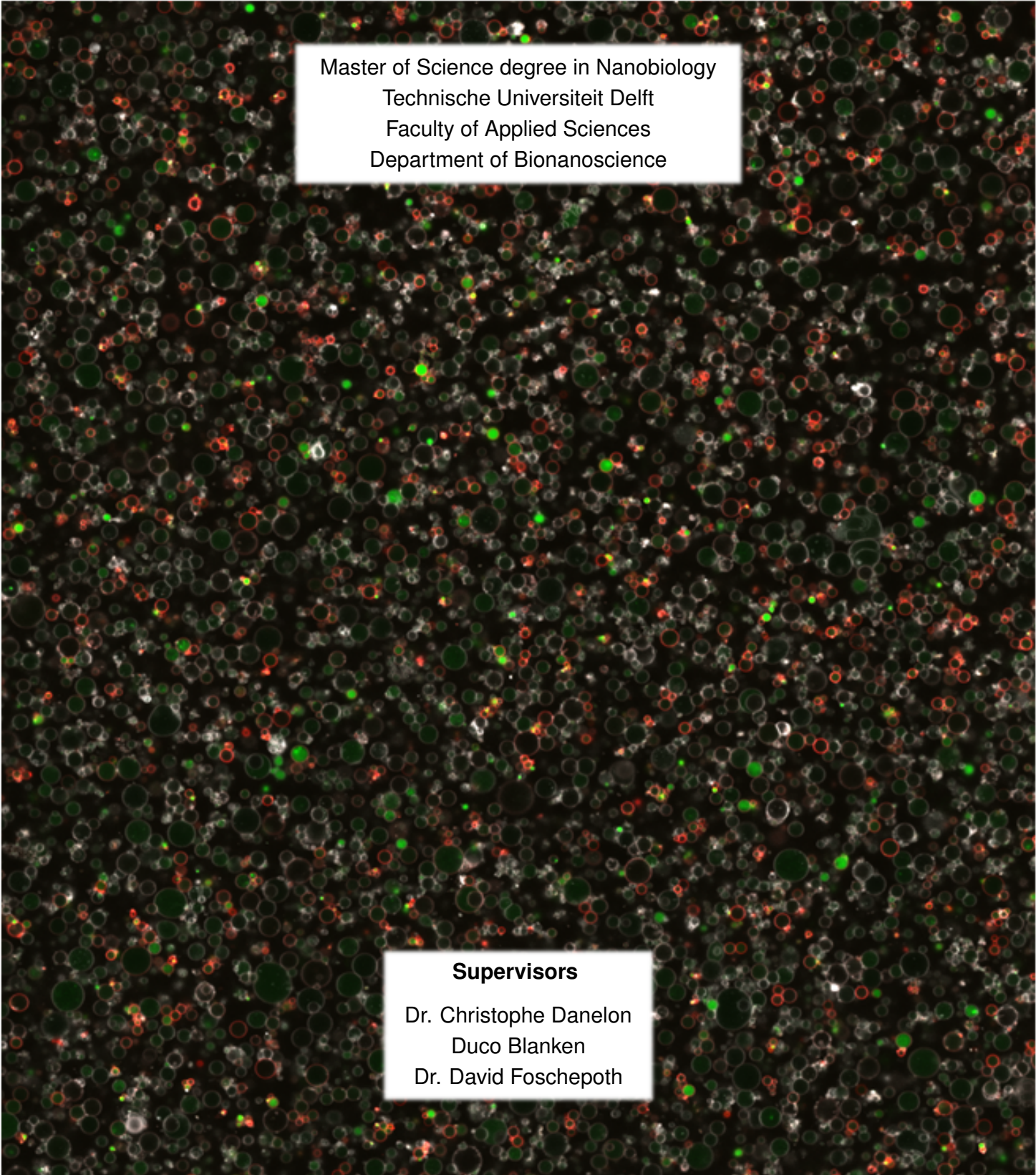


# Complicating things: how to combine lipid synthesis with DNA replication in liposomes

Mats van Tongeren

September 2020

A microscopic image showing a dense population of liposomes. The liposomes appear as small, circular structures of varying sizes, some with bright green or red fluorescence, and others appearing as faint outlines. They are distributed across a dark background.

Master of Science degree in Nanobiology  
Technische Universiteit Delft  
Faculty of Applied Sciences  
Department of Bionanoscience

## Supervisors

Dr. Christophe Danelon  
Duco Blanken  
Dr. David Foscipoth



# Abstract

The construction of a bottom-up minimal cell requires the eventual integration of reconstituted functional modules, like lipid synthesis and DNA replication. These modules have separately been reconstituted inside liposomes by using the protein synthesis using recombinant elements (PURE) cell-free gene expression system to express the *E. coli* Kennedy lipid synthesis pathway and the genes of the phage  $\Phi$ 29 DNA replication machinery. However, simultaneous expression has yet to be attempted. In this project, we aim to achieve module integration and describe the functional behavior of the integrated modules. To test this, the genes of the two systems are co-encapsulated with the PURE system in liposomes under various experimental conditions. The liposomes were incubated with the LactC2-mCherry fluorescent probe and dsGreen intercalating dye. LactC2-mCherry binds specifically to phosphatidylserine (PS). Using confocal imaging, colocalization of LactC2-mCherry on the liposome membrane and increased dsGreen intensity inside the liposome was used to indicate successful expression of the lipid synthesis genes and successful DNA amplification respectively. SMELDit, a new image analysis tool enabling flow cytometry-like data analysis, was developed to quantify the small fraction of liposomes exhibiting the activity of both modules simultaneously. We detected simultaneous DNA replication and lipid synthesis in individual liposomes. We also characterized the expression in liposomes of the pMAR2 plasmid, which contained the DNA replication and lipid synthesis genes with  $\Phi$ 29 origins of replication. We found conditions under which pMAR2 showed both lipid synthesis as well as DNA amplification, confirmed that pMAR2 is capable of self-replication and showed that the expression of the two modules could be tuned by varying the RNAP concentrations. We believe this study demonstrates the successful integration and regulation of two functional modules inside the liposome-based minimal cell, using newly developed image analysis software. It also reveals several new challenges that emerge from the interactions of the two modules.

# Acknowledgements

I would like to thank Christophe for the opportunity to work in this group and his feedback and guidance during the project. Your feedback, especially in the last couple months, helped me improve my skills as a researcher. You kept your door open for questions and were always available and ready with thoughts and advice when I needed it.

I would like to thank Duco for being a great supervisor. Without your advice, motivation and ability to play the forward role in a kicker team, I would not have made it through this project. I wish you the best for your upcoming PhD defence and your future career outside this lab.

I would like to thank David for helping me out with some practical lab skills in the beginning of my project. I would like to also acknowledge that pMAR2 was made by David and Marit, without it my project would have been very different.

I would like to thank the other members of the lab as well, Anne for helping me with the proteomic measurements, Zhanar for teaching me how to use the qPCR machine and Elisa, Ana and Ilja for helping me out many times. I had a great time with all of you!

Finally, I would like to thank all the people I shared an office with throughout my project. We might not have had the best luck in taking care of plants, but it was always fun environment and I enjoyed being a part of it.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The minimal cell . . . . .	1
1.2	The Danelon lab implementation of a minimal cell . . . . .	2
1.3	$\Phi$ 29 DNA replication . . . . .	3
1.4	Kennedy metabolic pathway and PS . . . . .	3
1.5	Lactadherin binding for PS detection . . . . .	5
1.6	Motivation for co-expression of DNA replication and lipid synthesis . . . . .	7
1.6.1	The strengths and weaknesses of the reconstituted modules . . . . .	7
1.6.2	Directed evolution as a future research direction . . . . .	8
1.7	Research Goals . . . . .	10
<b>2</b>	<b>Materials and Methods</b>	<b>11</b>
2.1	DNA constructs used in this project . . . . .	11
2.1.1	G340 . . . . .	11
2.1.2	pGEMM7.0 . . . . .	11
2.1.3	pMAR2 . . . . .	11
2.2	GUV production . . . . .	12
2.2.1	Lipid-coated beads . . . . .	12
2.2.2	Plasmid amplification . . . . .	13
2.2.3	Restriction enzyme digestion . . . . .	13
2.2.4	PCR amplification of oriLR- <i>p2-p3</i> from G340 plasmid . . . . .	13
2.2.5	IVTT in liposomes . . . . .	13
2.2.6	Lipid precursor film . . . . .	14
2.2.7	Natural swelling to produce GUVs . . . . .	14
2.3	Sample imaging . . . . .	14
2.3.1	Production of imaging chambers . . . . .	14
2.3.2	Imaging chamber preparation . . . . .	14
2.3.3	Confocal microscopy . . . . .	14
2.4	Auxiliary sample analysis . . . . .	15
2.4.1	Lipidomics using liquid chromatography mass spectrometry . . . . .	15
2.4.2	Proteomics using QconCAT . . . . .	16
2.4.3	Trypsin digest . . . . .	16
2.4.4	LC-MS/MS analysis . . . . .	16
2.4.5	Quantitative polymerase chain reaction . . . . .	16
2.4.6	qPCR primers . . . . .	17
2.4.7	DNA analysis on agarose gel . . . . .	17
<b>3</b>	<b>SMELDit, a newly developed image analysis tool</b>	<b>19</b>
3.1	Motivation . . . . .	19
3.2	Function . . . . .	19
<b>4</b>	<b>Results</b>	<b>21</b>
4.1	SMELDit can identify DNA amplification phenotypes . . . . .	21
4.2	Combining LactC2-mCherry probe with Cy5 membrane staining . . . . .	22
4.3	Combining DNA replication with lipid synthesis . . . . .	23
4.4	Amplifying the lipid synthesis genes . . . . .	24

4.5	Expression of pMAR2 can be tuned using varying RNAP concentrations in the $\Delta$ T7 PURE system . . . . .	26
4.5.1	pMAR2 is capable of self-replication in the absence T7 RNAP . . . . .	28
4.6	Bulk analysis of pMAR2 expression shows little to no activity . . . . .	29
<b>5</b>	<b>Discussion, Conclusion and Outlook</b>	<b>31</b>
5.1	Combining the expression of two systems has a cost . . . . .	31
5.2	Replication of pMAR2 fails at times with no clear cause . . . . .	31
5.3	Adjusting the T7 RNAP concentration affects lipid synthesis and DNA amplification . . . . .	32
5.4	Expression from pMAR2 in liposomes is different from bulk . . . . .	33
5.5	Replication of pMAR2 could be limited by its length . . . . .	34
5.6	Future pMAR constructs . . . . .	35
5.7	dsGreen could stain RNA . . . . .	35
5.8	Outlook on directed evolution as a future research direction . . . . .	36
5.8.1	Luminex flow cytometry . . . . .	36
5.8.2	Deep learning-guided directed evolution . . . . .	37
<b>A</b>	<b>Appendix A</b>	<b>A1</b>
A.1	SYBR Gold stained gel failed to detect the amplified pMAR2 . . . . .	A1
A.2	Standard curves for six-gene qPCR analysis . . . . .	A2
A.3	Investigation of the effect of dNTPs on lipid synthesis was likely detecting technical error . . . . .	A3
A.4	Bulk experiments show no amplification in qPCR measurements . . . . .	A4
<b>B</b>	<b>Appendix B</b>	<b>B1</b>
B.1	SMELDit Manual . . . . .	B1
B.1.1	SMELDit setup . . . . .	B1
B.1.2	Performing the image processing . . . . .	B1
B.1.3	Loading, Analysing and Plotting the liposomes . . . . .	B1
B.2	SMELDit Code . . . . .	B3
B.2.1	Splitting the channels of the input image automatically . . . . .	B3
B.2.2	Analysing, Indexing and Saving Liposomes . . . . .	B4
B.2.3	Loading and Plotting indexed liposomes . . . . .	B7

# Acronyms

## Acronyms

**Acyl-CoA** Acyl Coenzyme A.

**CDP-DAG** Cytidine diphosphate-diacylglycerol.

**CdsA** Cytidine diphosphate -diacylglycerol synthase A.

**CL** Cardiolipin.

**CTP** Cytidine Triphosphate.

**DNAP** DNA polymerase.

**DOPC** 1,2-dioleoyl-sn-glycero-3-phosphocholine.

**DOPE** 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine.

**DOPG** 1,2-dioleoyl-sn-glycero-3-phospho-(1,-rac-glycerol.

**DOPS** 1,2-dioleoyl-sn-glycero-3-phospho-L-serine.

**DSB** Double Strand Binding protein.

**DSPE PEG(2000)-biotin** 1,2-distearoyl-sn-glycero-3-phosphoethanolamine-N-[biotinyl(polyethylene glycol)-2000].

**DSPE PEG(2000)-Cy5** 1,2-distearoyl-sn-glycero-3-phosphoethanolamine-N-[amino(polyethylene glycol)-2000]-N-(Cyanine 5).

**EDTA** Ethylenediaminetetraacetic Acid.

**G3P** Glycerol-3-Phosphate.

**GPAT** G3P Acyl Transferase.

**GUV** Giant Unilamellar Vesicle.

**IVTT** *In Vitro* Transcription and Translation.

**LactC2** Lactadherin C2 domain.

**LC-MS** liquid chromatography mass spectrometry.

**LPAAT** Lysophosphatidic Acid Acyl Transferase.

**PA** Phosphatidyl Acid.

**PE** Phosphatidylethanolamine.

**PG** Phosphatidylglycerol.

**Pgpa** Phosphatidylglycerophosphatase A.

**PgsA** Phosphatidylglycerophosphate synthase A.

**PS** Phosphatidylserine.

**Psd** Phosphatidylserine decarboxylase.

**PssA** Phosphatidylserine synthase A.

**PURE system** Protein synthesis Using Recombinant Elements.

**qPCR** quantitative Polymerase Chain Reaction.

**RNAP** RNA polymerase.

**SMELDiT** Show Me Example Liposomes Damn it.

**SSB** Single Strand Binding protein.

**TP** Terminal Protein.

# 1 Introduction

Thinking about life is foundational to the human experience. Therefore, the question "What does a living being need to do to be alive?" has been asked in many different contexts. The nature of what we learn from (partially) answering this question can range from the societal and philosophical to the strictly mechanistic.

For the mechanistic case, the difficulty lies not so much in producing a reasonable answer to this question, but rather in minimalizing and generalizing the answer. After all, if a simpler set of requirements produce life on their own, any more expansive list of requirements would be too strict. The smallest and simplest living unit found in nature is the cell. So, if the goal is to describe the simplest form of life, the practical task will become finding the simplest living cell, a minimal cell. In the next section, the conceptual basis of the minimal cell will be examined and approaches to its construction addressed.

## 1.1 The minimal cell

A minimal cell is a system consisting of the minimal yet sufficient molecular components to be defined as alive [1]. The constraints to what system qualifies as a minimal cell rest in the definition of "alive". In other words: what does a living being need to do to be alive? Fortunately, the Danelon lab does not need to find a new answer, as there are some well established notions in defining what is considered alive. In their work, Luisi et al. [1] identify three main requirements a cell needs to meet in order to be considered alive: **metabolic homeostasis**, **reproduction**, and **evolution**. With the goalposts for being alive now set, the definition of the minimal cell can be rephrased to be: a system consisting of the minimal yet sufficient molecular components to maintain metabolic homeostasis, reproduce itself and evolve over time.

With a functional definition of "alive" set, the effort towards a minimal cell now becomes an optimization problem. What strategy will lead to the most simple cell? At first it might seem logical to find the simplest living thing in nature, but there are some complications with that approach. For one, natural organisms generally have genes that can be deleted without losing any of the functional requirements necessary for life. The presence of a gene in a living organism on its own is therefore not enough proof that this gene is essential for life. So, most genomes encode more function than is essential for life. A second problem would be that many of the organisms with extremely small genomes are obligatory symbiotes, meaning these organisms do not meet the requirements to be considered alive outside their host environment. Nevertheless, nature does provide a starting point from which a minimal cell could be built by altering existing organisms. Synthetic biology makes use of existing biological elements to engineer new functions. Synthetic biology can be split into two approaches: top-down and bottom-up [2].

A top-down synthetic biology approach to a minimal cell starts with an existing organism and removes its non-essential parts until any other simplification would render it unable to sustain its life. Generally, this kind of research translates to finding a minimal genome by successive rounds of gene deletion and screening for viability. The state of the art in top-down minimal cells is the minimized genome of *Mycoplasma mycoides*, produced by Hutchison et al. [3]. After successive rounds of deletions, their synthetic genome JCVI-syn3.0 has 473 protein-expressing genes left from the 1016 of the wild type *Mycoplasma mycoides*, while maintaining independent viability. With 473 genes, JCVI-syn3.0 is smaller than all known genomes of natural free-living organisms. However, 149 of those 473 genes could not be assigned a function. This lack of understanding of the minimized system is not a fluke, but rather an inherent problem with this top-down approach to the minimal cell. The successive rounds of deletion are not based on *a priori* knowledge about the system, but rather on the direct effect of the genes, deletion on overall viability of the cell. This function-agnostic approach speeds up the process of genome minimization, at the cost of mechanistic understanding of what the minimal cell is physically doing to survive. A mechanistic understanding of the minimal cell might be formed through thorough analysis of the system after the cell has already been minimized using the top-down approach, but this results in problems of functional disentanglement. If a cell component is embedded inside a wider unknown system, it can become nearly impossible to attribute a cell function to that component.

The straightforward path to mechanistic understanding is to describe the function of each component in isolation. These components could then be added back together in a fully synthetic minimal cell. The bottom-up synthetic biology approach to the minimal cell is exactly that strategy of isolating cell components, studying their function and then merging them back together into a completely new minimal cell,

without using an existing organism as a scaffold for cell function. A bottom-up minimal cell therefore does not need to source its components from a single organism. Consequently, sourcing components from different organisms or even viruses allows these components to be chosen to be as efficient and simple as possible, at the cost of the possibility of unknown negative interactions between individual components. Currently, the bottom-up project is constrained by the available knowledge and characterization of cell components. The current state of the art on the bottom-up minimal cell is therefore more concerned by studying individual reconstituted systems rather than full-scale synthetic cell production. Nevertheless, the field has matured to a phase where research has started to focus on the merging of these reconstituted functional systems. To get into the specifics of what this means for the research in the minimal cell project, the next section will describe the philosophy, implementation and current state of the minimal cell project in the Danelon lab.

## 1.2 The Danelon lab implementation of a minimal cell

The Danelon lab approaches building a minimal cell bottom-up with a combination of cell-free gene expression and function from purified proteins. The overarching philosophy of the lab is that all functions necessary for life fall into three categories: **compartmentalization** of cellular function, a functional **flow of information** and **proliferation** of the cell. It should be noted that these functional groupings do not map one-to-one on the requirements of life, nor are they necessarily the most minimal functional groupings, as they are derived empirically out of commonalities in the function of known organisms. There is a possibility that some unknown organisms can live without one of these categories of cell functions.

**Compartmentalization** is the set of functions responsible for instantiating and maintaining the cell compartment. The cell compartment provides a sectioned off environment with a controlled inflow and outflow of nutrients and is therefore an important requirement of homeostasis. The cell compartment also separates the genome of the minimal cell from the genomes of other cells. This isolated expression of a genome allows for distinction between cell phenotypes, where phenotype is defined as the interaction of the genome's expression with its environment. This makes the cell compartment crucial for forming an evolvable unit, as without phenotype there could be no phenotype-based natural selection.

In the Danelon lab, compartmentalization is achieved by co-encapsulating cell components in liposomes. A liposome is a spherical vesicle consisting of at least one phospholipid bilayer. Phospholipids are amphiphilic, with a hydrophilic head group and long hydrophobic tails. In aqueous solution they are able to self-organize into lipid bilayers, with the hydrophilic head group facing outward, thereby minimizing the interaction of the non-polar lipid tail of the phospholipid and the polar water molecules outside the bilayer. Liposomes may not necessarily be the least complex form of cellular compartment, but phospholipid cell membranes occur widely in nature and are often necessary for the function of membrane proteins. The liposomes used for encapsulation in the Danelon lab are made to be unilamellar to match the natural conditions of the cell membrane. With their radius of around three micron they can be classified as Giant Unilamellar Vesicles (GUVs) [4].

**Flow of information** is the set of cell functions responsible for expressing the genetic information of the minimal cell's genome as proteins. This reconstitution of the central dogma is the scaffold upon which other cell function can be reconstituted. In the Danelon lab, the flow of information from DNA to RNA to proteins is reconstituted using the Protein synthesis Using Recombinant Elements (PURE) system. The PURE system is in itself a minimal system for accomplishing *in vitro* transcription and translation (IVTT), being able to synthesize proteins from a given DNA template in a cell-free environment [5]. All together, the PURE system contains RNA polymerase (RNAP), ribosomes, nucleotides, amino acids, tRNA, transcription factors and energy storage and regeneration machinery. The PURE system contains purified components from mostly *Escherichia coli*, but also components from other sources like phage T7 RNAP.

**Proliferation** is the set of functions responsible for replicating the cell. This includes the growth and eventual division of the cell compartment, regeneration and replication of the cellular machinery involved in the flow of information and replication of the minimal cell's genome. Realizing full proliferation of the minimal cell while maintaining compartmentalization of cell functions and the flow of information would make the Danelon minimal cell meet the three requirements of life.

Reconstituting proliferation is an ongoing field of research. Recently, many of the functions of proliferation have been individually reconstituted in the cell-free environment such as modest steps to compartment

growth [6], an ongoing effort to reconstitute the Min system involved in cell division in *E. coli* [7] and cell-free self-replication of DNA [8].

In this project, the combination of two functional systems using components from both *Escherichia coli* and phage  $\Phi$ 29 is studied with the intention of describing the interactions of these systems. In the next sections of the introduction I will address the phage  $\Phi$ 29 DNA replication machinery and the *E. coli* Kennedy lipid metabolism pathway used in the experiments. Subsequently, the choice of these two systems as candidates for early integration will be motivated. Finally, the research goals of this project will be addressed.

### 1.3 $\Phi$ 29 DNA replication

A minimal cell needs to replicate its own genetic material to proliferate. Since the genomic material used in the PURE system is DNA, the Danelon lab has focused on the reconstitution of DNA replication. DNA replication system should be isothermal, since the PURE system also has a narrow temperature range in which it is active. The Danelon lab has landed on the use of the isothermal DNA replication system of phage  $\Phi$ 29 [8].

The  $\Phi$ 29 DNA replication system was first reconstituted by Blanco and Salas [9]. The reconstituted  $\Phi$ 29 DNA replication system consists of only four genes: the DNA polymerase (DNAP, *p2*), a terminal protein (TP, *p3*) and two auxiliary proteins that bind to single stranded DNA (SSB, *p5*) and to double stranded DNA (DSB, *p6*) [10]. The auxiliary proteins are not required for DNA replication but increase the efficiency of the process. The terminal protein (TP) functions by binding to  $\Phi$ 29 polymerase (DNAP) when present in solution with ammonium cations [11]. The TP-DNAP complex will then be recruited to the origins of replication at the ends of a linear DNA by parental TP (fig. 1). Parental TP is the TP left from a previous round of replication. If no parental TP is present, the DNAP-TP complex will be recruited to the origins of replication at a lower efficiency. The DNA replication is then initiated by DNAP which will start to unwind and process along the DNA, leaving the TP bound at the origin of replication for priming the next round of replication. This protein-primed nature of the  $\Phi$ 29 DNA replication machinery requires no primer oligonucleotides or temperature changes during replication. Protein-primed replication also preserves all of the sequence information during a round of replication, which is important for sustainable self-replication.

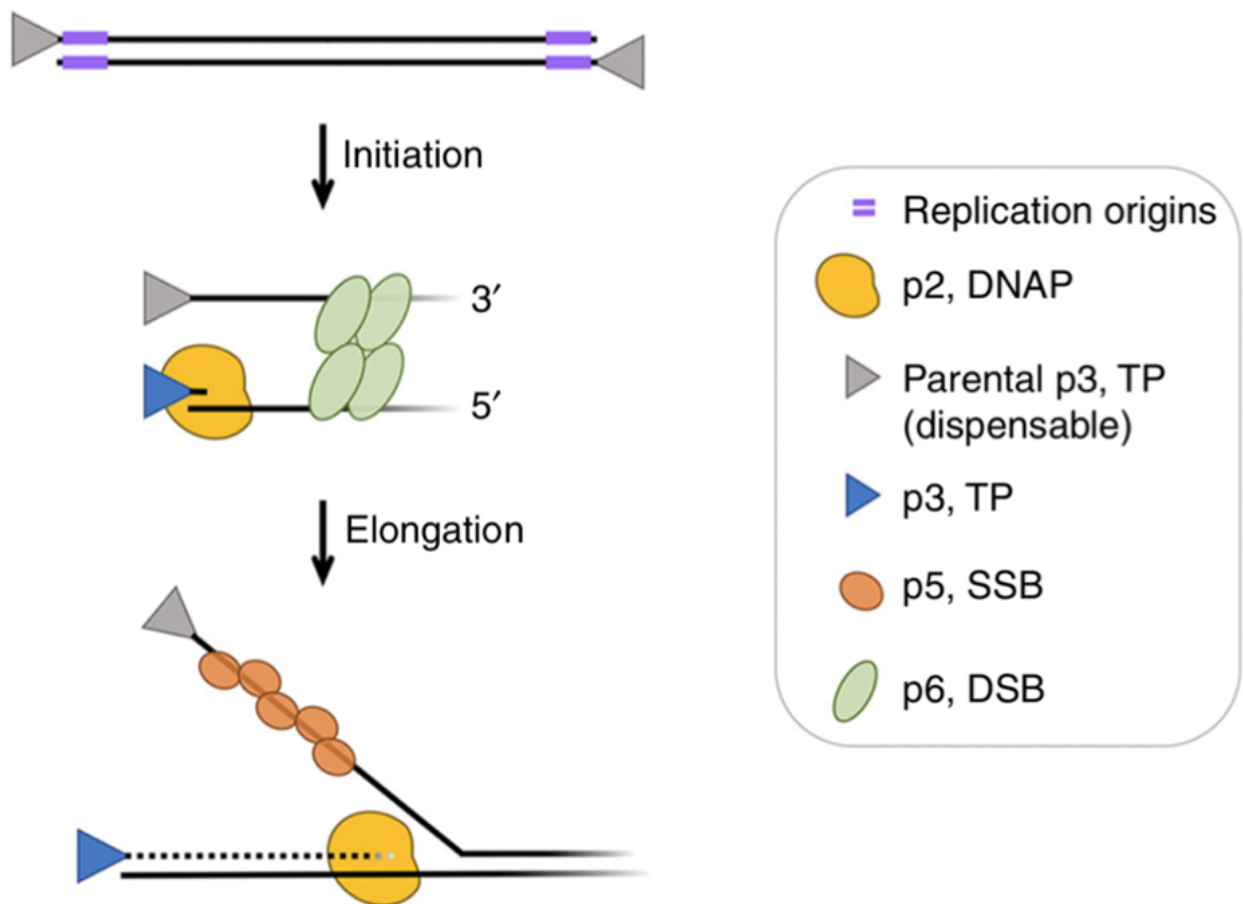
In addition to the need for proliferation, the minimal cell needs to also be evolvable. DNA replication plays a key role here as well. The minimal cell should be capable of self-replication of its genome, in order to be considered evolvable. The replication machinery of  $\Phi$ 29 is well suited to this task, as its components can all be expressed using the PURE system. It has been previously shown in Nies et al. [8] that, self-replication of template DNA encoding the  $\Phi$ 29 DNA replication machinery can only be achieved when the auxiliary proteins are added in purified form.

In this project,  $\Phi$ 29 DNA replication machinery is expressed to replicate template DNA flanked by origins of replication. In some cases the replicated DNA was also the DNA encoding the DNAP and TP genes. For the purposes of this project this is considered self-replication of the DNA template.

### 1.4 Kennedy metabolic pathway and PS

A minimal cell needs to be able to grow and divide to proliferate. Sustainable cell growth implies a doubling of the compartment membrane between division events. Therefore, in order to avoid death by dilution, a minimal cell should also be able to grow its compartment. Since the cell compartment is a liposome, compartment growth takes the form of phospholipid synthesis. However, it is not enough to produce phospholipids. The lipid synthesis of the minimal cell should also match the variety of phospholipids of the lipid composition of its membrane. This is important because different phospholipids have different headgroups, which differ in charge and size and can affect protein binding. The length and saturation of the carbon chain is also a factor as it changes the structure of the membrane [12] affecting permeability, melting temperature and intrinsic curvature [13].

The Kennedy pathway is the metabolic pathway responsible for the biosynthesis of the most abundant plasma membrane lipids in *E. coli* (fig. 2), which includes phosphatidylethanolamine (PE), phosphatidylglycerol (PG), and cardiolipin (CL). The Kennedy pathway has two main branches, one responsible for the synthesis of PE and the other for PG and CL. The pathway starts with long-chain acyl coenzyme A



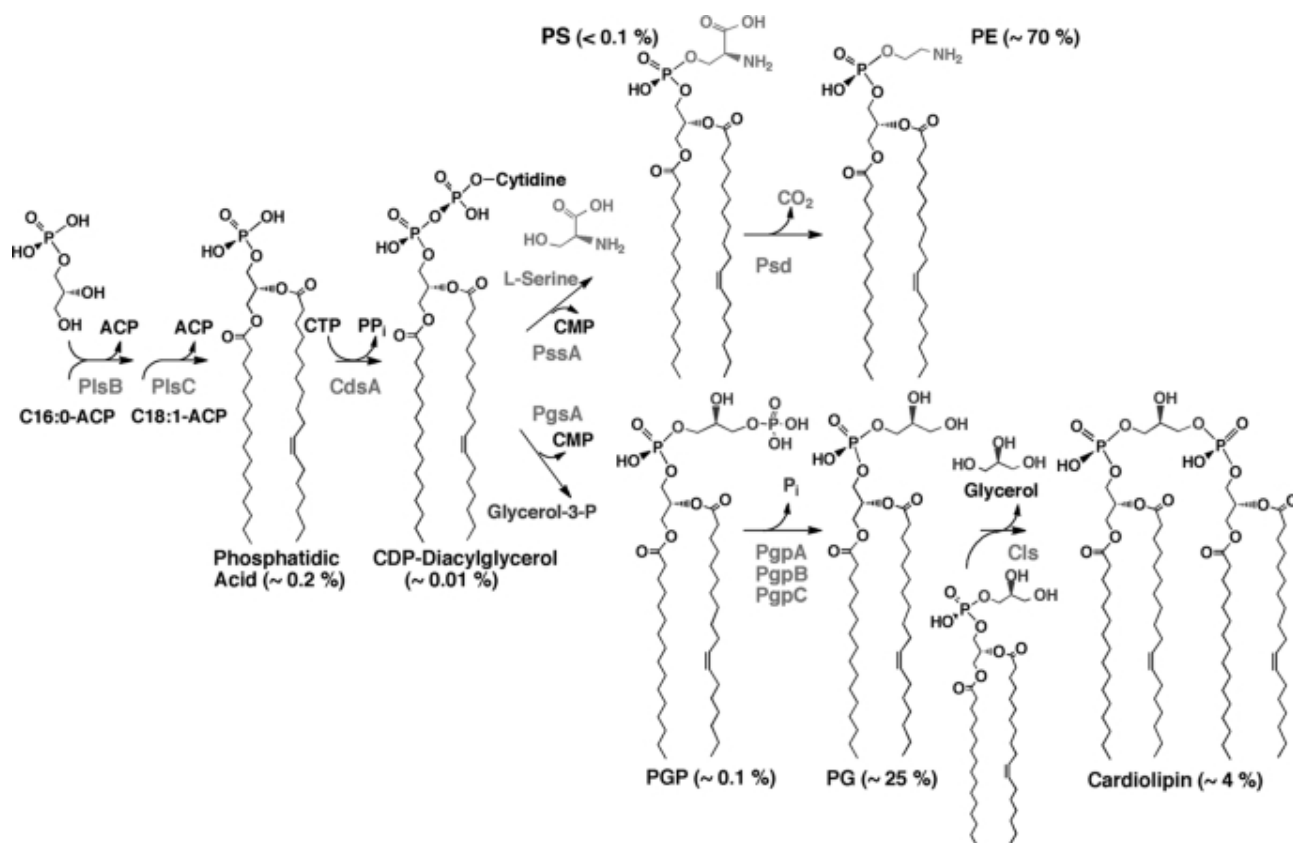
**Figure 1: Schematic representation of the  $\Phi 29$  DNA replication initiation and elongation.** DNA replication is initiated by parental TP recruiting the DNAP-TP complex to the origins of replication, which is made more efficient by the presence of DSB protein. DNAP then elongates the DNA until the replication is complete. While the replication fork moves, the single-stranded DNA is stabilized by SSB proteins. Adapted from Nies et al. [8].

(Acyl-CoA) and glycerol-3-phosphate (G3P), which are metabolized into phosphatidic acid (PA) by G3P acyl transferase (GPAT) and lysophosphatidic acid acyl transferase (LPAAT). The respective genes for these enzymes are *plsB* and *plsC*. PA is the universal precursor to all other lipids in the Kennedy pathway. It can accumulate in the plasma membrane both *in vivo* and in liposomes [6]. However, PA can also be efficiently converted to other commonly occurring lipids. PA is converted to its downstream metabolites by modification of its head group. Cytidine diphosphate -diacylglycerol synthase A (CdsA) catalyses the activation of PA with cytidine triphosphate (CTP) into cytidine diphosphate-diacylglycerol (CDP-DAG).

CDP-DAG itself is an intermediate metabolite that makes up a small fraction of the lipid composition of the plasma membrane, but is nevertheless central to not only the Kennedy pathway but also other pathways responsible for the synthesis of different phospholipids such as phosphatidylcholine (PC) [14].

The Kennedy pathway splits into its two branches after the production of CDP-DAG (fig. 2). The branch producing PG starts with the phosphatidylglycerophosphate synthase A (PgsA) enzyme followed up by the phosphatidylglycerophosphatase A, B or C (PgpA, B or C) enzyme. The branch towards PE starts with the phosphatidylserine synthase A (PssA) enzyme producing phosphatidylserine (PS) from CDP-DAG and L-serine. The PE synthesis branch of the pathway is more relevant for the rest of this project, which is centred around the synthesis of PS.

PS is an intermediate in the synthesis of PE. It is possible in some species of bacteria to have accumulation of PS in the plasma membrane, but the conversion of PS to PE by phosphatidylserine decarboxylase (Psd) is usually too efficient for PS accumulation. The focus on PS synthesis might therefore be seen as counter-intuitive as its biological relevance is limited. However, the efficiency of Psd also places the metabolic bottleneck of the Kennedy pathway with other enzymes. Therefore, detecting the level of PS



**Figure 2: The Kennedy pathway is responsible for the synthesis of the most abundant phospholipids in the *E. coli* plasma membrane.** The genes responsible for each step indicated in grey in this figure. The lipid composition fraction of each phospholipid in *E. coli* is included within parentheses. It should be noted that these are averages. Lipid composition can change with environmental conditions and varies between organisms [15]. Figure from Lu et al. [16].

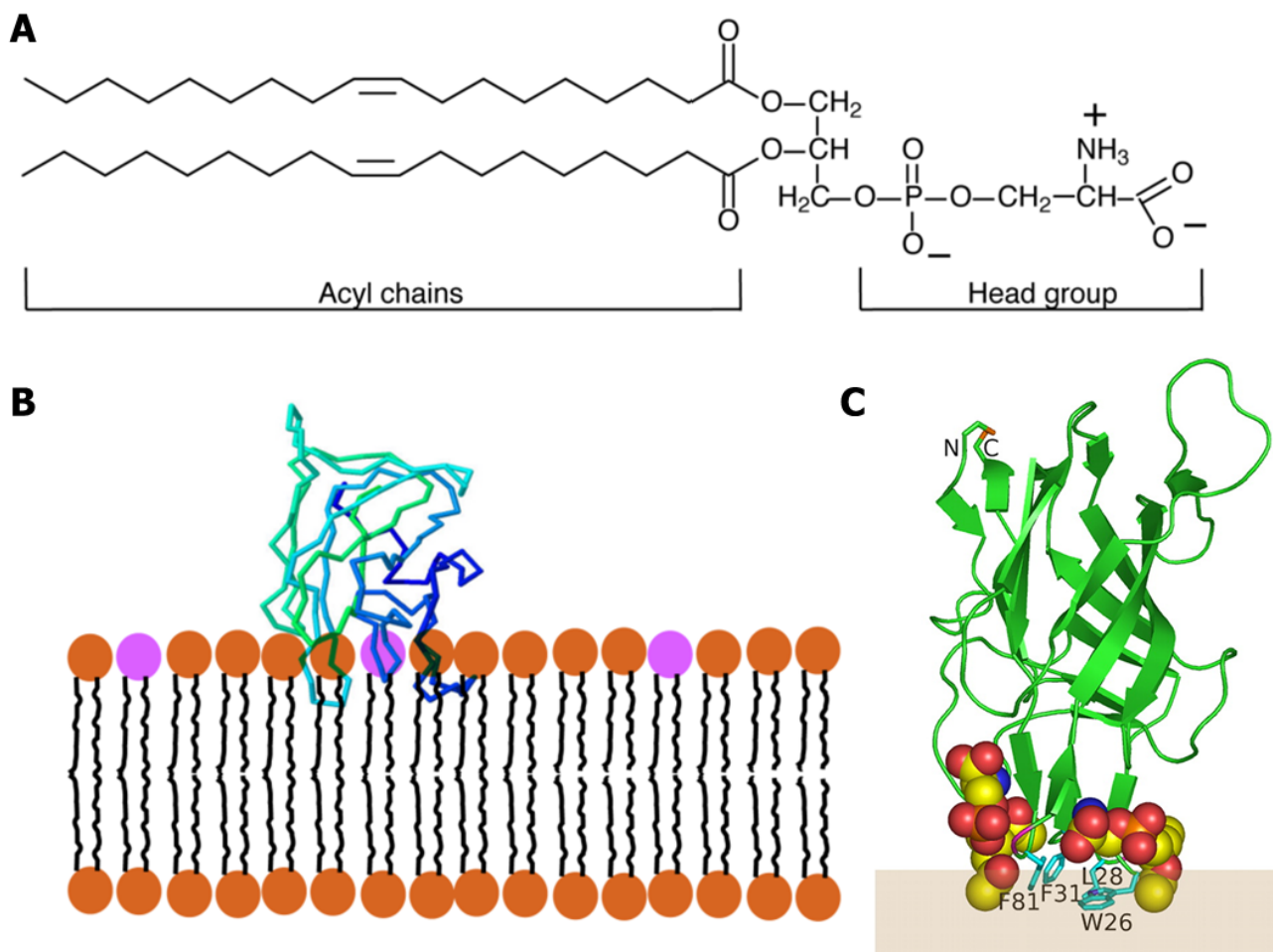
accumulation in the membrane in the absence of Psd is functionally very similar to detecting the accumulation of newly synthesized PE in the presence of Psd. The most important practical reason for synthesizing PS is the availability of well-described fluorescent probes that bind to PS specifically [17]. This project is specifically interested in describing the behavior of co-expression of lipid synthesis with DNA replication. An important part of this description is studying co-expression of the two functions within the same liposome. Fluorescent probes can indicate lipid synthesis on the level of individual liposomes. While PE can be detected and differentiated from the initially present PE by isotopic labelling in lipidomics measurements, however these measurements are fundamentally bulk measurements. PS synthesis is both measurable on the single liposome level and representative for the more biologically relevant PE synthesis. So, this project will focus on PS synthesis to represent the function of the wider Kennedy pathway.

PS synthesis can be reconstituted inside liposomes by expressing the four lipid synthesis genes using the PURE system (*plsB*, *plsC*, *cdsA* and *pssA*) with all the necessary lipid precursors present in solution. With these components present, the ability of the minimal cell to grow its own compartment can be quantified by measuring the PS accumulation in the lipid membrane of the liposome.

## 1.5 Lactadherin binding for PS detection

The PS detection probe used in this project was a fusion protein of the fluorophore mCherry and lactadherin C2 (LactC2) domain. LactC2-mCherry has been characterized in previous work in this lab [20]. In this project, LactC2-mCherry is used to report PS synthesis on the level of individual liposomes (fig. 4).

Lactadherin was initially discovered in cow milk as a protein bound to the outside of phospholipid bilayers surrounding triglyceride droplets, with the apparent goal of stabilizing said bilayers [21]. Bovine lactadherin has four domains: two epithelial growth factor-like (EGF-like) domains and two C domains with homology to the C1 and C2 domains of blood clotting factor V and factor VIII [22]. The lactadherin C2

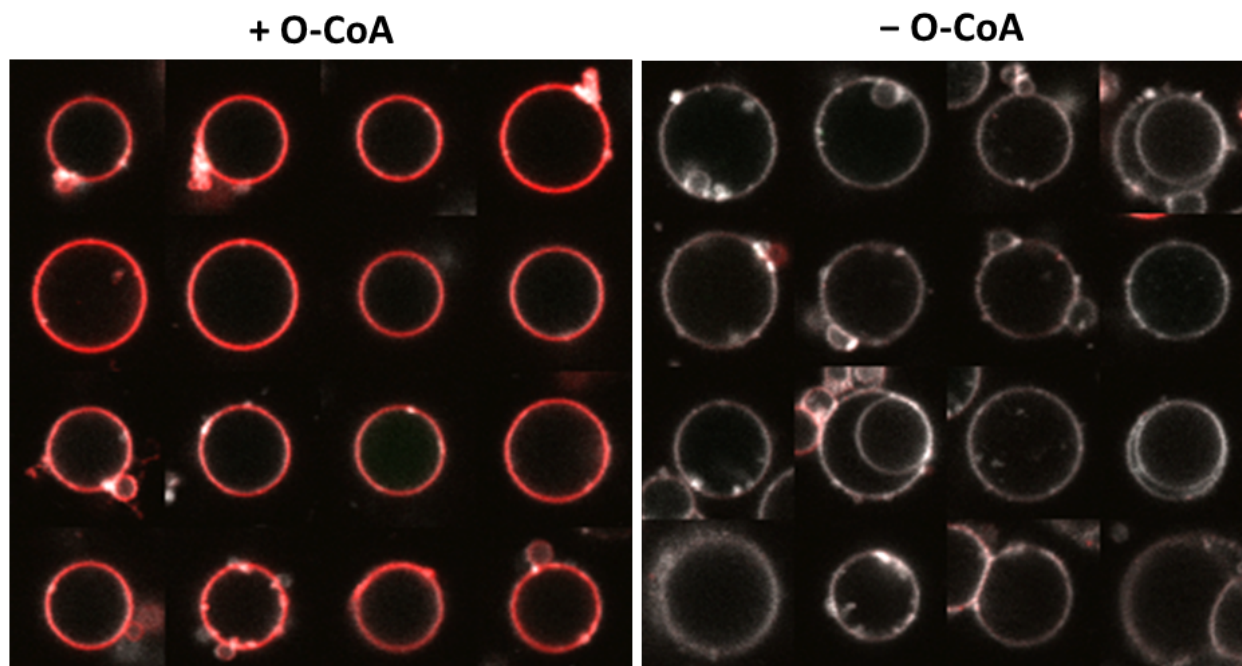


**Figure 3: Structure of LactC2 bound to PS.** (A) Diagram of PS with its anionic head group and two (18:1) oleoyl chains. (B) PS binding of the C2 domain of lactadherin. The PS headgroups are colored purple, other headgroups are colored brown. The C2 domain is depicted as a stick model with its hydrophobic domains sunken into the lipid bilayer. The LactC2 crystal structure is from the RSCB Protein Data Bank. (C) model of LactC2 (ribbon representation) and its hydrophobic residues (stick representation) and the two L-serine head groups it is bound to (volume representation). Figure (A&B) adapted from Kay and Grinstein [18] and (C) from Shao et al. [19].

(LactC2) domain has a similar function to the blood clotting factors by binding specifically to PS L-serine headgroups in lipid bilayers (fig. 3). Since LactC2 binds specifically to PS it can be fused to a fluorescent protein such as Green Fluorescent Protein (GFP), to create a fluorescent probe that is specific to PS [23]. The emission/excitation spectral properties of mCherry are more suitable than those of eGFP when used simultaneously with the other fluorophores in this project (fig. 6).

PS might be a small fraction of the lipid composition in bacteria, but in eukaryotes PS is the most abundant anionic phospholipid in the plasma membrane [24]. In healthy eukaryotic cells, PS is almost exclusively found on the inner (cytoplasmic-facing) leaflet of the plasma membrane. This asymmetric distribution over the leaflets is not a result of passive self-organization but rather an active process by ATP-dependent aminophospholipid flippases [25], a flippase being a protein transporting a phospholipid from the outer leaflet to the cytoplasmic-facing leaflet. Not to be mistaken for a floppase, which is a protein that transports phospholipids from the cytoplasmic-facing leaflet to the outer leaflet.

The presence of PS in the outer leaflet in blood platelets is a signal to start blood coagulation [26] mediated by the binding of several clotting factors. Similarly, a loss of PS asymmetry between the leaflets can be an indicator of apoptosis and will mark an apoptotic cell for phagocytosis mediated through PS binding factors like lactadherin [27].



**Figure 4: Example of liposomes stained with LactC2-mCherry** The liposomes depicted here contained DNA expressing the lipid synthesis genes in the presence of LactC2-mCherry with and without the necessary Oleoyl-CoA (O-CoA) lipid precursor. The mCherry signal is shown in red and the Cy5 signal is depicted as white. Co-localization of the red signal at the membrane indicates PS accumulation in the membrane.

## 1.6 Motivation for co-expression of DNA replication and lipid synthesis

The direct motivation for combining two functional modules is the current lack of a proper description of the consequences of co-expression. There are immediate predicted problems with the co-expression of the two systems. For instance, the gene expression achievable with the PURE*flex* system is currently limited by the depletion of translational machinery [28]. It is to be expected that co-expression of two systems will have a negative effect on the overall expression. Both modules will have to be expressed using the same limited resource pool, now shared. In previous work it has been shown that other proteins could be expressed without compromising the yield of synthesized fluorescent proteins, so this resource-limited view of expression is not universally true [29]. However, the resource-sharing behavior of the two modules used in this project is yet to be described. This project aims to describe this resource-sharing behavior by measuring the activity of both systems at various experimental conditions.

However, to understand the specific choice to combine lipid synthesis with DNA replication inside liposomes, the context of this project in the effort toward a bottom-up minimal cell needs to be described. After all, other cell function modules have also been reconstituted within the framework of the bottom-up minimal cell, so why combine lipid synthesis and DNA replication specifically?

DNA replication and lipid synthesis are function modules with complementary strengths and weaknesses. Additionally, the combination of DNA replication with another functional module opens up a new avenue for future research in the form of directed evolution.

### 1.6.1 The strengths and weaknesses of the reconstituted modules

What can be predicted about the function of the two modules on the scale of a hypothetical future minimal cell? Of the two, DNA replication can be seen as the senior partner in terms of successful reconstitution. In bulk measurements of DNA amplification by the expression of  $\Phi$ 29 DNA replication machinery the measured fold increase in DNA copies can reach  $\geq 50$  within the time span in which the PURE system remains active [8]. Experiments with lipid synthesis in liposomes encapsulating the same PURE system within the same time span never reach levels high enough to observe compartment growth using an optical microscope [6]. From the holistic perspective of the minimal cell, the bottleneck in the hypothetical doubling

time would be, between the two, the compartment growth, since both the compartment and the DNA need to double between division events.

However, the reality of combining the two modules could be entirely different. The limiting factor of the lipid synthesis is likely the solubility of the acyl-CoA in aqueous solution. This would mean that the reduction in produced enzymes caused by the co-expression does not necessarily translate to a drop in synthesized lipids. Where the lipid synthesis is enzymatic, the TP needs to stay bound to each replicated DNA copy. This means that a substantially larger number of DNA replication machinery proteins need to be produced to sustain function over time, compared to the lipid synthesis enzymes that can keep performing their function with new substrates. Therefore with equally shared resources, lipid synthesis might be less hindered in its function than DNA replication.

Another consideration is the length of the amplified DNA. With the introduction of four lipid synthesis genes the doubling time of the DNA might substantially increase. In previous work, the complete  $\Phi$ 29 viral genome has been replicated by expressed *p2* and *p3* in the PURE system [30]. The  $\Phi$ 29 genome has a length of about 20 kb, whereas the oriLR-*p2-p3* used in the experiments of Nies et al. [8] has a length of 3.2 kb. The resulting amplification fold was in fact up to an order of magnitude lower when amplifying the full genome. This would suggest that, when the genome of the minimal cell will get scaled up to encode more functions, the speed of the DNAP, not the production of TP, limits the achievable amplification fold within the time frame in which the PURE system is active.

To put it succinctly, DNA replication is currently producing high amplification folds, but might be hampered by amplification of longer DNA constructs. Lipid synthesis on the other hand, is currently not efficient enough to double the cell compartment before the PURE system becomes inactive likely due to low solubility of acyl-CoA. Moreover, the lipid synthesis enzymes are much less efficient at converting more soluble precursors. Hence, these current functioning modules can best be seen as proves of concept that need to be replaced with better suited alternatives in a fully fledged bottom-up minimal cell. However, it may not be necessary to replace the current modules with homologous modules from nature. Instead, the high amplification fold of the  $\Phi$ 29 DNA replication machinery can possibly be leveraged to improve the currently used modules through directed evolution.

### 1.6.2 Directed evolution as a future research direction

Beyond the direct value of describing the co-expression of two functional modules in the bottom-up minimal cell, this project can also be seen as enabling future research directions. The most promising of these directions is the possibility of engineering the modules of the minimal cell using directed evolution.

Directed evolution is the process of optimization of fitness by successive rounds of mutation, screening and selection [31][32]. Directed evolution can be understood as mimicking natural evolution *in vitro* with artificial selection pressure to coax towards a desired end state. In the case of modules of a bottom-up minimal cell for example, the selection could be towards a DNAP with faster DNA replication or lipid synthesis enzymes with a higher efficiency in converting acyl chain substrates other than acyl-CoA. Directed evolution is particularly useful to the bottom-up minimal cell project as a way to improve cell module function [33].

Directed evolution does not require *a priori* knowledge about the cause of increased function of the module. In contrast, each new cell component introduced to the bottom-up minimal cell needs to be characterized extensively. This process can be accelerated by instead evolving new cell components that meet the needs of the minimal cell from cell components that are already characterized. This is especially true when co-expressing two modules, since a new combination of modules will always run the risk of negatively interacting with each other. Components co-evolved using directed evolution, on the other hand, are likely to work well together.

In practice, directed evolution of the minimal cell would take the form of constructing libraries of gene variants, screening these for their function and selecting the best gene variants to be the starting points of a new library. The optimization cycle is limited by what can be screened for and the throughput of the screening. Some of the most high-throughput screening methods involve the use of flow-cytometry cell sorting techniques to automate the screening and selection step. Fluorescence-activated cell sorting (FACS) has been used for directed evolution [34], with both high throughput and fluorescence as a convenient way of screening for function. With FACS the fluorescent probes necessary for screening do not

need to be more complex than the ones used in this project, making it a promising future direction of this project. The recent advances in the artificial intelligence based image-activated cell sorting (IACS) [35] could take this concept one step further by sorting the cells based on more complex features in the images, like membrane deformation.

The LactC2-mCherry probe could allow future directed evolution experiments to target the PS synthesis using these or similar high-throughput screening methods. Such a directed evolution experiment could be used to engineer a variant of *p/sB* that more efficiently converts acyl-chain substrates with better solubility in water.

DNA replication as a module of the minimal cell is useful in two ways to directed evolution. Firstly, genes with high replication activity will become a larger fraction of the gene pool with each amplification cycle. So their function can be selected for by simply re-encapsulating the DNA of the previous cycle into new liposomes and repeating the process until only the best performing DNA replication genes are left. Secondly, the inclusion of error-prone DNA replication could be a way to automate the library creation step of the directed evolution process.

This project brings the Danelon lab one step closer to a future directed evolution experiment by describing the co-expression of two modules that are prime candidates for improvement through directed evolution. This project also aims to describe the resource-sharing behavior of the two modules as well as the effect on DNA replication of amplifying a larger DNA template containing the lipid synthesis and DNA replication genes. Therefore, this project is a necessary and important step towards applying directed evolution to the improvement of the modules of the bottom-up minimal cell.

## 1.7 Research Goals

The main aim of this project is to characterize the behavior of combined lipid synthesis with DNA replication inside liposomes. This aim leads to the following key objectives and sub-objectives for the project:

1. Validate compatibility of the established lipid synthesis and  $\Phi$ 29 DNA replication system fluorescent probes with Cy5 membrane staining separately.
  - Validate the newly developed image analysis software, SMELDiT, can be used for its intended goal of producing individual liposome phenotypes from image analysis.
  - Image both modules expressed individually in liposomes that contain DSPE PEG(2000)-Cy5 with both dsGreen and LactC2-mCherry present.
2. Characterize the co-expression of the two functional modules from the same DNA template, called pMAR2, a construct containing lipid synthesis genes and the  $\Phi$ 29 *p2* and *p3* genes under orthogonal promoters.
  - Characterize the expression of lipid synthesis enzymes from pMAR2 in liposomes.
  - Characterize the expression of  $\Phi$ 29 DNA replication machinery from pMAR2 and test the feasibility of pMAR2 self-replication.
3. Establish experimental conditions under which expression of both modules is detected.
  - Establish experimental conditions under which lipid synthesis and DNA replication can be detected in an individual liposome
  - Establish experimental conditions under which the lipid synthesis enzymes are expressed and the lipid synthesis genes are replicated.
4. Quantify performance of the combined modules and describe the resource-sharing behavior by varying the levels of expression of the two.
  - Use different combinations of DNA constructs to vary the levels of expression of both modules.
  - Use various concentrations of T7 RNAP to modulate the expression of the lipid synthesis enzymes.

In this project, we show how pMAR2 can be used to reconstitute both lipid synthesis and DNA replication simultaneously inside liposomes, explore and establish conditions under which expression can be biased towards one of the two modules and confirm the full self-replication of pMAR2.

## 2 Materials and Methods

### 2.1 DNA constructs used in this project

#### 2.1.1 G340

The G340 plasmid contains the *p2* and *p3* genes under T7 promoters as well as the OriR194 and OriL191  $\Phi$ 29 origins of replication (fig. 5A). The plasmid has a length of 5,935 bp, but it can also be linearized by partial PCR amplification. The resulting linear PCR product (oriLR-*p2-p3*) has a length of 3,213 bp and contains the *p2* and *p3* genes while leaving the origins of replication on the flanks of the DNA. These exposed origins allow for TP to bind which enables the initiation replication.

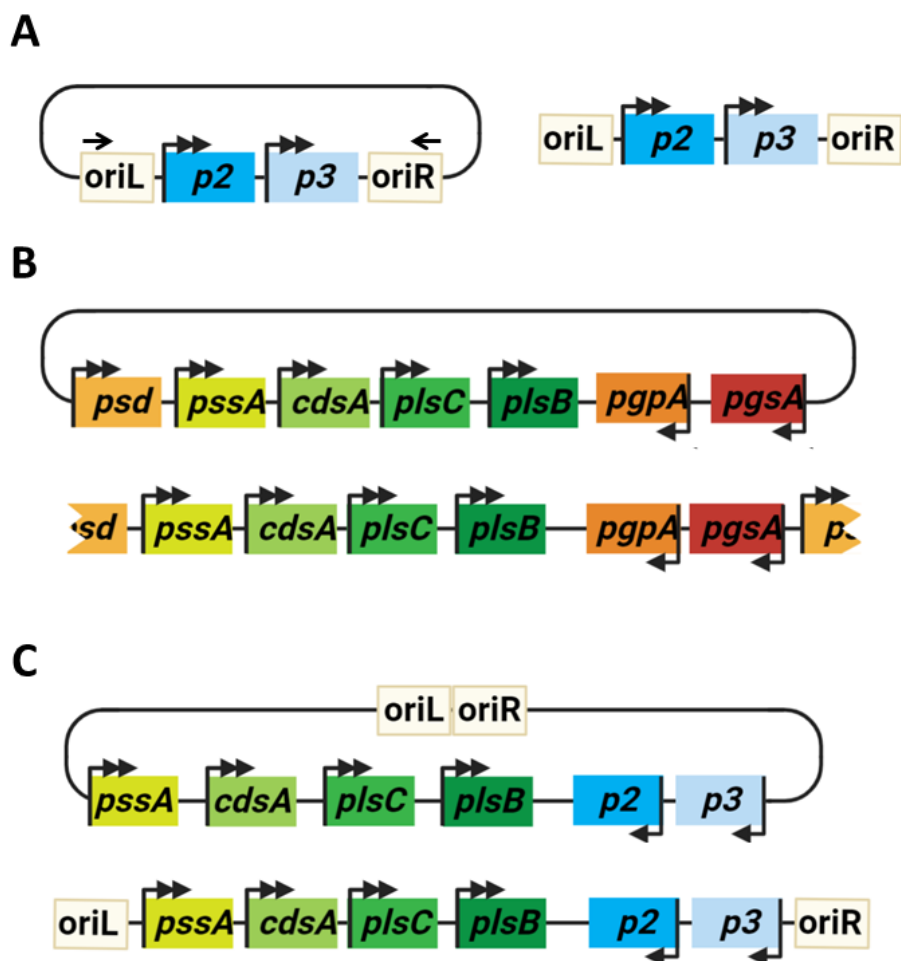
#### 2.1.2 pGEMM7.0

The pGEMM7.0 plasmid has a length of 11,105 bp and contains the seven genes of the Kennedy pathway: *plsB*, *plsC*, *cdsA* and *pssA* under a T7 promoter and *psd*, *pgsA* and *pgpA* under a SP6 promoter (fig. 5B). The genes with an SP6 promoter were placed on the opposite strand of the plasmid to prevent read-through transcription.

When pGEMM7.0 is linearized using *EcoRI* restriction enzyme, the *psd* gene is inactivated. The linear product (pGEMM7.0( $\Delta psd$ )) therefore does not express the Psd enzyme, which results in PS as the final product of the lipid synthesis.

#### 2.1.3 pMAR2

The pMAR2 plasmid has a length of 11,601 bp and contains the four genes of the Kennedy pathway responsible for phospholipid synthesis up to PS: *plsB*, *plsC*, *cdsA* and *pssA* under a T7 promoter and the  $\Phi$ 29 DNA replication genes *p2* and *p3* under an SP6 promoter (fig. 5C). Similarly to pGEMM7.0 the genes with the SP6 promoters were placed on the opposite strand of the genes with a T7 promoter to prevent transcriptional read-through. The pMAR2 plasmid also contained the OriR194 and OriL191  $\Phi$ 29 origins of replication. The pMAR2 plasmid could be linearized by digestion using the *PmeI* restriction enzyme. The resulting linear product contained the two origins of replication at the flanks of the DNA construct. These exposed origins of replication should allow the amplification of full length pMAR2 by the  $\Phi$ 29 DNA replication machinery.



**Figure 5: Schematic representation of the DNA constructs used. (A)** G340 and its linear PCR product *oriLR-p2-p3*. **(B)** The pGEMM7.0 plasmid and the pGEMM7.0( $\Delta psd$ ) linear digestion product. **(c)** The pMAR2 plasmid and its linearized digestion product.

## 2.2 GUV production

### 2.2.1 Lipid-coated beads

Lipid-coated beads are produced using a 10 ml round-bottom flask. First the flask was rinsed with chloroform. Then, 50 mol% 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC), 36 mol% 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine (DOPE), 12 mol% 1,2-dioleoyl-sn-glycero-3-phospho-(1,-rac-glycerol) (DOPG), 2 mol% 1,3-bis[1,2-dimyristoyl-sn-glycero-3-phospho]-glycerol (CL), 1 mass% 1,2-distearoyl-sn-glycero-3-phosphoethanolamine-N-[amino(polyethylene glycol)-2000]-N-(Cyanine 5) (DSPE PEG(2000)-Cy5) and 1 mass% 1,2-distearoyl-sn-glycero-3-phosphoethanolamine-N-[biotinyl(polyethylene glycol)-2000] (DSPE PEG(2000)-biotin) were added to the flask. After each addition the tip was flushed out into the flask twice with an equal volume chloroform for more accurate lipid concentrations. 100 mM rhamnose in methanol was added such that the final volume ratio of chloroform : rhamnose-in-methanol was 5 : 2. 600 mg of 212-300  $\mu$ m glass beads were weighed off and added to the flask. The flask was placed in the rotary evaporator and slowly brought to a pressure of 200 mbar. It was then evaporated for 2 hours while being rotated at 35 rpm. The flask was slowly returned to atmospheric pressure and removed from the device. The beads were gently removed using a micro spatula and placed into aliquots. The aliquots were then desiccated overnight and stored under argon.

In some instances the 12 mol% DOPG was replaced with 12 mol% 1,2-dioleoyl-sn-glycero-3-phospho-L-serine (DOPS) to serve as a positive control for the synthesis of PS. These two types of beads will be denoted as PG-beads and PS-beads if they contained DOPG or DOPS respectively.

### 2.2.2 Plasmid amplification

Plasmids were amplified by growing transformed One Shot TOP10 chemically competent *E. coli* (Thermo Fisher Scientific) cells from glycerol stock. The cells were plated out on 50 µg/ml ampicillin LB agar plates. The plates were incubated overnight at 37 °C. The grown colonies were transferred to separate falcon tubes containing 1:1000 ampicillin : LB medium and incubated at 37 °C while being agitated at 200 rpm overnight. The plasmid was then extracted using the PureYield Plasmid Miniprep System (Promega).

The concentration of the plasmid was determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific) by comparing the 260 and 280 nm wavelength absorption to determine sample purity.

### 2.2.3 Restriction enzyme digestion

In the case of pGEMM 7.0, the *psd* gene needed to be deactivated by restriction of the DNA by *EcoRI* (NEB). For every µg of DNA, 1 unit of *EcoRI* was added to the digestion mix. The mix was incubated for 1 hour at 37 °C and heat-inactivated for 20 minutes at 65 °C. Finally, the digestion mix was cleaned up using Wizard SV Gel and PCR Clean-Up System (Promega).

*PmeI* restriction enzyme was used to linearize pMAR2. For every µg of DNA, 1 unit of *PmeI* (NEB) was added to the digestion mix. The mix was incubated for 3 hours at 37 °C and heat-inactivated for 20 minutes at 65 °C. The digestion mix was cleaned up by DNA precipitation. The heat-inactivated digestion mix was added to a DNA precipitation mix consisting of 16.8 µl of 5 M sodium acetate and 700 µl cooled 100% ethanol. The solution was gently mixed and left to sediment at -20 °C overnight. The precipitation mix was centrifuged at 16,000 g at 4 °C for 20 min. The supernatant was then carefully pipetted out, taking care to not disturb the translucent pellet. The sample was then washed with 70% ethanol three times, each time centrifuging the sample for 3 minutes at 16,000 g before pipetting out the ethanol. Finally, the pellet was left to dry for 10 min, resuspended in Milli-Q water and incubated at 37 °C for 20 min.

The digestion was visualized by loading the plasmid on a 1% TAE agarose gel. The digested DNA was then compared to undigested plasmid to confirm the linearization. The concentration of the digested plasmid was determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific).

### 2.2.4 PCR amplification of oriLR-p2-p3 from G340 plasmid

A PCR mix was prepared containing 5 ng of the G340 plasmid DNA, 0.2 µM primer 491 ChD, 0.2 µM primer 492 ChD, 200 µM dNTP Mix, 0.02 U/µl Phusion High-Fidelity DNA Polymerase (Promega) and 5× Phusion HF buffer in a total volume of 50 µl. The PCR program had a initial heating step of 30 seconds 98 °C, followed by 25 cycles of 10 seconds at 98 °C, 15 seconds at 70 °C and two minutes at 72 °C. Finally, the sample was heated for eight minutes at 72 °C. The PCR product was then cleaned using the MinElute PCR purification kit (QIAGEN).

### 2.2.5 IVTT in liposomes

The PURE<sub>flex</sub> 2.0 system (GeneFrontier Corporation) was used as the encapsulated IVTT system encapsulated in the liposomes. The PURE<sub>flex</sub> 2.0 kit consists of three solutions. Solution I is the feeding solution containing amino acids, NTPs, tRNAs and substrates for enzymes. Solution II is the enzyme solution. It contains the T7 RNA polymerase, transcription initiation factors and enzymes necessary for energy regeneration and aminoacylation. Solution III is the ribosome solution, containing ribosomes at a concentration of 20 µM. In the case of ΔT7 PURE<sub>flex</sub> 2.0, solution II contained no T7 RNA polymerase. The other solutions of the ΔT7 were the same as used in the unaltered PURE<sub>flex</sub> 2.0 kit. The purified T7 RNAP that was added back in the ΔT7 experiments was mutant P266L/I810N T7 RNAP [36].

A swelling solution of 20 µl would contain 10 µl of solution I, 1 µl of solution II and 2 µl of solution III. The remaining 7 µl contained, depending on the nature of the experiment, the DNA (pMAR2, pGEMM7.0, oriLR-p2-p3 or G340), auxiliary purified P5 and P6, ammonium sulfate, dNTPs and lipid precursors. The DNA constructs were present at a concentration of 2 nM each, unless indicated otherwise.

The DNA amplification experiments required 300 µM dNTP and 20 mM Ammonium sulfate, purified P5 at 750 µg/ml and purified P6 at 210 µg/ml in PURE<sub>flex</sub>2.0.

The PS synthesis in liposomes experiments required both the expression of the lipid synthesis enzymes as well as the presence of lipid precursors in solution. The precursors for DOPS synthesis in this project were 100  $\mu$ M of Oleoyl-CoA, 0.5 mM G3P, 1 mM CTP and 0.5 mM L-Serine. Additionally, to prevent oxidation of the acyl chains, 5 mM of beta-mercaptoethanol is added to the swelling solution.

### **2.2.6 Lipid precursor film**

In order to add lipid precursors to the samples, 1.45  $\mu$ l of Oleoyl-CoA (Avanti Polar Lipids) dissolved in chloroform at a concentration of 1 mg/ml, using Microman positive displacement pipettes from Gilson. The solution was left to evaporate in the PCR tube at room temperature for 2 hours, depositing the oleoyl-CoA as a film in the tube. Afterwards, 15  $\mu$ l of sample is added to the PCR tube. The final concentration of oleoyl-CoA in the sample was 100  $\mu$ M.

### **2.2.7 Natural swelling to produce GUVs**

GUVs were produced by adding 5 mg of lipid coated beads to a swelling solution of 13.3  $\mu$ l inside a 0.5 ml Eppendorf tube. The beads were left to swell on ice for 2 hours. During the swelling the beads were gently tumbled in their tube every 30 min. Afterwards they were tumbled again, plunged into liquid nitrogen and left to thaw at ambient temperature. Four of these freeze-thaw cycles were performed for each sample. 9  $\mu$ l of sample was then transferred to PCR tubes that were prepared to have lipid precursor films. Finally 1.5  $\mu$ l of 22.5  $\mu$ M LactC2-mCherry and 0.75  $\mu$ l of DNase I (NEB) were added to the liposomes. The sample was then incubated overnight at 37 °C.

## **2.3 Sample imaging**

### **2.3.1 Production of imaging chambers**

Three glass microscopy slides (1 mm thickness) were glued together using NOA 61 glue (Norland Products) and hardened under UV light. Then using a diamond drill bit, three (3 mm diameter) holes were bored through the glass slides. A glass cover slip was glued to the back of the glass slides, covering the bore holes and forming the chambers. Then the chambers were cleaned using piranha cleaning. Afterwards, the chambers were tested for leakage by filling them with 15  $\mu$ l of water and leaving them out on a tissue paper for one hour. If no moisture was visible on the tissue paper, the chambers were considered functional. Any chamber set that was damaged or considered unreliable after re-use was discarded and replaced.

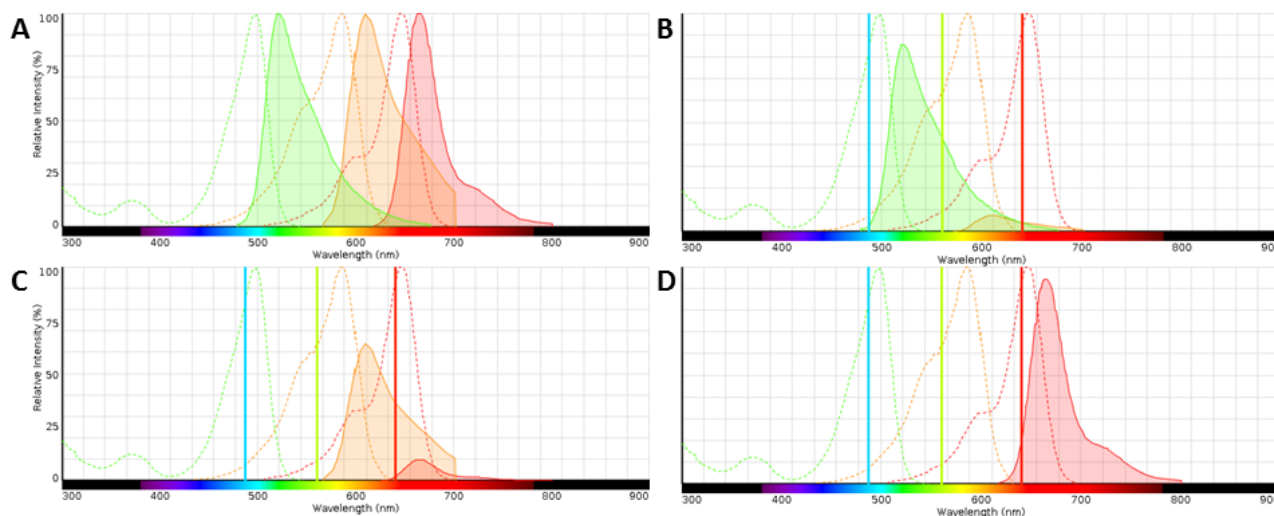
### **2.3.2 Imaging chamber preparation**

Before imaging, the chambers were cleaned by successive rounds of 10 minute sonication in: 1:1 chloroform : methanol, 2 vol% Hellmanex III, 1M KOH, Milli-Q water and 100% ethanol. Next, the chambers were dried and a square silicone border was placed around all three chamber. The chambers were incubated with BSA:BSA-biotin (1:1 at 1 mg/ml) for 10 min, washed with Milli-Q water two times and incubated with Neutravidin for another 10 min. The chambers were then washed 5 times using PURE buffer (PB) and 13  $\mu$ l of staining solution was added to each chamber. The staining solution consisted of 4:3:23 22.5  $\mu$ M LactC2-mCherry, 1000 times diluted dsGreen Gel Staining Solution, 10000 $\times$  (lumiprobe) and PB. Then 2  $\mu$ l of liposome sample was carefully added to the each chamber with a cut pipette tip. Afterwards, the chambers were sealed by placing a glass coverslip on top of the silicone border and left to sediment on a benchtop for at least 2 hours.

### **2.3.3 Confocal microscopy**

The Nikon A1R confocal microscope was used for imaging. The samples were imaged using a 100 $\times$  magnification oil immersion lens. Each capture consisted of a stitch of six by six fields of view, each 512 by 512 pixels with 12% overlap. The pixel size is 0.25  $\mu$ m. The laser scanning confocal microscope was set to image across three fluorescent channels: 488 nm laser set to 1% and 10 gain, the 580 nm to 3% and 30 gain and 660 nm was set to 10% and 90 gain. The laser wavelengths specifically excited one of the three

fluorophores present: dsGreen was excited by the 488 nm wavelength laser (Fig. 6B), LactC2-mCherry was excited by the 561 nm wavelength laser (Fig. 6C) and Cy5 was excited by the 640 nm wavelength laser (Fig. 6D). The pinhole diameter was set to 1.2 AU for 640 nm wavelength light. The emission filters were set to be appropriate for the emission spectra of mCherry, dsGreen and Cy5.



**Figure 6: Excitation and emission spectra of the used fluorophores and their predicted emission spectra given the choice of lasers.** Made using Fluorescence SpectraViewer (Thermofischer Scientific). **(A)** Emission and excitation spectra for SYBER green, a dsGreen homologue (green), mCherry (orange) and Cy5 (red). **(B)** Emission spectra (solid lines) and excitation spectra (dashed lines) for the three fluorophores when excited with a 488 nm wavelength laser. **(C)** Emission and excitation spectra for the three fluorophores when excited with a 561 nm wavelength laser. **(D)** Emission spectra and excitation for the three fluorophores when excited with a 640 nm wavelength laser.

For each of these captures the focus plane was set to be a few micron above the bottom of the chamber to match the height of the mid-section of an average-sized liposome attached to the bottom of the chamber. Imaging sessions consisted of at least four captures at each of the three chambers. The height of the sample was adjusted manually each time based on the liposomes present in each chamber.

## 2.4 Auxiliary sample analysis

### 2.4.1 Lipidomics using liquid chromatography mass spectrometry

The lipidomics liquid chromatography mass spectrometry (LC-MS) samples were prepared by ten times diluting liposome samples in sample prep solution. The sample prep solution consisted of 98:1:1, 100% methanol : 200 mM acetylacetone : 0.5 M Ethylenediaminetetraacetic acid (EDTA). The diluted samples were then sonicated for 10 minutes and centrifuged for 5 minutes at 16,000 g. The supernatant was diluted 10 $\times$  further in the sample prep solution. The prepped samples were loaded onto a CSH C18 2.1 $\times$ 50mm, 1.7  $\mu$ m liquid chromatography column (ACQUITY UPLC), mobile phase A (water with 0.05% ammonium hydroxide and 2 mM acetylacetone), and mobile phase B (% 2-propanol, 20% acetonitrile, 0.05% ammonium hydroxide and 2 mM acetylacetone).

The mass spectrometer had a triple-quad configuration. Measurements included pure standards of both DOPS and DOPC for quantification. The LC-MS was program was set to detect transitions corresponding to PS, PC, PA, PG, PE, CDP-DAG and CL. The transitions were established in previous work. In experiments where PS was already present before the start of lipid synthesis, instead of G3P, the isotopically labelled variant  $^{13}$ G3P was added to the swelling solution to introduce a 3 Da shift between newly synthesised PS and the PS that was already present.

Each sample was injected twice. The data analysis was performed using the Agilent Masshunter Quantitative analysis software. This software uses a method to produce the integrated peak values of specified transitions from raw data. The method used in this project was established in previous work. The integrated peak values were then averaged over the two injections for each sample and quantified using

MATLAB code, which was developed in previous work.

#### 2.4.2 Proteomics using QconCAT

LC-MS proteomic analysis was used in this project to quantify the relative expression of different proteins in PURE system samples by measuring several peptides that belong to our proteins of interest. Each peptide measured requires a standard at a known concentration for quantification. QconCAT is a tool to produce many isotopically labelled peptides that correspond to the peptides of interest in the sample. In essence, QconCAT produces a long  $^{15}\text{N}$  labelled peptide chain containing all these peptides of interest. Sometimes, multiple peptides corresponding to a single protein are added for additional accuracy in the quantification. The concentration of this large chain can accurately be quantified. Therefore, when the chain is digested with trypsin into its constituent peptides of interest, the concentration of these peptides will be known. The QconCAT used in this project was designed and produced for previous work in the lab, but contained several peptides that were of interest to this study. The QconCAT used was eluted in 10 mM Tris-HCl pH 8.0, 100 mM KCl buffer.

#### 2.4.3 Trypsin digest

Per LC-MS injection, 1.5  $\mu\text{L}$  of PURE system reaction was mixed with 3  $\mu\text{L}$  of 100 mM Tris-HCl pH8.0, 0.3  $\mu\text{L}$  of 20 mM  $\text{CaCl}_2$ , and 0.8  $\mu\text{L}$  MilliQ water. Samples were incubated at 90 °C for 10 minutes to stop the reaction. Then, 0.6  $\mu\text{L}$  of QconCAT (0.3 mg  $\text{mL}^{-1}$ ) was added, the sample was incubated again at 90 °C for 10 minutes and after cooling to room temperature 0.3  $\mu\text{L}$  of 1 mg  $\text{mL}^{-1}$  trypsin (trypsin-ultra, MS-grade, New England Biolabs) was added. Samples were then incubated at 37 °C overnight. After addition of 0.7  $\mu\text{L}$  10% trifluoroacetic acid, samples were centrifuged in a table-top centrifuge (5415R, Eppendorf) for 10 minutes at maximum speed. The supernatant was transferred to a glass vial with small-volume insert for LC-MS/MS analysis.

#### 2.4.4 LC-MS/MS analysis

LC-MS/MS analysis was performed on a 6460 Triple Quad LCMS system (Agilent Technologies, USA) using Skyline software. 7  $\mu\text{L}$  of sample was injected per run to an ACQUITY UPLC Peptide CSH C18 Column (Waters Corporation, USA). The peptides were separated in a gradient of buffer A (25 mM formic acid in MilliQ water) and buffer B (50 mM formic acid in acetonitrile) at a flow rate of 500  $\mu\text{L}$  per minute and at a column temperature of 40 °C. The column was equilibrated with 98% buffer A. After injection, the gradient was changed linearly over 20 minutes to 70% buffer A, over the next 4 minutes to 60% buffer A, and over the next 30 seconds to 20% buffer A. This ratio was held for another 30 seconds and the column was finally flushed with 98% buffer A to equilibrate for the next run. Selected peptides were measured by multiple reaction monitoring (MRM). For TP, SSB, DSB, PssA and PlsB, two peptides were present in the QconCAT. For DNAP and PlsC one peptide was included. In addition, two peptides from ribosomal proteins were also measured as control.

#### 2.4.5 Quantitative polymerase chain reaction

Quantitative polymerase chain reaction (qPCR) was used to determine the concentration of DNA present in the liposome samples. All liposome samples analysed with qPCR were prepared by undergoing a heat-inactivation at 75 °C for 10 minutes and then diluted 100 times. The qPCR samples could then be stored at -20 °C to be tested later with samples from other experiments.

Samples were loaded on a 96-well qPCR plate. Each sample was loaded into three different wells to perform the measurement in triplicate. Eight quantification standards were present, each in triplicate. The standards ranged from a pMAR2 concentration of 1 nM to 0.1 fM, with each standard decreasing with a factor 10. Each well contained 1  $\mu\text{L}$  of sample or standard and 9  $\mu\text{L}$  of mastermix. The 9  $\mu\text{L}$  mastermix consisted of 5  $\mu\text{L}$  of PowerUp SYBR Green Master Mix (Thermo Fisher Scientific), 0.4  $\mu\text{L}$  of 10  $\mu\text{M}$  forward and reverse qPCR primer and 3.2  $\mu\text{L}$  of Milli-Q water. After pipetting the plate was sealed using plastic sheets.

The prepared plate was then placed in the QuantStudio 5 Real-Time PCR system (Thermo Fisher Scientific). The machine was programmed to heat the wells to 50 °C for 2 minutes and 94 °C for 5 minutes and then cycle through 15 seconds at 94 °C, 15 seconds at 56 °C and 30 seconds at 68 °C 45 times. The programmed ended with heating the wells to 68 °C for 5 min. The results of the run were then analysed using the QuantStudio Design and Analysis software. The  $C_t$  of the standards was used to fit a standard curve. The standard curve was then used to quantify the DNA concentration of each of the tested samples.

#### 2.4.6 qPCR primers

The qPCR primers used in this project were designed to target all six of the genes present on pMAR2. Their sequences and reference numbers can be found in table 1.

**Table 1: Targets and sequences of the qPCR primers used in this project.**

Primer Number	Target	Orientation	Sequence
976 ChD	<i>p2</i>	forward	GGATGAAGACTACCCGCTGC
977 ChD	<i>p2</i>	reverse	ACAGGTCTGCGATTTACCG
980 ChD	<i>p3</i>	forward	ACGGCTGAAATTGACATCCCG
981 ChD	<i>p3</i>	reverse	CCAGGCGTTGAACTTCTTTGG
1119 ChD	<i>plsB</i>	forward	TCTCCCGCGACGTATTGATG
1120 ChD	<i>plsB</i>	reverse	AATAACGTCCGGCAACTCGT
1125 ChD	<i>pssA</i>	forward	AACAGGATGACGGTGGCAA
1126 ChD	<i>pssA</i>	reverse	GGAACATCTACGCCCGGATT
1225 ChD	<i>plsC</i>	forward	TATGACATGGTGACAGCA
1226 ChD	<i>plsC</i>	reverse	TTGCCGGTTAACCAGTA
1227 ChD	<i>cdsA</i>	forward	ATTAAGGACAGCGGTCAT
1228 ChD	<i>cdsA</i>	reverse	GCGTCCTGAATACCAGTA

#### 2.4.7 DNA analysis on agarose gel

DNA was loaded on a 1% agarose TAE gel and stained using ethidium bromide. Each DNA sample was loaded into the wells with Purple Gel Loading Dye (6×) (NEB). DNA length was visualised using the BenchTop 1-kb DNA ladder from Promega. The gels were run at 90 V for 45 min.



## 3 SMELDit, a newly developed image analysis tool

### 3.1 Motivation

Screening large images for specific phenotypes is time-consuming and a mostly qualitative mode of analysis. Traditional automated image analysis gives sample-wide statistics, whereas this project is interested in the co-expression of the two modules within individual liposomes. This is an important distinction to make if the objective is to identify co-occurring phenotypes. To find the fraction of liposomes that shows a combination of multiple phenotypes requires indexing of the recognized liposomes. When the liposome entries are indexed, the statistics produced by the image analysis can be assigned to individual liposomes. Conversely, individual liposomes can be categorized as belonging to multiple pre-defined groups based on these statistics.

This is not a new idea. This way of indexing individual cases is already in use in flow cytometry, where often the fluorescence intensities of multiple channels is combined to assign a phenotype to a measured cell. This is typically done by plotting a heatmap of all measured cells with a different variable on each of the two axes and then defining a gate. A gate is a specific area on the heatmap defined by the user to correspond to a specific phenotype. The relative occurrence of the phenotype can be defined as the fraction of cells found to be inside the gating.

In the case of this project, the starting point is not a rapid fluorescence measurement of a cell like in flow cytometry. Rather, the starting point will be large field of view confocal images of liposomes. This means that the liposomes in the images need to first be recognized and isolated before they can be analysed individually. The upside of this approach is that, when indexed, these images can be coupled back to the statistics they produce. Hence, a phenotype defined by a gating on two axes can be verified qualitatively by producing example images of liposomes that fall within this gating. These example images can be used to support the often subjective and highly important choice of gate definition. It is with this in mind that Show Me Example Liposomes Damn it (SMELDit) was developed.

### 3.2 Function

The software used for image analysis was newly developed for this project in MATLAB. SMELDit was designed to automatically extract single liposome features like in bulk analysis (fig. 7A) while indexing each analysed liposome. The indexed liposomes could then be assigned phenotypes based on their features. SMELDit could then be used to adjust the selection criteria of these phenotypes and produce examples of the defined phenotypes in real-time (fig. 7B).

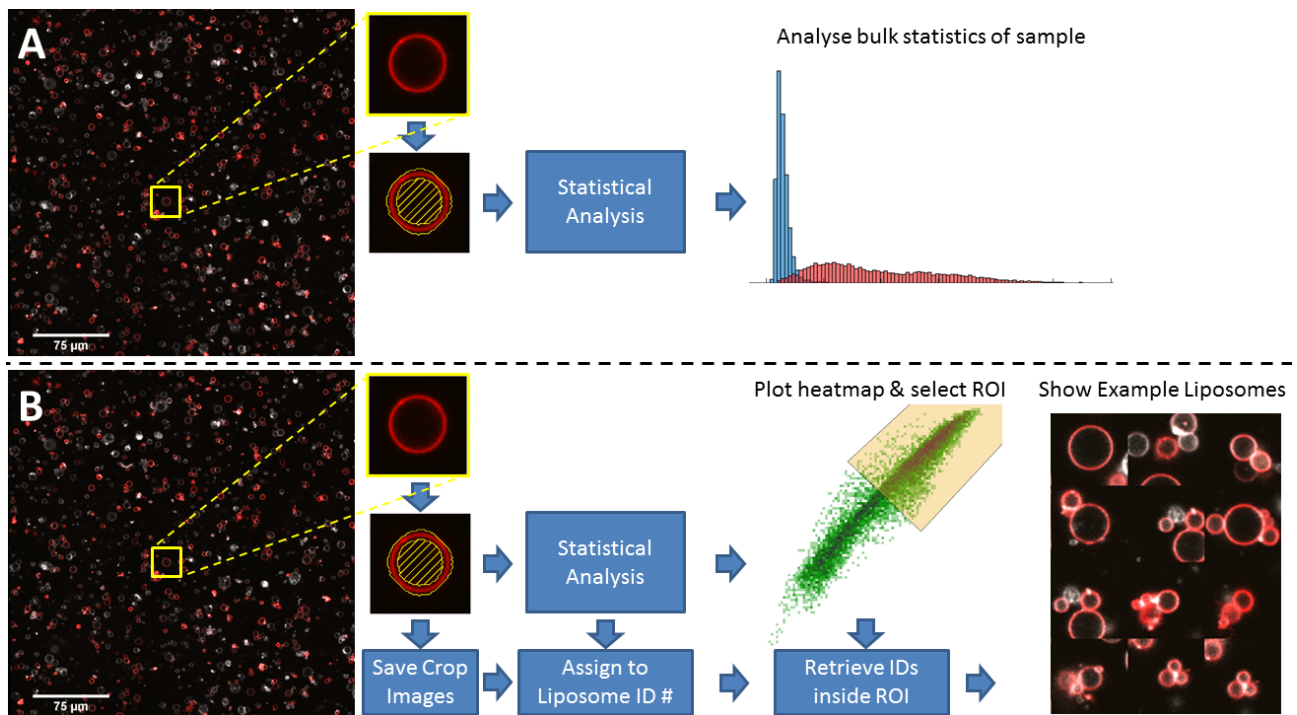
The liposome recognition and segmentation is performed in the same way as in the bulk analysis workflow. First, the Cy5 and mCherry channels of the image were added together and subsequently convolved with a laplacian filter kernel to produce high signal at membrane boundaries. What pixels belonged to the inside of each liposome was determined by a flood filling and binarization step. The binarization step was performed with a cutoff set to be 200. The processed binary image then underwent a filling step followed by an erosion step. The resulting binary image contained disconnected segments that should correspond to the lumen of the liposomes.

A selection step was performed to filter out segments corresponding to lipid aggregates and other forms of noise. All segments were individually analysed to assess their circularity  $C$  defined by equation 1 with  $P$  being the perimeter length and  $A$  being the area of the segment.

$$C = \frac{P^2}{4\pi A} \quad (1)$$

Any segments with a circularity lower than 0.5 or higher than 2 were rejected and not considered liposomes. Aggregates could be filtered out by rejecting segments with an average mCherry intensity inside the lumen of over 800, as mCherry does generally not accumulate on the inside of liposomes. The segments that passed the selection were considered liposomes and underwent further image analysis. All of these liposomes were stored as 60 by 60 pixel cropped images and given an ID number.

A segment corresponding to the membrane rim is constructed for each liposome by diluting the lumen segments (fig. 7). For each individual liposome SMELDit measures the apparent radius, average Cy5



**Figure 7: Workflow comparison between bulk analysis and SMELDit.** **(A)** The workflow of the sample-wide analysis of liposomes. The liposomes are recognized from a wider field of view. The liposome is segmented in its lumen (hatched area) and rim (area between the the outer yellow line and the hatched area) for further analysis. The analysis can then determine the mean intensity, intensity variance for each channel in both the lumen and the rim. The radius and area of the recognized liposomes could also be assigned. The statistical analysis is repeated for each recognized liposome in the sample. The result is data about the sample-wide distribution of these variables. **(B)** The workflow of SMELDit starts with the same individual statistical analysis of each recognized liposome. However, in the SMELDit workflow, a cropped image of each liposome is saved and assigned a unique liposome ID number. Then the distributions of the sample-wide variables could be plotted against each other in heatmaps. Regions of interest on these heatmaps could be manually defined by the user. After the ROIs were defined, the IDs of the liposomes inside the ROI could be retrieved. Examples of these liposomes were then immediately plotted in a montage to give immediate feedback on the choice of ROI.

intensity, average mCherry intensity and mCherry intensity variance from the membrane and the average dsGreen intensity and dsGreen intensity variance of the lumen. SMELDit can then be used to plot all liposomes in a sample on a two dimensional histogram comparing two of these variables. A region of interest (ROI) can be specified on these axes to represent a liposome phenotype. These ROIs can be saved for use across multiple samples. Once a phenotype is defined, SMELDit calculates what fraction of the liposomes in the sample displayed this phenotype.

Finally, SMELDit produces the list of liposome IDs inside the ROI and makes a montage of the corresponding cropped liposome images. This montage is displayed using the *imdisp* image processing tool [37], which allows the user to scroll through all of the liposome images present in the selected ROI. This allows the user to select a region of interest that matched the qualitative phenotype of interest. Finally, SMELDit automatically calculates what fraction of the total population of liposomes was present inside the ROI.

## 4 Results

The experimental approach of this project was split into three phases. In the first phase the experimental design was focused on establishing the conditions in which co-expression of both modules could be detected simultaneously. This included experimentally validating the use of SMELDit and the simultaneous use of the three fluorescent probes. In the second phase, the detection methods were characterized and several possible combinations of DNA constructs were screened for co-expression of both modules. This included experimental conditions in which the lipid synthesis genes were also replicated. In the third phase of the experimental approach the relative expression of the two modules is tuned. The tuning of module expression was facilitated by varying the concentrations of the RNAPs encapsulated in the liposomes.

### 4.1 SMELDit can identify DNA amplification phenotypes

In this project, DNA amplification was reported by staining the liposomes with dsGreen. Previous work of this lab has shown that, when stained, liposomes in DNA replication experiments will have highly stained aggregates in their lumen. This is not universally true for the entire population of liposomes as differing conditions between individual liposomes can cause some liposomes to fail to show any replication. This can be interpreted as the liposome sample having two distinct 'phenotypes' of liposome, the stained aggregates and the empty liposomes.

However, this project finds that DNA replication experiments result in liposomes that can be qualitatively assigned one of **three** phenotypes: either the liposome shows no staining, even staining or a stained aggregate in the lumen. The third evenly stained phenotype was discovered in this project by comparing both the mean dsGreen intensity and the dsGreen variance inside the lumen of each liposome. With SMELDit, assignment of each phenotype can be automated based on the dsGreen intensity in the lumen and the dsGreen intensity variance of the liposomes. This allows for a quantification of the relative representation of each phenotype in the imaged sample.

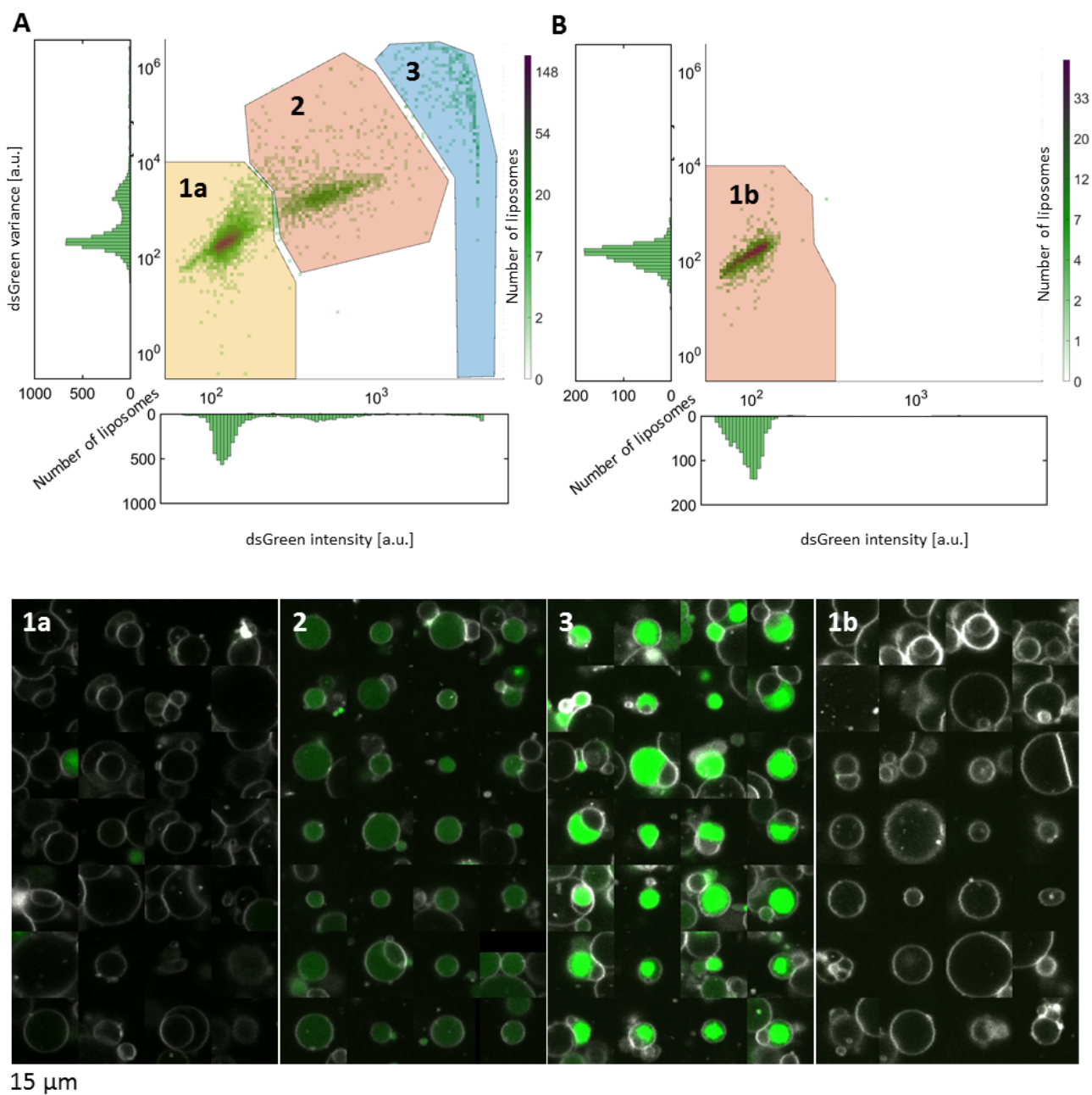
To demonstrate the ability of SMELDit to distinguish between these three phenotypes, *oriLR-p2-p3* was amplified inside liposomes. DsGreen was added to the staining solution before imaging and left to incubate for at least 30 minutes. The resulting images were then analysed using SMELDit (Fig. 8A) and compared to a negative sample that did not contain dNTPs necessary for DNA amplification.

In this particular sample, out of the 5,716 liposomes screened, SMELDit found that 72% could be classified as having no staining (Fig. 8.1a), 23% of liposomes could be classified as having even staining (Fig. 8.2) and 5% of liposomes could be classified as having stained aggregates in their lumen (Fig. 8.3). In the negative control sample, SMELDit classifies over 99% of the liposomes as having the no staining phenotype (Fig. 8.1b) using the same gating as in the positive sample. These results demonstrate that the quantification as performed by SMELDit remains consistent across multiple samples in an experiment and can be used to indicate successful DNA amplification.

The montages show the cropped images of individual liposomes. It is possible that a crop of 60 by 60 pixels contains more than one liposome. In these cases it is always the lumen and rim of the central-most liposome that has been analysed.

The gates for each phenotype were manually drawn. In this case the three different phenotypes were distinct enough on the heatmap for the gates to be drawn from the sample heatmap. Each experiment requires its own unique gating based on a negative control since the fluorescence of dsGreen and the laser power can fluctuate between experiments. The new gating can then be determined by drawing the non-stained phenotype gate around the population of liposomes in the negative control and then using this gating to see how many liposomes in the sample fall outside of this gating. All liposomes with a higher dsGreen intensity than the highest dsGreen intensity in the negative control can then be said to have elevated levels of dsGreen signal.

For most applications the distinction between the DNA aggregate and evenly stained phenotypes is not necessary. This catch-all phenotype of elevated dsGreen intensity is therefore the preferred method of describing the relative level of DNA amplification as it does not require a third gating to be drawn. Nevertheless, the ability of SMELDit to distinguish multiple phenotypes in a single sample could be useful in future experiments.

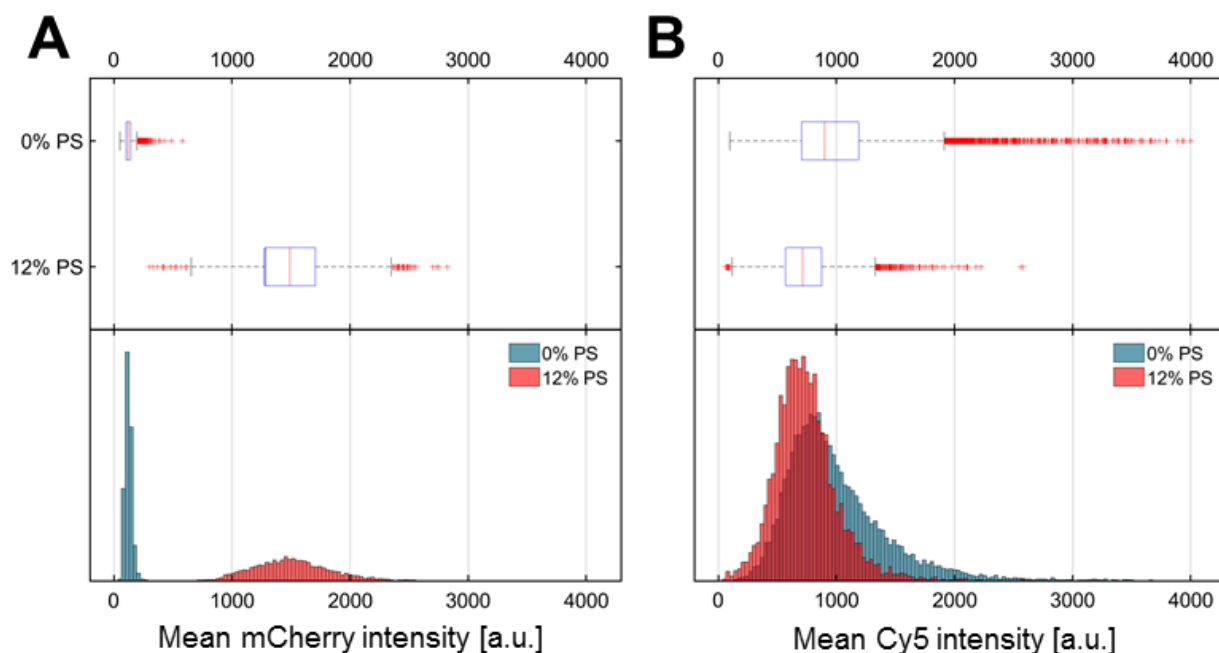


**Figure 8: Different phenotypes of DNA amplification stained by dsGreen in liposomes as identified using SMELDiT.** (A) shows a heatmap of all liposomes plotted against the mean dsGreen intensity in the lumen and the dsGreen intensity variance in the lumen. The three different areas on the heatmap (1a, 2 and 3) were chosen to represent the three phenotypes of DNA amplification present in the sample. Montages of example liposomes from these phenotype as produced by SMELDiT are displayed below the heatmap. (B) shows the result of the same analysis performed on a negative control. This sample had no dNTP added and was therefore expected to not show amplification. SMELDiT shows no liposomes are present of types 2 or 3. This is then further confirmed by the SMELDiT produced montage of example images (1b).

## 4.2 Combining LactC2-mCherry probe with Cy5 membrane staining

Synthesized PS in the liposome membrane will be fluorescently stained by addition of the LactC2-mCherry probe. DOPS could be added to the lipid mixture on the lipid coated beads to test if the purified LactC2-mCherry is active. The relative fluorescence intensity of liposomes with and without PS can then be compared between liposomes created using beads with 12 mol% PS and beads with 0 mol% PS. The results of imaging these liposomes in a buffer containing 3  $\mu$ M LactC2-mCherry are shown in figure 9.

The images were analysed using SMELDit. Figure 9A shows the increased mCherry rim intensity of the PS containing liposomes, while figure 9B shows similar Cy5 membrane dye intensities across the two samples.



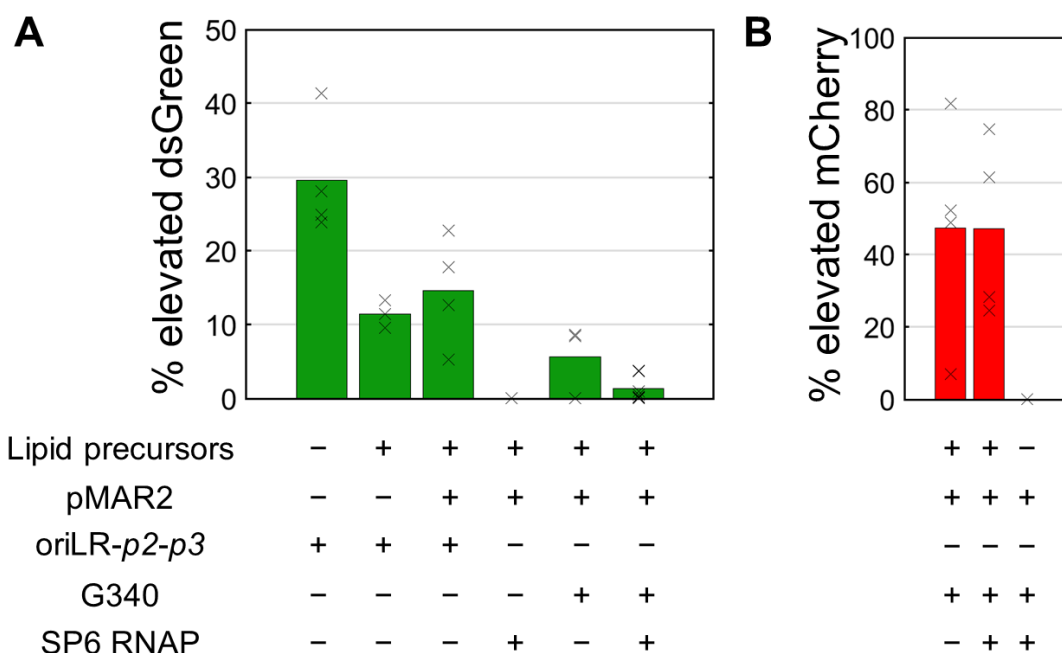
**Figure 9: Histograms and boxplots fluorescence intensity across 12 mol% PS and 0 mol% PS liposomes.** The boxplots show the median of the distribution as a red line, the 25 to 75% range of the distribution is indicated by the blue box, the 9% and 91% boundaries are indicated by the black lines and the outliers are plotted as red crosses. **(A)** The mCherry rim intensity across 12 mol% PS and 0 mol% PS liposomes. **(B)** The Cy5 intensity across 12 mol% PS and 0 mol% PS liposomes. The pictured experiment was one of several similar experiments. This one was chosen to represent the binding of LactC2-mCherry because it was the latest example and therefore the closest in experimental approach to several key experiments.

### 4.3 Combining DNA replication with lipid synthesis

The first approach to test the feasibility of combining both DNA replication and lipid synthesis inside liposomes was to co-encapsulate both modules with the exact same components used in previous work, where each module was studied separately. For DNA replication, this includes the oriLR-*p2-p3* construct with dNTPs. For lipid synthesis, the lipid precursors need to be present with pMAR2. All components were co-encapsulated with the PURE system inside liposomes, incubated, stained and then imaged. SMELDit was used to extract the percentages of liposomes showing elevated levels of the two fluorescent probes. With this metric, the relative performance of the two modules could be compared across several experimental conditions.

When coexpressed with lipid synthesis genes,  $\Phi$ 29 DNA amplification is less efficient, as signified by lower levels of dsGreen fluorescence (fig. 10A). Even when no lipid synthesis enzymes are expressed, the DNA replication is observed to diminish in the presence of lipid synthesis precursors. The expression of lipid synthesis genes with lipid precursors halves the percentage of liposomes showing an elevated level of dsGreen from nearly 30% in oriLR-*p2-p3* only experiments, to just below 15% in the combined experiment (fig. 10A). The dsGreen level of experiments with oriLR-*p2-p3* with only lipid precursors showed a larger reduction, with only 11% of liposomes showing an elevated level of dsGreen.

The experiments with these three conditions never displayed differential LactC2-mCherry staining. This means that there was no useful measurement of PS synthesis in liposomes that co-encapsulated pMAR2 with oriLR-*p2-p3*.



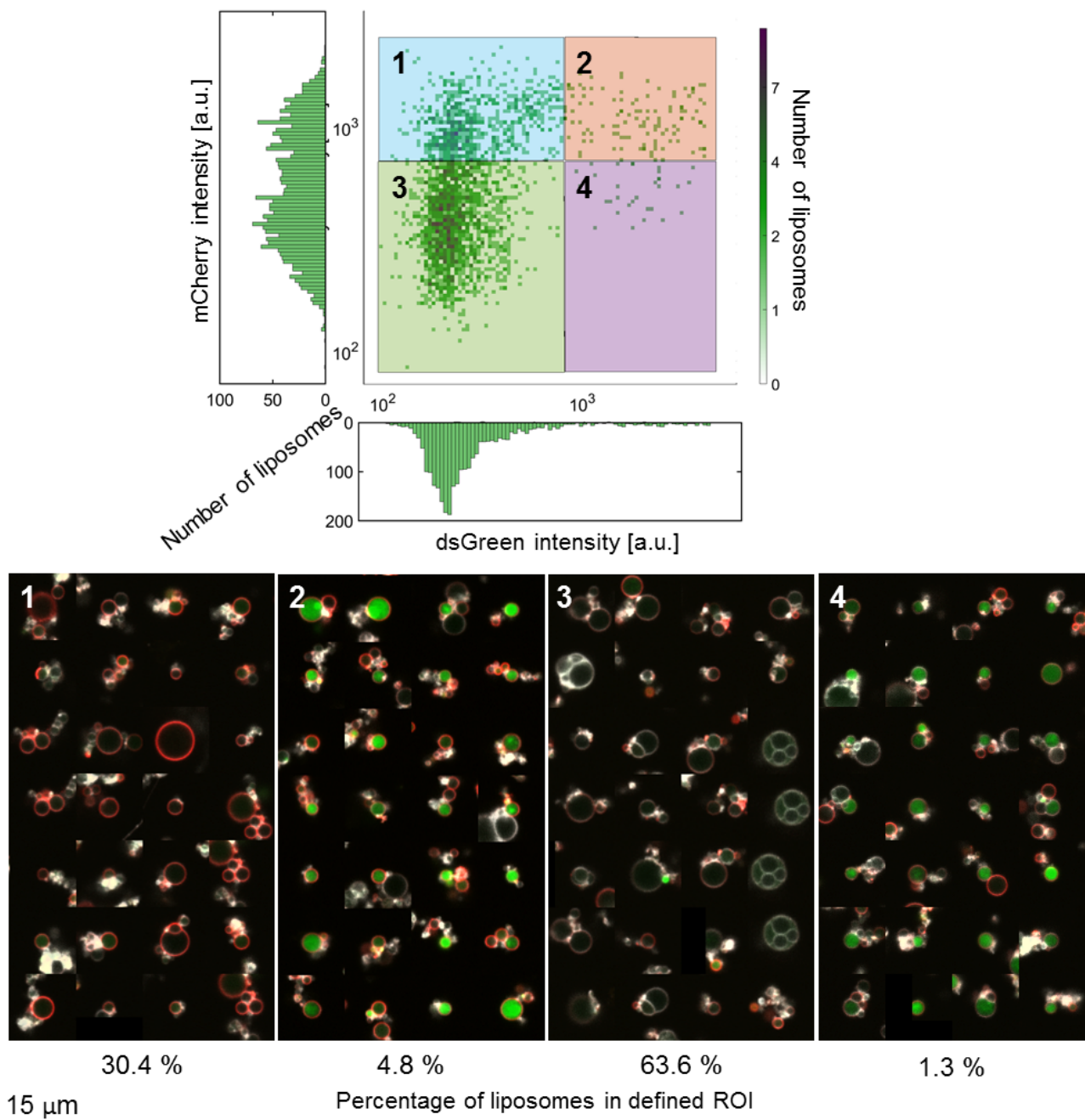
**Figure 10: Comparison of percentage of liposomes with elevated dsGreen and mCherry levels for differing experimental conditions.** (A) Comparison of percentage of liposomes with elevated dsGreen levels for differing experimental conditions. Each marker represents a separate experimental repeat. (B) Comparison of percentage of liposomes with elevated mCherry levels for the pMAR2 and G340 combination experimental conditions with a negative control that omitted oleoyl-CoA.

#### 4.4 Amplifying the lipid synthesis genes

The second stage of experiments aimed to not only combine DNA replication with lipid synthesis, but also to include the lipid synthesis genes as targets of the replication machinery. This is achieved by replicating linear pMAR2 either by the expression of p2 and p3 from pMAR2 alone by including SP6 RNAP or by expressing p2 and p3 from the G340 plasmid.

Replacing the oriLR-p2-p3 with the G340 plasmid resulted in a lowered amplification (fig. 10A), with 5.7% of liposomes showing elevated levels of dsGreen. In contrast, expressing the p2 and p3 genes directly from linear pMAR2 by adding SP6 RNAP resulted in no observable amplification. The failure of pMAR2 to amplify itself in the presence of SP6 RNAP lead to the consideration that SP6 RNAP might have a relatively lower activity compared to the T7 RNAP in the current system. One way to tune down the expression of the T7 promoter lipid synthesis genes is to introduce the G340 plasmid. The T7 RNAP will transcribe the p2 and p3 genes on the G340 plasmid while also reducing the transcription of the lipid synthesis genes, thereby shifting the expression balance more towards the DNA replication machinery. When SP6 RNAP was added to the swelling solution containing both linear pMAR2 and the G340 plasmid, a modest amount of liposomes showed elevated dsGreen, with an average of 1.3% (fig. 10).

The combination of pMAR2 and G340 also clearly shows elevated levels of LactC2-mCherry levels over a negative control that did not contain oleoyl-CoA and was therefore unable to synthesize PS (fig. 10B). The threshold for a liposome to be considered to have elevated levels of mCherry intensity was set to be higher than the highest mCherry intensity of the negative control. Therefore, 0% of the negative control is considered to have elevated levels of mCherry. The average of the +O-CoA samples was 47.4% in the sample without SP6 RNAP and 47.2% in the sample with SP6 RNAP. So, while adding SP6 RNAP to the pMAR2 and G340 plasmid combination diminishes the percentage of liposomes showing elevated levels of dsGreen, the percentage of liposomes with elevated levels of LactC2-mCherry staining does not show a significant change.



**Figure 11: The combination of pMAR2 and G340 results in liposomes displaying both LactC2-mCherry binding as well as elevated levels of dsGreen.** SMELDit can be used to select a ROI that corresponds to liposomes that have elevated levels of dsGreen as well as elevated levels of mCherry. The liposomes pictured contained pMAR2 and the G340 plasmid and without SP6 RNAP.

Using SMELDit, a ROI can be defined that selects for liposomes with elevated levels for both dsGreen and mCherry intensity (fig. 11). The corresponding montage of example images shows the co-expression of DNA replication and PS synthesis within individual liposomes (fig. 11.2). The example images show the presence of typical DNA aggregates in the lumen of the liposomes that are stained with LactC2-mCherry. The contrast of the dsGreen signal has been adjusted to show the relatively low intensity values inside liposomes that are still considered elevated by the image analysis.

Liposomes that showed elevated levels of LactC2-mCherry were about 8 times as likely to also exhibit elevated levels of dsGreen. This would suggest that on an individual liposome level the expression of the two modules is not only possible, but also somewhat correlated. The fact that it was only found to happen in a small fraction of the liposomes on the other hand would mean that this is a relatively rare phenotype. This in turn suggest that there are underlying differences in liposomes displaying this double-

positive phenotype compared to the general liposome population. It also suggests that a high-throughput screening method, like SMELDit, is necessary to accurately describe the activity in similar samples as the phenotype might become more rare under different experimental conditions.

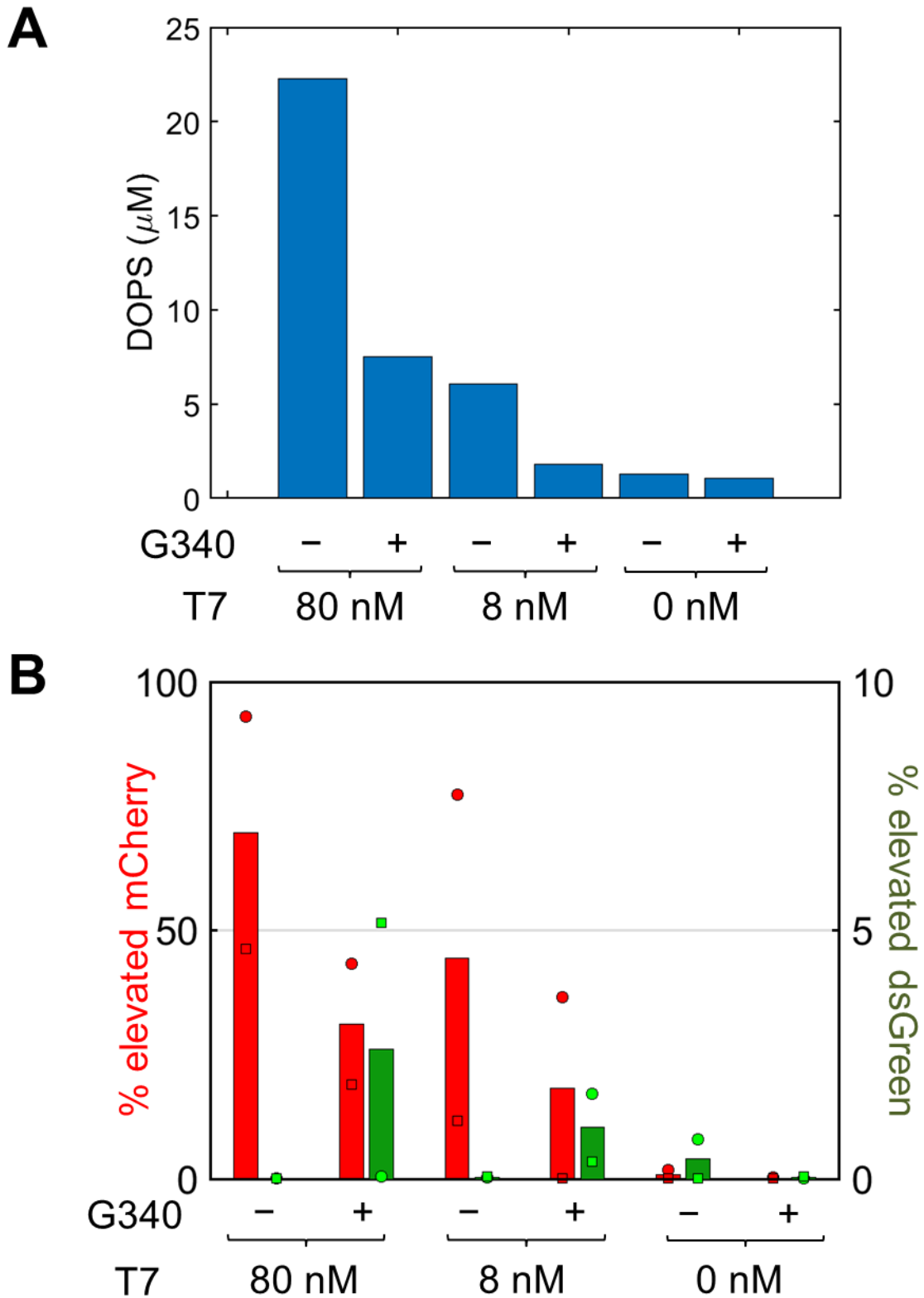
#### 4.5 Expression of pMAR2 can be tuned using varying RNAP concentrations in the $\Delta$ T7 PURE system

The PURE system contains T7 RNAP, which is responsible for transcribing genes under a T7 promoter. The *p2* and *p3* genes on pMAR2 are under an SP6 promoter. SP6 RNAP is not present in PURE<sub>frex2.0</sub>, so it is co-encapsulated with pMAR2 inside the liposomes in order to express *p2* and *p3*. Due to the limited volume of the swelling solution and the stock buffer of SP6 RNAP there is a limit to the concentration of SP6 RNAP that can be encapsulated inside liposomes. The standard concentration used in this project of 1 U/ $\mu$ l is already close to this limit. So instead of increasing the relative expression of the genes under an SP6 promoter by increasing the SP6 RNAP concentration, the relative expression of the genes under a T7 promoter can be decreased by decreasing the T7 RNAP concentration. This is achieved by use of  $\Delta$ T7 PURE<sub>frex2.0</sub>, in which solution II contains no T7 RNAP. Then the concentration of T7 RNAP can be varied by adding back different concentrations of purified T7 RNAP.

Although the exact concentration of T7 RNAP in PURE<sub>frex2.0</sub> is unknown, previous unpublished work showed that a final concentration of 80 nM of T7 RNAP has similar levels of expression to the standard PURE system. The experiments conducted in the  $\Delta$ T7 PURE system had three different concentrations of T7 RNAP: 80 nM, 8 nM and 0 nM. This should give results similar to the standard PURE system (80 nM), results of only SP6 expression (0 nM) and an in-between condition where T7 transcription is reduced but not completely disabled (8 nM).

The ideal condition of module integration would be to express both systems from the same DNA construct. However, under standard PURE system conditions the expression of *p2* and *p3* from pMAR2 alone is not enough to produce measurable DNA amplification (fig. 10). Lowering the T7 RNAP concentration could lead to an increase in the expression of *p2* and *p3* from pMAR2 as these genes have an SP6 promoter. This could enable the self-replication of pMAR2. Therefore, for each of the three T7 RNAP concentration conditions the liposome experiment was repeated with and without the G340 plasmid present. Each experiment was conducted with lipid precursors and dNTPs present at the standard concentrations. The samples were analysed using LC-MS to quantify the concentration of PS present in the lipid membranes (fig. 12A). Liposomes of the same samples were also imaged and analysed using SMELDit (fig. 12B).

The DOPS concentration measured by the LC-MS was around 22  $\mu$ M for the sample without G340 and 8  $\mu$ M for the sample with G340 at 80 nM of T7 RNAP, around 6.1 and 1.8  $\mu$ M for the samples with 8 nM of T7 RNAP, and the DOPS concentrations in the 0 nM T7 RNAP samples were 1.3 and 1.1  $\mu$ M (fig. 12A). Overall, the presence of the G340 plasmid results in a lower measured DOPS concentration in each T7 RNAP concentration. This suggests that decreasing fraction of T7 promoter genes that encode for lipid synthesis enzymes decreases the amount of PS synthesized, as the *p2* and *p3* genes on the G340 plasmid also have a T7 promoter.



**Figure 12: T7 RNAP concentration and inclusion of G340 impact PS synthesis.** All conditions in this figure also included SP6 RNAP. **(A)** LC-MS analysis showed that lowering the concentration of T7 RNAP will decreased the amount of PS synthesized. The co-expression of the T7 promoter genes on G340 also lowered the amount of synthesized PS. **(B)** The percentages of liposomes that showed elevated levels of mCherry (red) and elevated levels of dsGreen (green).

Image analysis showed that the lower T7 RNAP concentration resulted in a decreased percentage of liposomes showing elevated levels of mCherry rim intensity across both the G340 and no G340 conditions (fig. 12B). The LC-MS and LactC2-mCherry results match closely on a qualitative level. Compared to the LC-MS results the drop between 80 and 8 nM T7 RNAP is proportionally smaller. The percentage of liposomes with elevated levels of mCherry rim intensity is also consistently lower for the samples that contained G340.

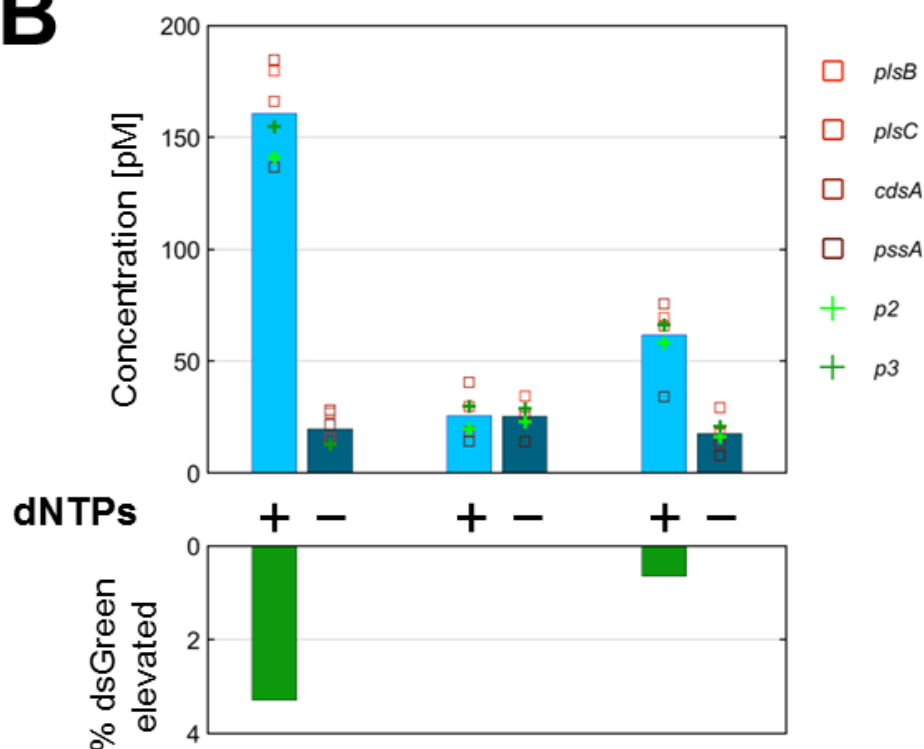
In the samples that contained G340 the percentage of liposomes showing elevated levels of dsGreen was the highest for the 80 nM T7 RNAP condition, with an average of 2.6% of liposomes showing elevated levels of dsGreen. The 8 nM T7 RNAP condition resulted in 1.0% of the liposomes showing an elevated level of dsGreen and the 0 nM T7 RNAP condition resulted in no liposomes with elevated dsGreen levels. Since the *p2* and *p3* genes are under a T7 promoter on the G340 plasmid, these results suggest that not pMAR2, but the G340 plasmid is the main source of transcription of *p2* and *p3* when present.

In contrast, the samples that contained only pMAR2 had the highest percentage, 0.4%, of liposomes showing elevated levels of dsGreen in the 0 nM T7 condition, with 0.04% of liposomes having elevated levels of dsGreen in the 8 nM T7 RNAP condition and 0% of liposomes being elevated in the 80 nM T7 RNAP condition (fig. 12B). This demonstrates that pMAR2 is capable of self-amplification under  $\Delta$ T7 PURE conditions. This self-amplification of pMAR2 is however small compared to the amplification of pMAR2 in the presence of G340. Suggesting the expression of *p2* and *p3* from pMAR2 alone is not sufficient or robust enough to produce DNA amplification.

#### 4.5.1 pMAR2 is capable of self-replication in the absence T7 RNAP

Under  $\Delta$ T7 conditions, the *p2* and *p3* genes on pMAR2 are expressed well enough to result in amplification detectable by dsGreen staining. With the first repeat showing the highest measured percentage of liposomes with an elevated level of dsGreen measured for these experimental conditions, with 3.3 % being higher than the highest repeat of figure 12B. However, this does not proof full-length replication of pMAR2. It could be that shorter parasitic products are being amplified. To test this, the DNA concentration of all six genes on pMAR2 (fig. 13A) was measured with qPCR using the appropriate primers as indicated in table 1. Before the qPCR-based approach was conceived of, several attempts were made to assess the length of the amplified DNA in liposomes by gel analysis. The problem with the a gel analysis approach is that the amount of DNA that is recovered from a liposome sample is typically between 100 pg and 1 ng. So, either an extremely large sample volume needs to be made or the gel analysis needs to be extremely sensitive to pick up this signal. An attempt was made using the ultra-sensitive SYBR Gold Nucleic Acid Gel Stain, which has been used to detect very small amounts of DNA in previous work in this lab. The result was however a gel with no band corresponding to pMAR2 (fig. A.1B). So, we opted to use a qPCR-based approach as it had the sensitivity we needed and was functional from the start.

The  $\Delta$ T7 condition amplification experiment was repeated in triplicate with negative controls that did not contain the dNTPs necessary for DNA replication. Of the three repeats only the first and third showed elevated levels of dsGreen during imaging (fig. 13B). The qPCR measurement showed the average measured DNA concentration of the six genes was 8.1 fold higher than the negative control in repeat one, equal to the negative control for repeat 2 and 3.5 fold higher for repeat 3. This brings the average amplification fold across the three repeats to 4.2, with the caveat that one of the repeats likely did not replicate at all. The DNA concentration measured in the samples was consistent across the six genes, which would suggest pMAR2 was replicated at full length.

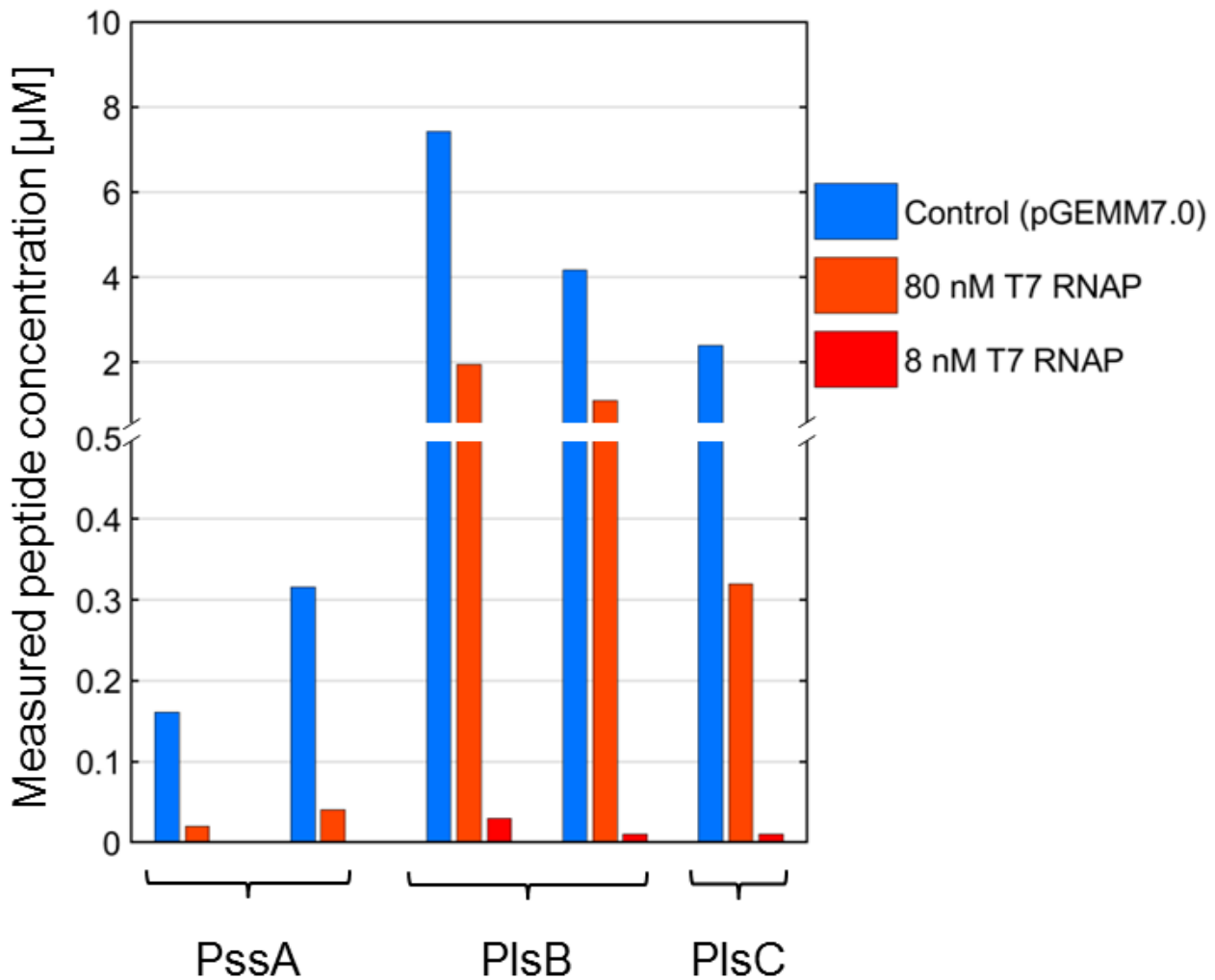
**A****B**

**Figure 13: Six primer qPCR measurement of pMAR2 amplification shows full-length self-replication in liposomes.** (A) Schematic depiction of linear pMAR2 showing all six genes present on the DNA construct. The corresponding standard curves can be found in figure A.2. (B) The measured concentration of the genes present on pMAR2 after amplification in liposomes compared to the negative controls across three repeats with the respective measured percentage of liposomes showing elevated levels of dsGreen.

#### 4.6 Bulk analysis of pMAR2 expression shows little to no activity

To characterize the expression of pMAR2 further, several experiments were performed outside liposomes. These bulk experiments were analysed with a targeted proteomics mass spectrometry method, with known concentrations of peptides from a QconCAT as an internal standard, to measure the concentrations of these peptides from the expressed proteins and qPCR to determine the DNA concentration.

The first bulk experiment was designed to accompany the pMAR2 self-replication under  $\Delta T7$  conditions (fig. 13). It compared a condition with and without dNTPs. The only difference with the  $\Delta T7$  conditions in liposomes was the absence of oleoyl-CoA in the outside solution. After overnight incubation at 37 °C, LC-MS proteomic analysis did not detect the presence of TP or DNAP in either of the samples. Similarly, qPCR analysis with the 1125/1126 ChD primers targeting the *pssA* gene on pMAR2 showed only a small difference in concentration between the dNTP containing sample and its negative control. This discrepancy between the sample activity in bulk versus in liposomes with otherwise identical experimental conditions is further explored in the discussion.



**Figure 14: Bulk experiment shows low concentration of lipid synthesis enzymes in presence of purified DNAP and TP.** The 80 nM and 8 nM T7 RNAP samples are compared to a typical bulk IVTT measurement of pGEMM7.0 in standard PURE system. LC-MS proteomic analysis with known concentrations of peptides from a QconCAT as an internal standard was used to measure two peptides for PssA and PlsB and one for PlsC.

The second bulk experiment measured the expression of lipid synthesis enzymes in 80, 8 and 0 nM of T7 RNAP. Because the previous bulk experiment had shown no expressed TP and DNAP present, TP and DNAP were added in purified form to the solution. One negative control containing no T7 RNAP or purified TP and DNAP was included, bringing the total to four samples.

LC-MS targeted proteomics with QconCAT peptide standards measured the peptide concentrations of both the  $\Delta$ T7 conditions. It detected no lipid synthesis enzymes, as expected. The 8 nM T7 RNAP sample was measured to have two orders of magnitude lower concentrations for PlsB and PlsC than the 80 nM T7 RNAP sample (fig. 14). PssA was not detected at all in the 8 nM T7 RNAP sample, since the PssA concentration is already low in the 80 nM sample the actual concentration might be under the detection limit. All measured peptide concentrations of the 80 nM T7 RNAP samples were lower than their concentrations in the pGEMM7.0 expression in standard PURE system, but do follow the same pattern of relative measured concentration.

## 5 Discussion, Conclusion and Outlook

Co-expression of lipid synthesis and DNA replication proteins has now been confirmed by simultaneous LactC2-mCherry and dsGreen staining. LC-MS lipidomics further reinforce the evidence for lipid synthesis by confirming the production of PS directly. With the support of qPCR data, the DNA amplification of pMAR2 in liposomes observed with dsGreen is confirmed to be full replication. This project has therefore shown that it is possible to co-express the lipid synthesis and DNA replication modules in liposomes, that the lipid synthesis genes can be replicated and expressed at the same time and that a DNA template that encodes both lipid synthesis genes and DNA replication genes is capable of fully replicating itself under the right conditions.

### 5.1 Combining the expression of two systems has a cost

OriLR-*p2-p3* amplifies in a lower fraction of liposomes when co-expressed with lipid synthesis enzymes (fig. 10). In this scenario, the two expressed modules share the resources inside the liposome. Therefore, the added expression of the lipid synthesis enzymes likely decreases the yield of expressed P2 and P3. If this is the cause for the drop in amplification between the two conditions, this would suggest that the availability of DNAP or TP are the limiting factors in the amplification of oriLR-*p2-p3*. Of the two proteins, the production of TP is more likely to be the bottleneck as new a pair of TP is necessary for every new DNA copy synthesized, whereas DNAP can stay functional for multiple rounds of replication. One possible way to test this would be to add purified TP to see if this increases the fraction of liposomes that show elevated levels of dsGreen.

However, the fraction of liposomes, encapsulating oriLR-*p2-p3* and showing an elevated level of dsGreen is already lowered from 30 to 11% by the introduction of the lipid precursors (fig. 10). At 15% average, the percentage of dsGreen elevated liposomes is actually higher for the co-encapsulation of pMAR2 with oriLR-*p2-p3*. This would suggest that the expression of the lipid synthesis genes actually benefits the DNA replication process in some way. One possible explanation for this counter-intuitive behavior is that one or multiple of the lipid precursors negatively affects DNA replication and that lipid synthesis enzymes will reduce the concentration of this precursor inside the liposome by metabolizing it. The most likely culprit in this case would be rCTP, as it could possibly stall the DNA elongation by incorporating into the DNA. This could be tested relatively simply by a lipid precursor titration experiment, in which the relative amplification fraction of oriLR-*p2-p3* in liposomes is compared. If for instance, a zero concentration of rCTP results in the percentage of dsGreen elevated liposomes returning to around 30%, the rCTP would be the likely cause for the drop in the DNA amplification.

The combined expression from pMAR2 and the G340 plasmid resulted in an even lower 5.7% average of liposomes exhibiting elevated levels of dsGreen (fig. 10). In contrast, pMAR2 failed to amplify in the presence of SP6 RNAP alone, suggesting the expression of the *p2* and *p3* genes of pMAR2 is substantially lower than that achieved by the genes on the G340 plasmid at the same concentration. However, only a 1.2% average of dsGreen elevated liposomes was measured in the pMAR2 and G340 with SP6 RNAP case (fig. 10), suggesting the SP6 RNAP is directly negatively affecting DNA amplification. The question of which underlying mechanisms cause this negative interaction between SP6 RNAP and DNA amplification is addressed in later sections. In the next section, the high variability between repeats of the experiments involving pMAR2 is examined.

### 5.2 Replication of pMAR2 fails at times with no clear cause

In the experiments with pMAR2 the level of DNA replication was inconsistent, not only between different experimental conditions, but also between repeats of the same experiment. Strikingly in two of the eight repeats in the pMAR2, G340 plasmid and SP6 RNAP condition (fig. 10) the measured percentage of liposomes showing elevated levels of dsGreen was zero. The same goes for one of the repeats of the pMAR2 self-replication experiment (fig. 13), where one of the three repeats was measured to have an amplification fold of one, with zero percent of liposomes showing elevated levels of dsGreen. This would suggest that reconstituting the DNA replication function has failed in these samples, while it was successfully reconstituted in other repeats of the same experiment.

Identical experimental design leading to different outcomes suggest the differentiating factor is the execution of the experiments. Technically the difference could also be stochastic, but since the number of processed liposomes regularly exceeds 10,000 per sample the measured percentage of liposomes with elevated dsGreen intensity is likely close to the true average. The remaining possibilities are human error or difference in activity between batches of components. Human error would specifically refer to pipetting inaccuracies, resulting in differing concentrations of components between experimental repeats. This would naturally lead to a larger variance in the measured activity of both modules between repeats. However, what combination of altered component concentrations could lead to complete inactivity remains unclear. Similarly, the difference in the components between batches could explain the sudden inactivity, as a similar effect on the lipid synthesis was found to be caused by a contaminated dNTP stock solution (fig. A.3), but what components could be responsible in the case of DNA replication remains unclear.

A potentially simple solution to both problems would be to make a master mix for multiple experimental repeats. This master mix could contain PURE<sub>flex</sub>2.0 solutions I and III, L-Serine, rCTP, G3P, Amonium sulfate, purified p5 and p6 and would be stored in aliquots corresponding to single experimental repeats. This master mix would increase the consistency across repeats while still giving flexibility in the choice and concentration of DNA, PURE<sub>flex</sub>2.0 solution II, dNTP and RNAP used. This might become especially necessary for future experiments with even more co-encapsulated components, as the noise caused by differing component concentrations between experimental repeats will increase. However, the risk with this approach would be that additional freeze-thawing of a mixture that contains purified proteins could lead to a lessened activity of the sample. The impact of this approach on the sample activity should be tested before the approach can be adopted.

### 5.3 Adjusting the T7 RNAP concentration affects lipid synthesis and DNA amplification

The expression of the lipid synthesis enzymes from pMAR2 can be decreased by lowering the concentration of T7 RNAP. This effect is observed directly by the measured enzyme concentrations in LC-MS proteomic analysis with QconCAT (fig. 14B) and indirectly by the LC-MS lipidomics measurement of PS concentration in liposomes (fig. 12A) and the measured percentage of liposomes with elevated LactC2-mCherry binding (fig. 12B) under different concentrations of T7 RNAP. We can therefore conclude that decreasing the T7 RNAP concentration lowers the expression of lipid synthesis enzymes, which in turn results in less PS accumulating in the liposome membranes. This lowered PS concentration can in turn be detected by the image analysis performed by SMELDit.

However, the detected concentration of lipid synthesis enzymes is lower for the 80 nM purified T7 RNAP case than the control of pGEMM7.0( $\Delta psd$ ) in PURE system with its 'native' T7 RNAP. This difference in expression is remarkable, since in previous work in this lab 80 nM of purified T7 RNAP was found to have similar protein yield to the native PURE system. However, it is possible that the active fraction of the purified T7 may have decreased with time.

Another interesting observation is that the relation between T7 RNAP concentration, measured lipid synthesis enzyme concentration and synthesized PS concentration is not necessarily linear. The concentration of the measured PlsB and PlsC peptides was on average 70 times higher in the 80 nM T7 RNAP sample compared to the 8 nM sample. The LC-MS measured PS concentration in the liposomes is only on average 4 times higher in the 80 nM T7 RNAP sample compared to the 8 nM sample. The fact that this relation is not directly linear is not necessarily surprising, especially since the proteomics were measured in bulk, whereas the lipidomics measured liposome samples. However, currently the relationship between T7 RNAP concentration and synthesized PS remains uncharacterized. Modeling this relation and screening the PS and lipid synthesis concentration for more T7 RNAP concentrations is a potential future project for the lab.

Decreasing the T7 RNAP concentration also decreases the percentage of liposomes with elevated levels of dsGreen in the samples with both pMAR2 and the G340 plasmid as expected (fig. 12B). However, in the samples without the G340 plasmid, the percentage of liposomes with elevated levels of dsGreen actually increases with a lowered T7 RNAP concentration. This would suggest that either the transcription of *p2* and *p3* from pMAR2 by the SP6 RNAP is negatively affected by the presence of T7 RNAP, or that the activity of the T7 RNAP compared to the SP6 RNAP is so much higher that the resource-sharing relation results in less overall production of the DNA replication machinery.

Another interesting result from the  $\Delta T7$  RNAP conditions is that 0% of liposomes showed elevated levels of dsGreen in the G340 sample, whereas the sample without G340 showed an average of 0.39% of liposomes with an elevated level of dsGreen. Although small, this difference is unexpected as both samples contain SP6 RNAP and an equal concentration of pMAR2, suggesting that the presence of G340 itself negatively affects the DNA amplification of pMAR2. However, these experiments only have two repeats, so the evidence for any such effect is weak at best. Nevertheless, it is possible that the SP6 RNAP is binding somewhere on the G340 plasmid, thereby effectively decreasing the activity of the SP6 RNAP in the G340-containing sample.

## 5.4 Expression from pMAR2 in liposomes is different from bulk

One of the major questions that remain unanswered in this project is the absence of DNAP and TP in the bulk expression of pMAR2 under  $\Delta T7$  RNAP conditions as measured by quantitative proteomic analysis. This absence of DNAP and TP is unexpected as the same experimental conditions inside liposomes result in DNA replication as evidenced by qPCR measurement (fig. 13). In contrast, the bulk experiment showed no DNA amplification in qPCR measurements (fig. A.4).

A possible explanation would be that, under  $\Delta T7$  conditions, the genes of pMAR2 under an SP6 promoter were expressed, but only at such a low level that the concentration of TP and DNAP peptides was under the detection limit of the LC-MS measurement. In liposomes, the local concentration of TP and DNAP could be higher due to stochasticity in the number of encapsulated DNA copies starting a positive feed back loop. The higher concentration of pMAR2 at the start of this positive feedback loop increases the TP and DNAP concentration inside the liposome, which increases the chance of the DNAP-TP complex forming, which in turn increases the chance to initiate DNA replication, increasing the concentration of pMAR2 further. In this scenario, only a small fraction of liposomes would end up in this self-reinforcing regime before the PURE system becomes inactive, which is consistent with the SMELDit analysis of the sample showing a relatively small fraction of liposomes with elevated levels of dsGreen. It has been previously established that rare phenotypes of liposomes can outperform the bulk yield of protein synthesis [29]. Hence, the idea that certain rare phenotypes of liposome could reach a regime in which DNA replication becomes detectable, while bulk reactions could not, has a precedent.

A straightforward way to test the hypothesis that local high concentrations of pMAR2 enable self-replication in liposomes under conditions that do not show amplification in bulk, is to increase the pMAR2 concentration in bulk experiments. If this were to result in a higher amplification fold, this would point to the stochasticity of DNA copy number in liposomes as the underlying mechanism for the difference between bulk and in liposome experiments.

There are some immediate problems that accompany this hypothesis. The DNA concentration at 2 nM is already quite high, so, while the DNA concentration is still the lowest of all components, the effects of stochasticity are questionable. If the average radius of a liposome is assumed to be 3  $\mu\text{m}$ , the average volume would be 113 fL. At 2 nM this would bring the average copy number of pMAR2 in liposomes to 136. Equation 2 shows how to calculate the chance of liposomes containing more than  $x$  copies of DNA using the cumulative Poisson distribution  $F(x, \lambda)$ , with  $\lambda$  being the expected number of DNA copies per liposomes, in this case  $\lambda = 136$ . This relation can also be used to find the minimal copy number corresponding to the top 1% of liposomes. Here  $P(X > 163) = 0.01077$  meaning that, if the measured fraction of liposomes showing elevated levels of dsGreen was entirely explained by the starting copy number and that about 1% of the liposomes showed elevated levels of dsGreen, the initial copy number of DNA needed to be in a replication regime would only be about 20% higher than the average. While not impossible, this would seem to be an extreme divergence in outcome for only a small difference in the initial conditions inside the liposomes.

$$P(X > x) = 1 - F(x, \lambda) = \sum_{i=0}^x \frac{e^{-\lambda} \lambda^i}{i!} \quad (2)$$

There are other factors not yet considered that could increase the effects of stochasticity of encapsulation. For a start, there are many other components co-encapsulated with the DNA molecules. Not only that, but the assumption of a Poisson distribution of partitioning components of a solution only holds

if these components are independently distributed. Protein or ribosome aggregates would break this assumption and have been found to lead to extreme concentrations in liposomes [38]. The effect of confinement stochasticity are also amplified by the transcriptional activity of the encapsulated DNA copies. Previous work has found that as low as 10% of encapsulated DNA molecules are transcriptionally active [39], effectively reducing the copy number of DNA encapsulated in the liposomes. In conclusion, although there are some mitigating factors, the high average copy number of the DNA encapsulation makes a purely confinement stochasticity-based explanation for the difference between bulk and in liposome experiments unlikely.

If confinement stochasticity does not sufficiently explain the difference between bulk and in liposome experiments, other explanations could be more technical in nature. For one the air-water interface in bulk experiments could affect protein activity. However, the discrepancy between bulk and in liposome experiments has not been observed in previous work in the lab with the  $\Phi$ 29 DNA replication machinery. So, it seems unlikely that the air-water interface has only now become a problem.

## 5.5 Replication of pMAR2 could be limited by its length

With the confirmation of the six-gene qPCR measurement, we can conclude that pMAR2 can self-replicate in  $\Delta$ T7 conditions. However, the measured amplification fold is lower than that of oriLR-*p2-p3* in previous work [8]. The percentage of liposomes showing elevated levels of dsGreen are substantially lower for experiments with pMAR2 compared to experiments with oriLR-*p2-p3*. A straightforward explanation for why pMAR2 might underperform is a lower expression of TP and DNAP, something which would be supported by the bulk proteomic measurements. However, while a low expression of TP and DNAP might explain less DNA amplification in the  $\Delta$ T7 conditions, it does not explain the similarly low amplification observed in the pMAR2, G340 and SP6 RNAP condition. The G340 plasmid should initially have similar levels of expression of TP and DNAP to oriLR-*p2-p3* as the transcribed sequences for G340 and oriLR-*p2-p3* are identical and they are both present at 2 nM concentration.

So, factors other than gene expression could be dominant in limiting the replication of pMAR2. Specifically, at 11,601 bp pMAR2 is significantly longer than oriLR-*p2-p3* at 3,213 bp. Replicating longer DNA molecules can limit the achievable replication fold in two ways.

Firstly, at equal starting concentrations, a longer DNA template will deplete more of the available dNTP concentration with each copy than a shorter starting template. Therefore, at an equal starting concentration of dNTPs the longer template would reach a lower amplification fold than the shorter template. Secondly, the system could be rate limited by the procession speed of the DNAP during elongation. If the time needed to replicate a DNA template scales linear with its length and the PURE system is only active for a limited period of time, shorter templates would undergo more doubling events than longer templates. Longer templates would therefore have both a lower amplification fold as well as a lower mass yield than short templates.

At a starting concentration of 2 nM pMAR2 and 300  $\mu$ M of each dNTP, if all dNTPs in solution were incorporated into new copies of pMAR2, 51.7 would be the maximum achievable amplification fold. For the shorter oriLR-*p2-p3* the same initial DNA and dNTP concentration would result in 187 as the maximum amplification fold. However, the yield of DNA in terms of mass would be the same in both cases. Since dsGreen binds to DNA agnostic of its sequence, both cases would therefore have a similar fluorescence intensity. Our experiments show that pMAR2 amplification results in a lower fraction of liposomes showing elevated levels of dsGreen, when compared to oriLR-*p2-p3*, suggesting that the mass yield of DNA is lower for pMAR2. This suggests that the depletion of the dNTPs in solution is not the main limiting factor under these conditions.

From literature it has been theorized that the recombinant  $\Phi$ 29 DNA replication becomes rate limited by the elongation time of the DNA template at a critical length of about 7000 bp [40]. So, while oriLR-*p2-p3* is rate limited by the initiation time, pMAR2 should theoretically be well within the elongation time limited regime. This would mean that given the same time frame of PURE system activity pMAR2 will be replicated less than oriLR-*p2-p3*. A possible way to test the idea that the length of pMAR2 is the factor that decreases its relative amplification fold is to clonally amplify oriLR-*p2-p3* using circular pMAR2 under  $\Delta$ T7 conditions. Under these experimental conditions the pMAR2 plasmid would be the only DNA expressing  $\Phi$ 29 DNA replication machinery, whereas the shorter oriLR-*p2-p3* would be the only replication target.

If this results in an amplification fold of oriLR-*p2-p3* similar to that of self-replicating pMAR2, this would suggest the DNA replication to be limited by the gene expression, whereas if these conditions reach a higher amplification fold this would suggest the length of the DNA was the limiting factor.

Finally, pMAR2 replication could be slowed by its higher number of genes.  $\Phi$ 29 DNAP can be stalled on the DNA by colliding with a transcription complex [41], hence additional genes on a contiguous DNA molecule will increase the probability of a collision of the DNAP with a transcription complex and therefore increase the stall rate of the DNAP. It is important to note that the direction of the collision matters, a head-on collision will stall the DNAP completely, whereas a co-directional collision slows the replication. This is important information for the design of our DNA template, as currently the pMAR2 template was designed to have the genes under the SP6 promoter on the opposite strand to prevent transcriptional readthrough, which may have inadvertently increased the chances of a stalling the replication fork. The most straightforward way to test the effect of collisions with transcription complexes on the replication of pMAR2 is to compare the replication fold of pMAR2 in different concentrations of T7 RNAP and SP6 RNAP, in the presence of a high concentration of purified DNAP and TP. According to the hypothesis that collisions with the transcription complexes is a limiting factor for the DNA replication, the amplification fold should be substantially lower in the presence of the RNAPs. Alternatively, new DNA templates with different promoters could be constructed and compared to pMAR2 to test the effect of collisions with the transcriptional machinery. In the next section we will address the planned new pMAR3 and pMAR4 constructs and what experiments these new constructs would enable.

## 5.6 Future pMAR constructs

Currently, the construction of two new pMAR versions (pMAR3 and pMAR4) is underway. The new versions will have promoter schemes different from that of pMAR2. The first new construct, pMAR3, is likely going to become available in the near future. The SP6 promoters of pMAR2 will be replaced with T7 promoters in pMAR3. This leaves pMAR3 with only T7 promoters. The second new construct, pMAR4, is adapted from pMAR3 to exchange the T7 promoters of the lipid synthesis genes with SP6 promoters.

The design of pMAR3 was motivated by the observation that the relative expression of genes under the SP6 promoters is low compared to genes under the T7 promoter. So, by exchanging the SP6 for T7 promoters, the *p2* and *p3* expression is expected to be increased compared to pMAR2, since the G340 and pMAR2 combination results in a higher percentage of liposomes with elevated levels of dsGreen in the absence of SP6 RNAP. The loss of the SP6 promoters will mean the loss of the orthogonal transcription from the construct. Without orthogonal transcription, it will be harder to influence the relative expression of the two modules. Nevertheless, the potential upside would be that pMAR3 could be the first construct capable of simultaneous self-replication and expression of lipid synthesis enzymes. If these conditions are met, pMAR3 would enable a directed evolution experiment with a dual-selection pressure for higher DNA replication folds and lipid synthesis yields.

The design of pMAR4 is aimed at regaining the orthogonal transcription with an increased *p2* and *p3* expression. This project has shown that PS synthesis is more robust in its reconstitution than the DNA replication module. So, if a future experiment requires the relative expression of the two modules to be adjusted, the lipid synthesis module would be a better target for downregulation. Placing the lipid synthesis genes under the SP6 promoter will likely reduce their base expression level. However, regulating the relative expression with the concentration of SP6 RNAP allows the use of the T7 RNAP 'native' to the PURE system, which is both convenient and results in a higher expression. Moreover, the experiments with lowered T7 RNAP concentration show that PS is synthesized, even at a lowered expression (fig. 12). So a lowered base expression of the lipid synthesis genes is likely to still result in a detectable level of PS synthesis. Similar to pMAR3, pMAR4 could be used in future directed evolution experiments with the added benefit of having the modules be individually adjustable.

## 5.7 dsGreen could stain RNA

During this project the fluorescent probes played a central role. It is therefore important to discuss potential points of failure in the use of these probes that have come to our attention during this project. Namely, dsGreen could potentially stain RNA. While dsGreen has a 13-fold higher fluorescence for dsDNA than for

ssDNA or RNA [42], it is possible that elevated levels of dsGreen are in part caused by high transcription and not replication. It should be noted that this could be especially difficult to disentangle as high transcription and replication activity are likely correlated. However, during this project the relative percentage of dsGreen elevated liposomes matched reasonably well with the relative amplification fold measured by qPCR (fig. 13). Moreover, liposomes with elevated levels of dsGreen were only observed in samples that were theoretically capable of expressing DNA replication machinery. So, for now the specificity of dsGreen to DNA would seem sufficient to accurately detect DNA replication inside liposomes.

## 5.8 Outlook on directed evolution as a future research direction

As stated in the introduction, this project is best understood as a step toward applying directed evolution to the minimal cell. After all, the potential of future directed evolution experiments has motivated the choice of functional modules, as DNA replication has several inherent benefits for DNA re-encapsulation and PS synthesis can be detected with a fluorescent probe, enabling high-throughput screening techniques. It is therefore important to address what such future directed evolution experiments might look like and what issues need to be solved to get to these experiments.

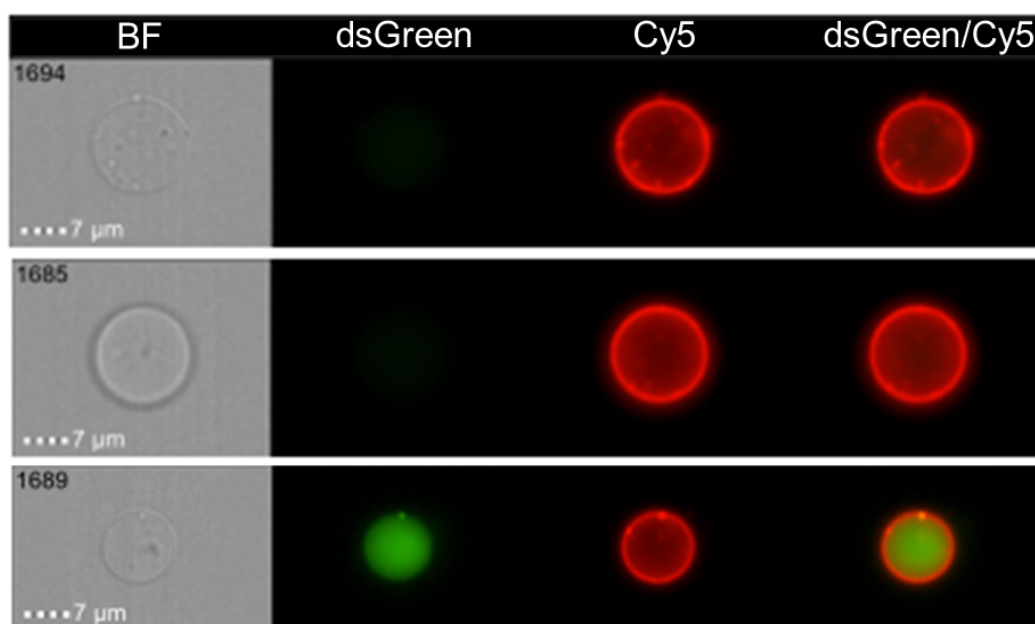
Before the methods of directed evolution can be addressed, we first need to lay out what this lab aims to achieve with directed evolution. The two most prominent aims for directed evolution experiments for a future project would be increasing the efficiency of PlsB in metabolizing acyl-[acyl-carrier protein] (acyl-ACP) as a lipid precursor and increasing the processivity of DNAP. Acyl-ACP is more water soluble than acyl-CoA and could therefore be used to reach higher yields of synthesized lipids. Conveniently, the relative yield of PS synthesized in the presence of acyl-ACP can be screened using LactC2-mCherry. The processivity of DNAP is highly likely to be a limiting factor of DNA replication, especially for larger DNA constructs like pMAR2. While it would be hard to screen the processivity directly, we could screen the DNA yield using dsGreen fluorescence in liposomes for a large target DNA template and assume that a higher yield translates to an increased processivity of the DNAP.

### 5.8.1 Luminex flow cytometry

The perspective of Abil and Danelon [33] gives a structured overview of the possible ways directed evolution could be used in the construction of a bottom-up minimal cell. One prominent direction pointed out is the use of flow cytometry high-throughput screening techniques like FACS and IACS. So, an important question to address is how the techniques used in this project can be combined with flow cytometry. Specifically, can the GUVs we use as our cell compartments survive the shear forces involved in flow cytometry without tearing or substantially deforming?

We tested the ability of our liposomes to hold up under pressure by injecting them into a Luminex LX200 flow cytometry system. This was done in collaboration with the Luminex Corporation who were interested in developing a technical manual for analysing GUVs using their system, whereas our lab was interested in describing the feasibility of analysing GUVs with similar techniques. The liposome samples were prepared and incubated overnight in Delft, then transported by icebox to the LX200 in the University Medical Center Utrecht, using public transit, showing a surprising resiliency to the stresses of such a transportation process.

Before injection the samples were diluted 100 times in PURE buffer. The actual imaging took place at a speed of between 5 and 70 liposomes per second using the luminex proprietary ImageStreamX Mark II and FlowSight software. One sample was diluted further to facilitate imaging. Each sample was run for a time span long enough to build a sufficient collection of individual liposome images. These images could then be analysed using their proprietary IDEAS Statistical Image Analysis Software. We also performed our own image analysis.



**Figure 15: ImageStream data shows that both the Cy5 membrane dye and the dsGreen DNA staining can be detected in flow cytometry.** The images shown here were created using the ImageStream software. They show three examples of detected unilamellar liposomes with their bright field (BF), dsGreen and Cy5 membrane dye channel.

The images showed no clear deformation of the GUVs (fig. 15). This collaboration also demonstrated that it is possible to detect the GUV membrane dye staining as well as dsGreen DNA staining in flow cytometry. This would suggest that screening GUVs for both LactC2-mCherry and dsGreen signal using flow cytometry methods should be possible. This would support the idea of FACS as a feasible future directed evolution screening method.

### 5.8.2 Deep learning-guided directed evolution

The speed at which directed evolution can be performed is largely dependent on the throughput of the screening method. Therefore, a high-throughput screening method like FACS can expand how far existing components can be modified within the time span of a directed evolution experiment. However, successive rounds of directed evolution can be time-intensive as well. So the time needed to create a library of possible protein or gene variants and the time needed to screen them also limit the 'reach' of a directed evolution experiment.

For applications like protein or nucleic acid evolution, this means that the amount of sequences sampled and therefore the amount of improvement we expect from directed evolution is limited by the available time and the throughput of the screening method. The time cost of applying directed evolution across the entire biological context of a bottom-up synthetic cell would currently be prohibitively large. Thus, in order to engineer a synthetic cellular context, directed evolution needs to be more efficient.

One way to speed up directed evolution, unique to synthetic cell, is to make directed evolution a continuous process [33]. Continuous directed evolution would forgo library construction or sequencing by driving genetic diversity through random mutation and proliferating the synthetic cells. Given a synthetic cell advanced enough to perform complete self-proliferation and division, the synthetic cells can be screened and selected based on function and subsequently diversified by random mutation while the population is grown back to the pre-selection size. With no sequencing or re-encapsulation necessary, the speed at which the directed evolution could be performed would be increased dramatically. However, the current minimal cell framework is not yet capable of division or complete self-proliferation. Therefore, in the absence of this more advanced minimal cell we need to look to other methods of accelerating directed evolution.

Recently, machine learning based functionality predictions have been combined with directed evolution to this end in an approach called machine-learning-guided directed evolution [43]. Machine learning algorithms can be used to predict fitness of RNA [44][45][46], non-coding DNA [47] and protein sequence

and structure [43][48][49]. In directed evolution, machine learning can be used to predict the fitness of biological sequences before their actual fitness is tested. This speeds up the overall directed evolution workflow, as machine learning can return a fitness prediction for a sequence much faster than any physical screening can be performed, therefore allowing to test orders of magnitude more sequences each step.

Machine learning is in essence a set of statistical techniques that learn their own decision boundaries from training on large existing datasets [50]. Training a machine learning algorithm to predict the function of a protein or gene based on its sequence therefore requires a large existing dataset of similar gene or protein sequences that have already been screened for their function. A trained machine learning model can then be used to make predictions about new input sequences. This can be thought of as extrapolating function to new sequences from the function of the known DNA or protein sequences. For the purposes of this lab, a sufficiently large dataset for either PlsB or  $\Phi$ 29 DNAP might not be available currently. However, after several rounds of conventional directed evolution, the screening and sequencing results of this lab itself could be leveraged to train such predictive models, making machine learning-guided directed evolution a feasible mid- to long-term approach to engineering new functionality in existing components. However, looking at the even longer-term approaches, there exist even more advanced ways of leveraging data to speed up directed evolution.

Deep learning [51] is a subset of machine learning methods. Distinct from other machine learning methods, deep learning learns to transform the input data into intermediate internal states to make its decisions instead of only learning from the direct input data. Deep learning stems from artificial neural networks [52]. Deep learning is called deep because the models consist of multiple layers of activation, each producing their own internal representation based on the representations of previous layers. Each successive layer therefore makes use of the learned feature extraction of the previous layers. In theory, this allows the models to learn different data processing techniques to help "make sense" of the input data. In practice, this results in deep learning performing better than other machine learning methods when given unprocessed close-to-nature input data.

Currently, deep learning is not yet the go-to machine learning technique in machine-learning-guided directed evolution. In contrast, deep learning has become state of the art in the related fields of classifying protein function [49], predicting functional mutants [53] or protein folding [54][55][56]. The successful application of deep learning in fields adjacent to guided directed evolution should encourage expansion of its application in directed evolution problems.

The most fundamental advantage of deep learning over other machine learning methods when applied to directed evolution is deep learning's ability to provide guidance along more points in the directed evolution workflow. In addition to predicting fitness in the screening process, deep learning can also be trained to produce new sequences using generative models. Generative models can be used to, in a single round of sequence generation, find fit sequences with more points of mutation than even machine-learning-guided can screen in a round. Generative models could therefore see use as a library construction and sampling tool, assisting the mutation process in directed evolution. The application of generative models to directed evolution might be a reality in the near future, with some directed evolution researchers starting their own generative model projects [57]. However, currently the use of generative models is limited to single-rounds of training and protein engineering, making full generative-model-guided directed evolution an as of yet unachieved goal.

## References

- [1] Pier Luigi Luisi, Thomas Oberholzer, and Antonio Lazcano. The notion of a dna minimal cell: a general discourse and some guidelines for an experimental approach. *Helvetica Chimica Acta*, 85(6): 1759–1777, 2002.
- [2] Pier Luigi Luisi. Toward the engineering of minimal living cells. *The Anatomical Record: An Official Publication of the American Association of Anatomists*, 268(3):208–214, 2002.
- [3] Clyde A Hutchison, Ray-Yuan Chuang, Vladimir N Noskov, Nacyra Assad-Garcia, Thomas J Deerinck, Mark H Ellisman, John Gill, Krishna Kannan, Bogumil J Karas, Li Ma, et al. Design and synthesis of a minimal bacterial genome. *Science*, 351(6280), 2016.
- [4] Abolfazl Akbarzadeh, Rogaie Rezaei-Sadabady, Soodabeh Davaran, Sang Woo Joo, Nosratollah Zarghami, Younes Hanifehpour, Mohammad Samiei, Mohammad Kouhi, and Kazem Nejati-Koshki. Liposome: classification, preparation, and applications. *Nanoscale research letters*, 8(1):102, 2013.
- [5] Yoshihiro Shimizu, Takashi Kanamori, and Takuya Ueda. Protein synthesis by pure translation systems. *Methods*, 36(3):299–304, 2005.
- [6] Duco Blanken, David Foschepoth, Adriana Calaña Serrão, and Christophe Danelon. Genetically controlled membrane synthesis in liposomes. *Nature Communications*, 11(1):4317, Aug 2020. ISSN 2041-1723.
- [7] Elisa Godino, Jonás Noguera López, David Foschepoth, Céline Cleij, Anne Doerr, Clara Ferrer Castellà, and Christophe Danelon. De novo synthesized min proteins drive oscillatory liposome deformation and regulate ftsa-ftszy cytoskeletal patterns. *Nature communications*, 10(1):1–12, 2019.
- [8] Pauline van Nies, Ilja Westerlaken, Duco Blanken, Margarita Salas, Mario Mencía, and Christophe Danelon. Self-replication of DNA by its encoded proteins in liposome-based synthetic cells. *Nature Communications*, 9(1):1583, 2018.
- [9] Luis Blanco and Margarita Salas. Replication of phage phi 29 dna with purified terminal protein and dna polymerase: synthesis of full-length phi 29 dna. *Proceedings of the National Academy of Sciences*, 82(19):6404–6408, 1985.
- [10] Mario Mencía, Pablo Gella, Ana Camacho, Miguel de Vega, and Margarita Salas. Terminal protein-primed amplification of heterologous dna with a minimal replication system based on phage  $\phi 29$ . *Proceedings of the National Academy of Sciences*, 108(46):18655–18660, 2011. ISSN 0027-8424.
- [11] Luis Blanco, Ignacio Prieto, Julio Gutiérrez, Antonio Bernad, José M Lázaro, José Miguel Hermoso, and Margarita Salas. Effect of  $\text{nh}_4^+$  ions on phi 29 dna-protein p3 replication: formation of a complex between the terminal protein and the dna polymerase. *Journal of virology*, 61(12):3983–3991, 1987.
- [12] W Dowhan. Molecular basis for membrane phospholipid diversity: why are there so many lipids? *Annual review of biochemistry*, 66(1):199–232, 1997.
- [13] Harvey T McMahon and Emmanuel Boucrot. Membrane curvature at a glance. *Journal of cell science*, 128(6):1065–1070, 2015.
- [14] Fernando Martínez-Morales, Max Schobert, Isabel M Lopez-Lara, and Otto Geiger. Pathways for phosphatidylcholine biosynthesis in bacteria. *Microbiology*, 149(12):3461–3471, 2003.
- [15] Christian Sohlenkamp and Otto Geiger. Bacterial membrane lipids: diversity in structures and pathways. *FEMS microbiology reviews*, 40(1):133–159, 2016.
- [16] Yi-Hsueh Lu, Ziqiang Guan, Jinshi Zhao, and Christian RH Raetz. Three phosphatidylglycerol-phosphate phosphatases in the inner membrane of escherichia coli. *Journal of Biological Chemistry*, 286(7):5506–5518, 2011.

- [17] Jason G Kay and Sergio Grinstein. Sensing phosphatidylserine in cellular membranes. *Sensors*, 11(2):1744–1755, 2011.
- [18] Jason G. Kay and Sergio Grinstein. Sensing phosphatidylserine in cellular membranes. *Sensors*, 11(2):1744–1755, 2011. ISSN 1424-8220.
- [19] Chenghua Shao, Valerie A Novakovic, James F Head, Barbara A Seaton, and Gary E Gilbert. Crystal structure of lactadherin c2 domain at 1.7 Å resolution with mutational and computational analyses of its membrane-binding motif. *Journal of Biological Chemistry*, 283(11):7230–7241, 2008.
- [20] Marijn van den Brink. Characterization of lactc2-mcherry as a probe for phosphatidylserine, 2019. Bachelor’s thesis, Delft University of Technology.
- [21] Jialan Shi, Christian W Heegaard, Jan T Rasmussen, and Gary E Gilbert. Lactadherin binds selectively to membranes containing phosphatidyl-l-serine and increased curvature. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1667(1):82–90, 2004.
- [22] Mikkel H Andersen, Helle Graversen, Sergey N Fedosov, Torben E Petersen, and Jan T Rasmussen. Functional analyses of two cellular binding domains of bovine lactadherin. *Biochemistry*, 39(20):6200–6206, 2000.
- [23] Tony Yeung, Gary E Gilbert, Jialan Shi, John Silviu, Andras Kapus, and Sergio Grinstein. Membrane phosphatidylserine regulates surface charge and protein localization. *Science*, 319(5860):210–213, 2008.
- [24] Peter A Leventis and Sergio Grinstein. The distribution and function of phosphatidylserine in cellular membranes. *Annual review of biophysics*, 39:407–427, 2010.
- [25] David L Daleke. Phospholipid flippases. *Journal of Biological Chemistry*, 282(2):821–825, 2007.
- [26] Robert FA Zwaal, Paul Comfurius, and Edouard M Bevers. Lipid–protein interactions in blood coagulation. *Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes*, 1376(3):433–453, 1998.
- [27] Marcel HAM Fens, Enrico Mastrobattista, Anko M De Graaff, Frits M Flesch, Anton Ultee, Jan T Rasmussen, Grietje Molema, Gert Storm, and Raymond M Schiffelers. Angiogenic endothelium shows lactadherin-dependent phagocytosis of aged erythrocytes and apoptotic cells. *Blood, The Journal of the American Society of Hematology*, 111(9):4542–4550, 2008.
- [28] Anne Doerr, Elise de Reus, Pauline van Nies, Mischa van der Haar, Katy Wei, Johannes Kattan, Aljoscha Wahl, and Christophe Danelon. Modelling cell-free rna and protein synthesis with minimal systems. *Physical biology*, 16(2):025001, 2019.
- [29] Duco Blanken, Pauline van Nies, and Christophe Danelon. Quantitative imaging of gene-expressing liposomes reveals rare favorable phenotypes. *Physical Biology*, 16(4):045002, 2019.
- [30] Pauline van Nies. *Virus-inspired dna replication coupled with gene expression in a minimal cell framework*. PhD thesis, 2017. Casimir PhD Series, Delft-Leiden 2017-02.
- [31] Frances H Arnold. Design by Directed Evolution. *Accounts of Chemical Research*, 31(3):125–131, mar 1998. ISSN 0001-4842.
- [32] Michael S Packer and David R Liu. Methods for the directed evolution of proteins. *Nature Reviews Genetics*, 16(7):379–394, 2015.
- [33] Zhanar Abil and Christophe Danelon. Roadmap to building a cell: An evolutionary approach. *Frontiers in Bioengineering and Biotechnology*, 8:927, 2020.
- [34] Guangyu Yang and Stephen G Withers. Ultrahigh-throughput facs-based screening for directed enzyme evolution. *ChemBioChem*, 10(17):2704–2715, 2009.

- [35] Nao Nitta, Takeaki Sugimura, Akihiro Isozaki, Hideharu Mikami, Kei Hiraki, Shinya Sakuma, Takanori Iino, Fumihito Arai, Taichiro Endo, Yasuhiro Fujiwaki, et al. Intelligent image-activated cell sorting. *Cell*, 175(1):266–276, 2018.
- [36] Madina B Iskakova, Witold Szaflarski, Marc Dreyfus, Jaanus Remme, and Knud H Nierhaus. Troubleshooting coupled in vitro transcription–translation system derived from *escherichia coli* cells: synthesis of high-yield fully active proteins. *Nucleic acids research*, 34(19):e135–e135, 2006.
- [37] Oliver Woodford. imdisp. MATLAB Central File Exchange, 2020. Version 1.12.0.0. URL <https://www.mathworks.com/matlabcentral/fileexchange/22387-imdisp>. Retrieved August 11, 2020.
- [38] Tereza Pereira de Souza, Alfred Fahr, Pier Luigi Luisi, and Pasquale Stano. Spontaneous encapsulation and concentration of biological macromolecules in liposomes: an intriguing phenomenon and its relevance in origins of life. *Journal of molecular evolution*, 79(5-6):179–192, 2014.
- [39] Zohreh Nourian and Christophe Danelon. Linking genotype and phenotype in protein synthesizing liposomes with external supply of resources. *ACS synthetic biology*, 2(4):186–193, 2013.
- [40] José A Esteban, Luis Blanco, Laurentino Villar, and Margarita Salas. In vitro evolution of terminal protein-containing genomes. *Proceedings of the National Academy of Sciences*, 94(7):2921–2926, 1997.
- [41] Montserrat Elías-Arnanz and Margarita Salas. Bacteriophage  $\phi$ 29 dna replication arrest caused by codirectional collisions with the transcription machinery. *The EMBO Journal*, 16(18):5775–5783, 1997.
- [42] Frank Vitzthum, Georg Geiger, Hans Bisswanger, Herwig Brunner, and Jürgen Bernhagen. A quantitative fluorescence-based microplate assay for the determination of double-stranded dna using sybr green i and a standard ultraviolet transilluminator gel imaging system. *Analytical Biochemistry*, 276(1):59–64, 1999.
- [43] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.
- [44] Devin Willmott, David Murrugarra, and Qiang Ye. Improving RNA secondary structure prediction via state inference with deep recurrent neural networks. jun 2019.
- [45] Jacqueline Valeri, Katherine M Collins, Bianca A Lepe, Timothy K Lu, and Diogo M Camacho. Sequence-to-function deep learning frameworks for synthetic biology. *bioRxiv*, page 870055, jan 2019.
- [46] Nicolaas M Angenent-Mari, Alexander S Garruss, Luis R Soenksen, George Church, and James J Collins. Deep Learning for RNA Synthetic Biology. *bioRxiv*, page 872077, jan 2019.
- [47] Nicholas Bogard, Johannes Linder, Alexander B Rosenberg, and Georg Seelig. A deep neural network for predicting and engineering alternative polyadenylation. *Cell*, 178(1):91–106, 2019.
- [48] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, 2018.
- [49] Yunan Luo, Lam Vo, Hantian Ding, Yufeng Su, Yang Liu, Wesley Wei Qian, Huimin Zhao, and Jian Peng. Evolutionary context-integrated deep sequence modeling for protein engineering. *bioRxiv*, page 2020.01.16.908509, jan 2020.
- [50] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.
- [51] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

- [52] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [53] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.
- [54] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, pages 1–5, 2020.
- [55] Andriy Kryshchuk, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020, dec 2019. ISSN 0887-3585.
- [56] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell systems*, 8(4):292–301, 2019.
- [57] Zachary Wu, Kevin K Yang, Michael J Liszka, Alycia Lee, Alina Batzilla, David Wernick, David P Weiner, and Frances H Arnold. Signal peptides generated by attention-based neural networks. *ACS Synthetic Biology*, 9(8):2154–2161, 2020.

## A Appendix A

### A.1 SYBR Gold stained gel failed to detect the amplified pMAR2

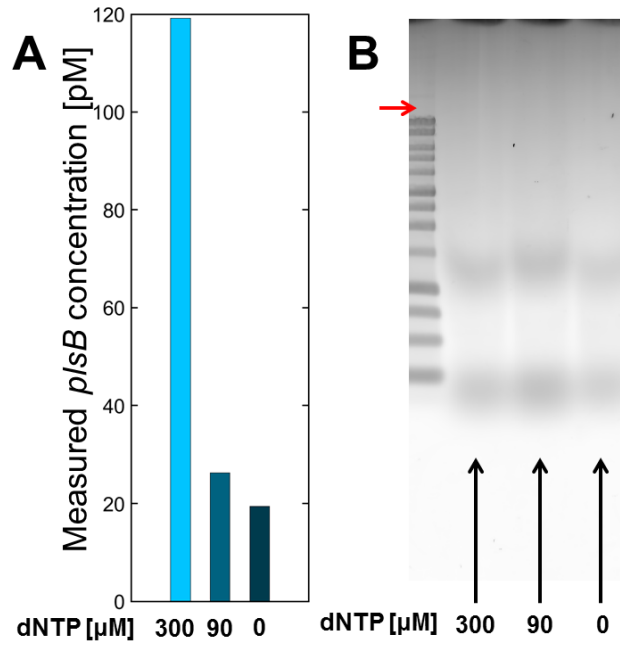
The qPCR measurement of a liposome sample containing pMAR2, G340 plasmid and SP6 RNAP was found to have a 6 fold *p/sB* amplification compared to its negative control (fig. A.1A). The samples were analysed on a gel to show the length of this amplification product. In order to load a sample from liposomes on a gel the samples were heat-inactivated at 75 °C for 10 min. Triton X-100 was added to a final concentration of 1 % to lyse the liposomes. The samples were further incubated with RNAase One and RNAase A, followed by a proteinase K treatment.

The pMAR2 construct is too long for the conventional post-IVTT clean-up protocol using RNeasy (QIAGEN), RNA column purification. In previous work in this lab [30], the full length replication of the  $\Phi$ 29 genome (~20 kb) with purified p2, p3, p5 and p6 proteins was confirmed by gel analysis. To perform clean-up, the  $\Phi$ 29 genome was loaded on a buffer exchange column with a cut-off of 100 kDa or lower. After centrifugation, the supernatant in the column will still contain the DNA template, but not the small peptides or oligonucleotides.

A 100 kb cut-off VIVACON 500 buffer exchange column (Sartorius Stedim Biotech) was used to purify pMAR2 from the sample. The samples were loaded onto the columns and centrifuged at 3,000 g for 10 minutes. The columns were then reversed and spun at 2,500 g for 2 min to retrieve the supernatant. The equivalent of 3.7 ng (using the *p/sB* concentration measured by qPCR and assuming it to be full-length pMAR2) of DNA was loaded onto the gel, with equal volume for each sample.

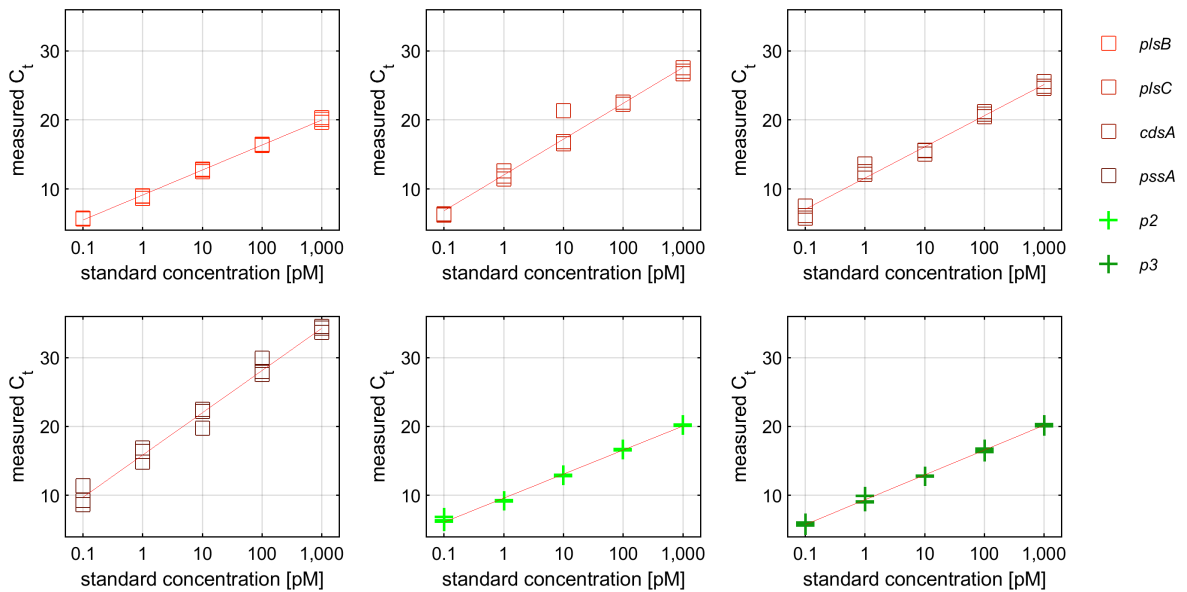
To get the sensitivity needed to pick up the small quantities of DNA loaded on this gel, the gel was stained using the, self-described, ultra-sensitive SYBR Gold Nucleic Acid Gel Stain (Thermo Fisher Scientific). The gel was stained with 10  $\mu$ l SYBR gold in 100 ml of TAE buffer and put on a shaker for 50 minutes. The resulting gel was analysed using a Amersham Typhoon Gel and Blot Imaging System with the Cy2 excitation laser (closest available for SYBR gold). The resulting gel images showed no bands at the length of pMAR2 for 400 PMV (fig. A.1B) and were overexposed for any higher laser exposure.

The gel showed no sign of a band corresponding to pMAR2. The laser exposure we could use was limited by the fact that the benchtop ladder was not diluted and would overexpose easily. However, even when the benchtop ladder was physically cut out of the gel the highest exposure still did not yield any pMAR2 bands. It is possible that the pMAR2 purification using buffer exchange columns has a very low efficiency, some DNase I might have still been active or that the actual amplification here was of a smaller *p/sB*-containing parasitic DNA molecule. Since there are many potential points of failure in this protocol it was abandoned for a qPCR-based approach to confirming full-length replication (fig. 13).



**Figure A.1: SYBR Gold stained gel showed no band corresponding to pMAR2 (A)** The *p/sB* concentration of the samples as detected by qPCR. The samples had three different dNTP concentrations. **(B)** SYBR gold stained gel with the expected height of pMAR2 indicated by the red arrow.

## A.2 Standard curves for six-gene qPCR analysis

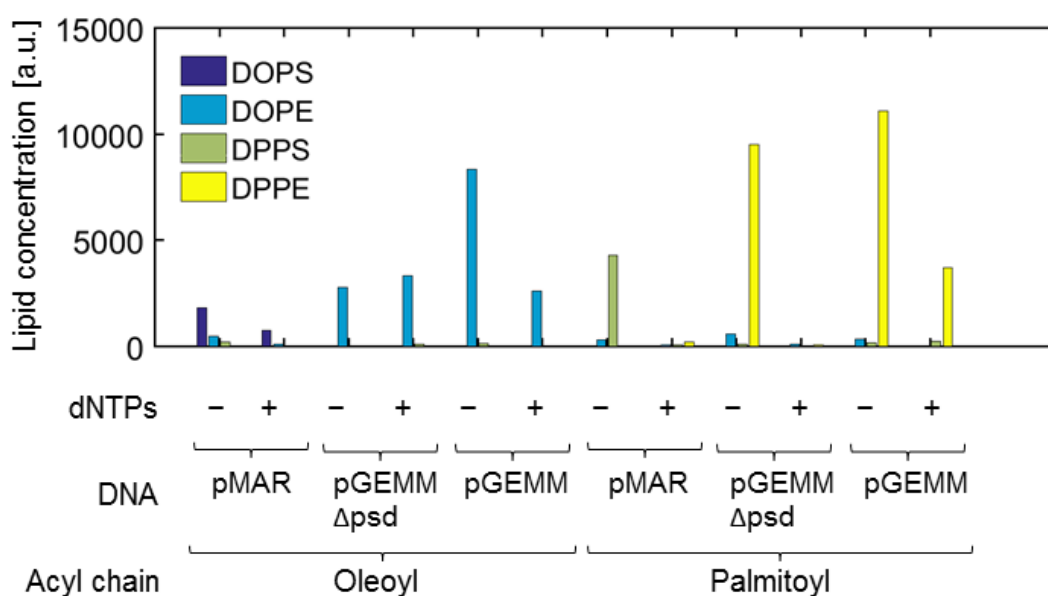


**Figure A.2: Standard curves for all of the six genes measured in the qPCR experiment.** The standards measured line up into a neat logarithmic relationship between the measured  $C_t$  and the standard concentration, although *plsC*, *cdsA* and *pssA* have a comparatively steep slope. Each standard had 3 technical repeats, similar to those of the sample measurements.

### A.3 Investigation of the effect of dNTPs on lipid synthesis was likely detecting technical error

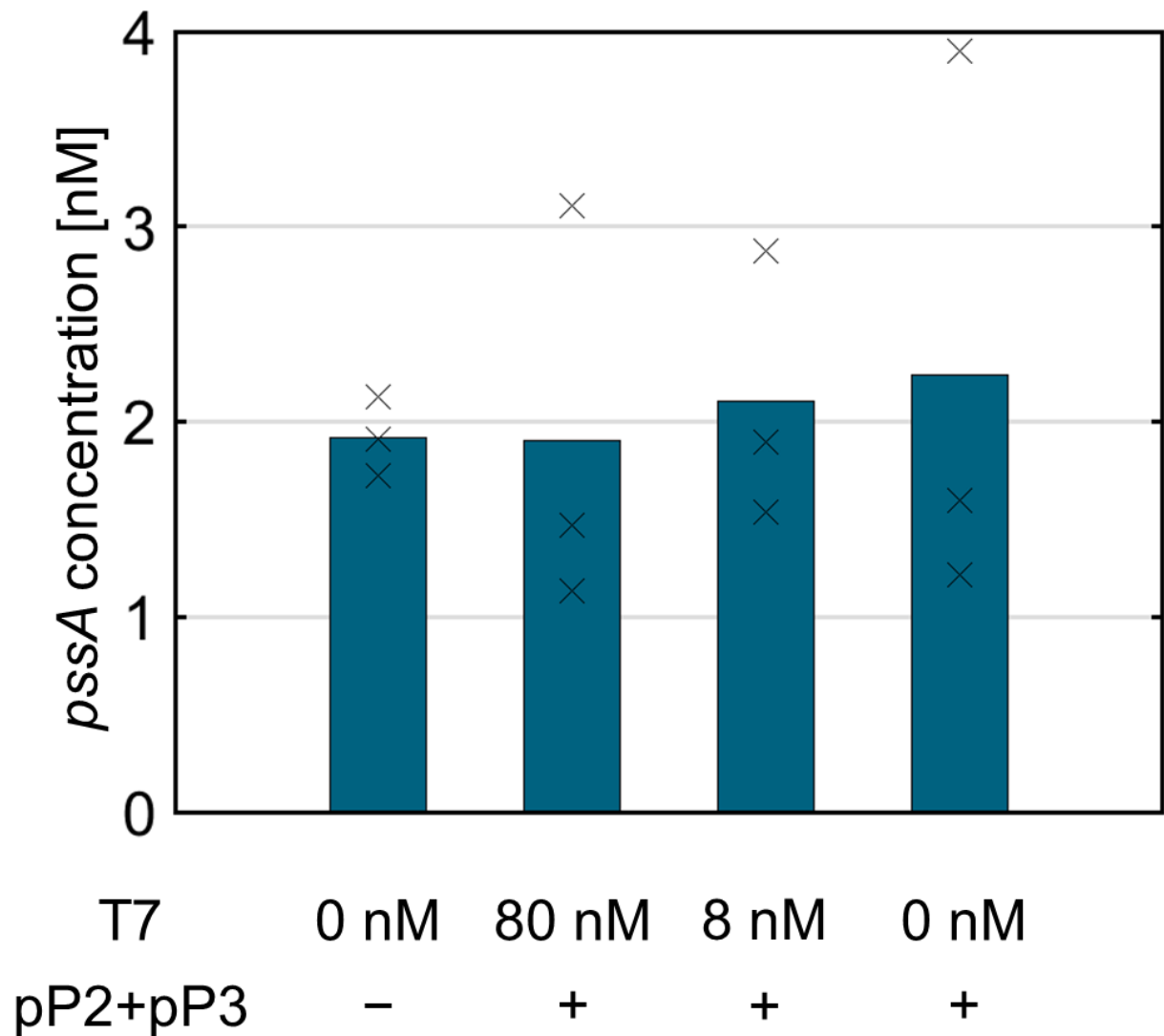
This project also featured a series of experiments measuring the effect of dNTP concentration on lipid synthesis, after preliminary experiments had shown that the LactC2-mCherry staining was higher in samples without dNTPs. To test whether this negative effect was truly attributable to the presence of dNTP a series of experiments were performed with small unilamellar vesicles (SUVs) using different DNA constructs and even different acyl-CoA lipid precursors. LC-MS lipidomics measurements showed a consistently higher lipid production in samples that did not contain dNTPs (fig. A.3). It also showed that the pGEMM7.0( $\Delta psd$ ) still mainly synthesized PE lipids suggesting the *psd* digestion was incomplete in our DNA.

The effect of dNTPs on the lipid synthesis was no longer detected when a new stock solution of dNTPs was used. This suggests that the effect observed was likely due to a unknown contaminant in the dNTPs rather than the presence of dNTPs themselves.



**Figure A.3: Unquantified lipid concentrations as measured by LC-MS lipidomics analysis with and without dNTPs.** These results showed a higher lipid yield without dNTPs across several DNA templates and lipid precursors. The negative effect of dNTPs stopped being detected after a different stock solution of dNTPs was used.

#### A.4 Bulk experiments show no amplification in qPCR measurements



**Figure A.4: The measured *pssA* concentrations of a bulk experiment that included purified p2 and p3 show no amplification.** Two of the measured samples (second and third here) are the same as those in the figure 14 proteomics measurements. The three repeats are technical in nature and come from the same sample. None of the samples showed any significant change from the input concentration of 2 nM of pMAR2.

## B Appendix B

### B.1 SMELDit Manual

The code for SMELDit was developed and run on MATLAB 2017B. The workflow of using SMELDit can be split into two parts: processing input images and analysing the resulting data.

#### B.1.1 SMELDit setup

Currently, the datastructure of the inputs of SMELDit needs to confirm to a rigid format of three image channels split into separate files. All images of an experiment should be present in the same folder with the naming convention `sample[sample number]_[3 digit number of image in sample]_[c1, c2 or c3 denoting the channel].tiff`. So, `sample2_004_c1.tiff` would be the first channel of the fourth image of the second sample of the experiment.

The laser scanning confocal microscope used in this project had an `.nd2` output file format. The `SplitChannels.ijm` was designed to automatically split these image files into their respective channels and save them as `.tiff` files. Before running the macro, make sure all `.nd2` files are present in the same folder with no other files. Then when running the macro, in the first prompt select the folder with the `.nd2` files. The next three prompts will ask the user in which directories the split channel images will be outputted. For use with SMELDit, the user should select the same output directory three times.

The code for SMELDit itself is split into three MATLAB files, `imdisp.m`, `Save_Recognized_liposomes.m` and `Load_Recognized_liposomes.m`. All three should be present in the same directory as the split channel image files.

#### B.1.2 Performing the image processing

The image processing part of SMELDit is handled by the `Save_Recognized_liposomes.m` code. This code will take the split channel images, recognize the liposomes in them, index these liposomes, save each liposome as its own cropped image and then store the indexed variables of these liposomes in a `.mat` datafile.

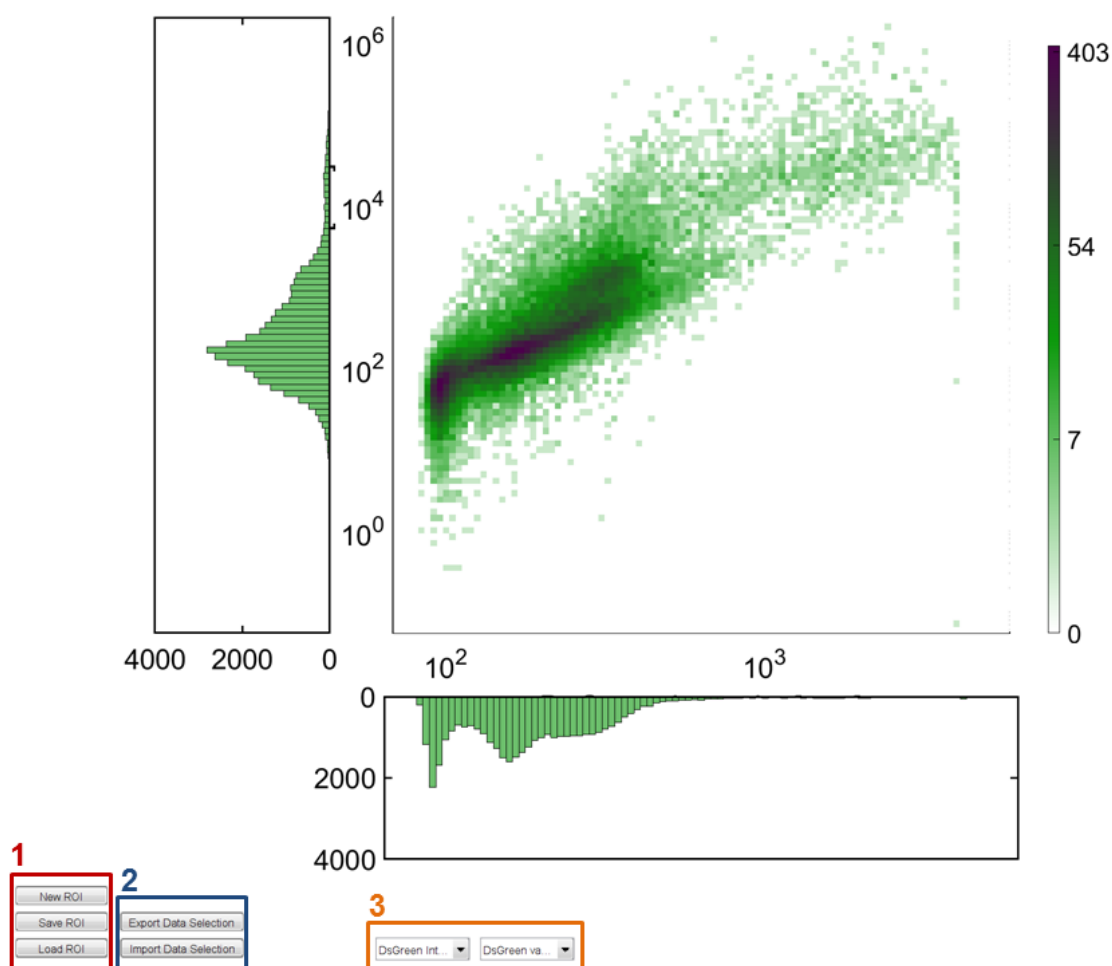
Before `Save_Recognized_liposomes.m` can be run several parameters need to be set manually for the experiment. First, `Nsample` should be set to the number of samples and `Nimg` should be a vector of length `Nsample` containing the number of images to process in each sample. Then the suffixes of the channel files should be set to their corresponding channels. Finally, the `folder` should be set to the current directory + `'/Recognized Liposomes/'` and `experimentDate` can be set to whatever the output data file should be labelled as.

Running the `Save_Recognized_liposomes.m` with the correct parameters will result in a data file and a directory with the cropped liposome images for each sample.

#### B.1.3 Loading, Analysing and Plotting the liposomes

The `Load_Recognized_liposomes.m` code is used to load and analyse existing data files and cropped images. Before the code can be run some parameters have to be set again. Namely, the `sampleNumber` needs to be set to the sample that the user wants to load. Finally, the `folder` and `experimentDate` need to be set again.

Once the parameters have been set running the code will result in a plotted heatmap with several UI features as depicted by figure B.1.



**Figure B.1: Example of a heatmap plot produced by SMELDiT** The axis labels are not plotted as these can often best be added later in an appropriate fontsize. In this particular example the x-axis was set to be dsGreen mean intensity and the y-axis was set to be dsGreen variance. The colorbar is logarithmically scaled and displays the number of liposome for each bin of the heatmap. The one dimensional histograms to the side show the distribution corresponding to the metric of that axis. (1) This section includes: the button that let the user define a new ROI, the button that lets the user save the current ROI settings as a `.mat` file and the button that loads these saved ROI settings. (2) This section includes: UI button that lets the user export the liposomes inside the current ROI as its own `.mat` dataset and the button to import existing datasets. (3) This section contains the selection scroll-lists of the metrics. Changing a metric here will redraw the heatmap to match its new axes. The left and right scroll-lists correspond to the x and y axis respectively.

The user is now ready to start their analysis by simply clicking on the **New ROI** button in section (1) and drawing a polygon across the heatmap. The vertices of the polygon are drawn by left mouse clicks and the polygon is closed by a single right mouse click. Then the `imdisp.m` file will be called to produce a montage of liposomes inside the ROI, all of which can be scrolled through by using the arrow keys. Finally, the percentage of liposomes inside the current defined ROI is displayed in the top left corner of the heatmap plot.

## B.2 SMELDit Code

The code for SMELDit is split into 1 ImageJ macro and 3 MATLAB files:

- SplitChannels.ijm
- imdisp.m<sup>†</sup> [37]
- Save\_Recognized\_liposomes.m
- Load\_Recognized\_liposomes.m

<sup>†</sup>Imdisp is part of a published MATLAB toolkit for displaying images and is therefore not listed in this appendix.

### B.2.1 Splitting the channels of the input image automatically

The code for SplitChannels.ijm was:

---

```
inputDir = getDirectory("Choose composite image directory! ");
chDir1 = getDirectory("Choose directory to save ch1! ");
chDir2 = getDirectory("Choose directory to save ch2! ");
chDir3 = getDirectory("Choose directory to save ch3! ");

fileList1 = getFileList(inputDir);

setBatchMode(true);
for (i = 0; i < fileList1.length; i++) {
    showProgress(i, fileList1.length);
    file1 = fileList1[i];
    run("Bio-Formats Importer", "open=" + inputDir + file1 + " color_mode=Default view=
        ↪ Hyperstack stack_order=XYZCT");
    id1 = getTitle();
    id1_base=replace(id1, ".nd2", "");
    run("Split Channels");

    chnumber= getList("image.titles");
    for (c = 0; c < chnumber.length; c++) {
        cn=c+1;
        selectWindow("C"+cn+"-"+id1);

        if (cn==1) {
            rename(id1_base+"_c3");
            idch1 = getTitle();
            saveAs("Tiff", chDir1 + "/" + idch1);}

        if (cn==2) {
            rename(id1_base+"_c2");
            idch3 = getTitle();
            saveAs("Tiff", chDir3 + "/" + idch3);}

        if (cn==3) {
            rename (id1_base+"_c1");
            idch2 = getTitle();
            saveAs("Tiff", chDir2 + "/" + idch2);}
    }
    run("Close All");
}

setBatchMode(false);
showMessage("Finished splitting channels!");
```

---

## B.2.2 Analysing, Indexing and Saving Liposomes

The code for `Save_Recognized_liposomes.m` was:

```
clear all
close all

Nsample=6; % Number of samples
Nimg=[5 5 5 5 5 5]; % Number of images in each sample (length should match the value of
    ↳ Nsample)

for sampleNumber = 1:6

    clearvars -except sampleNumber Nsample Nimg

    % variables you need to set manually for each experiment
    Cy5_channel='c1'; %suffix for Cy5 channel
    dsGreen_channel='c3'; %suffix for dsGreen channel
    mCherry_channel='c2'; %suffix for mCherry channel

    Nimages=Nimg(sampleNumber); %assigns the number of images in this sample as defined
    ↳ above
    sampleName=['sample' num2str(sampleNumber)]; %name of the sample

    %folder in which the cropped liposome images will be saved
    %adjust the directory to 'yourfilepath/foldercontainingthecodefile/Recognized
    ↳ Liposomes/'
    folder = ['UR:/filepath/' yourdirectory '/Recognized Liposomes/' sampleName];
    experimentDate='yyyy_mm_dd'; %date of experiment is used to name the datafile

    %%rest of the code

    if exist(folder, 'dir')
    else
        mkdir(folder)
    end

    StoreRecognizedLiposomes= true; %save cropped liposome images (default to true unless
    ↳ already present)
    CropSize=30; %radius of the cropped images (setting to 30 means 60x60 pixel image size
    ↳ )
    %Values that are used in saving the data from image analysis
    Metrics={'Area','Radius','DsGreen Intensity','DsGreen variance','mCherry intensity','
    ↳ mCherry variance','Cy5 intensity'};
    varNames={'Area','R','dsGreen','dsGreen_var','mCherry','mCherry_var','Cy5'};
    Units={'# pixels','\num', 'a.u.', 'a.u.', 'a.u.', 'a.u.', 'a.u.'};
    PlotScaleLog=[false false true true true true false];
    f = waitbar(0, ['Processed ' num2str(0) ' out of ' num2str(Nimages) ' images.']);
    for i = 1:Nimages

        I_dsGreen = imread([sampleName '_' num2str(i,'%03.f') '_' dsGreen_channel '.tif']);
        I_mCherry = imread([sampleName '_' num2str(i,'%03.f') '_' mCherry_channel '.tif']);
        I_Cy5 = imread([sampleName '_' num2str(i,'%03.f') '_' Cy5_channel '.tif']);

        [Area{i}, dsGreen{i}, dsGreen_var{i}, centroids{i}, mCherry{i}, mCherry_var{i}, Cy5{i
        ↳ }] = image_analysis(I_Cy5, I_mCherry, I_dsGreen);
        ImInx{i}=ones(1,size(centroids{i},2)).*i;

        % Make the images for each recognized liposome
        if StoreRecognizedLiposomes
            padded_dsGreen = padarray(I_dsGreen, [CropSize CropSize], 0);
            padded_Cy5 = padarray(I_Cy5, [CropSize CropSize], 0);
            padded_mCherry = padarray(I_mCherry, [CropSize CropSize], 0);
            centroids{i}=round(centroids{i});
            for jj = 1:size(centroids{i},2)
```

```

IC_dsGreen=imcrop(padded_dsGreen,[centroids{i}(:,jj)' CropSize*2-1 CropSize
    ↳ *2-1]).*2^4;
IC_Cy5=imcrop(padded_Cy5,[centroids{i}(:,jj)' CropSize*2-1 CropSize*2-1]).*2^4;
IC_mCherry=imcrop(padded_mCherry,[centroids{i}(:,jj)' CropSize*2-1 CropSize
    ↳ *2-1]).*2^4;
IC = cat(3, IC_mCherry, IC_dsGreen, zeros(2*CropSize,2*CropSize));
IC2 = cat(3, IC_Cy5,IC_Cy5,IC_Cy5);
IC=imfuse(IC,IC2,'blend');
baseFileName=[sampleName '_00' num2str(i) '_' sprintf( '%05d', jj ) '.tif'];
fullFileName = fullfile(folder, baseFileName);
imwrite(IC, fullFileName);

end
end
waitbar(i/Nimages,f, ['Processed ' num2str(i) ' out of ' num2str(Nimages) ' images.'])
end

Area= cell2mat(Area);
dsGreen= cell2mat(dsGreen);
mCherry= cell2mat(mCherry);
Cy5= cell2mat(Cy5);
mCherry_var= cell2mat(mCherry_var);
dsGreen_var= cell2mat(dsGreen_var);
ImInx= cell2mat(ImInx);
R = (2+sqrt(3*Area./(2*pi)))*0.25; %erosion disk 2

save(['data_' experimentDate '_' sampleName], 'Area', 'R', 'dsGreen', 'dsGreen_var',
    ↳ 'mCherry', 'mCherry_var', 'Cy5', 'centroids', 'ImInx', 'Nimages', 'folder', 'sampleName'
    ↳ )
end

%% image analysis

function [Area, dsGreen, dsGreen_var, Centroids, mCherry, mCherry_var, Cy5] =
    ↳ image_analysis(I_Cy5,I_mCherry, I_dsGreen)

h = [-1 -1 -1; -1 12 -1; -1 -1 -1]./4;
I_sharp = imfilter(I_Cy5+I_mCherry,h);
I_filled = imfill(I_sharp);
I_inside = I_filled-I_sharp;
fgm = zeros(size(I_Cy5));
fgm(I_inside<=200)= 0;
fgm(I_inside>200)= 1;

fgm = imfill(fgm,'holes');
se = strel('disk',2);
fgm = imerode(fgm,se);
fgm2=zeros(size(fgm));

CC = bwconncomp(fgm);
pixels = CC.PixelIdxList;
props = regionprops(CC, 'PixelIdxList', 'Perimeter', 'Area', 'Centroid');
passed = logical(ones(1,length(pixels)));

% select for kind of circular things
for j = 1:length(pixels)
    P(j) = props(j).Perimeter;
    A(j) = props(j).Area;

    C(j) = P(j)^2/(4*pi*A(j));
    if C(j) >2
        fgm(CC.PixelIdxList{j}) = 0;
        passed(j)=false;
    elseif C(j) < 0.5
        fgm(CC.PixelIdxList{j}) = 0;

```

```

        passed(j)=false;
    elseif mean(I_mCherry(CC.PixelIdxList{j}))>500
        fgm(CC.PixelIdxList{j}) = 0;
        passed(j)=false;
    end

end

clear CC
clear pixels

% determine observables
CC = bwconncomp(fgm);
x_size=size(fgm,1);
y_size=size(fgm,2);
pixels = CC.PixelIdxList;
pixels2 = pixels;

props=props(passed);
Area=zeros(1,length(pixels));
dsGreen=zeros(1,length(pixels));
mCherry=zeros(1,length(pixels));
dsGreen_var=zeros(1,length(pixels));
Centroids=zeros(2,length(pixels));

for j = 1:length(pixels)
    Area(j) = length(pixels{j});
    dsGreen(j) = mean(I_dsGreen(pixels{j}));
    dsGreen_var(j)= var(single(I_dsGreen(pixels{j})));
    Centroids(:,j)=props(j).Centroid;

    %expand each element kk=6 pixels away from border
    pix=cell2mat(pixels{j});
    for kk=1:6
        isnotborder=(pix>x_size) & (pix<x_size*(y_size-1)+1 & ((rem(pix,x_size)~=0)&(rem(
            ↪ pix,x_size)~=1)));
        pix_notborder=pix(isnotborder);
        connect=[pix_notborder-x_size; pix_notborder-1; pix_notborder+1; pix_notborder+
            ↪ x_size];
        pix=[pix; connect];
        pix=unique(pix);
    end
    pixels2{j}=pix;
    pix=cell2mat(pixels2{j});
    pixborder{j}=pix(~ismember(pix,cell2mat(pixels{j})));
    mCherry(j) = mean(I_mCherry(pixborder{j}));
    Cy5(j) = mean(I_Cy5(pixborder{j}));
    mCherry_var(j) = var(single(I_mCherry(pixborder{j})));
end
end

```

---

### B.2.3 Loading and Plotting indexed liposomes

The code for Load\_Recognized\_liposomes.m was:

```
clear all
close all
set(groot,{'DefaultAxesXColor','DefaultAxesYColor','DefaultAxesZColor',
    ↳ DefaultAxesLineWidth'},{'k','k','k',1})

%% Parameters to set
experimentDate='yyyy_mm_dd'; %Needed to load the right data file

sampleNumber=4; %The number of the sample to load
sampleName=['sample' num2str(sampleNumber)]; %Sample name

%folder that corresponds to the directory that contains the saved cropped liposome images
folder = ['UR:/filepath/' yourdirectory '/Recognized Liposomes/' sampleName];

%% Body of the code
load(['data_' experimentDate '_' sampleName'],'Area','R','dsGreen','dsGreen_var','mCherry'
    ↳ , 'mCherry_var','Cy5','centroids','ImInx','Nimages')
Metrics={'Area','Radius','DsGreen Intensity','DsGreen variance','mCherry intensity','
    ↳ mCherry variance','Cy5 intensity'};
varNames={'Area','R','dsGreen','dsGreen_var','mCherry','mCherry_var','Cy5'};
Units={'# pixels','\mum','a.u.','a.u.','a.u.','a.u.','a.u.'};
PlotScaleLog=[false false true true true true];

Nmontage=64; %Number of images in a single montage
Nbins=100; %Bins in the 2d histogram of the heatmap

x_metric =3; y_metric=4; %The metrics that are displayed initially

figure(1)
plotHeatmap(x_metric,y_metric,Nbins)
pause(0.00001);
frame_h = get(handle(gcf),'JavaFrame');
set(gca,'FontSize',20);
set(frame_h,'Maximized',1);
% set(gcf, 'Units', 'Normalized', 'OuterPosition', [0, 0.04, 9/16, 1]);

%% UI buttons defined

c = uicontrol;
c.String = 'New ROI';
c.Callback = @selectROI;
c.Position = [5 65 90 20];

c2 = uicontrol;
c2.String = 'Save ROI';
c2.Callback = @saveROI;
c2.Position = [5 40 90 20];

c3 = uicontrol;
c3.String = 'Load ROI';
c3.Callback = @loadROI;
c3.Position = [5 15 90 20];

c4 = uicontrol;
c4.String = 'Export Data Selection';
c4.Callback = @exportData;
c4.Position = [105 40 120 20];

c5 = uicontrol;
c5.String = 'Import Data Selection';
c5.Callback = @importData;
c5.Position = [105 15 120 20];
```

```

c_x = uicontrol('Style','popupmenu');
c_x.String = Metrics;
c_x.Position = [350 15 90 20];
c_x.Callback = @selection_x;

c_y = uicontrol('Style','popupmenu');
c_y.String = Metrics;
c_y.Position = [450 15 90 20];
c_y.Callback = @selection_y;

set(c_y,'Value',y_metric)
set(c_y,'Callback',{@selection_y,c_x,c_y,x_metric,Nbins})

set(c_x,'Value',x_metric)
set(c_x,'Callback',{@selection_x,c_x,c_y,y_metric,Nbins})

%% UI functions

function selection_x(src,event,c_x,c_y,y_metric,Nbins)
    x_metric = c_x.Value;
    plotsurf=evalin('base','plotsurf');
    set(c_x,'Callback',{@selection_x,c_x,c_y,y_metric,Nbins})
    set(c_y,'Callback',{@selection_y,c_x,c_y,x_metric,Nbins})
    assignin('base','x_metric',x_metric)
    plotHeatmap(x_metric,y_metric,Nbins,plotsurf)
end

function selection_y(src,event,c_x,c_y,x_metric,Nbins)
    y_metric = c_y.Value;
    plotsurf=evalin('base','plotsurf');
    set(c_x,'Callback',{@selection_x,c_x,c_y,y_metric,Nbins})
    set(c_y,'Callback',{@selection_y,c_x,c_y,x_metric,Nbins})
    assignin('base','y_metric',y_metric)
    plotHeatmap(x_metric,y_metric,Nbins,plotsurf)
end

function loadROI(src,event)
    pgon=evalin('base','pgon');
    delete(pgon);
    [ROIfile,ROIpath] = uigetfile;
    load([ROIpath ROIfile],'roiMask','position','edgeX','edgeY','x_metric','y_metric')

    plotsurf=evalin('base','plotsurf');
    Nbins=evalin('base','Nbins');
    plotHeatmap(x_metric,y_metric,Nbins,plotsurf,edgeX,edgeY)

    pgon = polyshape(position);
    figure(1),hold on, pgon=plot(pgon);
    assignin('base','pgon',pgon)

    c_x=evalin('base','c_x');
    c_y=evalin('base','c_y');

    set(c_x,'Value',x_metric)
    set(c_y,'Value',y_metric)
    set(c_x,'Callback',{@selection_x,c_x,c_y,y_metric,Nbins})
    set(c_y,'Callback',{@selection_y,c_x,c_y,x_metric,Nbins})

    assignin('base','c_x',c_x)
    assignin('base','c_y',c_y)

    makeMontage(roiMask)
end

function saveROI(src,event)
    [ROIfile,ROIpath] = uiputfile('*.mat','Workspace File','ROIfile.mat');

```

```

roiMask=evalin('base', 'roiMask');
position=evalin('base', 'position');
edgeX=evalin('base', 'edgeX');
edgeY=evalin('base', 'edgeY');
x_metric=evalin('base', 'x_metric');
y_metric=evalin('base', 'y_metric');

save([ROIpath ROIfile], 'roiMask', 'position', 'edgeX', 'edgeY', 'x_metric', 'y_metric')
end

function exportData(src,event)
[datafile,datapath] = uiputfile('*.mat','Workspace File','dataSelection.mat');
Area=evalin('base', 'Area(testtot)');
R=evalin('base', 'R(testtot)');
dsGreen=evalin('base', 'dsGreen(testtot)');
dsGreen_var=evalin('base', 'dsGreen_var(testtot)');
mCherry=evalin('base', 'mCherry(testtot)');
mCherry_var=evalin('base', 'mCherry_var(testtot)');
Cy5=evalin('base', 'Cy5(testtot)');
centroids=evalin('base', 'centroids');
ImInx=evalin('base', 'ImInx');
Nimages=evalin('base', 'Nimages');
testtot=evalin('base', 'testtot');
inx=evalin('base', 'inx');
folder=evalin('base', 'folder');
sampleName=evalin('base', 'sampleName');
start=1;

save([datapath datafile], 'Area', 'R', 'dsGreen', 'dsGreen_var', 'mCherry', 'mCherry_var', '
    ↳ Cy5', 'centroids', 'ImInx', 'Nimages', 'inx', 'sampleName', 'folder')
end

function importData(src,event)
[datafile,datapath] = uigetfile;
load([datapath datafile], 'Area', 'R', 'dsGreen', 'dsGreen_var', 'mCherry', 'mCherry_var', '
    ↳ Cy5', 'centroids', 'ImInx', 'Nimages', 'inx', 'sampleName', 'folder')

assignin('base', 'Area', Area);
assignin('base', 'R', R);
assignin('base', 'dsGreen', dsGreen);
assignin('base', 'dsGreen_var', dsGreen_var);
assignin('base', 'mCherry', mCherry);
assignin('base', 'mCherry_var', mCherry_var);
assignin('base', 'Cy5', Cy5);
assignin('base', 'centroids', centroids);
assignin('base', 'ImInx', ImInx);
assignin('base', 'Nimages', Nimages);
assignin('base', 'inxBase', inx);
assignin('base', 'folder', folder);
assignin('base', 'sampleName', sampleName);

x_metric=evalin('base', 'x_metric');
y_metric=evalin('base', 'y_metric');
Nbins=evalin('base', 'Nbins');
plotsurf=evalin('base', 'plotsurf');
plotHeatmap(x_metric,y_metric,Nbins,plotsurf)
end

%% -- auxilliary functions

function selectROI(src,event)
ise = evalin('base', 'exist(''pgon'', ''var'') == 1' );
if ise
    pgon = evalin('base', 'pgon');
    delete(pgon);
end

```

```

roi = impoly;
position = getPosition(roi);
delete(roi)

Nbins=evalin('base', 'Nbins');
edgeX=evalin('base', 'edgeX');
edgeY=evalin('base', 'edgeY');
position(position(:,1)<min(edgeX),1)=min(edgeX);
position(position(:,2)<min(edgeY),2)=min(edgeY);
position(position(:,1)>max(edgeX),1)=max(edgeX);
position(position(:,2)>max(edgeY),2)=max(edgeY);

pgon_out = polyshape(position);
hold on, pgon_out=plot(pgon_out);
assignin('base','pgon',pgon_out)
assignin('base','position',position)
[~,~,polyX,polyY] = histcounts2(position(:,1),position(:,2),edgeX,edgeY);
polyX(end+1)=polyX(1);
polyY(end+1)=polyY(1);
roiMask = poly2mask(polyX,polyY,Nbins,Nbins);
roiMask = roiMask';
SE = strel('square',3);
roiMask = imdilate(roiMask,SE);
assignin('base','roiMask',roiMask)
makeMontage(roiMask);

end

function plotHeatmap(x_metric,y_metric,Nbins,plotsurf,edgeX,edgeY)

varNames=evalin('base', 'varNames');
Metrics=evalin('base', 'Metrics');
Units=evalin('base', 'Units');
PlotScaleLog=evalin('base', 'PlotScaleLog');

X = [evalin('base', varNames{x_metric}); evalin('base', varNames{y_metric})];
assignin('base','X', X)

if exist('edgeX','var') == 0
    if PlotScaleLog(x_metric)
        edgeX=logspace(min(log(X(1,X(1,:)>0)/1.2)/log(10)),max(log(X(1,:)*1.5)/log(10)),
            ↪ Nbins+1);
    else
        edgeX=linspace(min(X(1,:)/1.05),max(X(1,:)*1.05),Nbins+1);
    end

    if PlotScaleLog(y_metric)
        edgeY=logspace(min(log(X(2,X(2,:)>0)/1.2)/log(10)),max(log(X(2,:)*1.5)/log(10)),
            ↪ Nbins+1);
    else
        edgeY=linspace(min(X(2,:)/1.05),max(X(2,:)*1.05),Nbins+1);
    end
end

assignin('base','edgeX',edgeX)
assignin('base','edgeY',edgeY)

[N,~,~,BinX,BinY]=histcounts2(X(1,:),X(2,:),edgeX,edgeY);
assignin('base','BinX',BinX)
assignin('base','BinY',BinY)

if exist('plotsurf','var')
    delete(plotsurf);
    cb=evalin('base', 'cb');
    delete(cb);
end

```

```

ise = evalin( 'base', 'exist(''pgon'', ''var'') == 1' );
if ise
    pgon = evalin('base', 'pgon');
    delete(pgon);
end

sub3 = subplot(4,4,[2 3 4 6 7 8 10 11 12]);
plotsurf = surf(edgeX(1:end-1),edgeY(1:end-1),log(N'+1));
assignin('base','plotsurf',plotsurf)

if PlotScaleLog(x_metric)
    set(gca,'xscale','log')
else
    set(gca,'xscale','linear')
end

if PlotScaleLog(y_metric)
    set(gca,'yscale','log')
else
    set(gca,'yscale','linear')
end

view(0,-90)
set(gca, 'Ydir', 'reverse')
set(gca,'FontSize',20);
shading flat
map=cat(2,linspace(1,0.05,100)',linspace(1,0.6,100)',linspace(1,0.05,100)');
map2=cat(2,linspace(0.05,0.3,100)',linspace(0.6,0,100)',linspace(0.05,0.3,100)');
map=cat(1,map,map2);
xlim([min(edgeX) max(edgeX(1:end-1))])
ylim([min(edgeY) max(edgeY(1:end-1))])
xlabel([Metrics{x_metric} ' [' Units{x_metric} ']' ])
ylabel([Metrics{y_metric} ' [' Units{y_metric} ']' ])
pbaspect([1 1 1])

subplot(4,4,[1 5 9]);
histogram(X(2,:),edgeY,'FaceColor',[0.05, 0.6, 0.05]),camroll(90)
set(gca,'FontSize',20);
xlim([min(edgeY) max(edgeY(1:end-1))])
xticks('');
if PlotScaleLog(y_metric)
    set(gca,'xscale','log')
else
    set(gca,'xscale','linear')
end
subplot(4,4,[14 15 16]);
histogram(X(1,:),edgeX,'FaceColor',[0.05, 0.6, 0.05]),camroll(180),set(gca, 'Ydir', '
    ↪ reverse')
set(gca,'FontSize',20);
xlim([min(edgeX) max(edgeX(1:end-1))])
xticks('');
if PlotScaleLog(x_metric)
    set(gca,'xscale','log')
else
    set(gca,'xscale','linear')
end
subplot(4,4,[2 3 4 6 7 8 10 11 12])
pos = get(sub3,'Position');
cb = colorbar('Position',[pos(1)+pos(4) pos(2) 0.01 pos(3)]);
colormap(gca,map)
assignin('base','cb',cb)
set(cb,'Ticks',cb.Ticks,...
    'TickLabels',ceil(exp(cb.Ticks))-1)

end

```

```

function plotButtonPushed(src,event,pgon,X,centroids,ImInx,c,folder,Nimages,sampleName,
    ↪ Nmontage,Nbins)
    if exist('pgon','var')
        delete(pgon);
    end
    roi = impoly;
    position = getPosition(roi);
    delete(roi)

    edgeX=evalin('base','edgeX');
    edgeY=evalin('base','edgeY');
    BinX=evalin('base','BinX');
    BinY=evalin('base','BinY');
    position(position(:,1)<min(edgeX),1)=min(edgeX);
    position(position(:,2)<min(edgeY),2)=min(edgeY);
    position(position(:,1)>max(edgeX),1)=max(edgeX);
    position(position(:,2)>max(edgeY),2)=max(edgeY);

    pgon_out = polyshape(position);
    hold on, pgon_out=plot(pgon_out);
    [~,~,~,polyX,polyY] = histcounts2(position(:,1),position(:,2),edgeX,edgeY);
    polyX(end+1)=polyX(1);
    polyY(end+1)=polyY(1);
    roiMask = poly2mask(polyX,polyY,Nbins,Nbins);
    roiMask = roiMask';
    for ii = 1:length(X(1,:))
        if (BinX(ii)==0)&&(BinY(ii)>0)
            testtot(ii) = roiMask(BinX(ii)+1,BinY(ii));
        elseif (BinY(ii)==0) && (BinX(ii)>0)
            testtot(ii) = roiMask(BinX(ii),BinY(ii)+1);
        elseif (BinY(ii)==0) && (BinX(ii)==0)
            testtot(ii) = roiMask(BinX(ii)+1,BinY(ii)+1);
        else
            testtot(ii) = roiMask(BinX(ii),BinY(ii));
        end
    end
    set(c,'Callback',{@plotButtonPushed,pgon_out,X,centroids,ImInx,c,folder,Nimages,
        ↪ sampleName,Nmontage,Nbins})
    assignin('base','c',c)

    inx=find(testtot);

    im_recognitions=[0];
    for ii=1:Nimages-1
        im_recognitions=[im_recognitions size(centroids{ii},2)];
    end
    im_offset=cumsum(im_recognitions);

    Nmontage_used=Nmontage;
    if Nmontage > length(inx)
        Nmontage_used = length(inx);
    end

    filenamesmontage=cell(1,Nmontage_used);
    for ii=1:Nmontage_used
        jj=ii;
        Ix=inx(jj);
        baseFileName=[sampleName '_00' num2str(ImInx(inx(jj))) '_' sprintf('%05d', Ix-
            ↪ im_offset(ImInx(inx(jj)))) '.tif'];
        fullFileName = fullfile(folder, baseFileName);
        filenamesmontage{ii}= fullFileName;
    end

    figure(5), montage(filenamesmontage), title(['Showing ' num2str(Nmontage_used) '
        ↪ example liposomes'])
end

```

```

function makeMontage(roiMask)
    X=evalin('base','X');
    centroids=evalin('base','centroids');
    ImInx=evalin('base','ImInx');
    folder=evalin('base','folder');
    Nimages=evalin('base','Nimages');
    sampleName=evalin('base','sampleName');
    BinX=evalin('base','BinX');
    BinY=evalin('base','BinY');

    for ii = 1:length(X(1,:))
        if (BinX(ii)==0)&&(BinY(ii)>0)
            testtot(ii) = roiMask(BinX(ii)+1,BinY(ii));
        elseif (BinY(ii)==0) && (BinX(ii)>0)
            testtot(ii) = roiMask(BinX(ii),BinY(ii)+1);
        elseif (BinY(ii)==0) && (BinX(ii)==0)
            testtot(ii) = roiMask(BinX(ii)+1,BinY(ii)+1);
        else
            testtot(ii) = roiMask(BinX(ii),BinY(ii));
        end
    end
    fl=figure(1);
    t = uicontrol(fl,'Style','text',...
        'String',['Percentage of liposomes inside Gate: ' num2str(sum(testtot)/length
            ↳ (testtot)*100)],...
        'Position',[5 870 80 60]);

    assignin('base','testtot',testtot)

    ise = evalin('base','exist(''inxBase'', ''var'') == 0 ');
    if ise
        inxBase=linspace(1,length(testtot),length(testtot));
    else
        inxBase=evalin('base','inxBase');
    end

    inx=inxBase(testtot);

    assignin('base','inx',inx)

    im_recognitions=[0];
    for ii=1:Nimages-1
        im_recognitions=[im_recognitions size(centroids{ii},2)];
    end
    im_offset=cumsum(im_recognitions);
    Nmontage_used = length(inx);
    filenamesmontage=cell(1,Nmontage_used);

    for ii=1:Nmontage_used
        jj=ii;
        Ix=inx(jj);
        baseFileName=[sampleName '_00' num2str(ImInx(Ix)) '_' sprintf('%05d', Ix-
            ↳ im_offset(ImInx(Ix))) '.tif'];
        fullFileName = fullfile(folder, baseFileName);
        filenamesmontage{ii}= fullFileName;
    end

    figure(2)
    imdisp(filenamesmontage, 'Size', [8 8]),title(['Showing ' num2str(Nmontage_used) '
        ↳ example liposomes'])

end

```