Towards a Synthetic Nucleus: Separating Transcription and Translation in Cell-Free Protein Expression Systems

Thesis by
Zoila E. Jurado Quiroga

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
Defended May 7th, 2024

Zoila E. Jurado Quiroga
ORCID: 0000-0003-4160-5068

ACKNOWLEDGEMENTS

I want to express my sincere appreciation to my supervisor, Richard Murray, for his support and patience throughout my Ph.D. journey and the writing of this thesis. He rules our lab group as a "benevolent dictator," embodying authority and compassion in his leadership style. His influence extends far beyond the confines of the research, and it was with Richard's guidance that I have grown as a scientist, personally and professionally.

I am also grateful to Austin Minnich, Rebecca Voorhees, and Ayush Pandey, my thesis committee members. Austin was my advisor when I entered Caltech and helped mentor me as a young, bright-eyed graduate student. As a mechanical engineer, he allowed and encouraged me to dabble in bioengineering, ultimately leading to this thesis. He has continued to show support throughout the years, and I will always appreciate him for this. Rebecca has provided helpful comments and insights over the years, opening up her lab and helping me put a nail in my first project. More importantly, she resonated with me, and our conversations helped me to continue during the most challenging times. Lastly, Ayush has always been a source of support. This thesis would not have been possible without his enthusiasm, patience, and time put towards teaching me the inferencing pipeline. Whether he knew it or not, he gave me the push I needed to bring this thesis to fruition.

I thank the National Science Foundation and Air Force Office of Scientific Research for the financial support that made this research possible. I also acknowledge ChatGPT for providing overarching summaries and helping with structuring sentences, Grammarly for additional spelling and grammar corrections, and BioRender.com, which was used to create many images throughout this thesis.

I have been honored to have worked alongside many inspiring colleagues, specifically Manisha, Andrey, William, Rory, Vipul, Zach, Masami, John, and Mengyi, who have provided encouragement, assistance, and valuable feedback throughout this journey. Furthermore, this journey has introduced me to individuals like Richard C., Karena, Marcus, Kimberley, Lukas, Kostas, Joaquín, Tomoyuki, and Saumya. They have helped me push past my insecurities and boundaries through their compassion, empathy, and friendship.

I have been fortunate to meet and be surrounded by intelligent and genuine people here, including Liz and Athena from the Caltech Y. They helped me explore outside of my research bubble and provided me with opportunities to succeed or fail safely. I am most grateful to my lab manager, Miki Yun, for teaching me everything I know, from holding a pipette to making lysate, and for her support when tasks fell outside her knowledge. Miki has been more than a lab manager; she has seen me grow in multiple ways and has helped keep me steady through my years. To my undergraduate advisor, Ali Gokirmak, thank you for pushing me to apply to graduate school and checking in on me.

I extend my heartfelt love and appreciation to my family for their encouragement and unconditional love. They know better than anyone how difficult this journey has been, the journey expanding past the years at Caltech. To Colton, I could not have imagined how long the journey would be, but I am so lucky to have had your support throughout it all. This thesis is as much my success as it is ours.

Finally, I would like to dedicate this thesis to my mom. I know she has sacrificed so much for me. Although I will never be able to heal the pain she has endured, which has enabled me to be where I am today, I hope she knows that without her, none of this would have been possible.

“Sí, se puede.” - Dolores Huerta

ABSTRACT

Synthetic cells represent the culmination of decades of research aimed at deciphering the intricacies of life at its most basic level. The result of the fusion of biology, chemistry, physics, and engineering, synthetic cells promise to revolutionize biotechnology, medicine, and beyond. This thesis focuses on the ramifications of incorporating a synthetic nucleus within a synthetic cell.

To experimentally study transcription and translation, we use a commercially available cell-free protein expression system comprising all the purified proteins essential for protein production (PURE), along with a fluorescent RNA aptamer–malachite green aptamer (MGapt), and a green fluorescent protein (deGFP). We observed that the chemical composition of the PURE system significantly impacts MGapt fluorescence, leading to inaccurate RNA calculations. We identify the reducing agent, dithiothreitol (DTT), to address this challenge as a crucial chemical affecting MGapt fluorescence. We propose a model that can reliably model MGapt measurements in commercial PURE. This investigation illuminates the intricate dynamics of MGapt in PURE and emphasizes the necessity of accounting for environmental factors in RNA measurements employing aptamers.

Subsequently, to advance our understanding of a synthetic nucleus and analyze the effects of separating transcription and translation in a cell-free protein expression, we propose and validate a chemical reaction network model for transcription (TX) in PURE. Additionally, we used open-source software to expand an existing translation (TL) model for any arbitrary DNA sequence to create a nearly complete model of TX-TL in PURE. Leveraging this model, we investigate the effect of introducing a synthetic nucleus by modulating the RNA diffusion rate and resource allocation. This detailed model showcases our capability to comprehensively model protein expression in PURE, enabling insights into the efficacy of segregating transcription and translation processes within the artificial cell environment. Lastly, we provide a perspective on the future of synthetic cells with an artificial nucleus and propose further steps to develop the proposed synthetic nucleus model.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Jurado, Zoila and Murray, Richard M. "Impact of chemical dynamics of commercial PURE systems on malachite green aptamer fluorescence." *ACS Synthetic Biology* (2024, in revision). Available as bioRxiv preprint. DOI: https://doi.org/10.1101/2023.08.14.553301.
Z. J. conceived the project, developed the mathematical model, conducted all experiments and analyzed results, fitting the model to the data collected, and participated in the writing of the manuscript.

[2] Jurado, Zoila, Pandey, Ayush, and Murray, Richard M. "A chemical reaction network model of PURE." *bioRxiv* (2023). DOI: https://doi.org/10.1101/2023.08.14.553301.
Z. J. conceived the project, developed the mathematical models, conducted all experiments, analyzed results, and participated in the writing of the manuscript. In addition to participating in the manuscript's writing, A.P. developed the modeling analysis pipeline to analyze the proposed models and parameter fitting.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

Synthetic biology, a rapidly evolving interdisciplinary field, aims to engineer life from scratch. Leading this engineering challenge is creating synthetic cells. Synthetic cells serve as invaluable tools for understanding the fundamental principles of life, enabling scientists to probe the origins of cellular complexity and evolutionary dynamics. One of the distinctive components of life is the nucleus found in eukaryotic cells. However, the creation of synthetic nuclei hinges on the ability to engineer both the physical structure and the biochemical processes within them, to build tools to model such chemical reactions, and to understand how to leverage the mechanisms to create and control them. In this thesis, we present new contributions in modeling cell-free protein expression systems, taking steps towards a synthetic nucleus. This thesis will break the central dogma apart by building new transcription and expanded translation models, leading to the first full detailed model of cell-free protein synthesis. Then, we will combine these models, leading to the understanding and modeling of a synthetic nucleus.

This chapter describes a method of attacking the problem of creating a synthetic cell, along with a general background to understand the field and the need for detailed modeling of cell-free protein synthesis systems. Chapter 2 presents a new finding and method to accurately interpret RNA production in commercial cell-free protein expression systems required to build a synthetic nucleus model. Following discoveries in Chapter 2, in Chapter 3 we propose a model for gene expression in cell-free systems from a DNA strand based on experimental data. Finally, leveraging Chapter 3, we separate the transcription and translation models to effectively simulate a synthetic nucleus in Chapter 4. We will discuss the future research directions in Chapter 5 of this thesis.

1.1 Building a Synthetic Cell: Starting from "Scratch"

The drive to understand the emergence of life from inert chemical elements and organic compounds has fueled extensive interdisciplinary research spanning many decades. Despite significant progress in unraveling the molecular intricacies of living systems, numerous fundamental questions persist regarding the origin and evolution of life—from prebiotic chemistry to the intricate cellular and multicel-

lular organisms we observe today. Fundamental mysteries remain, such as the molecular mechanisms underlying the evolution of life, including phenomena like self-organization, emergence, self-replication, autopoiesis, compartmentalization, and metabolism [1]. Synthetic biology emerges as an interdisciplinary frontier aiming to establish a structured framework for designing and/or redesigning living biological systems. This field embraces both "top-down" and "bottom-up" approaches, employing modular parts to systematically reconstruct complex biological systems and cells, as outlined in Figure 1.1.

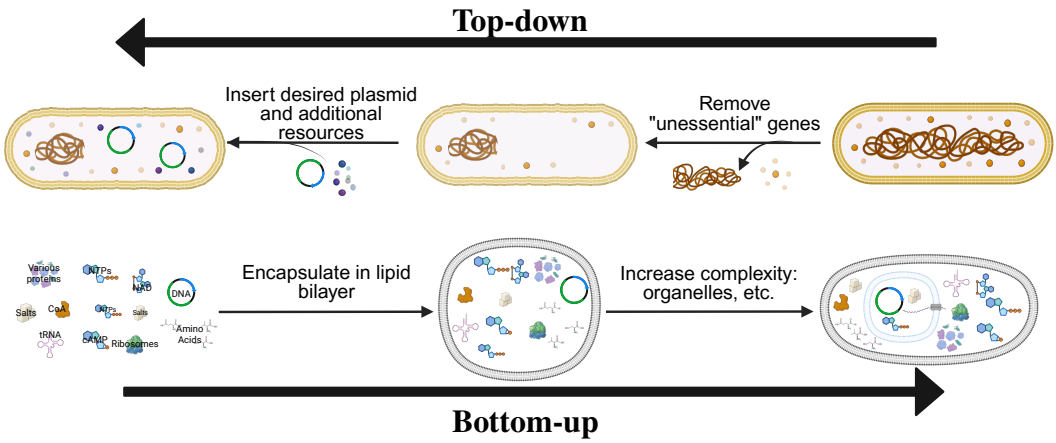

Figure 1.1: **Approaches to building a synthetic cell.** An illustration of the two main approaches to creating and designing a synthetic cell. (**Top**) The "top-down" approach starts with an existing cell and is altered only to contain essential genes. (**Bottom**) The "bottom-up" approach begins with chemicals and purified proteins encapsulated to form a basic cell with the potential to increase complexity by incorporating organelles or transmembrane proteins. Created with `BioRender.com`.

The "top-down" method embraces a proactive design strategy, drawing inspiration from engineering principles found in living systems. However, biological systems present inherent challenges—they are complex, nonlinear, functionally context-dependent, and stochastic, making them challenging to rational engineering [2, 3]. In contrast, the "bottom-up" approach seeks to reconstruct purified biochemical components into synthetic cells to replicate the fundamental aspects of living systems. By doing so, researchers gain insights into the chemistry and biochemistry of cellular processes. This method facilitates the assembly of controllable biochemical properties, often within lipid vesicles or other synthetic compartments, allowing for mechanistic understanding through the construction of characterized components [4].

1.2 Cell-Free Protein Synthesis Systems

The "bottom-up" approach relies on cell-free protein synthesis systems. Cell-free protein synthesis is a biochemical technique that produces proteins *in vitro*. Cell-free protein synthesis offers several advantages over in vivo protein synthesis, including rapid production, the ability to control reaction conditions, and the capacity to synthesize proteins that might be toxic or difficult to express in living cells [5]. It is also valuable for studying fundamental aspects of translation and for applications in biotechnology, such as producing therapeutic proteins, industrial enzymes, or synthetic biology components. Cell-free protein synthesis systems can be derived from various sources, including bacteria, yeast, plants, and mammalian cells, each offering unique advantages and capabilities for protein synthesis. However, cell-free protein synthesis can generally be divided into two broad categories, highlighted in Figure 1.2. Unlike traditional protein synthesis that occurs within living cells, cell-free protein synthesis systems utilize cell lysates or extracts containing the necessary molecular machinery for protein synthesis, including ribosomes, amino acids, tRNAs, initiation, elongation, and termination factors, as well as energy sources (e.g., ATP, GTP).

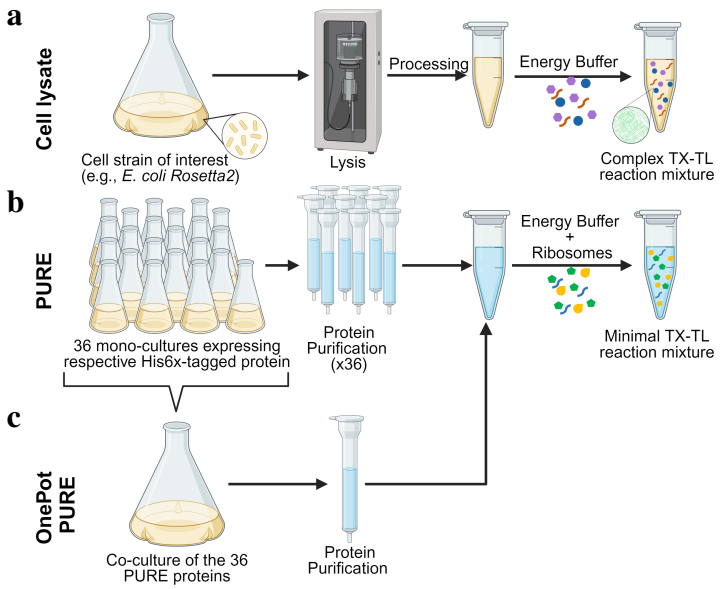

Figure 1.2: **Overview on *E. coli* based cell-free protein expression systems.** (**a**) In cell lysate, necessary molecular machinery for protein synthesis is primarily harvested from *E. coli* cells. (**b**) In PURE, individually purified transcription and translation machinery are combined to rebuild the 'central dogma.' (**c**) OnePot PURE is a variation of PURE, where transcription and translation machinery are co-cultured and purified. Created with `BioRender.com`.

The first widely used cell-free protein synthesis is based on cell lysate, first implemented in the 1960s to express synthetic RNAs to decipher the genetic code [6]. Cell lysate-based systems or TX-TL (transcription and translation) systems utilize cellular machinery harvested from the cell. After multiple stages of growth, lysis, and clarifying spins, an energy buffer with amino acids is added to the lysate to create a protein expression system [7, 8], as shown in Figure 1.2a. The widespread use of TX-TL, now commercially available, has its limitations. The research field is limited by batch-to-batch variability, affecting lifetime and total protein expression [9]. The batch-to-batch variability can result from multiple variables such as cell strain, optical density (OD) at the time of harvest, lysis method, energy mixture composition, and reagents batches. The ability to achieve 'design–build–test' cycles, similar to those found in other traditional engineering fields, is thus ultimately limited by the cell-free protein synthesis system's predictability.

The second category of cell-free protein synthesis utilizes purified components of transcription and translation machinery to reconstruct the "central dogma," as shown in Figure 1.2b. The first reported attempt of using purified proteins to express functional proteins was in 1977 [10], having limited success. Weissbach's group provided a starting point to which Ganoza *et al.* [11] and Pavlov *et al.* [12, 13] attempted to use precharged aminoacyl-tRNAs or partially purified aminoacyl-tRNA synthetase alongside purified proteins. Shortly after, in 2001, Shimizu *et al.* [14] achieved successful protein production using PURE — Protein synthesis Using purified Recombinant Elements. The PURE system contains all transcription and translation proteins required for protein production at known concentrations. The known composition PURE can help circumvent batch-to-batch and inter-laboratory variability problems seen in extract-based systems. A variant of PURE is OnePot PURE, illustrated in Figure 1.2c. In 2019, Lavickova and Maerkl introduced OnePot PURE [15], a method for co-culturing and purifying all 36 proteins together in a single workflow that can be completed within a week. This was further developed by Grasemann *et al.* in 2021 [16]. An advantage of OnePot PURE is that it is theoretically more accessible and reproducible. However, reproducibility and direct comparison to commercial PURE products have been challenging. Additionally, similar to cell-lysate protein concentrations in OnePot PURE are not accurately known unless analyzed by matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) or liquid chromatography-tandem mass spectrometry (LC-MS/MS) [17].

1.3 Engineering of a Synthetic Nucleus

The primary difference between prokaryotic and eukaryotic cells is their lack of compartmentalization, such as a distinct nucleus or other organelles encased in internal membranes. As a result of these compartments, many new functions arose; more notably, with the creation of a nucleus, transcription and translation were separated, allowing for greater control and regulation of the cell's activities. Furthermore, little is known about the spatial organization of RNA in bacterial cells; thus, a systematic approach to studying the compartmentalization of transcription free of cell limitations is highly desirable.

Historically, when *E. coli* based cell-free protein synthesis is encapsulated in a lipid bilayer through various techniques such as the oil-emulsion method, and microfluidics [18, 19] transcription and translation remained coupled. Ideally, the synthetic nucleus would encompass transcription machinery, including components such as NTPs, RNA polymerases, and DNA, all contained within a larger synthetic cell. The translation machinery would surround the synthetic nucleus mimicking the cytoplasm, as depicted in Figure 1.3a.

As cell-free protein synthesis systems and RNA synthesis technologies like PURE and, notably, OnePot PURE continued to advance, isolating transcription from translation has become increasingly feasible. This enables a more practical separation of transcription and translation mechanisms, facilitating their individual and holistic analysis. However, it is crucial to separate transcription from translation while maintaining some interconnectedness as must be made available to the translation machinery, highlighted in Figure 1.3b. Prior works that separated transcription from translations measured the RNA produced in the transcription system. However, in these studies, transcription and translation interactions were not explored. For instance, transcription was tested under various conditions, such as the use of crowding agents at different concentrations, on the production of Fluc mRNA using the Quant-iT RiboGreen RNA reagent over time [20], but different translation conditions were not explored.

The interaction between transcription and translation in a segregated system has yet to be fully explored, either empirically or numerically. The major limitation is the communication between the subsystems primarily through RNA transport or diffusion of other smaller chemicals. Potential RNA delivery mechanisms may draw from several different groups and are illustrated in Figure 1.3c: nuclear lysis, nuclear fusion, and active translocation of RNA out of the nucleus. Work from

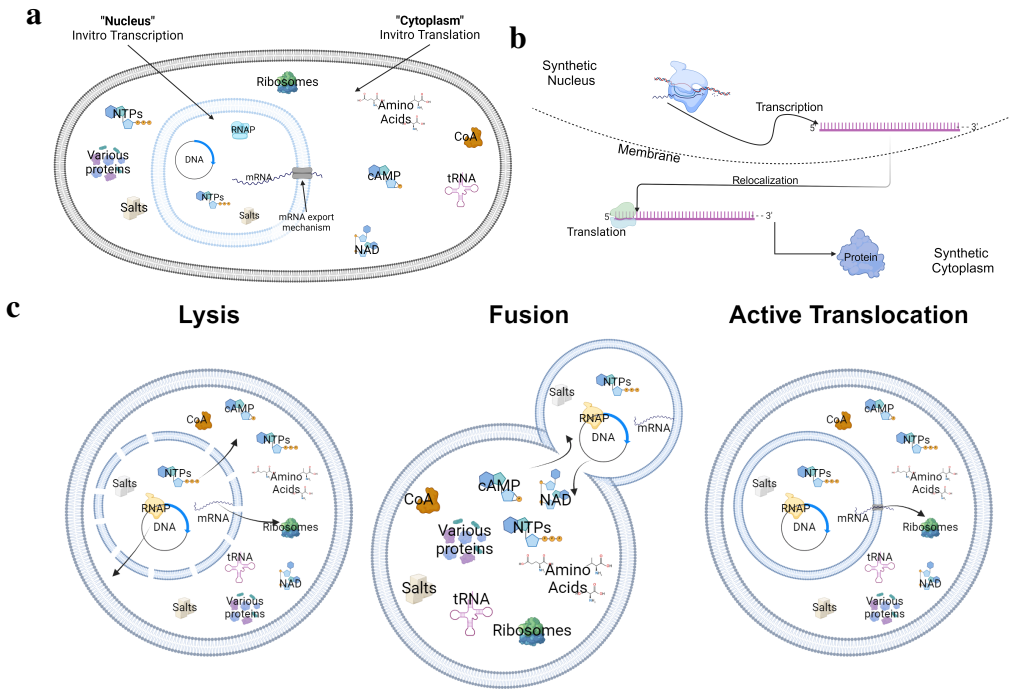

Figure 1.3: **Synthetic nucleus and RNA delivery mechanism.** (**a**) Illustration of a synthetic cell with a nucleus enclosed by a bilipid membrane, only containing components needed to produce an RNA strand. (**b**) Zoomed-in illustration of the membrane barrier separating the nucleus and cytosol. (**c**) Three possible RNA delivery mechanisms from the synthetic nucleus. (1) Lysis of the synthetic nucleus using synthetic lipids. (2) Fusion of two vesicles, one containing the transcription machinery and the second with the translation machinery. (3) Active and unidirectional translation of RNA through the use of cell-penetrating peptide. Created with `BioRender.com`.

the Elani group at Imperial College London demonstrated the creation of a stimuli-responsive vesicle capable of releasing signaling molecules through the utilization of synthetic lipids [21]. Furthermore, in the Kamat group at Northeastern University, genetic material was isolated from the protein expression machinery, and two liposomes containing DNA or the cell-lysate-based cell-free protein synthesis were then combined by fusing the liposome [22], showing the delay of expression. Lastly, Adamala's group at the University of Minnesota showed a promising approach to RNA translocation from a liposome using RNA-binding proteins and cell-penetrating peptides [23].

To delve into the evolutionary path of the nucleus and the ensuing diversification of life, we require a more profound comprehension of the underlying principles governing the construction of our synthetic cell or nucleus—namely, the transcription and

translation systems. This requires establishing dependable and accurate techniques for monitoring transcription, as well as the creation of modeling tools essential for designing and constructing a synthetic eukaryotic cell.

Monitoring Transcription

Transcription is often considered more challenging to measure due to RNA's inherent properties and technical considerations: instability, lower abundance, structural complexity, and calibration challenges. Proper measurement of RNA in cell-free systems is critical for directly reflecting gene expression levels, understanding regulatory mechanisms, quality control of expression, and future optimization of synthetic biology applications such as metabolic engineering, protein synthesis, and biosensor development. While translation can be measured using fluorescent proteins, monitoring transcription relies on RNA aptamers. The malachite green aptamer (MGapt) is an example of an RNA aptamer used to measure RNA production in lysate-based cell-free protein synthesis [24, 25], though not visibly present in literature using PURE-based cell-free protein synthesis.

MGapt was initially developed as an alternative to the chromophore-assisted laser inactivation (CALI) technique [26, 27], used in molecular biology to study the functions of specific genes. RNA aptamers, characterized by their short length and single-stranded nature, are oligonucleotides that selectively bind to specific target molecules. The unique structural attributes of single-stranded oligonucleotides contribute to their high affinity and specificity in recognizing and interacting with their designated targets. MGapt was derived through the systematic evolution of ligands by exponential enrichment (SELEX) [28–30] by Grate and Wilson in 1999. Starting from a random pool of 5×10^{15} RNA molecules, Grate and Wilson isolated and characterized an *in vitro* MG-binding RNA motif exhibiting a high affinity and specificity for binding to the triphenyl-methane dye, malachite green (MG_{dye}) [31].

Malachite green dye is a cationic triphenylmethane dye used since the 1930s as a fungicide, ectoparasiticide, and antimicrobial in aquaculture [32]. Commercially, MG_{dye} is prepared as the chloride or oxalate salt of its dye cation [33]. When dissolved independently, MG_{dye}has almost no fluorescence, with a quantum yield of 7.9×10^{-5}; however, in the presence of MGapt, the fluorescence of MG_{dye} and close relatives increases approximately 2360-fold [34]. The complex formed between MGapt and MG_{dye} relies solely on stacking and electrostatic interactions [35, 36]. In addition to the MGapt adaptive binding to the ligand, the uneven charge distribu-

tion of the MGapt binding pocket is the primary driving force producing structural changes in the ligand in addition to the MGapt adaptive binding to the ligand [37]. The process by which the MG_{dye} molecule undergoes conformational changes during binding to the MGapt results in a red-shift of the absorbance frequency. The strength of the MG_{dye}-MGapt bond has been reported as having a dissociation constant K_d between 100 nM and 800 nM [31, 34, 38, 39] and is contingent on the formation of a coplanar configuration [40].

Modeling of Cell-Free Protein Synthesis Reactions

Models of cell-free protein production are indispensable for advancing our understanding of cellular processes, engineering biological systems, and developing innovative solutions in fields ranging from biotechnology to medicine. In constructing these models, we utilize the chemical reaction network (CRN) formalism to create a detailed mechanistic model, employing a CRN compiler named BioCRNpyler [41]. The BioCRNpyler tool generates models in the Systems Biology Markup Language (SBML) [42], a standard format for biological modeling. These SBML files can be simulated using any compatible SBML simulator. We utilize the Bioscrape Python package [43] for simulation. Bioscrape converts the Chemical Reaction Network (CRN) model to ordinary differential equations (ODEs) and employs Python's `odeint` to solve them based on specified initial conditions. Each reaction rate in the CRN is expressed using mass-action propensity [44] for this conversion. We opt for Bioscrape because it supports sensitivity analysis, Bayesian inference tools, and model simulations. Local sensitivity analysis is conducted for each SBML model to determine the sensitivity of measured species to all parameters over time. Subsequently, we identify the most sensitive parameters using experimental data. Parameter identification is achieved through a Bayesian inference algorithm implemented in Bioscrape, utilizing the emcee Python package [45]. By incorporating experimental data, we obtain probability distributions for each identified parameter through Bayesian inference. Model simulations, using parameter values sampled from these posterior probability distributions, are compared against experimental data to assess model prediction quality. These posterior probability distributions also quantify the uncertainty in the data, highlighting a significant advantage of Bayesian inference methods.

Chapter 2

IMPACT OF CHEMICAL DYNAMICS OF COMMERCIAL PURE SYSTEMS ON MALACHITE GREEN APTAMER FLUORESCENCE

The contents of this chapter are reproduced from

[1] Jurado, Zoila and Murray, Richard M. "Impact of chemical dynamics of commercial PURE systems on malachite green aptamer fluorescence." *ACS Synthetic Biology* (2024, in revision). Available as bioRxiv preprint. DOI: https://doi.org/10.1101/2023.08.14.553301.

2.1 Introduction

Cell-free protein synthesis systems can be broadly categorized into two types. The first and most commonly used cell-free protein synthesis system is based on cell lysate. The cell lysate-based system uses cellular machinery harvested from the cell [7]. The second category contains all the necessary transcription and translation proteins for *E. coli*, each cultured and purified individually and combined at known concentrations, known as PURE — Protein synthesis Using purified Recombinant Elements [14]. A variant of PURE is OnePot PURE, where all 36 proteins are co-cultured and purified together [15]. Using PURE facilitates the modeling of the CPFS system, allowing for the development of a complete model and techniques to seamlessly integrate it into the "design-build-test" pipeline for genetic circuit construction or synthetic cell assembly.

Models in recent years have modeled the translation of peptides for PURE using chemical reaction networks [46, 47], and our previous work has added to these models by expanding the user peptide and the addition of transcription [48]. Validation of all models requires accurate transcription and translation monitoring. Proper measurement of RNA in cell-free systems is critical for directly reflecting gene expression levels, understanding regulatory mechanisms, quality control of expression, and future optimization of synthetic biology applications such as metabolic engineering, protein synthesis, and biosensor development. While translation can be measured using fluorescent proteins, transcription is more challenging, leading to the reliance on RNA aptamers, such as malachite green aptamer.

Malachite green aptamer (MGapt) was initially used to control gene expression in *S. cerevisiae* [49], MGapt has been used as a means to study RNA production, RNA dynamics, and investigating trade-offs between transcription and translation in cell-free protein systems with protein expression of deGFP [25, 50, 51]. Studies using MGapt to measure RNA production in cell-free protein synthesis have been used in lysates [24, 25]. Our use of MGapt in PURExpress reveals surprising dynamics, which may explain why the MGapt measurements appear absent from the literature on PURE cell-free protein synthesis. Generally, the exceptional specificity of aptamers allows the discrimination between closely related isoforms or different conformational states of the same target molecule [52, 53]. However, MGapt has been known to bind to other triphenylmethane dyes such as crystal violet (CV), tetramethylrhodamine (TMR), and Pyronin Y (PY) [34, 39]. Furthermore, it has been found that free MGapt reduces the amount of RNA folded in the correct binding conformation, and metal ions, though not required for high-affinity bind [54], stabilize the complexes with non-native ligands. In contrast, the complex with the original selection target is stable at low salt and without divalent metal ions [38]. The destabilization of MGapt through the use of organic solvent, the addition of Mg^{2+}, DTT, and other ions has been described in the past [33, 39, 55–58].

In this chapter, we demonstrate how the chemical composition of commercial PURE may destabilize MGapt, leading to different aptamer states corresponding to different fluorescence levels. We will first describe the observed MGapt expression in commercially available PURE and inconsistencies between what we could predict and what was observed, such as saturation time and dynamics of MGapt when no transcription is involved. We then discuss the potential effects of DTT in the system and propose a model that uses DTT as a driving force. Finally, we provide experimental validation of the measured MGapt model of the PURE system accounting for DTT's impact on the state of MGapt, allowing for accurate RNA calculations.

2.2 Results and Discussion

Expression and fluorescence of MGapt in different PURE systems. We initially evaluated the consistency of MGapt dynamics across two distinct PURE systems to ascertain whether this phenomenon is inherent to PURE. In commercial PURE systems, the protocol recommends incubating the reaction for 2-4 hours [59, 60], during which protein production saturates. RNA production would also be expected to saturate around the 2-hour mark to be consistent with the reaction lifespan. However, MGapt fluorescence does not saturate at 2 hours. In comparison in PUREx-

press and OnePot PURE, both expressing MGapt from plasmid DNA with construct pT7-MGapt-tT7 (light blue) and pT7-MGapt-UTR1-deGFP-tT7 (dark blue), MGapt fluorescence dynamics were striking, shown in Figure 2.1.

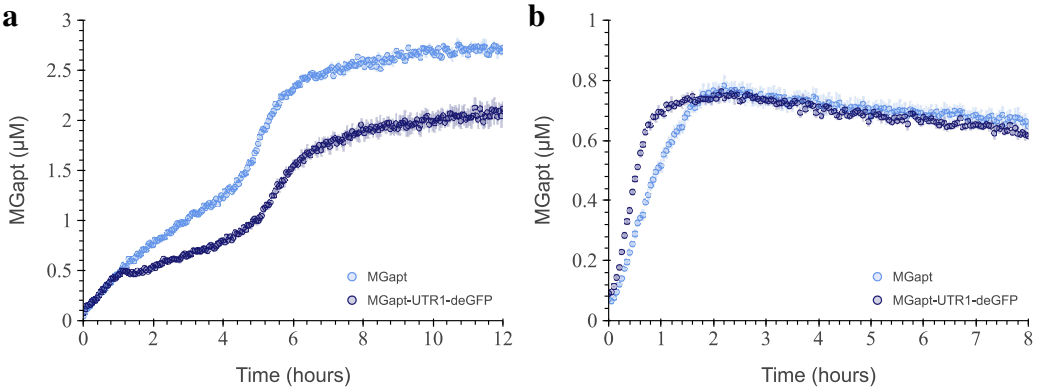

Figure 2.1: **MGapt concentration in PURExpress and OnePot-PURE.** The MGapt concentration over time for transcriptions of pT7-MGapt-tT7 (light blue) and pT7-MGapt-UTR1-deGFP-tT7 (dark blue) at 5 nM, in (**a**) PURExpress and (**b**) OnePot, made in lab. Experimental data consisted of three replicates (circles and respective error bars).

Experimental results indicate that the two PURE systems had different MGapt dynamics, irrespective of the total MGapt production. While each system exhibited a monotonic upward trend of MGapt fluorescence, the PURExpress reaction (Figure 2.1a) showed that MGapt fluorescence saturated at around six hours. On the other hand, the OnePot PURE reaction (Figure 2.1b) showed that MGapt fluorescence saturated at around two hours. The MGapt saturation could indicate that MGapt production continues after two hours in PURExpress, effectively out-living OnePot PURE's RNA production. It did not appear logical for the transcription process to persist beyond the translation. Furthermore, if MGapt fluorescence is considered an accurate reflection of RNA production, it would also suggest that the RNA production rate changes around four hours. The significant disparity in MGapt production dynamics suggested MGapt fluorescence might increase due to chemical differences between PURExpress and OnePot PURE.

Effects of buffering conditions MGapt fluorescence. Upon investigating the difference, we first compare the two systems chemical compositions [14, 16]; shown in Table 2.1. We discovered numerous salts in one system but not the other, which we concluded should not significantly impact MGapt fluorescence. The one major difference in the two systems was the reducing agent used, tris(2-carboxyethyl)phosphine (TCEP) versus dithiothreitol (DTT).

Table 2.1: Reported chemical composition of PURE systems.

Chemical	PURExpress	OnePot PURE
Magnesium acetate	9 mM	11.8 mM
Potassium phosphate	5 mM	-
Potassium glutamate	95 mM	100 mM
Ammonium chloride	5 mM	-
Calcium chloride	0.5 mM	-
Spermidine	1 mM	2 mM
Creatine phosphate	10 mM	20 mM
Putrescine	8 mM	-
Dithiothreitol (DTT)	1 mM	-
Tris(2-carboxyethyl)phosphine (TCEP)	-	1 mM

PURExpress uses 1 mM of DTT [61] while OnePot PURE uses 1 mM of TCEP [15]. DTT and TCEP both reduce disulfide bonds, but TCEP has the advantages of being significantly more stable in the absence of a metal chelator and less inhibitive in labeling with maleimide [62]. Additionally, compared to DTT, TCEP is more effective, non-volatile, and does not readily oxidize above pH 7.5 [63] the pH at which the PURE reaction occurs, as shown in Figure 2.2. In Figure 2.2a, we can see that regardless of the expression condition, the PURE reaction's pH for a 12 hour read fluctuates between pH of 7.75 and 8.0. The pH of the PURE reaction was measured using SNARF^{TM}-5F detailed in Section 2.4 (see Figure 2.14) and does not negatively affect protein expression as illustrated in Figure 2.2b.

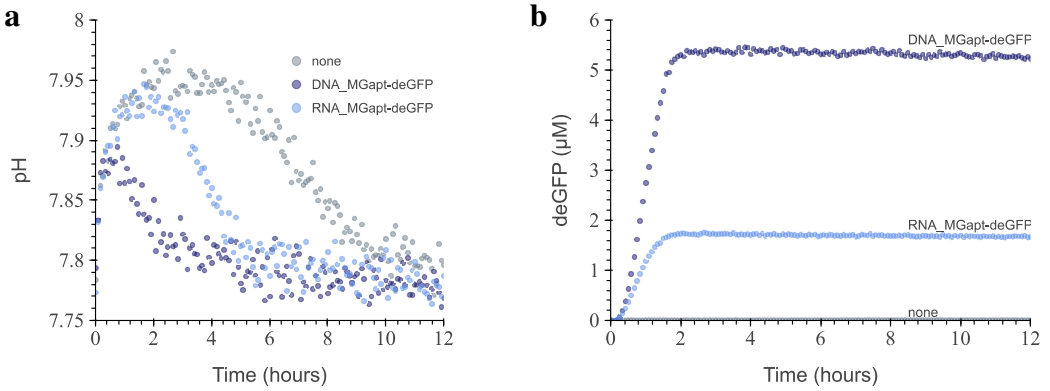

Figure 2.2: **Measurement of pH of the PURExpress over time. (a)** The pH of PURExpress under three different expression conditions: none-without DNA or RNA (gray circles), DNA plasmid of pT7-MGapt-UTR1-deGFP-tT7 (light blue circles), and RNA of pT7-MGapt-UTR1-deGFP (dark blue circles). **(b)** Production of deGFP over pH assay.

MGapt fluorescent was measured under four different chemical conditions to measure the effect of the reducing agent on MGapt fluorescence in separate *in vitro* reactions. Three different reducing agents were tested: TCEP, glutathione (GSH), and DTT), and the waste product pyrophosphate (PPi). The experiment consisted of three replicates of 10 µL reactions with the respective additives (TCEP, GSH, DTT, and PPi) at experimentally relevant concentrations and purified RNA of MGapt-UTR1-deGFP at 0.51 µM. MGapt fluorescence was read in BioTek plate reader (610/650) for six hours at 37 °C and calibrated to µM using the calibration curve shown in Figure 2.13. The full-time course measurements are depicted in Figure 2.3.

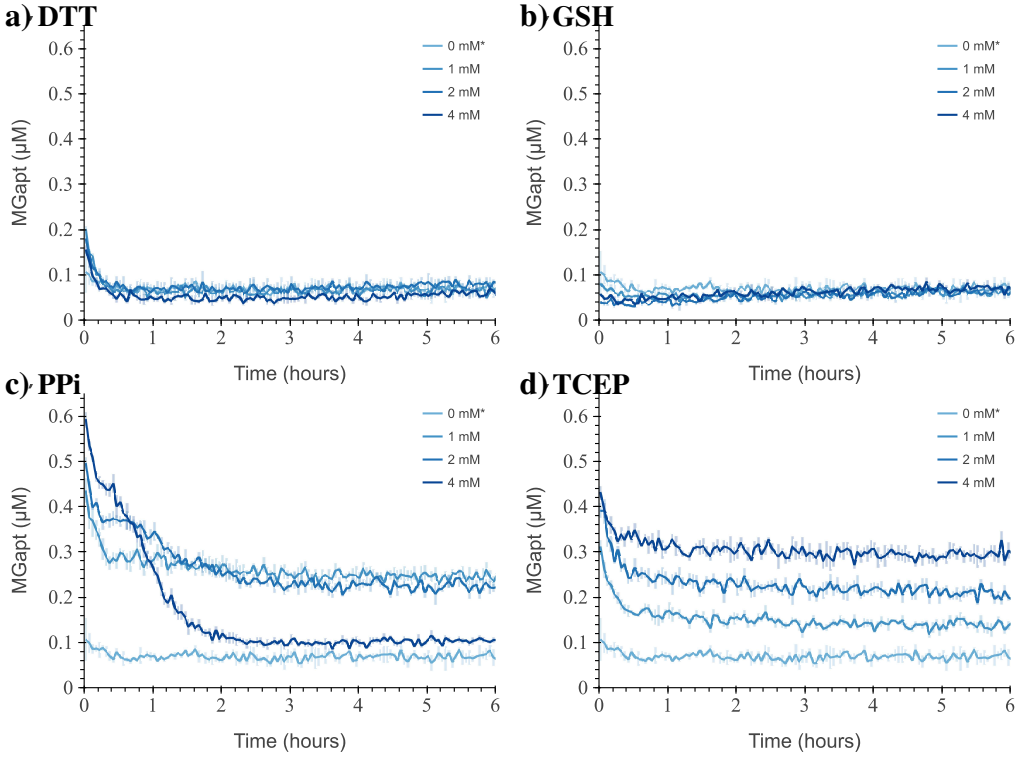

Figure 2.3: **MGapt measurement in different buffers.** Measurements of MGapt concentration of purified RNA for MGapt-UTR1-deGFP at 0.51 µM in different reducing agents and waste chemicals over four hours. Each subplot is titled with the respective buffer used: (**a**) dithiothreitol (DTT), (**b**) glutathione (GSH), (**c**) pyrophosphate (PPi) and (**d**) tris(2-carboxyethyl)phosphine (TCEP). The chemical concentrations are in different shades of blue, respectively. The plot shows the average of the three replicated with error bars; negative control without MGapt was subtracted.

The results presented in Figure 2.3 indicate that MGapt fluorescence is stable after the first 30 min when MGapt is added to a buffer solution containing all tested reducing agents. The MGapt concentration was averaged over six hours to compare differences under various conditions, as shown in Figure 2.4.

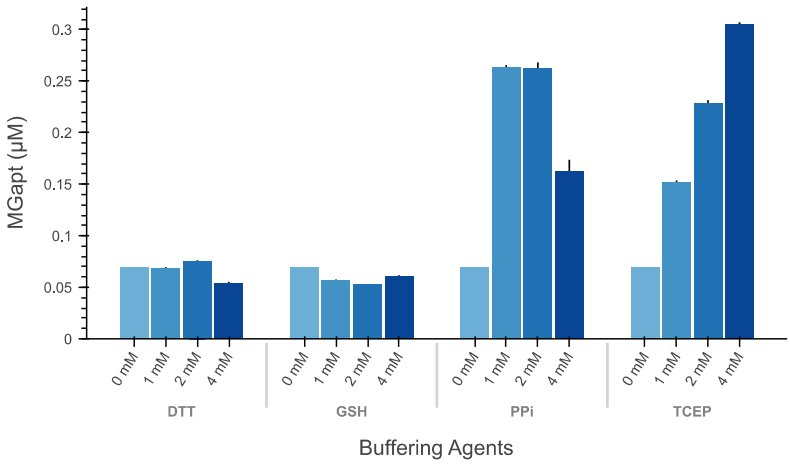

Figure 2.4: **Measured MGapt concentration under different buffer conditions.** Average measured MGapt concentration of purified RNA coding for MGapt-UTR1-deGFP in different reducing agents and waste chemicals over six hours. Each section of the bar plot is titled with the respective buffer used, and the increasing chemical concentrations are designated with increasing intensity of blue. The negative control was subtracted from the test conditions, and standard deviations of triplicates are represented by black bars.

Only TCEP and PPi samples showed concentration effects on the total MGapt fluorescence. Higher concentrations of TCEP and PPi resulted in higher MGapt fluorescence, except at 4 mM PPi, where MGapt fluorescence decreased. The results from Figure 2.4 did not provide conclusive evidence that the reducing agent, DTT, directly affected MGapt fluorescence nor that other reducing agents (TCEP or GSH) would be a better alternative. However, these standard titration experiments involving buffers and MGapt do not capture the chemical dynamics in the PURE reaction system.

The fluorescence of MGapt is affected by the chemical properties of the PURE system. While Figure 2.4 implies that the fluctuating concentration of PPi might lead to increased MGapt fluorescence, it is important to note that waste production would be directly correlated with system expression. Therefore, we would not anticipate a greater or quicker production of PPi in the PURExpress than in OnePot PURE. To avoid any possible interference of transcription products with MGapt, purified RNA of MGapt-UTR1-deGFP was added, removing the largest PPi production source, RNA strand elongation. Additionally, to isolate the effect of the PURE reaction on MGapt fluorescence, the translation reactions were minimized by using RNA containing only the MGapt. We added 0.49 μM of MGapt-deGFP-RNA and 0.53 μM

of MGapt-RNA independently into a PURExpress reaction and measured MGapt fluorescence over 12 hours. In Figure 2.5a, upon first observation in both cases with 0.49 µM of MGapt-deGFP-RNA (crosses) or 0.53 µM of MGapt-RNA (circles), the measured MGapt concentration is not constant and that the RNA only containing MGapt measures lower than the RNA of MGapt-UTR1-deGFP. However, when we normalize each result by its respective maximum, Figure 2.5b, the concentration of MGapt appears to increase gradually over time despite the absence of transcription. Interestingly, the environmental effect seemed similar for both the TL-null and TL-only conditions, indicating that the additional ribosome binding site and deGFP sequences could not solely affect MGapt fluorescence.

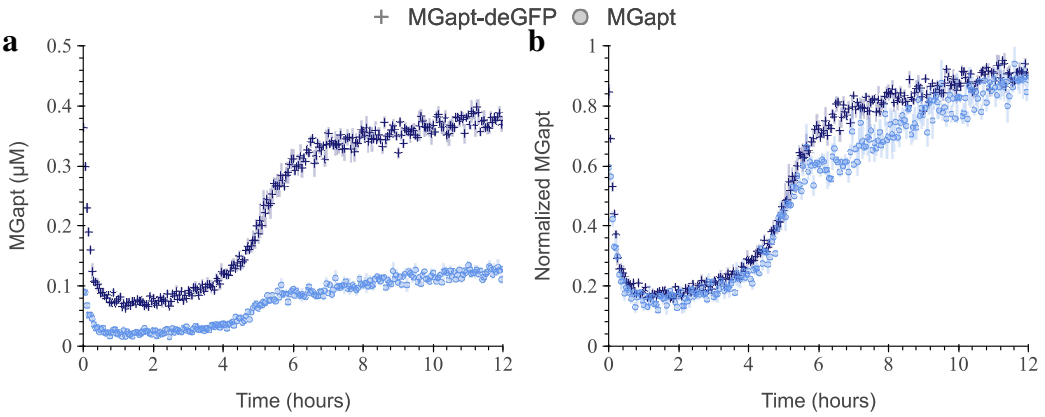

Figure 2.5: **Measured MGapt concentration dynamics in PURExpress starting from RNA.** PURExpress reactions of 10 µL with three technical replicates containing 10 µM of Malachite Green dye and 8 units of RNAse inhibitor. (**a**) Starting from purified RNA of MGapt at 0.53 µM (light blue crosses) and MGapt-UTR1-deGFP at 0.49 µM (dark blue circles) and after data is normalized (**b**).

To further demonstrate that the measured MGapt concentration over time was not dependent on the production of RNA or protein, additional PURE reactions were performed at different DNA concentrations, Figure 2.6a. The experiment consisted of three replicates of 10 µL with DNA plasmid pT7-MGapt-UTR1-deGFP-tT7 at six different DNA concentrations dispensed using an Echo 525 Acoustic Liquid Handler and the final DNA concentration was calculated on the rounded amount of DNA that was dispensed. To account for total RNA production, the measured MGapt concentrations were normalized in Figure 2.6b. The initial RNA production rate varied based on DNA concentrations, but all normalized MGapt concentration measurements converged. Notably, all concavity transitions occurred around the 4-hour mark regardless of RNA production rate, indicating that MGapt dynamics are unaffected by DNA concentration. The increase in MGapt concentration over time

further supports that MGapt fluorescence is affected by the chemical environment of the commercial PURE system, independent of transcription and translation.

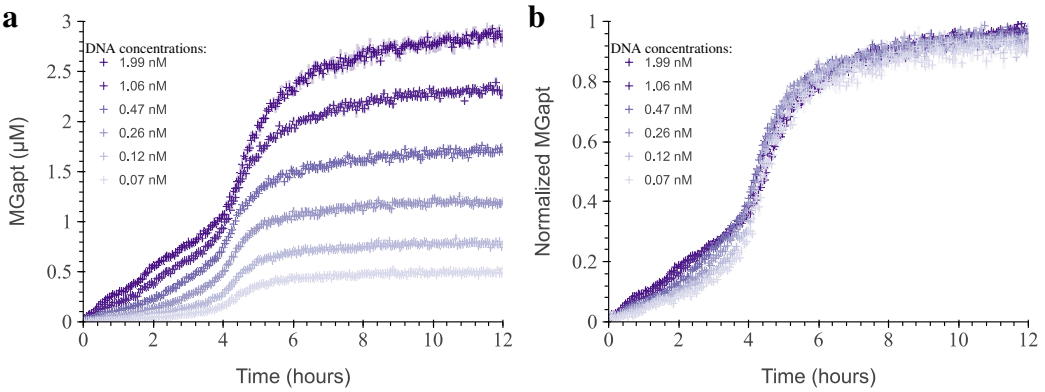

Figure 2.6: **Measured MGapt concentration dynamics in PURExpress at various DNA concentrations.** PURExpress reactions of 10 µL with three technical replicates containing 10 µM of Malachite Green dye and 8 units of RNAse inhibitor. (**a**) Expressing plasmid pT7-MGapt-UTR1-deGFP-tT7 at given DNA concentrations and after data is normalized (**b**).

The observed inflection point around 4 hours, in Figure 2.5 and Figure 2.6, suggests the presence of system repression by specific chemical substrates that degrade over time. As illustrated in a previously published model for PURE translation [46], there are 483 reactions with a nonzero rate, and 278 reactions occur without the presence of DNA: tRNA charging, NTP degradation, and energy recycling. These reactions also occur in OnePot PURE, indicating that this MGapt response would not be exclusive to one system. This leads back to the distinctions between OnePot PURE and commercial PURE, particularly their selection of reducing agents—DTT versus TCEP. While the reducing agent may not directly impact MGapt fluorescence, it could potentially influence other chemical reactions that affect MGapt fluorescence.

To test the effects of DTT on measured MGapt concentration, DTT was added to PURExpress expressing DNA plasmids of pT7-MGapt-UTR1-deGFP-tT7 and pT7-MGapt-tT7 at 5 nM (see Figure 2.7a). DTT was added to reach a final added DTT concentration of 1 mM, 4 mM,and 9 mM. Following data normalization, Figure 2.7b showed that increasing DTT concentrations shifted the inflection point of measured MGapt to the right. Additionally, we observed that the measured MGapt concentrations for both plasmids overlap, indicating that MGapt effects are primarily due to the addition of DTT.

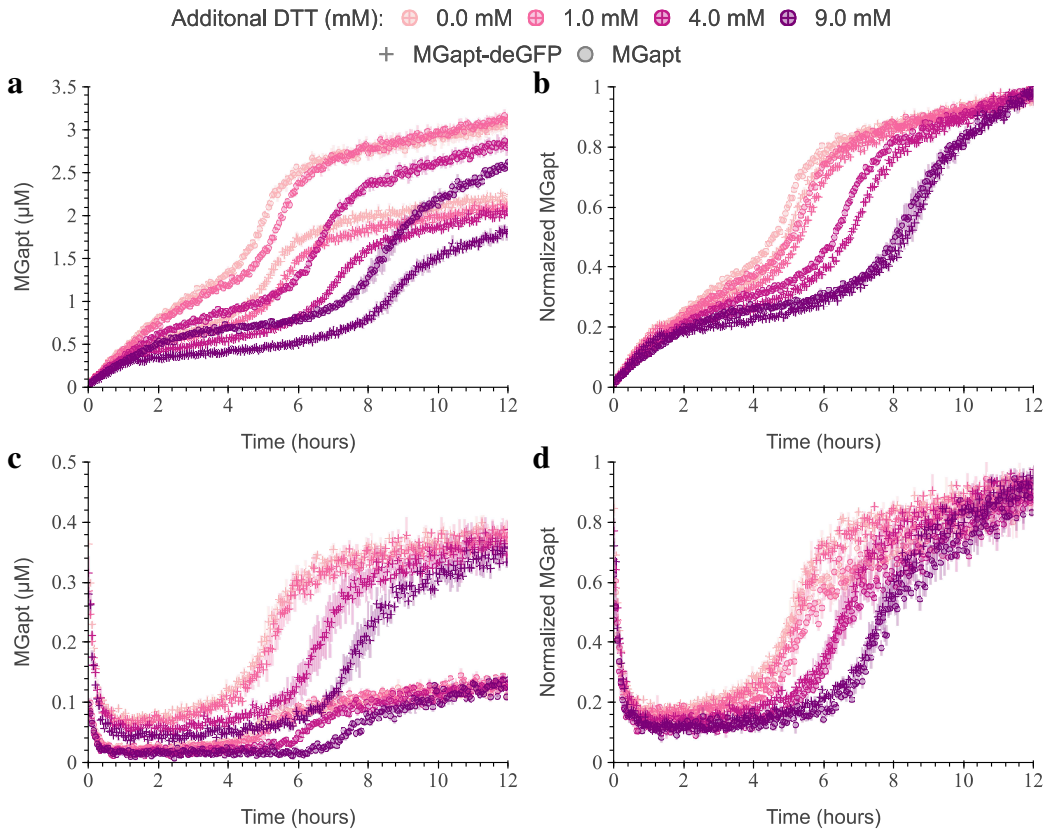

Figure 2.7: **Measured MGapt concentration dynamics in PURExpress under different DTT concentrations.** PURExpress reactions of 10 µL with three technical replicates containing 10 µM of Malachite Green dye and 8 units of RNAse inhibitor. (**a**) Expression of plasmid pT7-MGapt-UTR1-deGFP-tT7 at 4.93 nM and pT7-MGapt-tT7 at 5.03 nM with increasing DTT concentration and after data is normalized (**b**). (**c**) Starting from purified RNA of MGapt-UTR1-deGFP at 0.49 µM and MGapt at 0.53 µM with increasing DTT concentration and after data is normalized (**d**).

Finally, the previous experiment was repeated with purified RNA to underscore the concept that DTT has a direct impact and can effectively represent other auxiliary reactions and chemicals affecting MGapt. In this iteration, purified RNA of MGapt-UTR1-deGFP and MGapt at approximately 0.5 µM concentration was combined at the stated concentrations of DTT (see Figure 2.7c). In the normalized data, Figure 2.7d, data collapse based on added DTT concentration. Consistent with previous findings, the additional DTT seemed to inhibit MGapt fluoresces as the DTT concentration increased. Based on these experimental results, we conclude that DTT acts as a direct or indirect suppressor of MGapt fluorescence, which degrades over time due to its volatile nature. This signifies that using MGapt measurements as an indicator for RNA is inaccurate as it does not account for DTT's effect.

MGapt transitions different aptamer states with varying levels of fluorescence. We assert that MGapt in the commercial PURE reaction goes through different equilibrium and intermediates states, as MGapt has been previously reported [39, 55, 58]. The concentration of DTT plays a crucial role in meditating, promoting, and/or facilitating the instability of MGapt and the conversion between the two chemical states, as demonstrated in Figure 2.7. Using DTT as a proxy for the full chemical reactions responsible for altering MGapt states and thus fluorescence in the commercial PURE reaction, we propose the model depicting MGapt inhibition by DTT, as illustrated in Figure 2.8.

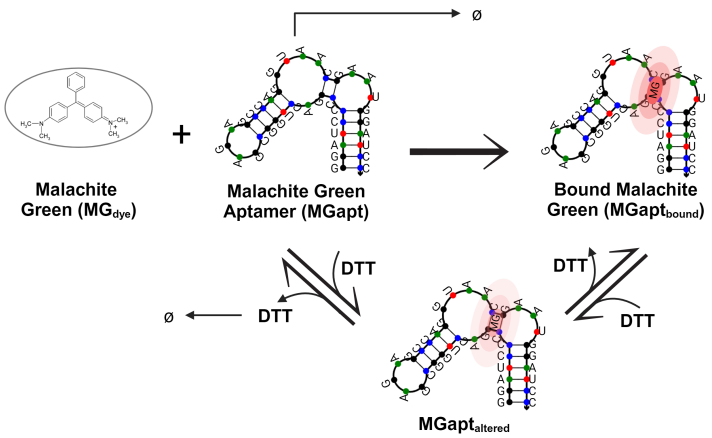

Figure 2.8: **Inhibition of malachite green aptamer by DTT.** Illustration of proposed interaction of DTT on bound fluorescent malachite green aptamer (MGapt). Adapted from chemical structure Stead *et al.* [55] and malachite green aptamer NuPack predicted secondary structure [64]. Created with `BioRender.com`.

To model the detailed step-by-step inhibition and conversion of MG_{dye} bounded to MGapt ($MGapt_{bound}$) to a less fluorescent $MGapt_{altered}$, we built a chemical reaction network (CRN) in a CRN compiler tool called BioCRNpyler [41]. We utilize the functionalities provided by BioCRNpyler model the reactions shown in Figure 2.8:

$$\text{MGapt} + \text{MG}_{\text{dye}} \xrightarrow{k_{\text{MGapt}}} \text{MGapt}_{\text{bound}} \tag{2.1}$$

$$\text{MGapt}_{\text{bound}} + \text{DTT} \underset{ku_{\text{DTT}}}{\overset{kb_{\text{DTT}}}{\rightleftharpoons}} \text{MGapt}_{\text{altered}} \tag{2.2}$$

$$\text{MGapt}_{\text{altered}} \underset{kb_{\text{MGapt}}}{\overset{ku_{\text{MGapt}}}{\rightleftharpoons}} \text{MGapt} + \text{DTT} + \text{MG}_{\text{dye}} \tag{2.3}$$

$$\text{MGapt} \xrightarrow{k_{\text{RNA}_{\text{deg}}}} \emptyset \tag{2.4}$$

$$\text{DTT} \xrightarrow{k_{\text{DTT}_{\text{deg}}}} \emptyset. \tag{2.5}$$

Beginning with transcribed and folded malachite green aptamer state (MGapt), malachite green oxalate (MG_{dye}) binds to MGapt to form fully fluorescent malachite green aptamer ($\text{MGapt}_{\text{bound}}$). Concentrations of DTT induce instability in $\text{MGapt}_{\text{bound}}$ through a reversible process, leading to an intermediate state with reduced fluorescence ($\text{MGapt}_{\text{altered}}$). The $\text{MGapt}_{\text{altered}}$ state can be further destabilized, causing dissociation into its three components: MG_{dye}, MGapt, and DTT. Finally, unbound states such as MGapt and DTT are susceptible to degradation. The introduction of the altered state, $\text{MGapt}_{\text{altered}}$, which exhibits lower fluorescence compared to $\text{MGapt}_{\text{bound}}$, implies that the total MGapt fluorescence is a linear combination of both states, $\text{MGapt}_{\text{measured}} = \text{MGapt}_{\text{bound}} + c_1\text{MGapt}_{\text{altered}}$. The $\text{MGapt}_{\text{measured}}$ represents the measured fluorescence detected by the BioTek plate reader or any other fluorescent reader. Consequently, the measured MGapt concentration in PURExpress is a misrepresentation of RNA production without considering or adjusting for additional MGapt states that may exhibit different fluorescence properties due to DTT interactions.

Analysis and identification of parameters in the proposed model. To analyze the model of the effects of DTT on MGapt measured, we measured the MGapt fluorescence in a 10 µL reaction done in triplicate using PURExpress. The reaction contained 0.22 µM of RNA of MGapt-UTR1-deGFP. To determine the coefficients of the linear combination, we utilized the lowest concentration measurement to indicate when 100 % of the aptamer was in the $\text{MGapt}_{\text{altered}}$ state. The coefficient c_1 was then determined based on the percentage of $\text{MGapt}_{\text{measured}}$ and total added RNA ($\text{MGapt}_{\text{total}}$), resulting in the final linear system:

$$\text{MGapt}_{\text{measured}} = \text{MGapt}_{\text{bound}} + 0.15\,\text{MGapt}_{\text{altered}} \tag{2.6}$$

$$\text{MGapt}_{\text{total}} = \text{MGapt}_{\text{bound}} + \text{MGapt}_{\text{altered}}. \tag{2.7}$$

To parameterize the CRN model constructed using BioCRNpyler, we initialized the parameters by hand-tuning the reactions for bounded conditions by steps. Due to the time required to prepare the reactions before measuring on BioTek, it was necessary to determine the initial conditions at the time of reading. Employing the system of equations, we computed the initial conditions based on the $\text{MGapt}_{\text{measured,t=0}}$ reading.

The model's initial conditions parameters depend on the total amount of RNA the user adds, where $\sum$ MGapt equals the total RNA containing MGapt. The initial conditions account for the formation of $\text{MGapt}_{\text{bound}}$ and $\text{MGapt}_{\text{altered}}$ before starting the BioTek read. Our calculations revealed that approximately 45 % and 55 % of the $\sum$ MGapt are in the two different states of $\text{MGapt}_{\text{bound}}$ and $\text{MGapt}_{\text{altered}}$, respectively. Therefore, $\text{MGapt}_{\text{t=0}}$ is zero and $\text{MGapt}_{\text{bound,t=0}} = 0.45 \sum \text{MGapt}$, $\text{MGapt}_{\text{altered,t=0}} = 0.55 \sum \text{MGapt}$, and $\text{MG}_{\text{dye,t=0}} = \text{MG}_{\text{dye,total}} - \sum \text{MGapt}$. The specific values of the initial conditions utilized for the training model with $\sum$ MGapt = 0.22 µM are outlined in Table 2.2.

Table 2.2: MGapt-DTT model initial conditions of training model.

Species	Value	Unit
MGapt	0	µM
$\text{MGapt}_{\text{bound}}$	0.099	µM
$\text{MGapt}_{\text{altered}}$	0.121	µM
DTT	1000	µM
MG_{dye}	9.78	µM

For accurate predictions of the measured MGapt concentration in PURE from the BioTek readings, we conducted model training using experimental data from the measured MGapt concentration of 0.22 µM RNA with the MGapt-UTR1-deGFP construct. Training this model using parameter identification was challenging due to the combined fluorescence of two distinct states in the measured data. With the ability to train only on one species, the effective concentration of $\text{MGapt}_{\text{bound}}$ was calculated using the set of linear equations. Ultimately, in addressing the inherent noise in the experimental data, we employ Bayesian inference to derive a distribution of potential parameter values based on the experimental data. To accomplish both tasks of evaluating identifiability and determining posterior parameter distributions, we utilize a biological data analysis pipeline [65] implemented with the Python package Bioscrape [43].

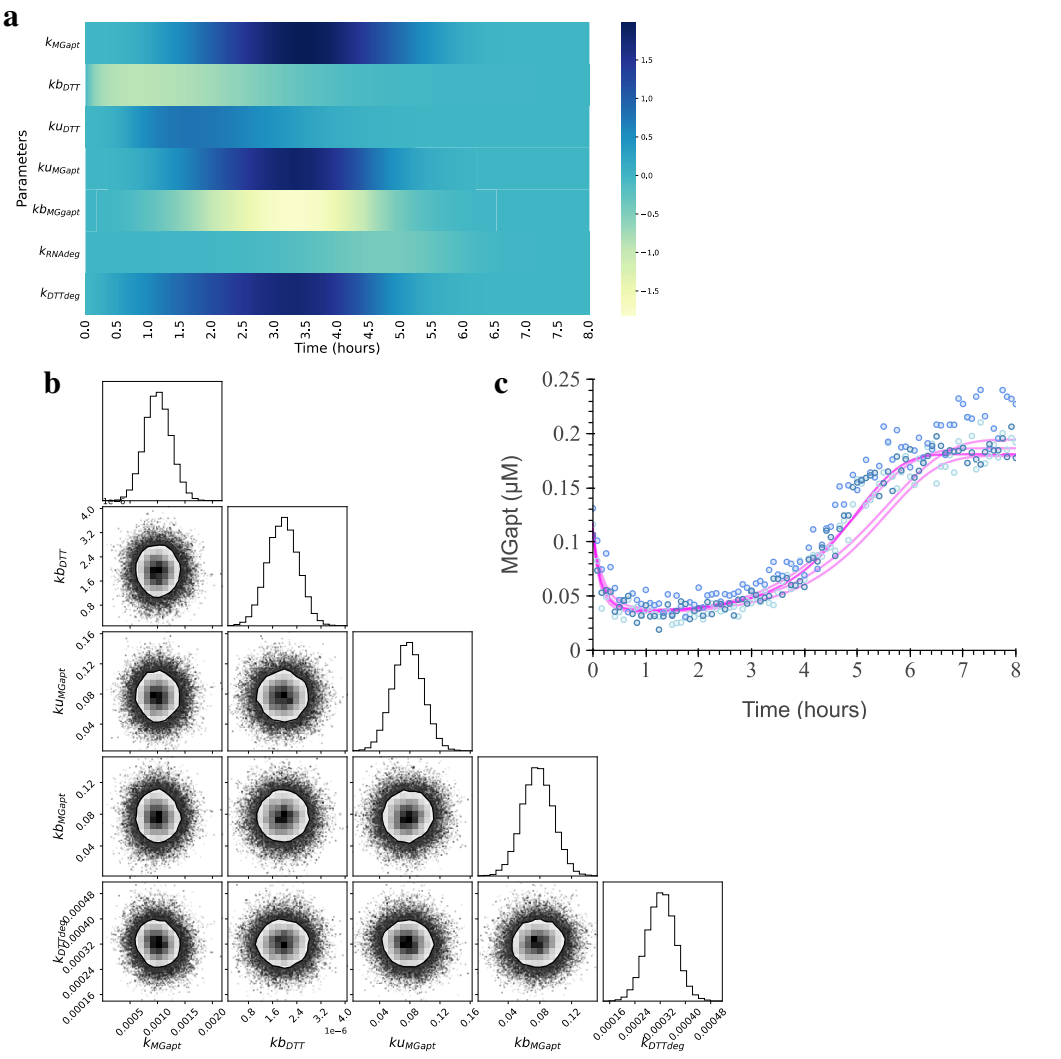

Figure 2.9: **Modeling, analysis, and parameter inferencing of DTT effects on MGapt model.** (**a**) The sensitivity of the MGapt to the CRN model parameters for the initial four hours. (**b**) The posterior distributions of parameters obtained after running Bayesian inference on k_{MGapt}, kb_{DTT}, ku_{MGapt}, kb_{MGapt}, and $k_{\mathrm{DTT_{deg}}}$. The corner plot depicts the covariance of the two parameters, with the contour showing the 75 % probability region for the parameter values. (**c**) With parameter values for k_{MGapt}, kb_{DTT}, ku_{MGapt}, kb_{MGapt}, and $k_{\mathrm{DTT_{deg}}}$ sampled from the posterior distributions, the five model simulations (magenta lines) are shown alongside the experimental data for three biological replicates (scattered blue points). Code for all data analysis, parameter inference, and related data are available on Github [66].

Parameters for identification were determined by conducting a local sensitivity analysis of all species in the model across all parameters and time. The sensitivity analysis heatmap for DTT's interaction with RNA of MGapt-UTR1-deGFP model is depicted in Figure 2.9a. We selected parameters with the highest sensitivity regarding MGapt output fluorescence. Among the seven reaction rates in equations

(2.1)-(2.5), we identify k_{MGapt}, kb_{DTT}, ku_{MGapt}, kb_{MGapt}, and $k_{\text{DTT}_{\text{deg}}}$ as the most sensitive parameters. These parameters correspond to the initiation of transcription and the formation rate of the RNAP-bound GDP and phosphate complex, respectively. Subsequently, we utilize Bayesian inference tools in Bioscrape to ascertain the posterior parameter distributions for these five parameters. The corner plot in Figure 2.9b illustrates the posterior parameter distributions and their covariance, providing a sampling distribution for predicting output using the fitted model. Model simulations using parameter values drawn from the posterior, alongside experimental data, are presented in Figure 2.9c. The final parameter values utilized are detailed in Table 2.3.

Table 2.3: MGapt-DTT model parameters values.

Parameter	Description	Value	Unit
k_{MGapt}	Formation of $\text{MGapt}_{\text{bound}}$ though the binding of MG_{dye} and MGapt	9.5×10^{-4}	$\mu\text{M}^{-1}\,\text{s}^{-1}$
$k_{\text{b}_{\text{DTT}}}$	Binding of DTT to $\text{MGapt}_{\text{bound}}$ to form $\text{MGapt}_{\text{altered}}$	2.0×10^{-6}	$\mu\text{M}^{-1}\,\text{s}^{-1}$
$k_{\text{u}_{\text{DTT}}}$	Unbinding of DTT from $\text{MGapt}_{\text{altered}}$	8.0×10^{-6}	s^{-1}
$k_{\text{u}_{\text{MGapt}}}$	Unbinding of DTT and MG_{dye} from $\text{MGapt}_{\text{altered}}$	0.0675	s^{-1}
$k_{\text{b}_{\text{MGapt}}}$	Rebinding of DTT and MG_{dye} and MGapt to reform $\text{MGapt}_{\text{altered}}$	0.0762	$\mu\text{M}^{-2}\,\text{s}^{-1}$
$k_{\text{RNA}_{\text{deg}}}$	Degradation of MGapt	3.2×10^{-4}	s^{-1}
$k_{\text{DTT}_{\text{deg}}}$	Degradation of DTT	3.2×10^{-4}	s^{-1}

The model's parameter value depends only on the amount of RNA added and the DTT concentration. Other auxiliary reactions, which may play a role in DTT interaction with MGapt, were not incorporated.

Validation and assessment of model capturing DTT's impact on MGapt fluorescence. To validate the model of the effects of DTT on MGapt measured, we conducted simultaneous tests using multiple RNA concentrations. The RNA of MGapt-UTR1-deGFP at final concentrations of 0.41 µM, 0.86 µM, 1.26 µM and 1.67 µM were read for 12 hours at 37 °C in a BioTek reader at 610/650 (ex/em). The initial concentrations of MGapt and $\text{MGapt}_{\text{altered}}$ at $t = 0$ were determined as previously outlined and are listed in Table 2.4. The initial conditions for MGapt and DTT are not affected by the change of RNA added by the user.

Table 2.4: MGapt-DTT model initial conditions for validation test.

$\sum$ MGapt (µM)	$MGapt_{bound}$ (µM)	$MGapt_{altered}$ (µM)	MG_{dye} (µM)
0.41	0.1845	0.2255	9.59
0.86	0.387	0.473	9.14
1.26	0.567	0.693	8.74
1.67	0.7515	0.9185	8.33

Finally, utilizing equation (2.6), the $MGapt_{measured}$ was computed and superimposed onto the experimental results using their respective colors shown in Figure 2.10. As seen in Figure 2.10a, the predicted $MGapt_{measured}$ concentration remains reasonably accurate throughout the measured time for concentrations below 1 µM. The increased error above below 1 µM of RNA is most likely due to changes in DTT degradation or increased stability of the $MGapt_{altered}$ state.

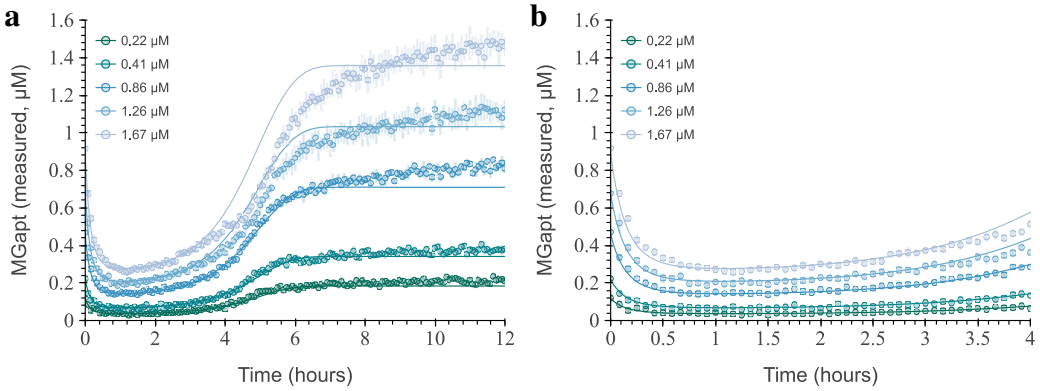

Figure 2.10: **Modeled $MGapt_{measured}$ at different initial RNA concentrations.** The modeled $MGapt_{measured}$ of BioTek at different RNA concentrations (solid line) overlaying with experimental data, three replicates (circles and error bars, of respective colors). (**a**) Full measured time of 12 hours. (**b**) Zoomed, reaction relevant time of 4 hours.

Regardless, the model does recapitulate the increase of the measured MGapt concentration and its overall dynamics. Furthermore, as indicated in the manuals of commercial PURE systems, the reaction is generally monitored for a 2 hour- 4 hour rather than 12 hour. Upon closer examination of the pertinent 4 hour reaction window, highlighted in Figure 2.10b, it becomes evident that the model accurately predicts the measured $MGapt_{measured}$. Moreover, Figure 2.11 shows that the model consistently predicts the measured MGapt within an average of 10 % margin of error within the initial 4 hour period for RNA concentration is greater than 1 µM, longer than the recommended reading time.

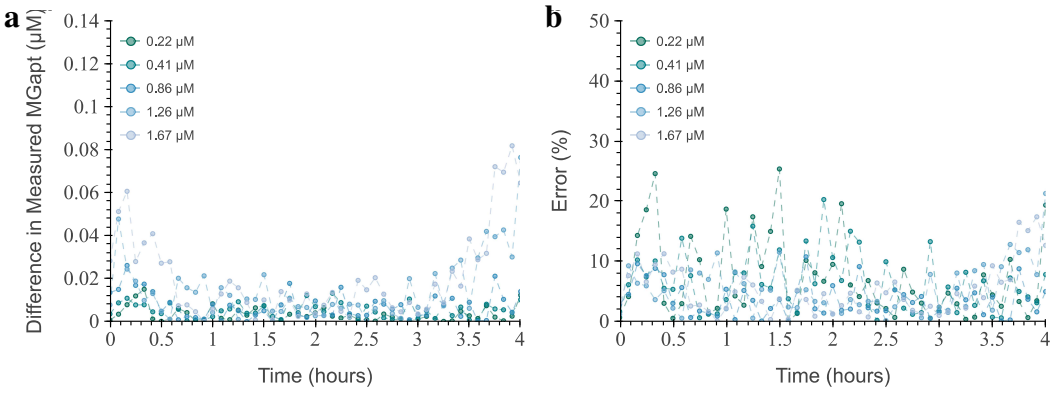

Figure 2.11: **Error of proposed BioCRNpyler model.** (**a**) Absolute error between the BioCRNpyler model and experimental data for different initial RNA concentrations. (**b**) Percent error between the BioCRNpyler model and experimental data for different initial RNA concentrations.The model compared to the mean of the experimental results for the respective RNA concentrations.

The accuracy and the capturing of the dynamics of measured MGapt fluorescence demonstrate that DTT can serve as a representative model for MGapt fluorescence, commonly employed in measuring RNA production. Moreover, the simulation can be used in reverse to calculate the total RNA in a system by calculating the total RNA ratio between measured and total to calculate back the amount of RNA; steps are illustrated in Figure 2.12.

To back-calculate the added RNA concentration from known RNA concentration measurements, we first start off with Figure 2.12a, the simulated MGapt fluorescence of the sample of 0.22 μM. Next, in Figure 2.12b, we calculated the ratio of MGapt between the sample concentration of 0.22 μM to the simulated measured concentrations to get the ratio of MGapt. By multiplying the dynamic calibration curve in Figure 2.12b and the measured MGapt fluorescence of RNA construct MGapt-UTR1-deGFP (see Figure 2.12c), we can approximately calculate the starting of 0.41 μM, 0.86 μM, 1.26 μM,and 1.67 μM of the added RNA. The final results of the back-calculation are shown in Figure 2.12d, where the calculated MGapt concentrations are overlayed with the true RNA concentration, displaying strong alignment and supporting the dual purpose of the MGapt and DTT models proposed.

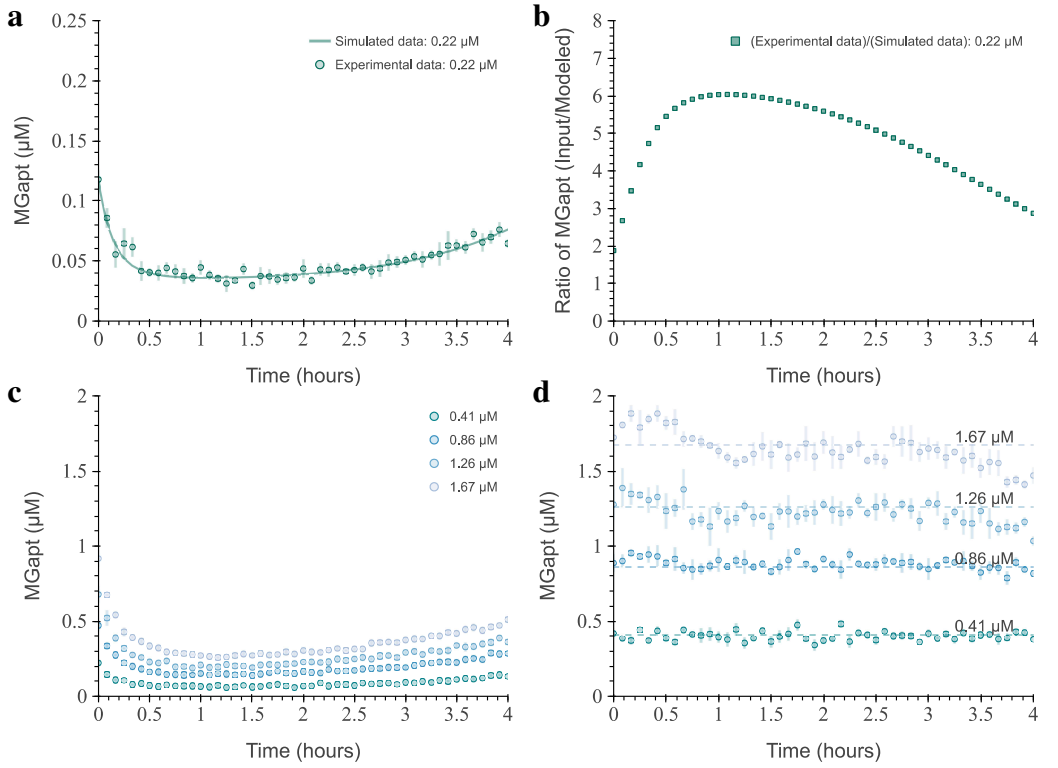

Figure 2.12: **Back calculation of MGapt concentration using proposed model.** (**a**) Measured MGapt concentration of 0.22 µM of RNA for MGapt-UTR1-deGFP (circles with error bars) and simulated MGapt measurements (solid line). (**b**) Proposed dynamic calibration curve for MGapt concentration measurements from data shown in (a). (**c**) Measured MGapt of RNA, MGapt-UTR1-deGFP, at concentrations: 0.41 µM, 0.86 µM, 1.26 µM,and 1.67 µM (circles with error bars, at respective color). (**d**) Calibrated experimental data from (c) with dynamic calibration curve in (b) (circles with error bars, at respective color), overlayed with the RNA concentration used (dashed lines).

2.3 Conclusion

Our experiments reveal that the chemical makeup of commercial PURE systems influences measured MGapt concentration. Specifically, we observe a correlation between the chemical properties of DTT, given that DTT is absent in OnePot PURE but present in other commercial PURE systems. Though DTT and TCEP are effective reducing agents commonly used in biochemical research, DTT is more sensitive to oxidation, less stable, and more violative. We hypothesize that the concentration of DTT suppresses the fluorescence of MGapt by reversibly converting it into alternative forms with reduced fluorescence, which are indistinguishable and challenging for readers to account for. The degradation of DTT over time, driven by its chemical properties, enables MGapt to return to its fully fluorescent state, dynamics observed using purified RNA.

We demonstrated that increasing the initial concentration of DTT in the expression of transcription, translation, or both could lengthen the suppression time. Moreover, given the inability to reduce the concentration of DTT in most commercial PURE systems, we introduce a model to predict the measured MGapt concentration by accounting for DTT's impact on the state of MGapt. We identified a distribution of possible parameters in the model with the experimental data at one concentration of RNA for MGapt-UTR1-deGFP. We validated our model by accurately predicting the measured MGapt for the same RNA construct at five different concentrations within the relevant 2 hour period.

This model is essential in comprehending and accurately quantifying transcription within cell-free expression systems, specifically commercial PURE systems. Employing our methodology enables the construction of mathematical models based on the measured MGapt concentration and can be utilized to retroactively calculate the total quantity of RNA used in the PURE reaction. More importantly, the realization that MGapt can transition between various fluorescence states in cell-free expression systems suggests potential inaccuracies in our understanding and assessment of RNA production. While MGapt is widely and extensively utilized as an aptamer, this phenomenon may not be exclusive. Exploring other aptamers and monitoring fluorescence changes over time could prove beneficial in identifying additional chemical reactions or environmental factors that might influence the measured RNA concentration.

2.4 Materials and Methods

All data analysis, parameter inference, and data presented in this chapter are available on Github [66].

PURE reactions and fluorescence measurements. The PURE reactions were prepared according to the NEB PURExpress (E6800) protocol, adapted for a 10 µL reaction volume, and incubated in a 384-well plate (Nunc) at 37 °C. A concentration of 5 nM DNA was utilized, unless otherwise specified, along with 8 units of RNAse inhibitor (NEB), and 10 µM malachite green oxalate added to each reaction. The RNA concentrations varied and were specified in the corresponding experimental setup description. DNA and RNA were added using an Echo 525 Acoustic Liquid Handler, and final concentrations were recalculated based on dispensed volumes. The remaining components were made into a master mix with 5 % excess, added to a 384-well plate using a repeater pipette. Fluorescence readings were obtained

using a BioTek Synergy H1 plate reader (BioTek) at 3-minute intervals for 12 hours at 37 °C. Excitation/emission wavelengths were set at 610/650 nm for MGapt with a gain of 150. All samples were analyzed using the same plate reader.

Standard MGapt RNA calibration. The fluorescence calibration curve for MGapt was generated using single-stranded RNA purchased from Integrated DNA Technologies (IDT), rArCrUrGrGrArUrCrCrCrGrArCrUrGrGrCrGrArGrArGrCrCrA-rGrGrUrArArCrGrArArUrGrGrArUrCrCrArArU. The sample arrived lyophilized in tube and weighed 64.5 nmol (0.92 mg). To achieve the concentration of 100 µM, 645 µL of nuclease-free water (NFW) was added. The sample was vortexed for several minutes and heated to 55 °C for 5 min before vortexing again. The stock concentration was measured by a Nanodrop 2000c before serial dilutions in 1X PBS. Next, respective dilutions were deposited onto the bottom of a Nunc 384 well plate using an Echo 525 Acoustic Liquid Handler. Each well contained a total volume of 10 µL with four technical replicates containing 10 µM of Malachite Green dye.

The Nunc 384 well plate was read using a BioTeK H1MF plate reader at 37 °C and at SI610/650nm (ex/em) and gain 150. Each point on each calibration curve represents the average of 20 points, and four replicates were read over 10 minutes at 2.5-minute intervals to generate 5 points per replicate. The points were all background-subtracted from the negative control such that the 0 µM samples had zero fluorescence. Points were fit using linear regression and were not forced to go through the origin. Fits for each calibration curve are indicated in Figure 2.13.

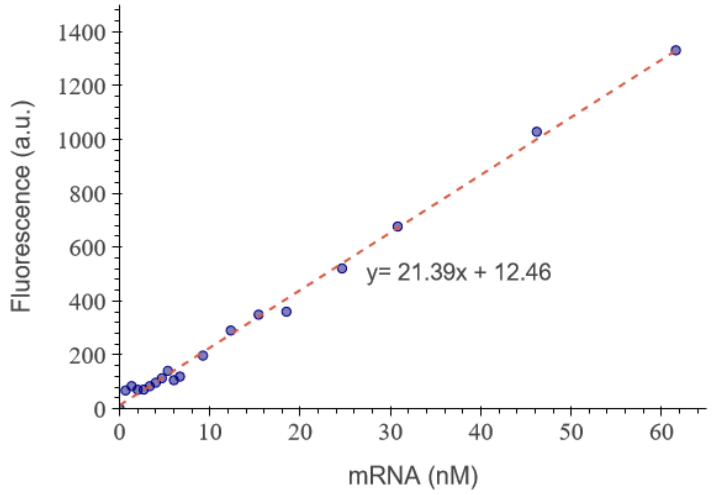

Figure 2.13: **Standard MGapt calibration curve.** The fluorescence calibration curve for MGapt is used to convert RFU to µM.

The pH of a PURE reaction using SNARF^{TM}-5F. The pH calibration curve in PURExpress was created using SNARF^{TM}-5F 5-(and-6)-carboxylic acid from Thermo Fisher Scientific (S23922). Eleven solutions with the following pH values were created: 3.01, 4.12, 5.08, 5.63, 5.97, 6.57, 7.02, 7.43, 8.07, 9.04, and 9.51 consisting of 50 mM HEPES, 1 mM spermidine, 350 mM K-Glutamate, 18 mM Mg-Glutamate, 50 mM creatine phosphate, and 5 mM dithiothreitol (DTT) [61]. Subsequently, 10 µL of SNARF^{TM}-5F at 100 µM was added to 90 µL of each pH solution. The PURE reaction samples were mixed as previously described, but for this read included 1 µL of SNARF^{TM}-5F before adding nuclease-free water to reach 10 µL.

Next, 10 µL of the final pH solutions with SNARF^{TM}-5F and PURE reaction were added to the Nunc 384 well plate and read using a BioTeK H1MF plate reader at 37 °C for 12 hours with gain 75 at (ex/em) 543/580 nm and 543/640 nm. The raw fluorescence units (RFU) at 580 nM was divided by RFU at 640 nM to attain Figure 2.14a. Finally, the pH calibration curve, shown in Figure 2.14b, was calculated by averaging the RFU in Figure 2.14a over the 12-hour read and plotting against the pH of the solution. Based on the ratio of 580 RFU over 640 RFU data of PURE samples, a linear regression line was fitted between pH 7.02 and pH 8.0 and used to calculate the pH over the PURE reaction time shown in Figure 2.2.

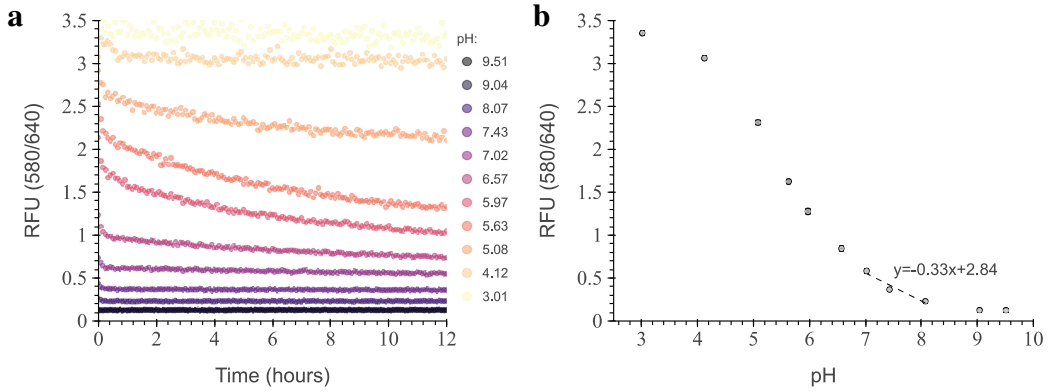

Figure 2.14: **pH calibration curves using SNARF^{TM}-5F. (a)** The pH was measured using SNARF-5F. The ratio of 580 RFU over 640 RFU data of solution, at respective pH, was used for pH calculations. **(b)** Calibration curve for pH calculations. Each point represents the average RFU (580/640) over the 12-hour read. A linear regression line was fitted between pH 7.02 and pH 8.07.

Incorporating MGapt into the desired plasmid. Primers used to clone MGapt (GGGATCCCGACTGGCGAGAGCCAGGTAACGAATGGATC) into DNA plasmid, pTXTL-T7p14-deGFP, originally obtained from myTXTL [67] and to linearize DNA for RNA purifications.

Table 2.5: List of primers used to make constructs.

Name	Sequence	Purpose
pT7_MGapt_FOR	GAGCCAGGTAACGAATGG ATCCAAT**AATTTTGTTTA ACTTTAAGAAGGAGATA TACCATG**	Cloning in MGapt to pTXTL-T7p14-deGFP
pT7_MGapt_REV	ATTGGATCCATTCGTTACC TGGCTCTCGCCAGTCGGG ATCCC**TCTAGAGGGAAA CCGTTG**	Cloning in MGapt to pTXTL-T7p14-deGFP
pPCR_MGapt_FOR	**GTGATGTCGGCGATATA GGC**	Linearize pTXTL-T7p14-mGapt
pPCR_MGapt_REV	**CACTATCGACTACGCGA TCATG**	Linearize pTXTL-T7p14-mGapt
pPCR_MGapt-UTR1 -deGFP_FOR	**GCGTAGAGGATCGAGAT CTCGATC**	Linearize modified pTXTL-T7p14-deGFP
pPCR_MGapt-UTR1 -deGFP_REV	**CTATCGACTACGCGATC ATGGC**	Linearize modified pTXTL-T7p14-deGFP

The bold text identifies the binding region of the plasmid.

Computational modeling and simulations. This model is based on MGapt different states effects by organic solvents found Stead *et al.*, Zhou *et al.*, and Da Costa *et al.* [39, 55, 58]. We use the chemical reaction network (CRN) formalism to create the detailed mechanistic model using a CRN compiler called BioCRNpyler [41]. The computational model was developed to take a total RNA concentration of MGapt-UTR1-deGFP and predict the measured MGapt concentration using a BioTek plate reader. Our CRN model consists of five distinct species and seven total reactions. Using BioCRNplyer, we identified five out of seven parameters using the experimental data. The final trained model predicts the BioTeK measurement of MGapt concentration over time for the RNA sequence of MGapt-UTR1-deGFP.

Chapter 3

A PURE CHEMICAL REACTION NETWORK OF PURE

The contents of this chapter are reproduced from

[1] Jurado, Zoila, Pandey, Ayush, and Murray, Richard M. "A chemical reaction network model of PURE." *bioRxiv* (2023). DOI: https://doi.org/10.1101/2023.08.14.553301.

3.1 Introduction

Over the last decade, multiple TX-TL protein expression models have been put forth and have shown to accurately model RNA and protein production [24, 25, 68, 69]. However, these models cannot predict expression without characterizing the models to their specific experimental data sets. Steps have been made towards more predictive models, modeling the behavior of whole circuits using a software toolbox and characterizing components of the entire model [70]. Though these TX-TL models can help understand phenomena or estimate unknown parameters, they continue to be constrained by the unknown composition of the cellular lysate.

One advantage of using cellular lysate is the retention of biological pathways of the cell strain, such as glycolysis, allowing for energy regeneration. The extent to which cellular processes remain functional is still undetermined. Characterization of lysate as the next step for TX-TL modeling relies on LCMS to measure small molecules, proteins, and lipids to understand how much of the core metabolism is active, potential side reactions, and waste generation effects [71]. However, measuring all proteins, small molecules, and lipids and mapping the chemical reactions associated with each in lysate is difficult. Thus, having a universal, batch-independent, and detailed TX-TL model is unlikely.

In contrast to cellular lysate-based cell-free protein synthesis, this allows for batch-to-batch and inter-laboratory repeatability. Consequently, the PURE system presents an opportunity for detailed modeling, allowing for reliable computational predictions. Without requiring recharacterization for every experimental run, these models could be integrated into pipelines to prototype larger circuits using the PURE system. Nonetheless, even with complete control and knowledge of the composition, existing PURE models fall back to the phenomenological modeling of transcription and

translation [25, 72–75] by grouping all NTPs as one variable, not modeling each step of protein production, and employing Hill functions instead of chemical reaction equations.

In 2017, Shimizu's group introduced a MATLAB model for the translation mechanisms in PURE [46, 47]. It is a detailed model that distinguishes between NTPs and incorporates each stage in translating an fMGG peptide. This computational model comprises 968 mass-action reactions and 241 species, including the 27 components that initialize the PURE system. Time courses of all components can be tracked in this model, which provides a valuable method to explore and systematically model the protein synthesis in PURE. But, fMGG is a small peptide, and extending this detailed model to the commonly used proteins in PURE is not straightforward. Explicitly writing each reaction and all possible species would be increasingly tedious as protein length increases. Moreover, since transcription is not modeled, this model is an incomplete description of the PURE system and cannot be experimentally validated. To address these limitations, in this chapter, we demonstrate (1) a detailed model of PURE for the transcription of arbitrary DNA sequences with mechanistic details for each step in the transcription; (2) a generalization of Shimizu's group translation model for arbitrary proteins; and (3) the experimental validation of a complete transcription and translation model of the PURE system.

3.2 Results and Discussion

Creating a chemical reaction network of transcription for any DNA sequence. To model the detailed step-by-step mechanistic process of transcription, we built a chemical reaction network (CRN). The chemical reactions of the transcription model were initially based on the reactions and rates proposed in the TX-TL model by Tuza *et al.* [76]. The user is required to input the DNA sequence of the desired protein, not including the promoter or terminator portions. To adapt and expand this model efficiently for any arbitrary RNA sequence, we used a CRN compiler tool called BioCRNpyler [41].

BioCRNpyler is a Python-based software package that can easily compile CRN models from simple descriptions of the parts of the system. For the transcription model, we include a detailed description of the process in such a way that the transcription mechanisms can be adaptable according to the RNA sequence being transcribed. BioCRNpyler also contains a library of parts and parameters that can be used to share parts of the model in other larger system models. We use the

features in BioCRNpyler to generate species and reactions depending on user input and to store these mechanisms in a way that can be used in larger models.

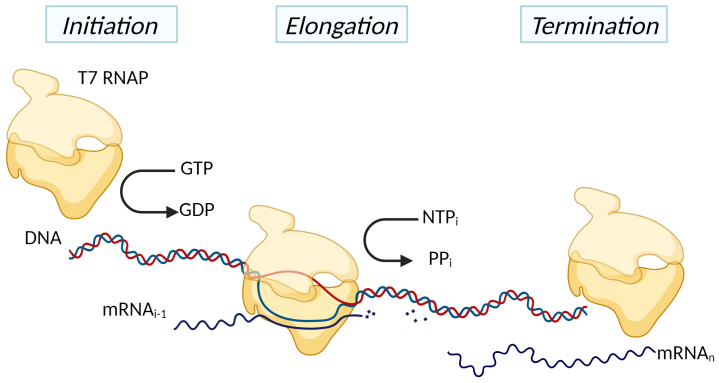

Figure 3.1: **Schematic of RNA synthesis with a reconstituted *E. coli* transcription system.** We split the transcription reactions into three sub-processes: initiation, elongation, and termination. Auxiliary reactions such as energy recycling or those explicitly related to translation are not included. Created with `BioRender.com`.

We split transcription into three groups: initiation, elongation, and termination, as illustrated in Figure 3.1 for $\mathrm{RNA}_{\mathrm{length}=n}$. We model the detailed mechanisms for each group to include all PURE components' interactions. Initiation steps include NTP degradation and a one-step GTP-dependent activation of T7 RNAP [77]:

$$\mathrm{RNAP} + \mathrm{DNA} + \mathrm{GTP} \xrightarrow{k_1} \mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{GTP} \tag{3.1}$$

$$\mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{GTP} \xrightarrow{k_2} \mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{GDP} \cdot \mathrm{P}_i \tag{3.2}$$

$$\mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{GDP} \cdot \mathrm{P}_i \xrightarrow{k_3} \mathrm{RNAP} \cdot \mathrm{DNA} + \mathrm{GDP} + \mathrm{P}_i \tag{3.3}$$

$$\mathrm{RNAP} \cdot \mathrm{DNA} \xrightarrow{k_{\mathrm{start}}} \mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{RNA}_0. \tag{3.4}$$

Elongation steps model each binding state separately for the addition of an NTP:

$$\mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{RNA}_0 + \mathrm{NTP}_1 \xrightarrow{k_{\mathrm{NTP_{bound}}}} \mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{RNA}_0 \cdot \mathrm{NTP}_1 \tag{3.5$_1$}$$

$$\mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{RNA}_0 \cdot \mathrm{NTP}_1 \xrightarrow{k_{\mathrm{NTP_{add}}}} \mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{RNA}_1 \cdot \mathrm{NMP}_1 \cdot \mathrm{PPi} \tag{3.6$_1$}$$

$$\mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{RNA}_1 \cdot \mathrm{NMP}_1 \cdot \mathrm{PPi} \xrightarrow{k_{\mathrm{NTP_{dis}}}} \mathrm{RNAP} \cdot \mathrm{DNA} \cdot \mathrm{RNA}_1 + \mathrm{NMP}_1 + \mathrm{PPi}. \tag{3.7$_1$}$$

The elongation steps repeat for each nucleic acid addition along the growing RNA chain of length n:

$$\mathrm{RNAP \cdot DNA \cdot RNA}_{n-1} + \mathrm{NTP}_n \xrightarrow{k_{\mathrm{NTP_{bound}}}} \mathrm{RNAP \cdot DNA \cdot RNA}_{n-1} \cdot \mathrm{NTP}_n \quad (3.5_n)$$

$$\mathrm{RNAP \cdot DNA \cdot RNA}_{n-1} \cdot \mathrm{NTP}_n \xrightarrow{k_{\mathrm{NTP_{add}}}} \mathrm{RNAP \cdot DNA \cdot RNA}_n \cdot \mathrm{NMP}_n \cdot \mathrm{PPi} \quad (3.6_n)$$

$$\mathrm{RNAP \cdot DNA \cdot RNA}_n \cdot \mathrm{NMP}_n \cdot \mathrm{PPi} \xrightarrow{k_{\mathrm{NTP_{dis}}}} \mathrm{RNAP \cdot DNA \cdot RNA}_n + \mathrm{NMP}_n + \mathrm{PPi}. \quad (3.7_n)$$

Finally, the termination step models the dissociation of the RNA_n, DNA, and T7 RNAP:

$$\mathrm{RNAP \cdot DNA \cdot RNA}_n \xrightarrow{k_{\mathrm{term}}} \mathrm{RNAP + DNA + RNA}_n. \quad (3.8)$$

In our transcription reactions, we do not explicitly account for multiple polymerases simultaneously bound on a singular DNA strand. However, the effects of simultaneous transcription can be incorporated into the reaction rates. Additionally, no auxiliary reactions, such as NTP recycling, are incorporated into the transcription model.

Parameter inference on the transcription model using pT7-MGapt-tT7 plasmid. We measured the T7 RNAP-driven transcription of malachite-green aptamer (MGapt) without a ribosome binding site (RBS) to parametrize the transcription models. The lack of RBS limits the reactions associated with translation, allowing the focus to be on transcription. A 10 µL reaction was done in triplicate using PURExpress® *In Vitro* Protein Synthesis Kit. The reaction contained 5 nM of plasmid DNA with construct pT7-MGapt-tT7, 10 µM malachite green oxalate, and 8 units of RNase inhibitor. The samples were mixed in PCR tubes with 5 % excess, then 10 µL was added to a 384-well plate and read for 3 hours at 37 °C in a BioTek H1MF plate reader. The total amount of RNA was calculated using dynamic calibration curves in Figure 3.15 for the respective construct.

Before starting the parameter inferencing, we needed to account for translation reactions independent of peptide synthesis, such as tRNA charging. As a result, we modified the initial conditions used in the Bayesian inference pipeline for the transcription model of DNA construct pT7-MGapt-tT7. The translation model was run without DNA to determine the amount of ATP and GTP consumed by these reactions within the first 15 min, giving us initial conditions for the transcription-only model given in Table 3.1.

Table 3.1: Initial conditions used in the Bayesian inference pipeline for the transcription-only model.

Species	Value	Unit
ATP	2206	µM
GTP	590	µM
CTP	1250	µM
UTP	1250	µM
DNA	5	nM
T7 RNAP	1	µM

To parameterize the CRN model constructed using BioCRNpyler, we initialized the parameters using the values from Tuza *et al.* [76]. The initial conditions were taken from PURE components [78] for NTPs and T7 RNAP concentrations. To get reliable predictions of transcription in PURE, we trained the model using experimental data for the MGapt fluorescence. However, this model training using parameter identification is not straightforward, as the CRN model we have constructed has many species and parameters. Since data is only available for one species in the model, it is crucial to assess the empirical identifiability of the model parameters. Finally, to account for the intrinsic noise observed in the experimental data, we use Bayesian inference to obtain a distribution of possible parameter values given the experimental data. To achieve both of these tasks (assessing identifiability and finding posterior parameter distributions), we use a biological data analysis pipeline [65] using the Python package Bioscrape [43]. This pipeline provides a practical interface to the BioCRNpyler model to run these analyses.

We use the local sensitivity analysis of all species in the model against all parameters and at all times to choose the parameters to identify. The sensitivity analysis heatmap for the pT7-MGapt-tT7 model is shown in Figure 3.2a. Based on the results using our initial parameters, Figure 3.2b, we initially choose to fit all eight reaction rates highlighted in equations (3.1)-(3.8) using a coarse tuning to search within the parameter space equal to twice the initial parameter given in Table 3A.1.

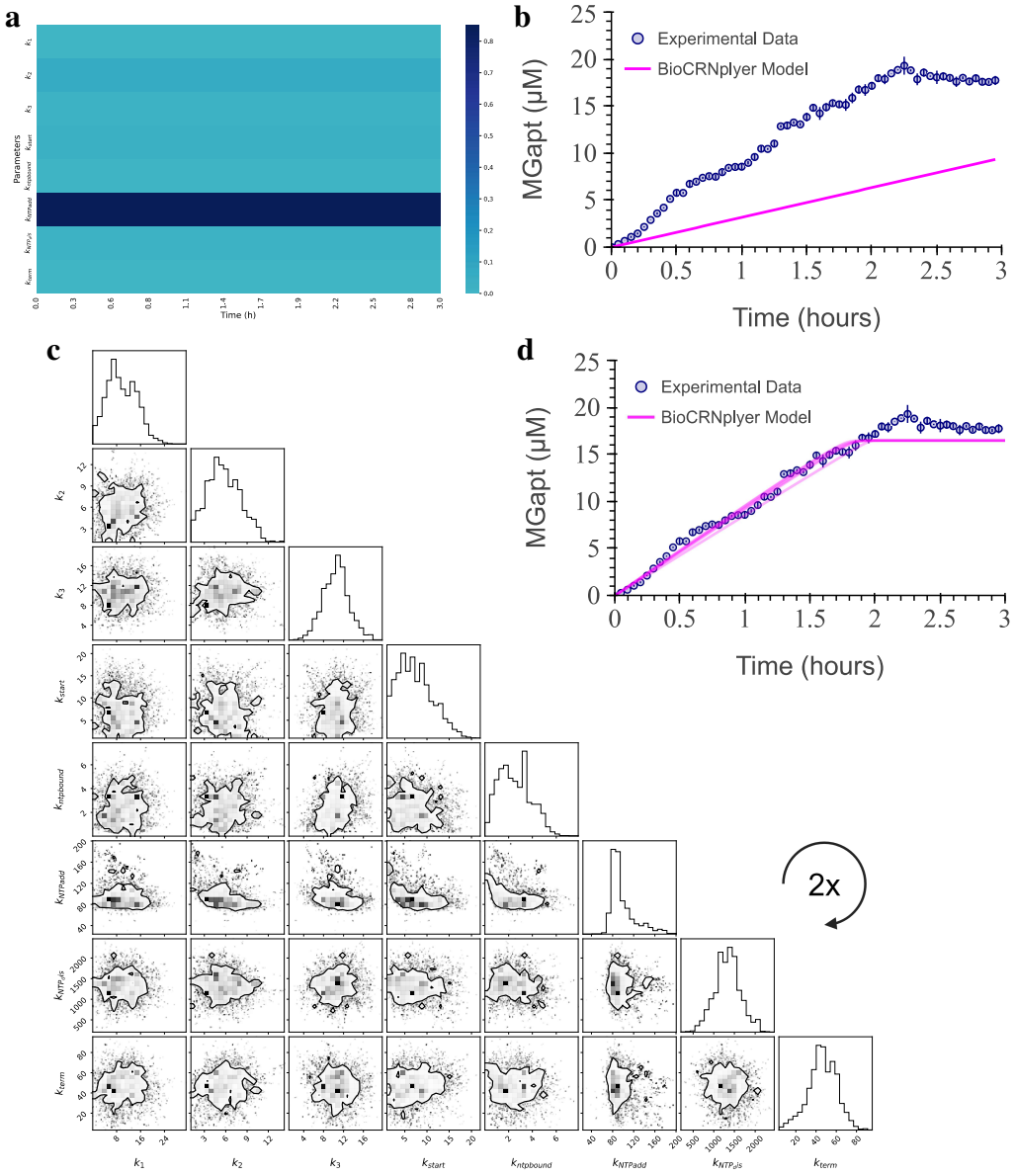

Figure 3.2: **Modeling, analysis, and parameter learning for the PURE transcription-only model.** (**a**) The sensitivity of the MGapt fluorescence to the CRN model parameters for all time. (**b**) The model simulations (in magenta) with the original reaction parameters and the experimental data for three biological replicates (blue circles with error bars). (**c**) The posterior distributions of parameters were obtained after running Bayesian inference on all eight reaction rates. The corner plot depicts the covariance of the eight parameters, with the contour showing the 75 % probability region for the parameter values. (**d**) With parameter values for all reaction rates sampled from the posterior distributions, the model simulations (in magenta) are shown alongside the experimental data for three biological replicates (blue circles with error bars).

The initial posterior distributions of parameters, shown in Figure 3A.1, were obtained using the Bayesian inference tools in Bioscrape on all the reaction rates. Subsequently, the model was re-trained using a narrower standard deviation around the results from the initial inference. The corner plot in Figure 3.2c shows the posterior parameter distributions and their covariance. This chart provides us with a distribution to sample from when predicting the output using the fitted model. The model simulations with parameter values drawn from the posterior along with the experimental data are shown in Figure 3.2d.

Table 3.2: Translation initial condition for the 36 proteins of PURE cell-free reaction model.

Species	Value	Unit
ATP	3750	µM
GTP	2500	µM
CTP	1250	µM
UTP	1250	µM
DNA	5	nM
CP	10	mM
FD	126.8498943	µM
T7 RNAP	1	µM
CK	10	µM
EFG	4.3	µM
EFTs	13	µM
EFTu	80	µM
IF1	99	µM
IF2	4.1	µM
IF3	4.9	µM
MK	5.6	µM
MTF	2.4	µM
NDK	1.8	µM
PPiase	0.16	µM
RF1	0.2	µM
RF2	0.2	µM
RF3	0.7	µM
RRF	16	µM

Species	Value	Unit
AAs	300	µM
AlaRS	3	µM
ArgRS	0.12	µM
AsnRS	1.7	µM
AspRS	0.49	µM
CysRS	0.1	µM
GlnRS	0.24	µM
GlyRS	0.35	µM
GluRS	0.9	µM
HisRS	0.34	µM
IleRS	1.5	µM
LeuRS	0.16	µM
LysRS	0.46	µM
MetRS	0.44	µM
PheRS	0.54	µM
ProRS	0.67	µM
SerRS	0.16	µM
ThrRS	0.34	µM
TrpRS	0.11	µM
TyrRS	0.03	µM
ValRS	0.07	µM
RS70S	3	µM

Once we identified the parameters for all eight reaction rates using transcription-only data and adjusted initial conditions, we selected the most sensitive parameters: k_2, k_3, k_{start}, and $k_{\text{NTP}_{\text{add}}}$, based on our sensitivity analysis in Figure 3.2a. These parameters were chosen to fully characterize the expression from pT7-MGapt-tT7, encompassing translation reactions independent of DNA presence. The initial conditions for the transcription and translation model are given in Table 3.2; based on Version 7 PURE concentration published in Table S1 by Kazuta *et al.* [61]

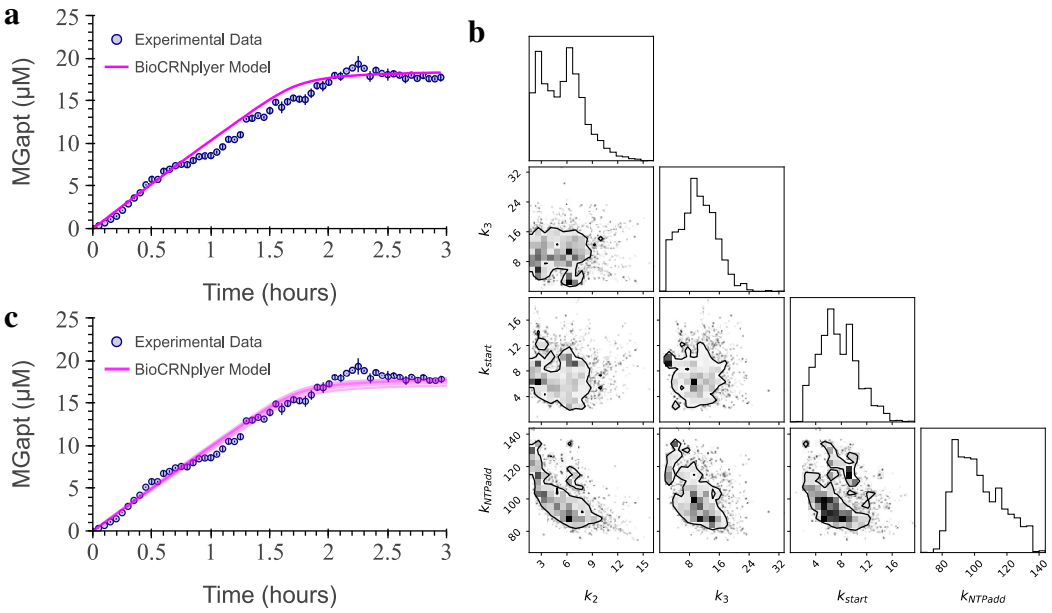

Figure 3.3: **Modeling, analysis, and parameter learning for the PURE model without protein production.** (**a**) The PURE model simulations (in magenta) with the previously fitted reaction parameters and the experimental data for three biological replicates (blue circles with error bars). This model includes translation without protein production. (**b**) We use the MGapt fluorescence to infer the posterior parameter distributions for k_2, k_3, k_{start} and $k_{\text{NTP}_{\text{add}}}$. The corner plot depicts the covariance of the four parameters, with the contour showing the 75 % probability region for the parameter values. (**c**) With parameter values drawn from the posterior distributions shown in (b), the model predictions (magenta) are shown alongside the experimental data. With parameter values for all reaction rates sampled from the posterior distributions, the model simulations (in magenta) are shown alongside the experimental data for three biological replicates (blue circles with error bars).

Finally, we use the Bayesian inference again to identify the posterior parameter distributions for these four parameters. The protein production independent PURE simulation with the previously referenced reaction parameters from the translation-only model is shown in Figure 3.3a. The corner plot in Figure 3.3b shows the posterior parameter distributions and their covariance. This chart provides a final

distribution to sample from when predicting the output using the fitted model. The model simulations with parameter values drawn from the posterior along with the experimental data are shown in Figure 3.3c. The final parameter values used are given in Table 3.3.

Table 3.3: Final transcription model parameters for PURE cell-free extract.

Parameter	Description	Value	Unit
k_1	Binding of RNAP and GTP to the DNA	9.41	$\mu M^{-2}\, s^{-1}$
k_2	Rate of formation of the RNAP bound GDP and phosphate complex on the DNA from RNAP bound GTP complex	5.06	s^{-1}
k_3	Unbinding of GDP and Phosphate from the RNAP and DNA complex	10.88	s^{-1}
k_{start}	Start of the initiation of the RNA transcript, (RNA_0) from the RNAP and DNA complex	7.64	s^{-1}
$k_{\text{NTP}_{\text{bound}}}$	Binding rate of NTP to the RNAP bound DNA, complex with initiated RNA transcript	2.68	$\mu M^{-1}\, s^{-1}$
$k_{\text{NTP}_{\text{add}}}$	Rate of elongation of the transcript	102.17	s^{-1}
$k_{\text{NTP}_{\text{dis}}}$	Unbinding rate of NMP and PPi from the open complex	1306.62	s^{-1}
k_{term}	Termination rate	45.99	s^{-1}

The model's number of parameters depends on the transcript sequence. For example, the transcription model for malachite green aptamer has 276 parameters.

Expansion of PURE translation model for an arbitrary set of amino acids. The MATLAB model for the PURE translation [46, 47] is limited to the case of the fMGG peptide. The model comprises 968 reactions and 241 species, all explicitly written out. As a result, the ability to change or extend the peptide is labor-some and futile. To make peptide variation more tractable, we first converted the MATLAB fMGG translation model to Python using BioCRNpyler [41]. The comparison between the BioCRNpyler model (magenta line) and MATLAB model (blue circles) can be seen in Figure 3.4a, with the error between the two in orange. While retaining all the reactions and parameters from the spreadsheet-based MATLAB model, the translation to Python enables user-friendly scripting and loops to iterate over an arbitrary length of peptide.

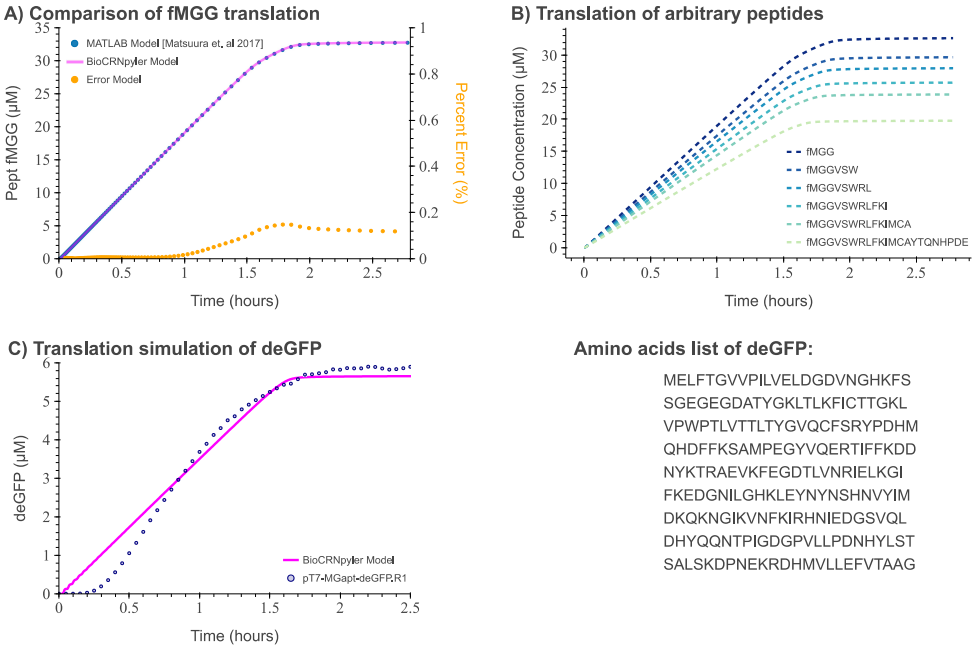

Figure 3.4: **Expansion the PURE-TL model.** (**a**) Comparison of original MATLAB model to BioCRNpyler model. BioCRNpyler model (magenta line) and MATLAB model (blue circles) overlap, and the difference between models in (orange circles) on a secondary axis. (**b**) Expansion of BioCRNpyler with different amino acids to the original fMGG peptide. (**c**) Translation model prediction of deGFP (magenta line) overlaying experimental result (blue circles). (**d**) Amino acid list used as input for the BioCRNpyler translation model.

The expanded PURE translation model in BioCRNpyler was built by adding different amino acids one by one until all 20 amino acids were incorporated (see Figure 3.4b). As expected, the amount of the final peptide decreases as the amino acid chain lengthens. Finally, the model was expanded for repeated amino acids and arbitrary amino acid sequences of length greater than 21. The translation of green fluorescent protein (deGFP) (RNA_n=805, $Pept_n$=226) was modeled using the BioCRNpyler translation-only model with the initial conditions adopted from the MATLAB model. Initial conditions for tRNAs and amino acids not associated with Met or Gly were absent in the MATLAB model. So, we set the initial conditions of the Gly-amino acids and tRNAs as previously given in Table 3.2. Modeling the translation of an experimentally relevant protein such as deGFP enabled us to compare the PURE models to experimental results (see Figure 3.4c, model in magenta overlays three experimental repeats in blue).

Using the extended translation model, we set RNA_n to 0.126 μM, such that the total deGFP expression is comparable to experimental results of 6 μM, as seen in

Figure 3.4c. The translation-only model does not accurately predict the first hour of protein expression. This discrepancy is expected since we start with a nonzero RNA. In the combined transcription and translation model, RNA would be produced, resulting in the delay of protein expression. Therefore, the immediate production of deGFP was expected, but further supports the need for a coupled transcription and translation model.

Validation of translation model of pT7-MGapt-UTR1-deGFP-tT7. To verify the extended translation model, we used additional 10 µL PURE reactions with purified RNA of MGapt-UTR1-deGFP at final concentrations of between 0.22 µM and 3.38 µM, initially introduced in Chapter 2. Based on the deGFP measurement, shown in Figure 3.5, we observed that the relationship between RNA added and deGFP production was non-linear.

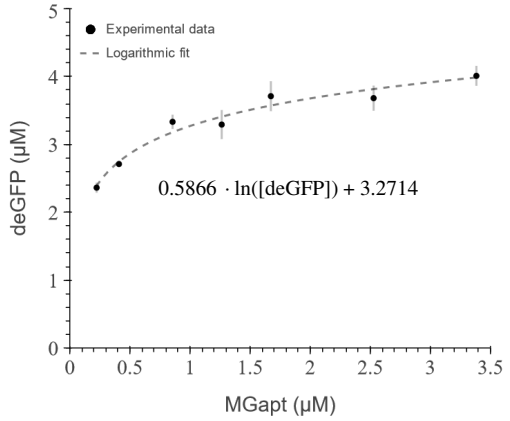

Figure 3.5: **Relationship between deGFP production and initial RNA concentrations.** Translation was initialized by the addition of RNA to achieve final concentrations: 0.22 µM, 0.41 µM, 0.86 µM, 1.23 µM, 1.67 µM, 2.53 µM, and 3.38 µM. The corresponding average deGFP production (black circles with associated error bars) was fit using a log fit due to RNA's apparent relationship with deGFP production. The fitted line overlays the data (dashed black line) with the equation displayed.

This suggests that there may be a limiting species, or the total RNA could be inhibiting translation, which is not fully captured in the translation model. We proposed that an effective RNA ($\mathrm{RNA}_{\mathrm{effective}}$) given by the equation,

$$\mathrm{RNA}_{\mathrm{effective}} = k\,(\mathrm{RNA}) \cdot \mathrm{RNA} \tag{3.9}$$

can be used to account for diminishing returns.

Due to the non-linear relationship between RNA and protein production, the RNA effective multiplication factor ($k\,(\mathrm{RNA})$) is a function of RNA. To calculate the

multiplication factor k(RNA) in equation (3.9) we identified the effective RNA ($\text{RNA}_{\text{effective}}$) that produces the corresponding deGFP for RNA concentrations at 0.22 µM and 3.38 µM (see Figure 3.6a). Next, we compute k(RNA) at 0.22 µM and 3.38 µM. Finally, shown in Figure 3.6b, we plot the computed k(RNA) against MGapt concentration and using a power trendline, we fit the points giving $k(\text{RNA}) = 0.1703\,\text{RNA}^{-0.801}$.

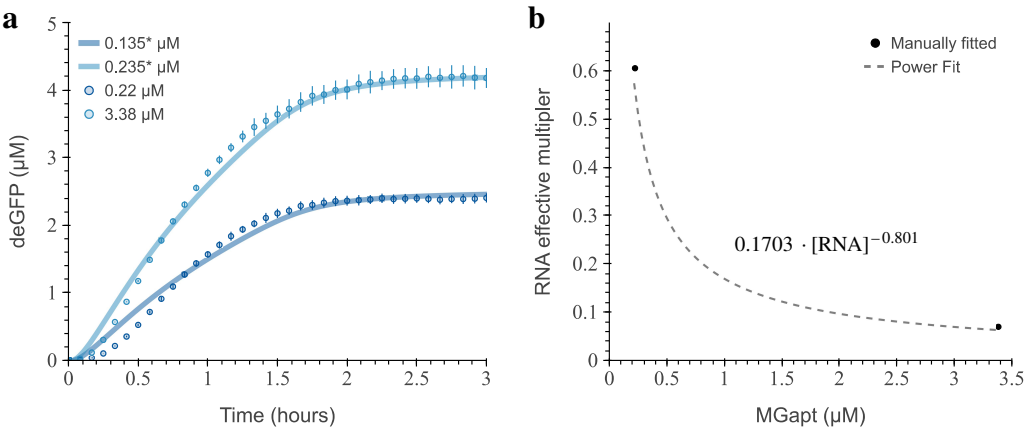

Figure 3.6: **Effective RNA required for the model to achieve corresponding deGFP production.** (**a**) Results from manually tuning the initial RNA concentrations to match the deGFP production from the two ends of RNA concentration experimentally tested: 0.22 µM, and 3.38 µM (circles with error bars). The model results with manually tuned RNA are shown by solid lines with colors corresponding to the experimental data it was tuned to. (**b**) The proposed RNA multiplier function to calculate the effective RNA concentrations to fit experimental deGFP production with the translation-only model.

Utilizing the RNA effective multiplication factor, we modeled the translation of deGFP using purified RNA at 0.41 µM, 0.86 µM, 1.26 µM, 1.67 , and 2.53 µM. Figure 3.7 shows the comparison between the simulated translation-only deGFP production model (solid lines) and the experimental data (circles with error bars) at various start RNA concentrations in corresponding colors.

The translation model does capture the final deGFP production, with the incorporation of $\text{RNA}_{\text{effective}}$, with error less than 6 % (see Table 3.4). However, the rate of deGFP production is not fully captured in the model despite accounting for the time delay between the start of the reaction and getting it into the plate reader. The disparity between the translation model and experimental data suggests that tuning the translation model parameters would be necessary. Unfortunately, due to the extensive size and time required for parameter fitting of all parameters, we opted to continue using it as proposed by Matsuura *et al.* [46].

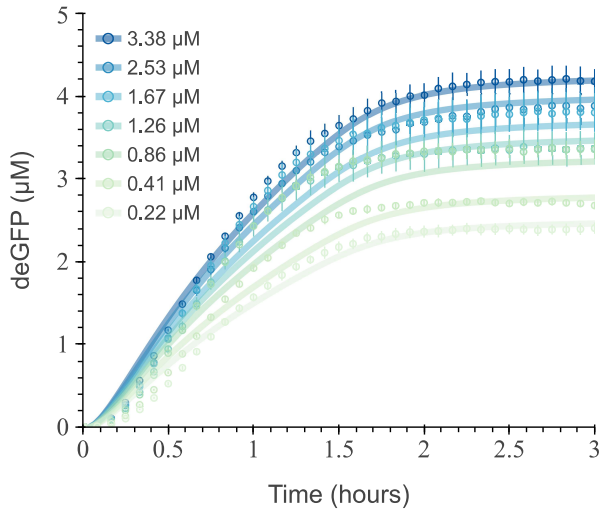

Figure 3.7: **Simulated deGFP production leveraging effective RNA calculation.** Modeled deGFP production in the translation model starting at various RNA concentrations (solid lines) overlaying experimental results (circles with error bars) at respective colors.

Table 3.4: Absolute error of the translation-only model compared to experimental results. The table lists the initial RNA conditions not used in calculating and determining the effective RNA multiplier function and the calculated effective RNA. The deGFP expressed reflects the concentrations at 2 h for both the simulation and experiment.

RNA_{added}	$RNA_{effective}$	$deGFP_{experimental}$	$deGFP_{simulated}$	Absolute Error
0.409 μM	1.153 μM	2.71 μM	2.65 μM	2.21 %
0.855 μM	0.178 μM	3.33 μM	3.08 μM	7.81 %
1.264 μM	0.192 μM	3.29 μM	3.31 μM	0.61 %
1.67 μM	0.204 μM	3.71 μM	3.50 μM	5.66 %
2.528 μM	0.221 μM	3.68 μM	3.78 μM	2.99 %

Combination transcription and translation model to achieve a detailed PURE model. With the successful construction of separate transcription and translation models for arbitrary sequences, BioCRNpyler easily allows for the combination of the two models. The reaction and species from both the translations and transcription models were compiled together, removing any duplicate species such as ATP, GTP, and PPi, etc. Next, the transcription model's output, RNA_n, and translation model's input, RNA, were linked with the uni-direction mass-action reaction:

$$RNA_n \xrightarrow{k_{linker}} RNA. \tag{3.10}$$

The parameter value of k_{linker} is arbitrarily set to $1000\,\text{s}^{-1}$, such that all of the RNA produced in the transcription model will be instantly available to the translation model. The inclusion of reaction in equation (3.10) can be omitted by utilizing the same species as the output and input of the respective models. However, it serves as a feature and/or replaceable reaction in case diffusion or RNA translations need to be incorporated. To account for protein folding, the following reaction was added:

$$\text{deGFP} \xrightarrow{k_{\text{folding}}} \text{deGFP}_m, \tag{3.11}$$

where $k_{\text{folding}} = 600s^{-1}$ [79], completing the combined PURE model. The total number of reactions of the combined transcription and translation model is 6988 with 6280 species. Unable to obtain the initial conditions of PURExpress from NEB and found contradicting concentrations across literature, the initial conditions were set to the Version 7 PURE concentration published in Table S1 [61]. The one exception is the small molecule creatine phosphate's (CP) initial concentration is 10 mM; see Table 3.2 for the initial conditions of all of the proteins and amino acids. Similar to the translation model by Matsuura *et al.*, the model tracks all species.

We initially found that the combined PURE model of deGFP production from DNA construct pT7-MGapt-UTR1-deGFP-tT7 at 5 nM did not accurately predict deGFP production despite the model's accuracy of RNA produced within the first 1 h. The over-prediction of deGFP may be the lingering non-linear correlation between RNA and deGFP production observed earlier. However, any inhibition of overloading the translation reaction with RNA would not be as significant, as RNA is being slowly produced. A more plausible alternative explanation is that incomplete proteins are produced in PURE cell-free protein synthesis systems compared to lysate-based cell-free protein synthesis proteins. The synthesis of incomplete peptides from the DNA construct pT7-UTR1-deGFP-tT7 can be observed using radiolabeled [^{35}S]-methionine [80] in two cell-free protein synthesis systems. The synthesis of full-length deGFP is approximately 28 kDa, and thus should measure between 26 kDa and 34 kDa. The completed translation of deGFP is denoted by the dark band observed in both BL21 (DE3) *E. coli* cell-lysate (Figure 3.8a) and NEB PURExpress (Figure 3.8b).

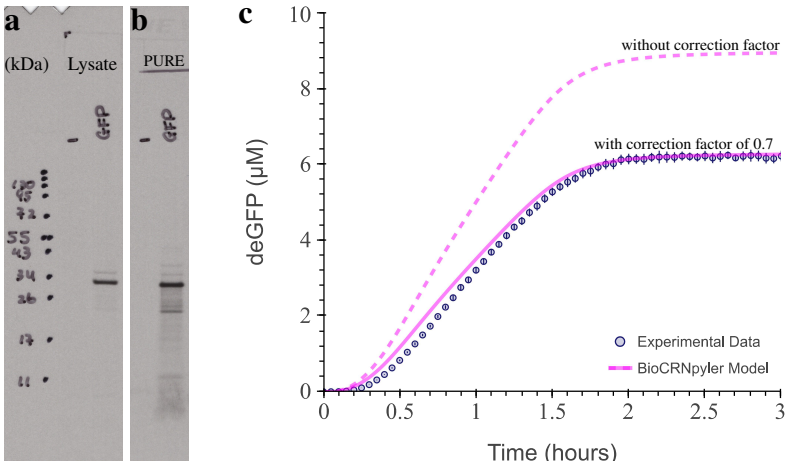

Figure 3.8: **[^{35}S]-methionine labeling of lysate-based and PURE cell-free protein synthesis systems.** Newly synthesized proteins from the DNA construct pT7-UTR1-deGFP-tT7 using BL21 (DE3) *E. coli* cell-lysate (**a**) and NEB PURExpress (**b**) using radiolabeled [^{35}S]-methionine. (**c**) Modeled deGFP expression in the combined models (magenta line) with correction factor of 0.7 (solid line) and without (dashed line) overlaying with experimental data, three replicates (blue circles and blue error bars). Figures (a) and (b) are courtesy of Masami Hazu, a PhD candidate in Prof. Voorhees' lab, who collaborated with us to conduct the experiment and contributed to the final results.

When comparing the newly synthesized proteins between the two cell-free protein synthesis systems, it is evident that PURExpress, Figure 3.8b, produces a higher quantity of incomplete proteins. Our combined model does not include early termination of translation or misfolding of deGFP. Therefore, to fit our deGFP results, we propose a 30 % reduction of deGFP production due to extenuating factors, depicted in Figure 3.8c. The final results of the combined model predictions are shown in Figure 3.9 overlaid with the experimental data (in blue shapes). The absolute errors of the combined BioCRNpyler model to the experimental data are depicted in Figure 3.10.

As seen in Figure 3.10a, the model of RNA production is consistent with experiments until approximately 1 h and RNA saturates approximately at the same time. However after 1 h the model over predicts MGapt production by 40 %. The difference after 1h in total MGapt produced between the predicted and experimental data may be attributed to MGapt degradation or inaccurately posed in the initial conditions. Additionally, the slight delay observed in the simulated MGapt and experimental MGapt is likely primarily due to the time elapsed from mixing the reaction and transferring the plate to the plate reader leading to the 20 % shown in Figure 3.10a.

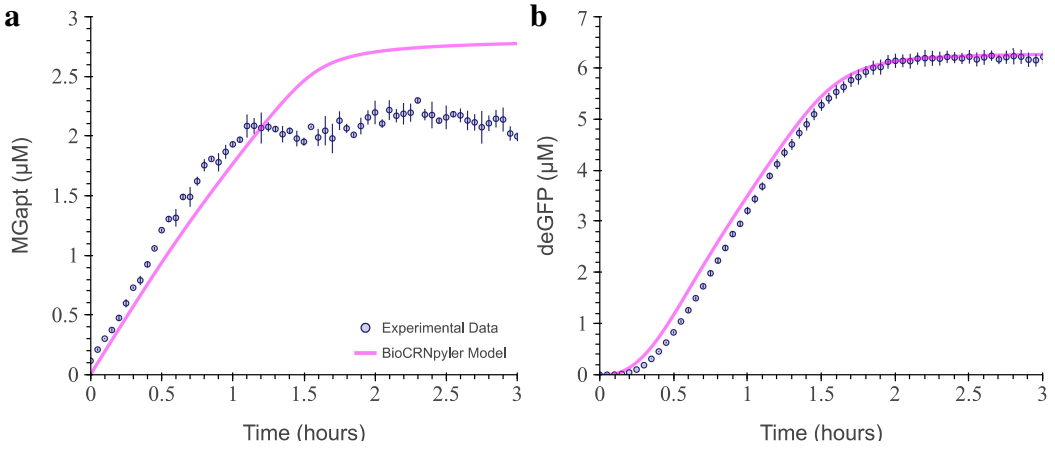

Figure 3.9: **The combined transcription and translation model for pT7-MGapt-UTR1-deGFP-tT7, DNA=5 nM, with experimental result in PURExpress.** (**a**) Modeled RNA production in the combined model (magenta line) overlaying with experimental data, three replicates (blue circles and blue error bars). (**b**) Modeled deGFP expression in the combined model (magenta line) overlaying with experimental data, three replicates (blue circles and blue error bars).

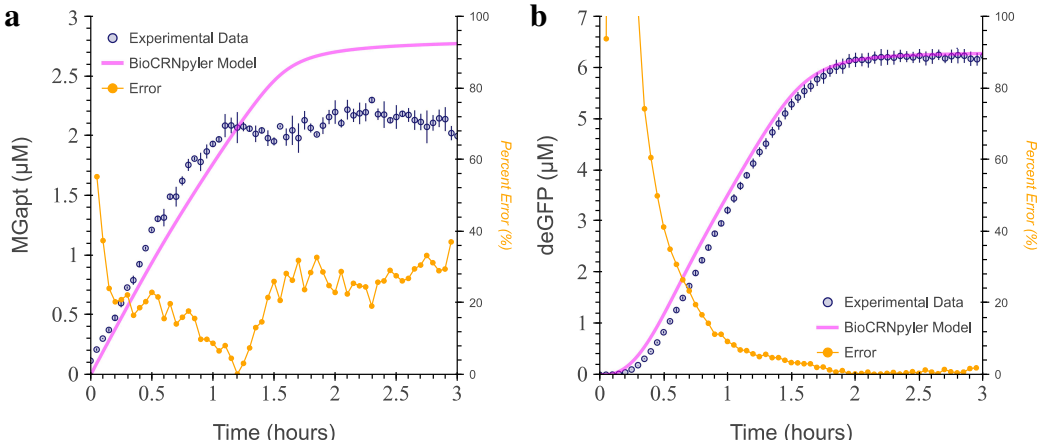

Figure 3.10: **Absolute error of combined BioCRNpyler model compared to experimental results.** (**a**) Modeled RNA production in the combined model (magenta line) overlayed with experimental data, three replicates (blue circles and blue error bars), and the absolute error (orange circles) on a secondary axis. (**b**) Modeled deGFP expression in the combined model (magenta line) overlayed with experimental data, three replicates (blue circles and blue error bars), and the absolute error (orange circles) on a secondary axis.

The expression pattern of deGFP more closely resembles actual experimental data regarding when expression ceases, compared to the translation-only model presented earlier. This is clear in comparing the "kink" in Figure 3.4c when the deGFP stops expressing with a much smoother transition in Figure 3.9b. By including the scaling factor of 0.70, the model accurately predicts total deGFP production throughout

the reaction's lifetime with a 5 % error at the end of the reaction, as shown in Figure 3.10b. Finally, due to the detailed nature of the model, we can track the concentrations of all proteins, amino acids, and energy carriers to begin exploring the limiting factors of PURE.

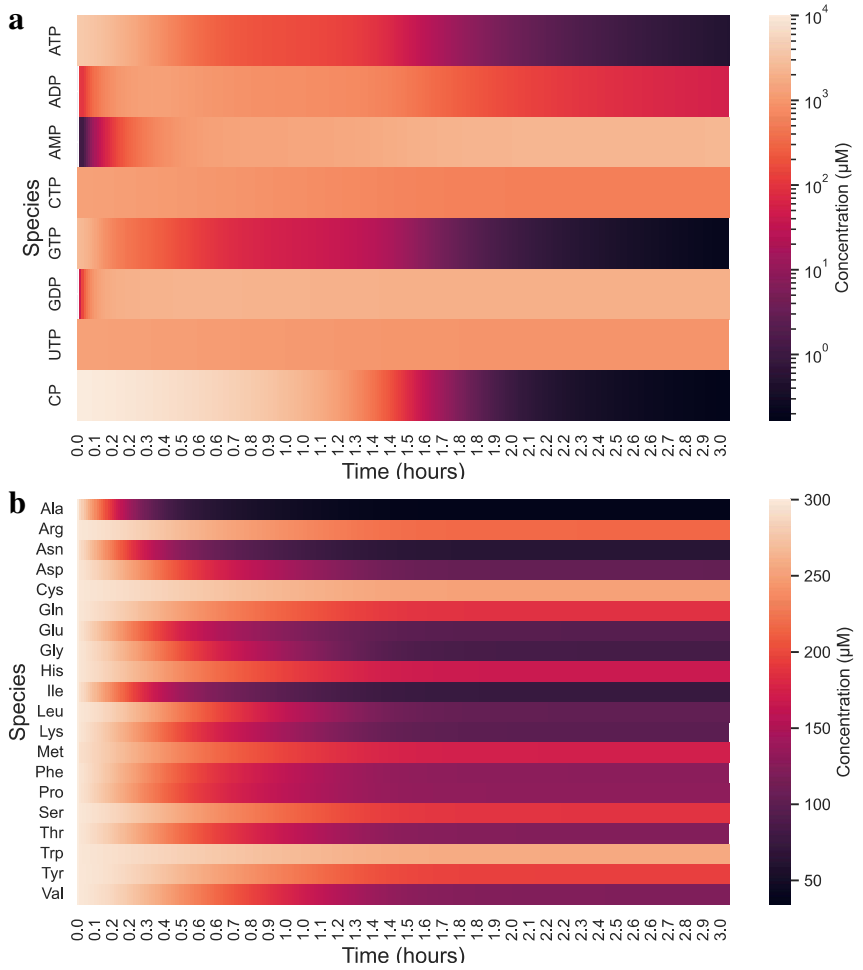

Figure 3.11: **The concentrations of NXPs and amino acids over time in the combined transcription and translation model for pT7-MGapt-UTR1-deGFP-tT7, DNA=** 5 nM**.** (**a**) Concentrations (μM) of ATP, ADP, AMP, CTP, GTP, GDP, and UTP over simulation time using log-normal scale. (**b**) Concentrations (μM) of all amino acids over simulation time.

By creating heatmaps of the concentrations of energy-related small molecules, in Figure 3.11a, we can infer that ATP, GTP, and CP are the likely limiting energy carriers in PURE—the consumption of GTP and ATP limits the model's RNA and protein production. The energy source, CP, serves as a phosphate donor for ATP synthesis and is fully depleted by 2 h. However, experimentally, it has been found that increasing CP does not increase the reaction lifespan and reduces total

protein yield, supporting the need for further development of energy generation and recycling in cell-free protein synthesis. Additionally, Figure 3.11b shows that amino acids are not limiting in PURE. Figure 3.11b includes the concentrations of only the free amino acids, those not in a complex with their respective tRNA. The excess of amino acids is apparent, with none of the 20 amino acid concentrations falling below 20 µM.

Validation of combined transcription and translation model of pT7-MGapt-UTR1-deGFP-tT7. To validate the complete detailed PURE model, which combines the transcription and translation model, we ran additional 10 µL PURE reactions with DNA of pT7-MGapt-UTR1-deGFP-tT7 at final concentrations of 0.07 nM, 0.12 nM, 0.26 nM, 0.47 nM, 1.06 nM, and 1.99 nM done in triplicates. By plotting the average deGFP production at 2 h against added DNA, we observed that the relationship between DNA added and deGFP production was non-linear (see Figure 3.12a). We observed diminishing protein production returns as DNA concentration exceeds 0.5 nM. To further investigate the origin of the nonlinearity, we plotted the relationship between DNA added to MGapt produced and MGapt produced to deGFP produced in Figure 3.12b and Figure 3.12c, respectively.

We observed that the relationship between DNA added to MGapt produced (Figure 3.12b) appears to be linear when the DNA concentration is above 0.5 nM, indicating a change of transcriptional regime at higher DNA concentrations. The shift in the transcription regime may indicate a transition from a DNA-limited regime. However, Figure 3.12c shows a diminishing return on protein production when the maximum concentration of RNA produced surpasses 1 µM. The diminishing return of deGFP production relative to RNA, previously observed in the translation-only experiments, may result from the saturation of translation proteins, production of inhibitory products, or reduced levels of ATP, GTP, and other small molecules not currently incorporated in the model.

Notably, in the translation-only results shown in Figure 3.5, when purified RNA was added at 0.855 µM and 2.528 µM, the respective deGFP expression was 3.08 µM and 3.78 µM. This is approximately 60 % lower than the deGFP expressed at comparable synthesized RNA concentrations in Figure 3.12c. These results suggest that starting with DNA in the PURE reaction is more effective for protein production than dosing in purified RNA. The reason for this remains to be explored, as our current model does not capture these experimental results. Consequently, similar to the validation of the translation-only model, we propose using an effective

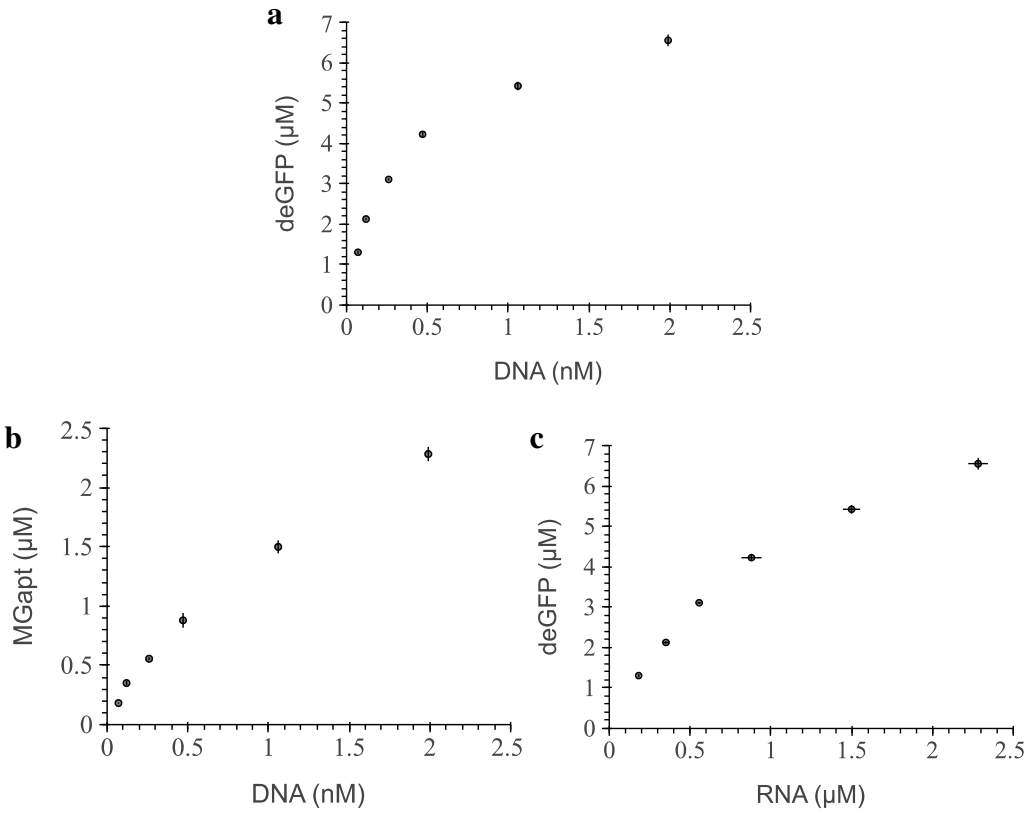

Figure 3.12: **Relationship between deGFP and RNA production at varying initial DNA concentrations.** Expression from the plasmids of pT7-MGapt-UTR1-deGFP-tT7 at DNA the following concentrations: 0.07 nM, 0.12 nM, 0.26 nM, 0.47 nM, 1.06 nM, and 1.99 nM. (**a**) The average overall deGFP production versus initial DNA concentration is depicted, followed by a breakdown into (**b**) RNA production relative to DNA and (**c**) deGFP production based on RNA production. The corresponding experimental data for each subplot is plotted in black circles with associated error bars.

DNA ($DNA_{effective}$) to recapitulate the maximum RNA synthesized experimentally to the model. Additionally, due to the sharp shift in the transcription regime in Figure 3.12c, we propose the effective DNA ($DNA_{effective}$) given by the equation,

$$DNA_{effective} = k\,(DNA) \cdot DNA. \tag{3.12}$$

To calculate the multiplication factor $k\,(DNA)$ in equation (3.12) we identified the effective DNA ($DNA_{effective}$) that produces the corresponding RNA for DNA concentrations at 0 nM, 0.12 nM, 0.47 nM, and 1.99 nM. Next, we plot the $DNA_{effective}$ against the experimental DNA concentration. We used a linear regression to fit the points when DNA was below 0.5 nM and a logarithmic trendline to fit the points above 0.5 nM, as shown in Figure 3.13.

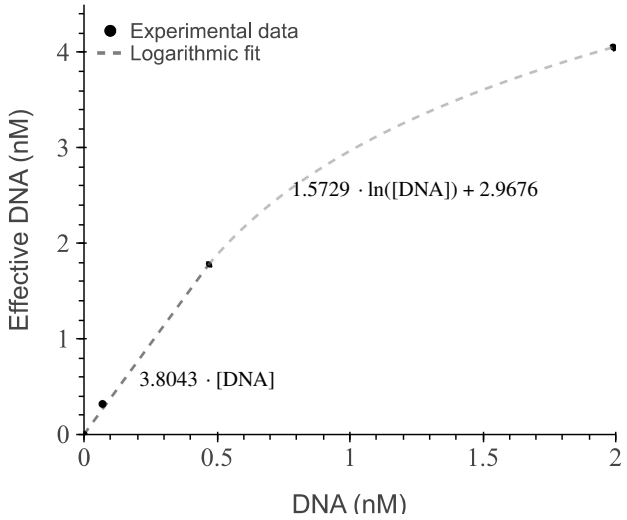

Figure 3.13: **Effective DNA required for the model to achieve corresponding RNA production.** The proposed piece-wise DNA function calculates effective DNA concentrations needed to fit experimental RNA production with the combined transcription and translation model.

Using the piece-wise DNA effective multiplication factor, we modeled the transcription and translation of deGFP at the different DNA concentrations. The simulation results are shown in Figure 3.14 and Table 3.5. In Figure 3.14, the simulation results (solid lines) are superimposed on experimental data (circles and error bars) with the varying DNA concentrations represented by different shades of purple in descending order.

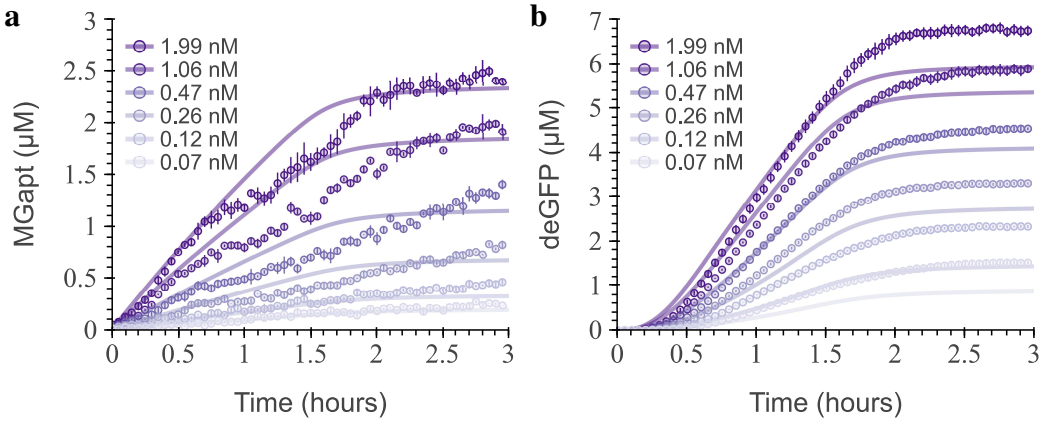

Figure 3.14: **The PURE model for pT7-MGapt-UTR1-deGFP-tT7, at different initial DNA concentrations, with experimental results in PURExpress.** The simulation results (solid lines) are overlayed with experimental data (circles and error bars) for the production of (**a**) RNA and (**b**) deGFP. The different DNA concentrations are reflected in the different shades of purple in decreasing order.

Table 3.5: Absolute error of MGapt and deGFP production of the PURE model over multiple DNA concentrations. The table includes the tested initial DNA conditions, corresponding effective DNA, MGapt synthesized, and protein production results from the model and experimental data. The MGapt and deGFP expressed reflect the concentrations at 2 h for both the simulation and experiment.

DNA_{added} (nM)	1.99	1.06	0.47	0.26	0.12	0.07
$DNA_{effective}$ (nM)	4.05	3.06	1.79	0.99	0.46	0.27
$MGapt_{exp.}$ (µM)	2.28	1.5	0.88	0.56	0.35	0.18
$MGapt_{model}$ (µM)	2.26	1.77	1.09	0.63	0.30	0.18
$MGapt_{error}$ (%)	0.88	18.0	23.9	12.5	14.3	0.0
$deGFP_{exp.}$ (nM)	6.55	5.42	4.23	3.11	2.13	1.31
$deGFP_{model}$ (µM)	5.77	5.20	3.91	2.55	1.31	0.79
$deGFP_{error}$ (%)	11.9	4.06	7.57	18.0	38.5	39.7

When comparing the simulated results to the experimental data, it becomes evident that while we can align the final MGapt production, the model fails to fully capture the MGapt synthesis's dynamics fully in Figure 3.14a. The model approximately aligns with the experimental results up to 45 min, after which there is a noticeable shift in the MGapt production rate. Furthermore, in Figure 3.14b, the model fails to capture the final deGFP expression, typically underestimating the final deGFP concentrations. Therefore, further fitting is required as the model parameters for the translation reactions were not fit to any data, and the transcription model was fit to experimental data at 5 nM DNA. Additionally, despite additional reactions that need to be identified and incorporated, the full PURE model remains useful to our understanding of PURE systems. The model can capture general trends and approximate the production of MGapt and deGFP.

3.3 Conclusion

This chapter proposes a complete transcription and translation model of cell-free protein expression for arbitrary proteins in the PURE system. The existing model of translation in PURE from Matsuura *et al.* [46] only modeled the translation of the fMGG peptide. We have developed a transcription model incorporating each step of the growing RNA strand and the specific NTP required to transcribe the RNA from any given DNA sequence. Further, we have expanded on the PURE translation model so that proteins with any given amino acid sequence can be modeled. By combining the transcription and translation models that we developed, we present a complete

model of protein expression in PURE cell-free systems. Using our approach, it is possible to create mathematical models of the expression of experimentally relevant proteins in PURE.

We validated the models using experimental data. We identified a distribution of possible parameters in the model with the experimental data for MGapt fluorescence. We showed that the validated model accurately predicts the MGapt fluorescence for two different plasmids — one where MGapt is expressed in an isolated manner and another where it is expressed together with deGFP. The combined model of the PURE cell-free system with the validated transcription model and the extended translation model was then used to predict deGFP expression. The model predictions agree with the experimental results.

This detailed model of the PURExpress system is a step towards cell-free protein synthesis predictability. It can be used as a platform to guide the design of future probing experiments with biological circuits in PURE. Utilizing this detailed model with OnePot PURE [15], a version of PURE where all 36 proteins are co-cultured and purified together, can help circumvent batch-to-batch and inter-laboratory variability problems seen in extract-based systems. Using quantitative proteomics [81] initial conditions of each OnePot batch can be measured and simulated to predict batch yield and improve the reproducibility of OnePot PURE. Beyond this, the model can be instrumental in identifying potential directions for further research, particularly regarding the coupling between the transcription and the translation mechanisms.

The model can be improved by incorporating reactions that model potential inhibitory interactions, such as inorganic phosphates, pyrophosphates, and pH. Additional reactions could include DNA polymerases for DNA replication and ribosomal loading reactions. Modeling of incomplete transcription and translation products [82] could also be beneficial in improving the accuracy of the predictions. We hope this model will combine multiple research groups by compiling data from fluorescent or LCMS measurements of different experimental conditions. Together, we can build a library of characterized parts for PURE or OnePot PURE to achieve a robust ‘design–build–test’ cycle.

3.4 Materials and Methods

All data analysis, parameter inference, and data presented in this chapter are available on Github [83].

Dynamic MGapt RNA calibration. The dynamic fluorescence calibration curve for MGapt was generated using purified RNA of MGapt at 0.522 µM and MGapt-UTR1-deGFP at 0.548 µM in a 10 µL PURExpress reaction done in three technical replicates. The PURE reaction with the purified RNA containing the MGapt was loaded to a Nunc 384 well plate and was read using a BioTeK H1MF plate reader at 37 °C and at SI610/650nm (ex/em) and gain 150. The relative fluorescence units (RFU) for each of the RNA units tested are shown in Figure 3.15a (MGapt) and Figure 3.15b (MGapt-UTR1-deGFP). Subsequently, dynamic calibration curves for RNA of MGapt (Figure 3.15c) and MGapt-UTR1-deGFP (Figure 3.15d) were obtained by dividing RFU measurements in Figure 3.15a and Figure 3.15b by their respective RNA concentrations, then smoothed. The dynamic fluorescence calibration curve for MGapt is specific to the PURE reaction and RNA measured.

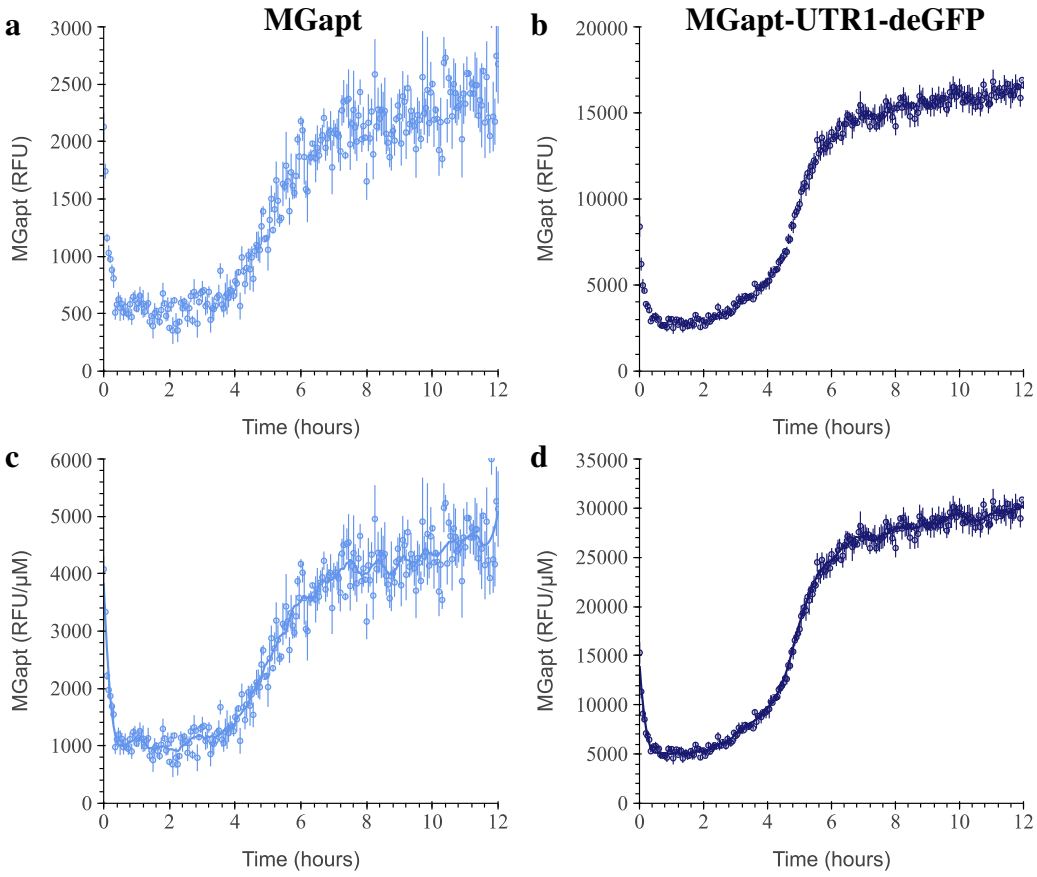

Figure 3.15: **Dynamic MGapt calibration curve.** Measured RFU of RNA of MGapt at 0.522 µM (**a**) and MGapt-UTR1-deGFP at 0.548 µM (**b**) done in triplicates (circles with error bars). Respective dynamic calibration curves for MGapt (**c**) and MGapt-UTR1-deGFP (**d**) RNA, calculated by dividing RFU measurement in (**a**) and (**b**) by respective RNA concentration then smooth.

Standard deGFP calibration curves. The fluorescence calibration curve for deGFP was generated using purchased purified eGFP from Cell Biolabs (STA-201). Samples were prepared as described in the myTXTL manual [67]. The 1 mg mL^{-1} eGFP (29.0 kDa was estimated to have a concentration of 34.483 µM). The eGFP stock was diluted in series in 1X PBS, and 10 µL of each dilution was pipetted onto the wall of a Nunc 384 well plate, spun down, and then sealed with a plastic film. The plate was allowed to sit for 45 min at room temperature before being read in a BioTek H1MF plate reader at 30 °C and at SI485/515nm (ex/em) and gain of 61. Each point on the calibration curve represents the average of 12 points; three replicates were read over 3 minutes at 1-minute intervals to generate 4 points per replicate. The points were all background-subtracted such that the PBS-only samples had zero fluorescence. Points were fit using linear regression and were not forced to go through the origin. Fits for each calibration curve are indicated in the 3.16.

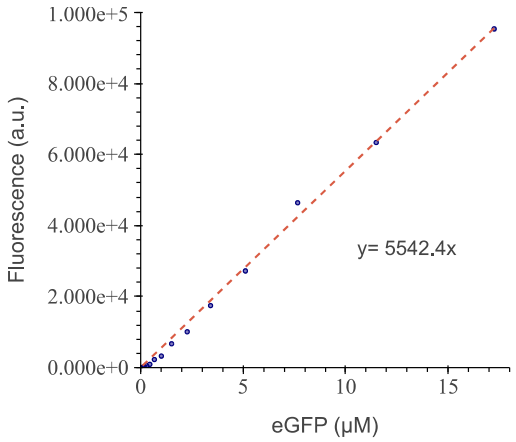

Figure 3.16: **Standard deGFP calibration curve.** The fluorescence calibration curve for deGFP is used to convert RFU to µM.

PURE reactions and fluorescence measurements. PURE reactions were mixed by following the protocol by PURExpress (E6800), adjusted for a 10 µL reaction, and allowed to run in a 384-well plate (Nunc) at 37 °C. DNA at 5 nM was used, unless otherwise stated, 0.8 units of RNAse inhibitor (NEB), and 10 µM of malachite-green dye was added to each reaction. Fluorescence measurements were read in a Synergy H1 plate reader (Biotek) at 3 min intervals using excitation/emission wavelengths set at 610/650 nm (MGapt) at gain 150 and 485/515 nm(deGFP) at gain 61. All samples were read in the same plate reader, and for deGFP relative fluorescence units (RFUs) were converted to nM of protein using a purified eGFP standard by following the protocol in paper [67]. Calibration curves for MGapt and deGFP are depicted in Figure 3.15 and Figure 3.16.

Computational modeling and simulations. BioCRNpyler outputs the model in the standard biological modeling language called the Systems Biology Markup Language (SBML) [42]. The exported SBML files can be simulated with any compatible SBML simulator. We use the Bioscrape [43] Python package to simulate the SBML models. The CRN model is converted to an ordinary differential equation by Bioscrape and solved using Python `odeint` for desired initial conditions. To convert the CRN to an ODE, each reaction rate is written using the mass-action propensity [44]. We use Bioscrape because it supports sensitivity analysis and Bayesian inference tools for SBML models and the model simulations. For each SBML model, we run the local sensitivity analysis to obtain the sensitivity of the measured species with all parameters and at all times. Then, we aim to identify the most sensitive parameters for the model using the experimental data. We perform the parameter identification using a Bayesian inference algorithm implemented in Bioscrape with emcee Python package [45]. Given the experimental data, we obtain a probability distribution for each identified parameter with Bayesian inference. Model simulations with parameter values sampled from these posterior probability distributions are then plotted against the experimental data to evaluate the quality of the model predictions. These posterior probability distributions also quantify the uncertainty in the data, which is an important advantage of Bayesian inference methods.

All calculations and plotting were performed using the standard stack of Python packages: NumPy [84], SciPy [85], Pandas [86], Matplotlib [87], Bokeh [88], and Seaborn [89]. The run time for each simulation depends on the model. The simulation time varies from less than a second to a couple of hours on a personal computer running AMD Ryzen 7 4700U 2.0 GHz with 16GB of RAM. Running the model simulations on a high-performance computing cluster can speed up the simulation times and Bayesian inference routines by around 10x.

3.A Appendix A

Initial transcription-only model parameters used in the transcription-only model. The initial parameter values for the transcription model are given in Table 3A.1. The initial chemical reaction rates of the transcription model were based initially on the TX-TL model by Tuza *et al.* [76].

Table 3A.1: Initial transcription model parameters for PURE cell-free expression.

Parameter	Description	Value	Unit
k_1	Binding of RNAP and GTP to the DNA	6.10	$\mu M^{-2}\,s^{-1}$
k_2	Rate of formation of the RNAP bound GDP and phosphate complex on the DNA from RNAP bound GTP complex	2.95	s^{-1}
k_3	Unbinding of GDP and Phosphate from the RNAP and DNA complex	7.82	s^{-1}
k_{start}	Start of the initiation of the RNA transcript, (RNA_0) from the RNAP and DNA complex	5.24	s^{-1}
$k_{\text{NTP}_{\text{bound}}}$	Binding rate of NTP to the RNAP bound DNA, complex with initiated RNA transcript	1.47	$\mu M^{-1}\,s^{-1}$
$k_{\text{NTP}_{\text{add}}}$	Rate of elongation of the transcript	23.59	s^{-1}
$k_{\text{NTP}_{\text{dis}}}$	Unbinding rate of NMP and PPi from the open complex	985.89	s^{-1}
k_{term}	Termination rate	32.38	s^{-1}

Bayesian inference posterior distributions from the coarse tuning of the transcription-only model.

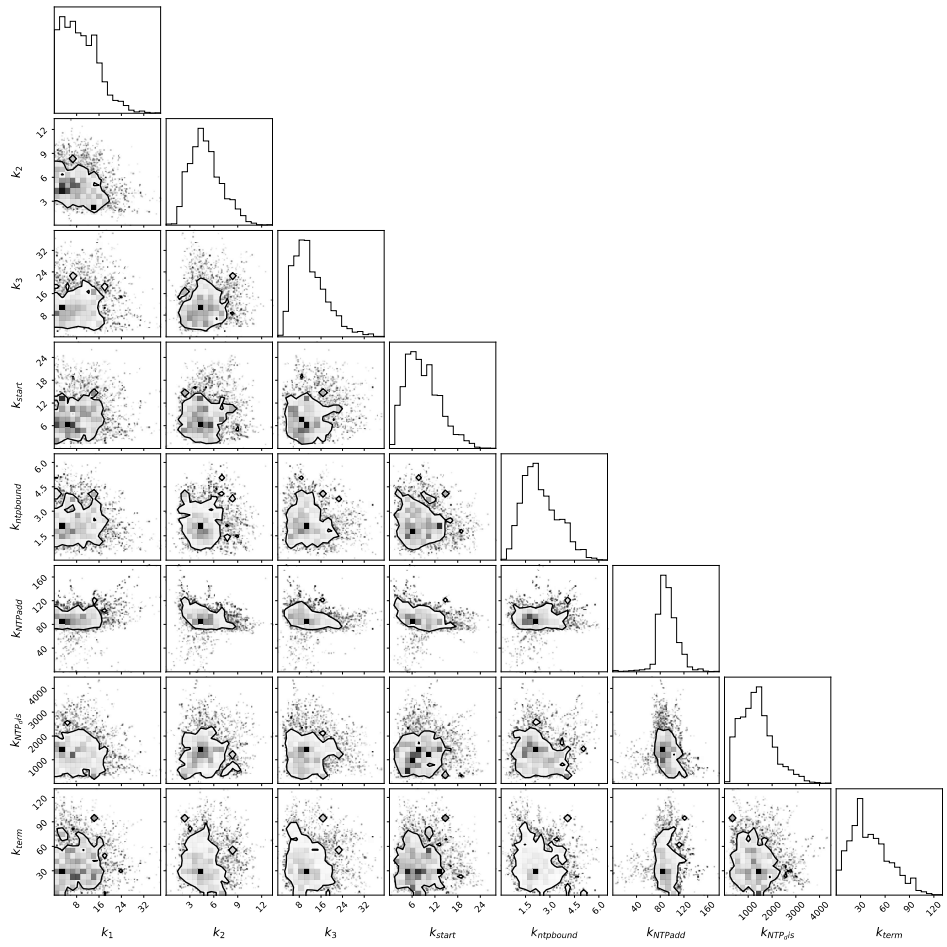

Figure 3A.1: **Initial coarse Bayesian inference posterior distributions.** The initial posterior distributions of parameters on all eight reaction rates. The corner plot depicts the covariance of the eight parameters, with the contour showing the 75 % probability region for the parameter values having limited information on accurate parameter values.

Chapter 4

SEPARATION OF TRANSCRIPTION AND TRANSLATION

4.1 Introduction

The primary distinction between prokaryotic and eukaryotic cells is the absence of compartmentalization, such as a distinct nucleus or other organelles enclosed in internal membranes. As a result of these compartments, many new functions arose. We focus on creating a membrane-bound organelle containing the cell's genetic information, the nucleus, an evolutionary milestone around 2.1 billion years ago as supported by the fossil record [90]. Regardless of the events leading to the ancestral eukaryotic cells acquiring a nucleus [91], transcription and translation were forever separated.

The separation of transcription and translation in eukaryotic cells allowed for greater control and regulation of gene expression through the development and use of post-translational modifications (e.g., splicing) and at the transcription level. Governed by natural selection, these finer gene expression controls must have provided them with beneficial traits, improving their survival and reproduction rate. To understand the evolution of the nucleus and its benefits, we studied the compartmentalization of transcription from translation. Using the transcription and expanded translation models introduced in Chapter 3, we have successfully developed an initial model of a synthetic cell with a nucleus. By isolating the transcription model, we will explore the benefits and limitations of separating transcription and translation measured by the total protein production and changes in dynamics such as duration and production rate.

Alongside simulating the synthesis of deGFP using synthetic nuclei, we have modeled the expression of the HiBiT peptide. The HiBiT peptide was promising in the initial experimental design proposals of the synthetic nucleus, given its short RNA strand of only 87 base pairs, which yields a peptide consisting of 12 amino acids. As a short peptide, HiBiT comprises a restricted set of amino acids. As illustrated in Figure 4.1a, depicting the frequency of each amino acid, not all amino acids are present within the peptide. Unlike deGFP, which requires the presence of all amino acids for expression, the most frequent amino acids in deGFP are glycine, leucine, lysine, aspartic acid, valine, threonine, and alanine as depicted in Figure 4.1b.

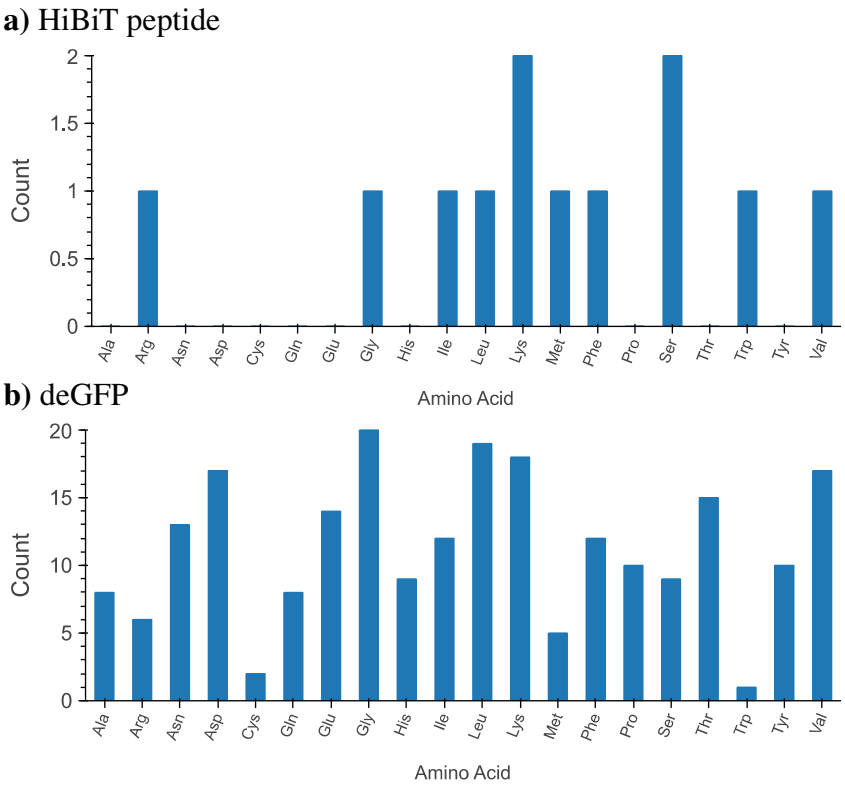

Figure 4.1: **Amino acid frequency for simulated protein.** A graph depicting the amount of each amino acid in the HiBiT peptide **(a)** and deGFP **(b)**.

The HiBiT peptide is part of the Nluc system that binds to the Large BiT (LgBiT), forming a complex. The completed enzyme consumes the substrate furimazine, generating a bright luminescent signal detectable with a luminescent reader [92, 93]. Despite encountering challenges in implementing these experimental designs, the numerical simulations still address broader questions regarding the benefits of a synthetic nucleus.

4.2 Results and Discussion

Modeling of a synthetic cell with a nucleus. The synthetic cell with an artificial nucleus is depicted in Figure 4.2 as two compartments. The "nucleus" (Figure 4.2b) is a self-contained system focused solely on transcription. The "cytoplasm" (Figure 4.2c) functions as the synthetic cytoplasm, accepting the RNA generated by the synthetic nucleus for protein synthesis. The "nucleus" and "cytoplasm" interact by a selectively permeable membrane. The membrane and its permeability to RNA are captured in the linker reaction illustrated in Figure 4.2.

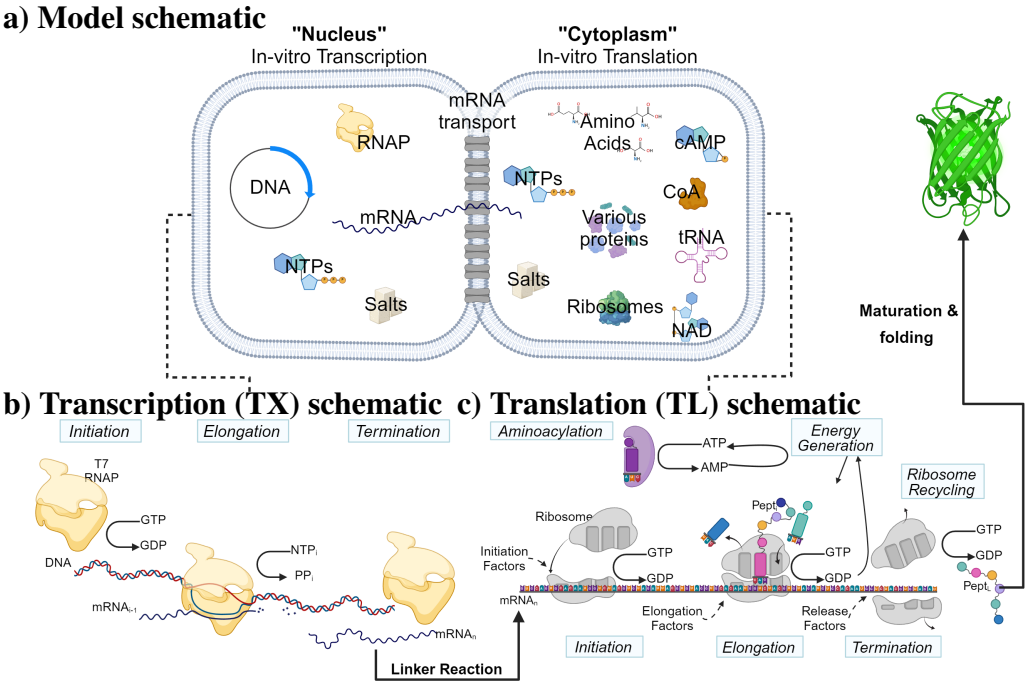

Figure 4.2: **Modeling a synthetic nucleus: combining transcription and translation system models.** (**a**) Model schematic of a transcription and translation system joined by a membrane system that allows diffusion of mRNA. (**b**) The transcription schematic and (**c**) the translation schematic, introduced in Chapter 3. The transcription and translation systems are joined by a membrane system that allows the diffusion of mRNA. Following peptide formation, additional folding and maturation reactions are included for deGFP. Illustration of translation model (c) was adapted from Matsuura *et al.* [46]. Created with `BioRender.com`.

The membrane only allows for the unidirectional translocation of RNA. Initially, it combines the two models seamlessly, such that all the RNA produced in the translation model is instantaneously available for the translation model through the reaction in equation (3.10); $\text{mRNA}_n^* \xrightarrow{k_{\text{linker}}} \text{mRNA}_n$. However, the membrane reactions can be modified to follow the diffusion of additional small molecules such as NTPs. Following peptide formation, additional folding and maturation reactions are included for deGFP. The depiction of the synthetic cell model with an artificial nucleus serves as an abstraction of the synthetic nucleus for reference, as the models do not currently incorporate explicit dependence on membrane surface area or volume.

Impact of membrane permeability with shared resources. In our initial investigation into the consequences of engineering a synthetic nucleus, we established a numerical experiment in which energy resources like GTP, CTP, UTP, and ATP

were shared between the transcription and translation models. In these models, no distinction is made between ATP and GTP available for use by any part of the model, reflecting a synthetic cell scenario where NTPs can freely pass through the membrane. The impact on protein production with variable membrane permeability was investigated and tested by decreasing the values of k_{linker} by orders of 10 from $1\,\text{s}^{-1}$ to $0.0001\,\text{s}^{-1}$. The k_{linker} parameter value of $1\,\text{s}^{-1}$ is equivalent to a non-compartmentalized system and will be distinguished by the color magenta. The synthetic cell simulation with shared resources was run with two different constructs: pT7-MGapt-UTR1-deGFP-tT7 and pT7-UTR1-HiBiT-tT7, at 5 nM of DNA shown in Figure 4.3.

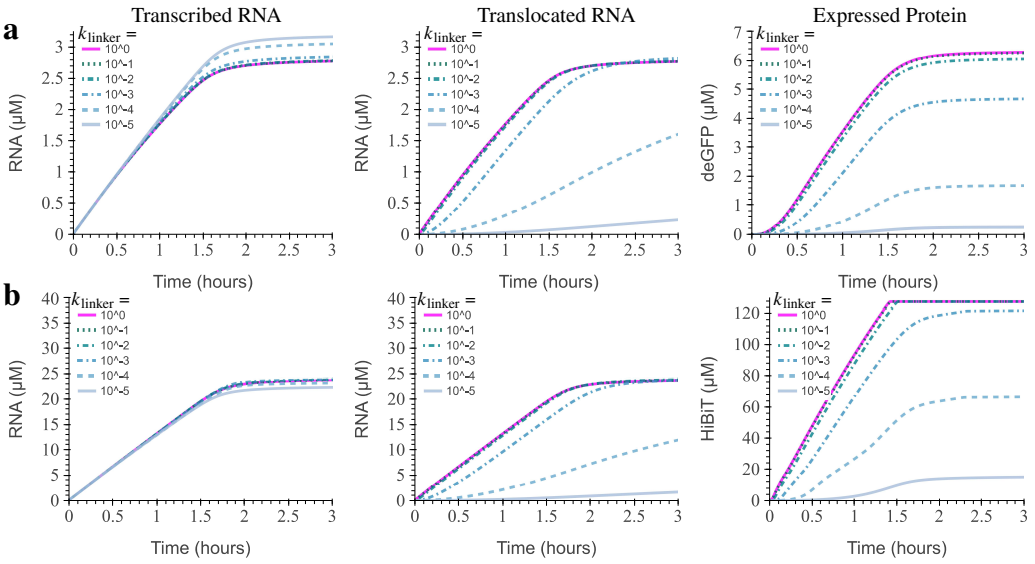

Figure 4.3: **Impact on protein expression in PURE model with shared resources and varying RNA permeability.** Simulation results of PURE-based cell-free protein expression of DNA constructs: (**a**) T7-MGapt-UTR1-deGFP-tT7 and (**b**) pT7-UTR1-HiBiT-tT7 at 5 nM with varying RNA permeability represented by different colors and dash type. The simulation results for each construct are divided into three columns: (**Left**) the total quantity of RNA transcribed by the "nucleus" (transcribed RNA), (**Middle**) the total quantity of RNA in the "cytoplasm" accessible to the translation machinery (translocated RNA), and (**Right**) the total quantity of protein produced (expressed protein).

The simulation results for each construct are split into three columns: the total amount of RNA transcribed by the "nucleus" (transcribed RNA), the total amount of RNA in the "cytoplasm" accessible to the translation machinery (translocated RNA), and the total amount of protein produced (expressed protein). Starting with the expression of deGFP, an observation emerged when comparing transcribed

RNA in Figure 4.3a: the total transcribed RNA varies with different membrane permeabilities. The observation that transcription increases with decreased RNA translocation suggest that resource competition may limit transcription. Inhibiting RNA translocation may allow for a greater allocation of energy to transcribing the RNA strand. This energy trade-off does not exhibit a one-to-one relationship, as depicted in Figure 4.3a. While protein yield decreases with reduced membrane permeability across all tested parameter values of k_{linker}, supporting the conclusion that severely hampering RNA permeability leads to a significant decline in protein expression.

The results for the expression of the HiBiT peptide are shown in Figure 4.3b. The HiBiT peptide is about 20× smaller than the deGFP protein. Consequently, our models show that the total RNA transcribed and protein translated is higher than that of deGFP in Figure 4.3a. However, we see that neither the production of RNA nor protein increases by 20×. The disproportional increase of RNA and DNA is likely due to the inefficient energy usage of the entire system.

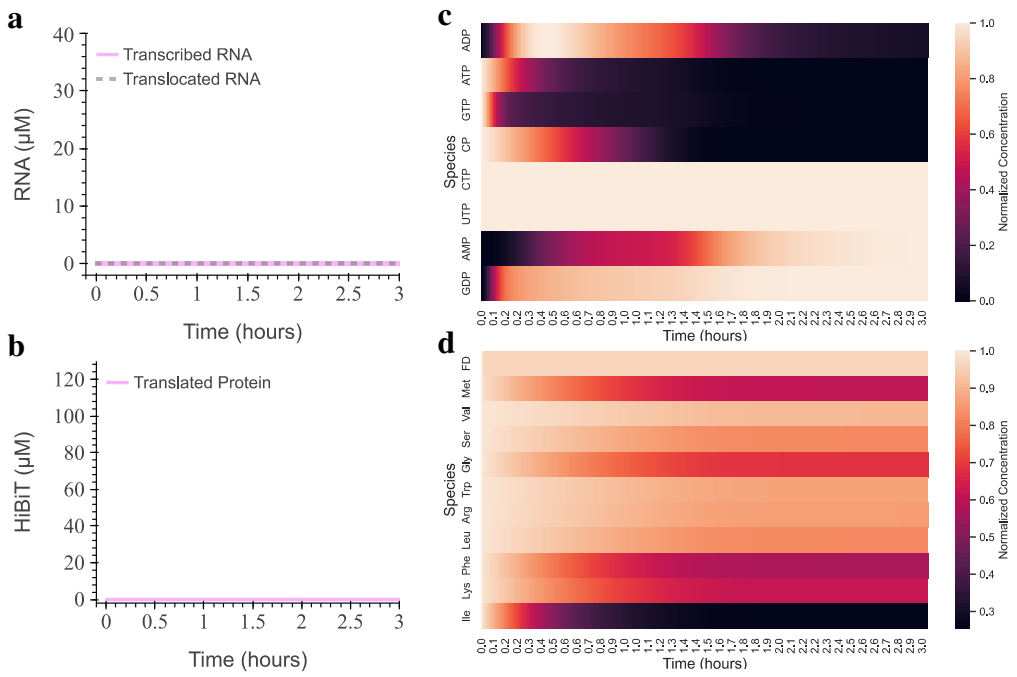

Figure 4.4: **Resource depletion in the PURE reaction without DNA or RNA.** Simulation results of PURE-based cell-free protein expression if run without DNA or RNA. (**a**) Plots of the transcribed RNA (magenta solid line) and the translocated RNA (grey dashed line) and (**b**) show the HiBiT peptide produced. **c**) Normalized concentrations of RNA, ATP, ADP, AMP, CTP, GTP, GDP, CP, and UTP over the simulation time. (**d**) Normalized concentrations of amino acids in the HiBiT peptide over the simulation time alongside FD and the HiBiT peptide.

For example, in Figure 4.4, when the DNA and RNA concentration was set to 0 nM despite no RNA synthesis (Figure 4.4a) or protein production (Figure 4.4b), ATP and GTP would continuously be used in auxiliary reactions (Figure 4.4c). As seen in Figure 4.4d, the normalized concentrations of unbound amino acids are below 1, indicating the charging of tRNAs without peptide formation. Additional auxiliary reactions include complex formation with elongation, initiation, and release factors. Figure 4.4 helps explain why the reaction's lifespan will never exceed 2 h when energy resources are shared.

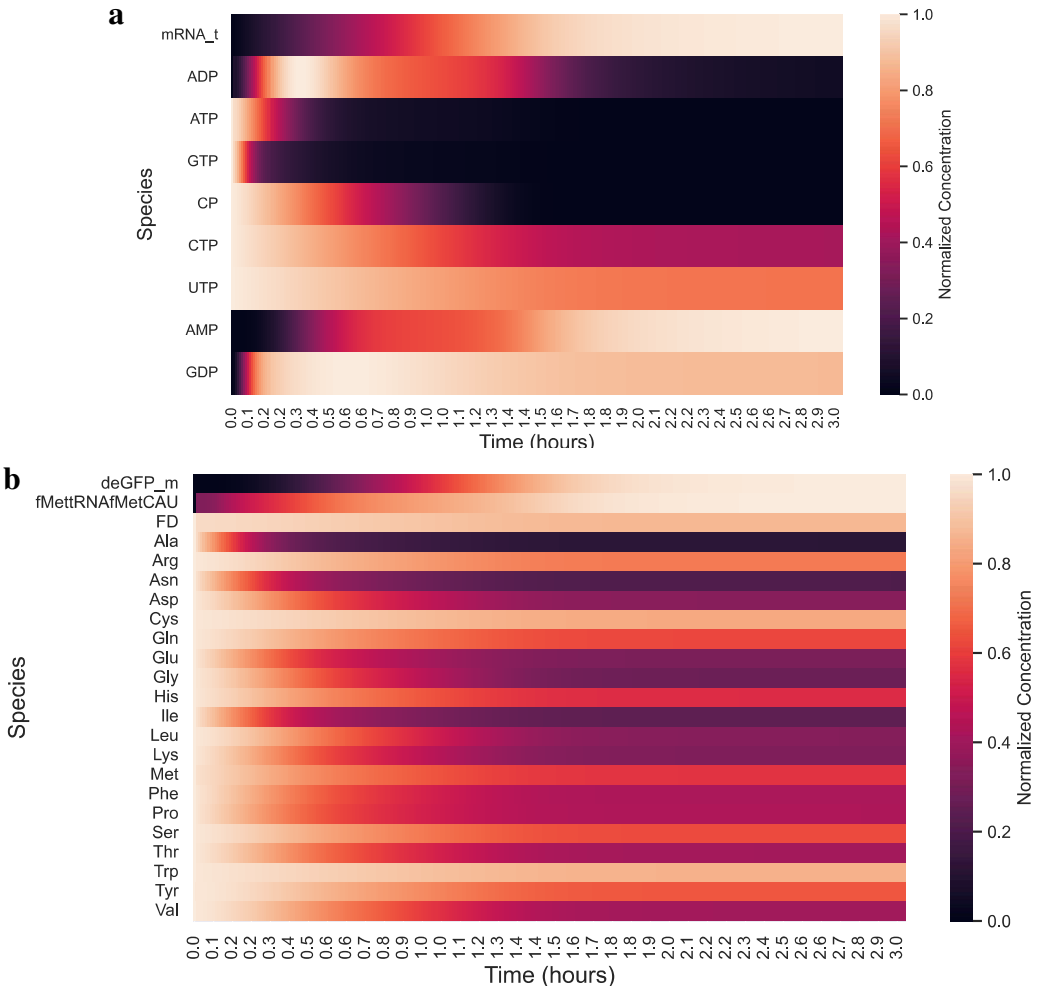

Figure 4.5: **The concentrations of NXPs and amino acids over time in the combined transcription and translation model expressing the deGFP.** (**a**) Normalized concentrations of RNA, ADP, ATP, GDP, CP, CTP, UTP, AMP, and GMP over the simulation time. (**b**) Normalized concentrations of amino acids over the simulation time alongside FD and deGFP peptide and k_{linker}= 1 s^{-1}.

Examining the heatmaps of the PURE simulation expressing deGFP in Figure 4.5, we observe that RNA and deGFP production saturation occurs around the same

time creatine phosphate (CP) and ADP are fully consumed. However, in Figure 4.6, when modeling the HiBiT peptide, the limiting resource appears to be 10-formyltetrahydrofolate (FD). FD is a formyl donor essential to producing the first amino acid, synthesis-formylmethionine (fMet).

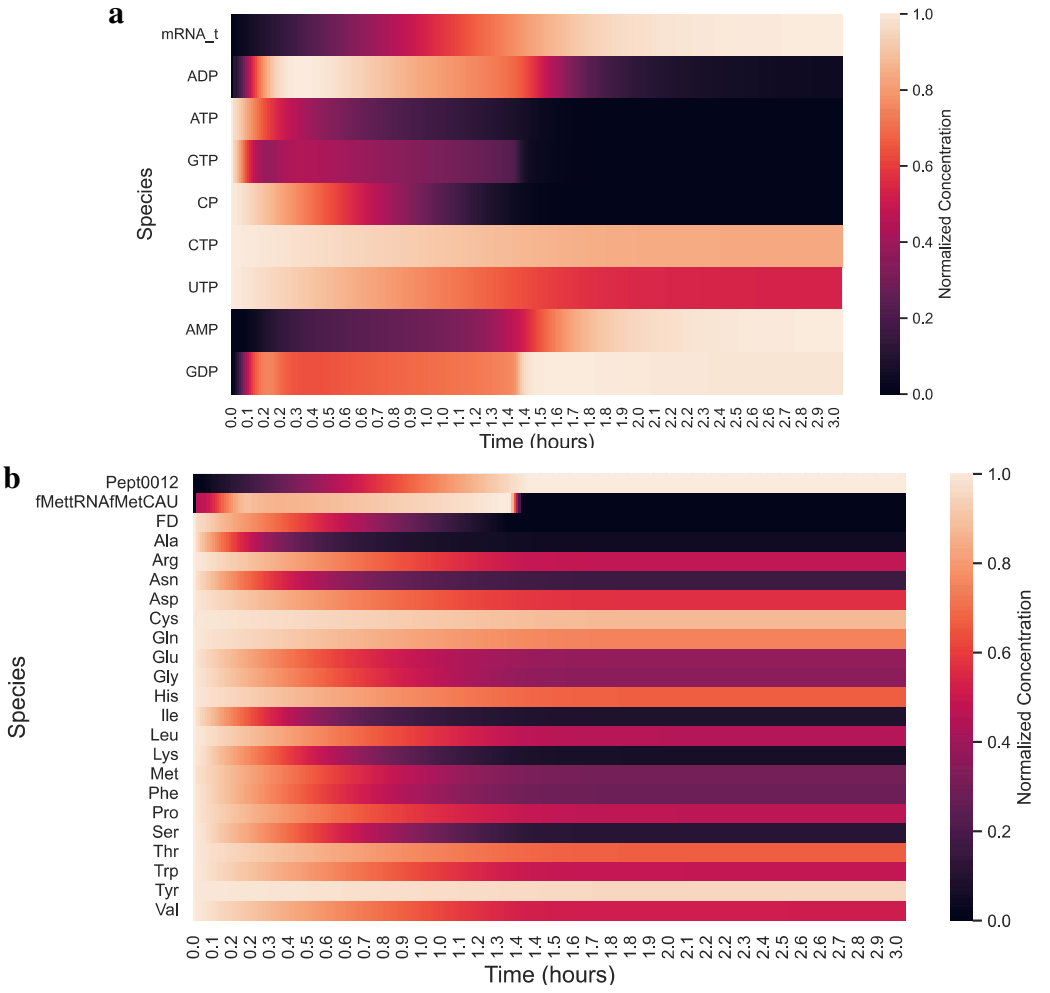

Figure 4.6: **The concentrations of NXPs and amino acids over time in the combined transcription and translation model expressing the HiBiT peptide.** (**a**) Normalized concentrations of RNA, ATP, ADP, AMP, CTP, GTP, GDP, CP, and UTP over the simulation time. (**b**) Normalized concentrations of amino acids in the HiBiT peptide over the simulation time and k_{linker}= 1 s^{-1}.

Finally, in both constructed models, decreased membrane permeability decreases total protein production. We observe decreasing transcribed RNA as membrane permeability decreases only when expressing the HiBiT peptide, which contradicts results from deGFP models. The conflicting observation suggests that transcription may exhibit self-inhibitory behavior by depleting itself from ADP and GDP.

Impact of membrane permeability with split resources. Specifically from Figure 4.3b, we also observe that unless RNA's accessibility to the transcription machinery is severely restricted, the total protein yield remains unchanged. Given that the ribosome concentration is fixed at 3 µM, any RNA expression beyond 3 µM is redundant and results in an inefficient utilization of total energy. Therefore, the transcription model's NTPs can be constricted without protein production loss, increasing the synthetic cell's total efficiency. Realistically, the membrane was made impermeable to NTP concentrations in the transcription model to constrict the concentrations of NTPs. As a result, the NTPs in the model were designated for the transcription or the translation model, further functionally splitting the "nucleus" and "cytoplasm" compartments.

Table 4.1: The allocation of energy resources of NTPs between the "nucleus" and "cytoplasm."

Nucleus	
Species	Concentration (µM)
ATP_{tx}	$3750\,p$
GTP_{tx}	$2500\,p$
CTP_{tx}	$1250\,p$
UTP_{tx}	$1250\,p$

Cytoplasm	
Species	Concentration (µM)
ATP_{tl}	3750 - ATP_{tx}
GTP_{tl}	2500 - GTP_{tx}

The NTP and its respective concentration associated with each compartment are identified with the subscript 'tx' for the "nucleus" or 'tx' for the "cytoplasm." The variable p indicates the explicit portion allocated to the transcriptions model; all NTPs are uniformly decreased by this factor. The combined amount of GTP and ATP remains constant across both systems.

The "nucleus" comprises only the essential components necessary for RNA strand synthesis: DNA, T7 RNAP, and NTPS. Meanwhile, the "cytoplasm" contains the remaining 35 essential PURE proteins along with ATP, GTP, CP, FD, ribosomes, amino acids, and tRNAs. To deepen our exploration and comprehension of energy usage in the model, alongside reducing membrane permeability to RNA, we decreased the concentrations of NTPs in the "nucleus" compartment while maintaining a constant total concentration within the synthetic cell. Therefore, except for the initial NTP conditions, all other species remain unchanged from the original model, as listed in Table 3.2. The initial concentrations of NTPs in the "nucleus" were established by proportionally dividing the concentration of the NTP by a factor of p, with the remainder allocated to the cytoplasm shown in Table 4.1. Three values

of the factor p were tested: $p = 50\,\%$, $p = 25\,\%$, and $p = 10\,\%$ for the DNA constructs expressing the HiBiT peptide or deGFP.

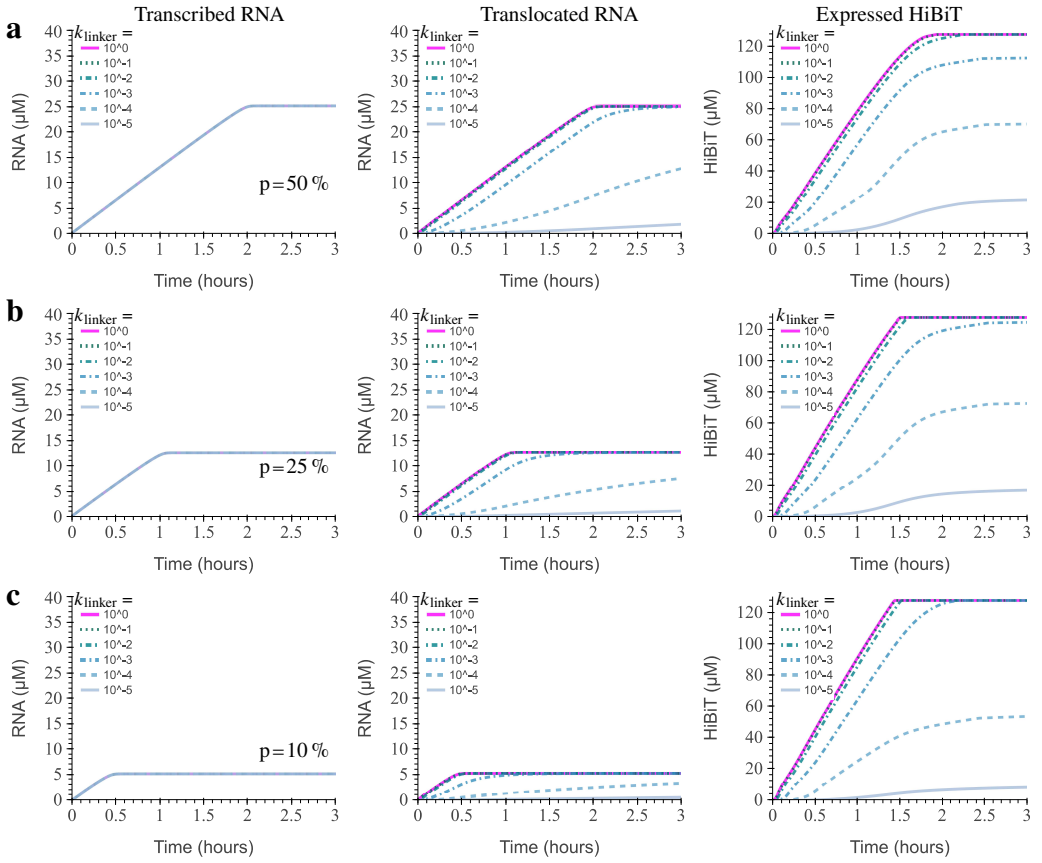

Figure 4.7: **Impact on the HiBiT peptide expression in PURE model with shared resources and varying RNA permeability.** Simulation results of PURE-based cell-free protein expression of DNA constructs, T7-UTR1-HiBiT-tT7 at 5 nM with varying RNA permeability (different colors and dash type) at different energy allocation. The transcription energy was reduced by the factor p at (**a**) 50 %, (**b**) 25 %, and (**c**) 10 % while maintaining the total energy across transcription and translation constant.

The simulation results depicting the expression of RNA and protein production from the pT7-UTR1-HiBiT-tT7 plasmid at 5 nM, with transcription further isolated from the translation at different allocated NTP concentrations are illustrated in Figure 4.7. As depicted in Figure 4.3b, when NTPs were shared resources, there was a decrease in protein yield with decreasing membrane permeability. Conversely, upon isolating transcription, we no longer observe diminishing transcribed RNA as the permeability of RNA decreases. The total amount of NTP in the compartment determines the maximum RNA transcribed, decreasing proportionally as the concentration of NTPs decreases. Furthermore, while increasing GTP and ATP slightly affects the rate of

HiBiT production, the maximum yield of the HiBiT peptide is unperturbed due to the limitation of FD or the rate of consumption of GTP and ATP. The effect of the limited ribosomes can be observed in the inflection point observed in Figure 4.7a and Figure 4.7b most noticeably when k_{linker} is 0.001 s^{-1}.

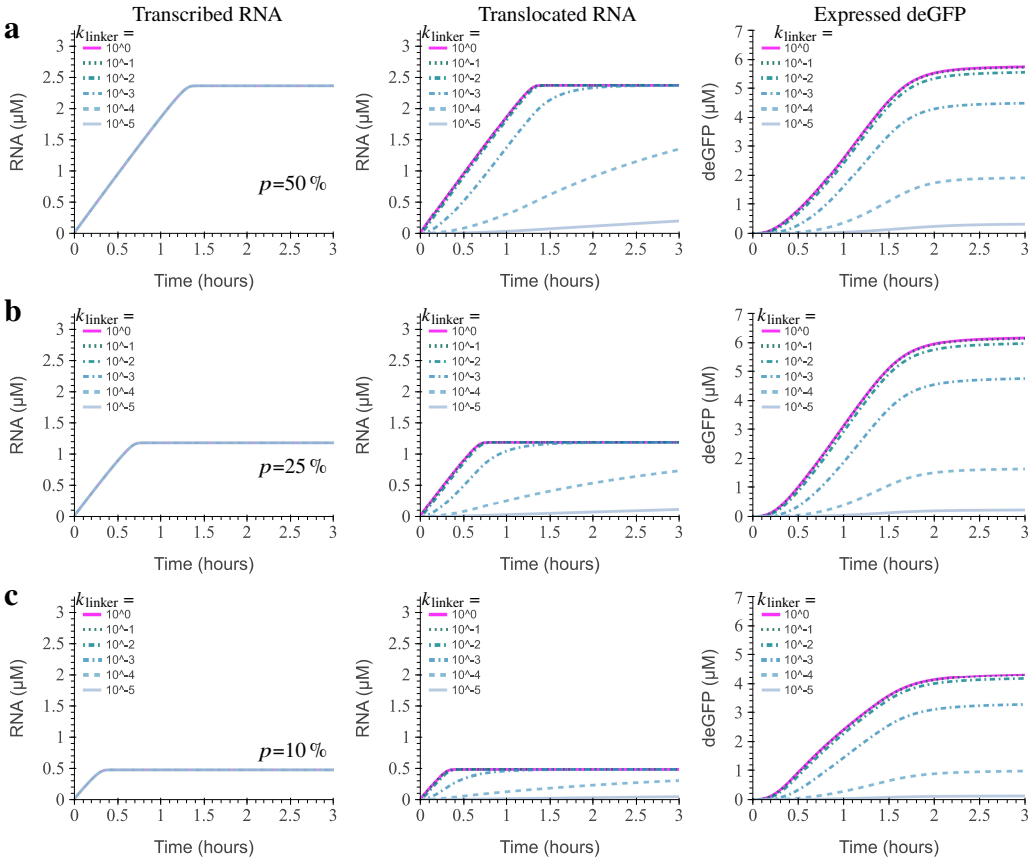

Figure 4.8: **Impact on deGFP expression in PURE model with shared resources and varying RNA permeability.** Simulation results of PURE-based cell-free protein expression of DNA constructs, T7-MGapt-UTR1-deGFP-tT7 at 5 nM with varying RNA permeability (different colors and dash type) at different energy allocation. The total energy of the transcription and translation system was kept constant as the transcription energy was reduced by the factor p at (**a**) 50 %, (**b**) 25 %, and (**c**) 10 %.

Running the numerical experiment with the larger pT7-MGapt-UTR1-deGFP-tT7 construct, as depicted in Figure 4.8, reveals both similarities and differences compared to the expression of pT7-UTR1-HiBiT-tT7. Similar to the expression of the HiBiT peptide, RNA transcription is independent of membrane permeability, while deGFP yield decreases with reduced permeability and is limited by NTP concentrations. However, the expression of pT7-MGapt-UTR1-deGFP-tT7 does not enter the ribosome-saturated regime when the RNA concentration is greater than 3 μM. As a result, an intriguing observation arises when comparing the production of deGFP

in Figure 4.8a and Figure 4.8b. We observe a trade-off between RNA and energy concentrations for deGFP, indicating that an increase in ATP and GTP can offset the potential decrease in protein production due to reduced RNA availability.

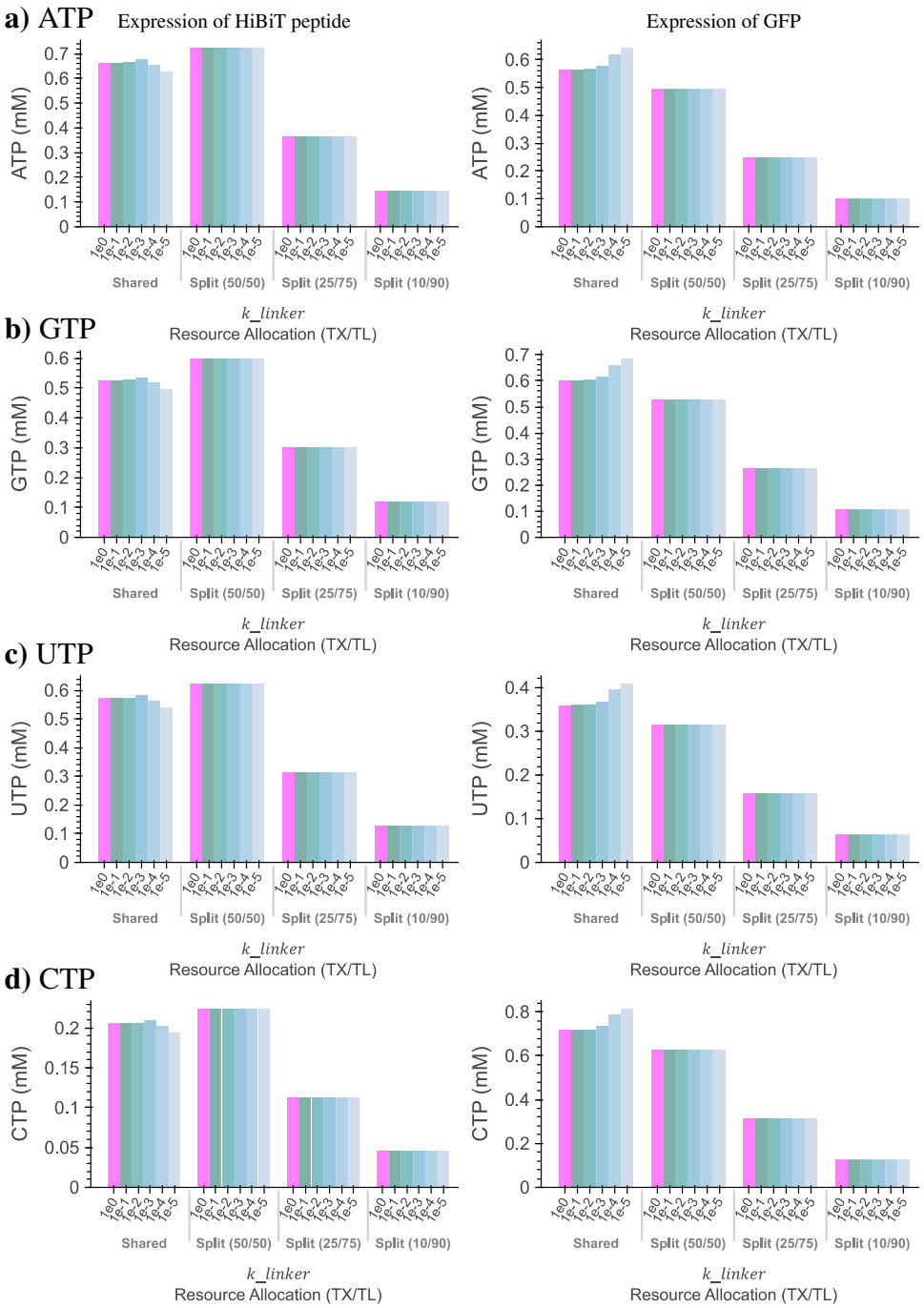

Figure 4.9: **Depletion of NTP by transcription during the expression of the HiBiT peptide and deGFP.** The depletion of (**a**) ATP, (**b**) GTP, (**c**) CTP, and (**d**) UTP in specified nucleus models with different allocations of energy between the "nucleus" and "cytoplasm" compartments.

Drawing from the findings presented in Figure 4.8 and Figure 4.7, we sought to compare energy utilization across the various membrane models under different energy allocations. To achieve this, we opted to quantify the energy consumed by transcription, recognizing that an increase in RNA production does not necessarily correlate with higher protein yields in either HiBiT peptide or deGFP expression. The energy expended on transcription was computed by calculating the difference between the start and end concentration of each NTP_{tx} in the NTP impermeable membrane model. For the model with shared resources, the difference between the starting and ending concentrations of UTP_{tx} and CTP_{tx} was calculated similarly. However, for ATP_{tx} and GTP_{tx}, the total RNA produced was multiplied by the required ATP or GTP to form the RNA strand. Although this method may underestimate the energy used for transcription, it still provides a basis for comparing energy consumption between the two processes. The results are illustrated in Figure 4.9.

In Figure 4.9, it is evident that whether simulating the expression of HiBiT peptide or deGFP, the energy consumed by the transcription process is approximately equivalent when the membrane is permeable, allowing shared NTPs, compared to when energy is evenly allocated, and the membrane is impermeable to NTPs. Furthermore, there is no notable difference in deGFP production when the system's energy is divided with 25 % allocated to the "nucleus" and 75 % to the "cytoplasm" compartments, respectively (see Figure 4.8). This suggests that in Figure 4.9, the difference in RNA produced in the membrane permeable model compared to the membrane permeable nucleus model, with energy divided between transcription and translation at 25 %and 75 % respectively, serves no apparent purpose and only competes for translation resources.

Expanding on this insight, we held the energy constant in the transcription process while adjusting the energy allocated to translation. In this simulation, we modeled the expression of deGFP with ATP and GTP at 25 %, 50 %, and 90 % of the maximum ATP and GTP of 3.75 mM and 2.5 mM respectively. The results are shown in Figure 4.10 with initial ATP and GTP values listed above each figure. The heatmap of selected transcription species and translation species for the expression of deGFP at the reduced translation energies are presented in Figure 4.11. Figure 4.10 and Figure 4.11 reveal that we can increase total protein production by increasing the initial concentrations of ATP and GTP in translation without maximizing RNA production, finding that at 90 % of maximum concentrations of ATP and GTP (Figure 4.10d and Figure 4.11d), we achieve the higher deGFP production of 6.78 µM.

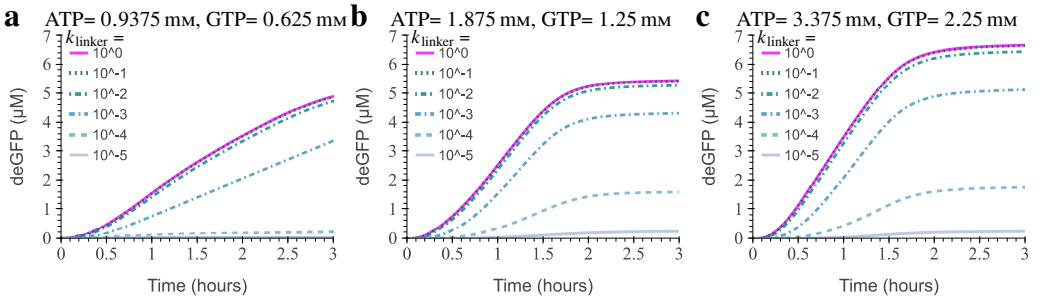

Figure 4.10: **Effect of total energy for translation on the expression of deGFP.** Simulation results of PURE-based cell-free protein expression of DNA constructs, T7-MGapt-UTR1-deGFP-tT7 at 5 nM with varying RNA permeability (different colors and dash type) at different energy allocation. The total energy between transcription and translation was not constant. While the transcription energy was reduced by 25 % for all simulations and the translation energy was (**a**) 25 %, (**b**) 50 %, and (**c**) 90 % of the original ATP and GTP concentrations of 3.75 mM and 2.5 mM, respectively.

If maximizing protein production is not the goal of the design goal, Figure 4.10 and Figure 4.11 also suggest that the lifespan of the reaction can be increased by reducing ATP and GTP concentrations. Lifespan is an important characterization of cell-free protein expression reaction, as it limits potential circuits or implementations of the cell-free protein expression system. We observe that Figure 4.10a and Figure 4.11b has the longest reaction lifespan, which can be increased further by tuning the membrane permeability. Unfortunately, the trade-off with increased lifespan is a reduction in protein yield. Translation efficiency continues to be affected by extraneous reactions constraining available energy.

Regardless, from our model, we find that by segregating transcription and translation, we can increase energy efficiency to make proteins. With transcription and translation isolated, we can design our synthetic cell to maximize the production of the HiBiT peptide while increasing energy efficiency by including ATP at 375 μM, GTP at 250 μM, UTP at 125 μM, and CTP at 125 μM in the "nucleus" compartment and ATP at 1875 μM, GTP at 1250 μM in the "cytoplasm" compartment; reducing overall energy by 40 %. Additionally, we observed in Figure 4.10b that by reducing the total energy of the synthetic cell by 25 %, our protein yield only decreases by approximately 11.5 % resulting in improved translation efficiency. These numerical experiments support the idea that establishing and isolating a "nucleus" can enhance translation efficiency and enable adjustable protein dynamics.

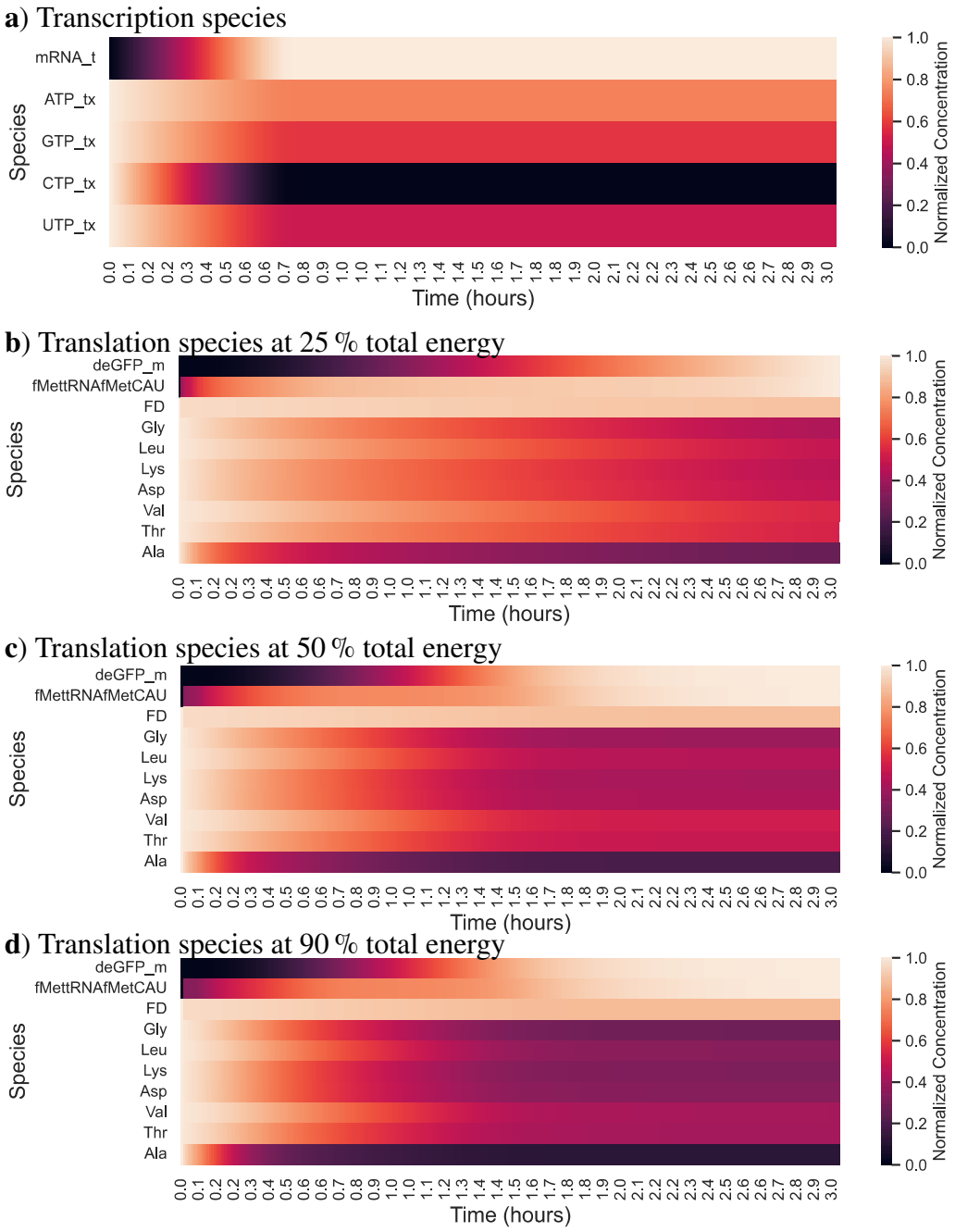

Figure 4.11: **The concentrations of NXPs and amino acids over time in the PURE model expressing the deGFP at different energies in the cytoplasm.** (**a**) Normalized concentrations of transcription species and selected translation species in deGFP at translation energies reduced to (**b**) 50 %, (**c**) 25 %, and (**d**) 10 % of the shared model and k_{linker}= 1 s^{-1}.

Utilization of resources between transcription and translation. The capability to decrease overall energy without sacrificing protein production in the HiBiT

simulations raises concerns regarding transcription's potential self-inhibition and inhibition of translation. Comparing our models regarding the ATP and GTP regeneration cycle with a "nucleus" possessing a membrane permeable to NTP or not, the reason may be evident. As depicted in Figure 4.12, PURE-based cell-free protein expression systems use the protein creatine kinase (CK) to catalyze creatine phosphate (CP) and ADP to produce creatine (Cr) and ATP. Nucleoside diphosphate kinase (NDK) then generates GTP from GDP and ATP. Both chemical reactions rely on the cycling between ADP ↔ ATP and GDP ↔ GTP, and consequently, the transcription process becomes a sink for GTP and ATP. Assuming CTP and UTP are not the limiting resources, in Figure 4.12a, we see that when transcription is not isolated from translation, all the ATP utilized in RNA strand synthesis leads to the production of pyrophosphate (PPi), as does most of the GTP. The thinner arrow indicating GDP production from transcription is specific only to the polymerase binding to the DNA and not to RNA strand elongation.

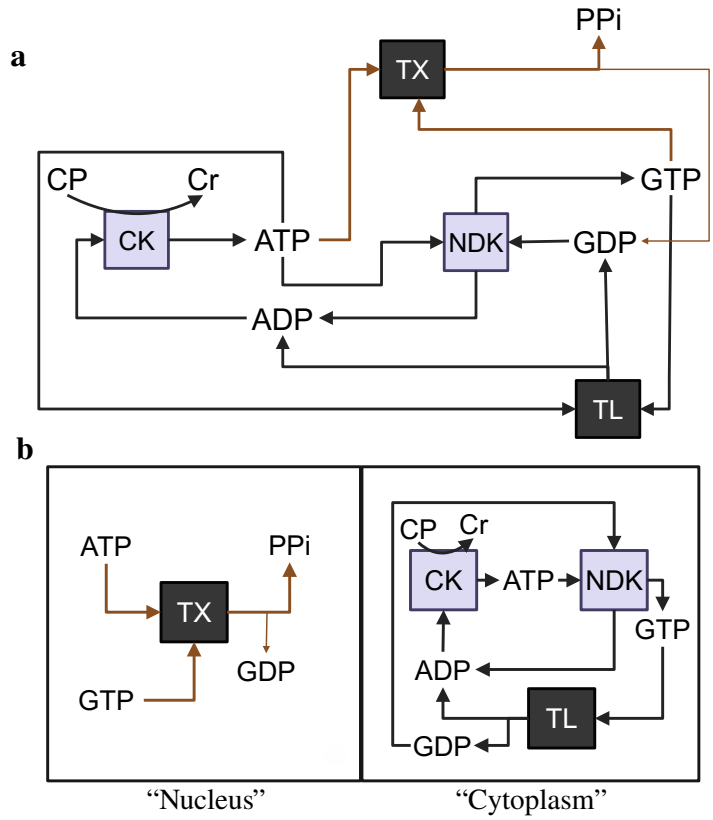

Figure 4.12: **Schematic of energy regeneration system use in PURE.** (**a**) Represents a synthetic cell where energy resources are shared or can freely defuse through the membrane, while (**b**) illustrates models energy generation when the membrane is impermeable to NTPs. Created with `BioRender.com`.

By implementing a membrane impermeable to NTPs, as depicted in Figure 4.12b, the recycling of ATP and GTP forms a closed loop, which persists until the creatine phosphate (CP) is consumed. Moreover, we can finely tune the RNA production while considering the RNA translocation rate to best suit the desired expression dynamics. Another method to achieve similar results would be to limit CTP and UTP concentrations, which are not shown. Limiting CTP and UTP would not neutralize transcription's effect on itself or translation nor enable modulation protein production as precisely as having a separate "nucleus" only permeable to RNA.

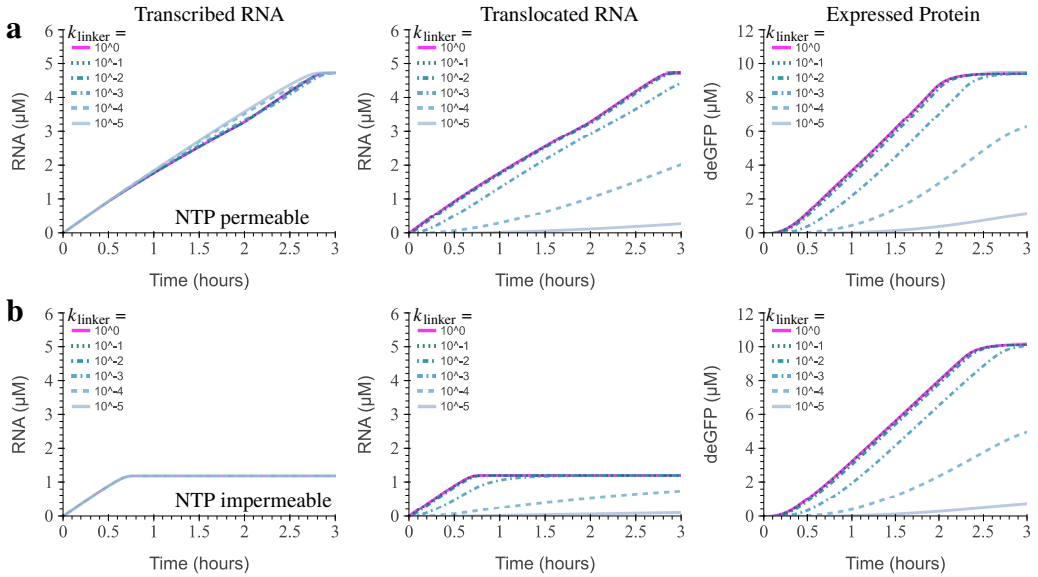

Figure 4.13: **Effect of doubling creatine phosphate (CP) on the expression of deGFP.** Simulation results of deGFP expression when "nucleus" that is (**a**) permeable or (**b**) impermeable to NTPs at different RNA translocating rates.

Expanding on the same premise, augmenting the quantity of creatine phosphate (CP) alone does not lead to higher protein expression without creating an isolated "nucleus" compartment. To underscore this, Figure 4.13 illustrates that doubling the creatine phosphate content in both nucleus models yields equivalent protein outputs, with the NTP-segregated model producing one-third of RNA. We observe that at twice the creatine phosphate initial concentration when the nucleus is NTP permeable (Figure 4.13a), RNA production is predicted to continue past the simulated time of 3 h, no longer NTP limited.

By generating heatmaps of the concentrations of selected species, as shown in Figure 4.14, we can conclude that under these conditions, deGFP production is limited by the translocation rate and the ribosome concentrations, and ultimately, it is capped by the availability of the amino acid glycine. These observations are

mirrored in the NTP impermeable "nucleus" model (Figure 4.13b), where the over-translation of deGFP is slower owing to the RNA presence. Consequently, this leads to an extended lifespan that is more perceptible at faster translocation rates.

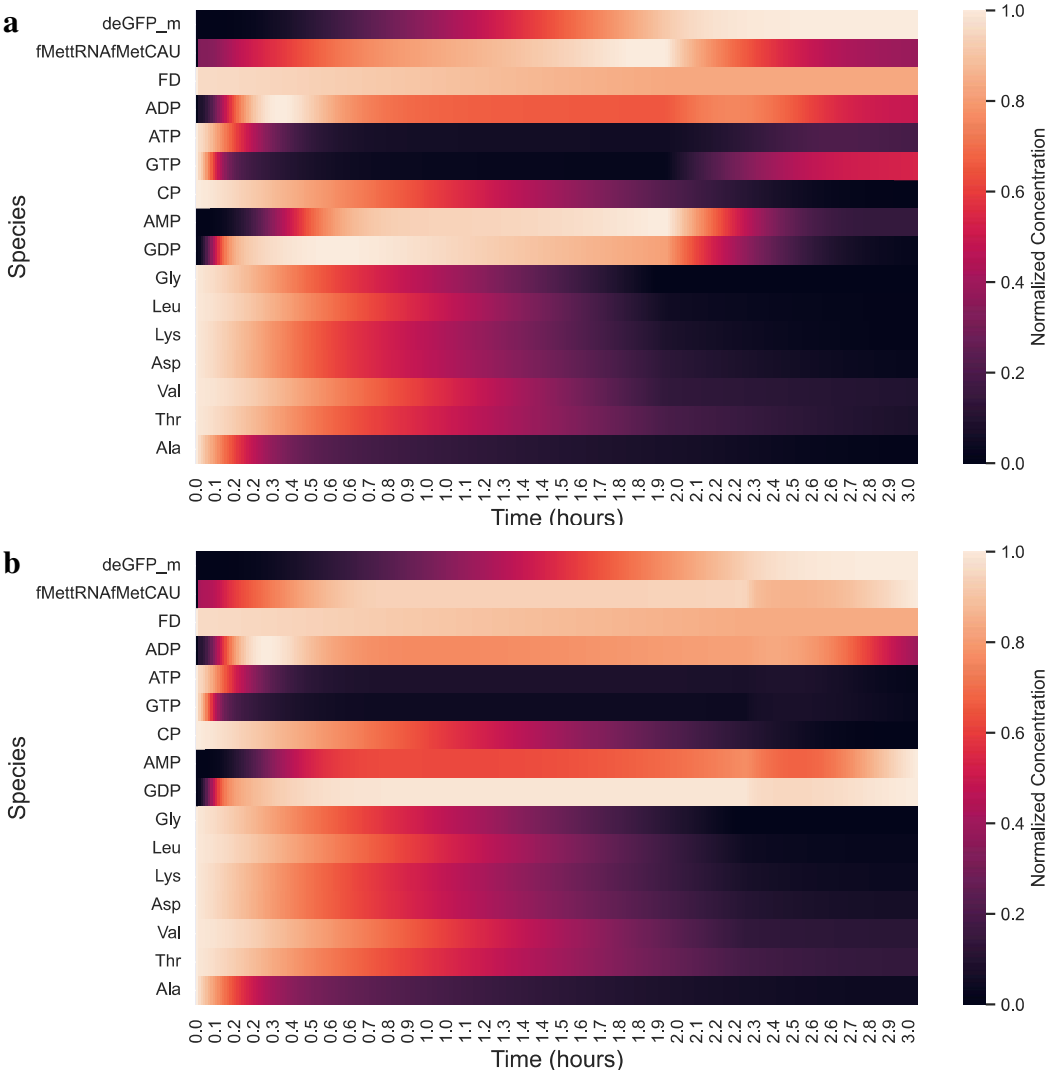

Figure 4.14: **Selected translation species concentrations of "nucleus" models expressing deGFP at double the creatine phosphate (CP) concentration.** Membrane permeable to NTPs (**a**) and membrane impermeable to NTPs (**b**) at double the original creatine phosphate (CP) concentration of 10 mM and k_{linker}= 1 s^{-1}.

For completeness, we reduced creatine phosphate concentration by half; results are plotted in Figure 4.15. Here, we see that the maximum transcribed RNA in the NTP permeable membrane, Figure 4.15a, reduces by approximately 60 % and saturates around 1 h. In Figure 4.15b, the transcribed RNA dynamics remain unchanged with the NTP impermeable membrane.

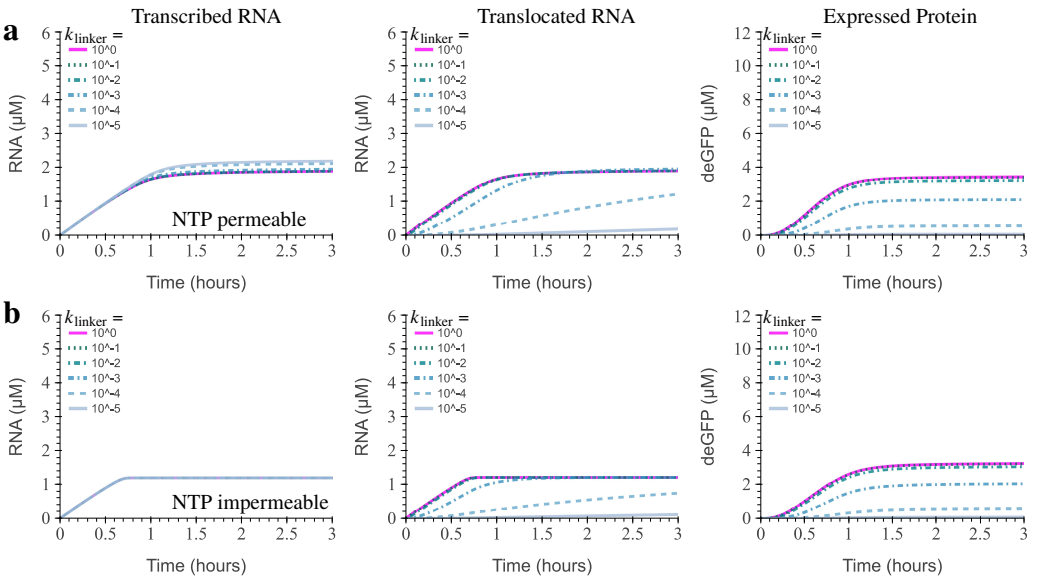

Figure 4.15: **Effect of halving creatine phosphate on the expression of deGFP.** Simulation results of deGFP expression when "nucleus" that is (**a**) permeable or (**b**) impermeable to NTPs at different RNA translocating rates.

Furthermore, the extended lifespan observed with doubled creatine phosphate concentration is no longer present, although the translation rate with the permeable membrane remains slightly faster. We can next investigate the limiting factor of deGFP production by generating heatmaps of the concentrations of selected species, as shown in Figure 4.16. We can identify creatine phosphate (CP) as the limiting factor in deGFP production, leading to the cessation of energy recycling. This is observed as deGFP production saturates as creatine phosphate is fully depleted in both membrane conditions.

We observe that the increased RNA production results in an increased depletion of the energy carriers ATP and GTP, which are necessary for the energy recycling pathway utilized in PURE. Comparing the effective removal of all NTP available for translation is demonstrated in Figure 4.17 under specified nucleus model conditions, considering NTP permeability or impermeability at specific initial creatine phosphate concentrations. Comparing the different nucleus models, it is clear that more NTPs are permanently consumed when the "nucleus" membrane is permeable to NTPs. Due to other limiting factors, we observe that granting unrestricted access to NTPs for transcription may accelerate RNA production, minimizing deGFP saturation time. This expedited process comes at an energy trade-off that is not balanced by the total protein expression.

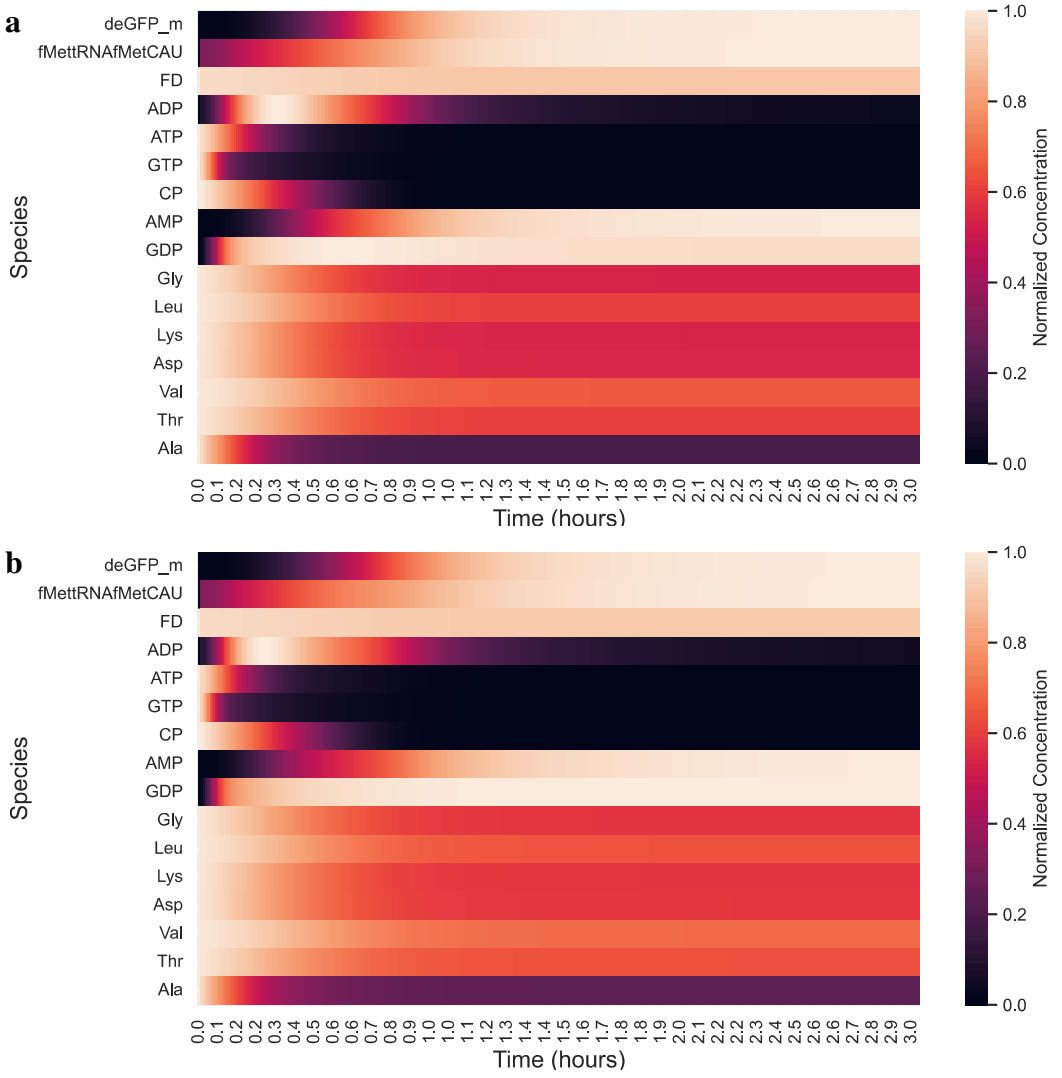

Figure 4.16: **Selected translation species concentrations of "nucleus" models expressing deGFP at various creatine phosphate concentrations.** Membrane permeable to NTPs (**a**) and membrane impermeable to NTPs (**b**) at half the original creatine phosphate concentration of 10 mM and k_{linker}= 1 s^{-1}.

Finally, as demonstrated previously, PURE reactions often yield incomplete translation. Utilizing the model and previous simulations, we can calculate the percentage of completed peptides as another metric to evaluate any advantages of incorporating a synthetic nucleus. This metric was calculated by dividing the concentrations of completed translation by the sum of incomplete peptides. The incomplete peptides can include those that terminated early, but in our model, they stall due to a lack of resources.

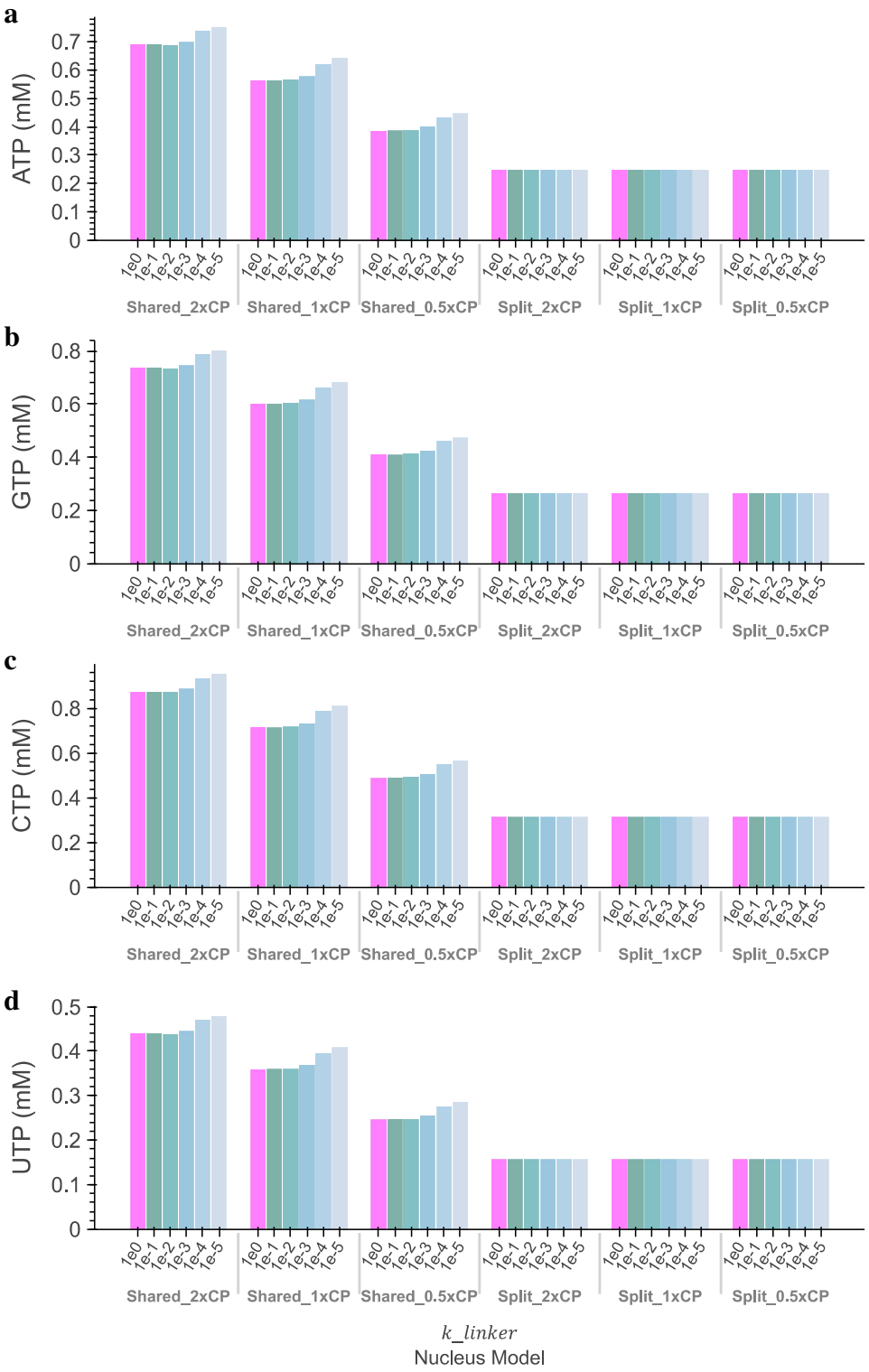

Figure 4.17: **Depletion of NTP by transcription.** The depletion of (**a**) ATP, (**b**) GTP, (**c**) CTP, and (**d**) UTP in specified nucleus models.

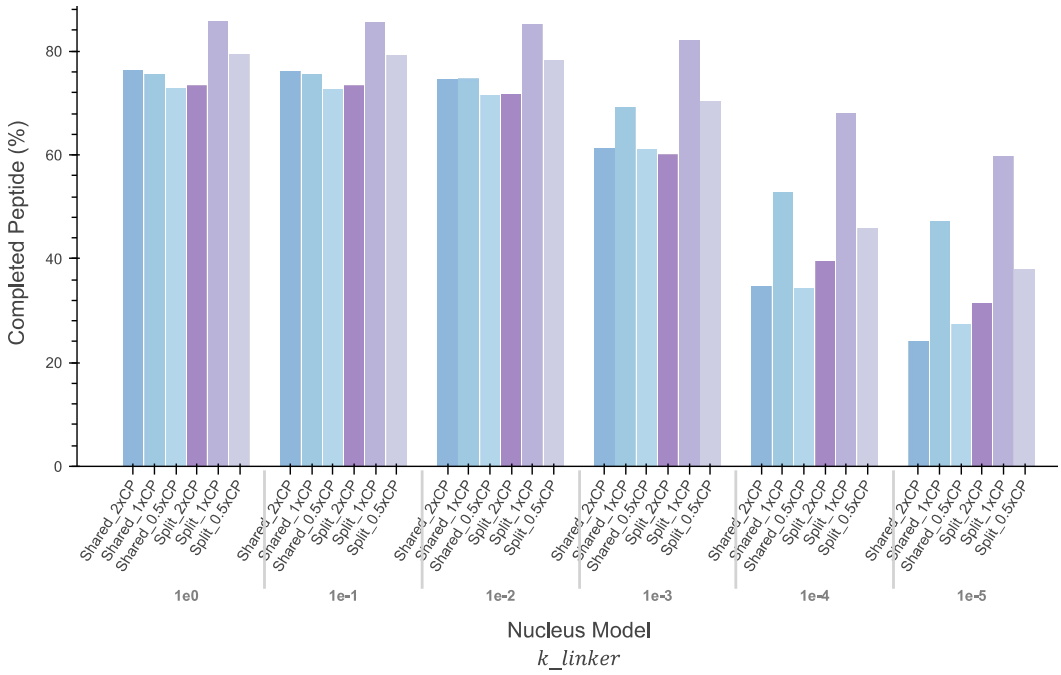

Figure 4.18: **The percentage of complete deGFP translation.** The percentage is calculated by taking the matured deGFP protein and deciding over the sum of remaining incomplete peptides at the end of the simulation in specified nucleus models.

In Figure 4.18, we see that across all model conditions, the fully segregated "nucleus" was more effective at producing fully translated protein when compared to its counterpart at the same creatine phosphate concentration. Furthermore, we more clearly see that increasing initial creatine phosphate concentration only yields more fully translated proteins when translocation rates are above $0.1\,\mathrm{s}^{-1}$ otherwise, translation must compete with the other reactions for resources.

4.3 Conclusion

In this chapter, we have leveraged models from Chapter 3 to explore the advantages and limitations of separating transcription and translation. We examined two separation conditions dictated by whether or not the membrane separating transcription from translation is permeable to NTPs. Additionally, we explored the effects of RNA translocation rates on RNA and protein production in both separation conditions. We further investigated how the incorporation of "nucleus" would impact the expression of two DNA constructs: pT7-MGapt-UTR1-deGFP-tT7 and pT7-UTR1-HiBiT-tT7, both at 5 nM. All models and results from the numerical experiments presented in this chapter are documented and accessible on GitHub [83].

These numerical experiments have uncovered evidence supporting an energy trade-off between transcription and translation. We observed that when the membrane is permeable to NTPs, inhibiting RNA translocation may allow for a greater allocation of energy to transcription, albeit up to a certain threshold. We noted a reduction in transcribed RNA for smaller transcripts as membrane permeability decreased, which may result from transcription self-inhibition. These models show that transcription may compete for energy resources against translation and itself. However, the internal resources allocation and adverse effects of transcription on RNA production and translation can be mitigated by limiting the initial concentration of CTP and UTP or negated by implementing a membrane impermeable to NTPs. Although a nuclear pore exclusively permeable to RNA seems conceptually implausible, in the current context of synthetic cell development with an artificial nucleus, this represents our current capability, given the absence of nuclear pore proteins or the formation of channels facilitating passive NTP transport.

Our results support that the complete separation of transcription and translation would increase protein production due to our energy generation system, which depends on limited creatine phosphate. The numerical experiments show that increasing creatine phosphate concentration does not constantly produce higher protein expression when shared or segregated resources. However, we can exclusively increase protein production in the NTP impermeable membrane condition by increasing creatine phosphate concentrations or energy allocation to the "cytoplasm" resulting in a more energy-efficient cell. By explicitly isolating transcription, we can also improve the longevity of the energy carriers GTP and ATP required to maximize protein production and/or the length of the lifespan of the reaction. Ultimately, we have found that adjusting one condition to optimize specific parameters often comes at a cost to another aspect of the system. The cost-benefit analysis ultimately lies in the hands of the cell engineers and depends on the system's intended purpose.

Chapter 5

CONCLUSIONS & FUTURE WORK

5.1 Conclusions

Cell-free expression systems provide a method for rapid DNA circuit prototyping and functional protein synthesis. While crude extracts remain a black box with many components carrying out unknown reactions, the PURE system contains only the required transcription and translation components for protein production. Theoretically, all proteins and small molecules are at known concentrations, allowing detailed modeling for reliable computational predictions. We utilized our knowledge and experience with cell-free systems to shed light on the transcription and translation processes and build a chemical reaction network to capture protein expression in PURE. Through our models, we were driven to explain phenomena not yet described in the literature.

In this thesis, we have demonstrated that protein production is not always directly correlated with RNA production due to the competition for energy between transcription and translation. Our models predict that transcription competes with translation for resources, depleting the total energy carriers needed for energy generation. However, faster RNA production can lead to higher protein expression, as translation also competes for resources against auxiliary reactions not directly leading to protein production. Although our models are not all-encompassing, they represent a step toward developing a synthetic nucleus, showing that we can build and leverage models to understand the relationship between transcription and translation in a cell-free protein expression system.

5.2 Potential PURE Model Improvements

Our model and conclusion position us on the verge of new insights, ready to be explored. Unfortunately, without the feasibility of easily adjusting the concentration of any energy carriers in the commercial PURE or chemically separating transcription from translation, we are prevented from physically engineering a synthetic nucleus. Our models and findings underscore the need for dependable, reproducible, and customizable PURE-based cell-free protein reaction systems. As such, systems are crucial for advancing our understanding and enabling experimental setups so our models can evolve and our comprehension of cellular processes can expand.

To continue improving and advancing our understanding of cell-free systems and to approach the complexity of a cellular nucleus, we propose the following:

1. Characterizing the effects of crowding, small molecules, proteins, temperature, and localization on transcription and translation.

2. Transitioning to the use of non-commercial PURE systems. Homemade PURE or OnePot grants the flexibility to modulate protein concentration and salt levels systematically, facilitating the exploration of various experimental conditions. Furthermore, the accurate compositions of the system can be measured and reported.

3. Incorporating of DNA replication. One of the many attributes of BioCRNpyler is the ease of creating chemical reaction networks that can be combined. Slowly, we can build and compile the model cellular mechanism, taking steps towards a virtual synthetic cell that can better inform us how to engineer one physically.

4. Creation of a PURE-based cell-free protein expression system data repository. Running all possible permutations of PURE would be time-consuming and costly; with so many groups studying cell-free protein expression systems to build a cell, creating a repository will allow for better modeling and understanding.

5.3 Unactualized Experiments

In our pursuit of exploring the synthetic nucleus's potential benefits, we initially focused on creating a synthetic nucleus. However, physically isolating transcription from translation while allowing RNA translocation proved more challenging than anticipated. Initially, we utilized *E. coli* Rosetta2 lysate-based cell-free protein expression systems and encountered minor hurdles, such as isolating transcription and translation reactions. These were addressed by employing DNA constructs driven by a T7 promoter. The *E. coli* Rosetta2 lysate-based cell-free protein expressions lacked native T7 RNAP, enabling us to use a commercial or in-house transcription system for transcription-only reactions and generate purified RNA for translation-only reactions. The translation-only reaction could be achieved by omitting DNA and adding purified T7 RNAP, resulting in separate and controllable transcription and translation reactions.

The major obstacle arose when attempting to compartmentalize the transcription and the translation reaction. Trying to restrict the diffusion to RNA only or RNA and other small molecules, as in our synthetic nucleus models in Chapter 4, between the "nucleus" and the "cytoplasm" compartment was not straightforward. Constrained by our capability to create vesicles and transport RNA between the compartments, we investigated two experimental setups to isolate transcription from translation: temporal separation using an Echo 525 Acoustic Liquid Handler to add the transcription to the translation reactions sequentially and spatial separation using dialysis cassettes of different sizes in a 24-well plate.

Temporal separation using an Echo 525 Acoustic Liquid Handler

Since the chemical separation of transcription and translation was not feasible, we opted for physical separation. We achieved this by running the transcription reaction initially and then adding the products to the crude cell lysate as our translation reaction. In this experiment, illustrated in Figure 5.1, we ran an *in vitro* transcription reaction (IVT_x) using a transcription mix made in-house. The transcription reaction was initiated at different times(t=−5 h, −4 h, −2 h, and 0 h) so that t=0 would correspond to inoculation of the translation reaction with the transcription reaction products. At each respective start times two 50 µL transcription reactions were freshly mixed, one containing 5 nM of DNA expressing deGFP and incubated at 37 °C.

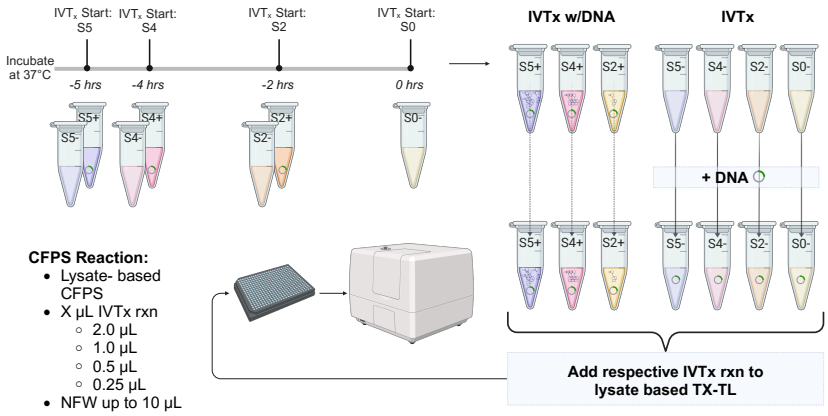

Figure 5.1: **Experimental schematic of deGFP expression with temporal separation of transcription and translation.** Overall experiment setup where the transcription reaction was run initially in separate tubes with and without DNA expressing deGFP. After variable incubation times at 37 °C, DNA was added to experimental samples that did not initially contain DNA. Next, the transcription reaction was added to lysate-based TX-TL at various volumes. The expression of deGFP was measured in a BioTek over 12 h at 29 °C. Created with `BioRender.com`.

Following incubation, at t=0 h, 5 nM of DNA expressing deGFP was added to the samples that did not already contain DNA. Next, using an Echo 525 Acoustic Liquid Handler, the entirety of the transcription was added to 7.5 µL lysate-based cell-free protein expression system in volumes of 0.25 µL, 0.5 µL, 1.0 µL,and 2.0 µL. Nucleus-free water (NFW) was then added to bring the total reaction volume to 10 µL. The expression of deGFP was measured using the BioTek at 29 °C, 485/515 nm (ex/em), and gain of 61 in triplicates. Figure 5.2 depicts our prediction and the final results. We predicted that in the control samples (see Figure 5.2a), where the transcription reaction was incubated without DNA, the total deGFP produced would remain unaffected regardless of the incubation time, or the amount of the transcription reaction added. However, in the samples incubated with DNA (see Figure 5.2b), we anticipated that longer incubation times would yield higher RNA concentrations, leading to increased protein production. Similarly, we predicted that adding larger volumes of the transcription reaction would enhance protein production.

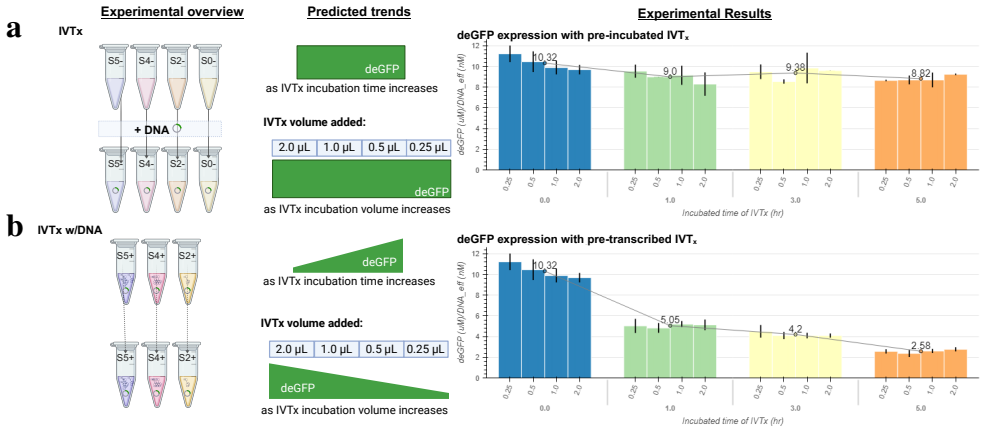

Figure 5.2: **Prediction and results of deGFP expression with temporal separation of transcription and translation.** Cartoon of the experimental conditions tested, alongside the predicted trends and results for **(a)** control samples, where the transcription reaction was incubated without DNA, and **(b)** test samples, where the transcription reaction was incubated with DNA. The deGFP concentration was normalized to the effective DNA (DNA_{eff}) in each well, and the results for incubation time of t= 0 h are repeated in (a) and (b). Created with `BioRender.com`.

Due to the varying amounts of DNA added to each test sample, the deGFP concentration was normalized to the effective DNA (DNA_{eff}) in each well. In the control samples, Figure 5.2a shows that there was an insignificant difference between all controls, with an average ratio of deGFP concentration (µM) to effective DNA (nM)

of $9.355\,\mu M\,nM^{-1}$ and a standard deviation of $0.48\,\mu M\,nM^{-1}$. Unexpectedly, the experimental samples in Figure 5.2b, contradict our prediction, showing that as incubation time increases, the ratio of deGFP concentration (μM) to effective DNA (nM) decreases and is unaffected by the volume of the transcription reaction added. However, the results from the control samples show that incubating the transcription mix alone was insufficient to affect deGFP expression. Therefore, the reduction in deGFP production suggests that transcription products may negatively affect translation. Waste's inhibition on translation has been previously reported in literature [71] but primarily focuses on metabolic trade-offs in cell-lysate cell-free protein synthesis systems [51, 71, 94] and not explicitly a trade-off between transcription and translation due to waste produced from transcription. The results suggested the possibility of inhibition due to waste accumulation from transcription, underscoring the potential advantages of the nucleus.

Nonetheless, an experimental setup that functionally isolates transcription from translation is necessary to investigate further the effects of the trade-off between transcription and translation. The temporal separation of transcription and translation is insufficient because DNA and T7 RNAP remain present in the final translation reaction. Distinguishing protein translated from previously transcribed RNA or directly from the DNA becomes impossible. The DNA and T7 RNAP must be contained away from the lysate-based cell-free protein synthesis system to guarantee that protein production solely stems from RNA.

Spatial separation using dialysis cassettes

To ensure translation exclusively relied on RNA produced in the transcription reaction rather than the DNA present and to prevent simultaneous transcription and translation, we transitioned to using dialysis cassettes as the "nucleus" in our synthetic nucleus experiments. The dialysis cassettes (0.5 μL Slide-A-Lyzer®TM MINI Dialysis Devices) from Thermo Scientific®TM were purchased in various sizes: 3.5K MWCO, 10K MWCO, 7K MWCO, and 20K MWCO and placed in a 24-well plate. Each dialysis cassette's top, just above the edge of the cassette, was trimmed using a razor. Subsequently, the cassettes were inserted into a 3D-printed holder, preventing the bottom from touching the plate's bottom depicted in Figure 5.3a.

Before inserting the dialysis cassette into the well, we added 300 μL of the translation reaction to the bottom of the 24-well plate. This experiment used crude extract (TX-TL) externally; due to the volume required to cover the bottom of the well, using any

PURE system would not have been financially feasible. Then 50 µL of the *in vitro* transcription reaction, containing T7 RNAP, NTPs, and salts, was added inside the dialysis cassettes along with DNA expressing the HiBiT and placed into the well. The HiBiT peptide, as previously mentioned, is a part of the Promega system [93]. In our setup, illustrated in Figure 5.3b, it was postulated that the DNA-HiBiT would produce mRNA-HiBiT and diffuse through the dialysis membrane at different rates. Upon reaching the other side, mRNA-HiBiT would be translated into a peptide that binds to an LgBiT protein, forming an enzyme. This enzyme would catalyze the supplied substrate, leading to luminescence detected by the BioTek reader, enabling measurement of total protein expression in the separated system.

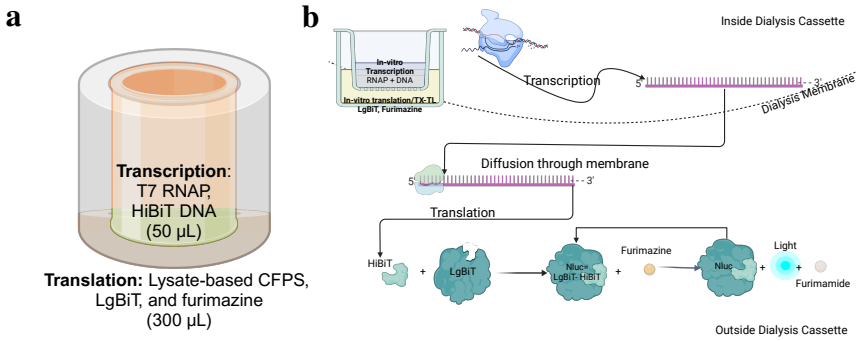

Figure 5.3: **Illustration of the experiment separating transcription from translation using the Nano-Glo® HiBiT system. (a)** An illustration depicting the suspended dialysis cassette in a 24-well plate, along with the corresponding compartments and their contents. **(b)** Cartoon of the transcription, translation, and enzymatic process of the HiBiT system. The transcription reaction is added to the inner compartment, and the translation mix is added to the outer compartment. The inner compartment transcribes DNA to mRNA. Then, RNA diffuses through the dialysis membrane pores and is translated into a peptide. The HiBiT peptide binds to the LgBiT protein in the outer compartment to complete an enzyme that consumes a substrate, creating luminescence. Created with `BioRender.com`.

We aimed to investigate the impact of dialysis size on RNA transport by evaluating protein expression dynamics and total protein yield through luminescence reading. Notably, the molecular weight of plasmid DNA and ribosomes is at least 10^3 greater than any dialysis cassettes used. Unfortunately, after multiple attempts, we found that achieving reproducible results was challenging: the bleed-through of luminescence to neighboring wells limited the number of replicates and conditions that could be run on one experimental run. Additionally, luminescence dynamics made quantifying total protein production arduous due to the multiple conditions needing calibration.

Looking into alternative systems, we moved to a split-GFP system (GFP1-10 and GFP11) [95]. Similar to the HiBiT system, the smaller subunit, GFP11, would be transcribed inside the cassette, next the mRNA-GFP11 would diffuse outside the cassette and be translated. The GFP11 peptide would bind with GFP1-10, purified separately, to form a completed GFP protein. The split-GFP system provides a stable and one-to-one protein qualification and removes any unknown components that come from using a commercial kit. However promising the split-GFP system appeared, we found that the purification of the GFP1-10 protein was not reliably reproducible due to the formation of inclusion bodies during the protein purification process. We also concluded that relying on crude extract would not allow the complete separation of transcription and translation. However, shifting to a fully in vitro commercial system would not be feasible due to the required volume and associated cost.

5.4 The Future of a Synthetic Nucleus

To continue our goals of constructing a cell and deepening our comprehension of cellular processes, we must prioritize the development of reliable, reproducible, and accessible systems. During the initial years of synthetic cell research, I was struck by the challenges I encountered in vesicle formation or the fickleness of lysate-based cell-free protein expression. While it is possible to separate transcription components from translation components using One-Pot PURE or a combination of T7-driven DNA constructs and *E. coli* Rosetta2 crude lysate, as we have demonstrated, a robust experimental system is still needed. One promising tool would be the implementation and expanded use of cell-penetrating peptides [23] to controllable translocate RNA from the "nucleus" compartment to the "cytoplasm." The utilization of cell-penetrating peptides still relies on consistent vesicle formation within the broader synthetic cell community, particularly the creation of nested vesicles. Demonstrating that constructing a synthetic cell, and even more so, a more complex synthetic cell with eukaryotic traits, necessitates building upon a strong foundation of constructing liposomes, cell-free protein synthesis systems, and tools for studying RNA and protein production. These foundational requirements must be fulfilled without prolonged troubleshooting for synthetic cells to realize their full potential. This approach should be universal and foster collaboration; otherwise, we only inhibit our goals.

BIBLIOGRAPHY

[1] Stein, Valentin et al. "Building synthetic cell–from the technology infrastructure to cellular entities." *ACS Synthetic Biology* 2.6 (2013), pp. 324–336.

[2] Sandberg, Troy E. et al. "Adaptive evolution of a minimal organism with a synthetic genome." *Iscience* 26.9 (2023).

[3] Moger-Reischer, Roy Z. et al. "Evolution of a minimal cell." *Nature* 620.7972 (2023), pp. 122–127.

[4] Gaut, Nathaniel J. and Adamala, Katarzyna P. "Reconstituting natural cell elements in synthetic cells." *Advanced Biology* 5.3 (2021), p. 2000188.

[5] Garenne, David et al. "Cell-free gene expression." *Nature Reviews Methods Primers* 1.1 (2021), p. 49.

[6] Nirenberg, Marshall W. and Matthaei, J. Heinrich. "The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides." *Proceedings of the National Academy of Sciences* 47.10 (1961), pp. 1588–1602.

[7] Sun, Zachary Z. et al. "Protocols for implementing an *Escherichia coli* based TX-TL cell-free expression system for synthetic biology." *Journal of Visualized Experiments* (2013), e50762. ISSN: 1940-087X. DOI: 10.3791/50762.

[8] Levine, Max Z. et al. "*Escherichia coli*-based cell-free protein synthesis: Protocols for a robust, flexible, and accessible platform technology." *Journal of Visualized Experiments* (2019), e58882.

[9] Cole, Stephanie D. et al. "Quantification of interlaboratory cell-free protein synthesis variability." *ACS Synthetic Biology* 8.9 (2019), pp. 2080–2091.

[10] Kung, Hsiang-Fu et al. "DNA-directed *in vitro* synthesis of beta-galactosidase. Studies with purified factors." *Journal of Biological Chemistry* 252.19 (1977), pp. 6889–6894.

[11] Ganoza, M. Clelia, Cunningham, Christina, and Green, Robert M. "Isolation and point of action of a factor from *Escherichia coli* required to reconstruct translation." *Proceedings of the National Academy of Sciences* 82.6 (1985), pp. 1648–1652.

[12] Pavlov, Michael Yu. and Ehrenberg, Måns. "Rate of translation of natural mRNAs in an optimized *in vitro* system." *Archives of Biochemistry and Biophysics* 328.1 (1996), pp. 9–16.

[13] Pavlov, Michael Yu et al. "Release factor RF3 abolishes competition between release factor RF1 and ribosome recycling factor (RRF) for a ribosome binding site." *Journal of Molecular Biology* 273.2 (1997), pp. 389–401.

[14] Shimizu, Yoshihiro et al. “Cell-free translation reconstituted with purified components.” *Nature Biotechnology* 19.8 (2001), pp. 751–755. ISSN: 1546-1696. DOI: 10.1038/90802.

[15] Lavickova, Barbora and Maerkl, Sebastian J. “A simple, robust, and low-cost method to produce the PURE cell-free system.” *ACS Synthetic Biology* 8.2 (2019), pp. 455–462. DOI: 10.1021/acssynbio.8b00427.

[16] Grasemann, Laura et al. “OnePot PURE cell-free system.” *Journal of Visualized Experiments* 172 (2021), e62625.

[17] Biemann, K. “Mass spectrometry of peptides and proteins.” *Annual Review of Biochemistry* 61.1 (1992), pp. 977–1010.

[18] Adamala, Katarzyna P. et al. “Engineering genetic circuit interactions within and between synthetic minimal cells.” *Nature Chemistry* 9.5 (2017), pp. 431–439.

[19] Yandrapalli, Naresh et al. “Surfactant-free production of biomimetic giant unilamellar vesicles using PDMS-based microfluidics.” *Communications Chemistry* 4.1 (2021), p. 100.

[20] Li, Jun et al. “Improved cell-free RNA and protein synthesis system.” *PLoS One* 9.9 (2014), e106232.

[21] Gispert, Ignacio et al. “Stimuli-responsive vesicles as distributed artificial organelles for bacterial activation.” *Proceedings of the National Academy of Sciences* 119.42 (2022), e2206563119.

[22] Peruzzi, Justin A. et al. “Barcoding biological reactions with DNA-functionalized vesicles.” *Angewandte Chemie* 131.51 (2019), pp. 18856–18863.

[23] Heili, Joseph M. et al. “Controlled exchange of protein and nucleic acid signals from and between synthetic minimal cells.” *Cell Systems* 15.1 (2024), pp. 49–62.

[24] Siegal-Gaskins, Dan et al. “Gene circuit performance characterization and resource usage in a cell-free ‘breadboard.’” *ACS Synthetic Biology* 3.6 (2014), pp. 416–425. DOI: 10.1021/sb400203p.

[25] Marshall, Ryan and Noireaux, Vincent. “Quantitative modeling of transcription and translation of an all-*E. coli* cell-free system.” *Scientific reports* 9.1 (2019), p. 11980.

[26] Jay, Daniel G. and Keshishian, Haig. “Laser inactivation of fasciclin I disrupts axon adhesion of grasshopper pioneer neurons.” *Nature* 348.6301 (1990), pp. 548–550.

[27] Liao, Joseph C., Roider, Johann, and Jay, Daniel G. “Chromophore-assisted laser inactivation of proteins is mediated by the photogeneration of free radicals.” *Proceedings of the National Academy of Sciences* 91.7 (1994), pp. 2659–2663.

[28] Ellington, Andrew D. and Szostak, Jack W. "*In vitro* selection of RNA molecules that bind specific ligands." *Nature* 346.6287 (1990), pp. 818–822.

[29] Tuerk, Craig and Gold, Larry. "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." *Science* 249.4968 (1990), pp. 505–510.

[30] Robertson, Debra L. and Joyce, Gerald F. "Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA." *Nature* 344.6265 (1990), pp. 467–468.

[31] Grate, Dilara and Wilson, Charles. "Laser-mediated, site-specific inactivation of RNA transcripts." *Proceedings of the National Academy of Sciences* 96.11 (1999), pp. 6131–6136.

[32] Foster, Fred J. and Woodbury, Lowell. "The use of malachite green as a fish fungicide and antiseptic." *The Progressive Fish-Culturist* 3.18 (1936), pp. 7–9.

[33] Plakas, Steven M., Doerge, Daniel R., and Turnipseed, Sherri B. "Disposition and metabolism of malachite green and other therapeutic dyes in fish." *Xenobiotics in Fish* (1999), p. 149.

[34] Babendure, Jeremy R., Adams, Stephen R., and Tsien, Roger Y. "Aptamers switch on fluorescence of triphenylmethane dyes." *Journal of the American Chemical Society* 125.48 (2003), pp. 14716–14717.

[35] Baugh, Christopher, Grate, Dilârâ, and Wilson, Charles. "2.8 Å crystal structure of the malachite green aptamer." *Journal of Molecular Biology* 301.1 (2000), pp. 117–128.

[36] Flinders, Jeremy et al. "Recognition of planar and nonplanar ligands in the malachite green–RNA aptamer complex." *ChemBioChem* 5.1 (2004), pp. 62–72.

[37] Nguyen, Dat H. et al. "Binding to an RNA aptamer changes the charge distribution and conformation of malachite green." *Journal of the American Chemical Society* 124.50 (2002), pp. 15081–15084.

[38] Da Costa, Jason B. and Dieckmann, Thorsten. "Entropy and Mg^{2+} control ligand affinity and specificity in the malachite green binding RNA aptamer." *Molecular BioSystems* 7.7 (2011), pp. 2156–2163.

[39] Da Costa, Jason B., Andreiev, Aurelia I., and Thorsten, Dieckmann. "Thermodynamics and kinetics of adaptive binding in the malachite green RNA aptamer." *Biochemistry* 52.38 (2013), pp. 6575–6583.

[40] Nguyen, Dat H. et al. "Dynamics studies of a malachite green–RNA complex revealing the origin of the red-shift and energetic contributions of stacking interactions." *The Journal of Physical Chemistry B* 108.4 (2004), pp. 1279–1286.

[41] Poole, William et al. "BioCRNpyler: Compiling chemical reaction networks from biomolecular parts in diverse contexts." *PLOS Computational Biology* 18.4 (2022), e1009987.

[42] Hucka, Michael et al. "The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models." *Bioinformatics* 19.4 (2003), pp. 524–531.

[43] Pandey, Ayush et al. "Fast and flexible simulation and parameter estimation for synthetic biology using bioscrape." *Journal of Open Source Software* 8.83 (2023), p. 5057.

[44] Del Vecchio, Domitilla and Murray, Richard M. *Biomolecular feedback systems.* Princeton University Press Princeton, NJ, 2015.

[45] Foreman-Mackey, Daniel et al. "emcee: The MCMC hammer." *Publications of the Astronomical Society of the Pacific* 125.925 (2013), p. 306.

[46] Matsuura, Tomoaki et al. "Reaction dynamics analysis of a reconstituted *Escherichia coli* protein translation system by computational modeling." *Proceedings of the National Academy of Sciences* 114.8 (2017), E1336–E1344. DOI: 10.1073/pnas.1615351114.

[47] Matsuura, Tomoaki, Hosoda, Kazufumi, and Shimizu, Yoshihiro. "Robustness of a reconstituted *Escherichia coli* protein translation system analyzed by computational modeling." *ACS Synthetic Biology* 7.8 (2018), pp. 1964–1972.

[48] Jurado, Zoila, Pandey, Ayush, and Murray, Richard M. "A chemical reaction network model of PURE." *bioRxiv* (2023). DOI: 10.1101/2023.08.14.553301.

[49] Grate, Dilara and Wilson, Charles. "Inducible regulation of the *S. cerevisiae* cell cycle mediated by an RNA aptamer–ligand complex." *Bioorganic & Medicinal Chemistry* 9.10 (2001), pp. 2565–2570.

[50] Franco, Elisa et al. "Timing molecular motion and production with a synthetic transcriptional clock." *Proceedings of the National Academy of Sciences* 108.40 (2011), E784–E793.

[51] Kapasiawala, Manisha and Murray, Richard M. "Metabolic perturbations to an *E. coli*-based cell-free system reveal a trade-off between transcription and translation." *bioRxiv* (2024).

[52] Conrad, Richard and Ellington, Andrew D. "Detecting immobilized protein kinase C isozymes with RNA aptamers." *Analytical Biochemistry* 242.2 (1996), pp. 261–265.

[53] Proske, Daniela et al. "A Y2 receptor mimetic aptamer directed against neuropeptide Y." *Journal of Biological Chemistry* 277.13 (2002), pp. 11416–11422.

[54] Carothers, James M. et al. “Selecting RNA aptamers for synthetic biology: Investigating magnesium dependence and predicting binding affinity.” *Nucleic Acids Research* 38.8 (2010), pp. 2736–2747.

[55] Stead, Sara L. et al. “An RNA-aptamer-based assay for the detection and analysis of malachite green and leucomalachite green residues in fish tissue.” *Analytical Chemistry* 82.7 (2010), pp. 2652–2660.

[56] Sokoloski, Joshua E., Dombrowski, Sarah E., and Bevilacqua, Philip C. “Thermodynamics of ligand binding to a heterogeneous RNA population in the malachite green aptamer.” *Biochemistry* 51.1 (2012), pp. 565–572.

[57] Pan, Tao et al. “Study on decolorization of triphenylmethane dyes by DTT.” *Huan Jing ke Xue= Huanjing Kexue* 33.3 (2012), pp. 866–870.

[58] Zhou, Yubin et al. “Organic additives stabilize RNA aptamer binding of malachite green.” *Talanta* 160 (2016), pp. 172–182.

[59] New England BioLabs. *PURExpress®* In Vitro *Protein Synthesis.* Ipswich, MA, 2020.

[60] PURExfrex. *PURExfrex® 1.0 Cell-free Protein Synthesis Kit.* Kashiwa, Chiba, Japan, 2024.

[61] Kazuta, Yasuaki et al. “Synthesis of milligram quantities of proteins using a reconstituted in vitro protein synthesis system.” *Journal of Bioscience and Bioengineering* 118.5 (2014), pp. 554–557. ISSN: 1389-1723. DOI: https://doi.org/10.1016/j.jbiosc.2014.04.019.

[62] Getz, Elise Burmeister et al. “A comparison between the sulfhydryl reductants tris(2-carboxyethyl)phosphine and dithiothreitol for use in protein biochemistry.” *Analytical Biochemistry* 273.1 (1999), pp. 73–80. DOI: https://doi.org/10.1006/abio.1999.4203.

[63] Han, Joan Christine and Han, Grace Yang. “A procedure for quantitative determination of tris (2-carboxyethyl) phosphine, an odorless reducing agent more stable and effective than dithiothreitol.” *Analytical Biochemistry* 220.1 (1994), pp. 5–10.

[64] Zadeh, Joseph N. et al. “NUPACK: Analysis and design of nucleic acid systems.” *Journal of Computational Chemistry* 32.1 (2011), pp. 170–173.

[65] Pandey, Ayush et al. “Characterization of integrase and excisionase activity in a cell-free protein expression system using a modeling and analysis pipeline.” *ACS Synthetic Biology* 12.2 (2023), pp. 511–523.

[66] Jurado, Zoila. *Source code for MGapt and DTT models*. https://github.com/zjuradoq/PURE_MGapt_and_DTT_model. 2024.

[67] Arbor Biosciences. *myTXTL T7 Expression Kit.* Michigan, United States of America, 2019.

[68] Karzbrun, Eyal et al. “Coarse-grained dynamics of protein synthesis in a cell-free system.” *Physical Review Letters* 106.4 (2011), p. 048104.

[69] Chizzolini, Fabio et al. “Cell-free translation is more variable than transcription.” *ACS Synthetic Biology* 6.4 (2017), pp. 638–647.

[70] Singhal, Vipul et al. “A MATLAB toolbox for modeling genetic circuits in cell-free systems.” *Synthetic Biology* 6.1 (2021), ysab007.

[71] Poole, William. “Compilation and inference with chemical reaction networks.” PhD thesis. Pasadena, CA USA: California Institute of Technology, 2022.

[72] Laohakunakorn, Nadanai et al. “Bottom-up construction of complex biomolecular systems with cell-free synthetic biology.” *Frontiers in Bioengineering and Biotechnology* 8 (2020). ISSN: 2296-4185.

[73] Doerr, Anne et al. “Modelling cell-free RNA and protein synthesis with minimal systems.” *Physical Biology* 16.2 (2019), p. 025001.

[74] Stögbauer, Tobias et al. “Experiment and mathematical modeling of gene expression dynamics in a cell-free system.” *Integrative Biology* 4.5 (2012), pp. 494–501.

[75] Mavelli, Fabio, Marangoni, Roberto, and Stano, Pasquale. “A simple protein synthesis model for the PURE system operation.” *Bulletin of Mathematical Biology* 77 (2015), pp. 1185–1212.

[76] Tuza, Zoltan A. et al. “Analysis-based parameter estimation of an *in vitro* transcription-translation system.” *Eur. Control Conf. 2015*. 2015 European Control Conference. 2015, pp. 1554–1560. DOI: 10.1109/ECC.2015.7330760.

[77] Jia, Yiping and Patel, Smita S. “Kinetic mechanism of GTP binding and RNA synthesis during transcription initiation by bacteriophage T7 RNA polymerase.” *Journal of Biological Chemistry* 272.48 (1997), pp. 30147–30153.

[78] Shimizu, Yoshihiro et al. “Cell-free translation systems for protein engineering.” *The FEBS Journal* 273.18 (2006), pp. 4133–4140. ISSN: 1742-4658. DOI: 10.1111/j.1742-4658.2006.05431.x.

[79] Dietz, Hendrik and Rief, Matthias. “Exploring the energy landscape of GFP by single-molecule mechanical experiments.” *Proceedings of the National Academy of Sciences,* 101.46 (2004), pp. 16192–16197. DOI: 10.1073/pnas.0404549101.

[80] Chen, Lin-Chi and Casadevall, Arturo. “Labeling of proteins with [^{35}S] methionine and/or [^{35}S] cysteine in the absence of cells.” *Analytical Biochemistry* 269.1 (1999), pp. 179–188.

[81] Rozanova, Svitlana et al. "Quantitative mass spectrometry-based proteomics: An overview." *Quantitative Methods in Proteomics.* New York, NY: Springer US, 2021, pp. 85–116. ISBN: 978-1-0716-1024-4. DOI: 10.1007/978-1-0716-1024-4_8.

[82] Li, Jun et al. "Dissecting limiting factors of the Protein synthesis Using Recombinant Elements (PURE) system." *Translation* 5.1 (2017), e1327006. DOI: 10.1080/21690731.2017.1327006.

[83] Jurado, Zoila. *Source code for PURE CRN models*. https://github.com/zjuradoq/PURE_CRN_models. 2024.

[84] Harris, Charles R. et al. "Array programming with NumPy." *Nature* 585.7825 (2020), pp. 357–362.

[85] Virtanen, Pauli et al. "SciPy 1.0: Fundamental algorithms for scientific computing in Python." *Nature Methods* 17.3 (2020), pp. 261–272.

[86] McKinney, Wes et al. "pandas: A foundational Python library for data analysis and statistics." *Python for High Performance and Scientific Computing* 14.9 (2011), pp. 1–9.

[87] Hunter, J. D. "Matplotlib: A 2D graphics environment." *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[88] Bokeh Development Team. *Bokeh: Python library for interactive visualization.* 2018. URL: http://www.bokeh.pydata.org.

[89] Waskom, Michael L. "Seaborn: Statistical data visualization." *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021.

[90] Cooper, Geoffrey M. and Ganem, Donald. "The cell: A molecular approach." *Nature Medicine* 3.9 (1997), pp. 1042–1042.

[91] McInerney, James, Pisani, Davide, and O'Connell, Mary J. "The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data." *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678 (2015), p. 20140323.

[92] Dixon, Andrew S. et al. "NanoLuc complementation reporter optimized for accurate measurement of protein interactions in cells." *ACS Chemical Biology* 11.2 (2016), pp. 400–408.

[93] *Nano-Glo® HiBiT Extracellular Detection System Technical Manual.* Promega. Madison, WI, 2023.

[94] Miguez, April M. et al. "Metabolic dynamics in *Escherichia coli*-based cell-free systems". *ACS synthetic biology* 10.9 (2021), pp. 2252–2265.

[95] Püllmann, P. et al. "A modular two yeast species secretion system for the production and preparative application of unspecific peroxygenases." *Communications Biology* 4 (1 2021).