significant flux through PyrB as compared to another competing enzyme, ArgI [2]. Instead, channelling was achieved only through additional co-clustering of CarB and PyrB into “functional agglomerates”. This finding
highlighted the complexity of metabolic regulation due to factors such as stoichiometric counts and enzyme kinetic properties [93, 159].

Beyond branch points, the global network analyses also revealed the presence of longer “channelling routes”, incorporating three or more consecutive enzymes. These routes may enable directed conversion of metabolites from one area of the metabolic network to specified areas in another region of the network. In the process of traversing these routes, metabolic barriers in the form of toxic and / or inhibitory intermediate compounds or thermodynamic barriers could be overcome to simplify the conversion of various metabolites. For example, several studies explored a multi-gene fusion segment of the shikimate synthesis pathway highlighted earlier (Figure 3e) and identified decreased transit times and a novel regulation mechanism in the fusion as potential benefits of the fusion [160-162]. However, such fusions do not always lead to successful channelling; a three-enzyme metabolon localized inside bacteriophage P22 particles was unable to channel intermediates to create glucose-6-phosphate from an initial lactose substrate [68]. While combinations of enzyme knockout, overexpression and gene fusion may be a solution for controlling metabolic flow, subsequent studies will be required to fully understand the factors influencing metabolic flow and achieve the desired channelling effect.

A major issue highlighted in this chapter is the lack of suitable visualization tools to display the complexity inherent in metabolic networks. Many metabolic representations in literature are based off of KEGG-defined pathways, which group a list of EC numbers and metabolites. A major drawback to utilizing KEGG pathways as reference points is the presence of arbitrary pathway borders, meaning large routes connecting multiple KEGG pathways are not well visualized. To demonstrate the complexity in a small portion of the metabolic network, a chevron-based visualization was initially utilized to illustrate reaction directionality associated with the thermodynamic enzyme pair predictions. While the enhanced chevron visualization provided additional levels of information for audiences, the addition of route-type information within the context of this KEGG-based pathway representation is not without challenges. Areas of the visualization with a number of over-lapping predicted channelling pairs require attention and careful design to clearly delineate the metabolites and enzymes involved in each pair. Channelling pairs involving enzymes that are not directly adjacent in the representation also present a visualization challenge with the potential of over-lapping chevrons. Although thin red
arcs were used to delineate overlapping chevrons, the diagram may be difficult to interpret when multiple arcs originate from one chevron.

To combat these issues, a network-based approach was adopted which utilize a combination of hypergraphs and arcs. Leveraging different edges to combine nodes in hypergraph representations offers an extra level of grouping organization that is not available in current network representations beyond manual modification of network diagrams using graphics software. For example, nodes could represent enzymes with hyperedges linking enzymes that are involved the in same metabolic process (e.g., photosynthesis, the citric acid cycle, glycolysis, etc.). The overlapping of hyperedges at specific nodes can highlight how different processes are linked and reveal routes for directing metabolites from metabolic process to another. The aforementioned arcs facilitate representation of channelling directionality while clearly delineating distinct combinations of enzymes and metabolites predicted to benefit from channelling. This approach of leveraging arcs and hyperedges facilitates easier interpretation of existing metabolic pathways than a KEGG representation, which cannot easily represent links across all metabolic networks and cannot display how metabolites travel through a network based on existing thermodynamic properties and regulation of network enzymes. The added layers of information will enable predictions of metabolic flow in a network, the potential routes a particular metabolite may undertake, and the enzymes in a network that may be modified (i.e., deleted, inhibited, activated) to redirect a metabolite towards a desired path.
Chapter 3: Exploring the Potential of the HK97 Capsid as a Compartment System

3.1 Introduction

An important hallmark of living organisms is the ability to co-ordinate a diverse array of biochemical reactions into reaction cascades. One strategy to maintain the efficiency of these cascades is metabolic channelling, where reaction intermediates are funneled into the next active site without releasing each intermediate into the cytosol. Of the various methods for introducing channelling as described in Chapter 1 (Sections 1.1, 1.2, 1.4 and 1.5), an intriguing possibility is the use of compartments. They have been identified across a range of organisms from bacteria to mammalian cells and been shown to effectively control the flow of metabolites between the cytosol and the compartment interior (Section 1.1 and 1.4).

In the field of biotechnology, there is a growing demand to develop ways to improve reaction cascades. Researchers have drawn inspiration from naturally-occurring systems, as illustrated through several examples found in nature to constrain enzymes together. These characterized systems being exploited in an attempt derive benefits of channeling through the constraint of enzymes. To date, several types of natural compartments have been used as reaction vessels, including bacterial nano-compartments known as encapsulins [7], and viral capsids such as the bacteriophage P22 capsid [46]. While each system has been shown to successfully encapsulate target enzymes that catalyze reactions in the capsid interior, certain weaknesses have been noted which limit their widespread use in downstream applications. Encapsulins are fairly small and their selective pores may limit the diffusion of certain substrates [7]. The P22 capsid has been reported to successfully encapsulate several target enzymes; however, it relies on an intermediate thrombin processing step to separate the naturally-targeted scaffolding protein from the target enzyme [76]. Previous studies have demonstrated that thrombin can degrade recombinant proteins in a non-specific manner [163], which may limit the amount of encapsulated intact enzyme.

In this chapter, I designed a capsid-based micro-compartment using the HK97 capsid in order to examine its suitability as a compartment. There are several advantages to using the
HK97 capsid. First, it is one of the most studied bacteriophage capsid systems, with its assembly steps and structures of stable assembly intermediates fully elucidated [11, 80, 164-169]. It is capable of self-assembly in vitro (Figure 4a) [11], creates capsids that are significantly larger (54 nm in diameter) than other capsid and nano-compartment systems, and contains numerous pores on its capsid shell which may be permissive to diffusion of a wide range of substrates [169]. The C-terminal portion of the HK97 maturation protease was found to be responsible for targeting the protease into the assembling capsid [79], providing a targeting mechanism to encapsulate proteins of interest. Together, these attributes make the HK97 capsid a strong candidate for use as a compartment system to introduce channelling.

Although the HK97 capsid is extremely well characterized (Section 1.5.3), its use for targeting heterologous enzymes with industrial applications has not been extensively explored. In order to understand the potential of the HK97 capsid as a compartment system, I examined several strategies by which the capsid and target cargo could be co-expressed together for simultaneous cargo encapsulation and capsid assembly. I confirmed the expression and assembly of the HK97 capsid, as well as its ability to compartmentalize cargo proteins through microscopy, SDS-PAGE and enzymatic activity assays. A key challenge was determining how to increase targeting into HK97 capsids, which was achieved through induced expression of a multi-plasmid system and the alteration of select residues within the capsid interior.

After demonstrating that the HK97 capsid can be a working capsular system, the remainder of the chapter is dedicated to characterizing its ability to encapsulate a number of different enzyme systems for potential applications involving one or more enzymes together. The abilities of the system to stabilize unstable proteins and encapsulate components in a reaction cascade to improve reaction kinetics were examined. Together, this work demonstrates the flexibility and potential of bacterial protein compartments for use in a number of industrial applications.
3.2 Materials and Methods

3.2.1 Engineering the HK97 Prohead and targeted enzyme constructs

To produce a construct that can create an assembled HK97 capsid, I cloned the key structural genes that have previously been shown to create assembled capsids in vitro [11]. Since phages are able to reproducibly create capsids of consistent size in extensive quantities, to make the capsid producing construct (termed “HKPro”), all elements required to produce a capsid (including transcriptional control elements) were amplified from HK97 genomic DNA with the forward primer 5’-TTGTATTTTCCAGGCTTACCGAACAGGGTGAC-3’ and the reverse primer 5’-CAAGCTTCGTCATCACATCCAGAACATCTATCGCC-3’ using Phusion High-Fidelity DNA Polymerase (New England Biolabs). The product was inserted into SacI-digested p15TV-L plasmid using the In-Fusion HD Cloning System (Clontech) and transformed into E. coli DH5α cells using heat shock. Heat shock transformation is a method by which plasmids can be taken up by bacterial cells. Typically, bacteria are grown until they reach a fast-growing phase (typically when the optical density at 600 nm is just below 0.4). Cells are then harvested and then treated with calcium chloride, which coats the bacteria’s exterior, including a component on the bacterial membrane known as lipopolysaccharide (LPS) [170]. Upon a short incubation (approximately 45 seconds to 1.5 minutes) at 42˚C, LPS separates from the membrane and creates small pores where foreign material such as DNA can enter the bacterial cell.

To execute the protease replacement strategy for encapsulating target proteins in HK97 capsids, the gene encoding the maturation protease (which catalyzes structural changes in the capsid prior to DNA loading) from the capsid-producing construct was replaced with the protein to be encapsulated. All genes, except the maturation protease, were amplified to act as a plasmid backbone to encapsulate target enzymes. A technology known as ligation-independent cloning (NEB) was leveraged to insert the gene encoding the target protein into the capsid-producing plasmid backbone. Briefly, the technology involves the creation of ~15 base pair 5’ overhangs on the plasmid backbone and the target gene, and subsequent annealing of complementary overhangs. Annealed constructs are transformed into E. coli before bacterial machinery repair nicks in the annealed construct and complete the cloning process. The forward primer 5’-TGAACTATACAAAGGCATTGAAACGATCCGTGATG-3’, the reverse primer 5’-
CTTTACTCATCACGGCC CCCATAAAAATTAAGC-3’, and Phusion High-Fidelity DNA Polymerase (NEB) were used to amplify required regions of HKPro construct (i.e., excluding protease-coding portion of the maturation protease). The C-terminal targeting sequence was deliberately left intact to facilitate encapsulation of target proteins. The first target protein to be fused into the backbone, green fluorescent protein (GFP), was amplified using the forward primer 5’-GGGGGCCGTGATGAGTAAAGGAGAAGAACTTTTCAC-3’ and the reverse primer 5’-CGTTTCAATGCCTTTGTATAGTTTCATCCATGCCATGTG-3’, along with Phusion High-Fidelity DNA Polymerase (NEB). Finally, GFP was fused to the backbone using the In-Fusion HD Cloning System (New England Biolabs).

To execute the co-transformation strategy of encapsulating target proteins in HK97 capsids, the capsid-expressing construct was co-expressed with a second plasmid containing the target protein fused to the protease C-terminal targeting sequence. In this scenario, the maturation protease and the target protein would theoretically be co-encapsulated in the same capsid with the protease catalyzing capsid maturation steps without cleaving the target protein in the process. While the maximum loading capacity of the targeted enzyme would theoretically be decreased, loading a protease may create a more stable capsid (Prohead II intermediate) in comparison to the protease replacement method. To create the plasmid to be co-expressed with the capsid-expressing construct, GFP with the targeting sequence was amplified with the forward primer 5’-GGAGATATACCATGGCAATGAGTAAAGGAGAAGAACTTTTCAC-3’ and the reverse primer 5’-TTCTTTACCAGACTCGAGTTATTTACCTAAGTTAGGAAGGAAAGAT-3’, using Phusion High-Fidelity DNA Polymerase (NEB) and GFP in the protease replacement construct as the DNA template. The insert was then digested with NcoI and XhoI (NEB) for 2 hours before ligating it to NcoI and XhoI digested pCDF using T4 DNA Ligase (NEB).

*E. coli* alkaline phosphatase was amplified with the forward primer 5’-TTT TTA TGGGG CCG TG ATG CCT GTT CTG GAA AAC CG-3’ and the reverse primer 5’-CGG ATC GGG CCG TG ATG CCT GTT CTG GAA AAC CG-3’ for insertion into the protease replacement vector as described for GFP. A co-transformation construct of alkaline phosphatase was also made in the same method as described for GFP. Briefly, alkaline phosphatase with the capsid targeting sequence was amplified using the forward primer 5’-GGA GAT ATA CCA TGG CAATG CCT GTT CTG GAA AAC CG-3’ and the reverse primer 5’-TTCTTTACCAGACTCGAGTTATTTACCTAAGTTAGGAAGGAAAGAT-3’ using the
alkaline phosphatase protease replacement construct as a template and inserted into pCDF as described for GFP.

Additional enzymes targeted into capsids in this chapter leveraged the In-Fusion HD Cloning method by amplifying the modified pCDF construct with the targeting sequence intact (created as part of the co-transformation method) and inserting target enzymes with homology to the modified pCDF construct. All constructs in this chapter were verified through Sanger sequencing at the Centre for Applied Genomics.

3.2.2 Protein expression

To express either enzymes or capsids for eventual purification, constructs harbouring the appropriate expression plasmids were transformed into *E. coli* BL21 ER2566 cells (NEB) and plated on LB agar with the appropriate selective antibiotic. Colonies from the transformation were grown in a 20 mL overnight culture with LB media and the appropriate antibiotic, and subsequently added to a 1L culture of LB media with the appropriate antibiotic for a large-scale purification. The 1L cultures are grown to an optical density of 0.35 at a wavelength of 600 nm before expression induction with 1 mL of 0.8 mM IPTG and overnight growth at 28°C. The cells were pelleted at 7,000 rpm for 10 minutes using a JLA 8.1 rotor after induction and stored at -80°C until protein purification.

3.2.3 Capsid purification from induced samples

Frozen pellets generated by inducing capsid expression in BL21 ER2566 cells as described in Section 3.2.2 were resuspended in lysis buffer (50 mM Tris-HCl pH 8, 5 mM EDTA, 10 mM DTT) and sonicated using a QSonica sonicator (30% amplitude, 5 minutes processing time, 10 seconds on, 10 seconds off). Cellular debris was removed using a JA25.5 rotor for 30 minutes at 15,000 rpm at 4°C. The supernatant was transferred to a new container and PEG 8000 and sodium chloride were added to final concentrations of 6% w/v and 0.5M respectively. The mixture was placed on ice for 45 minutes to allow for the precipitation of capsids before centrifugation at 10,000 rpm for 15 minutes in a JA 25.5 rotor to concentrate the precipitated capsids in a pellet. The pellet was resuspended in Prohead buffer (20 mM Tris-HCL pH 7.5, 50 mM NaCl) to keep the capsids stable in solution before a clarification spin at 10,000
rpm for 10 minutes to remove un-resuspended material. The supernatant from the clarification spin was centrifuged in a Ti-75 rotor at 35,000 rpm for 2 hours at 4°C to concentrate assembled capsids present in the solution. The resultant pellet was resuspended in Prohead buffer before a clarification spin at 15,000 rpm for 20 minutes. The supernatant of the clarification spin was then loaded onto a 10-30% glycerol step-gradient and centrifuged using a SW27 rotor at 27,000 rpm for 1.5 hours at 4°C. Concentrated regions of protein which appear in the middle of the gradient was extracted using a syringe and a 21-gauge needle. The extracted material was dialyzed overnight in Prohead buffer and further purified using ion-exchange chromatography (MonoQ column, Amersham Pharmacia Biotech AB). Peak fractions were centrifuged in a Beckman Ti-75 rotor at 35,000 rpm for 2 hours at 4°C to concentrate the purified capsids.

3.2.4 Purification of His-tagged enzymes

Frozen bacterial pellets which contain His-tagged enzymes generated by inducing protein expression in BL21 ER2566 cells as described in Section 3.2.2 are resuspended and lysed by sonication in 10 mM imidazole, 300 mM NaCl, and 50mM NaH₂PO₄•H2O pH 8. The resulting supernatant was incubated with Ni-NTA agarose beads (Qiagen), pre-equilibrated in lysis buffer, for 1 hour at 4 °C to enable binding of His-tagged enzymes to the Ni-NTA beads. Subsequent to binding, the resin was washed three times with 10 ml of wash buffer (20mM imidazole, 300mM NaCl, and 50mM NaH₂PO₄•H2O pH 8) and transferred to a column. Protein was eluted from the beads in the column with 250mM imidazole, 300mM NaCl, and 50mM NaH₂PO₄•H2O pH 8. The proteins were then dialyzed overnight in 4L of 10mM imidazole, 250mM KCl, 0.2mM EDTA, 10mM Tris pH 8 to reduce imidazole concentrations in the purified sample solution.

3.2.5 SDS-PAGE and Western blot

Protein samples were boiled in 2x loading buffer containing β-mercaptoethanol in a heating block at 100°C for 7 minutes. Samples were then spun down at 13,000 rpm for 1 minute to concentrate debris and loaded onto a polyacrylamide gel consisting of a 5% acrylamide stacking gel and a 15% resolving gel. The gel was run at a constant voltage of 130V for approximately 1.5 hours. To perform Western blotting of samples, the polyacrylamide gel (containing the same), blotting paper and nitrocellulose membrane were soaked in Western
Transfer buffer (25 mM Tris, 190 mM glycine, 20% methanol, pH 8.3) for 15 minutes. Proteins were transferred onto a membrane using a Trans-Blot SD cell semi-dry transfer apparatus (BioRad) at 10V for 50 minutes. The membrane was subsequently blocked using a 5% milk w/v in TBS-Tween (50 mM Tris, 150 mM NaCl, 0.1% Tween 20, pH 7.5), and incubated for 30 minutes in TBS-Tween (2 rounds) at 4°C. The blocked and washed membrane is incubated overnight using either anti-GFP or anti-phoA antibody (Santa Cruz). Following overnight incubation, the membrane is incubated in three rounds of TBS-Tween prior to incubation with a goat anti-mouse IgG horseradish peroxidase antibody (Bioshop) diluted 1:5000 in 5% skim milk for 1 hour at room temperature. Excess secondary antibody is washed off with 3 additional rounds of TBS-Tween before visualization using ECL Western blot detection reagents (Amersham) using their standard protocol.

3.2.6 Gel densitometry and average protein copy number per capsid

Coomassie-stained protein gels were scanned using the BioRad Gel Doc system and analyzed using the “Analyze” functionality within ImageJ software. Given that a capsid has 420 copies of the capsid subunit, the intensity ratio between the protein of interest and the capsid subunit can provide a rough estimate to the average number of target protein encapsulated per capsid. For example, if the band intensity ratio of a target protein and the capsid protein band is approximately 1:6, one can estimate that 70 copies of the target protein is encapsulated per capsid.

3.2.7 Enzyme activity assays

Alkaline Phosphatase Activity Assay

The activity of alkaline phosphatase in purified samples was measured using a colorimetric assay. In brief, a 100 µL reaction mixture containing 10 µL of purified sample, 73 µL of 1M Tris pH 8.0 and 12 µL of 500 µM p-nitrophenol phosphate (New England Biolabs) was monitored for the production of p-nitrophenol at 405 nm by using a M200-Pro machine (Tecan Trading AG). Kinetic parameters of alkaline phosphatase was determined using a non-linear regression of the standard Michaelis-Menten model in GraphPad Prism 5.0 (GraphPad Software, USA).
Reductive Dehalogenase Activity Assay

Reductive dehalogenases were tested for their ability to dechlorinate chloroform and trichloroethene. Concentrations of dechlorination products in lysates incubated with a chloroform assay mix was measured by gas chromatography with headspace analysis calibrated with external standards as described in Grostern and Edwards [171].

β-galactosidase and galactose dehydrogenase activity assay

Protocols based off of Ljungcrantz et al. [89] were leveraged to assay β-galactosidase and galactose dehydrogenase activity. β-galactosidase activity was assayed in a solution of 0.8 g/L ONPG, 0.1M NaH2PO4 pH 7.0 and 1mM MgSO4. Reaction progress was tracked by the formation of o-nitrophenol, which is detected spectrophotometrically at 420 nm. Galactose dehydrogenase activity is assayed in a solution of 16.6 mM galactose, 0.09M Tris-HCl pH 8.5 and 0.5 mM NAD+. Reaction progress is tracked by the formation of NADH which is detected spectrophotometrically at 340 nm. To track potential channelling between the two enzymes, the production of NADH was monitored using a M200-Pro machine (Tecan Trading AG) in a buffer containing 0.09M Tris-HCl pH 7.5, 0.06 MgCl2, 38 mM lactose and 4 mM NAD+. Arabinofuranosidase activity assays

Activity assays to monitor arabinofuranosidase activity in purified samples were performed using the Nelson-Somogyi procedure [172]. Non-encapsulated and encapsulated arabinofuranosidase (in HK97 capsids) samples were added to a reaction solution containing arabinoxylan substrate in 100 mM HEPES buffer (pH 6.5). The reaction was incubated at 37°C for 30 minutes followed by a 20-minute incubation at 100°C to terminate arabinofuranosidase activity. Reaction progress was tracked by the formation of reducing sugars from the arabinoxylan substrate, which is detected spectrophotometrically at 620 nm.

3.2.7 Transmission Electron Microscopy to visualize capsids

The capsid samples were examined using the transmission electron microscope (TEM). For each sample, five microliters was loaded onto a 400-mesh TEM carbon coated, glow discharged copper grid. Each grid was washed three times in deionized water and stained with 2% uranyl acetate for five seconds. Excess stain was blotted off with filter paper before
visualization using a Hitachi H-7000 TEM. The number of capsids was estimated by calculating
the averaging the number of capsids in several fields of view and scaling the observed count over
the area of the full mesh grid and the level of sample dilution.

3.3 Results

3.3.1 Assembled HK97 capsids can successfully encapsulate target proteins

To demonstrate the HK97 capsid’s utility as a compartment system, I first aimed to
engineer a targeting mechanism that reliably encapsulates target proteins of choice inside of
assembled HK97 capsids. The first step involved generating capsids in a reproducible, high
quantity manner. I created a capsid-producing construct consisting of the portal, protease, and
major capsid encoding genes from the HK97 genome that have been previously shown to be
sufficient for capsid formation [11]. The assembly of intact capsids was verified by purifying
capsids after expression of the construct in E. coli and visualizing purified samples with
transmission electron microscopy (Figure 4b). Next, I devised two strategies to test whether the
C-terminal protease targeting residues could also target heterologous proteins such as GFP inside
of assembling capsids. The first strategy, termed the protease replacement method, involves
replacing the protease coding region in the capsid construct with GFP such that GFP and the C-
terminal putative targeting sequence are fused in frame (Figure 12). With this method, GFP
would potentially co-assemble with the portal protein and the major capsid protein to form a
Prohead I HK97 capsid structure with encapsulated GFP. The second strategy, termed the co-
transformation method, involves co-transforming the capsid construct with a separate construct
where GFP is fused in frame with the putative targeting sequence (Figure 12). With this co-
transformation method, GFP would potentially co-assemble with the portal, protease and major
capsid proteins. The protease would then activate within the assembled capsid and cleave itself,
as well as the delta-domain of the major capsid proteins, without cleaving GFP itself. This
would result in generation of the second structural capsid intermediate, Prohead II, with GFP
localized inside. This intermediate is known to be more structurally stable than the Prohead I
structural intermediate [11, 80], which may confer additional stability to encapsulated proteins.
Figure 12. Two potential methods of encapsulating target proteins into HK97 capsids.

In the *protease replacement* method (top row), the protease coding region (red) is replaced by GFP (green) using ligation-independent cloning, with homology to a region 5′ of the protease start codon and the 5′ end of the C-terminal targeting sequence (pink). Upon expression, the proteins will theoretically assemble into a Prohead I structural intermediate with no protease being expressed or encapsulated. In the *co-expression* method (bottom row), the capsid construct and GFP fused to the C-terminal targeting sequence are expressed from separate plasmids. Upon assembly, the maturation protease and GFP would co-assemble into the same capsid. The protease would then activate and process the capsid into the Prohead II intermediate without cleaving GFP in the process.

After inducing the expression of the constructs for these two methods, lysates were processed using the capsid purification procedure (described in Section 3.2.3) to concentrate putative assembled capsid. Samples were visualized using transmission electron microscopy to confirm capsid assembly and further analyzed using Coomassie-stained SDS PAGE gels and Western blot analysis to determine whether GFP was encapsulated in the samples (Figure 13a). The two encapsulation strategies yielded different results with respect GFP encapsulation. GFP was successfully encapsulated with the capsids using the protease replacement method, with gel densitometry analysis estimating that 10 copies of GFP were encapsulated per capsid. However, the *co-expression* method yielded capsids with no co-purified GFP, despite a significantly higher level of GFP expression in crude lysates as compared to lysates generated through the *protease replacement* method (Figure 13a). This signifies that GFP is either excluded from the capsid (i.e., outcompeted by the protease), or is degraded by protease upon co-encapsulation. Previous studies have identified a peptide sequence which is recognized and cleaved by the HK97 maturation protease [79]- this sequence is present within GFP. The *co-expression* method also revealed that GFP does not bind to the outside of the HK97 capsid, as despite its abundance, it did not co-purify with the capsid sample. This observation signifies that bands corresponding to GFP in the SDS-PAGE gel from purified samples (generated using the *protease replacement*
method) represent encapsulated GFP, rather than GFP simply sticking to the outside of the capsid shell. Subjecting an induced lysate over-expressing GFP subjected to the capsid purification procedure resulted in all GFP being removed during purification (Figure 13a, lane 8). This result demonstrates that unencapsulated GFP is not purified using this procedure.

To verify that the targeting sequence is necessary for GFP encapsulation, stop codons were inserted between GFP and the targeting sequence in the protease replacement construct. Purification of capsids created from this construct and subsequent analysis on SDS-PAGE revealed no presence of purified GFP (Figure 13a, lane 12), illustrating that the targeting sequence was necessary for GFP encapsulation. A control experiment with the major capsid protein removed from the protease replacement construct resulted in no assembled capsids or GFP being recovered (Figure 13a, lane 11). Analysis of the various purified capsid samples using fluorescence microscopy was performed to verify that GFP was encapsulated in a correctly folded form. Imaging of the protease replacement GFP sample revealed the presence of distinct foci (Figure 13c). These foci were not visible when visualizing purified samples generated from the co-transformation method and the GFP only samples. In combination with the SDS-PAGE results, these results demonstrated that GFP is targeted into assembled capsids.

Given the similarity in molecular weight of GFP (27 kDa) [173] and the capsid protease (25 kDa) [11], it is expected that GFP would be encapsulated at a similar efficiency to the protease. Although targeting GFP into capsids was successful with the protease replacement strategy, the estimated targeting levels were significantly lower than that of the native protease. As highlighted in Section 1.5.3, the protease has previously been estimated to be encapsulated at 60 copies per capsid [11]. Based on gel densitometry analysis, only 5-10 copies of GFP were predicted to be encapsulated. In an attempt to increase GFP targeting levels, I fused GFP to the C-terminal targeting sequence and cloned the fusion product into an over-expression plasmid. This would theoretically serve to increase levels of GFP produced in hopes of increasing targeting efficiency. The HK97 maturation protease is naturally expressed at low levels, meaning that GFP expressed using the protease replacement strategy would also be expressed at low quantities. By expressing the capsid protein on one plasmid and the GFP fused to the C-terminal targeting peptide on a separate plasmid, targeting levels of GFP was estimated to be approximately 60 copies. This matches encapsulation levels of the naturally encapsulated
maturation protease (Figure 13b). Thus, it appears over-expressing target enzymes is a viable strategy for increasing the amount of target protein encapsulated per capsid.

**Figure 13. GFP is successfully encapsulated into HK97 Prohead I capsids.**

(a) Induced cell lysates and samples processed through the capsid purification procedure were analyzed on SDS-PAGE stained with Coomassie Blue dye and also by Western Blot using anti-GFP antibody (Santa Cruz). The uncleaved major capsid protein (42 kDa), the cleaved version (31 kDa) and GFP (27 kDa) are highlighted by corresponding arrows.
(b) GFP encapsulation efficiency in HK97 Prohead I capsids was increased through GFP overexpression. GFP fused to the C-terminal targeting sequence was co-expressed with another plasmid expressing the HK97 capsid protein. Capsid samples from this over-expression method and the protease replacement method were obtained from induced lysates that underwent the capsid purification procedure. Targeting levels were determined using gel densitometry.

(c) Fluorescence microscopy images samples containing Prohead I capsids alone, Prohead I capsids co-expressed with GFP, and over-expressed GFP after processing through the capsid purification procedure. The FITC channel detects fluorescence at a wavelength of 525 nm to detect emission by GFP, while TRITC detects fluorescence at 600 nm to differentiate any GFP-specific fluorescence (FITC-only fluorescence) from autofluorescence (FITC and TRITC fluorescence). Electron microscopy (EM) was performed on the same samples to demonstrate the presence of assembled capsids in the “Prohead I capsids” and “Prohead I capsids co-expressed with GFP” samples. Scale bars represent 100 nm in the EM images.

3.3.2 Enzymes display activity when targeted inside of HK97 capsids

Having shown that enzymes can be targeted inside capsids, I next examined whether targeting enzymes into HK97 capsids has an effect on the encapsulated enzyme’s activity. This was explored using *E. coli* alkaline phosphatase (PhoA), a stable enzyme for which a well-defined assay has been defined [174]. PhoA was amplified from *E. coli* BL21 DNA and targeted into HK97 capsids using the protease replacement strategy. Analysis of the purified sample with SDS-PAGE Coomassie-stained gels and Western blotting with anti-PhoA antibody confirmed that PhoA was successfully targeted inside of capsids (Figure 14a). As seen with the GFP targeting experiments, PhoA was not purified with capsids when the co-transformation strategy was used. Since PhoA does not contain any predicted cleavage sites for the protease [79], it suggests that the maturation protease is preferentially encapsulated over both GFP and PhoA, despite much lower levels of expression. This raised an interesting question as to what factors influenced encapsulation, which will be addressed later in this chapter.

To determine whether the encapsulated PhoA is active and substrates can diffuse inside of assembled capsids, a well-established alkaline phosphatase assay involving the conversion of *p*-nitrophenyl phosphate into *p*-nitrophenol was leveraged. A purified sample containing encapsulated alkaline phosphatase was able to generate *p*-nitrophenol (Figure 14b). On the other hand, a sample of over-expressed PhoA alone subjected to the capsid purification procedure yielded no activity (Figure 14b). A separate control experiment where empty Prohead I capsids were subjected to the alkaline phosphatase activity assay also did not generate *p*-nitrophenol. In
all, these results showed that alkaline phosphatase was active upon encapsulation in HK97 capsids and that the substrate could diffuse into the capsid.

To understand whether the portal complex was needed for diffusion of substrates into the capsid or whether pores on the capsid shell would suffice, a stop codon was introduced into the portal protein in the protease replacement construct at the beginning of the protein coding region. The portal protein is a 12-membered ring complex through which DNA is fed into Prohead II capsid intermediates to form the fully matured capsid head [165, 175]. The introduction of the stop codon served to knock out portal protein expression and subsequent incorporation into the capsid. Removing the portal protein from the PhoA protease replacement construct resulted in no significant differences in enzymatic activity as compared to capsids containing the portal complex and encapsulated PhoA (Figure 14c). This implies that pores in the capsid shell are sufficient for substrate diffusion and the larger pore formed by the portal is not required for substrate diffusion into the capsid interior.
Figure 14. Alkaline phosphatase is encapsulated and active inside of HK97 capsids.

(a) Induced cell lysates and samples processed with the capsid purification procedure were analyzed on SDS-PAGE with Coomassie Blue dye, and by Western Blot using anti-PhoA antibody (Santa Cruz).

(b) The alkaline phosphatase activity of various samples after undergoing the capsid purification procedure was measured by detecting $p$-nitrophenol production at 405 nm. Unencapsulated PhoA and

<table>
<thead>
<tr>
<th>Capsid</th>
<th>Activity&lt;br&gt;$p$-nitrophenol productions&lt;br&gt;(uM/s/ug enzyme)</th>
<th>$K_m$&lt;br&gt;(mM of p-NPP)</th>
<th>$k_{cat}$&lt;br&gt;(s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unencapsulated PhoA</td>
<td>0.65</td>
<td>37.04</td>
<td>3.3</td>
</tr>
<tr>
<td>Encapsulated PhoA</td>
<td>0.3225</td>
<td>79.27</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Capsid-encapsulated PhoA
His-tag purified PhoA
PhoA subjected to the capsid purification procedure
Empty capsids subjected to the capsid purification procedure
Capsid without portal complex
Full capsid

(d)
empty capsids exhibited no activity following the purification procedure, while encapsulated PhoA and purified His-tagged PhoA demonstrate activity during the assay.

(c) Removal of the portal complex resulted in no significant difference in activity of encapsulated alkaline phosphatase. This indicates that capsid shell pores are sufficient for substrate diffusion into the capsid.

(d) Michaelis-Menten kinetics of encapsulated and unencapsulated PhoA. \( p \)-nitrophenol production was monitored for 3 minutes after addition of \( p \)-nitrophenyl phosphate. Reaction rate is measured as the concentration of \( p \)-nitrophenol per second of the reaction per µg of enzyme in the reaction.

Next, the kinetics (\( V_{\text{max}} \) and \( K_m \)) of encapsulated and unencapsulated PhoA were determined to understand whether encapsulation of PhoA has an effect on the kinetics associated with converting \( p \)-nitrophenyl phosphate to \( p \)-nitrophenol. The \( K_m \) values were similar between the two samples (Figure 14d) however, the \( k_{\text{cat}} \) of the encapsulated sample was roughly 8-fold lower than the unencapsulated PhoA. One potential reason for this observation is that diffusion of substrates is limited by the capsid shell. To test whether substrate diffusion is limiting PhoA activity, a partial digestion of capsids was performed to increase porosity of the capsid (Figure 15a). While activity of PhoA in the partially digested PhoA capsid sample was increased in comparison to the undigested PhoA capsid sample (Figure 15b), electron microscopy analysis of the partially digested sample revealed that digestion led to capsid disintegration, rather than the formation of larger pores on intact capsids (Figure 15c). This result suggests that confinement within capsids and subsequent decreased substrate diffusion is responsible for decreased activity, rather than modification of the enzyme by attaching the C-terminal targeting sequence.
Figure 15. Decreased PhoA activity in capsids is likely due to substrate diffusion limitation.

Encapsulated PhoA was either digested with limited amounts of proteinase K or left undigested for 15 minutes. Digested / undigested samples were obtained every 5 minutes and analyzed (a) using SDS PAGE and (b) a p-nitrophenyl phosphate degradation assay to determine intact PhoA and total PhoA activity at each time point. (c) Capsid samples digested with 2 μg/μL proteinase K were visualized using EM after 5 minute and 15 minute degradation periods to determine whether degradation created larger pores on intact capsids or disintegrated capsids.

### 3.3.3 Insoluble proteins can be isolated within purified capsids

Having demonstrated that proteins can be targeted into HK97 capsids in an active form, I was interested in understanding whether the HK97 capsid could be used for various downstream applications. One of the applications included the potential stabilization of insoluble or unstable enzymes for biochemical studies, which were previously not amenable to such studies. To examine whether HK97 capsids could be used for such an application, three different classes of insoluble enzymes with potential industrial importance were targeted into capsids to explore whether they can be rescued in a soluble and active form. The first class of enzymes, reductive dehalogenases, are found in anaerobic organisms and allow host organisms to utilize halogenated compounds as carbon sources [176]. The ability of this class of enzymes to degrade toxic
halogenated compounds in the environment highlights their potential for bioremediation in natural environments. However, previous attempts to express these enzymes in *E. coli* resulted in the formation of insoluble, inactive proteins [177]. The second class of enzymes, laccases, are iron-sulfur cluster containing enzymes that catalyze oxidation-reduction reactions and have potential uses in a number of applications in the food, textile and paper industries [178]. Initial attempts to characterize putative laccases identified by BLAST searches in Dr. Alexei Savchenko’s group (University of Toronto) were unsuccessful as they could not be purified in an active form. The final class of enzymes targeted were arabinofuranosidases, which cleave glycosidic bonds in oligosaccharides typically found in plant cell walls. Plant matter is a rich source of complex polysaccharides that can be converted into biofuel [179]. The use of low specificity chemical catalysts to decompose plant matter into a simpler form is financially prohibitive, making natural degradative agents such as arabinofuranosidases an attractive alternative. However, a lack of understanding regarding the mechanism of action of arabinofuranosidase and the inability to purify certain laccases in a soluble form limits their potential effectiveness in producing biofuels.

In collaboration with Olivia Molenda from Dr. Elizabeth Edwards’ lab, three reductive dehalogenases (TceA, CfrA and VcrA) identified by Dr. Elizabeth Edward’s group at the University of Toronto were targeted into capsids to determine whether encapsulation can solubilize each enzyme and enable biochemical and structural studies. The enzymes were expressed and successfully targeted within capsids (Appendix A, Supplemental Figure 1). However, activity was not recovered in the encapsulated dehalogenase samples despite sample growth in various conditions with the addition of a number of supplements. In collaboration with Veronica Yim and Dr. Robert Flick, sixteen putative laccases were targeted into capsids and also were successfully encapsulated upon purification of capsids (a subset of which are shown in Figure 16). A known soluble and active laccase was also targeted as a control and its activity was found to be nearly 1000-fold slower as compared to its unencapsulated form (Data not shown). Testing of all the putative laccases that were successfully encapsulated in a panel of 23 substrates in 9 pH conditions however yielded no detectable activity.
Laccases fused to the C-terminal targeting sequence were co-expressed with the capsid protein before subsequent purification of capsids from induced samples. BSU1539, SC1976 and LPG1332 are approximately 38, 32, and 34 kDa in size, while SM23454 and TBD0811 are 36 kDa and 30 kDa in size respectively.

Finally, in collaboration with Cody Sarch of Emma Master’s group at the University of Toronto, putative arabinofuranosidases which could not be isolated in a soluble and active form were targeted into HK97 capsids to explore whether activity could be recovered. 4 putative and 2 previously characterized arabinofuranosidases were successfully targeted into capsids. Similar to the other enzyme classes, the putative arabinofuranosidases yielded no activity. The previously characterized enzymes also showed a significantly lower catalytic activity when encapsulated. These results demonstrate the possibility of targeting insoluble and unstable enzymes into capsids rather than aggregating in the cytosol without encapsulation into capsids. However, encapsulation does not seem to prevent the proteins from being in an inactive state, possibly due to the formation of non-functional aggregates within the capsid shell, the inability to incorporate co-factors required for enzyme activity, or due to the capsid's tendency to reduce kinetic activity.
3.3.4 Altering capsid charge interior can affect activity and targeting levels of targeted enzymes

Having shown that a range of enzymes can be encapsulated, the next step is to explore whether the activity or targeting levels of encapsulated enzymes can be modulated. Previous studies have altered a compartment’s porosity, volume and targeting sequences to investigate the effects on enzyme targeting levels [46, 51, 180]. While increasing porosity and internal volume of the P22 capsid individually had no effect on activity of encapsulated alcohol dehydrogenase, combining both alterations led to an increase the activity of alcohol dehydrogenase similar to non-encapsulated alcohol dehydrogenase levels [46]. In order to encapsulate enzymes for biotechnology applications, a better understanding of how the targeting levels of enzymes can be manipulated is required. One potential area of alteration that has not been reported is the alteration of interior charge. To examine whether this is a viable strategy to modulate targeting and/or enzyme activity in single and multi-enzyme systems, a charged residue (D136) was chosen to be the target residue for manipulation given its orientation in the capsid interior and its distance away from capsid protein interfaces, minimizing the chance to disrupt capsid assembly. Since there are 420 subunits that make up the Prohead I capsid, creating a 1-charge difference mutation (D136) and a 2-charge difference mutation (D136R) would lead to a 420 and 840 positive charge change per capsid respectively.

PhoA was co-expressed with either mutated or wild-type capsid and the amount of each encapsulated enzyme was quantified by gel densitometry to determine the impact of interior capsid charge on enzyme targeting and/or activity. While no change was observed for alkaline phosphatase targeting in the capsid mutants (Figure 17a), a difference in enzyme activity was observed. The D136R mutation had 2-fold decreased alkaline phosphatase activity as compared to the wild type and D136A capsid (Figure 17b). Together, these results demonstrate that altering capsid charge can affect enzyme activity within capsid shells and can serve as a viable strategy to optimize performance of capsid systems in future applications.
Figure 17. Capsid mutations affect alkaline phosphatase activity but not targeting levels.

(a) Induced and purified samples of PhoA encapsulated in either wild-type or D136A/D136R capsid mutants were visualized on SDS-PAGE. No significant differences in targeting levels of PhoA were observed.

(b) Alkaline phosphatase activity was assayed in the various capsid mutants after normalizing for the amount of PhoA across each sample. Activity in the D136R capsid mutant was significantly lower than either the wild-type or the D136A mutant. A negative control of NHA GH62 in wild-type capsids showed no activity.
3.3.5 Encapsulating multiple enzymes into HK97 Proheads

To determine whether reaction cascades can be created within capsids, I attempted to target multiple enzymes into the HK97 capsid and detect the presence of a coupled reaction. Since the HK97 capsid preferentially encapsulated its protease over PhoA and GFP (Section 3.3.1 and 3.3.2), I was interested in understanding potential mechanisms which influence encapsulation (i.e., what factors enable an enzyme to be preferentially encapsulated over another enzyme). To explore the phenomenon of preferential encapsulation, I selected β-galactosidase (β-gal) and galactose dehydrogenase (GalD), which were both fused to the C-terminal capsid targeting sequence on separate plasmids and over-expressed with the capsid-producing construct. Both enzymes were successfully encapsulated (Figure 18a), however, coupling of the two reactions was not detected. Gel densitometry estimations of targeting levels revealed that amount of β-gal within the capsid was lower than the amount of encapsulated GalD (approximately 18 β-gal and 19 GalD per capsid). When β-gal and GalD were targeted into the charge-mutated capsids described earlier, there were slight differences in targeting of the β-gal and GalD between the mutant and wild-type capsids. Based on gel densitometry analysis performed on multiple dilutions of capsid samples, ~1 copy of GalD was encapsulated per copy of β-gal in the wild type capsid, with D136A capsids containing approximately 1.2 copies of GalD per copy of β-gal, and D136R capsids containing approximately 1.5 copies of GalD per copy of β-gal (Figure 18a). In combination with the observation in Sections 3.3.1 and 3.3.2 that the protease is preferentially encapsulated, this result raises the question as to what factors affect encapsulation efficiency when multiple enzymes are being targeted into capsids.
(Left) Induced and purified samples of β-gal and GalD targeted into either wild-type or D136A/D136R capsid mutants. Both enzymes were successfully targeted into the three capsid variants.

(Right) Samples of β-gal and GalD targeted into either wild-type or D136A/D136R capsid mutants. Dilution of the samples revealed β-gal is targeted at slightly higher levels in the D136R mutant.

There are several potential possibilities which may influence preferential targeting including enzyme surface charge, enzyme hydrophobic patches, or important residues at specific regions of the target enzyme (excluding the targeting sequence). Targeting different combinations of multiple enzymes (e.g., β-gal, GalD, laccases, etc.) may provide insight into biochemical factors which influence the preferential encapsulation. However, the lack of available protein structures for a significant protein of successfully targeted enzymes (i.e., laccases, reductive dehalogenases) makes the evaluation of potential preferential targeting based on factors such as enzyme surface charge and hydrophobic patch extremely difficult. Attempting to modify various residues of target enzymes to identify mutations affecting targeting may not provide sufficient insights to warrant the effort involved; any changes in targeting may be due to enzyme structural changes rather than important interactions between the enzyme and capsid. I decided to explore biochemical properties of the targeted enzymes given its relative simplicity in hypothesizing and testing preferential encapsulation.
When analyzing various properties of β-gal and GalD to explain the preferential encapsulation of GalD, one hypothesis that arose relates to protein charge. Given the earlier result where β-gal encapsulation relative to GalD decreased with increasing positive charge of the capsid, I hypothesized that enzymes with a greater positive charge will be preferentially encapsulated in the wild-type capsid. While the overall isoelectric point (pl) of the two enzymes were similar (GalD: 5.38; β-gal: 5.30), there were notable differences in the N-terminal and C-terminal pl of both enzymes, with the N-terminus defined as the first 30 or 50 amino acids and the C-terminus defined as the last 30 or 50 amino acids (Table 5). β-gal has a higher isoelectric point than GalD when comparing their 50 residue N-termini, 50 residue C-termini and 30 residue C-termini. Given that the E. coli lumen pH is approximately ~7.2 – 7.8 [181], it is possible for β-gal to be preferentially encapsulated due to the repulsion of GalD with a lower pl (when pH > pl) or due to higher β-gal terminus positive charge (when pH < pl).

Table 5. Select properties of enzymes in multi-enzyme targeting experiments.

<table>
<thead>
<tr>
<th>Targeted Enzyme</th>
<th>Protein Size</th>
<th>N-terminal pl (50 residues)</th>
<th>N-terminal pl (30 residues)</th>
<th>C-terminal pl (50 residues)</th>
<th>C-terminal pl (30 residues)</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-galactosidase</td>
<td>117 kDa</td>
<td>9.17</td>
<td>5.82</td>
<td>5.24</td>
<td>5.38</td>
</tr>
<tr>
<td>Galactose dehydrogenase</td>
<td>33.1 kDa</td>
<td>7.96</td>
<td>8.36</td>
<td>4.35</td>
<td>4.36</td>
</tr>
<tr>
<td>Galactokinase</td>
<td>41.4 kDa</td>
<td>5.34</td>
<td>9.7</td>
<td>4.65</td>
<td>6.21</td>
</tr>
<tr>
<td>Laccase EC6170</td>
<td>51.9 kDa</td>
<td>11.54</td>
<td>10.86</td>
<td>4.43</td>
<td>5.32</td>
</tr>
<tr>
<td>Laccase NE0267</td>
<td>21.3 kDa</td>
<td>8.34</td>
<td>9.98</td>
<td>4.23</td>
<td>3.85</td>
</tr>
<tr>
<td>Laccase SC1976</td>
<td>25.1 kDa</td>
<td>4.38</td>
<td>4.75</td>
<td>7.02</td>
<td>7.95</td>
</tr>
<tr>
<td>Laccase TBD0811</td>
<td>23.6 kDa</td>
<td>5.55</td>
<td>4.13</td>
<td>9.19</td>
<td>6.25</td>
</tr>
</tbody>
</table>

To determine if the targeted enzyme N- or C-terminus pl is responsible for differences in encapsulation, four putative laccases (EC6170, NE0267, SC1976 and TBD0811) were co-targeted with GalD or galactokinase (GalK) into wild-type HK97 capsids. The amount of each co-targeted enzyme was measured through gel densitometry analysis and compared to predictions of enzyme encapsulation, generated based on the 50 residue N-termini, 50 residue C-termini and 30 residue C-termini pl (Figure 19).
β-galactosidase (β-gal) and galactose dehydrogenase (GalD) or galactose dehydrogenase or galactokinase (GalK) co-targeted with select laccases (EC6170, SC1976, TBD0811, NE0267) fused to HK97 capsid targeting sequence were co-expressed from separate plasmids along with the capsid-producing construct and purified within capsids using the capsid purification procedure. Enzyme levels were estimated by diluting samples over several dilutions and measuring band intensities using gel densitometry.

Table 6. Predictions and targeting results of various enzyme pair combinations associated with multi-enzyme targeting experiments. Yellow shaded boxes represent predictions which match experimentally-observed targeting result.

<table>
<thead>
<tr>
<th>Targeted Enzyme #1</th>
<th>Targeted Enzyme #2</th>
<th>N-terminal pi (50 AA)</th>
<th>C-terminal pi (50 AA)</th>
<th>C-terminal pi (30 AA)</th>
<th>Estimated Targeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galactose dehydrogenase</td>
<td>EC6170</td>
<td>EC6170</td>
<td>Preferentially encapsulated</td>
<td>Similar levels of both enzymes encapsulated</td>
<td><strong>Nearly 100% EC6170 (~48 copies)</strong></td>
</tr>
<tr>
<td>Galactose dehydrogenase</td>
<td>NE0267</td>
<td>NE0267</td>
<td>slight preferential encapsulation</td>
<td>Similar levels of both enzymes encapsulated</td>
<td><strong>20 GalD</strong> <strong>14 NE0267</strong></td>
</tr>
<tr>
<td>Galactose dehydrogenase</td>
<td>TBD0811</td>
<td>GalD</td>
<td>Preferentially encapsulated</td>
<td>TBD0811</td>
<td><strong>10 GalD</strong> <strong>52 TBD0811</strong></td>
</tr>
<tr>
<td>Galactokinase</td>
<td>EC6170</td>
<td>EC6170</td>
<td>Preferentially encapsulated</td>
<td>Similar levels of both enzymes encapsulated</td>
<td><strong>56 GalK</strong> <strong>27 EC6170</strong></td>
</tr>
<tr>
<td>Galactokinase</td>
<td>SC1976</td>
<td>Similar levels of both enzymes encapsulated</td>
<td>SC1976</td>
<td>Preferentially encapsulated</td>
<td><strong>14 GalK</strong> <strong>38 SC1976</strong></td>
</tr>
<tr>
<td>Galactokinase</td>
<td>TBD0811</td>
<td>Similar levels of both enzymes encapsulated</td>
<td>TBD0811</td>
<td>Preferentially encapsulated</td>
<td><strong>8 GalK</strong> <strong>76 TBD0811</strong></td>
</tr>
</tbody>
</table>
The observed targeting results overlapped at varying levels with my encapsulation predictions (Table 6). Predictions based on 30 residues at an enzyme’s C-terminus had the highest overlap with the observed results (4 out of 6). Meanwhile, predictions based on 50 residues at an enzyme’s C-terminus had 3 overlaps with observed results and predictions based on 50 residues at an enzyme’s N-terminus only had one overlap with observed results. Although the results of the co-targeting experiments do not specifically pinpoint one terminus as the key determinant of preferential encapsulation, results from the earlier capsid charge mutation experiments (involving β-gal and GalD) indicate that protein charge can affect encapsulation efficiencies in multi-enzyme systems. This may include surface charge on target proteins, which further assist in making initial contacts with the interior capsid proteins as it assembles into an enclosed capsid.

### 3.3.6 Capsids can stabilize enzymes at higher temperatures

The final application explored was the HK97 capsid’s ability to stabilize encapsulated enzymes and improve their activity when exposed to previously non-activity-permitting conditions. Purified capsids containing a previously characterized arabinofuranosidase from the fungal plant pathogen *Nectria haematococca* (NHA) were subjected to a number of treatments such as altered salt concentrations, pH and temperature conditions. As a control, unencapsulated NHA was also subjected to the same treatments, and the relative activity prior to and after a treatment was determined for both the encapsulated and unencapsulated forms. Alteration of salt concentrations or pH revealed no significant difference in relative activities (data not shown). Interestingly, incubation of the unencapsulated and encapsulated forms of NHA at a different temperature had a drastic effect (Figure 19a). Within 5 minutes of incubation at 47°C, the activity of the unencapsulated NHA decreased to 30% of its initial levels with activity almost abolished after 40 minutes post incubation. Incubation at this temperature also led to the formation of a visible precipitate in the sample (Figure 19b). After normalizing encapsulated NHA concentrations with unencapsulated NHA concentrations (using gel densitometry) and assaying NHA activity at 47°C, encapsulated NHA levels were stable 5 minutes post incubation at 47°C and retained over 80% of its original activity 40 minutes post incubation. Incubation at 60°C, where Prohead I capsids have previously been shown to disassemble, resulted in significantly reduced activity after 5 minutes that correlated with the formation of precipitate as
previously described. These results suggest that assembled capsids can stabilize encapsulated proteins and extend its half-life in conditions that previously abolished activity. This finding has implications for industrial processes, where enzymes can be stabilized in non-optimal reaction conditions to enhance enzyme stability and increase production of desired compounds.

Figure 19. The activity of NHA GH62 samples with different temperature treatments.

(a) The samples were extracted from temperature incubation at 5, 10, 20, 40 and 80 minutes post-incubation and mixed with the arabinofuranosidase substrate for 3 hours. The resultant mixture was assayed with the arabinose assay kit and the change in absorbance was recorded. Values were normalized to the absorbance change observed just as the samples were incubated at their respective temperatures. A capsid alone negative control was run with this experiment and resulted in no absorbance change.

(b) Physical states of NHA GH62 arabinofuranosidase samples before and after incubation at 47°C for 45 minutes. The encapsulated NHA GH62 sample remains unchanged while unencapsulated NHA GH62 has fully precipitated out of solution.
3.4 Discussion

Encapsulation offers much potential for engineering applications such as engineering reaction cascades for bioremediation and stabilizing insoluble enzymes for further biochemical studies. To date, several different micro-compartments have been engineered (including the P22 capsid, the CCMV capsid, and encapsulins) and tested for their potential uses in improving enzyme cascades. The compartments were shown to encapsulate heterologous cargo protein and enable enzymatic reactions to occur in the compartment lumen. However, challenges associated with using the aforementioned compartment systems for a range of industrial applications still remain. The limitations of the current systems include issues with substrate size and system complexity, as well as the limitations in the understanding of features that affect encapsulation and activity of encapsulated enzymes. In this study, I examine the use of the HK97 capsid as a compartment system as it does not suffer from the aforementioned limitations and is extremely well characterized, which will facilitate modifications to improve enzyme targeting and activity in downstream applications. I demonstrated that enzymes can be targeted into the HK97 capsid by fusion to a C-terminal targeting sequence on a construct containing the portal and major capsid genes. The enzymes are likely encapsulated in the capsid rather than attaching to the exterior of the capsid, as when capsids are purified using the co-targeting method, GFP is not co-purified with the capsid. If GFP adhered to the exterior of the capsid, it would have been purified upon mixing of GFP-expressing and capsid-expressing lysates, as well as in the co-transformation sample. The mechanism as to how the targeting peptide successfully interacts with the capsid interior is not fully known. Previous studies suggest that it might interact with the capsid delta domain [79], though there is currently no evidence directly confirming this interaction. Studies associated with the Eut and glycyl-radical micro-compartment systems revealed that targeted enzymes are successfully localized through peptides found in the Eut and glycyl-radical compartments containing a conserved hydrophobic motif, with levels of targeting being modulated by both the sequence used and the expression levels of target proteins [61, 182]. In this chapter, I demonstrated that modifying expression levels also led to changes to encapsulation quantities of GFP and PhoA. Hydrophobic regions of targeting peptides were postulated to be important for the correct assembly of the carboxysome and facilitating the initial interactions with the interior shell [183]. A more detailed understanding of targeting peptide
interactions within different compartment systems will be required in order to customize encapsulation, particularly for multi-enzyme systems.

Meanwhile, the targeting sequence for the native HK97 capsid has been shown to be required to target the protease into the capsid. It appears that other features may also contribute to capsid targeting. From my experiments I was able to show that some enzymes may be preferentially targeted for encapsulation, despite the presence of the same targeting peptide sequence. In this thesis, I observed preferential targeting of the native protease despite over-expression of GFP and PhoA with the same targeting sequence (through the co-transformation method). To date, there are very limited studies focused on the nature of the interaction between the protease and the components of the capsid interior (e.g., delta domain). Most research efforts were focused on the biochemical characteristics of targeting sequences [25]. While my current work has highlighted the protease’s ability to outcompete over-expressed enzymes with the targeting sequence, additional characterization of the protease / capsid interaction is necessary to understand how to either target proteins to the capsid efficiently or modulate targeting of enzymes to specific levels. In this chapter, I demonstrated that modifying charges in the capsid shell interior resulted in varying effects on the final quantities of enzyme pairs found within the capsid. This observation is even more surprising given that the protease interacts with the delta domain of the HK97 capsid protein (which is extended from the capsid shell interior) and not the capsid interior directly [79]. However, electron micrograph images in the study show that enzymes are packed close to the capsid shell (Figure 20), so it is not surprising that attraction or repulsion forces between the capsid interior and target enzymes play a role in dictating enzyme targeting levels. This suggests that considering a combination of hydrophobic and hydrophilic contacts will be important for future design of both single enzyme and multi-enzyme encapsulation systems, with the hydrophobic contacts from a targeting peptide making initial contacts and hydrophilic contacts enhancing both the capsid to targeted protein contact as well as capsid stability.
Figure 20. The interaction of the maturation protease with the capsid shell of an HK97 capsid.

A difference map between a HK97 Prohead I capsid encapsulating an inactivated protease and a HK97 Prohead I capsid without an encapsulated protease was generated to visualize the interaction of the protease and the capsid shell. The green circle represents the capsid shell proteins in the difference map. The blue line is an approximation of interaction area between the capsid shell and the protease, while the blue shaded area within the line represents the protease in the difference map. The delta domain of the capsid protein extends perpendicular from the capsid shell (red extensions from green circle towards the blue lines), with the protease interacting with the delta domain (red extensions from blue circle towards the blue line). The figure was adapted with permission from Veesler et al. [184]

Beyond the modulation of targeting, I was able to demonstrate that enzymes can be active when encapsulated within the HK97 capsid. One of the interesting observations from my activity assays was that activity of encapsulated alkaline phosphatase activity was significantly lower than its unencapsulated form. Several other studies have also observed this effect. Alcohol dehydrogenase D encapsulated within the P22 capsid had significantly lower activity [77]. This activity was recovered only when the P22 capsid was structurally altered to possess both greatly increased pore sizes and internal volume [46]. Alkaline phosphatase encapsulated within bacteriophage MS2 capsids exhibited slightly decreased the kinetic efficiency as compared to its unencapsulated form [75]. As highlighted in Section 1.4.5, a hypothesis as to why enzymes experience lower kinetic efficiency upon encapsulation is that increased encapsulation of enzymes negatively impacts substrate diffusion within the capsid due to molecular crowding, which leads to additional time required for substrates to diffuse to an enzyme’s active site. The enzymes I successfully targeted within the HK97 capsid are encapsulated at higher copy numbers than previous studies. The reduced activity of encapsulated PhoA as compared to unencapsulated PhoA from my studies is consistent with the aforementioned hypothesis. In the
context of metabolic channelling, having larger pores in the capsid shell to increase substrate diffusion and enzyme kinetic efficiency counteracts the principle of channelling, where substrate diffusion is meant to be limited. Results from my experiments also showed that altering the capsid interior charge can affect activity of encapsulated enzymes. To my knowledge, this phenomenon has not been reported for protein containers and represents a novel method to control both targeting levels and activity of multi-enzyme systems. More investigation into the mechanisms to facilitate increased enzyme activity in smaller capsid systems will be required to understand how containers can be adapted to maximize for metabolic channelling of multi-enzyme systems.

A significant challenge with utilizing protein compartments for engineering multi-enzyme systems in industrial or pharmaceutical applications is the ability to targeting a wide range of enzymes, particularly in high throughput studies which test different combinations of enzymes. I demonstrated that multiple enzymes could be targeted into capsids without prior fusion to each other, which has been the method of choice in previous studies to target multiple enzymes into a container *in vivo*. This is the first example of artificial co-encapsulation inside of a bacterial expression cell line of multiple enzymes without prior tethering. The homology-dependent cloning technique described in this thesis increases the throughput at which enzymes can be encapsulated inside the HK97 capsid. Any enzyme could be theoretically targeted into capsids by attachment of the C-terminal targeting peptide and co-expression with the capsid-producing plasmid. Using standard purification methods, large amounts of the encapsulated enzymes can be readily retrieved. Initial steps to characterize the encapsulation preferences of multi-enzyme systems shows that size and charge (of the capsid and target protein) are important factors; however, it is clear that there may be other factors influencing targeting that were not explored (e.g., surface charge of the targeted enzymes). Understanding all factors will be critical for controlling the targeting levels and activities of any encapsulated enzyme combinations.

While it is uncertain whether the same number of each enzyme are encapsulated in every capsid, it is likely not exclusive encapsulation of one enzyme, as demonstrated in the study where GFP and PalB were encapsulated in modified CCMV capsids *in vitro* [65]. Dual-colour fluorescence cross-correlation spectroscopy experiments of capsid samples where both GFP and PalB were encapsulated *in vitro* showed evidence of cross-correlation fluorescence, indicating that the two enzymes are present in close proximity within the same capsid. Cross-correlation
fluorescence was absent in capsid samples encapsulating only EGFP, only PalB and a mixture of EGFP only and PalB only capsid samples. Purified samples of both proteins mixed together did not yield any energy transfer. Together, these results indicated that both targets are co-encapsulated in the same capsid. Given that the natural maturation protease can outcompete proteins of a similar size that are expressed in much higher quantities and harbour the same targeting sequence, further analysis of the protease could reveal characteristics that influence targeting. A range of effects on different enzyme systems in this study was also observed upon mutating the capsid interior charge. Based on this observation, it is unlikely that any single method to elicit the same targeting and activity effects across all systems. These findings have important implications in compartment design and require further study in order to engineer and control multi-enzyme systems.

In all, I demonstrated that 22 insoluble target enzymes could be successfully encapsulated within HK97 capsids; however, encapsulation did not rescue activity for these proteins. While I was unable to detect activity of these enzymes upon encapsulation, another study was able to solubilize partially soluble, partially active α-galactosidase inside of P22 capsids in a form that yielded greater overall activity [46]. This illustrates the possibility of using capsids to help rescue enzyme activity. Combining the results of my thesis and the P22 study signified that encapsulation and purification of insoluble proteins does not necessarily recover full activity within the capsid. Target proteins will need to be able to fold in an active form in its unencapsulated state in order to demonstrate activity upon encapsulation. In addition, the ability of HK97 capsids to encapsulate a number of different enzymes may be beneficial for large scale encapsulation experiments. Current in vitro encapsulation methodologies, the complexities associated with micro-compartment assembly, and the requirement of subsequent processing encapsulated enzymes for the P22 capsid limit the ease in which one can target and purify a large collection of encapsulated enzymes. The ligation-independent cloning methodology and simple purification procedure described in this thesis will enable large scale encapsulation and downstream screening of activity in enzyme cascade systems, or detection of encapsulated enzyme activity for downstream biochemical characterization.

Although the HK97 capsid was unable to rescue activity from enzymes which form insoluble aggregates, the capsid did impart enzyme stability for active enzymes at higher temperatures. This phenomenon has been recently reported in the P22 capsid, where [NiFe]-
hydrogenase sequestered in the P22 capsid was found to be more stable at elevated temperatures and allowed for the production of hydrogen at previously non-permissive temperatures [51]. Beyond this thesis and the P22 capsid stabilization result, this phenomenon has not been observed for other protein containers. One comparable example involves the use of silica matrices, where confinement of lysozyme, α-lactalbumin, and metmyoglobin in silica prevented denaturation of the three enzymes, even in conditions of extreme pH, salt and temperatures that normally unfold proteins in free solution [49]. Capsids may likewise restrict the unfolding of tightly packed, encapsulated proteins, allowing for activity at temperatures that typically unfold the proteins. Since this phenomenon has not been reported for large protein compartments like micro-compartments, it may only apply to systems where enzymes are tightly packed.

In summary, I showed that HK97 can encapsulate targets in an active form and can also target combinations of multiple enzymes within the capsid interior. The HK97 capsid can successfully isolate insoluble enzymes and also stabilize enzymes at temperatures that precipitate the encapsulated enzyme. The results from my studies also highlight that certain factors of protein containers, such as internal container charge, can affect enzymatic activity and/or targeting of enzymes into the container. While some earlier predictions on the encapsulation pattern of multiple enzymes in this chapter were accurate, the varying accuracy of the predictions demonstrate the additional complexities that influence targeting. It also highlights the need to undertake further studies to reveal what additional factors must be considered to balance encapsulation levels of multi-enzyme systems. Once these factors are identified and fully understood, intelligent design of protein containers can be conducted to optimize the performance of future enzyme systems for downstream applications (e.g., reaction vessel, isolating unstable enzymes in an active form).
Chapter 4 – Future Directions

4.1 Summary

In this thesis, I investigated the HK97 capsid as a system for introducing metabolic channelling into enzyme systems of interest. Compartmentalization offers an opportunity to not only co-localize enzymes but also prevent metabolic intermediates from diffusing into the cytosol. This combination can improve upon existing systems currently used to engineer channelling, such as gene fusions and protein scaffolds, which may be exploited in industrial or pharmaceutical applications.

In Chapter 2, a computational survey was performed to build upon a previous metabolic channeling study performed by a previous lab member, Chris Sanford. A survey of enzyme pairings that share a common metabolite based on the KEGG database revealed over 9,000 unique enzyme pairs that could benefit from the introduction of metabolic channelling. The pairs were identified based on criteria which were discovered through either computational simulations or characterized systems in nature - the alleviation of thermodynamic constraints, the isolation of toxic intermediate compounds (as seen in glucosamine-6-phosphate synthase and imidazole glycerol phosphate synthase, where toxic ammonia is isolated through an intramolecular tunnel [185, 186]) and the isolation of inhibitory compounds. A comparison of the predicted channelling list to channelling reactions identified in nature, generated from gene fusion and protein-protein interaction data revealed a very small level of overlap. This could signify that gene fusions and protein complexes are not the optimal method for introducing channelling into enzyme systems in nature, and exploration of other mechanisms such as compartmentalization may yield more channelling candidates.

To evaluate whether channelling is isolated in nature or pervasive across species, gene fusion data was collated and analyzed to determine the array of phyla that contain gene fusions. Fusion were found across a wide spectrum of phyla, with clusters of fusions typically found within one phylum rather than across multiple phyla. Further analysis revealed that fusions are also commonly found in non-essential biochemical pathways, which reduces the likelihood of introducing detrimental / lethal changes for the host organism. Examining channelling
predictions and gene fusions across the global metabolic network revealed superpathways (sequences of consecutive enzymes linked through gene fusions and/or predicted to thermodynamically benefit from coupling) that signify potential areas for controlled metabolic flow across different areas of the network. Finally, as a solution to the lack of tools capable of displaying various layers of complexity in metabolic networks, the visualization tool Hyperscape was leveraged to show additional complexities inherent within metabolic pathways, such as the interplay between predicted channelling enzyme pairs and sequenced gene fusions. Hypergraphs and arced edges were able to better depict the directionality and complexity within these routes, presenting a viable option for effectively communicating pathway intricacies hidden within metabolic systems.

In Chapter 3, I characterized the potential of using the capsid from bacteriophage HK97 as a compartment. I demonstrated that fusing 63 residues from the C-terminus of the HK97 maturation protease with a protein or enzyme of choice resulted in the encapsulation of protein. Over-expression of the target enzyme resulted in near maximum levels of targeting based on the results of previous studies investigating the targeting of the natural maturation protease. Enzymes encapsulated in the HK97 capsid were active, but at a significantly reduced level in comparison to its unencapsulated form. This signified that substrates can diffuse into the capsid interior but likely at a significantly reduced rate. Insoluble enzymes were also successfully encapsulated but in an inactive form, indicating that capsids cannot enable proper folding unless the enzyme can fold into an active form on its own.

Following experiments with single enzyme systems, I attempted to target multiple enzymes into the same capsid and found that enzymes could be successfully encapsulated without prior tethering. Upon additional exploration of the targeted enzyme pairs, I observed that enzymes are encapsulated at varying efficiencies with protein charge a potential factor; however, this could not be definitively determined given the likely role of other factors driving the targeting variation. Alteration of the HK97 capsid’s interior charge (namely charge reversal of negatively charged residues) lead to changes in the targeting levels and activities of encapsulated enzymes. Finally, the HK97 capsid was shown to be able to protect enzymes (e.g. glycoside hydrolase) at temperatures previously known to denature target proteins. This highlights a potential application for capsids to impart added stability that can be exploited for use in possible industrial processes occurring at elevated temperatures/other non-native conditions.
4.2 Multiple capsid alterations to optimize multi-enzyme system encapsulation and activity for metabolic engineering

In Chapter 3, I demonstrated that the HK97 capsid can be altered to increase its targeting capabilities by modifying interior capsid charges. To improve existing targeting capabilities, an immediate next step is to simultaneously generate multiple mutations in the capsid interior. Along with the D136R mutation that was shown to influence the targeting levels of enzymes and enzymatic activity, additional mutations would be introduced on other residues that are negatively charged, neutral or positively charged, and not located near inter-capsid interaction interfaces. By making these mutations and targeting in alkaline phosphatase as well as a multi-enzyme system such as β-galactosidase and galactose dehydrogenase (chosen based on characterization of this enzyme system in a previous system [89]), a better understanding of the effects of multiple mutations can be used to fine-tune compartment systems in future metabolic engineering projects.

Another target for mutations is the C-terminal targeting peptide attached to targeted enzymes to be targeted in the HK97 capsid. Given that the targeting peptide is important for localization inside of the capsid, altering its sequence may alter targeting efficiency. Previous studies have shown that a shorter C-terminal stretch of 46 residues successfully targeted the protein GpD from bacteriophage lambda into HK97 capsids [79]. Since the exact mechanism of the interaction between capsid and the targeting sequence is not known, I would perform alanine scanning mutagenesis of the charged residues in the shorter targeting sequence to identify residues that are important in mediating the interaction. If mutations of individual residues do not lead to significant changes in targeting, alanine will be performed 5 residues at a time (i.e., mutating first 5 residues, mutating residues 6-10, etc.) One caveat to maximizing enzyme targeting the decrease in enzymatic activity, as overcrowding of enzymes within a capsid was previously shown to decrease kinetic efficiency [65]. Identifying residues that can increase or decrease targeting can be useful for optimizing enzyme stoichiometries and the kinetic efficiency of multi-enzyme systems to offset the effects of overcrowding.

A third set of mutations that can be generated are at the three-fold contacts of the capsid shell that make up pores. Closer analysis of the crystal structure revealed that short loops from the capsid protein protrude into the pores on the capsid shell, which may restrict the diffusion of
metabolites through the capsid shell [80]. To see whether these loops play a role in affecting substrate diffusion into the capsid, residues in the loops will be removed from the capsid-producing construct, prior to co-transformation and in vivo expression with PhoA fused to the C-terminal capsid targeting sequence. After confirmation that the altered capsid construct can successfully assemble PhoA encapsulated, PhoA activity in the altered constructs will be compared with PhoA activity in the wild-type capsid to assess whether removal of the capsid pore loops can increase activity (and hence, increase substrate diffusion through the capsid shell). In addition to removing the loops, residues near the pore may also play a role in substrate diffusion. Following a similar theme as the previously described charge mutation experiment (Section 3.3.5), select charged residues on the capsid protein located near/in the pore would be mutated to a neutral or an opposite-charged residue. Activities of PhoA targeted into the charge-altered and wild type capsids to evaluate whether pore charge alternations affect encapsulated enzyme activity. If enzymatic activity can be improved through capsid alternations, altering capsid loops and charges will represent viable options for maximizing desired output in future compartment systems.

4.3 Isolation of insoluble proteins purified within HK97 capsids for biochemical studies

In Section 3.3.3, I demonstrated that while many proteins can be encapsulated within HK97 capsids through attachment of the C-terminal targeting sequence and co-expression with the capsid protein, a protein that is not naturally active (soluble or insoluble) will not be encapsulated in an active form. I hypothesized that capsids can rescue enzymes in an active form if they can independently fold in an active form prior to aggregation, as recently demonstrated in the P22 capsid study where α-galactosidase was solubilized from the insoluble pellet in E. coli. To investigate whether the scope of enzymes rescued also includes enzymes that can be refolded following isolation from insoluble pellets, I plan to target 5-methylthioribose (MTR) kinase inside of the HK97 capsid. MTR kinase is important for the recycling of methionine, which plays an important role in protein synthesis and in methylation in various organisms [187]. Initial characterization of A. thaliana MTR kinase action required multiple refolding steps during the purification process. There is also the possibility that while the refolded enzyme was active, it was not refolded in its native form which limited its kinetic activity. If MTR kinase could be
captured in an active, native form within HK97 capsids, it would represent an important advancement in the simplification of structural and biochemical characterization of enzymes. After fusing MTR kinase to the C-terminal HK97 capsid targeting sequence and co-expressing it with the capsid protein, I plan to test for enzymatic activity by coupling the reaction of 5’ methylthioadenosine nucleosidase (MTAN) with MTR kinase. MTAN catalyzes the reaction of 5’ methylthioadenosine to 5’ methylribose, which is phosphorylated by MTR kinase using ATP [188]. The depletion of ATP can be tracked spectrophotometrically to determine whether MTR kinase is active when encapsulated by capsid in vivo.

Another potential application of HK97 capsids to explore is its utilization in isolating insoluble proteins for biochemical studies. While compartment systems such as the P22 capsid can successfully encapsulate insoluble enzymes, they are structurally stable and difficult to break apart [164]. As a result, breaking apart the capsid to isolate the encapsulated enzyme will likely disrupt the enzyme’s native structure and limit downstream investigation of the purified enzyme (e.g., structural studies). Given its relative instability, using Prohead I of the HK97 capsid can be more conducive for isolating the encapsulated protein without significant damage to the protein. It has previously been shown that purified HK97 Prohead I will naturally dissociate over a period of several weeks, indicating that Prohead I is not a long-term, structurally stable capsid form [11].

One potential method to recover encapsulated enzymes within isolated HK97 Prohead I capsids is by conducting short incubations of capsid samples at elevated temperatures. The HK97 capsids is also known to dissociate at elevated temperatures and cannot be reassembled at lower temperatures [11]. Given this observation, an encapsulated enzyme may potentially be isolated by incubation at an elevated temperature that does not completely denature the enzyme but facilitates dissociation of the capsid. Another method for dissociating capsids is by changing solvent conditions (e.g., pH, salt). The addition of compounds such as PEG, DMSO and glycerol have previously been shown to affect HK97 Prohead I assembly and may facilitate purification of the encapsulated enzyme [11, 164]. Assembly can be monitored using an agarose gel, where assembled capsids and capsomers from dissociated capsids have been shown to be easily distinguished upon visualization on a gel [11]. After conditions have been identified to dissociate capsids, α-galactosidase will be C-terminally tagged for encapsulation into HK97 and His-tagged for subsequent isolation from dissociated capsids. Capsids purified with α-galactosidase will be
subjected to conditions that successfully dissociate capsids before dilution in buffer for Ni-NTA purification. Once α-galactosidase is isolated from Ni-NTA beads, an activity assay measuring the hydrolysis of 4-nitrophenyl-a-D-galactopyranoside will be performed to determine the activity levels of the isolated enzyme. Activity in the purified enzyme samples will represent the feasibility in using capsid for isolating poorly behaved proteins that can be subjected to structural studies.

4.4 Identifying additional reaction pairings that can benefit from metabolic channelling for subsequent experimental testing and industrial application

As highlighted in Chapter 2, there are several sources of data that can be leveraged to predict additional pairs which would benefit from channelling. One area that was not considered when predicting enzyme pairs that would benefit from metabolic channelling was reactions that occur within micro-compartments. There is increasing evidence of the widespread nature of micro-compartments as well as new micro-compartments being identified which encapsulate reactions that have not been associated with compartments previously [32]. This new information provides new classes of reactions that may benefit from compartmentalization, such as enzymes that generate free radicals and glycyl radical enzymes, which utilize radicals of glycine and cysteine during catalysis [40]. Increased characterization of enzymes localized in compartments will drive progress in identifying drivers that necessitate their encapsulation. Understanding these drivers will enable the encapsulation of new enzyme combinations to drive channelling. Prioritizing predicted enzyme pairs in Chapter 2 for validation experiments will also be an important method for understanding drivers of channelling. By evaluating whether a kinetic efficiency increase is observed and the magnitude of any detected efficiency increase, the current understanding of channelling drivers can be verified based on the number of predicted channelling pairs experimentally confirmed to benefit from channelling. The priority pairs to target would be the highest scoring pairs as identified on Tables 1 and 2, as the large magnitude difference between the free energies of the reactions will be expected to generate the greatest increase in kinetic efficiency upon encapsulation.

Another potential source of data that could reveal enzymes that benefit from channelling is looking at gene fusions and protein interaction data that did not overlap with the predicted
channelling pairs. As discussed in Chapter 2, only a very small portion of gene fusion EC pairs and protein-protein interaction pairs, which depict complexes, were found to overlap with enzyme pairs predicted to benefit from channelling. By taking a closer look at the gene fusion pairs, new commonalities could be identified which may lead to the discovery of new channelling mechanisms. Some aspects of these pairs to analyze include sharing disregarded non-currency compounds (i.e., electrons) and their occurrence with transport systems, which were omitted from this study. A third source of data for identifying new instances of channelling are metagenomic datasets, generated by sequencing environmental samples and performing genomic analyses on the aggregated pool of genomic information [189]. The key advantage with using metagenomic datasets is the large amount of sequence data available due to the presence of multiple organisms – this provides an opportunity to identify new gene fusions and new compartments are that not present within currently sequenced genomes.

Finally, a solution for visualizing multiple features within a metabolic network has been proposed by utilizing hypergraphs within a metabolic network. As seen in Figure 10b, I have been able to integrate information about gene fusions, predicted channelling direction, and metabolic reaction direction to describe the pentose phosphate pathway. While traditional methods of visualization would have created significant difficulties interpreting all the layers of information together, the use of hyperedges and arcs allow for clear delineation of information despite significant overlap in the enzymes and metabolites involved in channelling and gene fusions. Evaluating areas of the metabolic network with large concentrations of hyperedges may reveal enzyme pairs that may benefit from channelling through bypassing metabolic flow from gene fusions or energetically favourable reactions).

4.5 Implications of Thesis Findings

In the content of general scientific literature, the results from computational and experimental studies presented in this thesis corroborate existing observations in the field and provide additional insight into areas such as network visualizations, multi-enzyme encapsulation and enzyme protection. Computational studies in this thesis generated a pool of enzyme pairs predicted to benefit from channelling that is twice as large as pools from previous studies and revealed a limited overlap of predicted channelling enzymes with enzyme fusions and enzyme
complexes. Enzyme fusions involving biochemically subsequent enzymes were pre-dominantly found in non-essential pathways and mostly limited to select phyla. Finally, leveraging a new network visualization alternative using hyperedges enabled easier rationalization/hypothesis generation of the purpose of certain genetic / biochemical phenomena (e.g., gene fusions). Given the limited number of comprehensive channelling studies present in literature [86, 126], the methodologies and associated findings discussed in this thesis represent advancements towards identifying instances of channelling and the biochemical reasons.

Results from experimental studies on enzyme targeting and the activity of encapsulated enzymes were also consistent with existing literature. At the time of discovery, the ability of HK97 capsids to stabilize glycoside hydrolases at elevated temperatures was the first example of protein compartments stabilizing enzymes at non-native temperatures. Prior to the discovery, silica gel pores were the only system which protected enzymes at elevated temperatures [49]. A subsequent publication revealed that P22 capsids provided the same protection to [Ni-Fe] hydrogenase when temperatures were elevated to 60°C [51]. While the modification of internal compartment charge affecting targeting levels is not a novel discovery [59], the change in activity as seen with alkaline phosphate has not been reported in literature to the best of my knowledge. Finally, the multi-enzyme targeting experiments represented the first instance of \textit{in vivo} targeting multiple enzymes into protein compartments without prior tethering. A subsequent publication was able to target multiple enzymes into the MS2 capsid using the SpyCatcher system [190].

Together, the advancements highlighted above will facilitate future studies of multi-enzyme systems which have a higher likelihood of biochemical improvements through encapsulation and provide an experimental protein compartment system that can be modified to alter enzyme encapsulation efficiencies and enzyme activity within the compartments. The ability to target any pair of enzymes into the capsid using the ligation-independent methodology described in this thesis enables large-scale targeting of enzymes for increasing metabolic efficiency, enzyme purification and characterization, and enzyme stabilization for downstream applications.
5.0 References


119

120. Hayasaka Y, Hayasaka S, Nagaki Y: Ocular changes after intravitreal injection of methanol, formaldehyde, or formate in rabbits. *Pharmacol Toxicol* 2001, **89**:74-78.


143. Rondon MR, Kazmierczak R, Escalante-Semerena JC: Glutathione is required for maximal transcription of the cobalamin biosynthetic and 1,2-propanediol utilization (cob/pdu) regulon
and for the catabolism of ethanolamine, 1,2-propanediol, and propionate in Salmonella typhimurium LT2. *J Bacteriol* 1995, **177**:5434-5439.


151. Yang YT, San KY, Bennett GN: *Redistribution of metabolic fluxes in Escherichia coli with fermentative lactate dehydrogenase overexpression and deletion.* *Metab Eng* 1999, **1**:141-152.


Appendix A – Supplemental Figures

Supplemental Figure 1. Encapsulation of insoluble dehalogenases within HK97 Capsids

Induced cell lysates and samples processed through the capsid purification procedure were analyzed on SDS-PAGE stained with Coomassie Blue dye. The uncleaved major capsid protein (42 kDa), portal protein (47 kDa), dehalogenase TceA (65 kDa) and dehalogenase CfaA (55 kDa) are highlighted using arrows.
Appendix B – List of Supplemental Tables

Supplemental Table 1. Enzyme pairs predicted to benefit from channelling by overcoming thermodynamic barriers.

Enzyme pair scores (column A) are taken calculated by taking the absolute value of the product of the free energy values of each reaction in a reaction pairing (column I and J). The sources for the free energy values of the substrates (column K and M) and products (columns L and N) are also listed. Henry represents values determined by the Mavrovouniotis group-contribution method while NIST represents values obtained from the NIST database.

Supplemental Table 2. Enzyme pairs predicted to benefit from channelling by sequestering toxic intermediates.

Supplemental Table 3. Enzyme pairs predicted to benefit from channelling by sequestering inhibitory intermediates.

Supplemental Table 4. Proteins containing multiple EC number designations identified in the SwissProt database.

Note: The supplemental Tables can be accessed on TSpace, U of T’s online thesis portal. The URL for TSpace is: https://tspace.library.utoronto.ca/