TOWARDS A WHOLE-CELL MODEL OF GROWTH RATE AND CELL SIZE CONTROL IN *ESCHERICHIA COLI*

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF CHEMICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Nicholas Anthony Ruggero
May 2017

© Copyright by Nicholas Anthony Ruggero 2017
All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Markus Covert) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Alexander Dunn)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Alfred Spormann)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(James Swartz)

Approved for the Stanford University Committee on Graduate Studies

Abstract

A central challenge in biology is understanding how systems level behaviors arise from their underlying molecular mechanisms. While significant progress has been made characterizing the mechanisms of individual cellular processes, a complete understanding of cell physiology remains an open challenge. Computational models of cell physiology that are both predictive and comprehensive are needed to integrate the scientific communitys diverse knowledge into a single computational theory.

Using an integrative modeling methodology we have made made significant progress towards a whole-cell computational model of *Escherichia coli* (*E. coli*). The model simulates the life cycle of a single *E. coli* cell growing exponentially, and includes representations of cellular processes like metabolism, transcription, translation, chromosome replication, and transcriptional regulation. Furthermore, the model is implemented with feedback mechanisms that control its cellular composition, growth rate, cell size, chromosome state, transcriptional expression, and metabolic capacity in response to its medium environment. The result is a model that grows for an arbitrary number of cell divisions, adapts to three different medium conditions, and simultaneously enables the observation of the abundance, activity, and interactions of every molecular species in *E. coli*.

We have used the *E. coli* model as a computational theory to validate diverse experimental datasets and to integrate them into a consistent modeling framework. In this way we were able to show that the varied knowledge of individual cellular processes largely integrate into a single consistent understanding of cell biology in *E. coli*. We have also demonstrated that the model can guide biological discovery including a new pattern of RNA expression in which the majority of genes in the *E. coli* model transcribed at a rate of less than once per cell cycle, which leads to proteins being expressed only once in many generations with a higher fold change.

This thesis presents a model that represents significant progress towards a whole-cell model of *E. coli*. We hope that expansions of this model will enable a more complete understanding of cell physiology, enable the construction of whole-cell models of other prokaryotes and higher organisms, and serve as a predictive theory to guide synthetic biology, engineering, and medicine.

Preface

Cell theory, the first true theory in biology, which predates Darwin's model of evolution and Mendel's genetics, is still an active field with big open questions. Many parts are understood in great detail in isolation, and highly successful attempts have been made to integrate them. In 2012 Karr *et al.* from our research group published the first whole-cell model of a single *Mycoplasma genitalium* (*M. genitalium*) that integrated diverse aspects of cellular physiology into a single comprehensive computational model [101]. The publication itself focused on predicting phenotype from genotype, but the aspect of this field I came to find the most fascinating was the coordination of growth across multiple scales required in order to ensure that a single cell was able to successfully complete its cell cycle. If *M. genitalium* was to double in size within a given amount of time, did it have enough ribosomes? Was it able to maintain a constant composition of DNA, RNA, and protein? Was it's genome replicated in time? Did it's rate of growth match experiment? These were all fundamental questions that the authors had to address in order to predict phenotype from genotype, and I would argue would not have been thought about in a rigorous manner without the goal of constructing such a comprehensive, and most importantly, integrated model.

These ideas and questions inspired my main contributions towards the second generation of whole-cell models and our efforts building a large-scale model of *Escherichia coli* (*E. coli*). How do bacteria like *E. coli* adapt to different nutritional environments and achieve an optimal growth rate and size? How do bacteria maintain a macromolecular composition that enables the control of growth rate and size? How does the cell maintain this control in the face of cellular stochasticity? And finally in my view the most important question, is our isolated understanding of the involved physiological processes sufficient to produce an integrated, self-consistent, and experimentally validated model?

I felt that investigating these ideas in the context of a whole-cell modeling framework would be a rigorous proving ground of the state of our knowledge. Furthermore constructing and integrating models of cellular growth control would enable the construction of more comprehensive whole-cell models that allowed for the simulation of multiple generations of cellular growth, across multiple environmental conditions, and including shifts between these conditions. In this thesis I focus on my contributions to the *E. coli* model I built as part of a team over the course of 6 years in the Covert research group.

Acknowledgments

Graduate school has been an extraordinary journey. During my time at Stanford I was able to further pave the way into a burgeoning field of science, develop new technologies, make contributions to scientific discoveries, and further mankind's understanding of biology. One thing that is immensely clear to me is that none of this would have been possible without the endless support of the Stanford faculty, my lab-mates, and my friends.

First, I would like to thank my advisor Markus Covert for being both an amazing scientific and professional resource and a boundless source of enthusiasm, support, and ideas. I could not have asked for a better advisor and mentor for my graduate career and I am incredible grateful for the privilege of working with him and the wonderful research group he has assembled.

Second, I would like to thank my colleagues in the Covert Lab. They were an indispensable source of knowledge, feedback, and inspiration and really made graduate school a pleasurable experience both intellectually and personally: Jayodita Sanghvi, Jonathan Karr, Jake Hughey, Miriam Gutschow, Sergi Regot, Keara Lane, Nate Maynard, Silvia Carrasco, Sajia Akhter, Stevan Jeknic, Katie Bodner, David Van Valen, Yu Tanouchi, John Mason, and Taka Kudo. I would like to specifically thank a few people. Elsa Birch for being my mentor during my first graduate school project and teaching me the importance of kerning and presentation. All of the members of the whole-cell team who have joined over the course of my PhD including Javier Carrera, Mialy DeFelice, Heejo Choi, and Travis Horst. The work I did would not have been possible without them. Finally, Derek Macklin for being my partner during the entire course of my PhD working on the *E. coli* model. I wouldn't have made it to the end without his incredible contributions and endless puns.

Finally, I would like to thank all of my friends who have supported me both directly and indirectly through graduate school. In particular, Ellen Casavant for her unwavering positivity and encouragement, for her wise advice, and for listening to more random details about whole-cell modeling than anyone should ever have to, always with a smile.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Background and thesis overview

A model of cellular physiology that is both predictive and comprehensive has been a longstanding goal in the field of biology [55, 134, 189]. Such a model would represent a major scientific advance in our fundamental understanding of biology, and have broad impacts on the fields of synthetic biology, engineering, and medicine. The construction, implementation, validation, and application of a model that includes the scope, detail, and predictive ability required to fulfill this goal has been a major focus of our research group.

In 2012 our research group published the first whole-cell model of *Mycoplasma genitalium* (*M. genitalium*) the smallest known free-living organism [101] along with a suite of software tools [97, 98, 99]. The model was gene-complete, meaning each of *M. genitalium*s 406 well annotated genes were functionally accounted for in the model. More than just a collection of individual gene functions and cellular processes, the model integrated our knowledge of bacterial physiology and cell biology into one seamless computer representation. Accomplishing this at a high level helped outline organizing principles that frame our intellectual understanding of biological systems.

From a technological standpoint the model was foundational and represented a major tour de force in computational biology and simulation. It utilized a novel multi-scale modeling methodology that allowed each aspect of cellular physiology to be represented in the most mathematically natural way given its network topology and available data. Each of the 406 well annotated gene functions was grouped into 28 different sub-models of cellular processes including chromosome replication, transcription, translation, metabolism and cytokinesis. These were assumed to act independently over a short enough timescale (1 second) and then their output was integrated together at the end of every time-step. Physiological processes were modeled based on the quantity and quality of data available. Mathematical representation used to represent them ranged from ordinary differential equations, to constraint based models, to boolean models going from more to less data. This

enabled the assimilation of heterogeneous and mathematically diverse datasets taken from over 900 publications culminating in over 1900 estimated parameters.

Given the sparsity of our knowledge about the individual and coordinated actions of components of a cell, this technical effort paid off and the simulation produced nontrivial, fundamentally new, and verifiable results and predictions. These included correctly identifying 284 of the known experimentally known 360 essential genes, and a novel mechanism for metabolic regulation of the cell cycle in *M. genitalium*. Most impressively, the model was used to accelerate discovery of new biology through model driven hypothesis testing. The kinetic rates of 3 enzymes were predicted by the model based on their simulated expression profile and growth rate, and then validated as correct experimentally [153].

Given this success it is easy to forget that the *M. genitalium* model was a proof of concept and considered by its authors to be a first draft. The scope and novelty of the effort left many avenues open for improvement in both the biology represented, and simulation architecture & workflow. For example, parameterizing and validating the model was particularly challenging due to a lack of *M. genitalium*-specific data, and the relative difficulty of working with *M. genitalium* experimentally [122]. Furthermore, the model was constructed from the bottom up starting at individual submodels, instead of top down starting with the simulation architecture, producing issues in model expansion and scalability.

With these challenges in mind my thesis work focused on building towards larger and more complex whole-cell models. This was a collaborative effort between myself, Derek Macklin, Javier Carrera, and our advisor Markus Covert as well as other graduate and rotation students (detailed below). Together we defined the scope and goals of this project.

The overarching engineering goal for the *M. genitalium* model was to include the function of every annotated gene. This goal was feasible as *M. genitalium* only has 525 genes in total, but this would prove a formidable barrier to entry when considering higher organisms. For example, *Saccharomyces cerevisiae* has 6,000 genes [121], and 20,000 in *Homo sapiens* [45]. We classified a gene-complete, whole-cell model in higher organisms as a long term goal, but not the most immediate task to further the field.

Instead of genes we decided to focus on improving and expanding our utilization of data. We chose to work towards a whole-cell model that integrated as much organism specific data as possible, preferably across multiple environments and growth conditions, in which results and predictions could be experimentally verified. Building larger, gene-complete models would be impossible without the innovation in parameter estimation, data integration, modeling framework extensibility, and feedback regulation this goal required.

With our goal chosen we had to select a model organism, and given that we wanted to include as much data as possible across multiple environments the options were limited. We choice to model *Escherichia coli* (*E. coli*) a gram-negative bacteria with 4,497 genes [104] that can grow in

varied medium environments both aerobically and anaerobically. *E. coli* is broadly used throughout academia and industry due to its experimental tractability and widespread acceptance. Much of what we know about molecular biology and bacterial physiology was originally discovered in *E. coli* and consequentially it is arguably the best studied organism with decades of published scientific research. Many genome scale, impinging datasets already existed in *E. coli* including its genome sequence and annotation, transcriptional expression, transcriptional regulatory network , mRNA and protein half lives, metabolic network, enzyme kinetics, and knowledge about how *E. coli* adapts to varying medium environments (see Chapter 5 Appendix C for full citation list). Furthermore, in *E. coli* it was possible for us to perform experimental data collection ourselves, and potentially validate simulation predictions experimentally. Building towards a whole-cell model of *E. coli* would expand and push the boundaries of the field.

Our reconstruction of *E. coli* represents a significant advance towards a whole-cell, gene-complete model of a larger, more complicated organism. The model stochastically simulates single-cell, exponential growth across three different medium environments including glucose minimal medium, glucose minimal medium without oxygen, and glucose minimal medium with 20 supplemental amino acids, producing diverse phenotypic behaviors. We account for every molecular species in the cell across all three conditions. The function of 1,218 genes are included (43% of the 4,559 annotated genes), and the resource costs and effects of producing all 4,497 genes is accounted for. These genes are implemented in 10 different cellular processes including metabolism, transcription, translation, chromosome replication, RNA & protein decay, and transcriptional regulation. The processes are integrated at a sub one second time scale using an adaptive time step. The data once integrated and reconstructed resulted in over 30,000 reconciled or estimated parameters. The simulation integrates huge volumes of data into a single consistent model, and enables the observation of the abundance, activity, and interactions of every molecular species in *E. coli*.

Beyond being a tool for data integration the *E. coli* model demonstrates many significant advances over the *M. genitalium* model implementation. We are able to simulate three different environments including anaerobic and defined rich media growth representing a spectrum of growth conditions. Furthermore, we can dynamically simulate shifts between these growth conditions and observe simulation dynamics of adaptation on a single cell level. In response to these shifts the *E. coli* model dynamically modulates its growth rate, cell size, and differential gene expression via mechanistic and semi-mechanistic feedback loops that respond to both internal and external signals of environmental changes. This adaptive regulation of cell size and growth rate enables the stable simulation of exponential growth for hundreds of generations. This allowed us to produce small populations of simulated cells, and sample a larger fraction of the distribution of cell-to-cell variation and stochastic inheritance effects. In comparison the *M. genitalium* model grew for a single cell cycle in one medium environment nearly without regulation. This drastic improvement is the consequence of more mechanistic sub-models of transcriptional regulation, chromosome replication,

cell division timing, and metabolism, as well as the improve quality and quantity of data available.

Our choice of model organism and focus on expanding the number of possible environments allowed us to integrate a massive amount of experimental data, checking its consistency using the simulation as a theoretical framework, and cross-validating hundreds of thousands of largely independent experimental measurements. This data integration coupled with computational theory suggested many fruitful lines of experimental inquiry including non-canonical protein expression dynamics, cell cycle regulation, and protein decay rates. Importantly, some of these predictions have been experimentally verified, and given the experimental tractability of *E. coli* potentially all could be.

The work described in this thesis suggests that building towards more refined, expanded, and data-inclusive whole-cell models promises to open a new paradigm in biology where predictive and comprehensive models serve as a theory to validate experiment. Such models not only mechanistically predict single-cell and collective phenotypes based on genotype and environment, but serve as a way to collate the communitys knowledge of cellular biology, and suggest future lies of inquiry. Furthermore, once whole-cell models exist for higher organisms, for example human macrophages, these models could have broad impacts both on engineering design and human health.

Thesis outline and contributions

The work presented in this thesis was performed in collaboration with a diverse and talented team. Here I summarize the content of each chapter, as well as detail my specific contributions to it. First and foremost, my advisor Markus Covert was the driving force and pioneer behind all of the projects I participated in during my graduate career. I am extremely grateful to him for his guidance, instruction, vision, encouragement, support, and boundless enthusiasm.

Chapter 2

Chapter 2 and Appendix A describe my first foray into computational systems biology and building integrated models. This project motivated my interest in building larger and more complex whole-cell model's of bacterial physiology. In Chapter 2 a computational approach is presented to assess what impact the host supply of metabolites has on viral replication. The model system used was an ordinary differential equation model of T7 bacteriophage infection coupled to a flux-balance analysis model of *E. coli* metabolism. The results of this modeling approach suggested that the rate at which a host's metabolism could supply small molecule metabolites was at least as severe a limitation as that imposed by limits on assembly by its macromolecular assembly machinery.

I constructed the model of T7 bacteriophage infecting *E. coli* under the guidance of and in collaboration with Dr. Elsa W. Birch. I contributed to the computer code required to run the model and helped with experimental validation. Reproduced with permission from Birch, E. W.,

Ruggero, N. A. & Covert, M. W. Determining Host Metabolic Limitations on Viral Replication via Integrated Modeling and Experimental Perturbation. *PLoS Comput Biol* 8, e1002746 (2012); copyright 2012 Elsa Birch, Nicholas Ruggero, and Markus Covert.

Chapter 3

Chapter 3 is a review Derek Macklin and I authored on the lessons learned by our research group from constructing the *M. genitalium* model [101]. We identified significant challenges for future whole-cell models in seven areas including experimental interrogation, data curation, model building and integration, accelerated computation, analysis and visualization, model validation, and collaboration and community development. This review was written prior to constructing the *E. coli* model and helped to guide our thinking on the process of building towards a more complicated whole-cell model.

The review was written in equal contribution with Derek Macklin. Reproduced with permission from Macklin, D. N., Ruggero, N. A. & Covert, M. W. The future of whole-cell modeling. *Curr. Opin. Biotechnol.* 28, 111115 (2014); copyright 2014 Derek Macklin, Nicholas Ruggero, and Markus Covert.

Chapter 4

The construction of the *E. coli* model presented in Chapter 5 was a collective effort (see below). Although I contributed broadly to the project, the portions of the project that are completely my own work are outlined in Chapter 4 and Appendix B. These contributions were motivated by my desire to build a more integrated model of *E. coli* that would dynamically adapt to its environment. The three sub-models described vary the concentration and elongation rate of ribosomes in response to environmental shifts, synchronize the chromosome replication and cell division cycle, and generate cell-to-cell variation in single cell growth rate. The results of this modeling produced an *E. coli* simulation that reproduced experimentally measured growth associated cell attributes like ribosome concentration, and number of chromosomes per cell, and enabled the *E. coli* simulation to adapt both its growth rate and cell size in response to its environment. Furthermore, the simulation correctly predicts the variation in cell division control (i.e. a cell dividing at a fixed mass vs after adding a constant mass) with growth rate without it being explicitly encoded in the model.

The work presented in this chapter is entirely my own but integrated extensively with the group contributions to the project. In addition this work would not have been possible without productive discussions with Yu Tanouchi.

Chapter 5

Chapter 5 and Appendix C describe our work towards building a whole-cell model of *E. coli*. The construction of this model required us to compile a massive, heterogeneous, and self-consistent

dataset that uses the model as a computational theory to cross-validates decades of *E. coli* research. The model integrates multiple interconnected sub-models of *E. coli* physiology including metabolism, transcription, translation, transcriptional regulation, chromosome replication, and RNA/protein decay. Simulation results are in excellent agreement with experimental data including RNA and protein expression, growth rate, macromolecular composition, and metabolic fluxes across three different growth conditions. Furthermore, the results suggest a new pattern of sub-generational protein expression where on average a protein is expressed less than once per generation, which produces significant cell-to-cell variability in this fraction of the proteome.

I constructed the *E. coli* model in a whole-cell modeling framework in equal contribution with Derek Macklin and Javier Carrera. Other key contributors include John Mason, Heejo Choi, Travis Walker, Morgan Paull, and numerous rotation students and other lab members.

Reproduced from Macklin, D., Ruggero, N., Carrera, J., Choi, H., Horst, T., Paull, M., Mason, J., DeFelice, M., Bray, S., Akhter, S., Kappel, K., Weaver, D., Karp, P., Covert, M. Cricks complete solution of E. coli, 40 years later. *In Submission.*

Chapter 2

Determining host metabolic limitations on viral replication via integrated modeling and experimental perturbation

Introduction and motivation

The majority of this thesis focuses on the effort of building towards a whole-cell model of *E. coli*. A large part of my motivation in pursuing this thesis topic was the work I did with Elsa Birch during the first year of my PhD investigating the interaction of *E. coli* metabolism and a bacteriophage infection. This work represented my first foray into systems biology, integrated computational modeling, *E. coli* biology, and model driven experimentation.

Beyond my own personal motivation, the work detailed in this chapter is a continuation of the foundational work from our research group integrating flux-balance analysis (FBA) models of *E. coli* metabolism with ordinary differential equation (ODE) models. The first attempt at this was an ODE model of central carbon metabolism integrated with an FBA model of all of *E. coli* metabolism published in 2008 [53]. The model presented in this chapter builds on this work by integrating an FBA model of *E. coli* metabolism with an ODE model of T7 bacteriophage infection. These FBA/ODE models (combining 2 mathematical modeling techniques), were the precursors to the whole-cell modeling methodology used by Karr *et al.* [101] (combining many mathematical modeling techniques). This whole-cell modeling methodology is the basis of the framework used for the work presented in this thesis in Chapter 5.

Background

Any virus is necessarily a metabolic product of its host, since viruses lack the macromolecule machinery and small molecule precursors required to replicate. This dependence has been underscored by recent screens to determine the host genes required for viral infection in a variety of species. The published sets of host-gene viral dependencies have consistently included metabolic genes - both enzymes and regulators - in systems ranging from phages T7 and lambda, to the human viruses HIV and influenza [127, 143, 25, 108, 200, 26, 95, 107]. In complementary findings, some bacterial viruses have recently been shown to encode components as well as direct modifiers of host metabolic machinery [24, 46]. Taken together, these studies emphasize the need to understand viral infection in the context of host metabolism [128].

Viral host dependency screens are useful for identifying individual host genes involved in the metabolic interplay of viral infection; however, studying any of these single points of connection is likely to reveal a complex network of host-viral interactions [129]. Understanding infection as a highly integrated system is therefore necessary to predict the outcome of viral infection following perturbations, such as changes to the host nutritional environment. Similarly, metabolism is a deeply interconnected network, and viral infection represents a dynamic perturbation of it. Achieving a systems-level understanding of host-viral metabolic interaction therefore requires, a strong set of computational tools coupled with quantitative dynamic measurements.

Given the challenge presented by developing such modeling tools and making the needed measurements, bacteria and their viruses, particularly *E. coli* and certain of its bacteriophages, are favorable candidate model systems for building a systems-level understanding of infection. These systems have a long history of study, individually and together, and as a result are associated with a wealth of well-established observations and experimental protocols. Additionally, the host-viral dependency screens involving *E. coli* identified sets of genes whose products were far better characterized and annotated than in any other screen [127, 143]. These systems also have industrial relevance: threatening large-scale cultures [35], and alternately providing highly specific disinfection tools [80].

Critically, *E. coli* and its phage are sufficiently understood to enable the construction of predictive computational models. Phage T7 replication has been described with structured ordinary differential equations (ODEs), that account for the dynamic production of molecular species that comprise the phage during infection[67] (Figure 2.1A right). This model was used to computationally predict the infection outcome of phage genome modifications [68, 66]. Separately, host *E. coli* metabolism has been most comprehensively modeled using Flux Balance Analysis (FBA), which uses linear optimization of an objective function to solve a system of steady-state mass balance ODEs [185]. FBA-based models have expanded to account for essentially all of the known metabolic functionality in *E. coli* (Figure 2.1A upper left) [147, 70, 138]; these models capture growth rates and nutrient exhaustion as well as the impact of genome perturbation and evolutionary outcomes over time

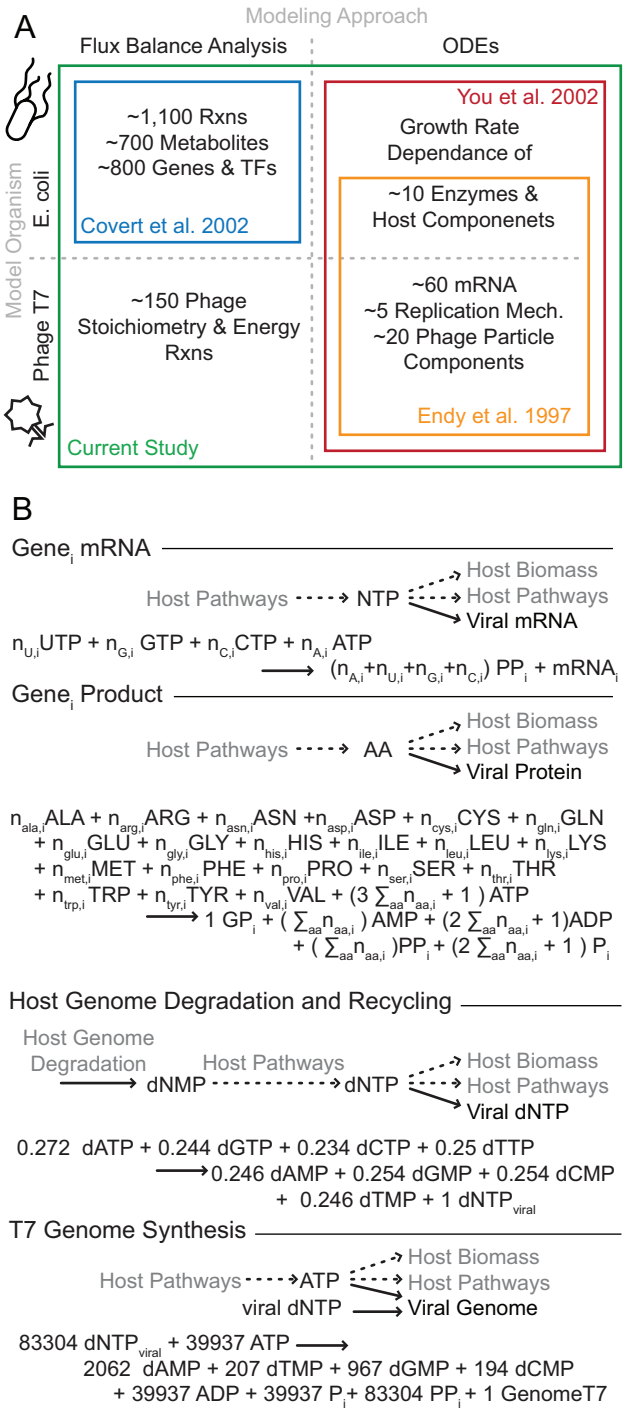


Figure 2.1: **Model approaches, scopes, and additions used in the current integration.** (A) The computational methods and the organisms represented by previous modeling efforts that are combined in this study. (B) The additional reactions constructed in this study for the purpose of translating T7 ODE reaction rates into host metabolite use. Shown at the top for each category is a schematic of metabolite connections to host metabolism, and under it the full stoichiometric reaction, which may be a formula based on nucleotide or amino acid sequence (the gene designations i taking both decimal and integer values in correspondence with the naming of T7 genes [170], a total of n=59 included). Assumptions made in formulating the reactions are expanded in Methods and SI, and the metabolite abbreviations used are consistent with the FBA model definition.

[52, 145, 140].

Two previous extensions of the *E. coli* FBA and T7 ODE models have attempted to encode some dependence of viral replication on host state. One effort was based on the *E. coli* FBA model, with metabolic reactions added to describe production of MS2 virions [90], thus demonstrating the fundamental translation of viral composition to host metabolic terms (the analogous translation for T7 is denoted in Figure 2.1A, lower left). The implemented FBA objective function assumed that the host optimized all of its resources toward viral production immediately upon infection, resulting in an overprediction of phage production. The other modeling effort added a set of correlations between the host growth rate and the availability of replication machinery for T7 processes [197], improving the model's predictions (Figure 2.1A upper right) to the T7 ODE model.

Both of these efforts strongly suggest that a comprehensive, detailed effort to integrate the host and virus into a single computational model will significantly advance our understanding of viral infection in its metabolic context. Ideally such an effort would build on previous work with this host-virus system, despite the different ODE and FBA modeling techniques. Integration of FBA and ODE-type models sets the flux values for a subset of reactions using available kinetic rate equations [53], providing a conceptual framework for combining the host and viral models as depicted in Figure 2.1A.

Here we present an integrated model that is based equally on *E. coli* FBA and the T7 ODEs. It includes a mathematical description of metabolic reactions and demand introduced by the virus, as well as a simulation algorithm that facilitates interaction between the two models throughout the entire course of infection. Our integrated modeling approach enables us to predict phage production changes as the host nutritional environment shifts, and provides insight into the underlying limiting factors in T7 infection.

2.1 An integrated model of *E. coli* and T7 infection

Our integration of the T7 ODEs and *E. coli* FBA (Figure 1A) began with a set of additions to each of the individual models. The *E. coli* FBA model stoichiometric matrix required new reactions to describe the routing of host precursors and energy towards viral synthesis. One reaction was constructed for the synthesis of each viral species represented in the ODE model: mRNA and protein for each of 59 viral genes, viral genome synthesis, and a reaction enforcing the recharge of nucleotide monophosphates (NMPs) released from host genome degradation (123 total reactions; Figure 1B and Methods). The T7 ODEs required one 'production only' reaction rate equation for each of the 123 phage reactions that consume the host metabolites that were added to the host FBA; the net concentration change rate for each molecular species in the original T7 ODEs consisted of production minus consumption terms. However, only the production rate term constrained the stoichiometric reaction in the FBA.

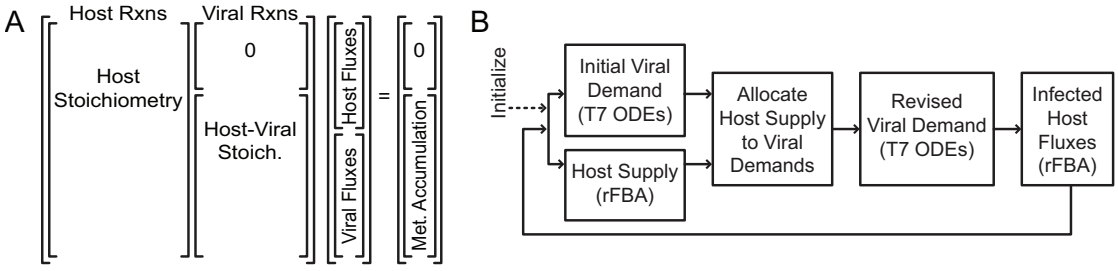


Figure 2.2: **Format and method for the integrated simulation.** (A) The combined host-viral form of the integrated FBA problem is a stoichiometric matrix (Stoich.) that can be considered as blocks: left, the independent host stoichiometric matrix; right, viral reactions consuming host metabolites. The combined matrix may be further organized by host metabolites that do not supply viral reactions (rows of the 0 matrix in the upper right) and host metabolites that are consumed by viral reactions (rows at the bottom aligned with Host-Viral Stoich). The vector of fluxes contains host reaction rates at the top and viral reaction fluxes at the bottom to multiply properly with the host-left and viral-right organization of reactions in the stoichiometric matrix. Accumulation is allowed at the intersections of host viral metabolism (Met. Accumulation; right), but the steady-state assumption is enforced for host-only metabolites (0). A simplified flowchart (B) of the algorithm for integrated simulations, where Initialize indicates the definition of media nutritional conditions and the start of iterations across time, simulating at each integration time point the individual T7 ODEs and *E. coli* FBA, then reconciling the viral rate metabolite demand with host network state supply (Allocate). Both models are then recalculated to incorporate information on their mutual constraint (Revised Viral Demand, and Infected Host Fluxes). Update of environmental information and regulatory constraints at the initiation of each integration step (not specifically denoted on figure) further constrains the host-viral system.

Furthermore, predictions based on the T7 ODEs are valid for a single infection cycle only, and lysis has not been modeled because knowledge of the proteins involved is still insufficient to inform a meaningful representation [110]. As a result we constrained the scope of the integrated model to one single infection cycle.

Next, we expanded the integrated-FBA approach beyond its original capacity to handle the viral demand for resources when these resource demands outpaced the host production capacity. The original implementation of integrated-FBA [53] included ODEs based on central metabolism, which were informed by the environmental state and thus remained within the capacity of host metabolism without any direct communication of host limitations. In contrast, the T7 ODEs do not encode variation in the environmental conditions or the corresponding changes in the host network state's supply of metabolites. As a result, conflicts between the viral metabolite demands and host metabolite supply can arise during the simulation. We therefore encoded communication of information about host limits to the T7 ODEs. This strategy was complicated by the fact that the kinetic formulation of the T7 ODEs is largely independent of small molecule concentrations, except for the nucleotides required for T7 genome synthesis. Furthermore, FBA does not provide concentration information.

Consequently, we devised a metabolite allocation-based approach to bounding reaction rates. Recognizing that the host-viral metabolic interface is the set of common metabolites used in macromolecule synthesis, we split the matrix formulation (Figure 2.2A) into a sum of metabolite rate vectors that represent the host supply and viral demand, where the former constrains the latter. Given a selected host flux distribution, we calculate a strict bound on viral metabolite use. Due to the lack of kinetic information about how the viral metabolic reactions contribute to the metabolite demand, we assume that all viral reactions have an equal and high affinity for precursor metabolites. After calculating rates for the viral reactions from the T7 ODEs to determine the demand for viral metabolites, we scale the rates of all reactions consuming a given metabolite by the same fraction such that total demand is brought within host supply. This method assures that while all reactions are limited evenly, no reaction is limited by a metabolite it does not consume; if amino acids are scarce but dNTPs are available, genome synthesis can proceed but translation cannot.

In summary, this allocation method converts the information about the host metabolic network state into constraints on the T7 ODEs. We implemented this method as part of an algorithm for T7 ODE and *E. coli* FBA integration with bidirectional information exchange and mutual constraint at each time step (Figure 2.2B). After initial specification of the host nutritional environment, the overall viral demand is calculated (without consideration of host limits) using the T7 ODEs, and the host capacity calculated using FBA. Host supply and viral demand are reconciled by calculating the upper bounds on viral production fluxes, after which the T7 ODEs are re-evaluated over the integration time step because metabolite limitation of one viral ODE may affect the ODE solution as a whole. Finally, the infected host flux distribution is calculated using optimization on the host

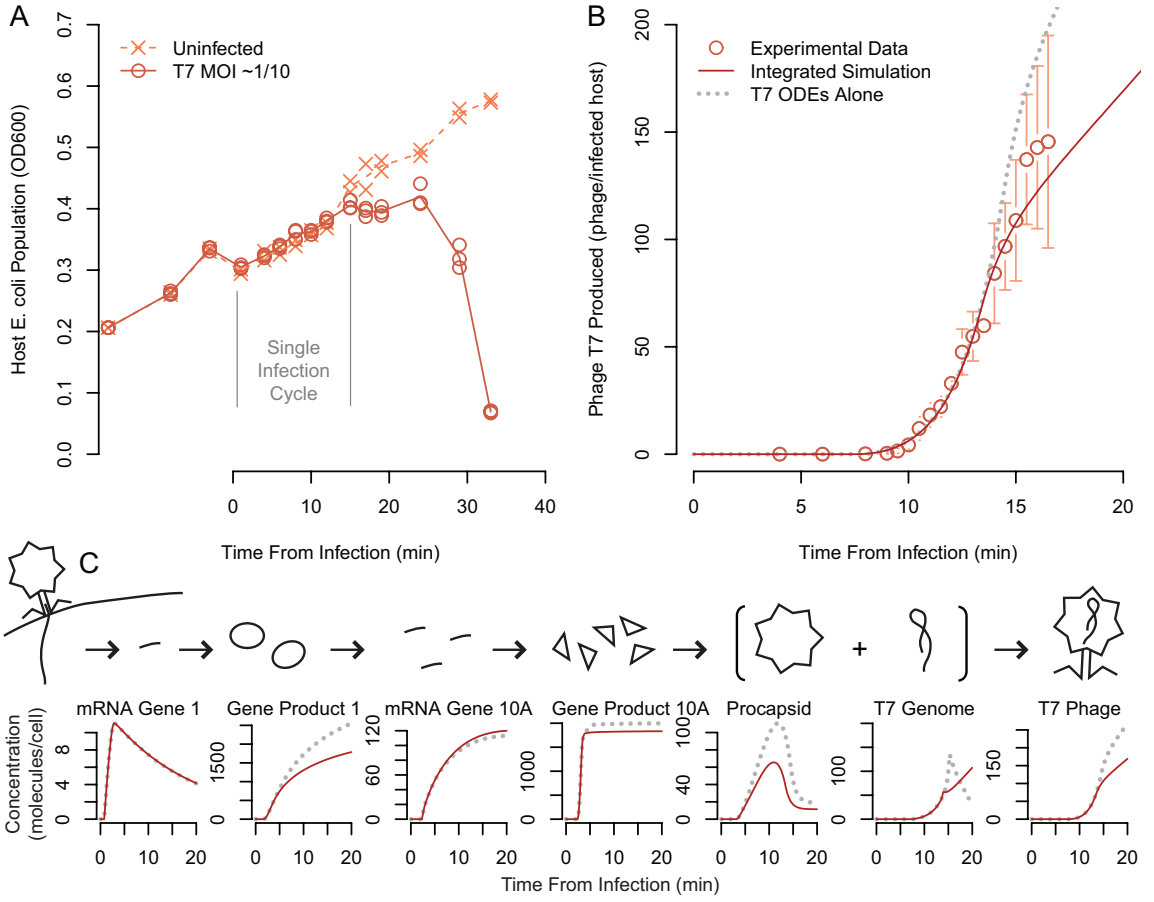


Figure 2.3: **Host population and phage population time courses.** (A) Dynamic time courses of experimental host population data uninfected (line is mean of n = 2) and infected cultures (line is mean of n = 3); an immediate drop in population density occurs when the solution of phage is added at $t = 0$, due to dilution. Initial infection multiplicity was 0.1. (B) Measured and simulated phage production per infected host in tryptone broth media (circles are mean, error bars shown are the standard deviation, n = 3) . Simulation presented for the integrated model and T7 ODEs alone simulated at $\mu = 1.5$ hour^{-1}. (C) Expanded comparison of the simulated concentrations of critical phage replication machinery and phage virion components compared to T7 ODEs alone. Gene Product 1 is the T7 RNA polymerase; Gene product 10A is the major capsid protein.

metabolic network, with viral fluxes bounded exactly to constrained T7 ODE reaction rate values.

2.2 Comparing model and experiment for tryptone media

To validate the ability of the model to predict infection phenotypes, we observed *E. coli* infection by T7 during growth on tryptone broth. We first measured the growth of *E. coli* cultures in the presence and absence of T7 (Figure 2.3A). The culture is cleared within 35 minutes, representing approximately two infection cycles at $37°C$. Experimental strain, media, and assay protocol detail for this and all following experimental results is in A.2.

Unfortunately, with standard OD resolution, the infected and uninfected cultures were not distinguishable from one another within the single infection cycle (Figure 2.3A) simulated by the model. Thus, differences in host growth rate were not a useful metric to assess the prediction performance of our computational model. We therefore returned to the traditional plaque assay-based approach to determine the number of phage produced per infected host cell during a single initial infection cycle, consistent with previous work with the T7 ODEs [67, 197] (Figure 2.3B). We observed rapid increases in the number of phage beginning around 10 minutes.

To compare model predictions to observations, we simulated phage production time courses under the same environmental conditions using our fully integrated model as well as the T7 ODEs alone. We found that the T7 ODEs alone substantially overpredicted the production of T7 phage over time (Figure 2.3B). This overprediction has been reported previously [67, 197]. The integrated model more accurately captured the phage production time course (Figure 2.3B), suggesting that the integrated model is limiting the production of T7 virions (detailed comparison across media given below).

To determine the cause of this limitation, we considered the model's predictions of phage production and host metabolism in more detail. We compared simulated intracellular concentrations of selected phage components for the integrated simulation to those during simulation of the T7 ODEs alone (Figure 2.3C). The model predicts that production of Gene Product (GP) 1 is limited at translation; GP 1 is the T7 RNA polymerase and is required to transcribe middle and late T7 genes. Despite reduced transcription capacity, sufficient mRNA for the major capsid protein (Gene 10A) is still produced. Major capsid protein production is metabolically limited at translation, and thus procapsid availability for phage assembly is decreased, resulting in fewer phage produced during late infection than predicted by the T7 ODEs alone. In the integrated simulation, although phage T7 genome is produced at the same rate as the T7 ODEs alone, it is not packaged as quickly, with a considerable fraction of the total genomes produced remaining unpackaged after assumed lysis. This excess phage T7 genome resulting from phage production limitation at the protein level is consistent with previous experimental observations [67]. The most prominent limitation by metabolism appears during the later steps of replication: mid and late gene product synthesis and genome production.

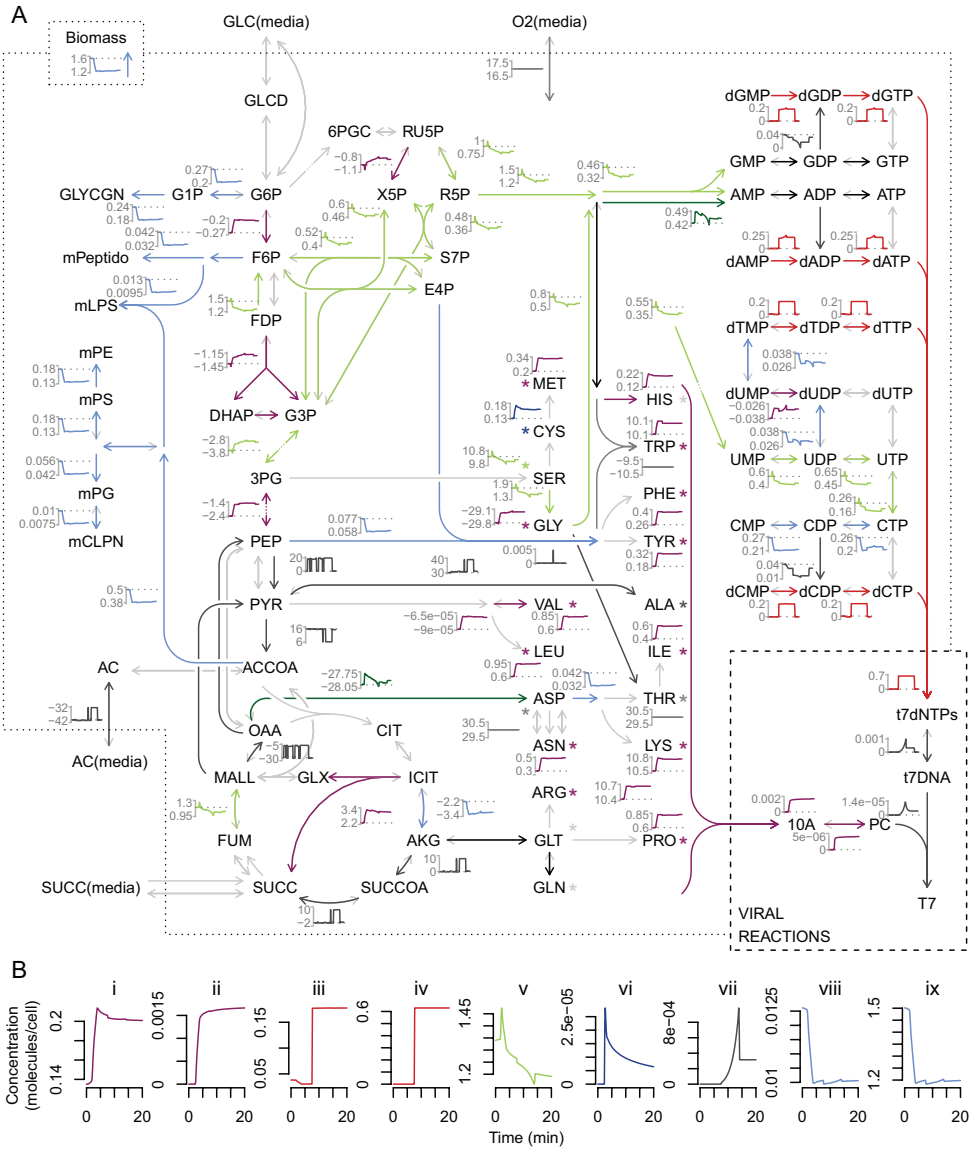


Figure 2.4: **Infected host fluxes on tryptone media.** (A) Flux dynamics are displayed for a subset of the metabolic network map. Arrows representing reactions and the subplots of flux through those reactions are colored according to clustering of flux dynamics. Positive flux values correspond to the reaction direction indicated by the colored arrowhead, negative flux direction is depicted with light grey barbs. Asterisks (*) represent an abbreviation of the arrow for uptake from media. Metabolite abbreviations are consistent with FBA model definition. For clustering, fluxes were treated as vectors with (1-correlation) as distance, and clustered using average hierarchical grouping with a cutoff height of 0.25. Clusters with fewer than ten members appear in black, and clusters with constant dynamics are highlighted in grey. All nonzero fluxes in any media (tryptone, glucose, succinate, and acetate) were included in the flux clustering so that cluster designation and color coding is consistent across media and figures. Maps for media other than tryptone are Figures A.1, A.2, and A.3. (B) Select flux dynamics expanded for clarity ordered to exemplify host flux changes driven by viral dynamics: (i) host amino acid synthesis, (ii) major viral capsid protein synthesis, (iii) host nucleotide phosphorylation, (iv) viral digestion of host genome to dNMPs, (v) purine biosynthesis, (vi) viral mRNA synthesis, (vii) viral genome synthesis, (viii) host cell envelope biosynthesis, (ix) host biomass accumulation.

In contrast, mRNA production is relatively unperturbed early in the simulation, suggesting that metabolic limitation varies in its impact over different periods during infection.

After considering the phage reaction changes in the integrated simulation, we used the model to investigate the changes in host metabolism during infection. The flux-balance component of the integrated model calculates a predicted flux distribution for *E. coli* growth on tryptone in the presence and absence of phage. Essentially all of the non-zero fluxes change dynamically over time in the presence of T7; a subset of these changes are shown alongside the underlying metabolic map (Figure 2.4). Many metabolic reactions experienced prominent flux changes that were coordinated during infection. Dynamic coordination of fluxes in time is not particularly surprising considering the underlying network structure of constraints. However, these similarities in addition to the sheer number of total fluxes that require consideration render unaided visual inspection of infection dynamic information rather uninformative. We found it useful to cluster the flux dynamics into broad categories, which facilitate interpretation of the interesting flux patterns in central and peripheral metabolism during viral replication.

The majority of the observed flux clusters are driven by viral flux requirements (Figure 2.4B). The increase in amino acid synthesis and uptake corresponds in time to the synthesis of viral proteins (Figure 2.4Bi-ii), and similarly flux through nucleotide phosphorylation is high during the period of host genome digestion to dNMPs and viral use of dNTPs (Figure 2.4Biii-iv). Increased nucleotide recharge and pooling is known to occur during phage T7 replication, due at least in part to interactions between phage gene products and host metabolic enzymes[143]. Some complex host flux dynamics result from multiple viral resource interactions (Figure 2.4Bv-vii); flux towards nucleotides first increases during rapid early viral mRNA production, and then decreases as viral genome synthesis occurs, corresponding to the presence of large quantities of nucleotides.

Flux towards host membrane components and cofactors decreases as the ability of the host to synthesize biomass is reduced by the viral draw on components (biomass flux decrease before 5 min) and energy (biomass flux decrease between 5 and 10 mins during dNTP recycling) (Figure 2.4Bvii-ix; light blue). This cluster is the largest of the nonzero flux clusters across and within media, and the sharp decrease in flux within 5 min represents the shutdown in processes that are not required by the virus. Interestingly, this shutdown is not explicitly encoded by either model and therefore represents an emergent property of the integrated model system. The detailed flux maps therefore provide potential for a deeper biological insight regarding the underlying metabolic changes that occur during viral infection.

2.3 Comparing model and experiment for minimal media

The T7 ODEs were originally parameterized to fit data where *E. coli* grew on tryptone broth or other rich media [67]. Later work incorporated correlations between available host machinery (e.g.,

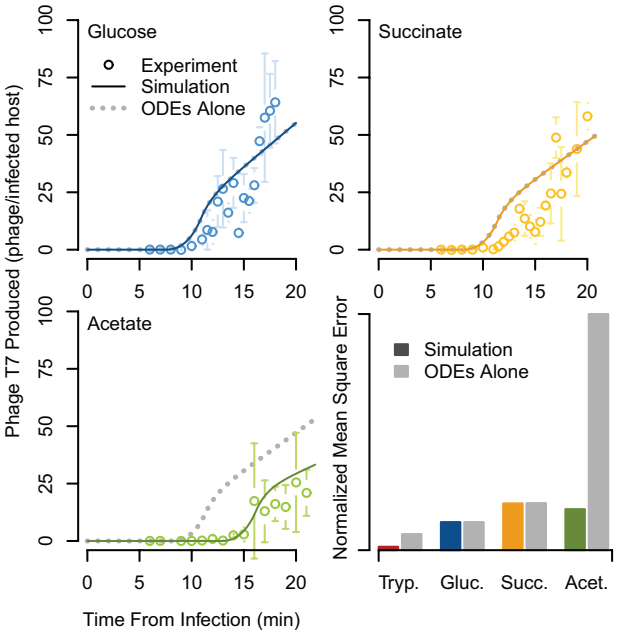


Figure 2.5: **Measured and simulated phage production.** Shown per infected host, across time, experiment compared to model predictions for integrated model system, and the T7 ODEs alone, on M9 minimal media with glucose, succinate, or acetate as carbon source (growth rates for T7 ODEs alone are $\mu = 0.66$ hour^{-1}, 0.45 hour^{-1}, 0.27 hour^{-1}, respectively). Error bars are standard deviation of n = 3. For glucose and succinate media the T7 ODEs time course is not visible because it falls directly beneath the integrated simulation line. The lower right panel quantifies the goodness of fit of the integrated simulation and the T7 ODEs alone to experimental observations using normalized mean squared error.

ribosomes) and host growth rate into the ODEs in order to account for the effect of growth rate on infection dynamics [197]. Host metabolism is encoded explicitly in our integrated host-virus model, and so instead of a given growth rate parameter, the integrated model requires only the environmental conditions as inputs.

Unlike either individual model, the integrated model is capable of predicting the viral infection dynamics for many different culture conditions. We tested model predictions for three previously unmodeled conditions: glucose, succinate, and acetate minimal media. In each case, we measured the phage production over time (Figure 2.5, bottom left and top panels). For glucose and succinate media, the models produced dynamics nearly identical to each other as well as similar to the experimental data. However, for infections on acetate minimal media, the integrated model was more accurate than the T7 ODEs alone. The two predicted time courses differ because the integrated model accounts for the slow growth and nutritional limitation of *E. coli* on acetate (roughly half of the growth rate on succinate). In particular, small decreases in gene product synthesis result in delayed achievement of the thresholds necessary for phage genome replication initiation. Furthermore, all of the simulations, from both the integrated model and the ODEs alone, deviate from the typical one-step-growth phage production trajectory. This is due to the rigid description of host DNA degradation and incorporation into viral genomes in the ODEs, which was originally characterized under a single environmental condition. Quantitative comparison of our observations to the model predictions verified that tryptone simulations were the most indicative of experiment, and that the tryptone and acetate integrated model simulations outperformed those of the ODEs alone (Figure 2.5, bottom right panel).

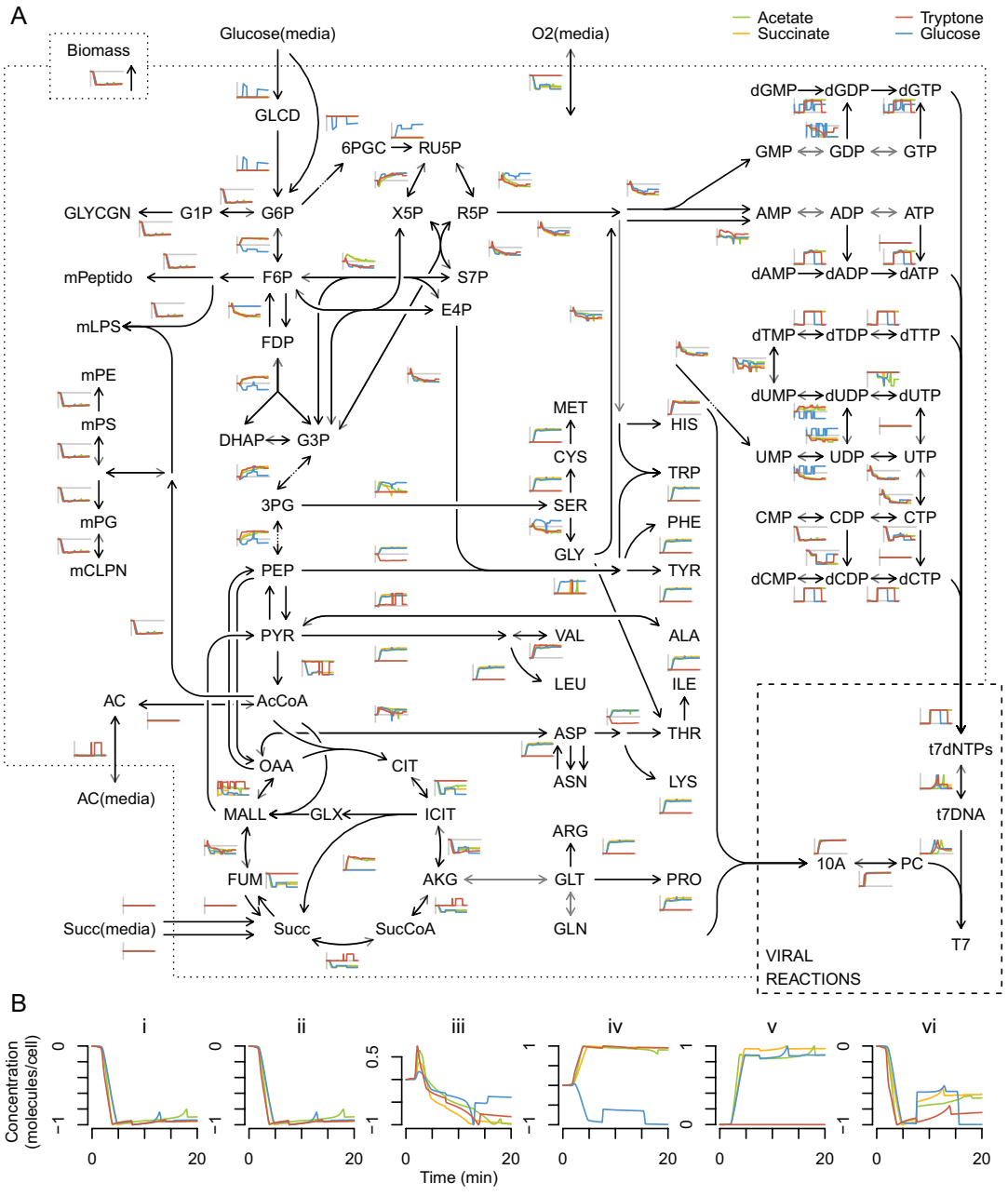


Figure 2.6: **Comparing normalized infected host flux dynamics spark-lines for all four media.** (A) Metabolic map and normalized flux dynamics for tryptone, glucose M9, succinate M9, and acetate M9 media. Flux values were shifted to the uninfected value ($t = 0$), and then normalized to their maximum magnitude on each medium; zero (initial) value is indicated by a grey horizontal line. Metabolite abbreviations are consistent with FBA model definition. (B) Expansion of a selected subset of normalized fluxes. Host cell envelope synthesis (i), and biomass accumulation (ii) decrease similarity across media. Purine synthesis (iii) exhibits dynamic similarity across media. Glycolysis (iv) is observed on glucose while gluconeogenisis occurs on other media. Amino acid synthesis (v) increases on minimal media but not on amino acid-rich tryptone; and the citric acid cycle (vi) demonstrates similarity in dynamic flux change timing, but differences in scaling and direction.

We next wanted to understand how the host and viral fluxes change under these different nutrient conditions. Detailed individual media flux maps analogous to 2.4 are provided for glucose, succinate, and acetate media in Figures A.1, A.2, and A.3 respectively. To generate a global evaluation of the host flux response to infection on varying media, we analyzed the aggregate similarity of the total flux distribution between pairs of media (Figure A.4). Generally this comparison indicated that the flux distribution for infection during growth on acetate was very similar to the distribution during growth on succinate, while there was more divergence between the tryptone and glucose flux distributions than for any other media pair.

Figure 2.6 displays the dynamic metabolic flux distribution for all four infection simulations, normalized to facilitate comparison. Of the fluxes that are non-zero in any of the media conditions, a large fraction show highly similar dynamics. These fluxes include critical biomass-related reactions such as those that contribute to membrane (Figure 2.6Bi-iii) or ribonuclotide biosynthesis. In some regions of the metabolic network, flux dynamics depend more on the media conditions; for example, in central metabolism the flux direction is often reversed between glucose and the other media because glycolysis is occurring rather than gluconeogenesis (Figure 2.6Biv). Reactions involved in amino acid synthesis also exhibit this phenomenon, as they increase in rate on all three minimal media, yet are zero on tryptone medium (Figure 2.6Bv), which contains amino acids. Another interesting example involves citric acid cycle activity, which is especially increased during the high energy demands of nucleotide recycling (Figure 2.6Bvi). One final subset, adjacent to key metabolites such as pyruvate (PYR), oxaloacetate (OAA), and succinate (SUCC), displayed erratic and rapid jumps between their extreme values, which results from equivalent optimal flux distributions calculated by FBA in highly interconnected sections of the metabolic network.

2.4 Limiting factors for phage production across conditions

Finally, we used our model results to address the issue of host-based limitation of viral infection. Many studies assume that phage infection of *E. coli* is limited by "machinery" – the number of ribosomes, RNA polymerases, and similar factors. Another possibility is that in some cases the host metabolic rates are limiting factors; however, decoupling this limitation is difficult due to the regulation of *E. coli* protein synthesis capacity by the availability and type of nutrients [32]. We sought to compare the effects of *E. coli* machinery- or metabolic-based limitation on T7 infection, an exploration enabled by our integrated simulation which can be perturbed in ways not practical experimentally. The detailed simulation output presented in Figure 2.3C indicates that metabolic limitation may be more prominent for certain phage processes and during specific periods of infection. As a summary output for comparison across conditions, we chose the phage production at seventeen minutes post infection. This point is shortly after which all cultures had begun to lyse, releasing phage, and thus making the bulk quantity relevant to phage propagation across generations within

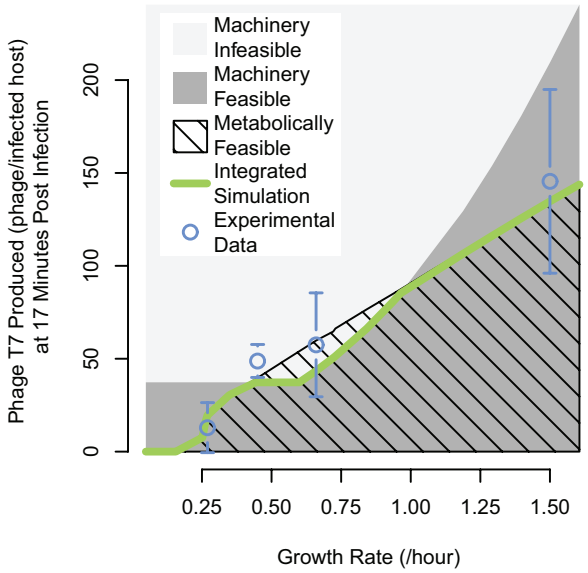


Figure 2.7: **Variation in the limiting factor for phage production across host growth rates.** Modeling results overlaid with experimental phage production measurements. The machinery-feasible region represents phage production values from T7 ODEs alone, with the growth rate supplied to correlations for availability of the host replication machinery; phage production values above the machinery-feasible boundary are considered machinery infeasible. The upper boundary of the metabolically feasible region was calculated using the integrated simulation, but with access to excess host replication factors, which we simulated by multiplying the host growth rate from FBA by a factor of 1.25 when it was passed to the T7 ODE host machinery correlations. Growth rate variation for calculating limitation boundaries and integrated simulation was evaluated with a set of modified flux bounds, with most growth rate sampling values simulated with both carbon and oxygen limitation, which produced essentially identical phage production predictions (resulting points lie within width of the line displayed). Error bars are standard deviation of n = 3.

a host population.

The boundary representing machinery limitations is provided by evaluation of the T7 ODEs alone across varied input growth rates (Figure 2.7). The region that falls below the model prediction is feasible (dark gray), and everything above is not (light gray). To calculate the bounding metabolic phage production limitation, we simulated the integrated model with the modification that excess host replication machinery components were provided to the ODE model (accomplished by passing a higher host growth rate to the ODEs than that predicted by FBA). This calculation was carried out for carbon- and oxygen-limited growth at each resulting growth rate, which resulted in uniform predictions of phage production at each growth rate. Metabolic feasibility here refers to the supply of small molecule metabolites needed to build phage virions; the metabolic limit increases smoothly with host growth rate because the phage is made of a subset of the metabolites included in the host

biomass reaction that represents FBA growth, and a state of host growth maximization is assumed for host supply. This context reveals the integrated model to be slightly mechanistically limited over the range of growth rates between approximately 0.4/hour and 1/hour, and more severely metabolically limited at higher and lower growth rates; however, simulations at very low growth rates do produce empty capsids, reflecting the strong repression of virion DNA production encoded in the ODEs. Metabolic limitation at high and low growth rates explains the better performance of the integrated model than the T7 ODEs alone in predicting phage production on acetate and tryptone media (Figure 2.3 and 2.5 respectively).

2.5 Discussion of host-viral metabolic interaction model

In summary, we investigated the role of host metabolism in viral infection. *E. coli* infection by T7 provided a unique opportunity to address this issue because each system had been modeled, parameterized, and tested independently. We integrated the host metabolic FBA and T7 ODE models and compared the resulting integrated model predictions with new experimental observations. We found that our integrated model was not only a better predictor of viral infection dynamics than either of the individual models for a range of experimental conditions, but also shed new insight on the interplay between virus and host during infection. Most of the active host metabolic pathways were highly impacted by the metabolic demand imposed by virion production. Moreover, we grouped and categorized these pathways by their dynamics; these groups were directly related to the timing of viral demand for key virion components.

It is commonly assumed that viral infection dynamics are predominantly limited by the amount of protein synthesis machinery in the host [79, 67]. In contrast, our results suggest that in many cases metabolic limitation is at least as severe as machinery limitation. This conclusion in turn implies that the wealth of available metabolic reconstructions may enable computational predictions on virion production even when detailed information about interaction with host macromolecules is lacking. More broadly, these results emphasize the importance of considering viral infections in the context of host metabolism.

Finally, we anticipate that models such as this integrated model may be used to rationally perturb the viral infection process by manipulating the host. The modeling and integration approaches developed here are general to a host flux-balance model and a set of viral ODEs, and by integrating the two it may be possible to predict key host metabolic factors whose absence would hamper infection, even as these factors depend on environmental conditions.

2.6 Model reimplementation and integration methods

We implemented the T7 ODEs in MATLAB (R2011a The MathWorks Inc.), informed by the equations presented in the initial publication [67] as well as the code available for the most recent version [197]. The T7 ODEs were originally compared to phage production data at 30°C having been simulated using parameters measured at either 30°C or 37°C [67, 197]. The published flux bounds and regulatory rules of FBA correspond to *E. coli* growth at 37°C, and therefore for consistency the T7 parameters were modified to 37°C where necessary (Table A.5). This modification included kinetic parameters and promotor strengths to maintain prediction constancy with the proportion of phage gene products produced [84], (Table A.4). A stiff solver (ode15s) was used for all solutions of T7 ODEs, as required by discontinuous rate definition equations.

The regulatory-FBA model reaction equations and metabolites are iMC1010v2 [51], with the minor change that a few reversible reactions were reversed for pathway direction consistency. Media definitions for simulated M9 minimal were consistent with past publications and tryptone media was approximated as amino acids (Table A.1); the short time of T7 infection meant that media components were in excess for all simulations with growth rate limitations resulted from flux bound constraints. Some regulatory rules were updated to permit growth on rich media (Table A.2). Flux bounds were mostly consistent with previous publications, with the exception of the relevant set used during growth on tryptone amino acids that were fit using growth rates we collected (Figure A.5).

Phage stoichiometry reactions were included in the FBA system (Figure 2.1B), one for each gene's mRNA and each gene product, as well as for phage genome synthesis and a reaction accounting for degraded host genome dNMP recycling to dNTPs. Included in these reactions are the precursor small molecules that make up each final macromolecule, as well as the energy required for transcription or translation. The FBA host biomass reaction energy requirements are typically phrased in terms of ATP only; to be consistent, the GTP used for energy in phage production processes is included in the reaction stoichiometry as ATP, and the energy requirements for the T7 DNA helicase, which is known to use dTTP preferentially [114] for energy, were also converted to ATP. A full list of assumptions and references for generating phage stoichiometry reactions is in Table A.3.

We added a production rate equation consisting of only the positive terms from the net rate equation for each molecular species in the original T7 ODE model, to bound the forward-only reaction fluxes in FBA. Furthermore, another ODE was added to account for the fraction of the host genome material remaining for degradation. A set of input arguments to the T7 ODEs was also introduced to pass limits on one or more of the production rates. If a production rate was limited, its value is accounted for in the net rate equation. Implicit in this implementation is the assumption that if an mRNA or gene product is degraded, the components are not available to metabolism during infection [196].

Integrated simulation algorithm

A simplified flowchart of the integrated simulation algorithm is shown in Figure 2.2B. The FBA and ODE numerical simulations interacted at every 10 seconds of simulation time. Since host lysis is not modeled by the T7 ODEs, there is not a single logical exit criterion for the simulation. Thus the simulation is run for a set time length slightly greater than what is expected to be the productive duration of infection. A text expansion of the integrated simulation algorithm flowchart shown in (Figure 2.2B) follows, with further detail presented in (Appendix A and Figure A.6):

1. **Specification:** Define media composition of nutrient concentrations, including those that are replenished (often O_2) and those that are exhaustible (usually carbon source).

2. **Initialization, Host:** Determine steady regulation state and growth rate in media, set all media to replenished, and run sequential rFBA simulations until convergence. Set initial time point host regulation state and pass growth rate to T7 ODEs.

3. **Initialization, Virus:** Evaluate T7 ODE host growth rate correlations to set model parameters for host machinery availability. Set initial concentration state of viral ODEs to 0, except for the variable representing host genome for degradation, which is set from growth rate correlations.

4. **Initial Viral Demand:** Evaluate T7 ODEs without any limits imposed for initial estimates of the amount of resources the virus will request from host metabolism. Many viral sub time steps are made within integration time step as determined by ODE solver.

5. **Host Supply:** Set flux bounds based on environmental availability, and regulatory rules referencing environment and host state. Evaluate host linear programming problem (maximize biomass flux in this case) to determine host resources feasibly available to viral reactions.

6. **Allocate Host Supply To Viral Demands:** Distribute metabolites to viral fluxes and set production reaction rate bounds (see expanded section that follows).

7. **Revised Viral Demand:** Evaluate T7 ODEs with production reaction limits. Many viral sub time steps within integration time step as determined by ODE solver.

8. **Infected Host Fluxes:** Set viral reaction fluxes in FBA vector to net viral production rate averaged over integration time step, and evaluate combined linear programming problem (maximize biomass flux) to arrive at overall flux distribution.

9. **Update States:** Consumption and excretion to/from the environment, flux distribution values, viral concentrations. Return to 4 or exit.

Metabolite distribution

Because the T7 ODE kinetic rates do not depend on small molecule concentrations, we bound the phage macromolecule production rates themselves to host production capacity. The method to determine rate limits relies first on an initial 'viral demand' which is based on an evaluation of the T7 ODEs without applied limits over the integration time step. Implementation of this strategy takes advantage of the divided matrix formulation of the problem shown in Figure 2.2A. We further split the matrix (detail in Section A) into summed terms representing the small metabolites provided by the host $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{\mathrm{host}}$, and those consumed by the viral production fluxes $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{\mathrm{viral}}$. In the resulting relationship, shown in Eq. 2.1 (consistent with convention of FBA intake to organism being negative flux), $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{\mathrm{host}}$ is the solution of the typical host FBA problem neglecting biomass exchange, taking advantage of the fact that host biomass is composed of a superset of the small metabolites consumed by viral reactions. The simplified form of this relationship is enabled by allowance of metabolite accumulation at the intersection of host and viral reactions.

$$\begin{bmatrix} 0 & \mathbf{S}_{\mathrm{HV}} \end{bmatrix} \begin{bmatrix} 0 \\ \mathbf{v}_{\mathrm{viral}} \end{bmatrix} = \left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{\mathrm{viral}} \geqslant -\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{\mathrm{host}} \tag{2.1}$$

Once a feasible host flux distribution is selected (by solving for a 'host supply' flux distribution), Eq. 2.1 provides a simple relation that must be obeyed by viral production flux rates in order to assure a solution exists to the combined host viral metabolic problem. The method devised to select a vector of maximal viral fluxes or rates (to pass to T7 ODEs) is detailed in Section A, but essentially allows the maximal evenly scaled flux through viral reactions consuming any given metabolic precursor. For example, allowing full production of viral DNA even if amino acid availability is limiting protein synthesis, yet restricting both if a shared reactant such as ATP is limiting.

Chapter 3

The future of whole-cell modeling

3.1 Introduction

Predictive and comprehensive models of cellular physiology are critical to understanding and engineering biological systems. Such whole-cell models have the potential to guide experiments in molecular biology, enable computer-aided design and simulation in synthetic biology, and inform personalized treatment in medicine. Constructing and validating models with sufficient scope, detail, and predictive power, for a variety of cells, will be a massive undertaking.

Beginning in the late 1970s [164], researchers began modeling cell physiology, primarily using ordinary differential equation (ODE) approaches, creating increasingly detailed models over the next three decades [59, 181, 165]. Later, other groups introduced frameworks that generally require fewer parameters than ODE systems including constraint-based [155, 186] and Boolean methods [57]. Combining these approaches for their respective benefits, our group developed a hybrid methodology: we modeled individual biological processes, each with its own mathematical representation, and merged their outputs to compute the overall state of the cell [54]. Using this approach, we simulated the life cycle of individual *Mycoplasma genitalium* cells, accounting for every molecule and representing the function of every annotated gene [102].

Several unforeseen obstacles arose during the modeling process, which should inform any future whole-cell modeling efforts. Specifically, modeling larger cells and more complex physiology presents challenges in (1) experimental interrogation, (2) data curation, (3) model building and integration, (4) accelerated computation, (5) analysis and visualization, (6) model validation, and (7) collaboration and community development, shown in Figure 3.1. No single research group can simultaneously innovate in all these areas. Rather, a broader community will need to coalesce to tackle these problems. We address this article to that community, discussing the challenges and highlighting notable progress in each area.

3.2 Experimental interrogation

Parameterizing and validating the *M. genitalium* whole-cell model was particularly challenging due to a lack of organism-specific data. Many values were estimated from measurements made in other species. Future efforts will ideally simulate well-characterized organisms, for example *Mycoplasma pneumoniae* [77, 109, 198, 125], *Escherichia coli* [88], and *Saccharomyces cerevisiae* [141, 133]. Because whole-cell models simulate the life-cycle of an individual cell, one would ideally use spatially-resolved, genome-scale, dynamic, single-cell measurements to parameterize and validate the models. However, many published measurements are static ensemble averages representing a population mean at a single time point [83, 192, 124, 74, 73]. This lack of data ultimately presents the modeler with a dilemma: either infer missing data, or create a less detailed model of a particular phenomenon. To create the *M. genitalium* model, we necessarily inferred some degree of dynamical behavior. Faced with a similar problem, others have found ways to incorporate static spatial data in their efforts to create dynamic 3D cell-scale simulations [148]. Promising work in advancing single-cell measurement techniques and technologies [175, 116, 174, 86] will ultimately drive more detailed and accurate modeling. To make these efforts even more impactful and useful, the experimental community could work to establish standardized conditions and place a higher value on consistent, reproducible measurements.

3.3 Data curation

No single technology exists which can chronically measure and record the entire state of a single cell. As a result, heterogeneous data sets must be combined and unified for model parameterization and validation. While efforts such as the BioCyc databases have sought to unify genomic and metabolic pathway information [40], separate databases contain functional parameters such as kinetic rates [159, 194] and expression levels [17]. To compile the data required to build the *M. genitalium* model, which we share via WholeCellKB [100], we had to download and synthesize parameters from these and other databases as well as the primary literature. For larger and more complex organisms, the sheer magnitude of data to collect, and the number of discrepancies to resolve, will present significant hurdles to parameterizing a model.

Since parameterization data increases with organism complexity and known physiology, a part-time manual curation effort will not be tenable. Researchers will need to exploit advances in natural language processing to extract information from the primary literature en masse [72], or outsource part of the effort. Formally interacting with domain experts, as has been done in the flux-balance analysis community [178], will be critical to assembling consensus data sets. Ultimately, a combination of computer-automated and human-augmented approaches will be necessary to gather and assemble the data for larger whole-cell models.

A collection of centralized, organism-specific databases similar to WholeCellKB will be required

for subsequent whole-cell modeling efforts. In the best case, researchers would go beyond including raw data for each figure in a paper [9] and would deposit their results to the appropriate database in a machine-readable format. Dedicated curators would update the database schemas to incorporate new types of information as needed. In addition, the databases would alert the community to significant discrepancies between parameters and flag them as critical issues to resolve. By providing these capabilities, the databases would link experimental evidence to whole-cell models.

3.4 Model building and integration

Comprehensively representing cell physiology in a single computational model requires integrating diverse phenomena over multiple length and time scales, handling the different levels of understanding associated with each phenomenon, and representing the state of the cell in sufficient detail. Our lab's approach to meeting these requirements relies on the notion of biological modularity [81], allowing us to divide the cell into independent state variables (e.g., representing metabolite counts or the functional state of macromolecules) and cellular processes (e.g., transcription, metabolism) [102]. We create sub-models of each cellular process using a mathematical representation informed by available data and current understanding. We assume that, over a small time step, each sub-model can independently execute and update a subset of the cell state variables. To meaningfully combine sub-models in this fashion, we must (1) establish and link common variables, and (2) ensure that the combined behavior is consistent with physical laws and biological phenotypes.

To avoid duplicating work, it is desirable to incorporate published models of particular biological processes into a whole-cell modeling framework. This often requires that the published models be modified to use the common whole-cell state variables, which may, for example, involve changing the published model's quantities from concentrations to counts, or linking its variables to the appropriate cell compartment in the whole-cell framework. Establishing mathematical methods for properly converting a spatially-resolved variable, used in a detailed sub-model, to a bulk quantity, or even to a Boolean value, used in a less-detailed sub-model, would ease the data interconversion between sub-models. Numerical analysis of these methods could be performed to examine factors which affect stability and accuracy of the simulations, and to quantify numerical uncertainty in model predictions.

With a collection of sub-models that properly interface with cell state variables, it must further be enforced that their aggregate behavior does not violate physical laws. For example, the aggregate action of multiple sub-models should not result in the consumption of more resources than are present. To avoid this situation, we developed a method to allocate cell state variables to biological processes proportional to each process's need. In the future, this top-down approach could be replaced with one more grounded in physical laws.

Furthermore, the aggregate behavior of a collection of sub-models should be consistent with biological phenotypes. For instance, the small molecule, RNA, protein, and DNA mass fractions, must approximately double over the exponentially-growing cell's life cycle. This requirement constrains certain sub-model parameters so that metabolism, for example, produces nucleotides and amino acids in the proportions needed by replication, transcription, and translation. The *M. genitalium* model performed this adjustment prior to simulation; however, new methods must be developed to update these loosely-coupled parameters during simulation. Importantly, this will enable proper incorporation of regulatory sub-models [23, 38] which modify the nucleotide and amino acid demands as the RNA and protein expression profiles change in response to perturbations.

3.5 Accelerated computation

Computational simulation is a powerful scientific and engineering tool because it enables rapid and inexpensive exploration of alternative scenarios and hypotheses, as well as design optimization. Such investigations, however, hinge on efficient computation in order to explore a sufficiently large portion of parameter space. The whole-cell simulations of *M. genitalium*, which each took approximately ten hours to run, do not meet this criteria. We can extrapolate that, without innovation in this area, simulations of more complex organisms will take considerably longer to execute. High-performance parallelized computing technologies, such as the Compute Unified Device Architecture (CUDA) [4] or Message Passing Interface (MPI) [5], or even custom hardware platforms [78], in the spirit of Anton [62] or Neurogrid [166], should be adapted and investigated for their abilities to speed-up the execution of whole-cell simulations.

3.6 Data analysis and visualization

Raw simulation data, like raw experimental data, typically requires extensive analysis to be adequately understood and communicated. Techniques from machine learning and dynamical systems analysis could be used to explore and interrogate simulated single-cell phenotypes. These analyses could suggest novel hypotheses about the dynamics of single cells that wouldn't emerge from static, population-averaged data.

To complement analysis technologies, advances are needed in large-data visualization. While our group released WholeCellViz to expose a portion of the *M. genitalium* data set [112], going forward more sophisticated tools must be developed, particularly for exploration, rather than just communication, of large data sets. This requires the development of not only new visual motifs for biological data, but also improvements in data processing and retrieval to enable interactive interfaces for manipulating entire data sets. Existing tools [3] offer these interactive exploratory interfaces, but generally operate on smaller data sets [6]. Fortunately, these problems are recognized

as pressing issues by the visualization community [195]. Preliminary work has begun to explore new visual motifs for biological data [131], [130], [132], and the high-performance computing community is supporting new techniques to improve data retrieval [12].

3.7 Model validation

Model predictions and experimental validation are linked by an iterative process in which each provides feedback on the other [105]. For the initial validation of the *M. genitalium* whole-cell model, we simply compared model predictions to as many heterogeneous data sets as possible that were withheld from model reconstruction. We have also used the model to predict the outcome of experiments which are performed subsequently [152]. Nevertheless, the validation process for the *M. genitalium* model has been guided more by intuition than by a systematic methodology. Ideally, a quantitative metric would exist to specify how much of a model has been validated and would point to data sets needed to improve the coverage of validation. More subtly, methods should be developed which can differentiate novel predictions (e.g., gene essentiality in the *M. genitalium* model) from outputs arising directly from parameter fitting (e.g., biomass composition in the *M. genitalium* model). These innovations would support more widespread model adoption by building trust in the predictions.

3.8 Collaboration and community development

Whole-cell models of more complex microbes and cell types will likely become community endeavors, particularly as the models grow in scope and detail. To facilitate interaction with the broader community, we released the entire code base for the *M. genitalium* whole-cell model under the MIT license [7], permitting open development and re-use. Going forward, we must engage the broader community in contributing to whole-cell model development. The interface between cell state variables and process sub-models must be explicitly documented in detail to lower the barrier to contribution. Furthermore, a formal plug-in system must be developed to simplify the incorporation of alternate sub-models for a particular process. At the project-management level, metrics to quantify contribution and guidelines for authorship need to be proposed and ratified. At the community level, workshops, conferences, and competitions [8] specifically focusing on whole-cell modeling need to be organized to engage the breadth of contributing researchers.

3.9 Conclusion

The need to address the aforementioned challenges provides a wealth of opportunities for interdisciplinary contribution by experimentalists, modelers, computer scientists, statisticians, bioinformaticians, and software engineers. We hope a community will form where scientists and engineers from diverse backgrounds can collaborate and innovate together to overcome these obstacles.

Whole-cell modeling can help researchers prioritize experiments by identifying knowledge gaps and by highlighting measurement discrepancies [152]. Additionally, the comprehensive scope of a whole-cell model enables predictions of the pleiotropic effects of perturbation [142], critical to the future of synthetic biology and personalized medicine. Addressing the issues discussed here will enable whole-cell modeling to realize its potential, and in the process make an impact on model-guided science, synthetic biology, and medicine.

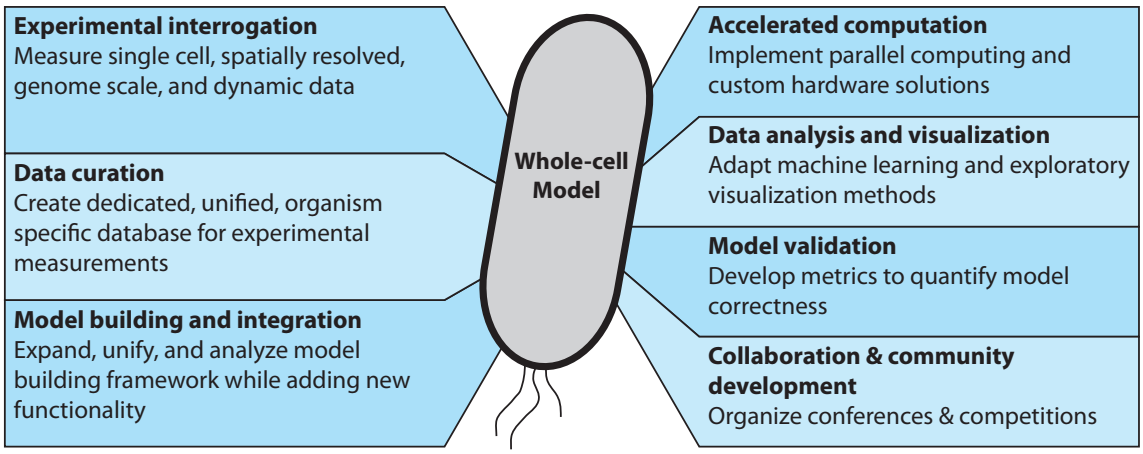


Figure 3.1: The interdisciplinary challenges faced by future whole-cell modeling efforts. A community of scientists and engineers will need to innovate together to surmount these challenges.

Chapter 4

Medium dependent control of bacterial growth rate, composition, and size

4.0.1 The macromolecular composition, growth rate, and size of bacterial cells varies with environment

Bacterial cells vary in size and composition in order to produce an optimal physiological state to maximize growth in a given environment. This phenomena was observed nearly 50 years ago by Schaechter *et al.* [156], and their findings are summarized here:

(i) the growth rate of a cell is related to its metabolic activity and that RNA-containing particles were involved in protein synthesis (which we now know are ribosomes).

(ii) cell size varied exponentially with growth rate, with faster growing cells being larger

(iii) the cellular composition of DNA, RNA, and protein only depend on growth rate, not on the media the bacteria were grown in.

(iv) the amounts of DNA, RNA, and protein are exponential functions of growth rate (i.e. $X = X_o \cdot e^{k \cdot \mu}$ where X is DNA, RNA, or protein and μ is the growth rate.).

(v) the exponents of these functions (k) are different for different macromolecules, which implies that the relative proportions of different macromolecules change with growth rate.

These findings imply that bacteria optimize their composition, size, and growth rate in order to maximize fitness in a given environment. This has also been both empirically validated [123] and is supported by theory [161].

Parameter	Symbol	Units	τ, 100 μ, 0.6	τ, 60 μ, 1.0	τ, 40 μ, 1.5	τ, 30 μ, 2.0	τ, 24 μ, 2.5
Protein/cell	P_C	$\mu g/10^9$ cells	100	156	234	340	450
RNA/cell	R_C	$\mu g/10^9$ cells	20	39	77	132	211
DNA/cell	G_C	$\mu g/10^9$ cells	7.6	9	11.3	14.4	18.3
Mass/cell	M_C	$\mu g/10^9$ cells	148	256	433	641	865
Sum $P_C + R_C + G_C$	-	$\mu g/10^9$ cells	127	204	322	486	697
RNA/Protein	R_C/P_C	$\mu g/\mu g$	0.2	0.25	0.32	0.39	0.47
DNA/Protein	G_C/P_C	$\mu g/\mu g$	0.076	0.057	0.048	0.042	0.041

Table 4.1: Macromolecular composition of exponentially growing *E. coli B/r* as a function of growth rate at 37°. τ is in minutes and μ is in doublings per hour. Reproduced from Bremer *et al.* [32].

Researchers in the following 50 years have carefully measured variation in macromolecular composition, cell size, and growth rate as a function of a cell's environment. Some of their results are summarized in Table 4.1. Although Table 4.1 is indexed by growth rate, in reality the independent variable is the nutritional quality of the medium that the culture was grown in. Media definitions can be found in a previous publication [163].

You can see from Table 4.1 that cell size measured in mass per cell (M_C) increases with growth rate with a cell that doubles in 24 minutes being nearly 6 fold larger than a cell that doubles in 100 minutes on average. Furthermore, the relative proportions of DNA, RNA, and protein vary with growth rate in a predictable way, and can be described by two ratios: (1) RNA/protein (R_C/P_C) increases with growth rate reflecting an increase in the concentration of ribosomes and higher rate of protein synthesis at higher growth rates, and (2) DNA/protein (G_C/P_C) decreases with growth rate reflecting control of DNA replication initiation. These ratios reflect average steady state values undergoing balanced growth (See Appendix B glossary for definition). The balance of the total cell mass that is not RNA, DNA, or protein is mostly comprised of membrane lipids and small molecules.

4.0.2 Computational modeling of the control of cellular composition, growth rate, and size in response to environment

The focus of my work in this chapter is on modeling how an *E. coli* cell varies and controls its composition, growth rate, and size in response to its environment. The interactions and range of attainable behaviors for each of these elements is encoded in an *E. coli* (or bacterial) strain's genetic background.

One of the difficulties in studying how these three elements of bacterial physiology interact is that there are multiple feedback loops present in the system, which makes it difficult to study or model any one in isolation. In order to make this complicated system tractable I use the assumption that each of these processes can be modeled independently assuming a sufficiently short time scale.

It is simplest to think of the direction of causality as follows:

1. Cellular environment determines cell composition (i.e. RNA / Protein, etc.)
2. Cellular composition determines growth rate
3. Growth rate determines the rate of cell division and chromosome replication
4. The balance between growth rate and cell division rate determine a cell's size

This simplified thought model will serve as a reference point for the rest of this chapter and enables the piecewise examination of these interconnected physiological processes (See Figure 4.1 for summary).

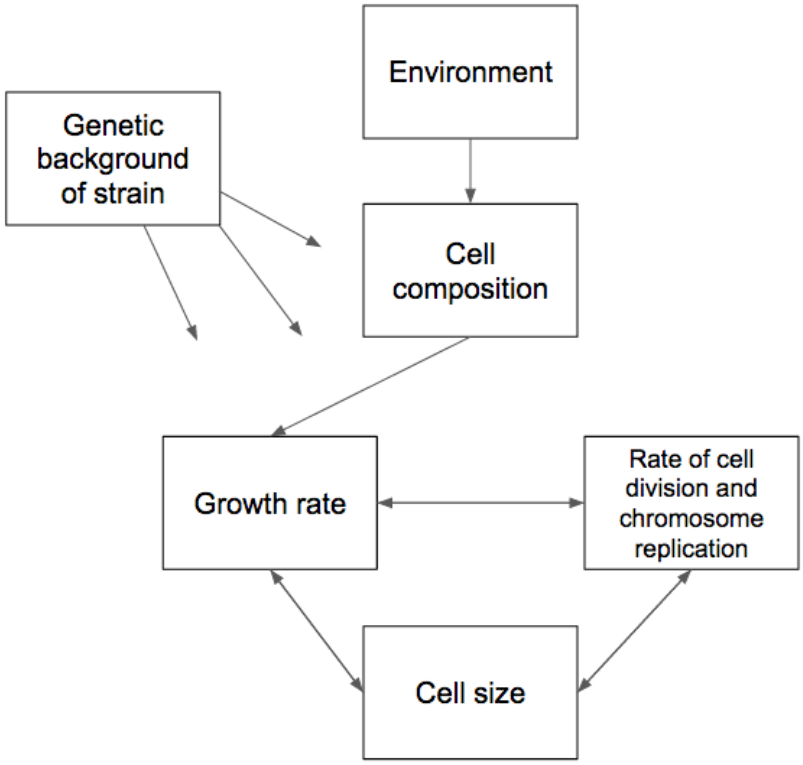


Figure 4.1: **Relationship between parameters related to the growth and macromolecular composition of cells** A schematic of the dependency relationships between different aspects of bacterial medium adaptation.

4.1 Model of macromolecular composition and growth rate response to media environment in *E. coli*

In *E. coli* there is an optimal growth rate that maximizes fitness for a given environment. In order to produce this optimal growth rate during exponential growth *E. coli* attempts to balance the supply of amino acids from metabolism with the rate of polymerization of amino acids by translation [161] by modulating its macromolecular composition.

Multiple feedback mechanisms exist to maintain homeostasis around the optimal growth rate and composition. Many are implicitly or explicitly incorporated into the whole-cell model. Here I describe the feedback mechanism that I implemented in the whole-cell modeling framework that controls *E. coli*'s growth rate by modifying ribosome concentration and elongation rate. This model was *not* included in the publication contained in Chapter 5 due to publication time constrains, although the results presented in this section may be implemented in future versions of the *E. coli* model.

4.1.1 Background: Dependence of growth rate on ribosome concentration and elongation rate

Bacterial (dry) mass is comprised mainly of proteins (50-70%), which carry out nearly all of the enzymatic and metabolic functions of a cell [32]. The process of protein translation in an *E. coli* cell accounts for more than two thirds of a cell's ATP consumption during rapid growth [150]. Therefore studies of bacterial growth have focused on the concentration and activity of the protein synthesis machinery namely ribosomes and their affiliated factors. Furthermore, the importance of ribosomes and their function to exponential growth in bacteria can be seen in the increasing ratio of RNA/protein with growth rate [32], by direct observation that polypeptide synthesis is limited by ribosome availability and function [188], and by evolutionary considerations of the cost of translation [168].

With a few simple relationships the effect that varying ribosome concentration and/or activity on cellular growth rate can be derived. The rate of change of polymerized amino acids in the cell is equal to the rate at which ribosomes synthesize them (decay for proteins is assumed to be negligible). This can be expressed in the form shown in Equation 4.1.

$$\frac{dP}{dt} = N_r \cdot c_p \tag{4.1}$$

Here P is in the units of polymerized amino acid residues per cell, N_r is the number of ribosomes per cell, and c_p is the rate of ribosome function and has units of amino acids polymerized per ribosome per time. Here I define c_p as the product of the elongation rate (e_r) of an average ribosome and the average fraction of time that ribosome is active (β_r) so that $c_p = e_r \cdot \beta_r$. The active fraction (β_r) is

roughly a constant 0.8 - 0.85 for all growth conditions and will be assumed to be so for the rest of this thesis [32].

Returning to the definition of balanced, exponential growth from Equation B.1 we can write another expression for the rate of change of polymerized amino acids.

$$\frac{dP}{dt} = \mu \cdot P \tag{4.2}$$

Which can be rearranged to solve for growth rate.

$$\mu = \frac{\frac{dP}{dt}}{P} \tag{4.3}$$

Substituting Equation 4.1 into Equation 4.3 yields the following relationship.

$$\mu = \frac{N_r}{P} \cdot \beta_r \cdot e_r \tag{4.4}$$

That can be rearranged to:

$$\frac{N_r}{P} = \frac{1}{\beta_r \cdot e_r} \cdot \mu \tag{4.5}$$

Equation 4.5 shows that the cytoplasmic concentration of ribosomes (N_r/P) is a linear function of the growth rate (μ) with the slope being determined by the reciprocal of the ribosome elongation rate. Plotting RNA/Protein from Table 4.2 and Scott *et al.* [161] the relationship in Equation 4.5 is born out as shown in Figure 4.2 with the slope of the line being proportional to the inverse of the maximal elongation rate.

The first conclusion from discussion is that although there are obviously other factors outside of the action of translation that affect a cell's growth rate in a given environment, *all of them* ultimately either affect the concentration and/or function of ribosomes.

The second conclusion is that examining Equation 4.4 it is apparent that multiple combinations of N_r/P and e can produce the same growth rate. This means that cell composition uniquely determines growth rate, but not vice versa. For example, two exponential cultures of *E. coli B/r* can be grown in (1) glycerol minimal medium or (2) succinate + 20 amino acids medium. Both are found to grow at roughly the same rate of 1.0 doubling per hour (at 37°C) [65]. The culture without all 20 amino acids will have a higher ribosome concentration (N_r/P) and a lower elongation rate per ribosome (lower e_r) so that the product of these two parameters, which equals the growth rate from Equation 4.4, is identical for both media despite their different values for N_r/P and e_r.

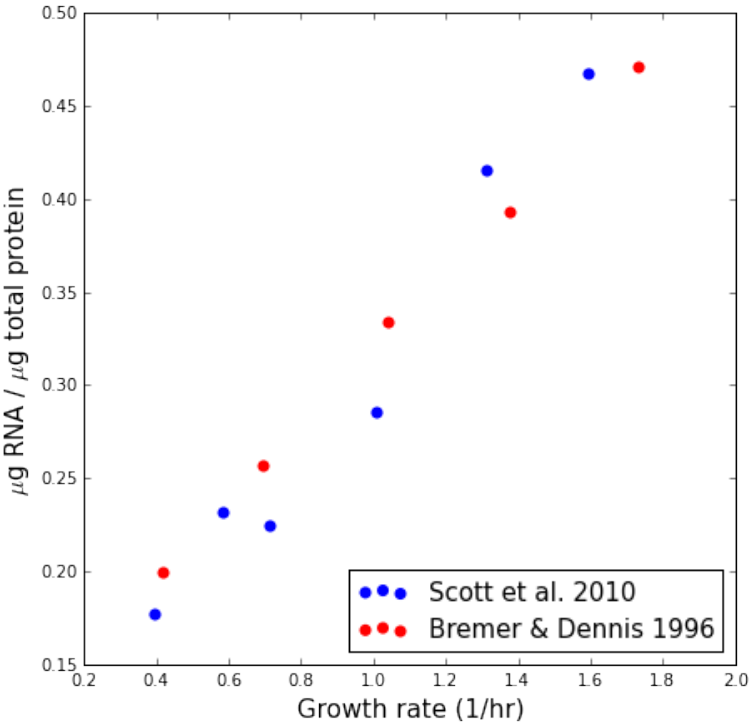


Figure 4.2: **Linear relationship between RNA concentration and growth rate** Empirical relationship between RNA concentration and growth rate in exponentially growing *E. coli*. In *E. coli* 84% is rRNA, 14% tRNA, and 2% mRNA [32], hence RNA content is used as a proxy for the cytoplasmic concentration of ribosomes. Adapted from Scott *et al.* (*E. coli* K-12) [161] and Bremer *et al.* (*E coli B/r*) [32]. An important note is that Equation 4.5 is an approximation because the elongation rate varies with growth rate, which complicates this relationship, but the final result is phenomenologically the same with deviations occurring at doubling times >85 minutes, hence this linear relationship is observed experimentally [106].

4.1.2 Background: Biological mechanism of feedback control on ribosome concentration and function in *E. coli*

In Section 4.1.1 I discussed how ribosome function and concentration determine a cell's growth rate. The next important question that arises is: How does an *E. coli* cell sense and control these two critical parameters in order to maximize its fitness?

Bremer *et al.* have proposed that in *E. coli* the parameter that is sensed and controlled using negative feedback is the average elongation rate of a ribosome (e_r) [29]. The exact mechanism used to sense e_r is unknown although the rest of the feedback loop is fairly well characterized and is described below.

When *E. coli* grows in rich media with all 20 amino acids their ribosomes have an average maximal elongation rate of 22 amino acids per second at 37°C. In minimal media the elongation

rate of ribosomes drops below this maximal value. For example, in succinate minimal media it is 13 amino acids per second (see Table 4.2 for more examples) [32].

During steady state exponential growth whenever the average elongation rate of ribosomes drops below its maximal value the signaling molecules guanosine pentaphosphate or tetraphosphate (collectively abbreviated as (p)ppGpp) are synthesized via (p)ppGpp synthetase II (PSII), which is one of the active products of the bi-functional gene *SpoT* gene in *E. coli* [135]. The other protein function encoded by the *SpoT* gene is a (p)ppGpp hydrolase. The levels of (p)ppGpp in an exponentially growing *E. coli* cell are determined by the balance of function of these two products of the *SpoT* gene. Exactly how the drop in ribosome elongation rate is sensed by either *SpoT* product is not known. Two plausible but unproven hypotheses are that uncharged tRNAs may inhibit the (p)ppGpp hydrolase activity and that PSII activity requires translation as it has a half life of 40 seconds [135, 29].

(p)ppGpp is a transcriptional regulator and interacts with RNA polymerase and enhances or decreases its affinity for a wide range of promoters as well as increasing the frequency of transcriptional pausing [34]. It causes a shift in gene expression that helps *E. coli* adapt to growing in a poorer nutritional environment and higher levels of (p)ppGpp are found in slower growing minimal media cultures (See Table 4.2).

Two of (p)ppGpp's most important effects target transcription at the P1/P2 promoters of *E. coli*'s *rrn* operons that contain the rRNA genes required to produce ribosomes. First, (p)ppGpp down regulates the P1 promoter decreasing the rate of initiation. Second, (p)ppGpp increases transcriptional stalling on all other transcripts resulting in a decrease in free RNA polymerase and a decrease in expression from the constitutive P2 promoters. Both of these effects serve to decrease the rate of production of rRNA and hence the rate of production of ribosomes [199, 34, 29]. The initiation rate of the P1/P2 promoters ultimately determine the growth rate of *E. coli* by directly changing the concentration of ribosomes (See Table 4.2) [65] and other co-regulated elongation factors, tRNAs, etc. [106].

This closes the feedback loop from average ribosome elongation rate. If an *E. coli* cell has supra-optimal levels of ribosomes for a given metabolic capacity, their average elongation rates will be lower, which produces more (p)ppGpp, inhibiting ribosome production and causing the concentration of ribosomes to drop and each proceed at a higher elongation rate to produce the same or slightly faster growth rate due to the increased concentration of translational machinery.

Parameter	Symbol	Units	τ, 100 μ, 0.6	τ, 60 μ, 1.0	τ, 40 μ, 1.5	τ, 30 μ, 2.0	τ, 24 μ, 2.5
Ribosomes/cell	N_r	10^3rib./cell	8	14.9	25.9	43.9	61.4
Ribosome elongation rate	e_r	aa/s-rib.	13	18	21	22	22
Ribosome activity	β_r	%	85	85	85	85	85
ppGpp concentration	ppGpp/P	pmol/10^{17}aa	8.5	6.6	4.2	2.9	2.0
RNAP synthesizing rRNA	ψ_r	%	21	31	48	59	68
Initiation rate at *rrn* gene	V_{rrn}	init/min/gene	4	10	23	39	58
rrn gene per cell[a]	N_{rrn}	gene/cell	12.4	15.1	20	26.9	35.9
rRNA doubling time[b]	τ_{rrn}	min	112	68	39	29	20

Table 4.2: Parameters relevant to macromolecular synthesis rates in exponentially growing *E. coli B/r* as a function of growth rate at 37°C. τ is in minutes and μ is in doublings per hour. Reproduced from Ehrenberg *et al.* and Bremeer *et al.* [65, 32]
[a] Haploid genome has 7 copies. This number is calculated based on the average number and location of replication forks based on the C period, D period, and doubling time.
[b] Calculated using $ln(2) \cdot (\frac{V_{rrn} \cdot N_{rrn}}{N_{rrn}})^{-1}$. The agreement with the expected doubling time is not perfect and likely arises from error due to rounding N_r.
.

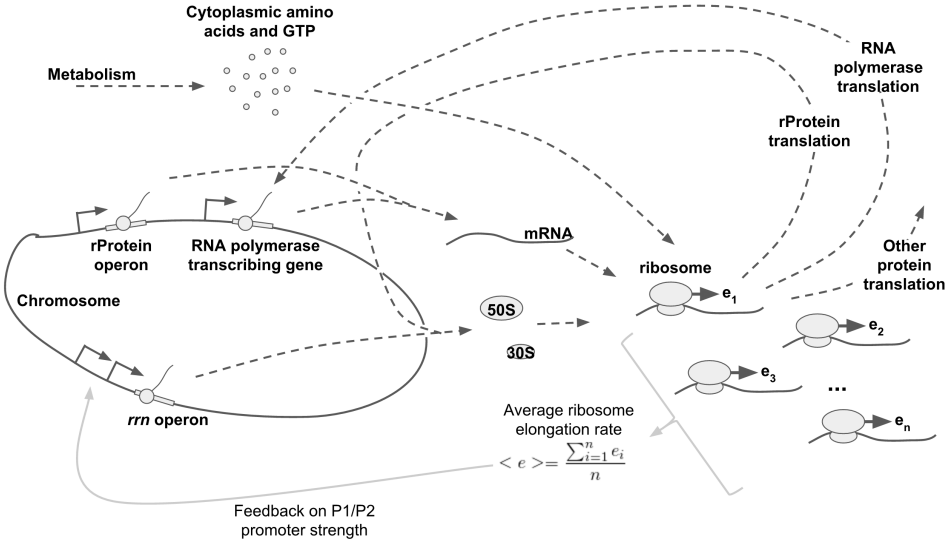


Figure 4.3: **Schematic of feedback on growth rate via proportional control of P1/P2 initiation rate from average ribosome elongation rate** Schematic of model implemented in Section 4.1. Ribosomes each individually elongation at rates $e_1, e_2, e_3, ...e_n$ determined by the rate of amino acid supply to translation by metabolism, maximal elongation rate, and mRNA transcript availability. The average elongation rate is calculated and used to feedback on the rate of rRNA transcription initiation.

4.1.3 Model construction and algorithm

Flux of amino acids through translation

In *E. coli* un-polymerized amino acids primarily exist as free pools in the cytoplasm [18]. Although these pools serve a number of metabolic, regulatory, and osmotic functions [18] their primary purpose during exponential growth is as precursors for translation and polypeptide synthesis.

During polypeptide synthesis free amino acids in the cytoplasm are bound to uncharged tRNAs by tRNA synthetases. These charged tRNAs are then chaperoned into actively elongating ribosomes by elongation factors where their acylated amino acids are polymerized into the growing polypeptide chain. The net turnover rate of all of these reactions can approach a net of 22 s^{-1} *in vivo* [29] and up to 160 s^{-1} *in vitro* [91].

In constructing the *E. coli* a number of modeling approximations were made in order to make this complicated system tractable. The turnover rate of tRNAs and elongation factors would rapidly deplete their pools during a simulation time step of 1 second. Three solutions exist: (1) use a shorter time step, which increases runtime, (2) artificially over express translation associated machinery, which will either perturb the mass or composition of the cell (as done in Karr *et al* [102]), or (3) incorporate kinetic (k_{cat}) and affinity (K_m) data to calculate net pseudo-reaction fluxes for each of the steps, which proved impossible as many critical parameters were not well bounded or were unknown. In order enable model construction we made the strong assumption that the associated translation elongation machinery of tRNAs, tRNA synthetases, and elongation factors are sufficiently abundant in the cell to not be rate limiting. Their function in translation is not explicitly modeled, although they are appropriately expressed as a metabolic load on the cell. Translating ribosomes in the model approximate their function by directly polymerizing amino acids while expending the correct number of high energy phosphate bonds per peptide bond formed.

Furthermore, the depletion of amino acid pools during starvation and the resulting decrease in concentration should limit the rate of translation *in vivo* by decreasing the concentration of charged tRNAs. The objective function implemented in the metabolic model of the *E. coli* model does an excellent job maintaining homeostasis but under perturbation conditions does not correctly reproduce the expected starvation phenotype of depleting amino acid pools (See Chapter 5 Appendix C or Chapter 6 conclusions for details). As a result the rate of amino acid supply to translation by metabolism, which should be a complex function of the environment, metabolic capacity, and translational capacity must be fit to match experimental doubling times instead of arising naturally from shifts in free amino acid pool concentrations.

Despite these modeling approximations and limitations, the model of environmental growth rate control via ribosome elongation rate feedback on rRNA initiation rates predictions fit the expected wild-type behavior closely across three media conditions as well as the dynamics of up-shifts between conditions and is detailed below.

The metabolic capacity for amino acid supply by metabolism is a fit parameter in the model

presented here. It is calculated based on the expected growth rate, protein expression profile, and cell mass. The rate is fit for a given growth rate as follows. First, the expected average distribution of protein expression in a given growth condition is calculated. This is covered in more detail in Chapter 5 Appendix C and will be simplified and summarized here.

The rate of change of a protein can be modeled using an ordinary differential equation:

$$\frac{dp_j}{dt} = \frac{e}{L_j} \cdot \alpha \cdot N_r \cdot m_j - (k_{d,j} + \frac{ln(2)}{\tau}) \cdot p_j \tag{4.6}$$

Where j is the protein index which varies from $1 \rightarrow$ number of coding genes. p_j is the count of the protein, e is the ribosome elongation rate in polymerized amino acids per ribosome per time, L_i is the length of the polypeptide in units of polymerized amino acids, α is the translational efficiency of the protein's mRNA, m_j is the counts of the proteins mRNA, $k_{d,j}$ is the first order degradation rate, and $\frac{ln(2)}{\tau}$ is the rate of dilution due to exponential growth.

At steady state Equation 4.6 can be solved for the steady state counts of p_j:

$$p_j = \frac{\frac{e}{L_j} \cdot \alpha \cdot N_r \cdot m_j}{(k_{d,j} + \frac{ln(2)}{\tau})} \tag{4.7}$$

For a given protein all of these parameters are known or also fit from experimental data simultaneously during an iterative calculation that converges to one final, self-consistent parameter set. From this the count of every protein in an average *E. coli* cell for a given environmental condition/doubling time can be calculated.

Given the nucleotide sequence of each protein p_j the corresponding amino acid sequence can be derived via the codon map for *E. coli*. The total count of a given polymerized amino acid in an average cell ($c_{aa,i}^{total}$) can be calculated by:

$$c_{aa,i}^{total} = \sum_{j=1}^{n_{protein}} p_j \cdot seq_{aa,i,j}^{count} \tag{4.8}$$

Where $seq_{aa,i,j}^{count}$ is the count of amino acid i in protein sequence j. This can be converted into a molar count by dividing by Avogadro's number, and a per unit mass quantity by dividing by the expected dry mass of the cell. By the definition of balanced, exponential growth (See Equation B.1) all components of an average cell must grow exponentially at the same rate. The molar per unit mass quantity can then be converted into a rate by multiplying by the growth rate (μ) for a given condition as follows:

$$S_i = \frac{c_{aa,i}^{total}}{N_{avogadro} \cdot m_{dry}} \cdot \mu \tag{4.9}$$

Where S_i is the expected molar rate of metabolic supply of amino acid i to translation per gram dry cell weight for a given growth condition. In the model of environmental growth rate control

Metabolic rate of supply to translation	Symbol	Units	τ, 100 μ, 0.6	τ, 60 μ, 1.0	τ, 40 μ, 1.5
Total	$\sum_i S_i$	aa/fg·min	**20,787**	**36,621**	**60,832**
Ala	S_1		2097	3691	6172
Arg	S_2		1123	1982	3412
Asn	S_3		908	1599	2618
Asp	S_4		1216	2141	3504
Cys	S_5		158	279	450
Glt	S_6		1385	2445	4079
Gln	S_7		828	1459	2389
Gly	S_8		1671	2943	4887
His	S_9		396	696	1155
Ile	S_{10}		1188	2095	3475
Leu	S_{11}		1771	3125	5132
Lys	S_{12}		1334	2348	4030
Met	S_{13}		533	937	1553
Phe	S_{14}		704	1239	2032
Pro	S_{15}		779	1373	2256
Ser	S_{16}		1084	1912	3148
Thr	S_{17}		1201	2115	3477
Trp	S_{18}		200	351	566
Tyr	S_{19}		589	1035	1667
Val	S_{20}		1615	2848	4820

Table 4.3: Calculated rate of metabolic supply to translation for 3 growth conditions specified in Chapter 5. 100, 40, an 24 minute doubling times correspond to glucose minimal media without oxygen, glucose minimal media, and glucose minimal + all 20 amino acids.
.

implemented here this rate of metabolic supply is fit for each growth condition using Equation 4.9 and enforced during simulation. The net throughput of translation (quantity of $\frac{N_r}{P} \cdot \beta \cdot e$) from Equation 4.4 is set for the expected growth rate and the *E. coli* model converges to the combination of concentration of ribosomes and the ribosome elongation rate to produce the net throughput. The fit, calculated values for the three conditions modeled in the *E. coli* model are tabulated in Table 4.3. Under conditions of severe metabolic duress if S_i is larger than the pool size for a given amino acid i in a time step then the rate supplied to translation is the minimum of these two quantities.

Average ribosome elongation rate determined by translation process

In the *E. coli* model translation is separated in two parts: initiation of 30S and 50S subunits on mRNA transcripts as 70S ribosomes, and elongation of ribosomes on mRNA transcripts polymerizing amino acids and hydrolyzing GTP up to resource limitations. I will discuss how this physiological process is modeled here, and how it has been improved over the version in Chapter 5.

Briefly, ribosomes are initiated on mRNA transcripts by sampling a multinomial distribution weighted by the number of mRNA transcripts for a given protein and that mRNA's translational efficiency. The details of this process implementation can be found in Appendix C to Chapter 5. Ribosomes on an mRNA transcript then process and polymerize amino acids into a polypeptide sequence while hydrolyzing GTP, and once the ribosome reaches the end of its transcript it terminates and dissociates back into 30S and 50S subunits releasing a newly formed protein. Details of this implementation can be found in Appendix B Algorithm 1.

The hypothesis put forth by Bremer *et al.* [29] is that the control variable sensed by the cell is the rate of polypeptide synthesis of ribosomes. The model I implemented calculates this statistic once per time step using Equation 4.10.

$$e_{average} = \frac{\sum_{i=1}^{N_r} n_{elongations,i}}{N_r} \tag{4.10}$$

Where $e_{average}$ is the average number of amino acids polymerized per ribosome, $n_{elongations,i}$ is the number of polymerized amino acids in a time step for ribosome i. This quantity is always less than or equal (and is typically equal to or very close) to the amount of amino acids made available to translation, shown in Equation 4.11.

$$e_{average} \leq \frac{m_{dry} \cdot \sum_{k=1}^{21} S_k}{N_r} \tag{4.11}$$

Where m_{dry} is the dry weight of the cell at a time step, and S_k is the rate of amino acid supply to translation per time step per dry weight from Equation 4.9.

Resource limitations of the rate at which metabolism supplies amino acids to translation are what determine each ribosome's elongation rate. Ribosomes are allowed to elongate up to the limits of amino acids, GTP, the maximal observed *in vivo* translation rate (22 aa/s/ribosome), and available mRNA template sequence.

Ribosome elongation rate feedback on P1/P2 promoter strength

As mentioned previously, under poorer environmental growth conditions the average elongation rate of ribosomes drops below its maximal value [29]. This drop in average elongation rate can be the result a supra-optimal concentration of ribosomes, insufficient pools of amino acid precursors, or insufficient GTP pools. Whenever the average elongation rate of ribosomes drops below its maximal value (p)ppGpp is synthesized by PSII in a translation dependent manner and (p)ppGpp degradation is inhibited by an unknown mechanism [135] . (p)ppGpp interacts with RNA polymerase and prevents RNA polymerase from binding to (p)ppGpp dependent promoters like the P1 promoters of stable RNA genes (rRNA, tRNA) [58]. This decrease in binding results in a decrease in the P1 promoter's average per mass initiation rates and the average rate of production of rRNA. Ribosome components like rRNA, rProteins, and other factors are regulated to be rRNA limited [58] so a

change in the production rate of rRNA results in a change of the production rate of 30S and 50S subunits, and the potential pool of elongating ribosomes.

Due to the numerous unknown parameters and mechanisms involved in the sequence of events described above I chose to look at the feedback of ribosome elongation rate of P1/P2 promoter strength from a control theory engineering perspective. Any control that involves the generation of a signal (average elongation rate) that results in an adjustment of the controlled parameter (ribosome concentration) can be referred to as feedback control. This is in contrast to a simple equilibrium reaction's adjustment back to steady state, which would be insufficient to maintain the non-equilibrium state of homeostasis. With this idea in mind, the pattern of sub-maximal ribosome elongation rates corresponding to sub-maximal growth rates resembles the steady state offset or error from a proportional control feedback system [117].

In a proportional control system the correction or strength of feedback is linearly proportional to the deviation of the control variable from its set point. This will inherently produce a steady state offset or error, which can be shown with the simple example illustrated in Figure 4.4.

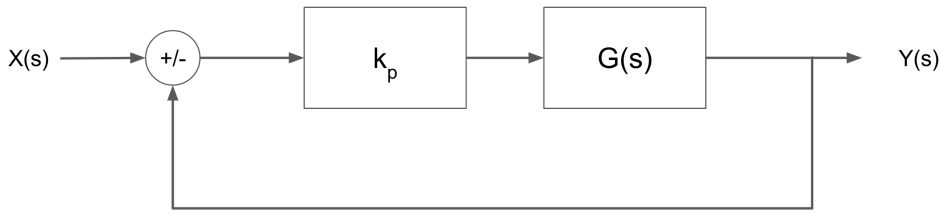


Figure 4.4: **Proportional feedback control** Diagram of proportional feedback control. $X(s)$ is the control variable value during system operation, $G(s)$ is the transfer function that describes the mathematical behavior of your controlled system, k_p is the gain on the proportional control, and $Y(s)$ is the set-point for the control variable

Given the system diagrammed in Figure 4.4 an equation for the output control variable of the system

$$X(s) = G(s) \cdot k_p \cdot (Y(s) - X(s)) \tag{4.12}$$

Where $X(s)$ is the control variable value during system operation, $G(s)$ is the transfer function that describes the mathematical behavior of your controlled system, k_p is the gain on the proportional control, and $Y(s)$ is the set-point for the control variable. Equation 4.12 can be rearranged to:

$$X(s) = Y(s) \cdot \frac{G(s) \cdot k_p}{1 + G(s) \cdot k_p} \tag{4.13}$$

From Equation 4.13 we can see that the value of the control variable during system operation ($X(s)$) will always be some fraction of the set-point ($Y(s)$) between 0 and 1 unless $k_p \to \infty$. How

large this offset is depends on the size of the proportional gain (k_p), and the properties of the controlled system ($G(s)$).

Connecting this example to the feedback control of ribosome elongation rate on ribosome production in the *E. coli* model is a simple analogy. I chose to model the set-point of the system as the maximal ribosome elongation rate (e_{max}), and the control variable as the average ribosome elongation rate ($e_{average}$). The transfer function for a whole-cell-like model ($G(t)$) is not analytically known but is computed dynamically during simulation run time as the time-evolved action of the rest of the model. The behavior I wanted to reproduce was the steady state offset seen *in vivo* between the maximally observed ribosome elongation rate and the average rates seen in minimal media.

The average ribosome elongation rate deviating from its maximal value linearly feeds back on the per mass rate of initiation of the P1/P2 promoters of *E. coli*'s *rrn* promoters. The value of k_p was semi-qualitativly fit to produce the correct steady state offset error in average ribosome elongation rate and doubling time. This modeling approach reduces many of the unknown biological mechanisms in this system to a "black box". The effects of PSII and (p)ppGpp hydrolase activity and expression, (p)ppGpp concentration and its interaction with RNA polymerase, as well as all of the specific regulatory proteins and effectors, and general conditions for transcription (such as superhelicity of DNA templates and NTP substrate concentrations) that would affect the strength of the P1/P2 promoters of stable RNA are all lumped into one linear function in Equation 4.14.

$$V_{initalize}^{P1/P2} = N_{RNAP,free} \cdot [P_{max} - k_p \cdot (e_{max} - e_{average})] \tag{4.14}$$

Where $V_{initalize}^{P1/P2}$ is the rate of initiation at the P1/P2 promoters, $N_{RNAP,free}$ is the number of free RNA polymerases during a time step (determined by rest of whole-cell model), P_{max} is the controller bias and corresponds to the maximal probability of RNA polymerase initiation on the P1/P2 promoters of stable RNA, k_p is the proportional gain, e_{max} is the maximal average elongation rate of ribosomes (22 aa/s/ribosome), and $e_{average}$ is the average elongation rate of ribosomes observed in the previous time step. The algorithm for how RNA polymerases are initialized on promoters can be found in Chapter 5 Appendix C.

Parameter	Symbol	Value	Units	Source
Proportional gain	k_p	0.005	unitless	Semi-quantitative fit
Max prob. of init. on *rrn* P1/P2 promoter	P_{max}	0.183	unitless	Fit from experimental data

Table 4.4: Parameters used in ribosome elongation rate feedback on P1/P2 promoter strength

Model initialization

The *E. coli* simulation begins immediately after cell division. We assume that the cell is sampled from a population of cells growing with steady state, balanced, exponential growth and populate the molecular species, attributes, etc. accordingly using a statistical model. There is nothing specific about the feedback model discussed here that requires special initialization outside of what the *E. coli* model already has implemented in the whole-cell modeling framework. Initialization is described in Chapter 5 Appendix C.

4.1.4 Model improvements, validation, and predictions

In this section I have described the feedback mechanism that I implemented in the whole-cell modeling framework to control *E. coli*'s growth rate by modifying its ribosome concentration using feedback control of ribosome production by the average ribosome elongation rate.

Model improvements

The model of translation elongation I have implemented differs from and improves upon the *E. coli* model in Chapter 5, and *M. genitalum* model in a number of important and biologically relevant ways. The cost of these improvements is a 10-20% slower run time per round of ribosome elongation events.

(1) Fit parameter reduction The number of parameters required to simulate the correct rate of RNA polymerase initiation at the P1/P2 promoters across three conditions is reduced to one parameter: k_p. This changes parameters like ribosome elongation rate and promoter strength from being set in each environment to dynamically encoded in the simulation semi-mechanistically. In short the model presented here is more general than that presented in Chapter 5.

(2) Ribosome elongation rate is determined by metabolic supply and can vary mechanically In the *M. genitalium* model the translation elongation process was optimized for speed of execution. The maximal elongation rate of each ribosome was set based on the known elongation rate in a given growth condition and was strictly enforced. Each ribosome was limited by a specific elongation rate determined by its environment. If there were amino acids or GTP available to proceed further along an mRNA transcript the ribosome would still stop when it reached the expected number of elongations for that growth condition. This limited to total throughput of translation elongation to $\leq N_r \cdot e_{expected}$ even if there was sufficient metabolic capacity to grow more quickly. In practice the throughput of translation was always $< N_r \cdot e_{expected}$ due to incomplete usage of the allocated resources from: (i) when ribosomes terminated they were not restarted in the same time step, or (ii) sequence/resource stochasticity where one limiting amino acid may prevent excess amino acids from being polymerized.

In the model presented in this section resource limitations are what determine each ribosome's elongation rate as opposed elongation rate limiting resource usage. The rate of metabolic supply of amino acids to translation (S_i) is enforced as described in the previous section. Ribosomes are allowed to elongate up to the limits of resources, the maximal observed *in vivo* translation rate (22 amino acids / s), and available mRNA template sequence. The result is that the number of polymerization reactions that each ribosome undergoes during each simulation time step can differ enzyme to enzyme and can either be greater than or less than the expected elongation rate for that growth condition.

This is an important difference. In the *M. genitalum* model in order to buffer against stochastic effects, and pool size limitations translation associated machinery was artificially over-expressed. Because the maximal limit on ribosome elongation was *exactly* the expected value, stochastic fluctuations in sequence and amino acid availability only allowed ribosomes to proceed less than the expected number of elongations (not more when possible). If ribosome components were not over expressed this strictly means that $N_r/P \cdot e \leq \mu_{expected}$. In the model I implemented stochastic variation in sequence and amino acid availability will allow ribosomes to proceed in a distribution around the expected elongation rate producing the correct average behavior without the need to over express ribosomal components. Furthermore, the average elongation rate of ribosomes will be a function of the number of ribosomes available for a given rate of metabolic supply. The same rate of supply with fewer ribosomes will produce a higher average elongation rate than one with more ribosomes on average. This behavior is critical for average elongation rate to feedback on ribosome concentration and also benefits from being more biologically mechanistic. Further details comparing the algorithms for the *M. genitalium* model and the *E. coli* model can be found in Appendix B Algorithm 2 and 1 respectively.

Model validation

In order to examine the dynamics of macromolecular composition, growth rate, and the feedback loop I implemented I generated simulation output under a number of growth conditions and perturbations. The approach detailed in this section qualitatively reproduces expected across all three modeled environments. As the quality of the medium improved, the growth rate of the cell increased to the experimentally measured value. This increase in growth rate was due to a higher concentration of ribosomes being produced, a higher ribosome elongation rate, and a higher rate of initiation at the *rrn* P1/P2 promoters (See Figure 4.5). Furthermore, the effect of changing the rate of metabolic supply of amino acids to translation on the average ribosome elongation rate is demonstrated with the expected outcome.

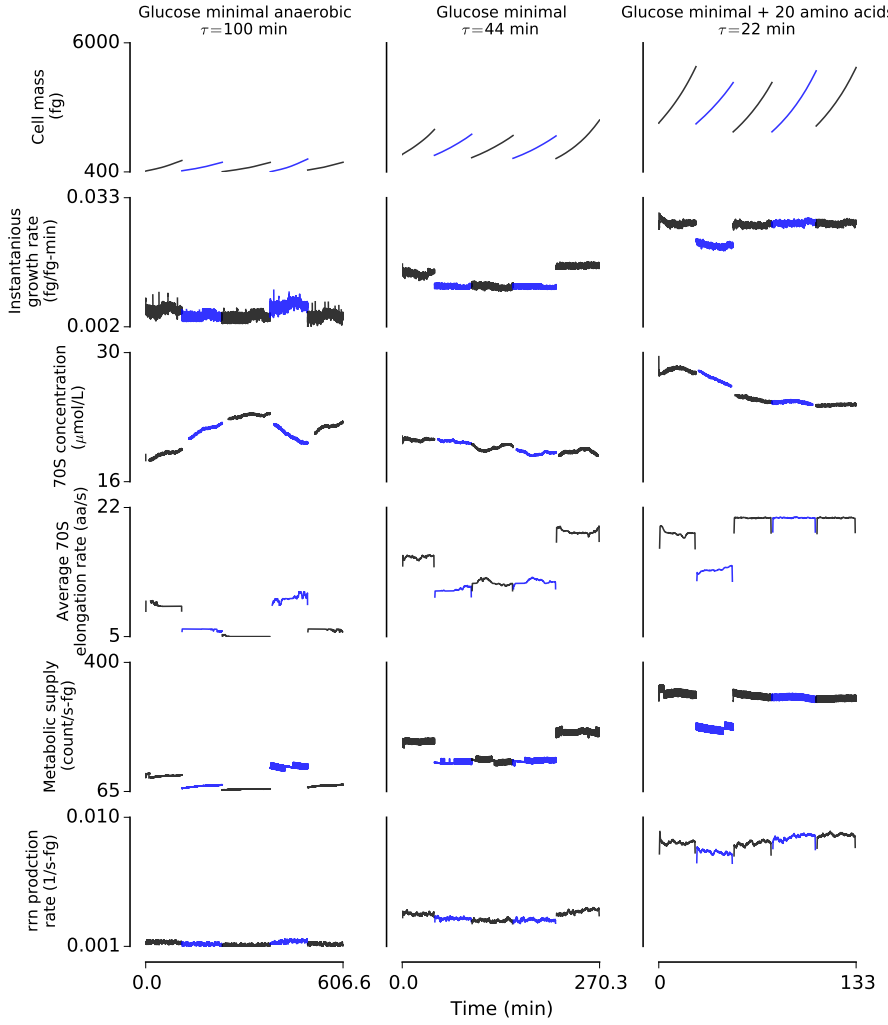


Figure 4.5: **Model of feedback control of macromolecular composition and growth rate across three growth conditions** Simulation output for three different growth conditions: (A) anaerobic glucose minimal medium (B) glucose minimal medium and (C) glucose minimal medium with all 20 amino acids. (Row 1) Cell total mass. (Row 2) Instantaneous growth rate showing cell-to-cell variability and simulation stochasticity. (Row 3) Concentration of active 70S ribosomes. (Row 4) Average elongation rate of 70S ribosomes. (Row 5) Rate of metabolic supply of amino acids. (Row 6) Rate of rRNA production.

Next, I examined the response of the simulation to an up-shift in medium conditions. The shift consisted of adding all 20 amino acids to glucose minimal medium. The up-shift first caused an increase in the rate of supply of amino acids to translation by metabolism, which mediates the

uptake of amino acids from the medium. Next, the average ribosome elongation rate increased to it's maximum, causing an increase in the rate of initiation at the *rrn* P1/P2 promoters. Finally, this increase in initiation rate caused the concentration of ribosomes to increase, a small increase in the average ribosome elongation rate, and a net increase in cell growth rate to a new steady state value. The dynamics of this can be seen in Figure 4.6.

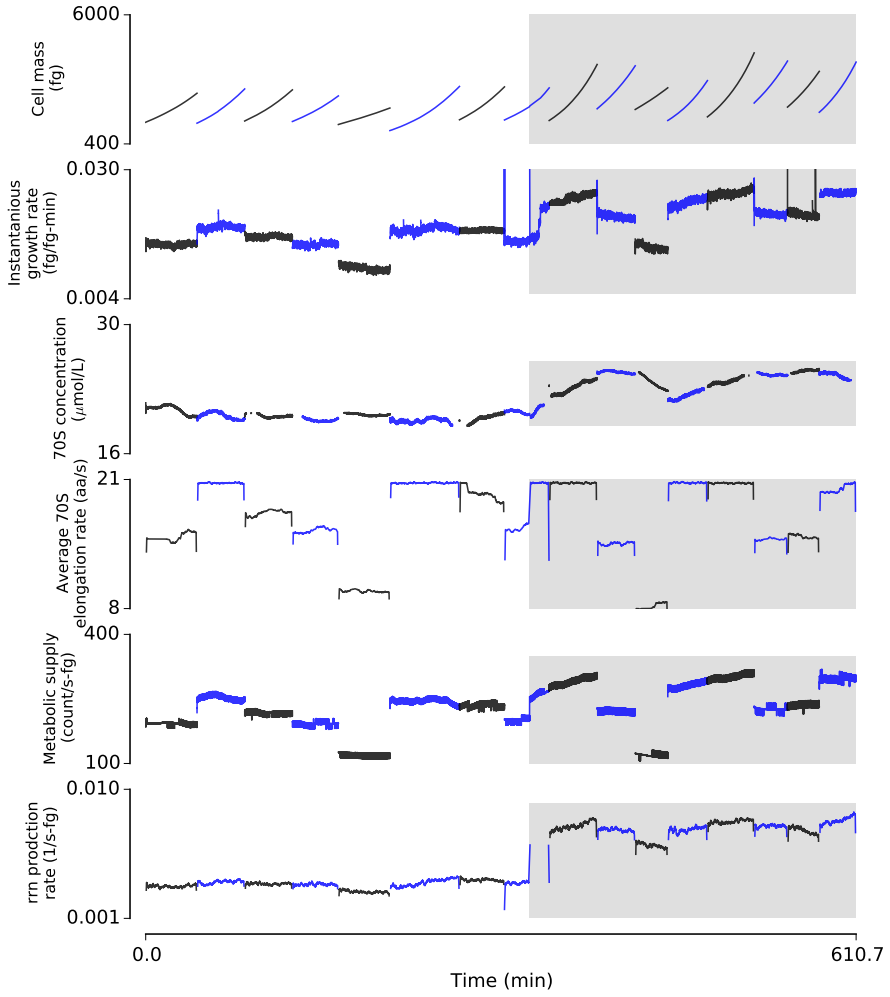


Figure 4.6: **Up-shift in medium with feedback control of macromolecular composition and growth rate in response to environment** Simulation output with shift in medium from glucose minimal to glucose minimal with all 20 amino acids supplemented (denoted by gray box). The shift in medium causes an increase in growth rate and automatic adjustment in the rate of chromosome replication initiation and coupled cell division. Importantly in the top traces you can see that the critical chromosome initiation mass shifts upwards as the size of the cell and increases relative to the number of origins of replication. See Figure 4.5 caption for explanation of traces.

As a final validation I compared measured population level attributes of *E. coli*'s RNA fraction (a proxy for number of ribosomes), ribosome elongation rate, and rate of *rrn* P1/P2 promoter initiation across three growth mediums and doubling times using small populations of simulated

E. coli cells. The simulated average RNA mass per cell, ribosome elongation rate, and rate of rRNA initiation were all in agreement with the trends shown by the experimental measured values. This demonstrates that the model of medium dependent control of macromolecular composition and growth rate reproduces expected cellular behavior across multiple growth conditions (Figure 4.7).

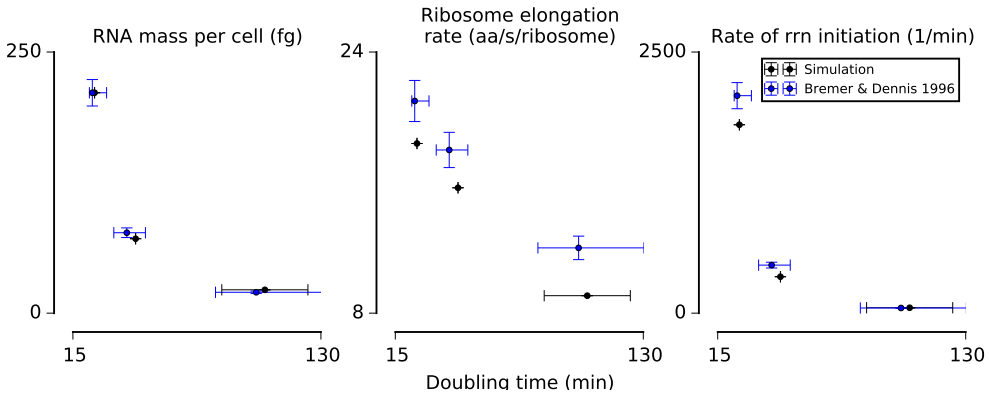


Figure 4.7: **Comparison of simulated and experimentally measured growth parameters in *E. coli*** Small populations of cells (n=8 per condition) were simulated for three growth conditions and compared by doubling time to data from Bremer and Dennis 1996 [32]. Dependent axis error bars on Bremer and Dennis data were estimated as 6% based on measurement error in RNA content [32]. Independent axis error bars on Bremer and Dennis data were estimated by interpolating data on single cell variability in growth rate from Wallden *et al.* [191].

Model predictions

The study of nutritional up-shifts and their effect on the macromolecular composition and growth rates of populations of bacterial cells have been studied has been studied for 50 years now [156]. It is well known that during a medium up-shift from minimal medium to synthetic rich medium the synthesis rate of rRNA and tRNA changes abruptly (within about a minute of the shift) to the new growth rate of the culture in the post-shift medium. The net growth rate of the culture lags behind and can take some hours to reach this new rate that was initially "set" by the rRNA and tRNA synthesis rates [65].

The dynamics of a medium shift on single-cell macromolecular composition and growth rate has never been studied due the experimental difficulty of determining single-cell mass compositions. With the *E. coli* simulation presented here it is possible to computationally investigate the effect of a medium up-shift on single *E. coli* cells.

The *E. coli* simulation predicts similar average behavior to the population level phenomena detailed above, but significant cell-to-cell variability in the dynamics of the response to a medium up-shift. As expected, the rRNA synthesis rate immediately jumps to the post up-shift growth rate, with the protein, and net cell growth rates converging more slowly (see Figure 4.8).

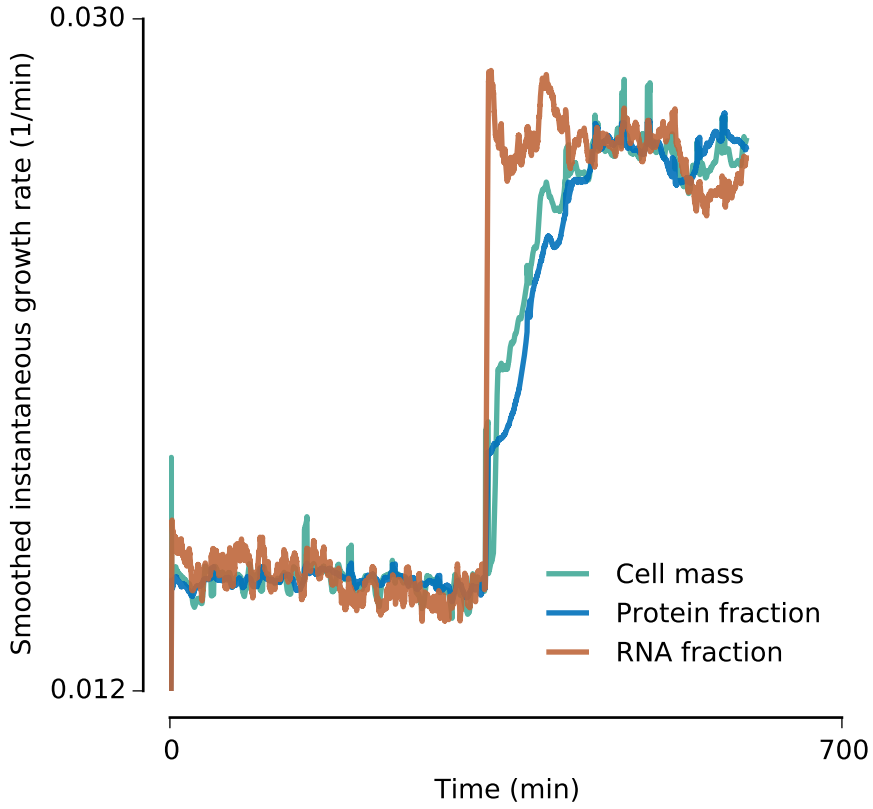


Figure 4.8: **Shift in macromolecular growth rates in response to medium environment change in *E. coli*** Response of the instantaneous growth rates of RNA, protein, and total cell mass to an up-shift in medium conditions from glucose minimal to glucose minimal plus amino acids. RNA, which is predominantly made of rRNA, immediately shifts to the new growth rate of the final medium, which drives the shift in macromolecular composition of the rest of the cell and sets the cell's final growth rate. These traces represent 16 generations of cell growth with the medium shift occurring halfway through generation 8.

4.2 Model of synchronization of chromosome replication and cell division in *E. coli*

The cell cycle is a fundamental aspect of cellular physiology. It consists of a series of events, including initiation of DNA replication and cell division, each of which takes place at an appropriate cell age and hence is coordinated with cell growth.

Under conditions of balanced, exponential growth the environment of an *E. coli* cell is constant, and therefore the coordination of cell cycle events cannot be triggered by external stimuli but must be in response to a set of internal signals that the cell uses to evaluate its age and readiness to proceed. A large body of evidence suggests that under normal balanced growth conditions the primary internal signal for cell division is the completion of a round of chromosome replication [60, 172, 191].

Here I describe the integrated model of chromosome replication and cell division that I implemented in the whole-cell modeling framework. This model was implemented in the *E. coli* model presented in Chapter 5.

4.2.1 Background: Chromosome replication initiation is regulated point in cell cycle

The replication and inheritance of genetic material is an essential process across all biological organisms. The completion of a round of chromosome replication makes sense as a requirement for cell division as cells must maintain a full complement of genes in order to grow and produce functional daughter cells. Furthermore, balanced bacterial growth as described in Equation B.1 requires that *E. coli*'s chromosome replication on average keeps pace with cell division events. A straightforward way to accomplish this goal is to require chromosome replication to complete prior to cell division.

This requirement at first seems impossible for *E. coli* to achieve under more favorable media conditions. The elongation rate of it's replication machinery limits the time it takes to duplicate its chromosome to $\geq\sim$ 40 minutes and *E. coli* can divide in as little as 20 minutes per cell cycle.

The solution to this apparent paradox was first described by Cooper and Helmstetter in 1968 through a series of publications. They defined the C period as the time for a replication fork to traverse the genome, and the D period as the time between the end of a round of replication and cell division [50]. They found that during balanced growth: (i) the C period in *E. coli* was roughly constant [82], (ii) cell division occurs at a constant amount of time after DNA replication is initiated (C+D period) [50, 49], and (iii) the number of origins of replication triggered simultaneously varies discontinuously with growth rate ($<$ 1 doubling/hr there is one origin, 1 - 2 doublings/hr there are two origins, and 2-3 doublings/hr there are four origins) [50].

In order to explain these results their model predicted "multiple replication forks during rapid growth" [50]. Regardless of the growth condition the C+D period was roughly constant, and the cell divided C+D time after chromosome replication initiation. This implied that the regulated point of both chromosome replication and the cell cycle is the initiation of chromosome replication. As long as on average a new round of chromosome replication was started per cell division event balanced growth is maintained (See Figure 4.9 for schematic of Cooper-Helmstetter model).

The question then arises: how does the cell trigger at least one round of chromosome replication per generation? A solution was proposed by Donachie also in 1968 [60] (it was a good year for

microbiology). Observing that the average cell size scales exponentially with doubling time [156] at roughly the same rate that the number of origins at chromosome replication initiation scales, Donachie concluded that chromosome replication initiation occurred at a constant cell mass per origin of replication. This cell mass is called here the critical mass ($m_{critical}$) and is constant for all growth rates. This criteria guarantees that on average the concentration of origins and DNA is roughly maintained while cell size may vary based on media condition dependent growth rate (See Figure 4.10 for schematic of Cooper-Helmstetter/Donachie model during cell growth).

In recent years, developments in microscopy, microfluidics, and image analysis have enabled single-cell experiments that have significantly improved our understanding of single cell growth, size, and heterogeneity, and chromosome dynamics in bacteria [177, 173, 36, 191]. In particular in the case of Donachie's predictions Taheri-Araghi *et al.* have been able to empirically corroborate his deductions 47 years later using microfluidics to correlate the time at cell division to a constant cell size C+D time prior where presumably chromosome replication initiation occurs [172]. Furthermore, Walladen *et al.* using fluorescently labeld DnaQ to mark sites of chromosome replication, were able to verify constant initiation masses at a single cell level with no inference of when chromosome replication initiation began like Taheri-Araghi *et al.* [191]. All these data are both reproducible and consistent with the model that chromosome replication initiation occurs at a constant critical cell mass per origin of replication across all growth conditions, and that cell division occurs C+D time after this event, completing the cell cycle.

Given the wealth of established data and the apparent invariance of the parameters required across growth conditions I constructed a phenomenological model of chromosome replication initiation in the model of *E. coli* detailed in Chapter 5. In this section I describe its construction, simulated output, and validation.

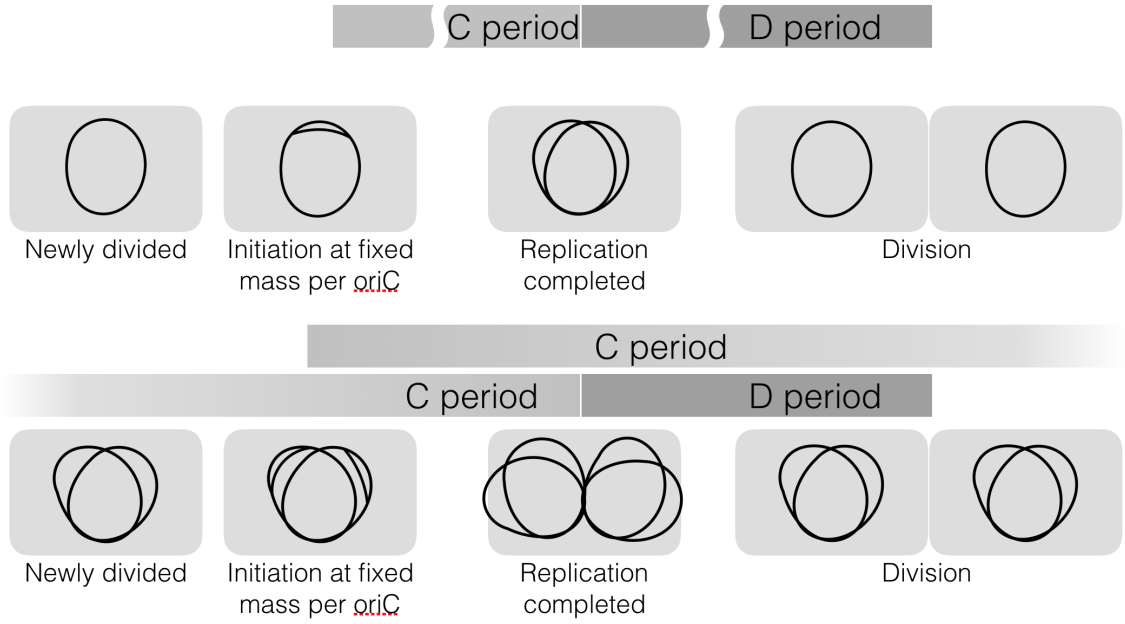


Figure 4.9: **Schematic of the Cooper-Helmstetter/Donachie model** A schematic of the Cooper-Helmstetter model of synchronized chromosome replication and cell division. Division events occur a fixed time after chromosome replication initiates. This is shown for two growth conditions: (top) slow growth and (bottom) intermediate growth. Chromosome replication initiation occurring once per generation results in overlapping cycles of replication at intermediate and fast growth.

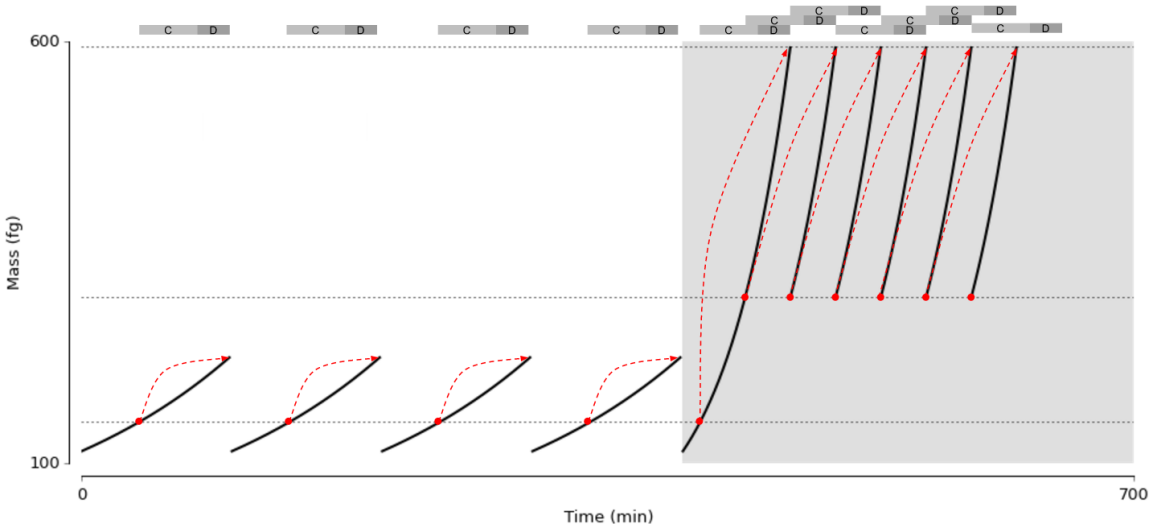


Figure 4.10: **Coordination of chromosome replication and cell division** Simulated cell growth and division for an idealized lineage of cells with no cell-to-cell variation in growth rate or initial size. The cell lineage goes through an up-shift in medium from a 100 to a 30 minute doubling time. Chromosome replication is initiated at a fixed volume per origin of replication (red dot), and the cells divide a fixed period of time (C+D = 60 min) later. The cell division event coupled to the chromosome replication initiation event is indicated by the dashed red arrow. C and D periods are indicated across the top of the plot. During slow growth rounds of chromosome replication are distinct and during intermediate growth they are overlapping.

4.2.2 Model construction and algorithm

The model I constructed is inspired by and adapted from Wallden *et al.*'s published model of chromosome replication initiation and cell division [191]. This adapted model was implemented in the whole-cell modeling framework built for *E. coli* presented in Chapter 5 with a few key and significant differences.

Chromosome replication initiation and enzymatic duplication

As mentioned in the background for this section, chromosome replication initiation occurs at a constant cell mass per origin of replication. I directly use this criteria in the *E. coli* model. This criteria can be summarized with the following relationship where $m_{critical}$ is constant for all cells across all growth rates.

$$\text{Trigger DNA replication initiation} = \begin{cases} True, & \text{if } \frac{m_{cell}(t)}{n_{origin}(t)} \geqslant m_{critical} \\ False, & \text{otherwise} \end{cases} \tag{4.15}$$

Where $m_{cell}(t)$ is the mass of the cell at time point t, $m_{critical}$ is the constant critical mass at

Parameter	Symbol	Value	Units	Source
Critical mass 100 min doubling time	$m_{critical}$	600	fg	Semi-quantitative fit
Critical mass 44 and 24 min doubling time	$m_{critical}$	975	fg	[60]
C period	C	40	min	[136]
D period	D	20	min	[136]
Doubling time	τ	24-100	min	[32]
Maximum replication fork elongation rate	e	967	nucleotide/sec-fork	[32]

Table 4.5: Parameters used in DNA replication initiation model

which chromosome replication initiation occurs, and $n_{origin}(t)$ is the number of origins of replication in the cell at time point t. Values used for $m_{critical}$ can be found in Table 4.5.

When chromosome replication initiation is triggered a pair of replication forks are formed around each origin of replication (one for the forward and one for the reverse stand). This doubles the number of origins and halves the quantity $\frac{m_{cell}(t)}{n_{origin}(t)}$ immediately after initiation occurs. In order for another round of replication to be initiated sufficient cell mass will have to accumulate again. This relatively simple criteria buffers against cellular stochasticity by de-coupling the point in a cell cycle that chromosome replication begins from the amount of time that has passed since the last cell division event. This is critical for a stable cell size distribution to form and ties growth (via accumulation of mass) to the chromosome replication cycle, and as we will see in the next subsection cell division. A detailed algorithm for chromosome replication initiation can be found in Appendix B Algorithm 17.

After chromosome replication initiation the dynamics of replication fork procession from the origin of replication to terminus is modeled as described in Chapter 5. Briefly, replication forks are modeled as point objects that polymerize deoxynucleotides based on *E. coli*'s DNA sequence. None of the enzymatic mechanisms of DNA polymerization are explicitly accounted for (for example, Okazaki fragments, helicases, etc.). The rate of replication forks procession is assumed to limited by a maximal experimentally observed elongation rate and the availability of deoxynucleotides (See Table 4.5 for elongation rate). Once a pair of forks terminates it is assumed that decatination occurs instantaneously.

Cell division

Cell division is the process by which a mother cell divides into two daughter cells, with each daughter receiving roughly half of the contents of the mother cell. The timing and frequency of cell division is critical for a population of cells maintaining it's optimal size distribution for a given environment.

Cooper & Helmstetter proposed in 1968 that cell division occurs at a fixed time period after chromosome replication initiation. This time period consists of the interval to replicate the chromosome (C period), followed by the interval for cytokinesis and cell division (D period) [50]. This

hypothesis was the basis for the model I implemented that determined the timing for cell division in the *E. coli* model.

The time for the C period naturally falls out of the *E. coli* model as the amount of time required for chromosome replication to occur. In the model it takes roughly 40 minutes for a pair of replication forks to traverse the chromosome assuming no deoxynucleotide limitation. This is close to what is experimentally measured [137].

The time for the D period was chosen to be a constant 20 minutes (D_{time}) [137]. Although the D period varies in length depending on growth rate, and is exponentially distributed [27], the molecular mechanisms that determine the length of the D period are not completely known and hence are not modeled. This did not significantly affect the simulated output and findings. Instead when a round of chromosome replication terminates the simulation time at which termination occurred is logged (t_{term}). When the simulation reaches the step which corresponds to the time $t_{term} + D_{time}$ the current cell divides. A detailed algorithm for this cell division trigger can be found in Chapter 5 Appendix C Algorithm 19 alongside the algorithm used to binomially divide the contents of an *E. coli* cell in Algorithm 20 (both implemented by me).

It is important to note that the C and D period interval does not necessarily need to occur within one cell cycle and can span multiple division events. During fast growth (doubling time 22 minutes) on defined rich media with glucose and all 20 amino acids an average replication initiation event in a cell will trigger cell division in it's granddaughter's generation. We will see in Section 4.4 this has important consequences.

Model initialization

In *E. coli* there are potentially multiple rounds of replication proceeding simultaneously at any point in the cell cycle. The simulation begins immediately after cell division and the number and position of any replication forks that are inherited from previous generations must be determined to correctly initialize the simulated cell.

First the number of rounds of replication that on average need to proceed simultaneously can be estimated in an average cell in a population using the C and D period as well as the expected doubling time given that environment. The number of simultaneous rounds (n_{limit}) can be calculated with Equation C.4 as the ratio of C+D period over the doubling time. Because we are considering a specific cell and not an average of a population of cells the number of rounds of replication needs to be an integer.

$$n_{limit} = floor(\frac{C + D}{\tau}) \tag{4.16}$$

For every round of replication proceeding there is a pair of replication forks and a pair of origins of replication. We are assuming that on average a cell after division has inherited one chromosome molecule (i.e. no more than one terC), and that may have more than one round of replication

proceeding on it (i.e. number of oriC ≥ 1). Therefor the number of origins of replication (n_{origin}) is defined by Equation C.5.

$$n_{origin} = 2^{n_{limit}} \tag{4.17}$$

Finally, the position between the oriC and the terC of each replication fork needs to be determined on average. This can be calculated with Equation C.6 where f is the fraction of length between the origin and terminus of replication that the replication fork has proceeded for the nth round of replication (where n can be any integer value between 0 and n_{limit}).

$$f = 1 - \frac{n \cdot \tau - D}{C} \tag{4.18}$$

The position in nucleotides (l) can then be calculated from Equation C.7 where L is the total length of the chromosome in *E. coli*.

$$l = f \cdot \frac{L}{2} \tag{4.19}$$

Proper initialization of the cell ensures the simulation begins close to the steady state of the system, although in practice the simulation is relatively stable. Perturbations in the ratio of cell mass to number of origins of replication quickly re-converge to steady state for a given environment. Values used for all parameters for initialization can be found in Table 4.5. Detailed algorithm for chromosome initialization can be found in Chapter 5 Appendix C Algorithm 5.

4.2.3 Model improvements, validation, and predictions

Model improvements over previous work

Comparison to Wallden *et al.* model As mentioned above Wallden *et al.* implemented a similar less integrated model of chromosome replication initiation and cell cycle control in *E. coli* [191]. A number of changes and improvements have been made which I detail here, all of which involve the more integrated nature of the whole-cell modeling frameowork used to develop the *E. coli* model detailed in Chapter 5.

In the previous work the accumulation of cell mass was approximated by an exponential function with a constant growth rate. The growth rate was varied cell to cell by sampling an experimentally measured distribution of single cell growth rates. This cell to cell variation in growth rate was a critical aspect of Wallden *et al.*'s model and produced cell-to-cell variation in division timing and cell size (discussed in greater detail in Section 4.3 and Section 4.4). In the current work cell mass is added via other physiological processes in the whole-cell modeling framework: metabolism, transcription, and translation, etc.. This allows other physiological processes to impinge on and affect the timing of replication initiation and cell division and adds additional stochasticity to the growth rate and mass

accumulation in a mechanistic fashion. For example, if a low copy number enzyme becomes rate limiting for growth after chromosome replication initiation occurs that cell will divide at a smaller final mass.

Furthermore, in the previous work the C and D period of the cell was a fixed number that varied cell to cell as a function of growth rate and didn't interact with any other physiological processes in the cell. In the current work the C period is explicitly modeled as the time required to polymerize the sequence of the chromosome from dNTPs produced by metabolism and includes variable rates of polymerization due to multiple replication forks. This allows the rate of supply of nucleotides and/or the availability of proteins that make up the replication machinery to affect the C period. If a required enzyme for dNTP biosynthesis becomes the rate limiting step in DNA replication (i.e. knockout, low expression, etc.) this will cause a mechanistically modeled increase in the C period length of the simulation.

Comparison to *M. genitalum* by Karr *et al.* model Karr *et al.* from our research group constructed the first whole-cell model in the organism *M. genitalum* [101], which represents the former state of the art in whole-cell modeling. While the model in Chapter 5 is not a complete whole-cell model, it does represent a significant advance in terms of cell cycle control and simulation stability. The *M. genitalium* model could only grow stably for a single generation due to not having a numerically stable and biologically motivated division criteria. The simulated *M. genitalum* cell divided when its mass reached twice it's initial value. This criteria essentially turns the population cell size distribution into a random walk where the variance grows as the number of generations increases. Since the *M. genitalium* model was publish a significant amount of work on single-cell division dynamics has been published [173, 191, 177] and is incorporated into the *E. coli* model. This enabled us to simulate thousands of *E. coli* cells over hundreds of generations producing stable, reproducible, and realistic cell size distributions (See Section 4.4).

Model validation

In order to examine the dynamics of simulated chromosome replication I generated simulation output under a number of growth conditions and perturbations. The approach detailed in this section qualitatively correctly reproduced expected fork dynamics and cell division timing. In medium environments that supported slow growth (anaerobic glucose minimal medium) a single round of chromosome replication occurred with long periods where no DNA synthesis took place. During intermediate growth periods where multiple replication forks existed as well as periods where no DNA polymerization both occurred (aerobic glucose minimal medium). Finally, under conditions of rapid growth multiple simultaneous rounds of chromosome replication proceeded simultaneously causing cell division events multiple generations after chromosome replication was initiated (all shown in Figure 4.11).

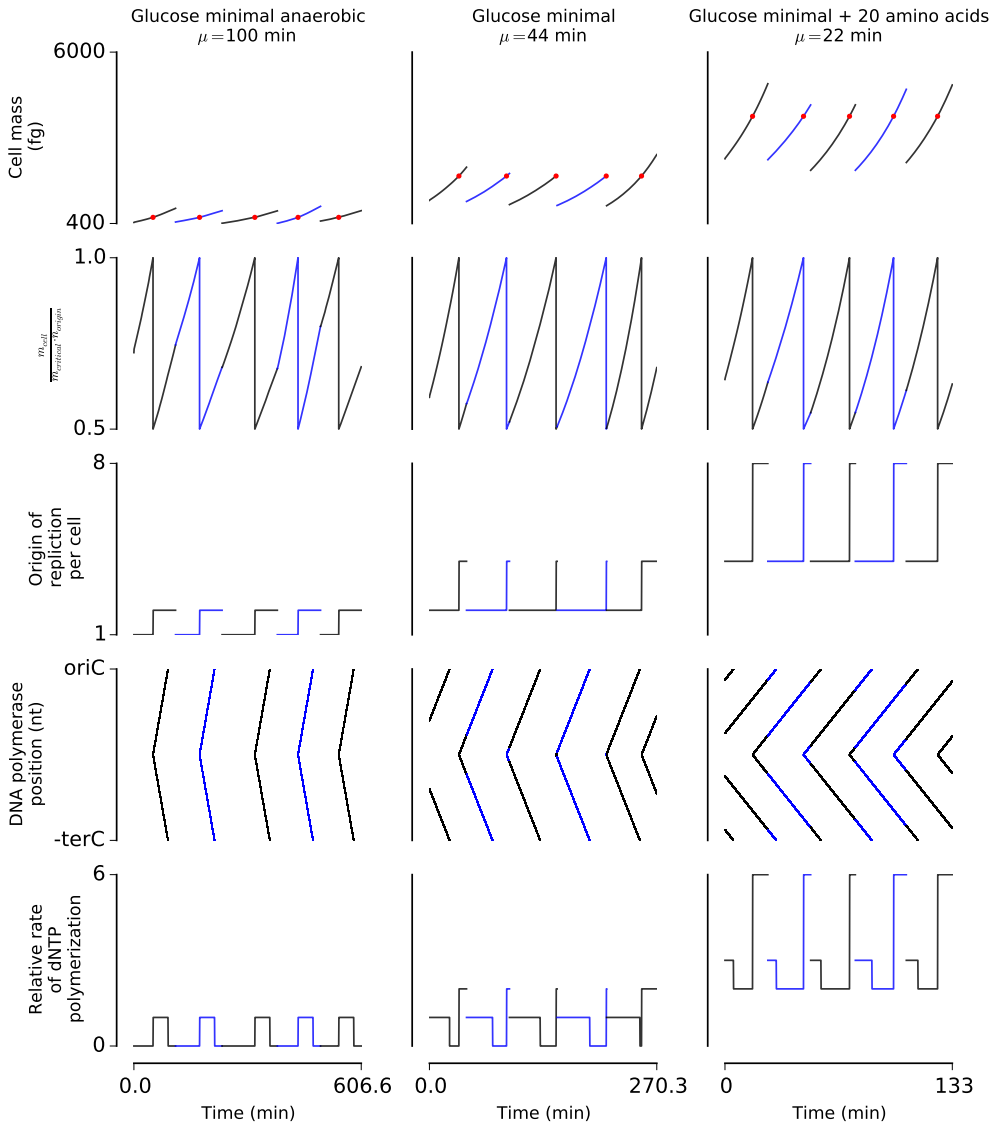


Figure 4.11: **Coupling of replication and cell division cycles in *E. coli* across three growth conditions** Simulation output for three different growth conditions: (A) anaerobic glucose minimal medium (B) glucose minimal medium and (C) glucose minimal medium with all 20 amino acids. (Row 1) Replication is initiated at a fixed mass per origin of replication (red circles), and cells then divide at a fixed interval of C+D period time later. (Row 2) Quantity from Equation 4.15, when it reaches 1 a round of replication is triggered and the number of origins is doubled, reducing it to 0.5. (Row 3) Number of origins of replication (Row 4) Position of replication forks. (Row 5) Number of pairs of replication forks.

Next, I examined the response of chromosome and cell division dynamics to an up-shift in medium conditions. This up-shift caused an increase in growth rate and also a stable increase in cell size due to the lag in chromosome replication initiation behind the increase in growth rate. The model correctly scaled the rate of chromosome replication initiation, and number of origins of replication per initiation event to match the growth rate both before and after the shift in environment. Furthermore, the critical chromosome initiation mass correctly shifts upwards as the size of the cell and increases relative to the number of origins of replication, which enables the model to correctly reproduce the expected average cell size for the more favorable media condition (Figure 4.12).

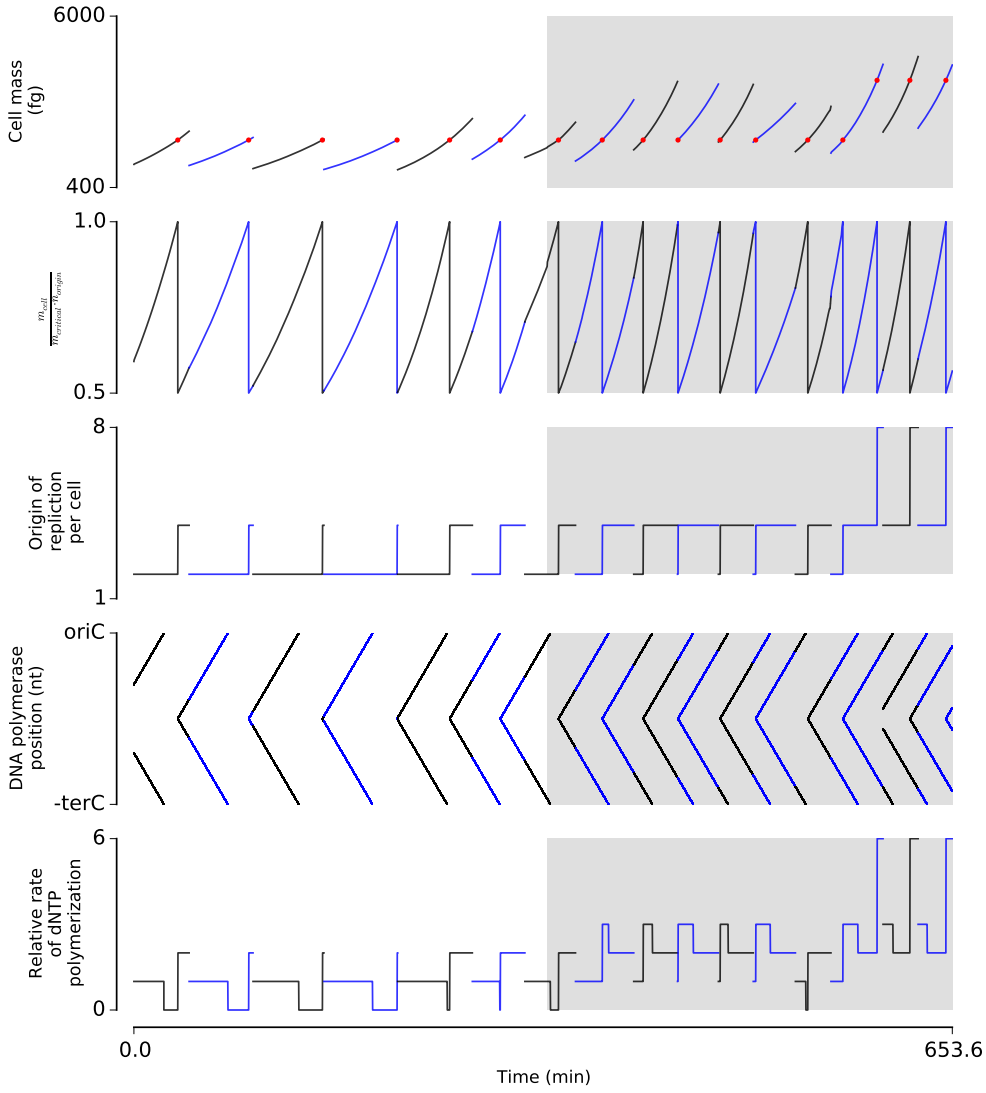


Figure 4.12: **Up-shift in medium with coupling of replication and cell division cycles in *E. coli*** Simulation output with shift in medium from glucose minimal to glucose minimal with all 20 amino acids supplemented (denoted by gray box). The shift in medium causes an increase in growth rate and automatic adjustment in the rate of chromosome replication initiation and coupled cell division. Importantly in the top traces you can see that the critical chromosome initiation mass shifts upwards as the size of the cell and increases relative to the number of origins of replication. See Figure 4.11 caption for explanation of traces.

As a final validation I compared measured population level cell cycle parameters to small populations of simulated *E. coli* cells. The average number of origins per cell, average number of replication forks per cell, average termini, and average number of origins are chromosome replication initiation all are in agreement with the trend shown for the experimentally measured values. This demonstrates that the model of replication initiation coupled to cell division coordinates doubling time and chromosome attributes across multiple growth conditions correctly (Figure 4.13).

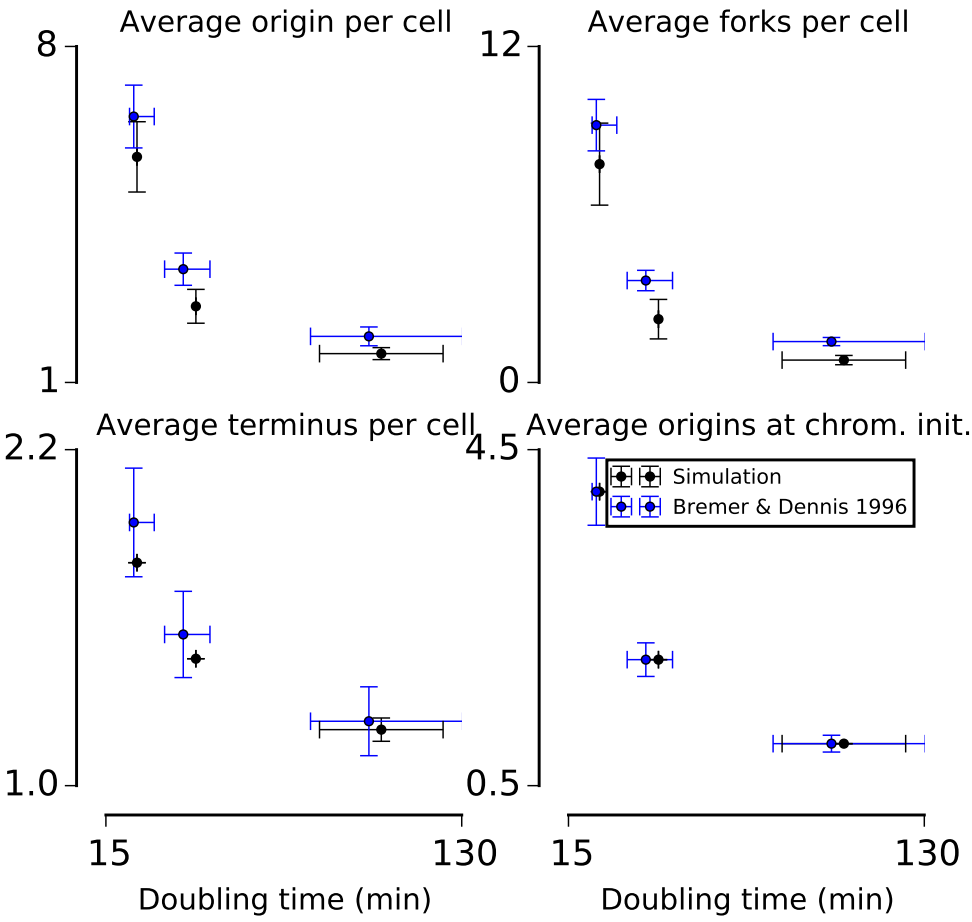


Figure 4.13: **Comparison of simulated and experimentally cell cycle parameters in *E. coli*** Small populations of cells (n=8 per condition) were simulated for three growth conditions and compared by doubling time to data from Bremer and Dennis 1996 [32]. Dependent axis error bars on Bremer and Dennis data were estimated as 10% based on measurement error in C and D period [32]. Independent axis error bars on Bremer and Dennis data were estimated by interpolating data on single cell variability in growth rate from Wallden *et al.* [191].

Model predictions

Balanced growth requires that on average at a population level there is one chromosome replication event per cell cycle. At the single cell level deviations from this criteria are possible and predicted by the *E. coli* simulation. Under certain growth conditions the *E. coli* simulation predicts stochastically varying numbers of chromosome replication initiation events per cell cycle with either zero, one, or two happening per cell cycle (see Figure 4.14). These results can be extended to hundreds of simulated cells under different growth conditions. In Figure 4.15 lineages of cells across three medium conditions and growth rates are shown. At faster growth rates, when more overlapping rounds of chromosome replication and cell division are occurring, a small but significant fraction of the population will have either zero or two initiation events per cell cycle. Previous experimental findings suggest that in synthetic rich medium (doubling time 22.5 min) up to 46% of cells showed multiple rounds of inititation per cell cycle, in glucose minimal medium (doubling time 37.7 min) up to 33%, and in sorbitol minimal medium (doubling time of 50.8 min) up to 6% [172]. Although the exact numbers do not match well with the *E. coli* model's predictions the trend still holds: at faster growth rates with multiple overlapping rounds of chromosome replication more cells will show multiple initiation events per cell cycle. Possible sources of the discrepancy between the *E. coli* model's findings and these inferred experimental results could be additional sources of noise in the chromosome replication initiation mechanism that were not accounted for in the model including noise around the initiation mass and D period.

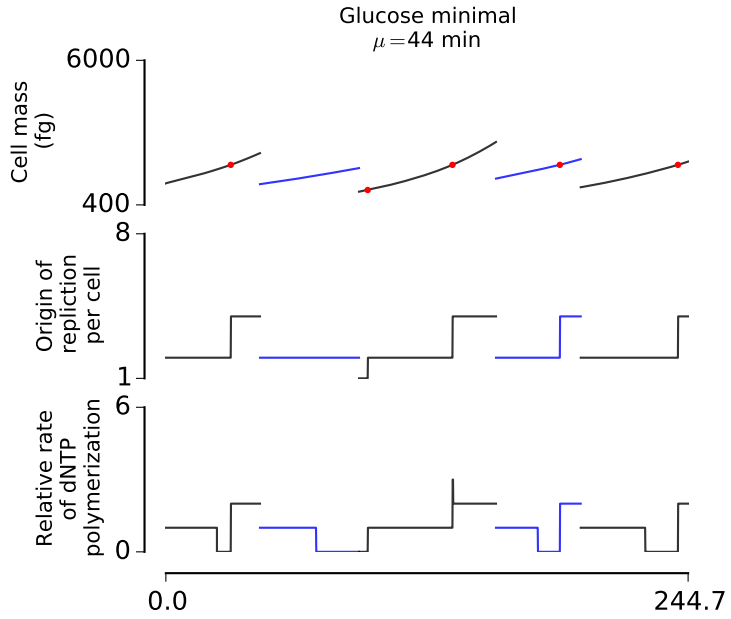


Figure 4.14: **Simulated *E. coli* cells show zero, one, and two rounds of chromosome replication initiation per cell cycle and demonstrate stability to cellular stochasticity** Simulated dynamics of chromosome replication synchronized with cell division predict some cells have multiple rounds of replication per cell cycle, and some cells have none. Cell-to-cell variability in growth rate (See Section 4.3) allows some cells to grow quickly enough to initiate two rounds of chromosome replication per cell cycle. Other cells grow slow enough after dividing that no rounds of chromosome replication begin. In this particularly noisy lineage of simulated cells the first generation initiates 1 rounds of chromosome replication, second generation 0, third generation 2, fourth generation 1, and fifth generation 1. Population level statistics on this phenomena are shown in Figure 4.15.

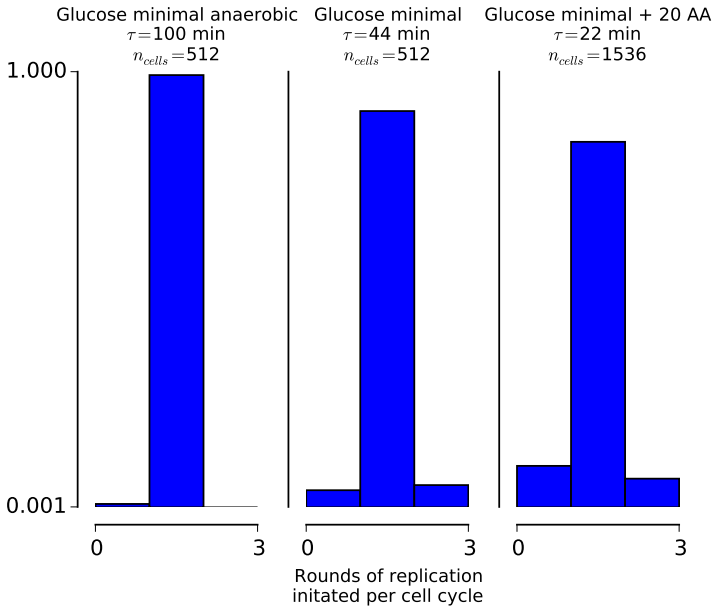


Figure 4.15: **Population level statistics of simulated *E. coli* cells show zero, one, and two rounds of chromosome replication initiation per cell cycle** Population level statistics of lineages like those shown in Figure 4.14 under three growth conditions. The *E. coli* model predicts that as the growth rate increases the chance that a cell will not perfectly synchronize its chromosome replication cycle with cell division increases.

4.3 Cell to cell variability in growth rate

4.3.1 Background: cell-to-cell variation in growth rate

Recent advances in microscopy, microfluidics, and computational image analysis have enabled quantitative, chronic, and high throughput observation of populations of single bacterial cells. These technological and experimental advances allow the direct characterization of cell sizes, lifetimes, and growth rates for exponentially growing, non-synchronized cells under well-controlled growth conditions [191, 177, 173]. Using these techniques cell-to-cell variation has been observed in cell size, division timing, division ratio, and growth rate within isogenetic populations. For example, in a population of *E. coli* in a constant environmental up to a 20% coefficient of variation in single cell growth rate has been measured [173, 191].

This cell-to-cell variability in growth rate is significant because it has been shown to be the driving source of noise in cell division timing and cell size [191] as well as presumably other physiological processes, and population level phenomena. The effect of cell-to-cell variation in growth rate on cell size distributions will be discussed further in Section 4.4. Here I consider the question: what

determines cell-to-cell variation in growth rate?

Why do not all genetically identical cells grow at the fastest possible rate for a specific growth environment, when cells with this growth rate will presumably out-compete their siblings and eventually take over the culture? The answer likely is that the cost of maintaining such tight, optimal control of cellular composition in response to the environment results in a high fitness cost. Furthermore, variation in growth rate, composition, and physiological state could be competitively advantageous to cells under conditions where persister formation [11] or adaptation to a varying environment [64] is favorable.

The growth rate correlation between mother and daughter cells is rapidly lost over generations [191]. Based on this data one hypothesis is that cell division itself causes suboptimal composition due to unequal partitioning of cellular components [85]. Naively one would assume that the contents of a cell is divided roughly binomially but experimental evidence and theory show that this is not the case and may be the source of suboptimal composition. In fact daughter cells unequally inherit important components such a ribosomes, cytoplasmic volume, and presumably others.

4.3.2 Background: hypotheses for cell division induced growth rate variation

Asymmetric inheritance of ribosomes at cell division has been observed experimentally multiple times [191, 41]. This arises from ribosomes having a skewed and noisy localization distributed on either side of the septal ring [41]. However when Wallden *et al.* tested whether the uneven inheritance of ribosomes correlated with cellular growth rate they found a very weak dependence that was not sufficient to explain the cell-to-cell variability in growth rate observed [191]. Interestingly, based on their data it is clear the stochastic, uneven partitioning of ribosomes between daughter cells is endemic and a common issue that bacteria must address and must be robust against. Even more interestingly, varying the concentration of ribosomes on the single cell level does not necessarily vary the growth rate of the cell. Using the material from Section 4.1 one possible hypothesis from these data is that if a cell inherits a supra-optimal concentration of ribosomes, the per ribosome elongation rate should drop assuming that the metabolic capacity of the mother cell is equally divided. Another possible hypothesis is that the average ribosome elongation rate varies due to single cell non-optimal ratios between ternary complexes of EF-Tu, amino-acylated tRNA, and GTP, which could arise due to asymmetric inheritance of ribosomes and more symmetric inheritance of ternary complexes [106]. Both diffuse slowly in the cytoplasm but ribosomes have a diffusion coefficient that is an order of magnitude lower than ternary complexes [94, 106]. None of these hypotheses have been experimentally tested and currently would be difficult to test or validate in the *E. coli* model as constructed.

Another potentially significant source of overlooked cell-to-cell variation is the unequal inheritance of cytoplasmic volume by daughter cells. This is particularly important when considering

the chromosome. At cell division, even if cytoplasmic volume is unequally inherited, each cell must inherit one full chromosome. Division ratios varying between 0.4 - 0.6 have been reported [177], and cells that inherit a smaller fraction of cytoplasm will have a higher concentration of chromosome material, promoters, DNA binding sites, etc. and those that inherit more will have lower concentrations. In particular promoter concentration can have a significant effect on transcription rates. The higher the concentration of promoters in a cell, the more RNA polymerase is promoter bound or transcribing, and the less is in the free cytoplasmic state. The rate of transcription initiation at a given promoter is a saturation function of the concentration of free RNA polymerase. Hence titrating the concentration of promoters titrates the rate of transcription initiation by titrating the concentration of free RNA polymerase [30]. Changes in both global and promoter specific transcription rates could potentially have a significant effect on growth rate and cell-to-cell variability. In addition to the chromosome concentration effects, the chromosome occludes a certain fraction of the cytoplasmic volume from ribosomes and translational machinery, which could also affect the rate of transcription, and the cellular growth rate.

4.3.3 Model construction and algorithm

Based on the experimental data available it is impossible to mechanistically say what causes cell-to-cell variation in composition or growth rate. Many different mechanisms were attempted in the *E. coli* model presented in Chapter 5 but none recapitulated the data observed or produced stable balanced growth on average. In light of this the most straightforward model was implemented.

After each cell division event the rate of metabolic supply of amino acids to translation (S from Equation 4.9) is varied randomly with a Gaussian distribution around the expected value for the medium environment. This can be written as follows:

$$S_{noise} = S \cdot \eta \tag{4.20}$$

Where S_{noise} is the random single cell value distributed around S the expected average, and η is a Gaussian distribution centered around unity. The variance was chosen to be 0.25, which was semi-quantitatively fit to produce the magnitude of experimentally observed cell-to-cell variation in growth rate.

4.3.4 Model results

The major result of this model is that each *E. coli* simulation now has its own, stochastically determined single cell growth rate. This can be seen in Figure 4.16. One set of simulations has its own individual, stochastically distributed around the expected, growth rate. The other set of simulations each single-cell growth rate is exactly what it is expected to be given the growth medium, with minor stochastic variation around this number. The consequences of this for recapitulating

experimentally observed phenomena will be more completely explained in Section 4.4.

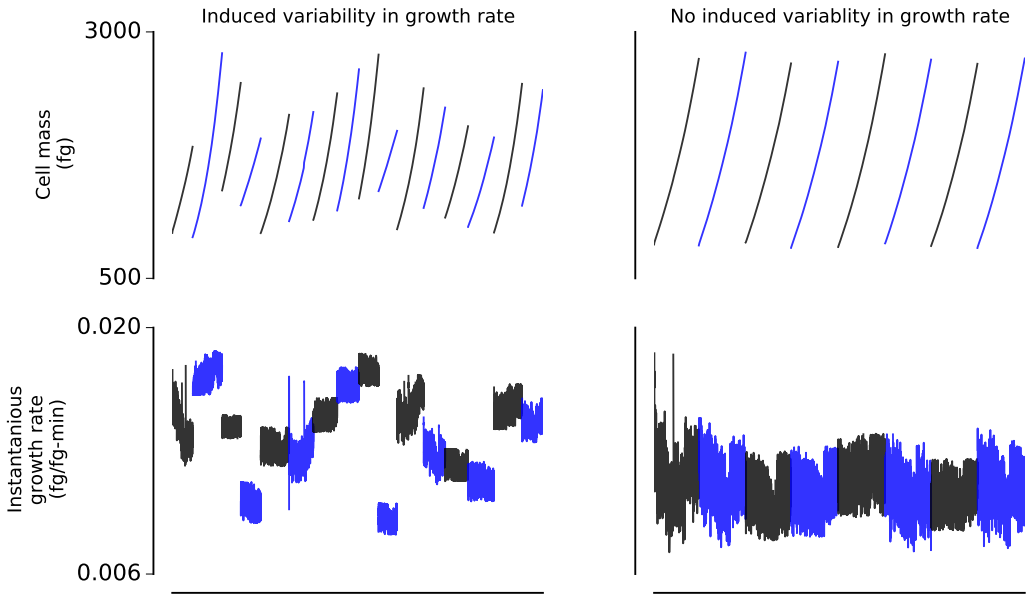


Figure 4.16: **Cell-to-cell variability in single cell growth rate** (A) Simulation traces of total cell mass with added noise in single cell growth rate. Within a single simulation the growth rate is roughly constant but each simulation has its own intrinsic growth rate. (B) Simulation traces of total cell mass with no added noise in single cell growth rate. The curvature of every trace is roughly constant and identical across all simulations.

4.4 Model of cell size maintenance

Cell size control reflects a balance between cellular growth and division. This process of cell growth and division is subjected to the stochastic perturbations of cellular noise, and thus mechanisms must be in place to ensure that cells narrowly distribute around a characteristic average cell size which for bacteria is generally determined by their environment.

For bacteria this regulation is most important in exponentially growing cells where smaller cells grow more slowly than larger cells (See Equation B.1) in absolute growth rate, thus any deviations will be compounded every generation. The existence of observed stable cell size distributions in exponentially growing bacterial cells demonstrates the presence of regulatory mechanisms that produce cell size homeostasis in culture.

4.4.1 Background: Cell size maintenance

Recent work has shown that single cells exhibit a variety of phenomenological division criteria [177, 93, 173, 191, 36]. Yu Tanouchi from our research group has demonstrated that the final size a cell reaches is a noisy linear mapping from its initial size [177]. This can be written as a simple equation:

$$m_{added} = a \cdot m_{init} + b + \eta \tag{4.21}$$

Where m_{added} is the mass added to a single cell between birth and division, m_{init} is the initial cell mass immediately after a cell division event, a and b are the slope and intercept respectively, and η is a Gaussian noise parameter.

Three regimes exist using this model [173]:

1. **Adder behavior**: Cells add a constant size between birth and division, irrespective of the birth size. This corresponds to a slope of $a = 0$. This cells larger than this uncorrelated added size will less than double and cells smaller will more than double in one cell cycle.

2. **Sizer behavior**: Cells add a size that is anti-correlated with their initial size, such that all cells divide at the same size. This corresponds to a slope of $a = -1$. Smaller cells add more mass and larger cells add less mass to achieve the same final mass.

3. **Timer behavior**: Cells grow for a constant amount of time between birth and division. This corresponds to a slope of $a = 1$. Smaller cells add less mass and larger cells add more mass due to all cells growing exponentially (i.e. proportional to their current mass) for a fixed amount of time per cell cycle.

No species or population of cells exhibit perfect adder, sizer, or timer behavior but exist on a continuum. For example, fission yeast like *Saccharomyces pombe* exhibit a slope of -0.76, and *Caulobacter crescentus* exhibit a slope of 0.25 [154]. As we will see below, even within one species *E. coli* exhibits adder behavior at fast growth rates and sizer behavior at slow growth rates [191].

Using a model similar to the one described in Section 4.2 Wallden *et al.* demonstrated that synchronizing chromosome replication initiation and cell division can recapitulate this linear mapping behavior. They were able to show that two regimes existed in *E. coli*: (1) in quickly growing cells replication is initiated one or two generations prior to its coupled cell division event. This means that regardless of a cells initial size, a cell will divide a fixed time after replication initiation, which results in a size expansion that is uncorrelated with its birth volume, and instead depends on growth rate (i.e. "adder" behavior). (2) During slower growing conditions replication initiation occurs during the same generation as its corresponding division event. Since chromosome replication initiation occurs at a fixed mass that is uncorrelated with birth size, and cells grow at roughly equal growth rate

within a condition, a constant amount of mass is added between chromosome replication initiation and cell division, and all cells divide at close to the same size (i.e. "sizer" behavior).

4.4.2 Model validation and results

The maintenance of stable cell size distributions in the *E. coli* model is the result of the combined effects of the growth rate, cell division, and growth rate noise models presented in Sections 4.1, 4.2 and 4.3 respectively. No further work was required to acquire these results.

To validate the *E. coli* model at least 1024 cells were simulated over 4 generations in glucose minimal, glucose minimal with supplemental amino acids, and glucose minimal without oxygen medium. These resulted in average doubling times of 44, 22, and 100 minutes respectively.

These data show that the *E. coli* model produces stable cell size distributions under each medium condition and growth rate that arise purely from the average growth rate in the medium, the average copy number of the chromosome, and the critical mass at which chromosome replication is initiated (Figure 4.17).

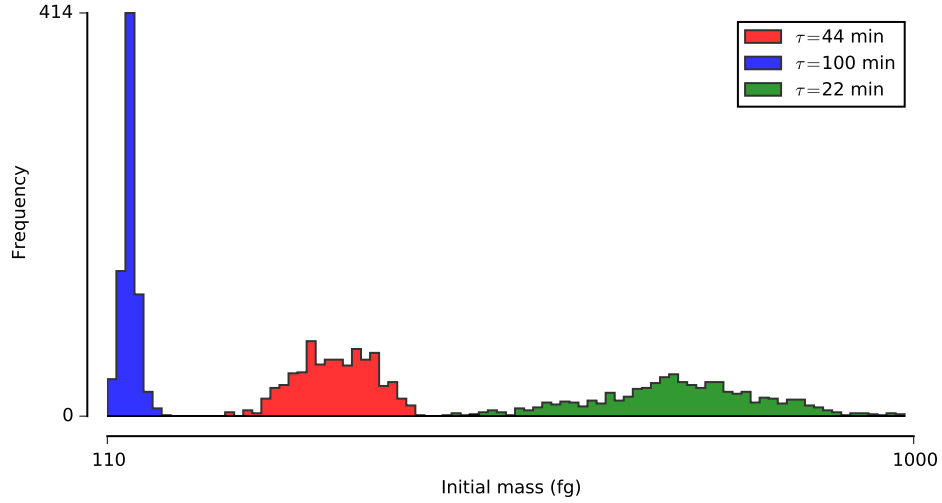


Figure 4.17: **Initial cell size distribution across three medium environments in *E. coli*** Simulated *E. coli* cells across three medium environments: glucose minimal without oxygen (100 min doubling time), glucose minimal medium (44 min doubling time), and glucose minimal medium with amino acids (22 min doubling time). For each condition over 1024 cells were simulated for 4 generations. The transition from sizer to adder behavior is evident from the distributions with slow growing cells exhibiting a narrower distribution and quickly growing cells exhibiting a broader distribution of cell size.

Furthermore the model demonstrated the same transition from sizer behavior under slower growing conditions to adder behavior under quickly growing conditions.

Plotting the data from each condition in the same form as Equation 4.21 (Figure 4.18) we see that the slope a transitions from -0.949 at a 100 min doubling time, to -0.532 at a 44 min doubling time, to 0.107 at a 22 min doubling time (not significantly different than a slope of zero with p-value of 0.117). The simulated cells exhibit a shift from more sizer-like behavior to more adder-like behavior as growth rate increases. This behavior can also qualitatively be observed in Figure 4.17 with slow growing cells exhibiting a narrower distribution and quickly growing cells exhibiting a broader distribution of cell size.

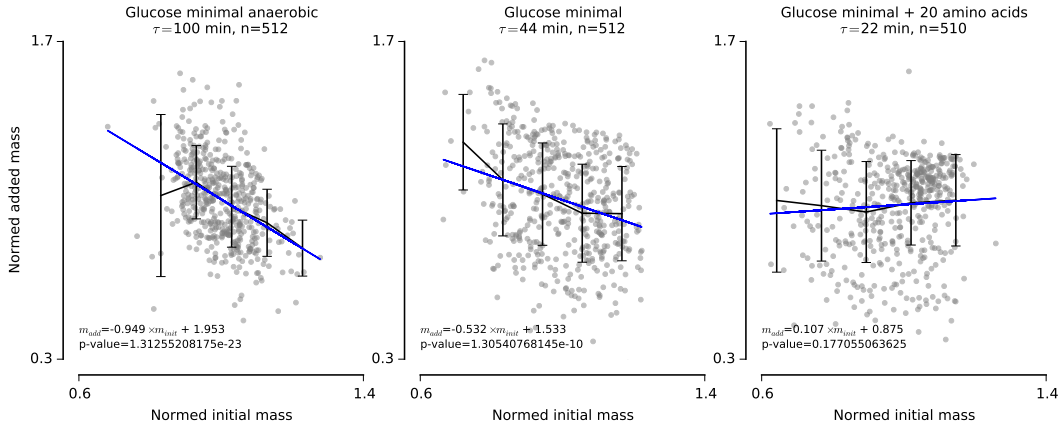


Figure 4.18: **Mass added per cell cycle vs initial mass across three medium conditions** Data showing transition from sizer to adder behavior as doubling time decreases. Each axis was normalized by its mean. Bins were added to points such that each bin had 100 cells on average. Linear regressions show the slope of the linear mapping shifting from -0.949 (95%CI -1.091 to -0.807), to -0.532 (95%CI -0.660 to -0.404), to 0.107 (95%CI -0.018 to 0.231) as the doubling time changes from 100, to 44, to 22 min respectively.

This behavior specifically arises from the interaction between the chromosome replication cycle and cell-to-cell variation in growth rate. Within a growth condition all cells initiate chromosome replication at approximately the same mass. Each chromosome replication initiation event triggers one cell division event after a fixed amount of time (60 min). These two coupled phenomena lead to sizer behavior at slow growth rates and adder at fast growth rates:

1. **Slow growth**: During slow growth chromosome replication is non-overlapping, and a paired replication initiation and cell division event occur within the same cell. It is easiest to think of this growth as occurring in two phases: birth to critical initiation mass, and then critical initiation mass to cell division. Smaller cells need to grow more, and larger cells need to grow less to reach their critical initiation mass. After reaching their critical initiation masses each cell adds roughly the same amount of mass with some variation due to cell-to-cell variability in growth rate. This leads to smaller cells adding more, and larger cells adding less mass (i.e. sizer behavior).

2. **Fast growth**: During fast growth chromosome replication is overlapping, and a paired replication initiation and cell division even might be separated by up to 2 cell division events and changes in growth rate. This means that the size that a cell divides at has little to do with when it reaches its critical initiation mass within its own cell cycle, and consequentially the amount of mass added over a cell cycle is uncorrelated with its initial size (i.e. adder behavior).

Chapter 5

Cricks complete solution of *E. coli*, 40 years later

Four decades ago, Francis Crick advocated for a coordinated worldwide scientific effort to determine a complete solution of *Escherichia coli* [55]. Since then, millions of measurements have been published using this organism. To what extent do these data constitute a solution? Here, we use large-scale modeling to simultaneously evaluate a massive set of heterogeneous experimental results in *E. coli*. We show that these data are strikingly self-consistent with respect to gene and protein expression and cellular metabolism as well as replication and growth physiology. Surprisingly, the totality of the data suggests that a clear majority of *E. coli* genes are expressed less than once per generation - including the genes associated with antibiotic resistance and persistence - and yet the cell is robust to this behavior. Our findings cross-validate experimental measurements made largely independently across the world and over many decades, and suggest several lines of fruitful inquiry for the future.

Crick recommended this coordinated effort be performed both for the intellectual satisfaction of having a single living cell completely explained, as well as in order to make major advances in biology in the most efficient way [55]. He suggested that the best way to accomplish this would be to establish a central laboratory which could coordinate work across nations, standardize biological reagents and materials, and produce vast libraries. Although such an approach was never adopted, the scientific community has performed and published millions of measurements, of many different kinds and in hundreds of laboratories, using E. coli over the intervening decades. To what extent, if any, do these data constitute a solution?

The answer to this question has implications beyond *E. coli* or even microbiology, as high-profile studies published recently have led this journal and others to question the reproducibility of scientific results in multiple scientific fields [48, 14]. If this is the case, then we should worry about whether what we know about *E. coli* is actually correct in two ways. First, are the data reproducible (i.e.,

does a repeated study produce the same measured outcomes)? Second, and more deeply, are the data cross-verifiable - meaning, does a multiplicity of heterogeneous data all point to the same conclusion?

This second and more important question requires us to consider all of the data together in their biological context, integrating and evaluating them as a whole - which in turn depends on theoretical approaches based on mathematical representations of known or inferred biological mechanisms. The application of theory to understand, consolidate, and even verify data is not new. The most famous expression of the latter point is attributed to Sir Arthur Eddington: I hope I shall not shock the experimental physicists too much if I add that it is also a good rule not to put overmuch confidence in the observational results that are put forward *until they are confirmed by theory* (italics his) [63]. In fact, this motto guided Watson and Cricks approach to developing a model of the double helical structure of DNA, for which Crick asserted that people dont realize that not only can data be wrong in science, it can be *misleading* (italics his) [92]. As a result, some data were set aside if they were inconsistent with the theory of the double helical structure.

Per Eddingtons remark, our goal is to use large-scale, integrative modeling of *E. coli*s fundamental biological processes as a theory, to confirm (or at least evaluate) the observational results that have been reported in *E. coli* over the past century. Obviously, the success of this effort depends on a mathematical approach which can both represent these biological processes mechanistically, as well as accommodate many millions of data points, which exhibit an extraordinary amount of heterogeneity. Attempts to mechanistically model cell behavior at the large scale also span several decades [164, 187, 149, 182, 47]. We recently demonstrated a large-scale modeling approach which was capable of integrating all of the known functions in the simplest culturable bacterium, *Mycoplasma genitalium* [101]. Notably, the model successfully reproduced many measured data, and even predicted previously unmeasured parameters which were subsequently verified experimentally [153].

Encouraged by this success, here we apply our approach to *E. coli*, arguably the best characterized organism. *E. coli* has nearly ten times more genes than *M. genitalium*, is comprised of roughly 50-100 times as many molecules that interact and react, can readily grow in a wide variety of environmental conditions, and exhibits extensive self-regulation and control, all of which pose significant challenges to whole-cell modeling. However, one of the most exciting aspects of modeling *E. coli* at the large scale is the enormous effort in data generation that has already been performed - and in many cases, stored in an accessible format [39]. Thus, whereas only 27.5% of the parameter values used in our *M. genitalium* model were actually measured in that organism, the model we describe here contains parameter values that are 100% measured in *E. coli*. This fact provided us an opportunity to assess the data against itself to a degree that would not be possible in any other organism.

In the Supplemental Materials (Appendix C), we describe in detail a model that fully integrates the Central Dogma together with carbon and energy metabolism, and in the context of balanced

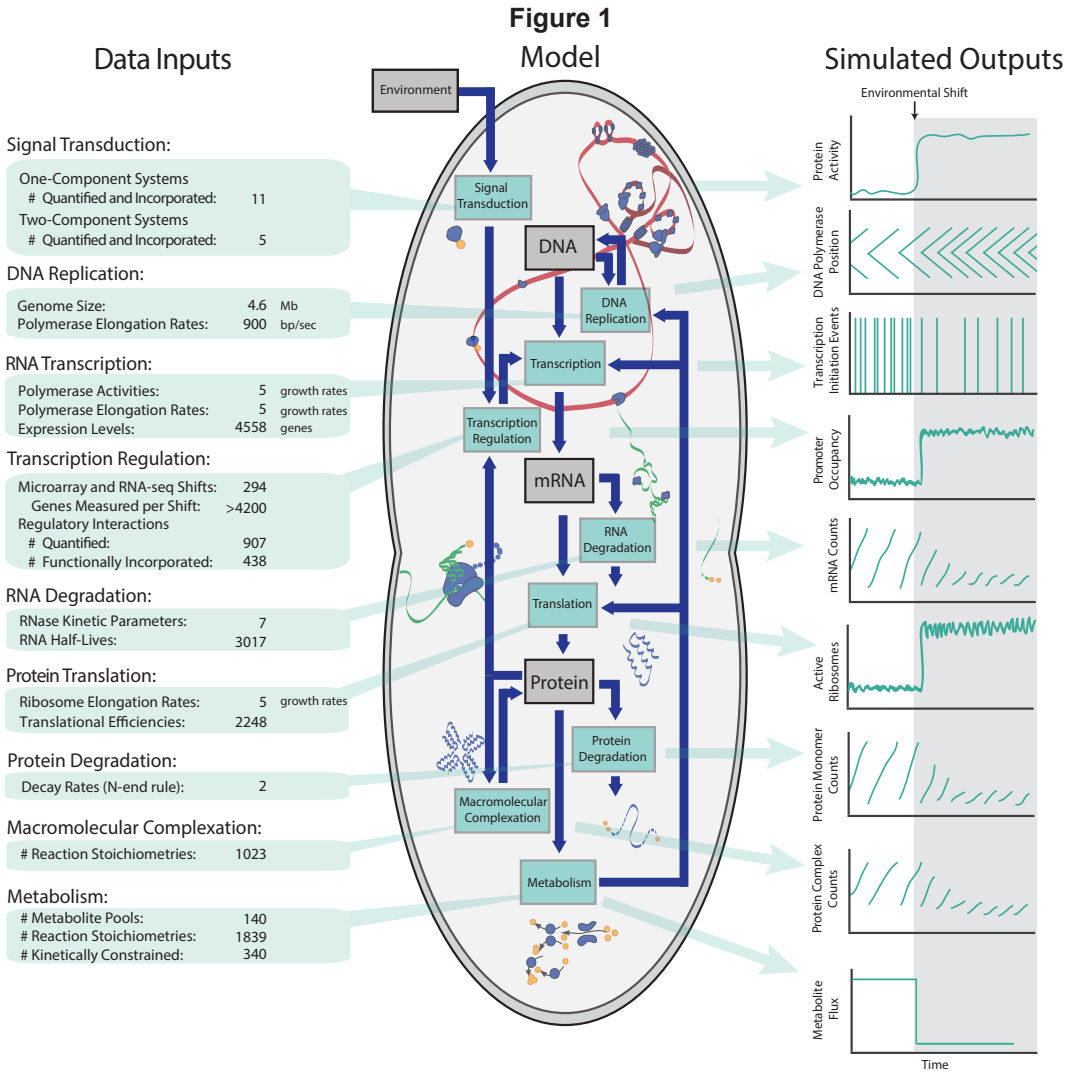


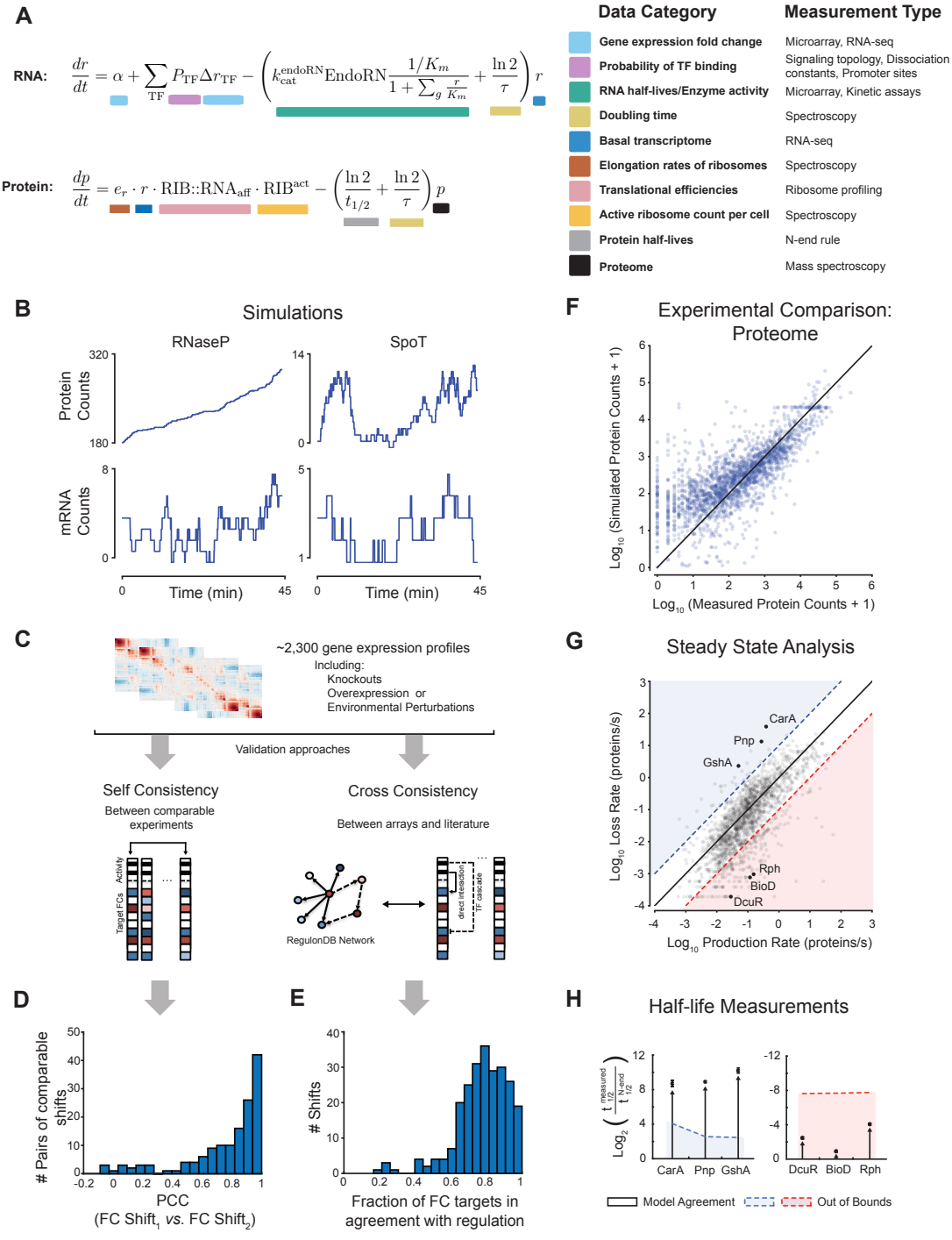
Figure 5.1: ***E. coli* model framework** A computational framework synthesizes several decades' worth of heterogeneous data sets from a single organism into an internally consistent model. The model can simulate the dynamics of multiple aspects of physiology across three environmental conditions.

growth. Functionally, 1,214 genes (or 43% of the well-annotated genes) have been included to represent these processes - which are also those for which the lions share of the existing E. coli data has been generated. The model has been optimized for three different environments, which are relatively well-characterized experimentally and exhibit diverse phenotypic behaviors: a minimal (M9 salts plus glucose, aerobic conditions), rich (minimal + all amino acids), and minimal anaerobic medium. From the totality of the data, over nineteen thousand parameters were identified and included in the model, while <1% were fit - which underscores that it is the data that are being tested by the theory, not vice versa. We then applied this model to assess the cross-verifiability of this massive dataset. We decided on three evaluative criteria: (1) *Equations* - can the data be encapsulated mathematically? (2) *Simulations* - does the model output make sense with respect to known cell behaviors? (3) *Validation* - are the data consistent with one another?

We first evaluated these criteria in terms of what Crick called the Central Dogma, namely the expression of genes in terms of mRNA and protein. Gene expression has been studied in extensive detail, with datasets that characterize mRNA expression under a variety of environmental conditions (including an RNA-Seq dataset that we generated for this study, see Supplemental Materials), mRNA and protein half lives, kinetics of RNase activity, experimentally-determined RNA-polymerase and transcription factor binding sites, dissociation constants for proteins bound to DNA binding sites or other cellular and environmental ligands, transcription unit structure, translational efficiencies of mRNA transcripts, and cellular growth rate and chemical composition measurements (see Supplemental Materials Appendix C for complete descriptions of all data used). In terms of our Equation criterion, we found that these datasets can indeed be integrated mathematically, beginning with a basis of thousands of ordinary differential equations (shown for a representative mRNA and protein in Figure 2A) that we then implemented as stochastic simulations (see Supplemental Materials Appendix C for a more detailed description of the modeling approach).

To address the Simulation criterion, we ran simulations of single cell life cycles and observed whether the parameters could yield plausible trajectories of mRNA and protein production. Representative traces for genes encoding both high- and low-stability mRNAs are shown in Figure 5.2B; these exhibited characteristic expression dynamics wherein mRNA was produced at random times, leading to a concomitant sharp increase in protein production, as has been observed experimentally [176].

Using the model, we also identified a number of ways in which these data could be cross-validated against each other (Figure 5.2C). We first compared expression fold changes that were determined from independent microarray and RNA-Seq datasets (based on an earlier compilation [37]), by identifying experiments that would be expected to have the same or similar gene expression outcomes (e.g., when an environmental stimulus experiment and a separate genetic perturbation experiment modulate the activity of the same transcription factor). In such cases, the mRNA fold changes were strongly correlated (72% of pair shifts have PCC > 0.7 or higher, Figure 5.2D). We then compared


A
RNA:
Protein:
Data Category
Measurement Type
Gene expression fold change
Microarray, RNA-seq
Probability of TF binding
Signaling topology, Dissociation constants, Promoter sites
RNA half-lives/Enzyme activity
Microarray, Kinetic assays
Doubling time
Spectroscopy
Basal transcriptome
RNA-seq
Elongation rates of ribosomes
Spectroscopy
Translational efficiencies
Ribosome profiling
Active ribosome count per cell
Spectroscopy
Protein half-lives
N-end rule
Proteome
Mass spectroscopy
B
Simulations
RNaseP
SpoT
Protein Counts
mRNA Counts
Time (min)
C
~2,300 gene expression profiles
Including:
Knockouts
Overexpression or
Environmental Perturbations
Validation approaches
Self Consistency
Between comparable experiments
Cross Consistency
Between arrays and literature
RegulonDB Network
D
Pairs of comparable shifts
PCC
(FC Shift1 vs. FC Shift2)
E
Shifts
Fraction of FC targets in agreement with regulation
F
Experimental Comparison: Proteome
Log10 (Simulated Protein Counts + 1)
Log10 (Measured Protein Counts + 1)
G
Steady State Analysis
Log10 Loss Rate (proteins/s)
Log10 Production Rate (proteins/s)
CarA
Pnp
GshA
Rph
BioD
DcuR
H
Half-life Measurements
CarA
Pnp
GshA
DcuR
BioD
Rph
Model Agreement
Out of Bounds

the same fold changes to the known regulatory topology extracted from the RegulonDB and EcoCyc databases [104, 75]. This analysis indicated that over 76% of the fold-changes determined from array or RNA-Seq data are consistent with the topology of the transcriptional regulatory and signaling networks (Figure 5.2E). For a third test, we compared the total simulation output for protein counts in a cell to an experimentally-determined proteome which was not used to parameterize the model [157], and found the agreement to be statistically significant (Figure 5.2F, PCC = 0.75, $p < 10^{-20}$, n = 2,233).

Finally, we recognized that under steady-state conditions, the rate of protein synthesis should equal the rate of decay. This proved to largely be the case in our simulations, with 85% of the production rates within an order of magnitude of the decay rate (Figure 5.2G). This was particularly surprising given that protein decay rates are very lightly characterized, and are thus usually estimated by the N-end rule, whereby protein half-lives are assumed to have values of 2 minutes or 10 hours, based on their N-terminal amino acid [13]. We wondered whether some of the outliers in our comparison might be due to a more nuanced or specific value for the protein half-life. To test this hypothesis, we determined the half-lives of six outlier proteins experimentally, and found that in all cases, the half-life predicted by our model was a better predictor of the data than the N-end rule (Figure 5.2H). In three of the cases, the new half-life was sufficient to explain the discrepancy between the protein production and decay rates; for the remaining three, other as-yet unknown factors are also likely to play a role. In total, all of our validation tests and follow-on experiments

Figure 5.2 *(preceding page)*: **Transcription and translation integration and validation** Model-driven analysis and cross-validation of the data associated with Central Dogma-related processes. (A) Representative equations and data sources describing RNA and protein expression for all of the RNAs and proteins in *E. coli*. These equations were the starting point for our stochastic single-cell simulations. (B) Simulated dynamics of mRNA and protein expression for two genes. (C) A pipeline for assessing the consistency of a set of 2,300 RNA expression profiles in terms of self-consistency and cross-consistency. The self-consistency was determined by identifying 144 pairs of experiments that would be expected to have the same or similar results, and comparing the gene expression shifts observed in each, while the cross-consistency was found by comparing the set of expression profiles with the published literature on transcription as compiled in RegulonDB [75], where a given transcription factor has been determined to activate or repress transcription of a given target gene. (D) Histogram showing the Pearson Correlation coefficients between paired experimental outcomes. (E) Histogram showing the fraction of gene expression fold-changes shown to be in agreement with the published literature. (F) A comparison of simulation and experimental results [157] with regard to the number of proteins expressed per cell for each gene. (G) A comparison of calculated protein production rates against protein synthesis rates for each gene. Six outliers are highlighted because we determined their protein decay rates experimentally. (H) Comparison of the N-end rule to new measurements of protein half-lives. The experimental data are represented as a log2 ratio (where the N-end rule half-life is the denominator). The top three outliers from Panel G (in the blue region) had half-lives that were many-fold higher than indicated by the N-end rule; the other three (red) had half-lives that were lower than indicated. In all cases, the model prediction (i.e., white region beyond the dashed line) was closer to the experimentally determined value.

suggested that the datasets included in this model are highly self-consistent with regard to gene expression and the Central Dogma.

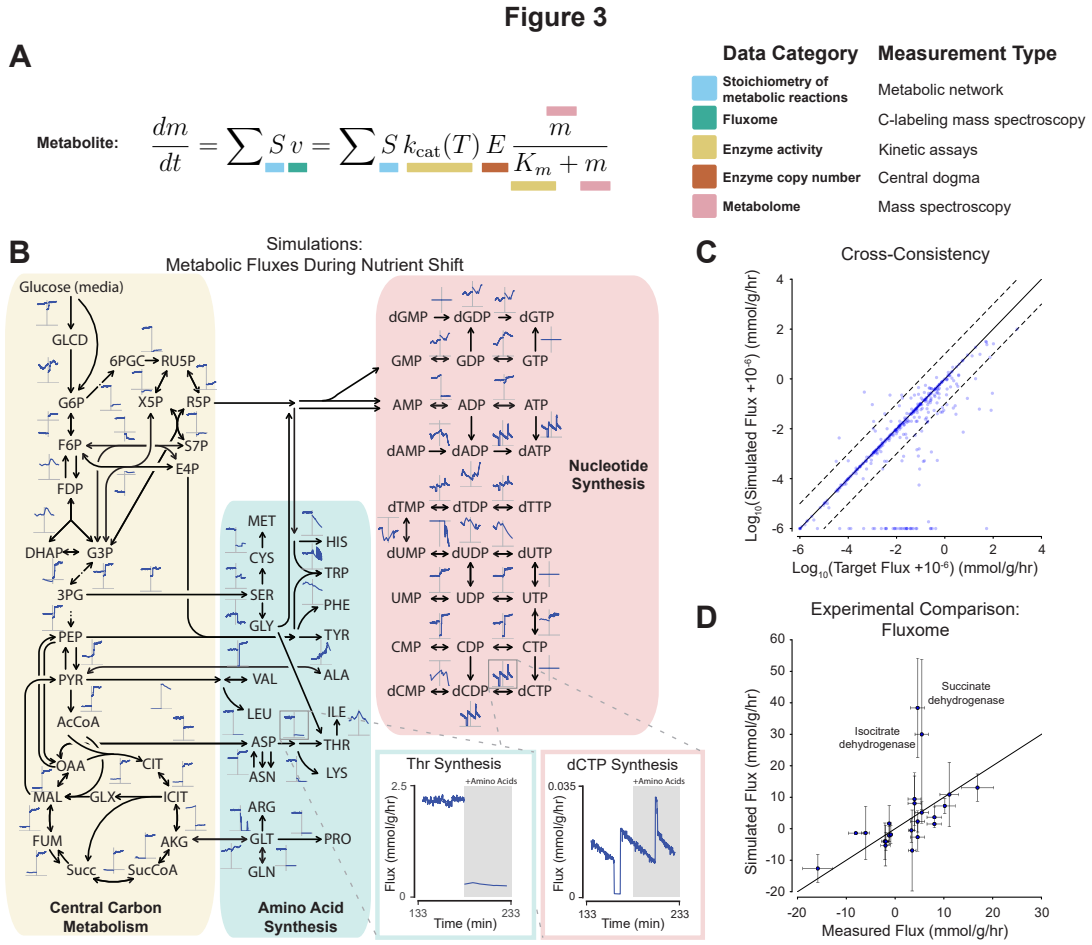
We next considered our evaluative criteria in terms of *E. coli*s metabolic network. This network accounts for 962 enzymes, and has been investigated as part of thousands of studies which determined the stoichiometry for each chemical reaction (as compiled by others [139, 193]), identified and characterized virtually all of the enzymes and transport proteins in the network, and measured concentrations of >100 small molecules internal to the cell [18], as well as physiological properties (e.g., chemical composition of the cell, maximum uptake and secretion rates of carbon sources and by-products), key fluxes in central carbon metabolism, and detailed kinetics for many enzymes. The processes associated with the Central Dogma are also deeply relevant here, as the concentrations of enzymes are used to determine kinetic constraints.

As before, it is possible to represent these networks mathematically by writing thousands more ordinary differential equations which describe the concentrations of small molecules over time (Figure 5.3A) and making subsequent simplifying assumptions in order to model flux through a metabolic network using linear optimization and flux balance analysis [140]. We have found that with significant modifications - particularly with regard to the incorporation of regulation of enzyme expression [53], the inclusion of enzyme kinetic constraints (e.g., based on metabolite concentrations and Michaelis-Menten parameters) [111, 101], and substantial innovation with regard to the objective function and bounds [21] (see Supplemental Material Appendix C for details) - this method can be incorporated into whole-cell modeling.

Referring to the Simulation criterion, we examined the output for the metabolic network. Figure 5.3B depicts a selection of model output (including the flux distribution over time for metabolic fluxes related to central carbon metabolism and amino acid biosynthesis) for the shift that occurs when cells initially growing in an aerobic glucose minimal medium are presented with the full range of amino acids. This perturbation leads to two major shifts in flux. First, the fluxes which normally synthesize amino acids are dramatically reduced. Second, the addition of amino acids leads to a higher growth rate (discussed in more detail below), which in turn drives another set of enzymes to produce more DNA precursors. Both these and the other model outputs are highly consistent with what we would expect.

We then performed validation tests of the metabolic data to determine whether the totality of the constraints imposed on the metabolic network by Michaelis-Menten constants and small molecule concentrations was internally consistent. This required a modification to our linear optimization approach which lightly penalized flux distributions which diverged from these kinetic constraints (see Supplemental Material Appendix C). We found that incorporation of these constraints (400 in total, constraining 340 metabolic reactions) led to a growth yield that was lower than has been measured experimentally. Further investigation identified two metabolic reactions from the Citric Acid cycle for which the incorporation of kinetic constraints reduced the yield (succinate dehydrogenase and

Figure 3


A
Metabolite:
$\frac{dm}{dt} = \sum S\,v = \sum S\,k_{\mathrm{cat}}(T)\,E\,\frac{m}{K_m + m}$
Data Category
Measurement Type
Stoichiometry of metabolic reactions
Metabolic network
Fluxome
C-labeling mass spectroscopy
Enzyme activity
Kinetic assays
Enzyme copy number
Central dogma
Metabolome
Mass spectroscopy
B
Simulations:
Metabolic Fluxes During Nutrient Shift
Glucose (media)
GLCD
6PGC
RU5P
G6P
X5P
R5P
F6P
S7P
E4P
FDP
DHAP
G3P
3PG
PEP
PYR
AcCoA
OAA
CIT
MAL
GLX
ICIT
FUM
AKG
Succ
SucCoA
Central Carbon Metabolism
MET
HIS
CYS
TRP
SER
PHE
GLY
TYR
VAL
ALA
LEU
ILE
ASP
THR
ASN
LYS
ARG
GLT
PRO
GLN
Amino Acid Synthesis
dGMP
dGDP
dGTP
GMP
GDP
GTP
AMP
ADP
ATP
dAMP
dADP
dATP
dTMP
dTDP
dTTP
dUMP
dUDP
dUTP
UMP
UDP
UTP
CMP
CDP
CTP
dCMP
dCDP
dCTP
Nucleotide Synthesis
Thr Synthesis
+Amino Acids
Flux (mmol/g/hr)
Time (min)
dCTP Synthesis
+Amino Acids
Flux (mmol/g/hr)
Time (min)
C
Cross-Consistency
Log10(Simulated Flux +10-6) (mmol/g/hr)
Log10(Target Flux +10-6) (mmol/g/hr)
D
Experimental Comparison:
Fluxome
Succinate dehydrogenase
Isocitrate dehydrogenase
Simulated Flux (mmol/g/hr)
Measured Flux (mmol/g/hr)

fumarate reductase, see Supplement Appendix C). With the constraints on these reactions relaxed, we ran simulations and compared the output metabolic fluxes to our kinetic constraints (Figure 5.3C). Twenty-eight of the fluxes were unused by the simulation (i.e., had a zero value), due to the fact that the model is not yet whole-cell. The remaining fluxes exhibited a remarkable degree of consistency, where 87% were within an order of magnitude of the constraint value (PCC = 0.86, $p < 10^{-100}$, n = 356). Next, we compared simulated results to another independent dataset that was withheld from our model parameterization (Figure 5.3D) [183]. With two exceptions, the simulation and data were highly correlated (PCC = 0.80, $p < 10^{-4}$, n = 21). Interestingly, succinate dehydrogenase was one of the proteins for which the kinetic constraint was relaxed as mentioned above, and the other (isocitrate dehydrogenase) is a near neighbor in the Citric Acid cycle, which invites further investigation. Overall, these results indicate a strong consistency between all of the datasets we compiled relative to metabolism.

Our final evaluation concerned measurements associated with cell growth and physiology. We were able to encapsulate data on cell physiology across multiple growth conditions in a set of equations describing the relationship between cell size, growth rate, ribosome concentration, ribosome elongation rate, the rate of DNA initiation, and other cell cycle parameters (Figure 5.4A). Considering our Simulation output and experimental Validation criteria, we examined the physiological response to the same environmental shift depicted in Figure 5.3B for metabolism (i.e., adding amino acids to a defined minimal medium), and found that this simulated up-shift in nutrient availability led to a steady-state increase in cell size, growth rate and ribosome concentration as expected (Figure 5.4B). The dynamics of the response also met expectations, with the RNA fractions instantaneous growth rate changing to the new expected growth rate relatively quickly, while other mass fractions and the total cell growth rate converged more slowly (Figure 5.4C), as others have observed [65]. These simulation outputs were also consistent with experimental measurements of the RNA mass per cell, ribosome elongation rate and stable RNA synthesis rate at three different doubling times (Figure 5.4D) [32].

Next, we examined the dynamics of simulated chromosome replication and its response to the same environmental shift. We modeled the initiation of chromosome replication as occurring at a

Figure 5.3 *(preceding page)*: **Metabolic network integration and validation**
Model-driven analysis and cross-validation of the data associated with metabolic processes. (A) Representative equations and data sources describing concentrations for all of the metabolites in our model. These equations were the foundation of a linear optimization-based model. (B) Simulated dynamics of selected metabolic fluxes during a shift from minimal media to minimal media supplemented with amino acids. Two representative plots of flux dynamics are enlarged for illustrative purposes. (C) A comparison of the models simulated flux results with flux values that were calculated from the kinetic data we curated. (D) A direct comparison of simulated flux values to experimental measurements [183]. Two outliers are highlighted for discussion in the main text. Error bars indicate standard deviation (n = 3 for measured flux and n = 32 for simulated flux).

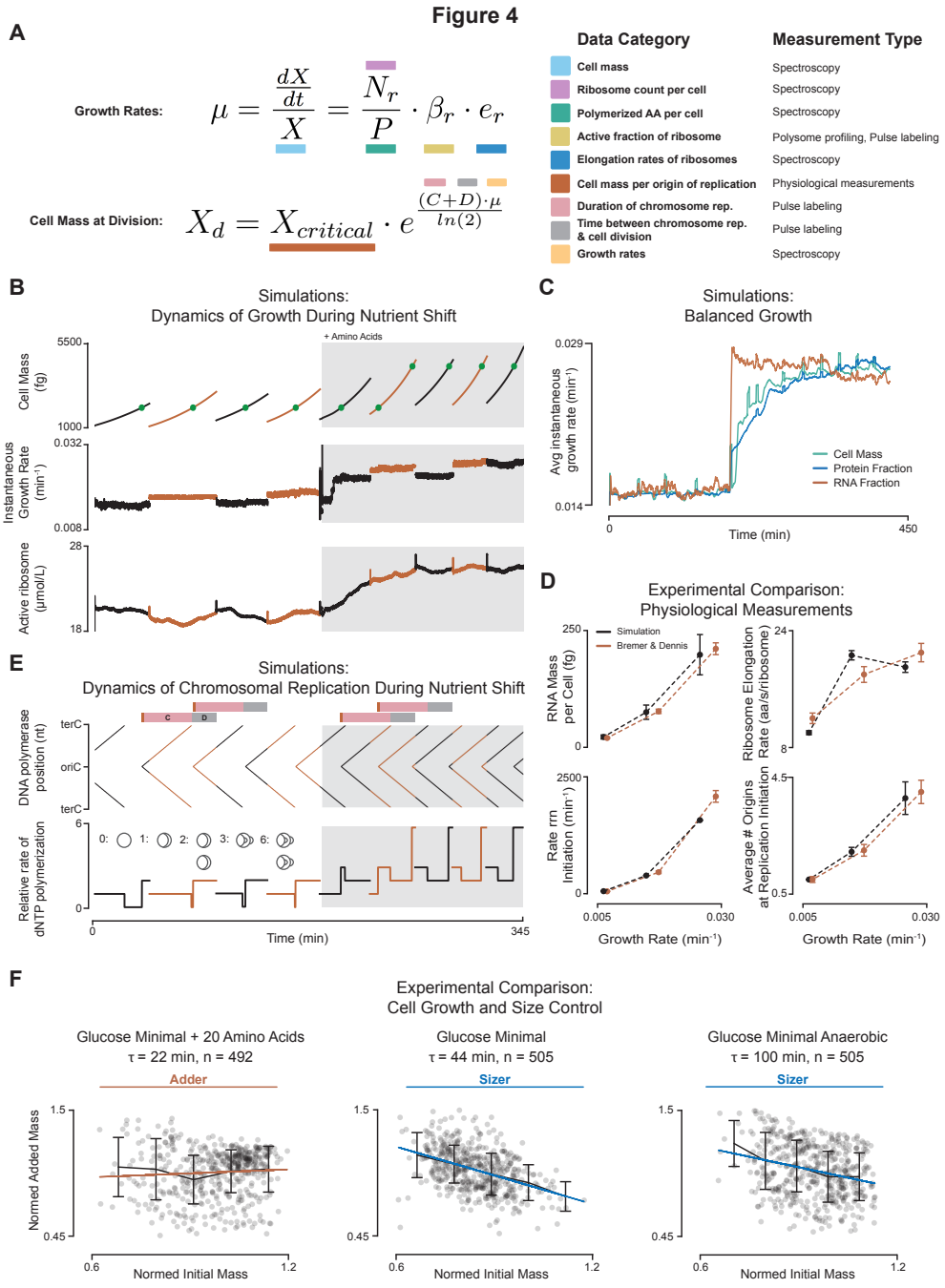


Figure 4
A
Growth Rates:
$\mu = \frac{\frac{dX}{dt}}{X} = \frac{N_r}{P} \cdot \beta_r \cdot e_r$
Cell Mass at Division:
$X_d = X_{critical} \cdot e^{\frac{(C+D)\cdot\mu}{ln(2)}}$
Data Category
Measurement Type
Cell mass
Spectroscopy
Ribosome count per cell
Spectroscopy
Polymerized AA per cell
Spectroscopy
Active fraction of ribosome
Polysome profiling, Pulse labeling
Elongation rates of ribosomes
Spectroscopy
Cell mass per origin of replication
Physiological measurements
Duration of chromosome rep.
Pulse labeling
Time between chromosome rep. & cell division
Pulse labeling
Growth rates
Spectroscopy
B
Simulations:
Dynamics of Growth During Nutrient Shift
+ Amino Acids
Cell Mass (fg)
5500
1000
Instantaneous Growth Rate (min⁻¹)
0.032
0.008
Active ribosome (μmol/L)
28
18
C
Simulations:
Balanced Growth
Avg instantaneous growth rate (min⁻¹)
0.029
0.014
Cell Mass
Protein Fraction
RNA Fraction
0
450
Time (min)
D
Experimental Comparison:
Physiological Measurements
Simulation
Bremer & Dennis
RNA Mass per Cell (fg)
250
0
Ribosome Elongation Rate (aa/s/ribosome)
24
8
Rate rrn Initiation (min⁻¹)
2500
0
Average # Origins at Replication Initiation
4.5
0.5
0.005
0.030
Growth Rate (min⁻¹)
E
Simulations:
Dynamics of Chromosomal Replication During Nutrient Shift
DNA polymerase position (nt)
terC
oriC
terC
C
D
Relative rate of dNTP polymerization
6
0
0: 1: 2: 3: 6:
0
345
Time (min)
F
Experimental Comparison:
Cell Growth and Size Control
Glucose Minimal + 20 Amino Acids
τ = 22 min, n = 492
Adder
Glucose Minimal
τ = 44 min, n = 505
Sizer
Glucose Minimal Anaerobic
τ = 100 min, n = 505
Sizer
Normed Added Mass
1.5
0.45
0.6
1.2
Normed Initial Mass

fixed ratio of cell mass per origin of replication [60], which produces multiple overlapping rounds of replication at faster growth rates [50]. This approach correctly reproduced expected replication fork dynamics, including periods where no DNA polymerization is occurring at slow growth rates, and multiple simultaneous rounds of replication at fast growth rates (Figure 5.4E). Furthermore, the rate of DNA replication initiation correctly scaled to match the growth rate both before and after the shift in environment, and was also confirmed by population measurements of the number of origins of replication per cell upon DNA replication initiation (Figure 5.4D).

We then coupled each chromosome replication initiation event to a cell division event which occurred after a fixed period of time for DNA replication and cytokinesis (occurring zero, one or two generations in the future depending on the growth rate), as inspired by recent work31 (Figure 5.4E). The interaction of (1) variation in the growth rate between individual cell simulations and (2) coupling cell division to a cell cycle event at a fixed mass led to two possible effects with regard to mass added per cell cycle, depending on the growth rate. In fast growing cells, the cell mass added over the life cycle was uncorrelated with the initial cell mass - a phenomenon referred to as adder behavior [177, 93], whereas for slower growing cells the added and initial cell masses were correlated (sizer behavior) [191] (Figure 5.4F). Integrating the responses of growth rate and chromosomal replication to media conditions therefore enabled us to simulate cellular behavior over many generations with stable cell size distributions - in dramatic contrast to our original *M. genitalium* model, which could only grow stably for one cell cycle11.

Finally, having access to all of these data simultaneously in a single integrated model enabled us to ask the question: can these data tell us something in toto that wouldnt have been obvious from

---

Figure 5.4 *(preceding page)*: **Cell growth and size integration and validation**
Model-driven analysis and cross-validation of the data associated with growth and DNA replication. (A) Representative equations and data sources describing key growth rate parameters and division requirements. These equations form the basis of the growth laws implemented in our simulations. (B) Simulated dynamics of cell mass, instantaneous growth rate, and active ribosome concentration over several generations (depicted using alternating colors of black and orange), both before and after supplementing a minimal medium with amino acids. The green circles indicate DNA replication initiation events. (C) Simulated average instantaneous growth rate of the overall cell mass as well as the RNA and protein fractions alone. After an environmental shift, the RNA mass fraction is the first to reach the new growth rate, but the other fractions also reach the new rate over time. (D) Comparison plots of cellular properties calculated from the simulations, together with their counterparts reported in the literature [32]. (E) DNA polymerase position and relative rate of dNTP polymerization, across generations and during a shift from minimal media to minimal media supplemented with amino acids. The pink and gray bars on top indicate the so-called C and D periods of replication (named for the processes of chromosome replication and cytokinesis, respectively) [191]; the inset drawings highlight the fact that the relative rate of dNTP polymerization can be thought of as analogous to the number of replication forks, as shown. (F) Comparison plots of the normed initial and added mass indicates that the simulations reproduce adder behavior in one condition, and sizer behavior in others, as would be expected [191, 177, 154].

any individual experiment? In this regard, the most striking observations we made relate to the fact that, although many genes are transcribed multiple times as a typical cell grows and divides, a clear majority of the genes in *E. coli* are transcribed at a rate of less than one per cell cycle (Figure 5.5A,B). In the case of growth on minimal glucose medium, 35.7% of the genes are transcribed at least once per cell cycle, and 4.6% are essentially never expressed in this environment. The remaining 59.7% are transcribed with a frequency that is distributed almost uniformly between 0% and 100% (Figure 5.5B). This leads to dramatic consequences for the dynamics of protein expression, where a protein might be expressed only once in ten generations - but with a much higher fold change - and otherwise simply be diluted in concentration as the cell grows and divides (Figure 5.5A,C,D).

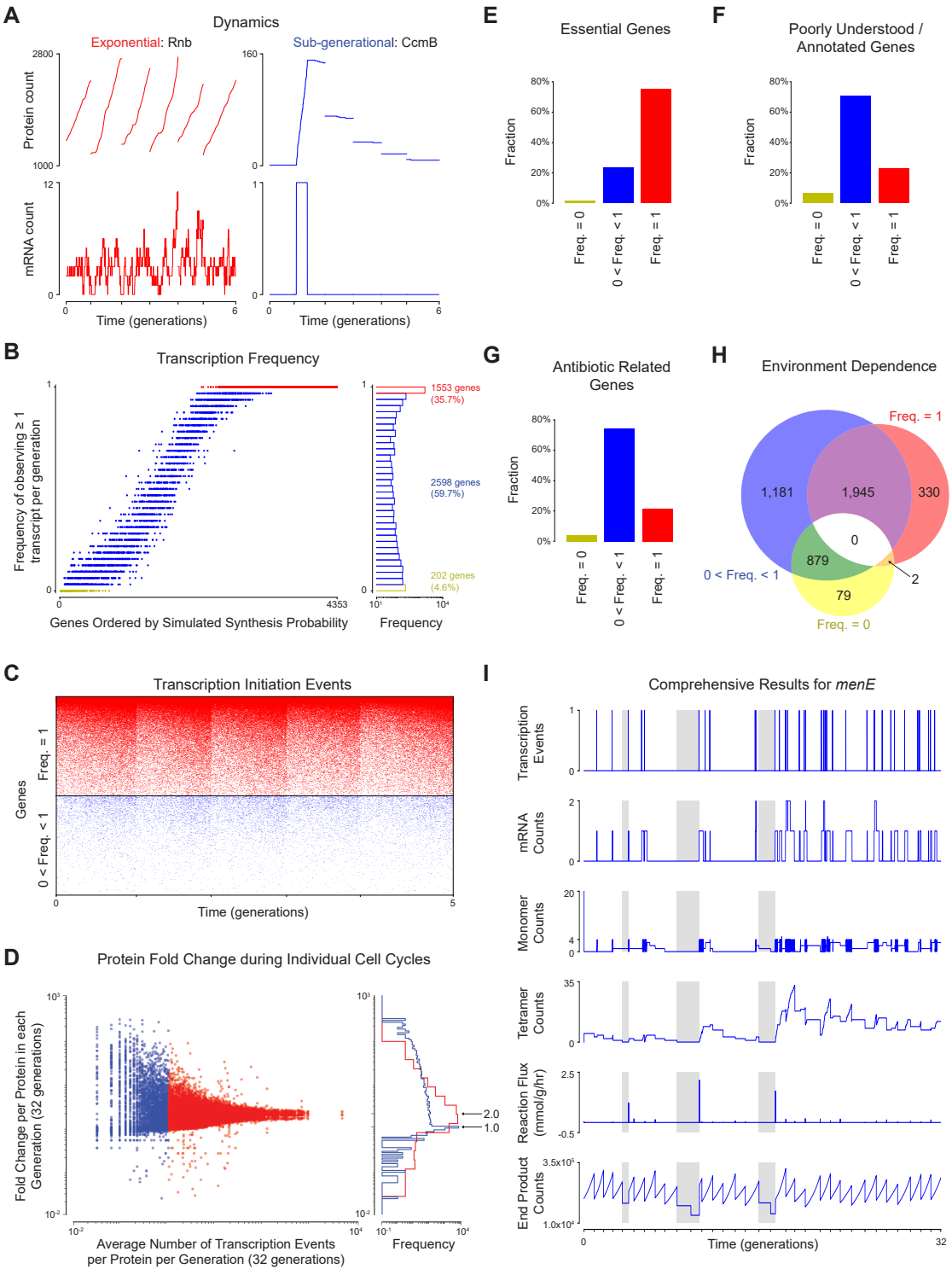
Investigating this phenomenon further, we found that the bulk of essential genes are found in the group with 100% cell cycle expression (Figure 5.5E), and most genes with unknown functionality or poor annotation are in the group of genes that were expressed less than once per cell cycle (Figure 5.5F). More surprising was the finding that certain functional categories are also predominantly in certain groups. For example, 74.3% of the genes associated with antibiotic resistance or persistence are found to be expressed less than once per cycle (Figure 5.5G), a compelling finding given that persistent bacteria appear to be phenotypically different from their siblings in culture [15, 10].

Considering these low-expression genes with respect to the rest of our compiled mRNA expression data, we were able to divide them into two categories. First, we found that one group of 1,945 genes was inducible - they were expressed at low expression levels in one or more environmental conditions, but at 100% cell cycle expression levels in other conditions. The other group of genes, 1,181 in total, were found to be constitutively expressed less than once per cycle, with a low probability of synthesis under all of the environmental conditions we considered (Figure 5.5H).

This finding implies that many proteins required for cell survival and growth may be absent from the cell for periods of time. We therefore wondered, how does the cell compensate for the temporary loss of an important enzyme due to very low expression rates? For example, O-succinylbenzoate-CoA ligase is an enzyme involved in menaquinone (sometimes known as Vitamin K2, important for respiration and electron transfer in bacteria) biosynthesis. This ligase is encoded by the *menE* gene, which is transcribed with a frequency of 1.1 times per cell cycle (Figure 5.5I), producing an average of 5.2 proteins per generation. The ligase is only active as a tetramer, and thus the average count of active complex in the cell is 7.2. Our simulations exhibit periods of time in which no tetramer exists (gray regions). During these MenE4 starvation phases, when the tetramer is completely absent, the internal concentration of menaquinone is reduced over time; however, following a new round of MenE4 expression, the menaquinone is rapidly re-synthesized. Sufficient internal metabolite pools and rapid enzyme kinetics is therefore one mechanism that can make cell growth robust to the loss of a key enzyme for periods of time.

In conclusion, the integration of experimental data using a large-scale theoretical framework allowed us to compile a massive, self-consistent dataset. Our findings cross-validate hundreds of

Figure 5


A
Dynamics
Exponential: Rnb
Sub-generational: CcmB
Protein count
mRNA count
Time (generations)
B
Transcription Frequency
Frequency of observing ≥ 1 transcript per generation
Genes Ordered by Simulated Synthesis Probability
1553 genes (35.7%)
2598 genes (59.7%)
202 genes (4.6%)
Frequency
C
Transcription Initiation Events
Genes
Freq. = 1
0 < Freq. < 1
Time (generations)
D
Protein Fold Change during Individual Cell Cycles
Fold Change per Protein in each Generation (32 generations)
Average Number of Transcription Events per Protein per Generation (32 generations)
2.0
1.0
Frequency
E
Essential Genes
Fraction
Freq. = 0
0 < Freq. < 1
Freq. = 1
F
Poorly Understood / Annotated Genes
G
Antibiotic Related Genes
H
Environment Dependence
Freq. = 1
1,181
1,945
330
0
879
0 < Freq. < 1
2
79
Freq. = 0
I
Comprehensive Results for menE
Transcription Events
mRNA Counts
Monomer Counts
Tetramer Counts
Reaction Flux (mmol/gDW/hr)
End Product Counts
Time (generations)

thousands of experimental measurements made largely independently across the world and over many decades. These findings stand in stark contrast to the reports on reproducibility cited above, and we believe that with respect to *E. coli* at least, we can have confidence in the scientific communitys output.

Importantly, this work depended on the use of theory to validate the experimental data, and not vice versa. Such an application of theory has been rare in the biological sciences - possibly because the large-scale integrated modeling approach we use is somewhat new - but this work demonstrates its potential, with regard to understanding, interpreting and cross-validating large data sets, and even suggesting lines of fruitful experimental inquiry for the future.

Returning to Cricks original monograph, does the totality of the data presented here, overlaid with a theoretical framework, constitute a complete solution? Crick gave a more detailed definition - By complete one means complete in the intellectual sense, implying that nothing appears to remain which further experiment could not easily explain using well-established facts and ideas [55]. By that metric, our work clearly falls short - not only because of the limited number of genes, molecules and environments that are accounted for, but also because the correlations and comparisons we show here, while strong, could yet be improved and expanded by directed future experimentation.

Though not complete, the data and model may still represent a solution of a certain kind. In

---

Figure 5.5 *(preceding page)*: **Sub-generational expression** A large fraction of *E. coli* genes are expressed less than once per cell cycle, which has physiological consequences. (A) Simulations of mRNA and protein expression over multiple generations for genes that are expressed at high (left, in red) and low levels (right, in blue; note that colors are conserved to preserve meaning throughout the figure) transcriptional frequencies. Counts are shown for a representative six-generation long window, with an arbitrarily chosen zeroth starting generation. (B) Frequency of observing at least one gene transcript per generation over a 32-generation simulation. Histograms show that 1,553 genes are transcribed at least once per cell cycle (red), 202 genes are essentially never expressed in this environment (yellow), and the remaining 2,598 genes are transcribed with a frequency between zero and one (blue). (C) Simulated transcription initiation events for each gene during the first five generations. (D) A plot of the fold change in protein count for each individual protein species over every cell cycle. This shows that proteins that are always expressed in every cell cycle usually double in number (histogram peak at 2), whereas proteins which are expressed less than once per cell cycle often remain at the same protein count (histogram peak at 1), but sometimes experience a very large fold change. (E) Expression frequency analysis of essential genes, (F) poorly understood or annotated genes, and (G) genes related to antibiotic resistance and persistence. (H) Venn diagram indicating genes that are constitutively expressed in every cell cycle regardless of environmental conditions (red), genes that are always expressed less than once per cell cycle (blue), or not expressed (yellow), as well as genes that can be environmentally induced from one group to another (purple, orange, green). (I) A set of plots that indicate the transcription, translation, complexation and metabolic activity of the MenE4 tetramer, which catalyzes a reaction responsible for producing menaquinone and demethylmenaquinone (represented in the final plot as a sum total). Each new generation is indicated with a tick mark along the x-axis; the gray areas highlight periods of time in which MenE4 is not present in the cell.

advanced mathematics, theorists recognize a distinction between particular and general solution types, and know that often the identification of particular solutions can lead to characterization of the general. In that context, we argue that this work represents a particular solution of *E. coli* - a solution that is only relevant to specific conditions, does not yet represent all known cellular processes, and incorporates a dataset that though extensive, is still a sample of all of the possible data which could be included - but which also provides a foundation upon which other functionalities, environments and datasets can readily be built. We therefore hope and anticipate that this work may provide an alternative to a central lab, which can encourage, motivate and support both theorists and experimental scientists - in the light of four decades of accomplishment - to reassess and eventually surmount Cricks grand challenge.

Chapter 6

Conclusion

The completion of the *E. coli* model has produced many exciting predictions and advanced the field of whole-cell modeling. Although we have made significant progress, by no means have we accomplished the goal set out in the introduction of a predictive and comprehensive cell model that demonstrates a theory which completely encapsulates cell biology. In this chapter I outline future lines of inquiry that I think are of particular interest to the field and worth pursuing, as well as the major challenges I see for advancing towards more complete and complicated whole-cell models.

6.1 Future experiments

Transcription unit structure

Despite multiple publications attempting to describe it with varying levels of success [44] and extensive databases of transcriptional elements and annotations [104, 1] the transcription unit structure of *E. coli* is not well known or characterized. Precisely defined transcription start and stop sites do not exist, which is highlighted by the fact that although there are 3,845 promoters listed in Ecocyc, there are only 283 annotated terminator sites (at the time of writing) [104]. In order to accommodate this lack of detail we chose to artificially express each gene in the *E. coli* model as its own transcription unit with its own artificial promoter and terminator. Undoubtedly this misses significant effects of the the co-transcription of genes, in particular for protein complexes [119], or the linking of RNA polymerase submit to ribosomal protein expression [65]. Using long-read sequencing technologies it should be possible to map the transcription unit structure of *E. coli* with sufficient level of detail to produce a whole-cell model.

Protein half lives

The half lives of only a handful of proteins in *E. coli* have been individually studied [126], in contrast to mRNA where >1,200 have been measured across multiple publications [162, 43]. The half life of a protein can significantly alter its transcriptional expression and rate of translation in the model to produce a given a desired level of protein abundance. In order to fill this gap we used the N-end rule [180], where whatever amino acid is at the N-terminus of the polypeptide terminates its degradation rate. Only two rates are specified: fast (∼2 min) or slow (∼10 hours). This lack of resolution has produced spurious model predictions that we identified and then measured ourselves using Western blots. These new measured half lives for 6 proteins brought their expression levels in the simulation significantly closer to expected values. This effort could be expanded to include a larger fraction of the proteome, further improving model predictions of protein expression.

Strain specific macromolecular composition

Although the *E. coli* model used data only measured in *E. coli*, we were not able to parameterize the model using a single strain. Different strains of *E. coli* can vary in growth rate, macromolecular composition, and gene expression. We relied heavily on data collected by Bremer & Dennis, which was conducted entire with the *E. coli* B rather than the currently popular K-12 MG1655 and BW25113 strains [32]. This resulted in a model that was an amalgam of both strains. The the growth rate and macromolecular composition was taken from *E. coli* B with almost all other data taken from K-12. While this represented a significant advance over the *M. genitalium* model, which only had 55% of its parameters actually taken from any *Mycoplasma* strain [101], ideally all data used in an *E. coli* reconstruction would be measured in a single strain. Measuring the growth rate, DNA, RNA, and protein content of an *E. coli* K-12 strain across multiple growth conditions would go a long way towards remedying this gap in knowledge.

Single cell instantaneous growth rate

Cell-to-cell variation in growth rate has been observed within an isogenetic population of *E. coli* cells in a constant environment. The factor which produces this variation has not been identified either through modeling or experimental interrogation. It has been hypothesized that the source of cell-to-cell variation that drives growth rate variation is unequal partitioning of an important cellular component at cell division, as opposed to noise in the control circuit [191]. The high throughput measurement of single cell growth rate either during multiple points in the cell cycle or continuously could either support or invalidate this hypothesis, and provide single cell data to guide future modeling efforts.

Currently single cell growth rate data has been collected using a microfluidic mother machine devices that enables steady state, exponential growth of cells under constant conditions. Cells are

imaged while growing and their area calculated through via computational image segmentation at regular intervals. Due to noise introduced by current methods of image segmentation the data is then fit with an exponential curve with its exponential constant being the averaged growth rate of a single cell. Reducing image segmentation noise such that the signal of cellular growth rate could be discerned at multiple points in the cell cycle should now be possible using the DeepCell image segmentation technique published by David Van Valen from our research group [184]. DeepCell uses a trained neural network to very accurately segment image pixels and measure the area of single *E. coli* cells. This should enable the determination of if a cells growth rate is truly constant over an entire cell cycle, or if it varies in some sort of predictable way with the cell cycle, cell division, or other measurable factors like chromosome replication dynamics [191], and ribosome clustering [41].

6.2 Future modeling

Metabolic modeling in whole-cell models

The *E. coli* metabolic model uses a derivative of flux balance analysis (FBA) with a multi-objective minimization for homeostatic metabolite composition and reaction kinetics. This alleviates a number of the significant issues encountered in the *M. genitalium* model with metabolite pooling, and it allowed us to use metabolite concentrations in sub-models, however it prevents the modeling of a number of important physiological behaviors of a cell. The most problematic of these is that intracellular metabolite concentrations do not vary dynamically as expected in response to perturbations. Instead of metabolite concentrations being the result of dynamic a balance of supply and demand, they are maintained at a fixed value by the metabolic objective function.

Feedback loops that involve sensing dynamic changes to intracellular concentrations cannot be implemented in a straightforward and biologically realistic way. For example, when *E. coli* shifts from growing in a rich medium to a minimal medium environment, intracellular amino acid pools are significantly depleted [65]. This depletion triggers a number of changes in the cell including inducible transcriptional regulation to up-regulate biosynthetic enzymes, that adapt the cell to the new environment [65, 32], and allow amino acid pools to recover. This sequence of events is not possible in the current *E. coli* model implementation.

One possible solution to this is removing central metabolism and amino acid production from the FBA metabolic model and instead implement these pathways as a set of ordinary differential equations much like Integrated FBA (iFBA) [53]. Sufficient kinetic parameters exist to constrain central carbon metabolism and many of the amino acid pathways, and even if they are not completely accurate, the model would qualitative produce the correct behavior in response to perturbations.

Global transcription control and effects of free RNA polymerase concentration

Growth rate dependent changes in the concentration of free cytoplasmic RNA polymerase ($[R_f]$) affect the transcription of all bacterial genes. The rate of transcript initiation at a given promoter (V_{init}) can be assumed to have a Michaelis-Menten type relationship with the concentration of free RNA polymerase.

$$V_{init} = \frac{V_{max}}{1 + \frac{K_m}{[R_f]}} \tag{6.1}$$

With V_{max} being the maximum rate of initiation at a promoter and K_m being the free RNA polymerase concentration at which half of the V_{max} is obtained. The parameters V_{max} and K_m are invariant for constitutive promoters and may be modified by transcription factors for regulated promoters. Despite their invariant parameters, the rate of initiation at constitutive promoters changes from environment to environment due to changes in growth rate. This is due to variation in the concentration of free RNA polymerase as a consequence of at least two effects:

1. Changes in the frequency of transcriptional pausing: At higher growth rates the lower concentration of (p)ppGpp reduces the rate of RNA polymerase pausing and decreases the residence time of an RNA polymerase on a gene. This increases the fraction of RNA polymerases which are free in the cytoplasm [30].

2. Changes in active promoter concentration: At the higher growth rates possible in richer media the number of active promoters is lower due to transcriptional repression of biosynthetic operons [65, 30]. This is in spite of the increase in per-gene copy number due to overlapping rounds of chromosome replication [32].

Together these effects can mechanistically account for up to 50% of the 3.2-fold increase in growth rate when shifting between succinate minimal and glucose-amino acid medium (0.67 to 2.14 doublings per hour). This is due to the effect that increased free RNA polymerase concentration has on the constitutive P2 promoter of the ribosomal operon in *E. coli* [65].

Furthermore, potentially each constitutive promoter has a different V_{max} and K_m parameter value, causing the proportional rate of transcription from each to vary with changing free RNA polymerase concentration without any transcription factor regulation. This has been shown for two constitutive promoters with known V_{max} and K_m values in Figure 6.1.

These mechanistic effects are all lost in the current *E. coli* model implementation, and are instead approximated by programmatically changing the rate of transcript initiation of some promoters with environment. In the work presented in Chapter 5 the concentration of free RNA polymerase has no effect on the rate of transcript initiation. Essentially for all values of $[R_f]$ the rate of transcript

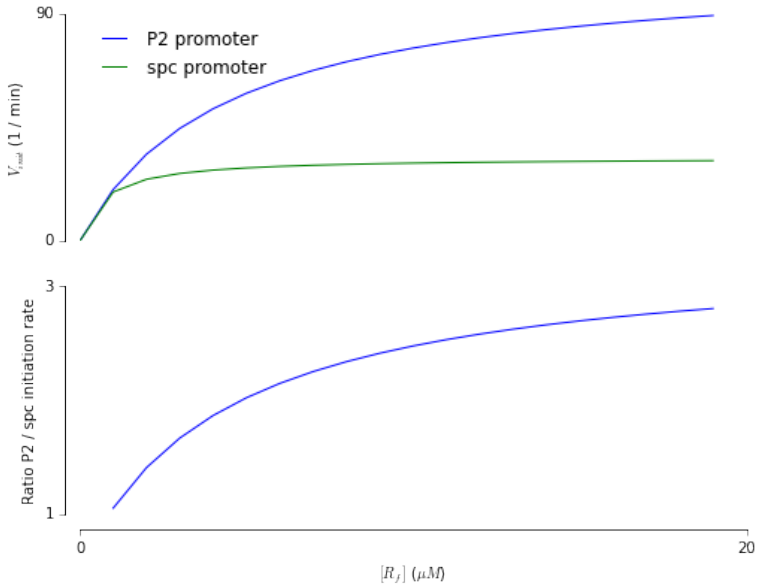


Figure 6.1: **Constitutive promoter rate of initiation varies with free RNA polymerase concentration** Rates of initiation for two constitutive promoters in *E. coli* B are shown. Due to their differing maximal rates of initiation and affinities for RNA polymerase, their ratio of initiation rates shifts with free RNA polymerase concentration. Kinetic values for P2 promoter ($V_{max} = 110$ 1/min, $K_m = 4.4\mu$M) and *spc* gene promoter ($V_{max} = 33$ 1/min, $K_m = 0.7\mu$M) taken from Bremer *et al.* [30]

initiation for a given promoter is constant unless the total counts of free RNA polymerase are very close to or exactly zero.

The main driving force behind this approximation was due to a lack of data and the increased complexity of reconciling and estimating parameters for this more detailed model of transcription initiation. Currently we estimate a lumped V_{max}/K_m parameter for each promoter and an orthogonal set of experimental data is necessary to separate these two parameters. With recent advances in high throughput DNA binding assays, like the Fordyce lab's work with Mechanically-Induced Trapping of Molecular Interactions (MITOMI), it may be possible to estimate a significant fraction of the K_m parameters for *E. coli*'s promoters. Even without this data, the rate of transcript initiation at a given promoter could be approximated at low $[R_f]$ as:

$$V_{init} \approx \frac{V_{max}}{K_m} \cdot [R_f] \tag{6.2}$$

Thus allowing the concentration of free RNA polymerase to have an effect on transcript initiation rates. This again increases the complexity of parameter estimation as $[R_f]$ will need to be estimated in each growth condition, but requires less new experimental data to be collected.

Chromosome concentration effects and cell division

When a mother cell divides into two daughter cells the septum forms on average half the distance between the two poles of the mother cell. There is significant cell-to-cell variation around this halfway point with a coefficient of variation of ∼20% [177]. Despite this unequal inheritance of cytoplasmic volume and its constituents, each cell *must* inherit a single chromosome copy. This produces cell-to-cell variability in DNA and promoter concentration, nucleoid occluded space, and potentially growth rate between the two daughter cells. To my knowledge this effect has not been investigated either experimentally or with modeling, although the idea is not completely novel [85].

Continuing the discussion in the previous section, varying promoter concentrations can cause variation in free RNA polymerase concentration, and hence alter both the total rate and proportional ratios of transcription from bacterial promoters. Assuming all other things are constant, increasing the concentration of promoters in a cell will have two effects: (1) it will increase the total rate of transcription from all promoters, and (2) it will decrease the rate of transcription from any individual promoter due to decreasing the concentration of free RNA polymerase. Incorporating this effect into future models of *E. coli* could potentially demonstrate that this is a significant source of cell-to-cell variation and may be the the cause of the observed cell-to-cell variation in growth rate.

Spatial organization and diffusion

In the current *E. coli* model we lack any explicit model of cell shape and assume a constant density to calculate a cell's volume from its mass. While constant density is a justified assumption under most

conditions, by not modeling cell shape and spatially oriented physiological processes we potentially miss out on a large fraction of observable phenotypes, a significant source of cell-to-cell variability, and growth rate limiting effects. For example, the *min* system, cytokinesis, and septum formation are not considered in the current *E. coli* model. Furthermore, in the *E. coli* model we assume that ribosome associated translational machinery is not rate-limiting, which may in fact be a poor assumption considering the aminoacyl-tRNA-GTP-Ef-TU complex is large enough that there is evidence that its diffuse may limit the rate of translation [106].

Phage modeling

Given the interest our research group has in bacteriophage an obvious extension of the whole-cell modeling framework would be to re-implement the model that Dr. Elsa Birch and I created of bacteriophage T7 infecting *E. coli* (detailed in Chapter 2) using the current *E. coli* model. Bacteriophage T7 is a purely lytic phage whose infection cycle is typically ¡20 minutes, which would make it straightforward to simply add T7s genes and reactions to the *E. coli* framework without implementing any extra layers of regulation. It would be interesting to see if this replicated the results we found using a less detailed and mechanistic FBA model, and see if any further experimentally verifiable predictions are suggested.

6.3 Future tools and framework

Parameter estimation

I've outlined a number of highly integrated model improvements that if implemented would produce a more adaptive model and better fulfill our goal of incorporating as much data as possible across multiple environments and perturbations. These are all exciting whole-cell modeling opportunities for the next generation of researchers to pursue, and I encourage them to do so. However, I would caution that often it is easier to conceive of a detailed, mechanistic computational sub-model than it is to actually implement it within a whole-cell model. This is evidenced by the fact that even though as a whole the *E. coli* model is hugely interconnected and complicated, taken individually most of the sub-models are very simple sketches of mechanistic biology.

Why is this? In my experience the biggest challenge in whole-cell modeling is *not* sub-model construction but *integration*. Why is this? There is no guarantee that a number of independently conceived and implemented sub-models will recreate the desired systems level behavior of exponential, balanced growth. This can arise due to any number of effects, the most common of which are an inappropriate model implementation (i.e. wrong modeling assumptions), or a correct model implementation with incorrectly estimated parameters. Typically the former is easier to identify and fix than the latter, and therefore the majority of the challenge lies in parameter reconciliation

and estimation. Even for simple bacteria a whole-cell model will contain thousands of parameters, most of which are not well characterized or simply unknown.

In theory parameters could be estimated by iteratively executing the model framework and minimizing a cost function. However, simulations are still computationally expensive to run with the *E. coli* model taking on the order of minutes to execute. Our current solution for estimating parameters is to build a set of "reduced" models based on heuristic approximations of the more detailed full model behavior. For example, these reduced models are less stochastic and will produce an "average" cell by making strong assumptions like fixed RNA polymerase and ribosome elongation rates, ignoring feedback loops, assuming steady state behavior has no variance, and by approximating stochastic gene and protein expression with statistical models. These reduced model simplifications work well when (1) the more detailed sub-models they are approximating on average don't violate their assumptions, and (2) a human can still conceptualize the whole system well enough to encapsulate all of its interactions in a reduced model. This is something that Derek Macklin and I discovered the difficult way when attempting to increase the sub-model complexity of the *E. coli* model and largely failing. Hence, using our current simulation and algorithms, only relatively simple sub-models can be implemented because their combined behavior is incredibly complex and difficult to approximate.

In my view the best way forward towards building more complex and integrated whole-cell models is by improving simulation execution time sufficiently to allow for numerical methods to estimate parameters using the actual sub-models in the simulation framework. This removes the burden on a human understanding the complete set of interactions and interconnected consequences of modeling assumptions during the design and implementation stage of model construction. We could compute the numerical gradient of a cost function that optimizes for balanced, exponential growth, expected gene expression, and potentially other input data depending on our confidence in its source. Using this numerical gradient we could find a minima that most likely would not be a unique solution but at least an acceptable solution. Fast simulation execution time, on the order of seconds, would be required to see convergence to a particular solution on practical timescales.

Data visualization

Visualizing the complex, interconnected, and detailed output of whole-cell simulations is absolutely critical for exploring output and gaining insight. Although tools like WholeCellViz currently exist for demonstrating the scope and complexity of whole-cell models [113], the state of the art in data exploration involves writing custom pieces of computer code to examine specific and aggregate simulation behavior. While this does provide a general solution to data visualization, it is cumbersome, requires knowledge of computer programming languages, and does not facilitate rapid data exploration. I believe that furthering the field of whole-cell model data visualization requires innovation in two areas.

First, new visual motives are needed to represent cellular physiology in an intuitive and recognizable way. What is required is something analogous to a map. People are now adept at understanding how to explore and understand data overlaid on a map motif thanks to the ubiquitous use of tools like Google Maps. At lower levels of zoom details are abstracted or omitted and an overall lay of the land is presented. At higher levels of zoom details emerge and the full representation of what is known about the geography and other overlaid attributes is shown. I believe a similar representation of cell physiology is possible. Examples of hugely detailed representations of biological networks abound and already exist to fill the need for a detailed high zoom picture. Where innovation is needed is high level representations of mass and energy flows, potentially using something like a Sankey diagram, and a set of low detail, high insight representations of a cells state, potentially using something like principle component analysis (PCA).

Second, new software tools are required to render these visual motifs from whole-cell model data. These would not be static interfaces like the exiting WholeCellViz but interactive data analysis tools closer to something like Tableau that would enable researchers visually interact with data and effortlessly transition between high level motifs and raw detailed data. Tools like these would increase the speed of the model development cycle, enable more detailed insights into situation output, and open the field of whole-cell modeling to researchers and enthusiasts who are not as computationally literate.

Appendix A

Additional results and methods for metabolic limits on viral replication

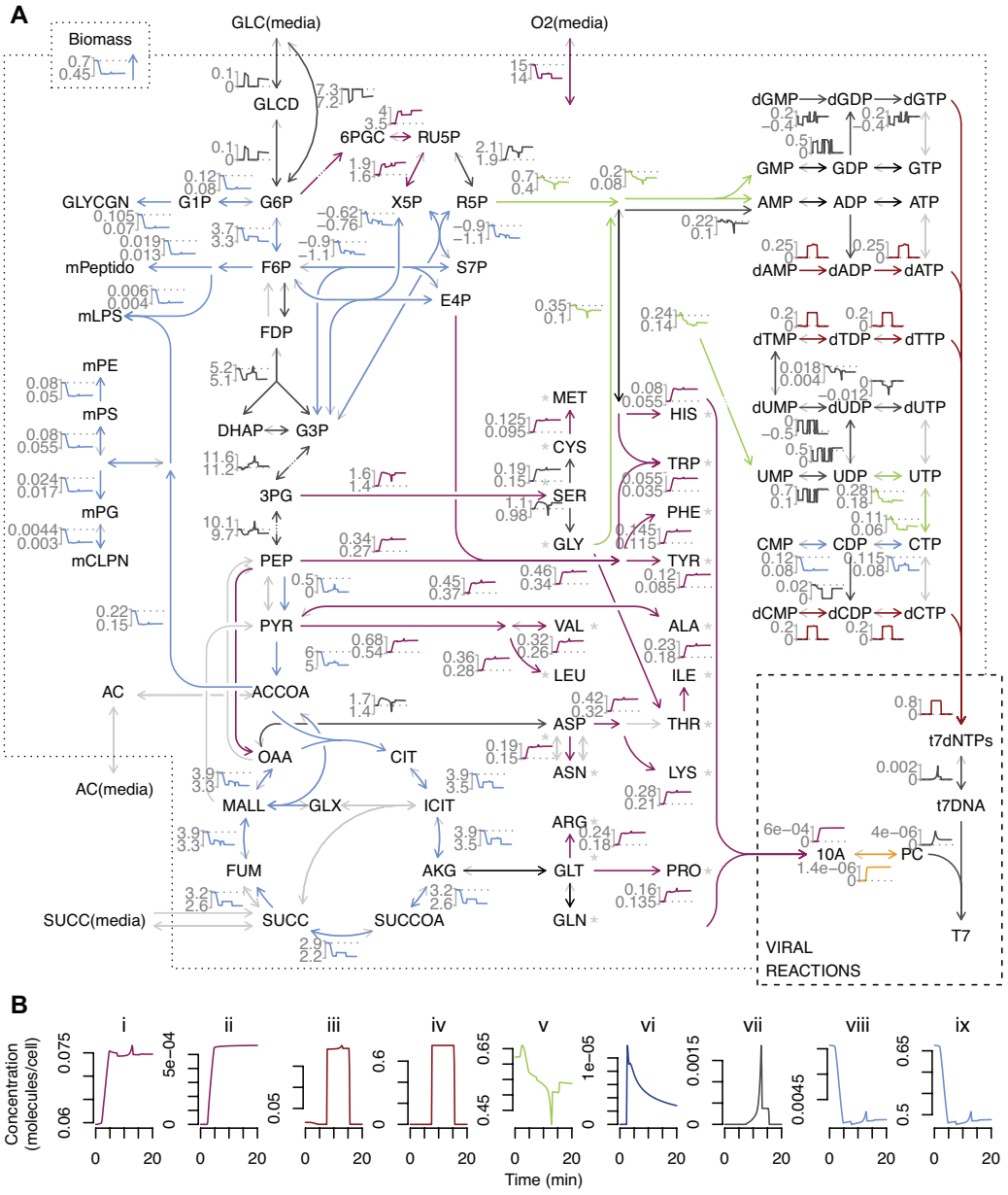
Here we include detailed information about the integrated model, as well as further experimental and computational data.

A.1 Further results-associated figures

To generate a more global evaluation of the host flux response to infection on varying media, we analyzed the aggregate similarity of the total flux distribution between pairs of media (Figure A.4). Overall, the flux distribution for infection during growth on acetate was very similar to the distribution during growth on succinate (green), largely because both carbon sources require gluconeogenesis for growth. The flux distributions for growth on glucose (blue) and tryptone (red) were more divergent because glucose is utilized by glycolysis and tryptone is unique among the four media for containing several amino acid carbon sources.

A.2 Bacterial strains, phages, media, and assays

The bacterial host strain used was *E. coli* K12 BW25113, and WT T7 phage (ATCC, BAA-1025-B2) was propagated according to established protocol [169]. Tryptone media contained 10 g/liter Tryptone (BD Bionutrients Bacto^{TM} Tryptone) and 5 g/liter NaCl consistent with previous T7 work [169, 67]. M9 minimal media contained 56.4 g/liter Difco M9 Minimal salts, with added 2 mM $MgSO_4$ and 0.1 mM CaCl; carbon sources glucose, succinate, and acetate were added at

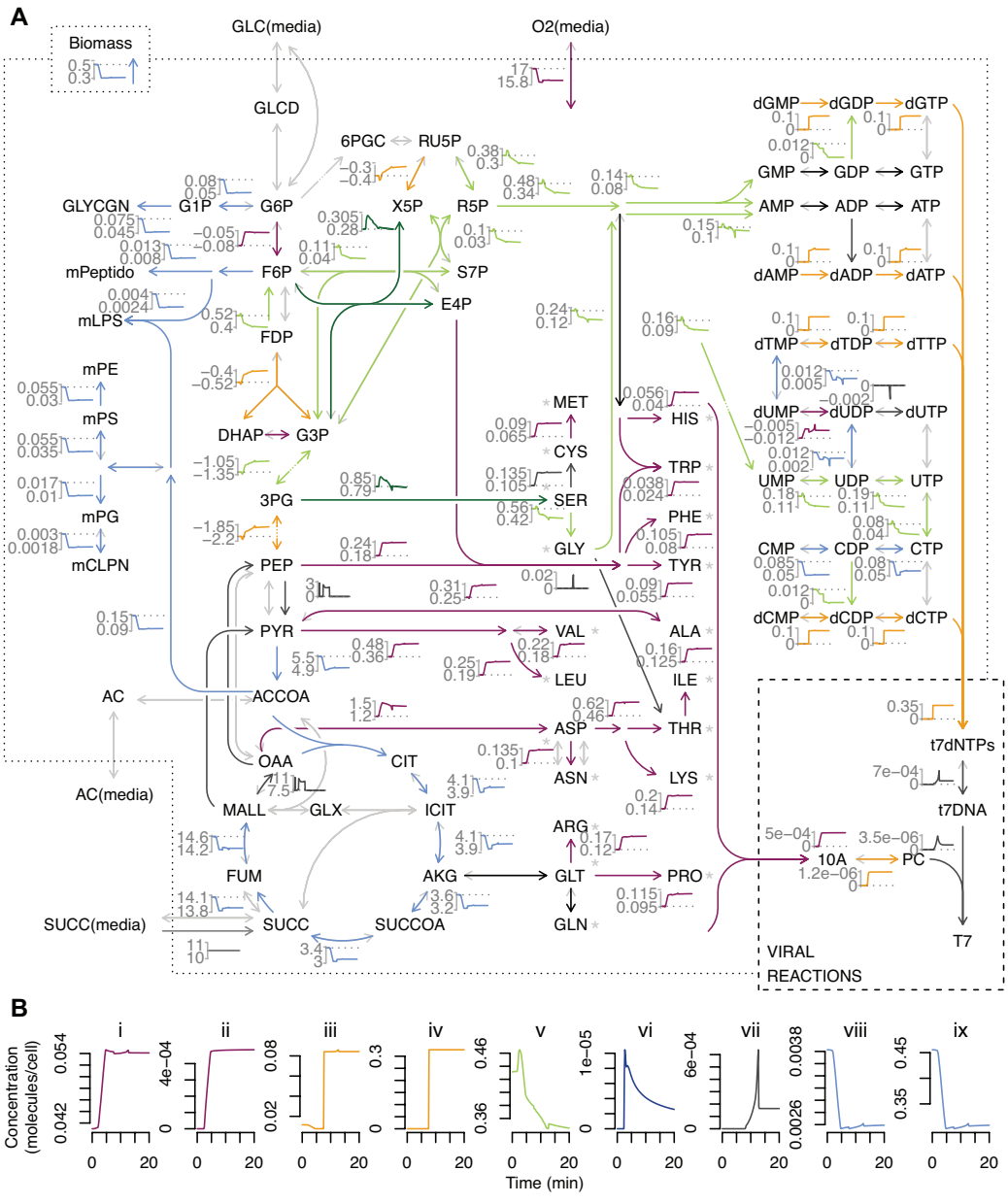

A
Biomass
GLC(media)
O2(media)
GLCD
6PGC
RU5P
GLYCGN
G1P
G6P
X5P
R5P
mPeptido
F6P
S7P
E4P
mLPS
FDP
mPE
mPS
mPG
mCLPN
DHAP
G3P
3PG
PEP
PYR
AC
ACCOA
AC(media)
OAA
CIT
MALL
GLX
ICIT
FUM
AKG
SUCC(media)
SUCC
SUCCOA
MET
CYS
SER
GLY
HIS
TRP
PHE
TYR
VAL
LEU
ALA
ILE
ASP
THR
ASN
LYS
ARG
GLT
PRO
GLN
dGMP
dGDP
dGTP
GMP
GDP
GTP
AMP
ADP
ATP
dAMP
dADP
dATP
dTMP
dTDP
dTTP
dUMP
dUDP
dUTP
UMP
UDP
UTP
CMP
CDP
CTP
dCMP
dCDP
dCTP
t7dNTPs
t7DNA
10A
PC
T7
VIRAL
REACTIONS
B
i
ii
iii
iv
v
vi
vii
viii
ix
Concentration (molecules/cell)
Time (min)

10 mM, 15 mM, and 30 mM to media preparations, respectively. All culture experiments were conducted at 37°C in a circulating water bath at a volume of 30 ml culture in a 250 ml flask that was magnetically stirred. Infections were at an initial MOI of 0.1 to assure hosts would only be infected once, and replicates were taken from separately infected flasks. Host population was measured as the optical density (OD) using a spectrophotometer at a wavelength of 595 nm. Phage dilution and storage was in SM phage buffer [197]. Measurements of phage titer were made by plating phage sample with 200 μl fresh bacterial culture at 1 OD from tryptone media in 3 ml tryptone broth with 0.7% agar atop tryptone broth 1% agar, and incubating the plate in an inverted position at 37°C for approximately 3 hr [169].

Phage time course assays

One-step phage growth experiments were conducted consistent with published protocols [197, 67, 110]. Prior to infection, bacterial hosts grew exponentially to a total density of 0.2 OD. Pilot experiments suggested that essentially all phage absorbed into the host cells within five minutes. Therefore, after 5 minutes of infection in the initial culture flask, a sample was diluted 1000-fold in warm shaken media into another flask of the same total culture volume (30 ml) to minimize adsorption of produced phage to new hosts. At 6 and 7 minutes (time points selected as just following complete phage absorption)infected hosts were counted. To count infected hosts 100 μl samples were transferred into ice-cold 900 ml aliquots of phage buffer, returned to ice, and plated less than 30 minutes later. At 6 and 7 minutes, as well as all other time points, 100 μl samples were transferred into room-temperature 900 ml aliquots phage buffer with 40 μl chloroform for host lysis. The chloroformed samples were incubated at room temperature for 30 minutes with periodic vortexing, then stored at 4°C until plating, usually within an hour. Phage from lysed samples at

Figure A.1 *(preceding page)*: **Infected host fluxes on glucose M9 minimal media.** (A) Flux dynamics are displayed for a subset of the metabolic network map. Arrows representing reactions and the subplots of flux through those reactions are colored according to clustering of flux dynamics. Positive flux values correspond to the reaction direction indicated by the colored arrowhead, negative flux direction is depicted with light grey barbs. Asterisks (*) represent an abbreviation of the arrow for uptake from media. Metabolite abbreviations are consistent with FBA model definition. For clustering, fluxes were treated as vectors with (1-correlation) as distance, and clustered using average hierarchical grouping with a cutoff height of 0.25. clusters with fewer than ten members appear in black, and clusters with constant dynamics are highlighted in grey. All nonzero fluxes in any media (tryptone, glucose, succinate, and acetate) were included in the flux clustering so that cluster designation and color coding is consistent across media and Figures A.1, A.2, and A.3. (B) Select flux dynamics expanded for clarity ordered to exemplify host flux changes driven by viral dynamics: (i) host amino acid synthesis, (ii) major viral capsid protein synthesis, (iii) host nucleotide phosphorylation, (iv) viral digestion of host genome to dNMPs, (v) purine biosynthesis, (vi) viral mRNA synthesis, (vii) viral genome synthesis, (viii) host cell envelope biosynthesis, (ix) host biomass accumulation.


A
Biomass
GLC(media)
O2(media)
GLCD
6PGC
RU5P
GLYCGN
G1P
G6P
X5P
R5P
mPeptido
F6P
S7P
E4P
mLPS
FDP
mPE
mPS
DHAP
G3P
3PG
mPG
mCLPN
PEP
PYR
AC
ACCOA
AC(media)
OAA
CIT
MALL
GLX
ICIT
FUM
AKG
GLT
SUCC(media)
SUCC
SUCCOA
GLN
MET
CYS
SER
GLY
HIS
TRP
PHE
TYR
VAL
ALA
LEU
ILE
ASP
THR
ASN
LYS
ARG
PRO
dGMP
dGDP
dGTP
GMP
GDP
GTP
AMP
ADP
ATP
dAMP
dADP
dATP
dTMP
dTDP
dTTP
dUMP
dUDP
dUTP
UMP
UDP
UTP
CMP
CDP
CTP
dCMP
dCDP
dCTP
t7dNTPs
t7DNA
10A
PC
T7
VIRAL
REACTIONS
B
i
ii
iii
iv
v
vi
vii
viii
ix
Concentration (molecules/cell)
Time (min)

later time points are reported normalized to the infected host count obtained by the difference of unlysed and lysed samples at 6 and 7 minutes.

A.3 Simulation parameters and component model updates

Media condition simulation

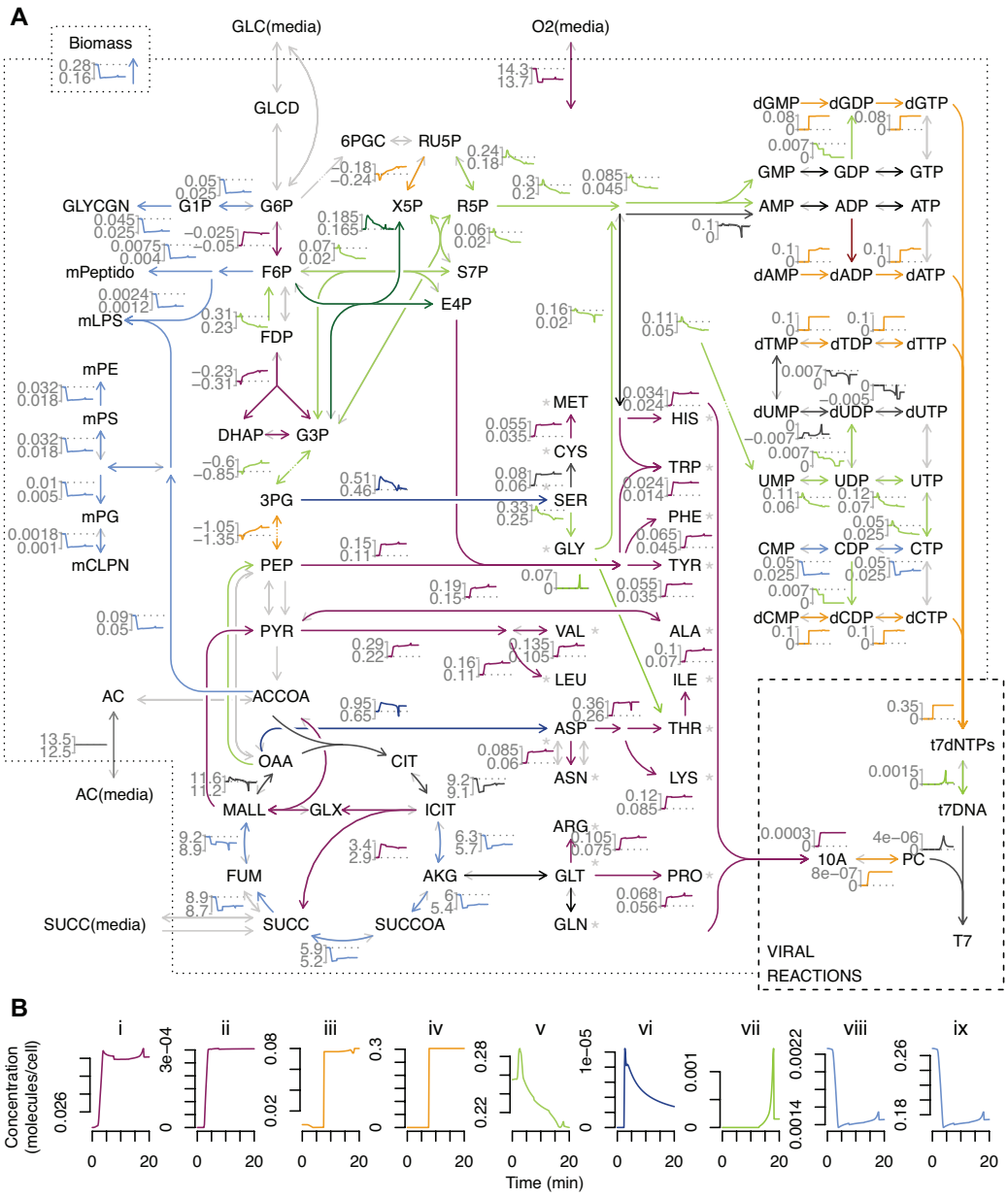
The media definitions used for simulations are given in Table A.1. The three minimal media are well-defined; the Tryptone media composition is based on the BD Bionutrients^{TM} Technical Manual. While most metabolite media components are considered exhaustible, some metabolites are present in excess, which is simulated by returning to a preset concentration at every FBA time step.

Certain modifications had to be made in the integrated model to account for different media conditions. For example, our experiments were performed at 37°C, but the original T7 ODEs were produced to model a 30°C environment. We therefore modified some of the parameters, as listed and explained in Table A.5. Equilibrium binding constants were assumed not to change with temperature. Parameters not listed in Table A.5 are consistent with previous implementations [197].

Furthermore, the original FBA regulatory rules identified in [51] had never been optimized for growth on rich media, and the model was therefore very limited in its ability to simulate reasonable growth rates. We therefore relaxed several of these rules to allow smooth growth on media containing only amino acid carbon sources in the combination present in our representation of tryptone. A list of relaxed regulatory rules is given in (Table A.2).

Some of the FBA uptake bounds were determined by fitting dynamic rFBA simulations to experimental measurements in the absence of virus. Experimental data for host growth in the different media are shown, together with fitted exponentials, in Figure A.5. The measured growth rates

Figure A.2 *(preceding page)*: **Infected host fluxes on succinate M9 minimal media.** (A) Flux dynamics are displayed for a subset of the metabolic network map. Arrows representing reactions and the subplots of flux through those reactions are colored according to clustering of flux dynamics. Positive flux values correspond to the reaction direction indicated by the colored arrowhead, negative flux direction is depicted with light grey barbs. Asterisks (*) represent an abbreviation of the arrow for uptake from media. Metabolite abbreviations are consistent with FBA model definition. For clustering, fluxes were treated as vectors with (1-correlation) as distance, and clustered using average hierarchical grouping with a cutoff height of 0.25. clusters with fewer than ten members appear in black, and clusters with constant dynamics are highlighted in grey. All nonzero fluxes in any media (tryptone, glucose, succinate, and acetate) were included in the flux clustering so that cluster designation and color coding is consistent across media and Figures A.1, A.2, and A.3. (B) Select flux dynamics expanded for clarity ordered to exemplify host flux changes driven by viral dynamics: (i) host amino acid synthesis, (ii) major viral capsid protein synthesis, (iii) host nucleotide phosphorylation, (iv) viral digestion of host genome to dNMPs, (v) purine biosynthesis, (vi) viral mRNA synthesis, (vii) viral genome synthesis, (viii) host cell envelope biosynthesis, (ix) host biomass accumulation.


A
Biomass
GLC(media)
O2(media)
GLCD
6PGC
RU5P
GLYCGN
G1P
G6P
X5P
R5P
mPeptido
F6P
S7P
E4P
mLPS
FDP
mPE
mPS
DHAP
G3P
3PG
mPG
mCLPN
PEP
PYR
AC
ACCOA
AC(media)
OAA
CIT
MALL
GLX
ICIT
FUM
AKG
GLT
SUCC(media)
SUCC
SUCCOA
GLN
MET
CYS
SER
GLY
HIS
TRP
PHE
TYR
VAL
ALA
LEU
ILE
ASP
THR
ASN
LYS
ARG
PRO
dGMP
dGDP
dGTP
GMP
GDP
GTP
AMP
ADP
ATP
dAMP
dADP
dATP
dTMP
dTDP
dTTP
dUMP
dUDP
dUTP
UMP
UDP
UTP
CMP
CDP
CTP
dCMP
dCDP
dCTP
t7dNTPs
t7DNA
10A
PC
T7
VIRAL
REACTIONS
B
i
ii
iii
iv
v
vi
vii
viii
ix
Concentration (molecules/cell)
Time (min)

were: tryptone $\mu = 1.5$ hour^{-1}, glucose $\mu = 0.66$ hour^{-1}, succinate $\mu = 0.45$ hour^{-1}, and acetate $\mu = 0.27$ hour^{-1}. Also shown are dynamic regulatory FBA simulations of host growth, the growth rates of which were fit via flux bound changes to the growth rate found by exponential fit. These fits were produced by estimation of key flux bounds, as follows:

for oxygen,

$$v_{min,O2} = -17\frac{mmol\ rxn}{gdcw \cdot hour},$$

for glucose

$$v_{min,GLC} = -7.3\frac{mmol\ rxn}{gdcw \cdot hour},$$

for succinate

$$v_{min,SUCC} = -10.5\frac{mmol\ rxn}{gdcw \cdot hour},$$

for acetate

$$v_{min,AC} = -13\frac{mmol\ rxn}{gdcw \cdot hour},$$

and for all amino acids

$$v_{min,AA} = -30\frac{mmol\ rxn}{gdcw \cdot hour}.$$

We include the integrated simulation output for glucose (Figure A.1), succinate (Figure A.2), and acetate (Figure A.3) M9 minimal media in flux maps analogous to that presented for tryptone in Figure 2.4. Color coding of flux arrows by cluster is consistent across figures, including tryptone.

Figure A.3 *(preceding page)*: **Infected host fluxes on acetate M9 minimal media.** (A) Flux dynamics are displayed for a subset of the metabolic network map. Arrows representing reactions and the subplots of flux through those reactions are colored according to clustering of flux dynamics. Positive flux values correspond to the reaction direction indicated by the colored arrowhead, negative flux direction is depicted with light grey barbs. Asterisks (*) represent an abbreviation of the arrow for uptake from media. Metabolite abbreviations are consistent with FBA model definition. For clustering, fluxes were treated as vectors with (1-correlation) as distance, and clustered using average hierarchical grouping with a cutoff height of 0.25. clusters with fewer than ten members appear in black, and clusters with constant dynamics are highlighted in grey. All nonzero fluxes in any media (tryptone, glucose, succinate, and acetate) were included in the flux clustering so that cluster designation and color coding is consistent across media and Figures A.1, A.2, and A.3. (B) Select flux dynamics expanded for clarity ordered to exemplify host flux changes driven by viral dynamics: (i) host amino acid synthesis, (ii) major viral capsid protein synthesis, (iii) host nucleotide phosphorylation, (iv) viral digestion of host genome to dNMPs, (v) purine biosynthesis, (vi) viral mRNA synthesis, (vii) viral genome synthesis, (viii) host cell envelope biosynthesis, (ix) host biomass accumulation.

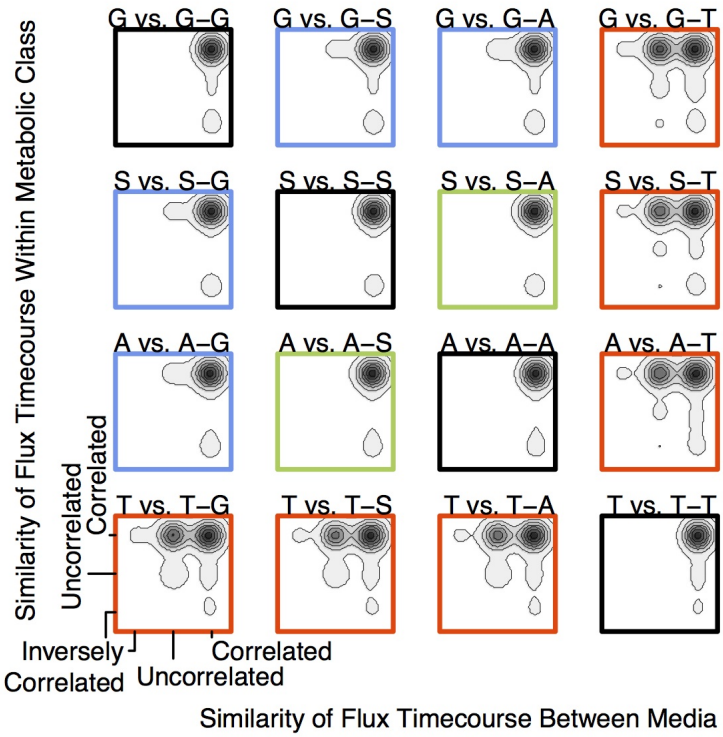


Figure A.4: **Similarity of flux dynamics within a media condition compared to similarity across different media conditions.** Correlation of each pair of fluxes within a metabolic class (Y), plotted against the correlation between a single flux between pair of media (Y-X). T, tryptone; G, glucose; S, succinate; A, acetate. Flux dynamics were treated as vectors to calculate the Pearson correlation. Only pairs that include one non-zero flux value were considered; for constant-constant pairs of flux dynamics a correlation of 1 was assigned, and for constant-varying pairs a correlation of 0 was assigned. Individual flux correlations were aggregated as density for plotting (darker as more dense), using kernel smoothing with a grid of 80 points and a bandwidth of 0.25. The density shading scale is not comparable between pairs. Similarity, as measured by positive correlation, of flux distributions indicated by high density on the right of the axis, and similarity within a single media indicated as density in upper regions. Centered density on either axis indicates dissimilarity, or lack of correlation. Panels highlighted with green are the highly similar flux dynamic distribution pair acetate succinate, most alike to the analysis of media with itself for reference, bounded in black. Panels highlighted with a blue border (glucose to succinate or acetate) are largely similar with some uncorrelated fluxes. Panels highlighter with red border (tryptone to any of the minimal media) are largely dissimilar. All media pairs display a large correlated fluxes because viral fluxes which are constrained by the T7 ODEs, which have dynamics that are similar except for scale across media. Inverse correlation within media and metabolic class potentially arises from the arbitrary directionality assigned to reversible reactions in the FBA definition.

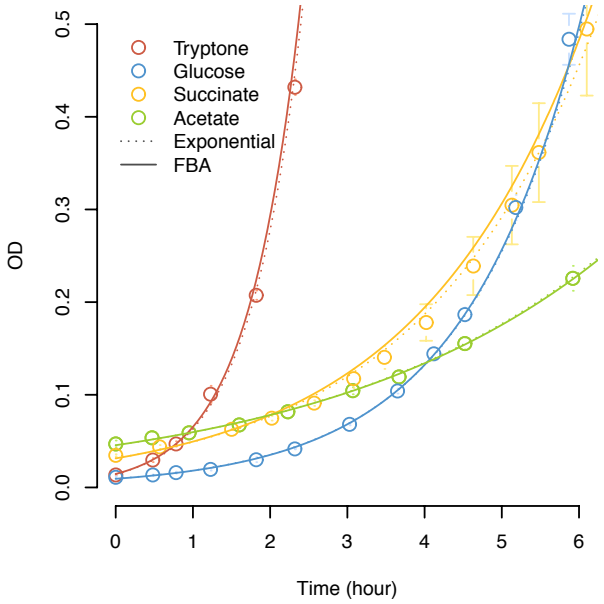


Figure A.5: **Time courses of uninfected** *E. coli* **growth on tryptone, succinate, glucose, and acetate.** Each time course was fit by a simple exponential (dotted) as well as using dynamic FBA (solid line), where initial conditions were determined by the first experimental measurement.

Updates to the T7 ODEs and FBA model

A number of additions were made to the T7 ODEs to reflect scientific advances and understanding subsequent to the model's original publications. Minor stoichiometry updates were made to reflect the most recent understanding of T7 virion composition, Table A.4. Stoichiometric multipliers are rounded up to integer values from value in citation. We also changed a subset of the promoter values for class II and III genes to adjust the ratio of gene product production to more closely correspond to the fraction in virion production and observed experimentally [84], Table A.4. We did not alter the early gene promotors or those for *E. coli* RNA polymerase in order to preserve critical feedback and inhibitory interactions. Table A.5 lists the parameter changes in detail.

The FBA model also had to be altered to account for viral-related reactions. The general reaction forms shown in Figure 2.1B were constructed using the assumptions in Table A.3. For each gene in the T7 genome, a viral mRNA and viral protein production reaction was added to FBA.

Degradation and recycling of the host genome by T7 are described as a single reaction accounting for the conversion of degraded host genome dNMPs through host pathways to directly produce viral dNTPs. A single reaction was used instead of multiple reactions (e.g., a separate degradation of host genome to dNMTs and uptake of dNTPs) in order to prevent the host from using these nucleotides, consistent with the pooling observed during T7 infection.

Another single reaction accounting for the synthesis of phage T7 genomes accounts for energy in

both strands of synthesis as well as for the energy and metabolites involved in proofreading.

A.4 Integrated simulation algorithm

The integrated simulation algorithm presented in Figure 2.2 is detailed in Figure A.6, and each step expanded below. Implementation was in MatLab and the code is available from the repository at http : //simtk.org/home/t7phagefba.

One note about the overall simulation algorithm is that the calculation of host supply and allocation across viral reactions is implemented without having first established whether the metabolite resources requested by the viral reactions exceed the possible host supply. This approach enforces our assumption that the virus is limited to the metabolic supply of the uninfected host state. If host supply calculation and resource allocation were completed only in the case that viral reaction metabolite use over-constrained the host problem, then the assumption of viral supply would be switched mid simulation – from viral maximization (which may not violate the bounds of the host but may exceed the metabolites available from a host biomass optimized flux distribution), to viral production being constrained by the host optimized metabolic state. This approach therefore maintains consistent application of our assumptions across time steps.

Specification (Start)

For each individual simulation run environmental perturbations are implemented in the media definition file, which specifies the media components and their concentrations in the initial Environment State.

Initialization (1)

Simulations are initialized using dynamic regulatory FBA simulation replenishing each media component. The steady state biomass production predicted by FBA is passed to the growth rate correlations given in the main and supplemental text and code of You *et al.* 2002 T7v2.5 [197].

Evaluate Viral Demand (2)

In this step, the T7 ODEs are evaluated as written with no limits applied using the viral state as initial condition concentrations and evaluated from t to $t + \Delta t_{integration}$, where $\Delta t_{integration}$ is the time step. The viral state is a vector of intercellular molecular concentrations separate from the FBA variables, which is stored at each time point of integration. The T7 ODE numerical solution output for the time period is in the format of concentration changes over time, and must be converted to a flux equivalent. Using the production-only term of the reaction rates associated with host fluxes $\frac{dP_j}{dt}$, the associated concentration variable changes over the time step can be directly averaged to

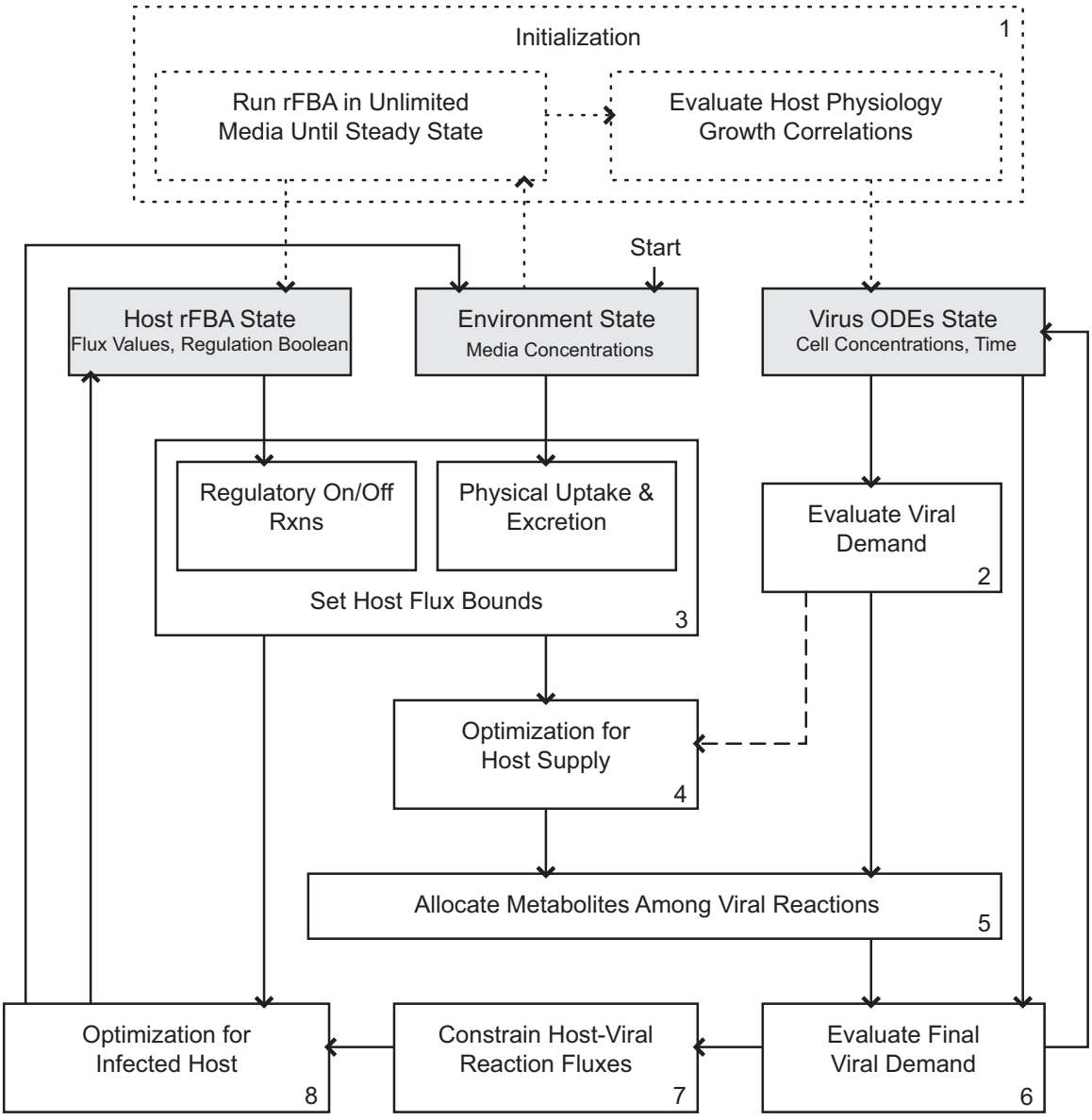


Figure A.6: **Detailed algorithm used for integrated simulation of *E. coli* FBA and T7 ODEs.** Solid lines are each integration time step, from beginning to end of iteration top to bottom. Shaded boxes are stored states describing model time course. Dotted lines completed for initialization. Dashed line connecting Viral Demand and Host Supply is needed or not needed depending on the optimization method being implemented in the latter. Full expansion of steps in Text S1.

flux unit values

$$\mathbf{v}_{request,ave} = \frac{([P_{j,t=(t+\Delta t_{int})}] - [P_{j,t=t}])}{\Delta t_{integration}}. \tag{A.1}$$

Several steps of the viral ODEs run between FBA steps, and we choose the maximum demand from all viral time steps. An average demand over all time steps was determined to be unsuitable because it unduly restrained viral production during the many time steps of rapid (and discontinuous) rate change based on phage genome translocation.

The maximum demand is evaluated as

$$\mathbf{v}_{request,max} = max_{(across\ t)} \left(\frac{\mathrm{d}[P_j]}{\mathrm{d}t} \right) \tag{A.2}$$

where the numerical approximation of rate

$$\frac{\mathrm{d}[P_j]}{\mathrm{d}t} \approx \frac{([P_{j,t=(t+\Delta t_{ODE})}] - [P_{j,t=t}])}{\Delta t_{ODE}} \tag{A.3}$$

is made for rates at each ODE evaluation time step.

This method to approximate viral demand has the potential disadvantage that the maximal flux value will lead to an over-request of some viral reactions, leading to an allocation of resources to this reaction that it will not use but is unavailable to other viral reactions. To mitigate this effect, we hold the integration time step below 15 sec.

Set Host Flux Bounds (3)

Next, we determine the bounds on all fluxes in the FBA model based on regulatory rules, environmental conditions, and viral demand. The Boolean regulatory relationships and substrate uptake bounds (except for those discussed earlier in the SI) were evaluated as in [51], based on time step $t_{integration}$ and biomass. These set the bounds on host reactions used in all subsequent linear optimization steps. Accumulated metabolites, which are allowed at the intersection of host and viral metabolism, based on the assumption that the developmental process of viral replication do violate steady state over single time-steps, and may be used in subsequent steps by the host (longer term accumulation is not observed). Mathematical accounting for the accumulation of metabolites as well as their consumption in later steps is directly analogous to exchange with the media, even in so far as concentrations are stored on a biomass independent basis. The model source code contains conversion factors between metabolite rates ($\mathbf{b}$) and accumulated concentrations, which are stored in the units equivalent to a media concentration. Viral demand bounds were obtained as defined in the previous section.

Optimization for Host Supply (4)

With the bounds determined, we used FBA to determine the host-virus flux distribution. FBA requires an objective function to calculate an optimal flux distribution. A common assumption in choosing an objective function is that the cell culture maximizes its growth rate, subject to the defined constraints [140]. For the very brief course of T7 phage infection (10-15 minutes), we assumed the state of host metabolism as set by the presence of enzymes remained relatively similar to the uninfected state. This enabled us to retain biomass maximization as the objective function. For comparison purposes, we also evaluated a strategy wherein the objective was to maximize viral reaction fluxes - to a maximum of $\mathbf{v}_{request}$. The simulation results for both objective functions were indistinguishable, due to the optimization of host after viral constraint (8).

We made two further modifications to the general FBA approach to enable modeling of T7 infection. First, T7 breaks down the host chromosome and the resulting free nucleotides are recycled and used for virion synthesis. Chromosome breakdown is therefore a major introduction of metabolites into the host metabolic network that are being modified, rather than synthesized from media sources. To account for chromosome breakdown, we added a reaction to enable host nucleotide recycling. In order to make the viral dNTPs available during the metabolite distribution step, we maximized recycling in a separate and subsequent optimization step. A separate recycling step is required because it is a draw on host energy resources would therefore not occur as a result of host biomass optimization. Furthermore, this approach approximates the concentration effects that are known to occur due to kinetics and T7 encoded gene product interactions with host metabolic enzymes, which lead to rapid host recharge of its own degraded genome nucleotides [144].

Our second modification to FBA involved the production of certain metabolites by the virus that could be used by the host. Most of the metabolite transfer in the simulation flows from host to virus, but some viral reactions include a return of metabolites to the host (e.g., ADP is produced when the virus uses ATP, or dNMPs result from proofreading of mistakes in T7 genome synthesis). These fluxes of metabolic resources into the host can then be used in other reactions after viral bounds are strictly set (8), but do not contribute during the calculation of the host supply upper bound.

Allocate Metabolites Among Viral Reactions (5)

Because the T7 ODE kinetic rates do not depend on small molecule concentrations, we bound phage macromolecule production rates themselves to host production capacity. The method to determine rate limits relies first on the 'initial demand' calculated as described above. All reaction rates consuming a given metabolite are then scaled to the availability of that metabolite. Implementation of this strategy takes advantage of the matrix formulation of the FBA problem, and splitting the combined Host-Viral stoichiometric matrix as shown in Figure 2.2a into $\mathbf{S}_H$ the host stoichiometry, $\mathbf{S}_{HV}$ host-viral stoichiometry of host metabolite consumption by viral reactions Eq. A.4.

$$\begin{bmatrix} \mathbf{S}_H & \mathbf{S}_{HV} \\ 0 & \mathbf{S}_V \end{bmatrix} \begin{bmatrix} \mathbf{v}_{host} \\ \mathbf{v}_{viral} \end{bmatrix} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} \tag{A.4}$$

$$\begin{bmatrix} \mathbf{S}_H & \mathbf{S}_{HV} \end{bmatrix} \begin{bmatrix} \mathbf{v}_{host} \\ \mathbf{v}_{viral} \end{bmatrix} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} \tag{A.5}$$

It should be noted that a matrix representation of the combined problem might include $\mathbf{S}_V$ which is viral metabolite stoichiometry, presumably predominantly macromolecules. It is not strictly necessary provided mass conservation is enforced by the ODEs, and so $\mathbf{S}_V$ is henceforth neglected Eq. A.5. By splitting the matrix and flux vector components, we represent the product of the combined metabolic system instead as a sum of metabolite rate vectors Eq. A.6 which are the net metabolites produced by the host, $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{host}$, and net small molecule metabolite consumption by viral reactions, $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{viral}$.

$$\begin{bmatrix} \mathbf{S}_H & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_{host} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & \mathbf{S}_{HV} \end{bmatrix} \begin{bmatrix} 0 \\ \mathbf{v}_{viral} \end{bmatrix} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$$

$$\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{host} + \left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{viral} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} \tag{A.6}$$

For a host flux distribution, $\mathbf{v}_{host}$, optimized towards a biomass production, the typical FBA formulation includes a biomass exchange reaction which renders $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{host}$ a steady-state vector of zeros and therefore trivial. However, if we consider the same maximized flux distribution without biomass exchange (denoted *) then this $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)^*_{host}$ is a positive metabolite rate vector representing the net small molecule precursors available for general macromolecular synthesis at that metabolic state. This metabolite rate can therefore be used to constrain viral metabolite consumption, for which we also assume that some accumulation can occur at the metabolite intersections of host and virus based on the developmental nature of virion replication Eq. A.7.

$$\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{host} + \left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{viral} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} \geqslant 0 \tag{A.7}$$

which, abbreviating $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$ as $\mathbf{b}$ simplifies to:

$$\mathbf{b}_{viral} \geqslant -\mathbf{b}^*_{host}. \tag{A.8}$$

We are left with the requirement only that the net viral metabolites $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{viral}$ not be less than $\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)^*_{host}$ as shown in Eq. A.8, consistent with convention of FBA intake to organism being negative flux).

Once a feasible host flux distribution is selected (Host Supply Step), this provides a simple relation that must be obeyed by viral production flux rates in order to assure a solution exists to the combined

host viral metabolic problem. The method devised to select a vector of maximal viral fluxes/rates (to pass to T7 ODEs) is detailed below. For the sake of space, $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$ is abbreviated $\dot{\mathbf{x}}$ in the following algorithm description. The inputs are the bounding metabolite-rate vector $\dot{\mathbf{x}}_b = -\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right)_{host}$, and the requested reaction rates from the unlimited ODEs $\mathbf{v}_{ode} = \mathbf{v}_{request}$. The values, vectors that are updated each iteration are denoted with a subscript i

1. Accept inputs $\dot{\mathbf{x}}_b$ and $\mathbf{v}_{ode}$. Set $\mathbf{v}_{i,ode} = \mathbf{v}_{1,ode} = \mathbf{v}_{ode}$, and $\dot{\mathbf{x}}_{i,b} = \dot{\mathbf{x}}_{1,b} = \dot{\mathbf{x}}_b$
2. Initialize variables used only within algorithm: The vector of allowed viral fluxes $\mathbf{v}_a = 0$ initialized to zero, the maximum multiplier of the requested viral fluxes $c_{i,max} = c_{1,max} = 1$ is initially unity.
3. Find the maximum c_i such that $\mathbf{S}_{HV}\mathbf{v}_{i,ode}c_i \geqslant \dot{\mathbf{x}}_b$, and subject to $0 \leqslant c_i \leqslant c_{i,max}$.
4. Update allowed viral flux vector, element-wise addition $\mathbf{v}_a = \mathbf{v}_a + \mathbf{v}_{i,ode}c_i$.
5. Update $\dot{\mathbf{x}}_{i+1,b} = \dot{\mathbf{x}}_{i,b} - \mathbf{S}_{HV}\mathbf{v}_{i,ode}c_i$ to account for resources allotted to viral fluxes this step.
6. Remove from subsequent iterations the viral fluxes that consume a resources that has been exhausted. For any element j where $\left(\dot{\mathbf{x}}_{i+1,b}\right)_j = 0$, then for all elements k for which $\left(\mathbf{S}_{HV}\right)_{j,k} < 0$ set $\left(\mathbf{v}_{i,ode}\right)_k = 0$
7. Update maximum fraction $c_{i+1,max} = c_{i,max} - c_i$
8. If ($i > 1$ AND $c_i = 0$) OR $\mathbf{v}_{i+1,ode} = 0$ exit, returning $\mathbf{v}_a$, no greater viral flux possible, all reactions limited by at least one reactant.
9. Else, update $i = i + 1$ return to step 3.

After exit $\mathbf{v}_a$ is the vector of maximum allowed viral reaction rates through production reactions.

Evaluate Final Viral Demand (6)

Once the maximum bounding rates for each viral reaction j have been determined, $\mathbf{v}_{a,j} = r_{j,prod,limit}$, these limits are applied to production rates for viral metabolites as follows:

1. Evaluate the production rate $r_{j,prod}$ for viral species j, which is the positive kinetic term(s) representing synthesis from metabolic precursors. Also evaluate remaining consumption terms of origional kinetic rate equation, the sum of which is $r_{j,cons}$
2. If a limit has been passed and $r_{j,prod} > r_{j,prod,limit}$ then reassign $r_{j,prod} = r_{j,prod,limit}$
3. Update and return $\frac{dC_j}{dt} = r_{j,prod} + r_{j,cons}$ used in the T7 ODEs and for viral concentration output, and the production only pseudo concentration $\frac{dP_j}{dt} = r_{j,prod}$ used in the constraint of viral reactions in the host-viral FBA problem.

The final concentration change over $\Delta t_{integration}$ is then passed to determine the infected host state as well as to update the Viral State of concentrations.

Constrain Host-Viral Reaction Fluxes (7)

After the viral demand reaction rates have been determined as host compatible values, the corresponding fluxes in FBA can be constrained. To constrain the host problem to the final infected state, we set flux bounds of host-viral reactions to the average production rates determined in the evaluation of final viral demand, $\mathbf{v}_{max} = \mathbf{v}_{min} = \mathbf{v}_{ave}$, where:

$$\mathbf{v}_{ave} = \frac{([P_{j,t=(t+\Delta t_{int})}] - [P_{j,t=t}])}{\Delta t_{integration}}. \tag{A.9}$$

The average conversion of viral reactions to fluxes is used here because it enforces conservation of mass, in contrast to the maximum request calculation used in previous steps to allow the virus access to the high instantaneous rates of metabolite use if possible. All host flux bounds are set as previously determined.

Optimization for Infected Host (8)

Following viral constraint, the host flux distribution is still underdetermined, and so maintaining the previous assumption that host pathways remain in the state of maximal host biomass production over the course of infection, the objective function evaluated after viral reactions are constrained is optimization of host biomass production. Finally after the infected host flux distribution is determined the host and environment state can be updated according to the established relationships and another iteration of the integrated simulation proceeds from step (2).

Table A.3: Assumptions and references for construction of phage stoichiometry reactions.

Viral mRNA production reactions	
Nucleotide stoichiometry for all 60 T7 bacteriophage mRNAs from genome sequence and to date publications.	GenBank (NC 001604)
Each added NTP was assumed to produce one pyrophosphate during polymerization.	[136]
Viral Gene Product production reactions	
Amino acid stoichiometry for all 60 T7 bacteriophage proteins.	GenBank
tRNA charging was assumed to hydrolyze 1 ATP to AMP and pyrophosphate.	[136]

Continued on next page

Table A.3 – continued from previous page

Aminoacyl-tRNA binding was assumed hydrolyze 1 GTP to GDP.	[136]
tRNA translocation during translation was assumed to hydrolyze 1 GTP to GDP.	[136]
Polymerization initiation was assumed to hydrolyze 1 GTP to GDP.	[136]
Host genome degradation and recycling reaction	
Nucleotide stoichiometry for T7 bacteriophage and E. coli genomes	GenBank (U00096)
During host genome degradation dNMPs were assumed to be produced in proportion to their content in the E. coli genome.	-
During viral genome production dNTPs were assumed to be used in proportion to their content in the T7 genome and converted into 1 viral dNTP.	-
T7 Genome synthesis reaction	
Nucleotide stoichiometry for T7 bacteriophage genome.	GenBank
Each added dNTP was assumed to produce one pyrophosphate during polymerization.	[136]
DNA helicase was assumed to hydrolyze 2 ATP to ADP per base pair of DNA unwound.	[114]
Lagging strand synthesis: RNAmer primers 4 nucleotides long were assumed to hydrolyze 2 high energy phosphate bonds per polymerized NTP (8 high energy phosphate bonds total)	[136, 69]
Lagging strand synthesis: Okazaki fragments were assumed to be 3,500 bases long on average.	[69]
Lagging strand synthesis: 2 high energy phosphate bonds were hydrolyzed to regenerate the NAD used in ligation.	[136]
Proofreading: Error rate in T7 DNA polymerase was assumed to be 10 percent for dATP, 5 percent for dTTP, 1 percent for dGTP and dCTP. These were counted as additional A,T,G,C incorporated into the genome and then hydrolyzed to dNMPs.	[136, 146]
In all cases high energy phosphate bonds were accounted for as ATP being hydrolyzed to ADP.	-

Table A.1: FBA simulation media definitions.

FBA Metabolite	Tryptone	Glucose	Succinate	Acetate
H20	55 mM	55 mM*	55 mM*	55 mM*
CO2	15 mM*	15 mM*	15 mM*	15 mM*
PI	15 mM*	15 mM*	15 mM*	15 mM*
H	10 mM*	10 mM*	10 mM*	10 mM*
SLF	10 mM*	10 mM*	10 mM*	10 mM*
O2	10 mM*	10 mM*	10 mM*	10 mM*
NH3	10 mM	10 mM	10 mM	10 mM
GLC	-	10 mM	-	-
SUCC	-	-	10 mM	-
AC	-	-	-	10 mM
ASP	5.7 mM	-	-	-
MET	1.6 mM	-	-	-
THR	3.7 mM	-	-	-
ILE	3.3 mM	-	-	-
SER	5.8 mM	-	-	-
LEU	6.3 mM	-	-	-
GLU	16.4 mM	-	-	-
TYR	1.6 mM	-	-	-
PRO	8.0 mM	-	-	-
PHE	2.7 mM	-	-	-
GLY	3.06 mM	-	-	-
HIS	1.6 mM	-	-	-
ALA	3.92 mM	-	-	-
LYS	4.9 mM	-	-	-
CYS	0.5 mM	-	-	-
ARG	1.9 mM	-	-	-
VAL	5.7 mM	-	-	-
TRP	0.5 mM	-	-	-
ASN	0.5 mM	-	-	-

Asterisk (*) indicates nutrients replenished to the given concentration at each time step.

Table A.2: List of FBA Rules Relaxed for Rich Media Growth.

Rxn	Original Rule
BCAAUP1R	b0401
BCAAUP2R	b0401
BCAAUP3R	b0401
ARGA	b2818
ARGB	b3959
ARGCR	b3958
ARGDR	b3359
ARGE1	b3957
ARGFR	b0273
ARGIR	b4254
ARGHR	b3960
SERC1	b0907
DAPB	b0031
LYSA	b2838
DEOB1R	b4383
DEOC	b4381
DEOD1R	b4384
DEOD2R	b4384

Changes relative to rules in iMC1010v2 [51]. Rules relaxed indicates regulatory boolean expression altered to TRUE.

Table A.4: Table of Major T7 ODEs Genome Definition Update.

Parameter	Explanation of Change	Value	Units	Reference
$n_{gp10A,PC}$	GP10A stoichiometry in procapsid. Updated to reflect combined inclusion of GP10A (415) and GP10B (16), treated as single species because T7 ODEs lack description of small percentage programed frameshift distinguishing between these proteins which are of very similar length and stoichiometry.	431	$\frac{molecules}{procapsid}$	[103]
$Gene10B$	GP10B genomic element removed during inclusion in 10A.	-	-	-
$n_{gp6.7,T7}$	GP6.7 DNA completion or ejection protein. Not in virion in original T7 ODEs.	18	$\frac{molecules}{phage}$	[103]
$n_{gp7.3,T7}$	GP7.3 Adsorption and/or ejection protein. Not in virion in original T7 ODEs.	33	$\frac{molecules}{phage}$	[103]
$n_{gp8,T7}$	GP8 head tail connector. Not changed from T7 ODEs.	12	$\frac{molecules}{phage}$	[103]
$n_{gp11,T7}$	GP11 tail protein. Stoichiometry updated.	12	$\frac{molecules}{phage}$	[103]
$n_{gp14,T7}$	GP14 core and/or ejected protein. Stoichiometry updated.	10	$\frac{molecules}{phage}$	[103]
$n_{gp15,T7}$	GP15 core and/or ejected protein. Stoichiometry updated.	8	$\frac{molecules}{phage}$	[103]
$n_{gp16,T7}$	GP16 core and/or ejected protein. Stoichiometry updated.	4	$\frac{molecules}{phage}$	[103]
$\phi_{4.7,4.3,4c,}$ ... $_{3.8,2.5,1.6,1.5}$	Promotor weighting	0.01	-	See text
$\phi_{6.5}$	Promotor weighting	0.05	-	See text
ϕ_{9}	Promotor weighting	0.2	-	See text
$\phi_{13,17}$	Promotor weighting	0.1	-	See text
$\eta_{T\phi}$	T7 terminator efficiency	0.85	Fraction	See text

Changes relative to genome used previously [66], other changes are based on any nucleotide/protein updates in current GenBank sequence for T7 the impact of which is negligible with respect to simulation predictions.

Table A.5: T7 ODEs Parameter updates, values, and references.

Parameter	Explanation of Change	Value	Units	Reference
$k_{T7,transloc}$	Phage T7 translocation rate. All rates were measured at 37C. Second and third rates are determined by E. coli and T7 RNA polymerase transcription rates respectively (see below).	[141, 55, 400]	bp/sec	[76, 31, 42]
$k_{Ec,RNAP}$	E. coli RNA polymerase transcription rate. Measured at 37C.	55	nt/sec/RNAP	[31]
$k_{T7,RNAP}$	T7 RNA polymerase transcription rate. Measured at 37C to be 2x faster than 200 nt/s/RNAP found at 30C.	400	nt/sec/RNAP	[42]
$k_{d,T7mRNA}$	T7 mRNA decay rate. Functional decay of early T7 mRNA found to be 6.5min at 30C, converted to rate, conservatively not adjusted for temperature since data is for early transcripts.	0.001	/sec	[196]
k_{RIBO}	Specific translation rate or ribosomes. Measured at 37C.	42	nt/sec/ribo	[56]
$k_{d,pro}$	T7 Protein Decay Rate. Adjusted from 2.8E-5 /sec at 30C to 37C using Arrhenius approximation.	3.92×10^{-5}	/sec	[115]
C_{repDNA}	T7 DNA Replication critical threshold of gp1 binding by gp3.5. Stated as arbitrarily set in T7v2.5 code.	5×10^{-7}	M	-
$k_{T7,DNAP}$	T7 DNA Replication rate. Found to be over 300 bp/sec/polymerase at 37C using techniques in Tabor et al. 1987.	475	$\frac{bp}{sec \cdot polymerase}$	[171]
$K_{M,DNA}$	DNA polymerase Km. Measured at 37C.	8668	nt/cell	[61]
$k_{T7,pack}$	T7 DNA Packaging rate. Adjusted from 0.702 /min at 30C to 37C using Arrhenius approximation.	0.983	/min	[167]

Continued on next page

Table A.5 – continued from previous page

Parameter	Explanation of Change	Value	Units	Reference
$C_{nuc,PC}$	Procapsid nucleation concentration. Changed to avoid numerical solver error, to single multiple of procapsid stoichiometry (see below). Original value used was for phage P22, understanding of T7 assembly mechanism recently updated.	431	$\frac{molecules}{cell}$	[103, 87]
$k_{d,EcDNA}$	Host genome degradation rate. As per original T7 ODEs, assuming rate consistent with 7.5 to 15 min degradation, 0.85 of host genomic material degraded, and n_{HG} the equivalent number of host genomes from correlation. Factor results from $\frac{\left(4.655\times10^{6}bp\right)\frac{2nt}{bp}}{7.5min\frac{60s}{min}}0.85$	$(0.000176)n_{HG}$	$\frac{nt}{s}$	[151]
$t_{d,EcDNA}$	Host genome degradation period. Original limits 7.5 to 15 min measured for 37C, allowed to continue past 15 min if degradation limited by metabolic interaction.	> 7.5	min	[151]

Appendix B

Supplement for growth and size control

B.1 Glossary of terms

I use a number of terms in this text that are worth defining clearly before continuing. These terms are the important factors to consider when modeling cellular growth and its modulation based on a cell's environment.

Cellular environment: A cell's environment for the purposes of this chapter is defined by the medium it is being grown in while in batch culture under conditions of balanced exponential growth (see below for definition). Medium can vary in nutritional quality with higher quality media yielding a faster rate of growth.

Macromolecular composition: Macromolecular composition refers to the amount of DNA, RNA, and protein and their ratios in a given cell or cell culture. In particular, an important macromolecular component when studying growth rate is the amount of rRNA or ribosomes present in a cell.

Growth rate and balanced growth: Under balanced, steady-state, log-phase growth the amount of every cellular component per unit volume of average cell (or rate of cell division) increases with the same exponential function of time (μ) giving it a doubling time of τ. Furthermore, every component in the system is at saturating, non-limiting concentrations, in contrast to chemostat growth where one component is limiting. Steady state also means that the bacteria have grown for at least 10 generations in a given environment (1000x increase in mass after dilution from an overnight culture) [58]. Under these conditions Equation B.1 applies:

$$\mu = \frac{ln(2)}{\tau} = \frac{\frac{dX}{dt}}{X} \tag{B.1}$$

Where X is any cellular intrinsic property. This means that the mass, count, volume, etc. of any cellular component grows at the same exponential rate and remain proportional to each other (i.e. balanced).

Rate of cell division: This is the frequency at which a bacterial cell divides. At a given growth rate a higher rate of cell division will result in smaller bacterial cells on average.

Cell size: Cell size is the volume or mass of a cell. This is controlled by the balance between cellular growth and division.

B.2 Algorithms

Input : m_{dry} dry mass of the simulated cell
Input : S_k rate of metabolic supply of amino acid k to translation where $k = 1 \rightarrow 21$ for each amino acid
Input : e maximal elongation rate of ribosome
Input : p_i position of ribosome on mRNA transcript $i = 1 \rightarrow n_{ribosome}$
Input : δt current time step
Input : c_{GTP} counts of GTP molecules
Input : L_j length of each mRNA $j = 1 \rightarrow n_{gene}$ for each coding gene.
`/* Calculate resources available for translation */`
Count of each amino acid is calculated available to translation is calculated where $k = 1 \rightarrow 21$ for each amino acid

$$c_{aa,k} = S_k \cdot \delta t \cdot m_{dry}$$

`/* Elongate ribosomes up to limits of sequence, amino acids, or energy */`
for *each ribosome i on mRNA transcript j* **do**
  1. Based on ribosome position p_i on mRNA transcript and maximal elongation rate e determine "stop condition" position (t_i) for ribosome assuming no amino acid limitation.

$$t_i = \texttt{min}(p_i + e \cdot \delta t,\ L_j)$$

  Stop condition is either maximal elongation rate scaled by the time step or the full length of sequence (i.e. the ribosome will terminate in this time step).
  2. Derive sequence between ribosome position (p_i) and stop condition (t_i).
  3. Based on derived sequence calculate the number of amino acids required to polymerize sequence $c^{req}_{aa,i}$ and number of GTP molecules required c^{req}_{GTP}.
  4. Elongate up to limits:
  if $\texttt{all}(c^{req}_{aa,k} < c_{aa,k})$ ***and*** $c^{req}_{GTP} < c_{GTP}$ **then**
    Update the position of each ribosome to stop position

$$p_i = t_i$$

  else
    Use "greedy" algorithm that attempts to equally process each ribosome. Update position of each ribosome to maximal position given the limitation of $c_{aa,k}$ and c_{GTP}.
  end
  5. Update counts of $c_{aa,k}$ and c_{GTP} to reflect polymerization usage.
end
`/* Terminate ribosomes that have reached the end of their mRNA transcript */`
for *each ribosome i on transcript j* **do**
  if $p_i == L_j$ **then**
    1. Increment count of protein that corresponds to elongating polypeptide that has terminated.
    2. Delete ribosome and increment 30S and 50S counts.
  end
end
Result: Each ribosome is elongated up to the limit of available mRNA sequence, maximal elongation rate, amino acid, or GTP limitation. Ribosomes that reach the end of their transcripts are terminated and released.

Algorithm 1: New algorithm for peptide chain elongation and termination

Input : $e_{expected}$ expected elongation rate of ribosome ($e_{expected} < e_{max}$)
Input : p_i position of ribosome on mRNA transcript $i = 1 \rightarrow n_{ribosome}$
Input : δt current time step
Input : c_{GTP} counts of GTP molecules
Input : L_j length of each mRNA $j = 1 \rightarrow n_{gene}$ for each coding gene.
`/* Calculate resources available for translation */`
for *each ribosome i on mRNA transcript j* **do**
  Derive amino acid sequence between ribosome position (p_i) and `min(`$p_i + e_{expected} \cdot \delta t$, L_j`)`. Sum each amino acid in derived sequence and add to cumulative sum $c_{aa,k}$, which is the total amount of each amino acid k required to polymerize every sequence in front of a ribosome up to the $e_{expected}$ elongation rate.
end
`/* Elongate ribosomes up to limits of sequence, amino acids, or energy */`
for *each ribosome i on mRNA transcript j* **do**
  1. Based on ribosome position p_i on mRNA transcript and expected elongation rate $e_{expected}$ determine "stop condition" position (t_i) for ribosome assuming no amino acid limitation.
    $t_i =$ `min(`$p_i + e_{expected} \cdot \delta t$, L_j`)`
  Stop condition is either maximal elongation rate scaled by the time step or the full length of sequence (i.e. the ribosome will terminate in this time step).
  2. Derive sequence between ribosome position (p_i) and stop condition (t_i).
  3. Based on derived sequence calculate the number of amino acids required to polymerize sequence $c^{req}_{aa,i}$ and number of GTP molecules required c^{req}_{GTP}.
  4. Elongate up to limits:
  if `all(`$c^{req}_{aa,k} < c_{aa,k}$`)` ***and*** $c^{req}_{GTP} < c_{GTP}$ **then**
    Update the position of each ribosome to stop position
      $p_i = t_i$
  else
    Use "greedy" algorithm that attempts to equally process each ribosome. Update position of each ribosome to maximal position given the limitation of $c_{aa,k}$ and c_{GTP}.
  end
  5. Update counts of $c_{aa,k}$ and c_{GTP} to reflect polymerization usage.
end
`/* Terminate ribosomes that have reached the end of their mRNA transcript */`
for *each ribosome i on transcript j* **do**
  if $p_i == L_j$ **then**
    1. Increment count of protein that corresponds to elongating polypeptide that has terminated.
    2. Delete ribosome and increment 30S and 50S counts.
  end
end
Result: Each ribosome is elongated up to the limit of available mRNA sequence, expected elongation rate, amino acid, or GTP limitation. Ribosomes that reach the end of their transcripts are terminated and released.

Algorithm 2: Currently published algorithm for peptide chain elongation and termination

Appendix C

Crick's complete solution of *E. coli*, 40 years later

C.1 Introduction

This document serves as a companion to our source code and simulation archive[1] in support of the main text. Similar to the structure of the supplement to the *M. genitalium* model [102], we first present an overview of our computational methods and then delve into the implementation details. We then provide a description of our experimental methods used to collect (1) expression data to initially parameterize the model and (2) protein decay rates used to refine the model.

Constructing a gene-complete whole-cell model of *E. coli* is a major undertaking. While the overarching engineering goal for the *M. genitalium* model was to include the function of every annotated gene, *E. coli* contains roughly ten times as many genes as *M. genitalium* and 50-100 times as many molecules that can interact, presenting us significant challenges in both modeling and computation. Although a gene-complete model of *E. coli* is our long-term goal, before focusing on increasing the number of genes, we decided to focus on improving and expanding our utilization of data. We chose to work towards a whole-cell model that integrated as much organism-specific data as possible, preferably across multiple environments and growth conditions, in which results and predictions could be experimentally verified. Building larger, gene-complete models would be impossible without the innovation in parameter estimation, data integration, modeling framework extensibility, and feedback regulation this goal required.

Perhaps the largest difference between *M. genitalium* and *E. coli* physiology is the extensive amount

[1] `https://simtk.org/projects/ecoli`. Updates and errata will be available at this location as well.

of regulation and control present in the latter. Whereas *M. genitalium* can only be cultured—and thus simulated—in a rich medium, *E. coli* can grow in a number of different environments, at a number of different growth rates (see Section C.2.6). This behavior is mediated by extensive regulatory mechanisms at the transcriptional and post-transcriptional levels that we can now simulate and that we describe in the following sections.

As discussed in the main text, while the model presented here is not gene-complete, it incorporates the biological processes for which the majority of high-throughput data is available. We spent considerable effort evaluating data sets and merging them into our framework. Ultimately, this enabled us to make the quantitative comparisons presented in the main text—comparisons that could not be made when modeling *M. genitalium*. However, this version of the *E. coli* model lacks several of the sub-models implemented in the *M. genitalium* model. Going forward, we plan to continue working toward a gene-complete model of *E. coli*.

For readers familiar with our *M. genitalium* work, we summarize our improvements over that model in terms of *Modeling* and *Computation*:

Modeling:

- We have a quantitative model of transcriptional regulation that incorporates the function of 22 transcription factors regulating 355 genes. This includes one- and two-component signaling processes to modulate transcription factor activity, as well as the modulation of RNA polymerase recruitment via TF-DNA binding interactions.
- The metabolic model is much more robust and includes detailed quantitative (Michaelis-Menten) parameters for 340 reactions. The metabolic model now maintains concentrations of metabolite pools subject to resource availability rather than producing metabolites in a fixed ratio at every time step. This enables the metabolic model to adjust to time-dependent/cell cycle-dependent behavior from other simulated processes while maintaining homeostasis.
- We now have an implementation of growth-rate control that enables our simulated *E. coli* cells to grow at different doubling times as a function of the environment. This improvement, supported by our model of DNA replication which can track multiple rounds of replication, is showcased in the main text and was not possible in *M. genitalium* simulations.
- Our model of translation uses translational efficiency data to inform ribosome binding to mRNA transcripts.
- We have a more detailed model of RNA decay that incorporates both the rates of degradation due to endonuclease-mediated cleavage and the rates of transcript digestion by exoRNases.

Computation:

- We have decreased simulation run-time by nearly two orders of magnitude. Whereas *M. genitalium* simulations took roughly 10 hours to run, *E. coli* simulations—which account for 50 times more molecules—take approximately 15 minutes to simulate the life cycle of an *E. coli* cell. We achieved this by (1) improving file I/O, (2) writing inner loops in Cython or C, and (3) warm-starting the linear solver in our metabolic model.
- This improvement in run-time enables us to reliably simulate multiple generations of cells (e.g., as shown in the main text), which was not possible with the *M. genitalium* simulations.
- Additionally, we have improved the whole-cell application programming interface (API). Code is much more readable, and on-boarding new researchers takes roughly 2 weeks rather than 6 months.

With this overview in mind, we now present our computational methods.

C.2 Computational methods

Figure C.1 summarizes the overall workflow of running and analyzing simulations. We begin with data sets from the primary literature, our own experiments, and databases (e.g., EcoCyc) which we unify into a `KnowledgeBase` (KB). We then reconcile parameters using a heuristic fitting procedure. Using these reconciled parameters, we run simulations and save their output for downstream analysis. In principle, as highlighted in Figure C.1, the first three procedures can be performed just once and the remaining procedures can be performed multiple times to explore the effects of different perturbations.

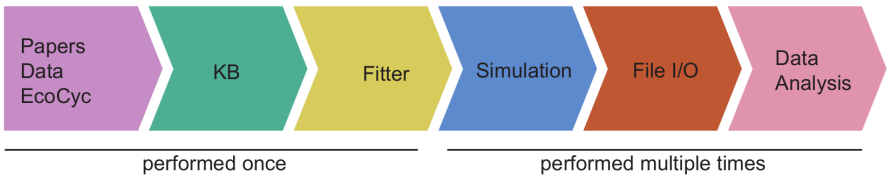


Figure C.1: **Overall workflow**
Starting with curated datasets, the `KnowledgeBase` is created. Using the `KnowledgeBase`, parameters are reconciled in the Fitter and used as initial conditions for the Simulation. For each simulation run, these preparatory steps are performed once. After all simulations are performed, visualizations are produced with analysis scripts.

C.2.1 Reconstruction and fitting

The *E. coli* model is based on a reconstruction of *E. coli* physiology that considers a multitude of data including both single-cell and population measurements. Some examples include average cell masses, macromolecular composition (e.g., RNA, protein, and DNA mass fractions), gene expression, protein and RNA half-lives, enzymatic rates, cell cycle parameters, and gene annotation. The bulk of the experimental data we used comes from three lab strains of *E. coli*: K-12 MG1655, B/r, and BW25113. The model can therefore be thought of as a composite strain which uses all of these data. In most cases, this assumption likely holds. However, in some cases these strains have different physiology with the most notable example being a difference in growth rate, under similar environmental conditions, between MG1655 (where most high-throughput data comes from) and B/r (where detailed composition data have been obtained). Our model was optimized using the B/r growth rates.

Due to the diverse nature of the data sources considered for the model, it was necessary to perform data reconciliation to create a single consistent parameter set. For example, cell mass must, on average, grow exponentially and double within the expected doubling time given all of the parameters. More generally, the reconciled parameter set, or `KnowledgeBase`, should produce balanced exponential growth while satisfying constraints set by cell theory and physiology.

Testing whether a set of parameters produced balanced exponential growth would ideally take place in the full large-scale cell simulation framework. However, even with >1 order of magnitude improvements in runtime over the *M. genitalium* model, this proves too computationally intensive. Therefore, to enable computationally tractable parameter fitting we created surrogate models that approximated the average behavior of the full-scale simulation using heuristic routines. For example, we approximate the output of a dynamic, stochastic simulation that produces a spectrum of proteins with a statistical distribution of protein counts. These surrogate models were executed in order to numerically fit parameters to our constraints (e.g., of cell theory).

In cases where parameters were inconsistent with the constraints imposed, or no value was known for a parameter, one of two operations was performed: (1) iteratively changing the parameter value (such as the expression of RNA polymerases, ribosomes, and metabolic enzymes) until constraints are satisfied or (2) calculating the parameter value from other known and reconciled data (such as using the dissociation constants of ligands binding their binding partners to calculate the anticipated intracellular concentration of ligands).

C.2.2 Estimating the number of parameters

We are often asked how many parameters are contained our models. Estimating this as a simple number belies the complex and heterogeneous nature of the model, where each sub-model is specified using a different mathematical formalism. At one extreme, every piece of data could be considered a parameter (e.g., we could count as a parameter every single nucleotide in the chromosome sequence), but we feel this isn't a helpful estimate. Likewise, the stoichiometric coefficients in the metabolic network, molecular masses of each mRNA and protein species, as well as all of the data in our RNA expression databases, could all be treated as parameters—but we don't take them into account when answering this question. While one could therefore count parameters in many different ways, Table C.1 provides the breakdown for the over nineteen thousand parameters that we state in the main text.

Estimating the number of fit parameters can also be ambiguous, as everything is linked, and in some cases it can be impossible to change a parameter in isolation. For example, given the fact that we account for (1) all molecules in a cell and (2) the total mass of the cell, changing counts of one molecule (by changing one parameter) means that we either have to update counts of other molecules (to maintain the total mass of the cell), or change the total mass of the cell (to account for the change in molecule counts). Not withstanding this difficulty, Table C.2 provides our breakdown of the major sets of parameters that were modified from their initial values.

Description	Parameter count
RNA Polymerase recruitment strengths (basal and TF-modulated)	4996
EndoRNase-RNA affinities (govern decay rate of each RNA)	4558
Translation efficiencies	4353
Protein half-lives	4353
Metabolic reaction constraints	616
Metabolite pools (basal condition)	140
External exchange flux bounds (basal condition)	54
Dissociation constants (e.g., for ligand-TF binding)	28
Reaction rates for two-component systems	21
Total	19119

Table C.1: Estimate of number of parameters in the model.

Description	Parameter count
RNA expression for the ribosome (58 genes)	58
TF-ligand affinities for TrpR, ArgP, Lrp, PutA, MetJ, CytR	6
RNA expression for *pabC*, *menH*, *cdsA*, *yibQ**, *atoB**	5
Translational efficiencies for *pabC*, *menH*, *yibQ**, *atoB**	4
RNA expression for RNA Polymerase (*rpoA*, *rpoB*, *rpoC*)	3
RNA degradation rate for RNA Polymerase (*rpoA*, *rpoB*, *rpoC*)	3
RNA degradation rate for *pabC*, *cdsA*	2
Protein decay rate of YibQ*	1
Total	82

Table C.2: Estimate of number of fit parameters in the model. *Denotes that it was necessary to modify this parameter for the anaerobic simulations to be viable.

C.2.3 Initial conditions

The state of the *E. coli* simulation is initialized immediately after cell division. Using the unified parameter set created during `Reconstruction` and stored in the `KnowledgeBase` the counts and properties of every species are set, using a statistical model to give each simulation a uniquely determined random initial state that, on average, fits experimental data.

Initializing RNA and protein counts The counts of RNA and protein molecules are initialized as follows. First, the total counts of RNA and protein molecules of each species are computed using Equation C.1.

$$M_{total} = \sum_i c_i \cdot MW_i / N_a \tag{C.1}$$

Where M_{total} is the total mass of RNA or protein, c_i, f_i, and MW_i are the counts, mass fraction, and molecular weight of RNA or protein i, and N_a is Avogadro's number. Substituting $c_i = c_{total} \cdot f_i$, where c_{total} is the total counts of RNA or protein species, and rearranging gives Equation C.2.

$$c_{total} = \frac{M_{total}}{\vec{f} \cdot \vec{MW} / N_a} \tag{C.2}$$

The total masses of RNA and protein per cell (M_{total}) as well as their expected distributions ($\vec{f}$) are all known from reconciled datasets in `KnowledgeBase`.

Total counts and the expected distribution are then used to sample a `multinomial` distribution to statistically compute the counts of each individual RNA or protein species using Equation C.3.

$$\vec{c} = \texttt{multinomial}(\vec{f}, c_{total}) \tag{C.3}$$

Details of this algorithm can be found in Algorithm 3.

Input : $M_{total}^{RNA}, M_{total}^{Protein}$ Total mass per cell of RNA and protein

Input : MW_i, MW_j Molar molecular weights of RNA $i = 1$ **to** n_{RNA} and protein $j = 1$ **to** $n_{protein}$

Input : f_i^{RNA} mass fraction based on RNA expression of RNA $i = 1$ **to** n_{RNA}

Input : $k_{d,j}$ degradation rate of protein $j = 1$ **to** $n_{protein}$

Input : N_a Avogadro's number

Input : ψ_j translational efficiencies of each mRNA $j = 1$ **to** $n_{protein}$

1. Calculate total counts of RNAs (C_{total}^{RNA}) based on total RNA mass (M_{total}^{RNA}) and distribution of expression ($\vec{f}^{RNA}$) from `KnowledgeBase`

$c_{total}^{RNA} = \frac{M_{total}^{RNA}}{\vec{f}^{RNA} \cdot \vec{MW}/N_a}$

2. Calculate counts of each RNA ($\vec{c}_{RNA}$) by sampling a `multinomial` distribution c_{total}^{RNA} times weighted by the expected distribution of expression ($\vec{f}^{RNA}$).

$\vec{c}_{RNA} = \texttt{multinomial}\ (\vec{f}^{RNA}, c_{total}^{RNA})$

3. Calculate expected distribution of protein counts ($\vec{f}^{protein}$) based on expected distribution of RNA counts ($\vec{f}^{RNA}$), translational efficiencies ($\vec{\psi}$), protein degradation rates ($\vec{k_d}$), and dilution using a steady state assumption.

$\vec{f}^{protein} = \frac{\vec{f}^{RNA} \cdot \vec{\psi}}{\frac{ln(2)}{\tau} + \vec{k_d}}$

4. Calculate total counts of proteins ($c_{total}^{protein}$) based on total protein mass ($M_{total}^{protein}$) and distribution of counts ($\vec{f}^{protein}$).

$c_{total}^{protein} = \frac{M_{total}^{protein}}{\vec{f}^{protein} \cdot \vec{MW}/N_a}$

5. Calculate counts of each protein ($\vec{c}_{protein}$) by sampling a `multinomial` distribution $c_{total}^{protein}$ times weighted by the expected distribution of expression ($\vec{f}^{protein}$).

$\vec{c}_{protein} = \texttt{multinomial}\ (\vec{f}^{protein}, c_{total}^{protein})$

Result: Counts of RNA are set at the beginning of the first generation of simulated cells

Algorithm 3: Algorithm for initializing counts of RNA and protein in *E. coli* model

Initializing small molecule counts The counts of small molecules such as cytoplasmic and membrane constituents are initialized as follows. Expected concentrations of small molecules are either known experimentally, or computed from an FBA biomass reaction and stored as a reconciled dataset in the `KnowledgeBase`. The volume of the cell is computed using its mass divided by its

density. Therefore adding counts of small molecules to the cell in order to match a concentration will necessarily change the volume of the cell. Using a system of linear equations, the counts of each small molecule and the new adjusted mass of the cell (adding the mass of the new small molecule counts) is calculated. The details of this calculation can be found in Algorithm 4.

Input : C_k^{SM} concentration of each small molecule $k = 1$ **to** n_{SM}

Input : ρ density of cell

Input : MW_k molecular weight of small molecule $k = 1$ **to** n_{SM}

Input : m_{init} Initial mass of the cell only considering RNA, protein, and DNA

1. Calculate masses of each metabolite to add (m_k) in order to achieve known metabolite concentration (C_k^{SM}) from `KnowledgeBase` assuming cell volume is calculated by dividing the cell mass by its density.

$$\begin{bmatrix} \frac{\rho}{C_1^{SM} \cdot MW_1} - 1 & -1 & \dots & -1 \\ -1 & \frac{\rho}{C_2^{SM} \cdot MW_2} - 1 & & \vdots \\ \vdots & & \ddots & -1 \\ -1 & \dots & -1 & \frac{\rho}{C_k^{SM} \cdot MW_k} - 1 \end{bmatrix} \cdot \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{bmatrix} = \begin{bmatrix} m_{init} \\ m_{init} \\ \vdots \\ m_{init} \end{bmatrix}$$

2. Calculate expected counts of each small molecule (c_k^{SM}).

$c_k^{SM} = \frac{m_k}{MW_k \cdot N_a}$

Result: Counts of each small molecule are calculated and set in state

Algorithm 4: Algorithm for initializing counts of small molecules in *E. coli* model

Initializing chromosome state In *E. coli* there are potentially multiple rounds of replication proceeding simultaneously at any point in the cell cycle. The simulation begins immediately after cell division and the number and position of any replication forks that are inherited from previous generations must be determined to correctly initialize the simulated cell. The number of origins of replication, replication forks, and their positions are initialized as follows.

First, the number of rounds of replication that on average need to proceed simultaneously can be estimated in an average cell in a population using the length of time required to replicate the chromosome (C period) and the length of time for cytokinesis (D period) as well as the expected doubling time given the environment (τ). The number of simultaneous rounds (n_{limit}) can be calculated with Equation C.4 as the ratio of C+D period over the doubling time [28]. Because we are considering a specific cell and not an average of a population of cells, the number of rounds of replication needs to be an integer, and we take the floor because a fractional round of chromosome initiation has not yet occurred.

$$n_{limit} = floor(\frac{C+D}{\tau}) \tag{C.4}$$

For every round of replication proceeding there are a pair of replication forks and a pair of origins of replication. We are assuming that on average a cell after division has inherited one chromosome molecule (i.e. no more than one terC), and that it may have more than one round of replication proceeding on it (i.e. number of oriC ≥ 1). Therefore the number of origins of replication (n_{origin}) is defined by Equation C.5.

$$n_{origin} = 2^{n_{limit}} \tag{C.5}$$

Finally, the position between the oriC and the terC of each replication fork needs to be determined on average. This can be calculated with Equation C.6 where f is the fraction of length between the origin and terminus of replication that the replication fork has proceeded for the nth round of replication (where n can be any integer value between 1 and n_{limit}).

$$f = 1 - \frac{n \cdot \tau - D}{C} \tag{C.6}$$

Where n is every integer value 1,2,...n_{limit}. The position in nucleotides (l) can then be calculated from Equation C.7 where L is the total length of the chromosome in *E. coli*.

$$l = f \cdot \frac{L}{2} \tag{C.7}$$

Proper initialization of the cell ensures the simulation begins close to the steady state of the system, and in practice the simulation is relatively stable. Perturbations in the ratio of cell mass to number of origins of replication quickly re-converge to steady state for a given environment. A detailed algorithm for chromosome initialization can be found in Algorithm 5.

Input : C length of C period
Input : D length of D period
Input : τ expected doubling time
Input : L length of chromosome in nucleotides
$n_{limit} = \texttt{floor}(\frac{C+D}{\tau})$
$n = 1$ **while** $n \leq n_{limit}$ **do**
  1 Determine initial number of forward and reverse replication forks ($n_{fork,f,init}$ and $n_{fork,r,init}$) for the given round of replication
  $n_{fork,f,init} = 2^{n-1}$
  $n_{fork,r,init} = 2^{n-1}$
  2 Determine position of each fork on forward and reverse strand as a fraction of total chromosome length (f)
  $f = 1 - \frac{n \cdot \tau - D}{C}$
  3 Calculate position of each fork on forward and reverse strand (l) in nucleotides and initialize $n_{fork,f,init}$ and $n_{fork,r,init}$ DNA polymerases at the calculated positions
  $l_{fork,f,init} = f \cdot \frac{L}{2}$
  $l_{fork,r,init} = f \cdot \frac{L}{2}$
  4 Increment round of replication that is being initialized
  $n = n + 1$
end
$n_{origin,init} = 2^{n_{limit}}$
Result: State of chromosome in cell is correctly initialized around the average of a population

Algorithm 5: Algorithm for initializing chromosome state in *E. coli* model

C.2.4 Simulation algorithm

A whole-cell model may be thought of as a system of ordinary differential equations (ODEs) where the cellular states are analogous to the ODEs' state variables and the cellular processes are analogous to the differential equations. Extending this analogy, the *E. coli* model is simulated using an algorithm that is comparable to those used to numerically integrate ODEs. The only significant difference from ODE numerical integration is that shared resources stored in cellular states must be partitioned to each cellular process in order to ensure mass conservation. Algorithm 6 and Figure C.2 summarize the simulation algorithm to execute a time step. The temporal evolution of the cell state is calculated on a short time scale (typically <1 second) by allocating cell state variables among processes (described in Algorithm 6 under Allocate shared resources), and executing the process code that updates counts in the state variables until the cell divides. Critically, we make the assumption that over a short time scale, each process acts independently.

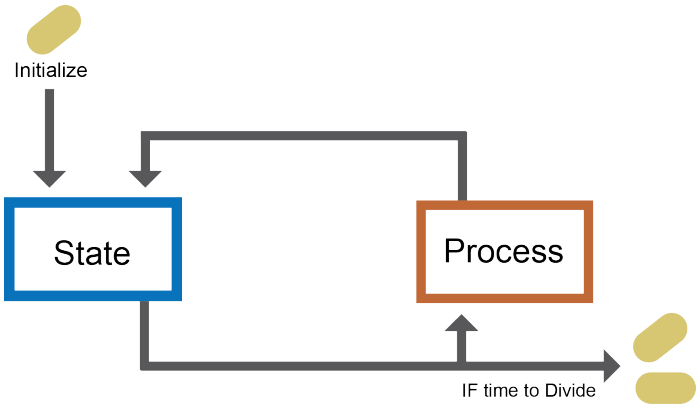


Figure C.2: **Schematic of whole-cell simulation algorithm**
The model takes in a set of initial conditions about a single cell and encodes this information as States, which contains information about each molecule. At the start of a time step these molecules are fed into Processes, while at the end of a time step the molecule information within States is updated. This sequence is iterated over the entire life cycle of the cell until it divides, which constitutes a single generation. Each of the daughter cells could then serve as the initial conditions for a new generation.

Initialize simulation states (described in Algorithm 3, 4, 5)
repeat
  `/* Allocate shared resources */`
  for *each molecule* i **do**
    for *each process* j **do**
      1. Calculate demand $d_{i,j}$ of process j for molecule i.
      2. Divide total count c_i of molecule i into partition $p_{i,j}$, for each process proportional to the demand such that $p_{i,j} = c_i \frac{d_{i,j}}{\sum_j d_{i,j}}$.
    end
  end
  `/* Calculate temporal evolution */`
  for *each process* j **do**
    1. Retrieve partitioned molecules $p_{i,j}$.
    2. Compute the contribution of process j to the temporal evolution of the partitioned molecules $\Delta p_{i,j}$.
    3. Update partitioned molecule counts $p_{i,j} = p_{i,j} + \Delta p_{i,j}$.
  end
  `/* Merge partitioned molecules */`
  for *each molecule* i **do**
    Update counts c_i based on updated partitions computed in each process, $c_i = \sum_j p_{i,j}$
  end
  Increment simulation step by 1
until *cell division*;
Result: Whole-cell model is executed for one cell cycle

Algorithm 6: Algorithm for whole-cell dynamic simulation

C.2.5 States and Processes

States

The simulation States are defined as the counts, locations, and attributes of every species in the model at a given time step, which are then operated on by Processes. There are two classes of States within the model - `BulkMolecules` and `UniqueMolecules`.

The `BulkMolecules` state tracks species in the simulation where individuals are not further distinguished from each other. For example, two ATP molecules in the cytoplasm are considered identical and tracked in `BulkMolecules`.

The `UniqueMolecules` state tracks species in the simulation where individuals are distinguishable

from each other by an attribute and cannot be interchanged without effect. For example, two ribosomes on different mRNA transcripts are uniquely identified by the transcript they are translating and their location on the transcript.

Processes

The simulation Processes update the simulation States from one time step to the next. Each `Process` represents an aspect of physiology of an *E. coli* cell. We discuss the implementation of each Process in detail in Section C.3.

C.2.6 Environments

We simulate different environments by adding or removing exchange flux bounds for the metabolic network (see Section C.3.2). Based on these bounds, the `Metabolism` process updates cellular concentrations of small molecules, and the `Transcription` and `Translation` processes modulate the activities and expression of RNA polymerases and ribosomes to globally shift cell composition. The transcriptional regulatory network (see Section C.3.1) responds to the new small molecule concentrations and adjusts gene expression appropriately.

In the main text we simulate 3 different environments (minimal media with glucose, minimal media with glucose supplemented with amino acids, and anaerobic - minimal media with no oxygen) which map to 3 different growth rates (doubling times of: 44 minutes, 25 minutes, and 100 minutes, respectively) to demonstrate the ability to shift cell composition. In benchmarking our simulations, we also simulated environments that would activate and inactivate each of the transcription factors (not shown).

C.2.7 Computational implementation and workflow

Programming language

The model is primarily implemented in Python, with Cython used for computationally intensive inner loops.

Workflow management

Workflows are defined, managed, and executed using FireWorks - a free open-source code for automating workflow execution which can be defined in Python (https://pythonhosted.org/FireWorks/) [89].

Model API

The *E. coli* model uses an organism independent whole-cell modeling application program interface (API) developed to facilitate model development, human readability, and consistent coding style. The API classes include I/O tools such as `Listeners` for recording data, `Views` for managing interactions between Processes and States, `Containers` for States, and other functions that are useful for writing Processes.

Listeners `Listeners` is a class that facilitates writing data to disk during simulation runtime. They create a human readable interface and reduce both the file size of simulation output and post-hoc computation by saving user-specified quantities computed during a simulation.

Views `Views` is a class that abstracts away potential issues with indexing into large matrices and provides a programmatic interface that allow States and Processes to interact cleanly during simulation.

C.3 Processes

The Processes of the *E. coli* model span several major areas of cellular physiology. We have clustered the Processes into groups that correspond to figures in the main text (Central Dogma, Metabolism, and Balanced Growth) and present these groups in that same order below. We modeled Processes using the most appropriate mathematics for their individual network topology and degree of experimental characterization. Each process is a computational representation of chemical reactions or transformations grouped by a physiological function. The actual division of reactions across processes is a modeling decision made during model construction, and the number of Processes does not reflect their complexity or scope. The inputs and outputs of each `Process` are the counts of metabolites or macromolecules and the catalytic capacity or configuration of the enzymes that catalyze the reactions in each `Process`. This section details the model implementation, computational algorithm, associated data, and relevant code for each `Process`.

C.3.1 Central dogma

Transcription

Model implementation

Transcription occurs through the action of two processes in the model: `TranscriptInitiation` and `TrancriptElongation`. `TranscriptInitiation` models the binding of RNA polymerase to each gene. The number of initiation events per gene is proportional to the number of free RNA polymerases weighted by each gene's synthesis probability. Details are in Algorithm 7.

`TranscriptElongation` models nucleotide polymerization into RNA molecules by RNA polymerases. Polymerization occurs across all polymerases simultaneously and resources are allocated to maximize the progress of all polymerases up to the limit of the expected polymerase elongation rate and available nucleotides. The termination of RNA elongation occurs once a RNA polymerase has reached the end of the annotated gene. Details are in Algorithm 8.

Difference from *M. genitalium* model

The *M. genitalium* model modeled RNA polymerase as existing in 4 states: free, non-specifically bound on a chromosome, bound to a promoter, and actively transcribing a gene. The *E. coli* model simplifies this by assuming RNA polymerase exists in two states: free and actively transcribing. Every time step, free RNA polymerase transitions to the actively transcribing state to maintain an experimentally-observed active fraction of RNA polymerase. The *E. coli* model does not yet include sigma, elongation or termination factors. The *E. coli* model also currently treats each gene as its own transcription unit.

Input : f_{act} fraction of RNA polymerases that are active

Input : r expected termination rate for active RNA polymerases

Input : $v_{\text{synth,i}}$ RNA synthesis probability for each gene where $i = 1$ **to** n_{gene}

Input : $c_{RNAP,f}$ count of free RNA polymerase

Input : `multinomial()` function that draws samples from a multinomial distribution

1. Calculate probability (p_{act}) of a free RNA polymerase binding to a gene.

$$p_{act} = \frac{f_{act} \cdot r}{1 - f_{act}}$$

2. Calculate the number of RNA polymerases that will bind and activate ($c_{RNAP,b}$).

$$c_{RNAP,b} = p_{act} \cdot c_{RNAP,f}$$

3 Sample multinomial distribution $c_{RNAP,b}$ times weighted by $v_{synth,i}$ to determine which genes receive a RNA polymerase and initiate ($n_{init,i}$).

$$n_{init,i} = \texttt{multinomial}(c_{RNAP,b}, v_{synth,i})$$

4 Assign $n_{init,i}$ RNA polymerases to gene i. Decrement free RNA polymerase counts.

Result: RNA polymerases bind to genes based on the number of free RNA polymerases and the synthesis probability for each gene.

Algorithm 7: Algorithm for RNA polymerase initiation on DNA

Input : e expected RNA polymerase elongation rate in given environment

Input : L_i length of each gene $i = 1$ **to** n_{gene} for each coding gene.

Input : p_j gene position of RNA polymerase $j = 1$ **to** n_{RNAP}

Input : $c_{nuc,k}$ counts of nucleotide $k = 1$ **to** 4

Input : δt length of current time step

`/* Elongate RNA transcripts up to limits of sequence or nucleotides */`

for *each RNA polymerase j on gene i* **do**

1. Based on RNA polymerase position p_j on a gene i and maximal elongation rate e determine stop condition (s_j) for RNA polymerase j assuming no nucleotide limitation.

$$s_j = \texttt{min}(p_j + e \cdot \delta t,\ L_i)$$

Stop condition is either maximal elongation rate scaled by the time step or the full length of sequence (i.e. the RNA polymerase will terminate in this time step).

2. Derive sequence between RNA polymerase position (p_j) and stop condition (s_j).

3. Based on derived sequence calculate the number of nucleotides required to polymerize sequence $c^{req}_{nuc,k}$.

4. Elongate up to limits:

if $\texttt{all}(c^{req}_{nuc,k} < c_{nuc,k})$ **then**

Update the position of each ribosome to stop position

$$p_j = s_j$$

else

4a. Attempt to elongate all RNA fragments.

4b. Update position of each polymerase to maximal position given the limitation of $c_{nuc,k}$.

end

5. Update counts of $c_{nuc,k}$ to reflect polymerization usage.

end

`/* Terminate RNA polymerases that have reached the end of their gene */`

for *each RNA polymerase j on gene i* **do**

if $p_j == L_i$ **then**

1. Increment count of RNA that corresponds to elongating RNA transcript that has terminated.

2. Increment free RNA polymerase counts.

end

end

Result: Each RNA transcript is elongated up to the limit of available gene sequence, expected elongation rate, or nucleotide limitation. RNA polymerases that reach the end of their genes are terminated and released.

Algorithm 8: Algorithm for mRNA elongation and termination

Associated data

Parameter	Symbol	Units	Value	Reference
Active fraction of RNAP	f_{act}	-	0.20 (growth-dependent)	[33]
RNA synthesis probability[(1)]	p_{synth}	-	[0, 0.015]	See Table C.4
RNAP elongation rate	e	nt/s	50 (growth-dependent)	[33]

Table C.3: Table of parameters for Transcript Initiation and Elongation processes.
[(1)]RNA synthesis probabilities were calculated as the relative fraction of RNA production (which is equal to the RNA degradation) for a given gene.

Associated files

`wcEcoli` Path	File	Type
`wcEcoli/models/ecoli/processes`	`transcript_initiation.py`	process
`wcEcoli/models/ecoli/processes`	`transcript_elongation.py`	process
`wcEcoli/reconstruction/ecoli/dataclasses/process`	`transcription.py`	data

Table C.4: Table of files for transcription.

Transcription regulation

Model implementation

There are two aspects to modeling transcriptional regulation: (1) modeling the activation or inhibition of a transcription factor (e.g., by a ligand), and (2) given an active transcription factor, modeling its effect on RNA polymerase recruitment to a promoter site. We address these topics sequentially below.

Modeling transcription factor activation

We consider three classes of transcription factors based on their mechanism of activation:

1. **One-component systems**: transcription factors that are directly activated or inhibited by a small molecule ligand. Examples of this class include the repressor TrpR which binds tryptophan, and the inducer AraC which binds arabinose.

2. **Two-component systems**: transcription factors that are paired with a separate sensing protein that responds to an environmental stimulus (these are simple analogs to the vast, complicated signaling networks that exist in eukaryotic cells). The sensing protein phosphorylates the cognate transcription factor in a condition-dependent fashion. Examples include

ArcA which is phosphorylated by its cognate ArcB in anaerobic conditions, and NarL which responds to the presence of nitrate when phosphorylated by its cognate sensor NarX.

3. **Zero-component systems**: transcription factors that are considered to be active whenever they are expressed. Examples include the Fis and Hns proteins. These two proteins, for instance, are important in maintaining higher-order DNA structure and likely have complex feedback loops modulating their activity. Because this complexity is not yet fully understood, we make the simplifying assumption that these proteins are always active unless they are knocked out.

One-component systems

For a transcription factor with concentration T whose activity is directly modulated by a ligand with concentration L that binds with stoichiometry n, we assume that the two species achieve equilibrium on a short time scale and that the affinity of the two molecules can be described by a dissociation constant K_d:

$$nL + T \leftrightarrow T^* \tag{C.8}$$

where T^* represents the concentration of the ligand-bound transcription factor.

With the dissociation constant K_d defined as:

$$K_d = \frac{L^n \cdot T}{T^*} \tag{C.9}$$

we have:

$$\frac{T^*}{T_T} = \frac{L^n}{L^n + K_d} \tag{C.10}$$

where T_T is the total concentration of the transcription factor, both ligand-bound and unbound. As we can see, the fraction of bound transcription factor is a function of ligand concentration and the dissociation constant. Importantly, if the ligand concentration is (approximately) constant over time, the fraction of bound transcription factor is (approximately) constant over time.

To computationally simulate this model we start with total counts of free transcription factor and ligand, completely dissociated from one another. We then form one molecule of the ligand-TF complex at a time and evaluate how close the ratio of $L^n \cdot T/T^*$ is to the actual K_d. We select the values of L, T and T^* that minimize the absolute difference between K_d and $L^n \cdot T/T^*$ (see Algorithm 9).

Two-component systems

For a transcription factor with concentration T; a cognate sensing protein with concentration S; a ligand with concentration L; subscripts f denoting a free (unbound) form of a molecule, b denoting a ligand-bound form of a molecule, and p denoting a phosphorylated form of a molecule; and ATP, ADP, H^+, and H_2O denoting concentrations of these molecules, we propose a system with the following:

Free (unbound) cognate sensing protein at equilibrium with ligand-bound cognate sensing protein, described by dissociation constant K_d:

$$L + S_f \leftrightarrow S_b \tag{C.11}$$

The autophosphorylation of a free (unbound) cognate sensing protein at a rate k_A:

$$S_f + ATP \overset{k_A}{\rightarrow} S_{fp} + ADP + H^+ \tag{C.12}$$

The autophosphorylation of a ligand-bound cognate sensing protein at a rate k_B:

$$S_b + ATP \overset{k_B}{\rightarrow} S_{bp} + ADP + H^+ \tag{C.13}$$

The phosphorylation of a transcription factor by its free, phosphorylated cognate sensing protein at a rate k_C:

$$S_{fp} + T \overset{k_C}{\rightarrow} S_f + T_p \tag{C.14}$$

The phosphorylation of a transcription factor by its bound, phosphorylated cognate sensing protein at a rate k_D:

$$S_{bp} + T \overset{k_D}{\rightarrow} S_b + T_p \tag{C.15}$$

The auto-phosphatase activity of a transcription factor at a rate k_E:

$$T_p + H_2O \overset{k_E}{\rightarrow} T + P_i \tag{C.16}$$

By assuming mass-action kinetics, we can represent this system mathematically using ordinary differential equations. Ligand binding is simulated in a fashion identical to the one-component systems and the rest of the sub-model is simulated using a numerical ODE integrator (see Algorithm 10).

Zero-component systems

We assume all transcription factors of this class will bind to available promoter sites.

Modeling the modulation of RNA polymerase recruitment

After modeling transcription factor activation, we need to model the probability that the transcription factor is bound to DNA, P_T, and, when the transcription factor is DNA-bound, its effect on RNA polymerase recruitment to the promoter site, Δr. Recalling the notation used in the *Transcription* section, we want to modulate the j^{th} entry in the v_{synth} vector of RNA polymerase initiation probabilities such that:

$$v_{\text{synth},j} = \alpha_j + \sum_i P_{T,i} \Delta r_{ij} \tag{C.17}$$

where α_j represents basal recruitment of RNA polymerase and the second term is dependent on transcription factor activity: the probability that the i^{th} transcription factor is DNA-bound is $P_{T,i}$, and the recruitment effect of the i^{th} transcription factor on the j^{th} gene is Δr_{ij}. The α and Δr values are computed prior to simulation based on gene expression values from conditions that modulate transcription factor activity. Values for P_T are calculated as described in Table C.5.

Transcription factor type	Promoter-bound probability
Zero-component system	$P_T = 1$ if TF is present, 0 otherwise
One-component system	$P_T = (T^*)/(T^* + T)$
Two-component system	$P_T = (T_p)/(T_p + T)$

Table C.5: Formulas used to compute the probability that a transcription factor is promoter-bound. T^* is the active form of a one-component system transcription factor, while T_p is the phosphorylated form of a two-component system transcription factor, and T is the inactive or unphosphorylated form of a transcription factor.

Input : c_m counts of molecules where $m = 1$ **to** $n_{molecules}$

Input : S matrix describing reaction stoichiometries where $S[i, j]$ describes the coefficient for the i^{th} molecule in the j^{th} reaction

Input : K_d^r dissociation constant where $r = 1$ **to** $n_{reactions}$

for *each ligand-binding reaction j* **do**

1. Dissociate all complexes in c formed by reaction j into constituent molecules

while *True* **do**

1. Form complex described by $S[:, j]$

if $\left| \frac{c_{reactant1}^{S[reactant1,j]} \cdot c_{reactant2}^{S[reactant2,j]} \cdot \ldots \cdot c_{reactantm}^{S[reactantm,j]}}{c_{complex}} - K_d \right|$ *has reached a minimum (i.e., the ratio of reactants to products is as close as possible to the dissociation constant)*

then

1. Set reactant and product values in c to these levels

2. Break out of while loop

end

end

end

Result: Ligands are bound to or unbound from their binding partners in a fashion that maintains equilibrium.

Algorithm 9: Algorithm for equilibrium binding

Input : Δt length of current time step
Input : c_m counts of molecules where $m = 1$ **to** $n_{molecules}$
Input : k_A rate of phosphorylation of free histidine kinase
Input : k_B rate of phosphorylation of ligand-bound histidine kinase
Input : k_C rate of phosphotransfer from phosphorylated free histidine kinase to response regulator
Input : k_D rate of phosphotransfer from phosphorylated ligand-bound histidine kinase to response regulator
Input : k_E rate of dephosphorylation of phosphorylated response regulator
Input : `solveToNextTimeStep()` function that solves two-component system ordinary differential equations to the next time step and returns the change in molecule counts (Δc_m)
1. Solve the ordinary differential equations describing phosphotransfer reactions to perform reactions to the next time step (Δt) using c_m, k_A, k_B, k_C, k_D and k_E.
Δc_m = `solveToNextTimeStep(`c_m, k_A, k_B, k_C, k_D, k_E, Δt`)`
2. Update molecule counts.
$c_m = c_m + \Delta c_m$
Result: Phosphate groups are transferred from histidine kinases to response regulators and back in response to counts of ligand stimulants.

Algorithm 10: Algorithm for two-component systems

Environment dependence: Constitutive and induced transcriptional frequency groups

In the main text, we categorize genes as expressed (1) at least once per cell cycle, (2) less than once per cell cycle, or (3) never expressed over a 32-generation simulation. We also note that for some genes, the categorization is environment-dependent, while for others, the categorization is constitutive (see Figure 5H). Careful readers will note that the model is only optimized for three conditions; this text explains how we could estimate gene expression for many conditions using the transcriptional regulatory data we compiled.

First, we observed that the synthesis probabilities (P_{synth}) and transcriptional frequencies (T_{freq}) of the genes that are expressed less than once per cell cycle are linearly correlated (Figure 5B). Outside of this linear range, the frequency of observing at least one transcript is bounded at zero (for very low synthesis probabilities) or one (for high synthesis probabilities). Using a piece-wise function interpolated from this linear behavior, we can infer the transcriptional frequency for a given gene

with a particular transcript synthesis probability:

$$T_{freq} = \begin{cases} 0 & P_{synth} \leq P_{synth,min} \\ \alpha + \beta \cdot P_{synth,g} & P_{synth,min} < P_{synth} < P_{synth,max} \\ 1 & P_{synth} \geq P_{synth,max} \end{cases} \tag{C.18}$$

Note that α and β are fit with the baseline data in Figure 5B.

The estimation of transcriptional frequencies (P^*_{synth}) under different environmental conditions and genetic perturbations was done as follows:

$$P^*_{synth,g} = FC_g \cdot P_{synth,g} \tag{C.19}$$

where $P_{synth,g}$ is the synthesis probability of a particular gene in our baseline condition, and FC_g is the gene expression fold change measured in the perturbed condition relative to the baseline.

The estimated synthesis probability for a given gene could then be used to compute the estimated transcription frequency (using Equation C.18) corresponding to a different environmental condition.

Using the above model allowed us to (1) use our collection of gene expression profile shifts (described in Associated data) to determine the maximal and minimal fold changes observed for each gene across all conditions, (2) calculate the fold change in expression with respect to our baseline condition (note that we restricted the experiments considered for this analysis to be exclusively those which were directly comparable to our baseline condition), (3) evaluate Equation C.19 to obtain $P^*_{synth,g}$, (4) and subsequently $T^*_{freq,g}$ using Equation C.18, and finally (5) compare the newly-computed $T^*_{freq,g}$ to the original value.

If the new $T^*_{freq,g}$ value is in a different category than the old value, then the gene's expression categorization is environment-dependent. As an example, if a gene is transcribed less than once per cell cycle in our baseline condition, but in another environment a $T^*_{freq,g}$ is calculated that is one or higher, this means that the gene can be induced to the category of more highly-expressed genes (i.e., moving from the blue group in Figure 5B to the red group).

While we recognize that this model might not be accurate for inferring transcriptional frequencies under conditions in which there are significant variations in cell physiology, the results of this analysis investigates how far genes can be induced into or out of the transcription frequency groups identified in Figure 5B. Genes that remained within the same transcription frequency group after being subjected to the linear transformations described are considered to be expressed constitutively

at their minimal media frequency (represented by the red, blue, and yellow regions in Figure 5H). On the other hand, genes that were able to be induced into different transcription frequency groups are represented by the purple, orange, and green regions (according to which frequency groups they were able to sample as a result of this analysis). A final note: the total sum of genes shown in Figure 5H (4416 counts) is greater than the total number of genes represented in Figure 5B (4353 genes). This is because 63 genes originated from the blue transcription frequency zone and were found to explore both the always transcribed and never transcribed groups under different environment conditions, and thus are represented twice: once in the purple region and once in the green region.

Associated files

`wcEcoli` Path	File	Type
`wcEcoli/models/ecoli/processes`	`equilibrium.py`	process
`wcEcoli/models/ecoli/processes`	`tf_binding.py`	process
`wcEcoli/models/ecoli/processes`	`two_component_system.py`	process
`wcEcoli/reconstruction/ecoli/dataclasses/process`	`equilibrium.py`	data
`wcEcoli/reconstruction/ecoli/dataclasses/process`	`transcription_regulation.py`	data
`wcEcoli/reconstruction/ecoli/dataclasses/process`	`two_component_system.py`	data

Table C.6: Table of files for transcription regulation.

Difference from *M. genitalium* model

The most significant difference from the *M. genitalium* model is the enhanced coverage of the regulatory network; 438 regulatory interactions are described by 22 transcription factors that regulate 355 genes. Accordingly, regulation is represented by three different classes of transcription regulators: two-component system, one-component system and zero-component systems. While the phosphotransfer reactions of two-component signaling pathways are modeled in `TwoComponentSystems`, one-component systems (which bind directly to the transcription factor) and zero-component systems (whose presence or absence determines activity) are modeled by the `EquilibriumBinding` and `TranscriptionFactorBinding` Processes.

Associated data

Parameter	Symbol	Units	Value
Ligand::TF dissociation constant	$k_d = k_r/k_f$	μM	[2e-15, 5e3]
Free HK phosphorylation rate	k_A	μM/s	[1e-4, 5e2]
Ligand::HK phosphorylation rate	k_B	μM/s	1.7e5
Phosphotransfer rate from free HK-P to TF	k_C	μM/s	1e8
Phosphotransfer rate from ligand::HK-P to TF	k_D	μM/s	1e8
Dephosphorylation rate of TF-P	k_E	μM/s	1e-2
DNA::TF dissociation constant	K_d	pM	[2e-4, 1.1e5]
Promoter sites	n	targets/chromosome	[1, 108]
Fold-change gene expression	FC	$log_2(a.u.)$	[-10.48, 9.73]
Gene expression profile shifts	-	shifts	294*

Table C.7: Table of parameters for equilibrium binding, two-component systems, and transcription factor binding Processes. HK: histidine kinase, TF: transcription factor, HK-P: phosphorylated histidine kinase, TF-P: phosphorylated transcription factor. *We found 144 pairs of comparable shifts (see Figure 2C). All parameters reference Supplemental Materials.

RNA degradation

Model Implementation

The RNA decay sub-model encodes a molecular simulation of RNA degradation and occurs via two steps that represent RNase-mediated mechanisms. It is implemented in the `RNAdegradation` process (detailed in Algorithm 11).

Endo-nucleolytic Cleavage First, the total counts of RNA degraded during a time step are computed as a fraction of the total capacity for endo-cleavage. Then, the total amount of RNA degraded is divided into different species (mRNA, tRNA, and rRNA) using known endoRNase::RNA affinities. Finally, non-functional RNA fragments are represented as an additional pseudo-metabolite in the `BulkMolecules` state.

Exo-nucleolytic Digestion The exoRNase enzymatic capacity is used to determine the fraction of RNA fragments that can be digested and converted to individual nucleotides that can be recycled by the `Metabolism` process.

Difference from *M. genitalium* model

The *E. coli* model provides a more detailed, mechanistic representation in the `RNADegradation` process compared to the *M. genitalium* model. Unlike the previous model, the gene functionality of endoRNase and exoRNase is mechanistically integrated to evaluate: (1) rates of RNA degradation due to endo-nucleolytic cleavage, and (2) rates of nucleotides digested by exoRNases.

Associated files

`wcEcoli` Path	File	Type
`wcEcoli/models/ecoli/processes`	`rna_degradation.py`	process
`wcEcoli/reconstruction/ecoli/dataclasses/process`	`rna_decay.py`	data

Table C.8: Table of files for RNA degradation.

Associated data

Parameter	Symbol	Units	Value	Reference
EndoRNase catalytic rate	$K_{cat,endo}$	non-functional RNA counts/s	0.10	Supp. Materials
ExoRNase catalytic rate	$K_{cat,exo}$	nt digested/s	50	Supp. Materials
mRNA half-lives[(1)]	τ_{mRNA}	min	[1.30, 31.40]	[20]
tRNA, rRNA half-lives	τ_{tRNA}, τ_{rRNA}	hour	48	[20]
Michaelis constant[(2)]	K_m	RNA counts	-	See Table C.8
RNAse mechanism of action	-	endo-/exo-RNAse	-	Supp. Materials
EndoRNase specificity[(3)]	-	(mRNA, tRNA, rRNA)/RNase	Boolean	Supp. Materials

Table C.9: Table of parameters for RNA degradation process.
[(1)]Non-measured mRNA half-lives were estimated as the average mRNA half-life (5.75 min).
[(2)]Michaelis constants were calculated by fitting the `RNADegradation` model to be equal to the first-order `RNADegradation` model, as follows:
$K_{cat,endo} \cdot c_{endo} \frac{c_{RNA,i}/K_{m,i}}{\sum_j c_{RNA,j}/K_{m,j}} = \frac{ln(2)}{\tau_{RNA,i}} \cdot c_{RNA,i}$

[(3)]Types of RNA that can be targeted by a given RNase.

Input : $K_{m,i}$ Michaelis constants of each mRNA transcript where $i = 1$ **to** n_{RNA}

Input : $K_{cat,endo}$, $K_{cat,exo}$ catalytic rate of endoRNase and exoRNase

Input : c_{endo}, c_{exo} count of endoRNase and exoRNase

Input : $c_{frag,i}$ count of non-functional RNA fragments where $i = 1$ **to** 4 for AMP, CMP, GMP, UMP

Input : c_{mRNA}, c_{tRNA}, c_{rRNA} count of each mRNA, tRNA and rRNA

Input : c_{molec} count of small molecules where $molec \rightarrow$ [H_2O, PPI, Proton, NMPs]

Input : `multinomial()` function that draws samples from a multinomial distribution

Input : `countNTs()` function that returns counts of AMP, CMP, GMP, and UMP for a given non-functional RNA fragment

Input : `lengthFragments()` function that returns the total number of bases of all RNA fragments

`/* Endo-nucleolytic cleveage */`

1. Calculate fraction of active endoRNases (f_i) that target each RNA where $i = 1$ **to** n_{gene}

$$f_i = \frac{\frac{c_{RNA,i}}{K_{m,i}}}{1+\sum \frac{c_{RNA}}{K_m}}$$

2. Calculate total counts of RNAs degraded (R)

$R_{mRNA} = \sum K_{cat,endo} \cdot c_{endo,mRNA} \cdot f_i$ where $i = 1$ **to** n_{mRNAs}

$R_{tRNA} = \sum K_{cat,endo} \cdot c_{endo,tRNA} \cdot f_i$ where $i = 1$ **to** n_{tRNAs}

$R_{rRNA} = \sum K_{cat,endo} \cdot c_{endo,rRNA} \cdot f_i$ where $i = 1$ **to** n_{rRNAs}

where $c_{endo,j}$: number of endoRNases targeting specific species considering endoRNase specificities, $j = 1$ **to** [mRNA, tRNA, rRNA]

3. Sample multinomial distribution D times weighted by endoRNase::RNA affinities to determine which RNAs are converted into non-functional RNAs (d_i)

$d_i =$ `multinomial(`R, $\frac{f_i}{\sum f}$`)`

4. Increase number of RNA fragments. Decrease RNA counts and amount of water required for RNA hydrolysis by endoRNases ($c_{H_2O,endo}$)

$c_{frag} = c_{frag} +$ `countNTs(`d_i`)`

$c_{RNA} = c_{RNA} - d_{RNA}$

$c_{H_2O} = c_{H_2O} - c_{H_2O,endo}$

$c_{PPi} = c_{PPi} + D$

Continued...

Algorithm 11: Algorithm for RNA degradation: endo-cleavage for transcripts, and exo-nucleolytic digestion

Continued from above...

`/* Exo-nucleolytic digestion */`

5. Compute exoRNase capacity (E)

$$E = K_{cat,exo} \cdot c_{exo}$$

if $E > \sum c_{frag,i}$ **then**

Update NMPs, water and proton counts

$$c_{NMP} = c_{NMP} + c_{frag}$$

$$c_{H_2O} = c_{H_2O} - \texttt{lengthFragments}(c_{frag})$$

$$c_{proton} = c_{proton} + \texttt{lengthFragments}(c_{frag})$$

Set counts of RNA fragments equal to zero ($c_{frag,i} = 0$)

else

Sample multinomial distribution c_{frag} with equal probability to determine which fragments are exo-digested ($c_{fragDig}$) and recycled

$$c_{fragDig,i} = \texttt{multinomial}(E, \tfrac{c_{frag,i}}{\sum c_{frag}})$$

Update NMPs, water, proton counts, and RNA fragments

$$c_{NMP} = c_{NMP} + c_{fragDig}$$

$$c_{H_2O} = c_{H_2O} - \texttt{lengthFragments}(c_{fragDig})$$

$$c_{proton} = c_{proton} + \texttt{lengthFragments}(c_{fragDig})$$

$$c_{frag} = c_{frag} - c_{fragDig}$$

end

Result: RNAs are selected and degraded by endoRNases, and non-functional RNA fragments are digested through exoRNases. During the process water is consumed, and amino acids are released.

Translation

Model implementation

Translation is the process by which the coding sequences of mRNA transcripts are translated by 70S ribosomes into polypeptides that then fold into proteins. This process accounts for more than two thirds of an *E. coli* cell's ATP consumption during rapid growth [150] and the majority of macromolecular mass accumulation. In the *E. coli* model translation occurs through the action of two processes in the model: `PolypeptideInitiation` and `PolypeptideElongation`.

`PolypeptideInitiation` models the complementation of 30S and 50S ribosomal subunits into 70S ribosomes on mRNA transcripts. Full 70S ribosomes are formed on mRNA transcripts by sampling a multinomial distribution with probabilistic weights calculated from the abundance of mRNA transcripts, and each transcript's translational efficiency (See Algorithm 12). Translational efficiencies were calculated from ribosomal profiling data [120].

`PolypeptideElongation` models the polymerization of amino acids into polypeptides by ribosomes using an mRNA transcript as a template, and the termination of elongation once a ribosome has reached the end of an mRNA transcript. This process is implemented assuming that tRNA charging by synthetases, ternary complex formation (GTP : EF-Tu : charged-tRNA), and ternary complex diffusion to elongating ribosomes are not rate limiting for polypeptide polymerization. Given this assumption this process directly polymerizes amino acids based on the codon sequence of the mRNA transcript. Polymerization occurs across all ribosomes simultaneously and resources are allocated to maximize the progress of all ribosomes up to the limit of the expected ribosome elongation rate in a medium, available amino acids, and available transcripts (see Algorithm 13).

Input : t_i translational efficiency of each mRNA transcript where $i = 1$ **to** n_{gene}

Input : $c_{mRNA,i}$ count of each mRNA transcript where $i = 1$ **to** n_{gene}

Input : c_{30S} count of free 30S ribosomal subunit

Input : c_{50S} count of free 50S ribosome subunit

Input : `multinomial()` function that draws samples from a multinomial distribution

1. Calculate probability (p_i) of forming a ribosome on each mRNA transcript weighted by the count and translational efficiency of the transcript.

$$p_i = \frac{c_{mRNA,i} \cdot t_i}{\sum_{i=1}^{n_{gene}} c_{mRNA,i} \cdot t_i}$$

2. Calculate maximal number of ribosomes that could be formed.

$$r_{max} = \texttt{min}(c_{30S}, c_{50S})$$

3 Sample multinomial distribution r_{max} times weighted by p_i to determine which transcripts receive a ribosome and initiate ($n_{init,i}$).

$$n_{init,i} = \texttt{multinomial}(r_{max}, p_i)$$

4 Assign $n_{init,i}$ ribosomes to mRNA transcript i. Decrement 30S and 50S counts.

$$c_{30S} = c_{30S} - \sum_{i=1}^{n_{gene}} n_{init,i}$$

$$c_{50S} = c_{50S} - \sum_{i=1}^{n_{gene}} n_{init,i}$$

Result: 70S ribosomes are formed from free 30S and 50S subunits on mRNA transcripts scaled by the count of the mRNA transcript and the transcript's translational efficiency.

Algorithm 12: Algorithm for ribosome initiation on mRNA transcripts

Input : $e_{expected}$ expected elongation rate of ribosome ($e_{expected} < e_{max}$)

Input : p_i position of ribosome on mRNA transcript $i = 1$ **to** $n_{ribosome}$

Input : δt length of current time step

Input : c_{GTP} counts of GTP molecules

Input : L_j length of each mRNA $j = 1$ **to** n_{gene} for each coding gene.

`/* Elongate polypeptides up to limits of sequence, amino acids, or energy */`

for *each ribosome i on mRNA transcript j* **do**

1. Based on ribosome position p_i on mRNA transcript and expected elongation rate $e_{expected}$ determine stop condition position (t_i) for ribosome assuming no amino acid limitation. Stop condition is either maximal elongation rate scaled by the time step or the full length of sequence (i.e. the ribosome will terminate in this time step).

$$t_i = \texttt{min}(p_i + e_{expected} \cdot \delta t,\ L_j)$$

2. Derive sequence between ribosome position (p_i) and stop condition (t_i).

3. Based on derived sequence calculate the number of amino acids required to polymerize sequence $c^{req}_{aa,i}$ and number of GTP molecules required c^{req}_{GTP}.

4. Elongate up to limits:

if $\texttt{all}(c^{req}_{aa,k} < c_{aa,k})$ ***and*** $c^{req}_{GTP} < c_{GTP}$ **then**

Update the position of each ribosome to stop position

$$p_i = t_i$$

else

4a. Attempt to elongate all polypeptide fragments.

4b. Update position of each ribosome to maximal position given the limitation of $c_{aa,k}$ and c_{GTP}.

end

5. Update counts of $c_{aa,k}$ and c_{GTP} to reflect polymerization usage.

end

`/* Terminate ribosomes that have reached the end of their mRNA transcript */`

for *each ribosome i on transcript j* **do**

if $p_i == L_j$ **then**

1. Increment count of protein that corresponds to elongating polypeptide that has terminated.

2. Dissociate ribosome and increment 30S and 50S counts.

end

end

Result: Each ribosome is elongated up to the limit of available mRNA sequence, expected elongation rate, amino acid, or GTP limitation. Ribosomes that reach the end of their transcripts are terminated and released.

Algorithm 13: Algorithm for peptide chain elongation and termination

Associated files

wcEcoli Path	File	Type
wcEcoli/models/ecoli/processes	polypeptide_initiation.py	process
wcEcoli/models/ecoli/processes	polypeptide_elongation.py	process
wcEcoli/reconstruction/ecoli/dataclasses/process	translation.py	data

Table C.10: Table of files for translation.

Difference from *M. genitalium* model

The PolypeptideInitiation process is implemented similarly in the *M. genitalium* with a few key differences. As the model of *E. coli* is not yet gene complete, the checks for initiation factors are not present. A major advance over the *M. genitalium* model is that the probability of ribosome initiation on a transcript is now proportional to the product of the mRNA count and its translational efficiency. In the *M. genitalium* model translational efficiency was not taken into account.

The PolypeptideElongation algorithm is implemented similarly to the *M. genitalium* model but again because the *E. coli* model is not yet gene complete, elongation factors are not accounted for. Additionally, tRNAs and their synthetases are not accounted for explicitly. Instead, the model directly polymerizes amino acids. This avoids computational issues with the simulation time step, tRNA pool size, and tRNA over expression that were present in the *M. genitalium* model. There is no implementation of ribosome stalling or tmRNAs. The polymerization resource allocation algorithm is the same as in *M. genitalium*.

Associated data

Parameter	Symbol	Units	Value	Reference
Translational efficiency[(1)]	t_i	RIB/mRNA	[0, 5.11]	[120]
Ribosome elongation rate	e	aa/s	18 (growth-dependent)	[33]
Protein counts (validation data)	$c_{protein}$	protein counts	[0, 250000]	[158]

Table C.11: Table of parameters for translation process.
[(1)]Non-measured translational efficiencies were estimated by the average translational efficiency (1.11 RIB/mRNA).

Protein degradation

Model Implementation

The `ProteinDegradation` process accounts for the degradation of protein monomers. It uses the N-end rule [179] to assign degradation rates for each protein, and selects proteins to be degraded as a Poisson process.

Input : $t_{1/2,i}$ Protein half-lives for each monomer where $i = 1$ **to** $n_{protein}$
Input : L_i length of each protein monomer where $i = 1$ **to** $n_{protein}$
Input : $c_{aa,i,j}$ count of each amino acid present in the protein monomer where $i = 1$ **to** $n_{protein}$ and $j = 1$ **to** 21 for each amino acid
Input : $c_{protein,i}$ the number of each protein present in the cell
1. Determine how many proteins to degrade based on the degradation rates and counts of each protein.

$$n_{protein,i} = \texttt{poisson}(\frac{ln(2)}{t_{1/2,i}} \cdot c_{protein,i} \cdot \Delta t)$$

2. Determine the number of hydrolysis reactions (n_{rxns}) that will need to occur.

$$n_{rxns} = \sum_i (L_i - 1) \cdot n_{proteins,i}$$

3. Determine the number of amino acids ($n_{aa,j}$) that will be released.

$$n_{aa,j} = \sum_i c_{aa,i,j} \cdot n_{proteins,i}$$

4. Degrade selected proteins, release amino acids from those proteins back into the cell, and consume H_2O that was required for hydrolysis reactions.
Result: Proteins are selected and degraded. During the process water is consumed, and amino acids are released.

Algorithm 14: Algorithm for Protein Degradation

Difference from *M. genitalium* model.

The *E. coli* model is not yet gene complete, hence this process does not take into account the activities of specific proteases and does not specifically target prematurely aborted polypeptides. In addition, protein unfolding and refolding by chaperones is not accounted for by this process.

Associated files

`wcEcoli` Path	File	Type
`wcEcoli/models/ecoli/processes`	`protein_degradation.py`	process
`wcEcoli/reconstruction/ecoli/dataclasses/process`	`translation.py`	data

Table C.12: Table of files for protein degradation.

Associated data

Parameter	Symbol	Units	Value	Reference
Protein half-lives	$t_{1/2}$	min	[2, 600]	[179]

Table C.13: Table of parameters for protein degradation process.

Complexation

Model implementation

This process models the formation of all macromolecular complexes except for 70S ribosome formation, which is handled by `Translation`. Macromolecular complexation is done by identifying complexation reactions that are possible (which are reactions that have sufficient counts of all subcomponents), performing one randomly chosen possible reaction, and re-identifying all possible complexation reactions. This process assumes that macromolecular complexes form spontaneously, and that complexation reactions are fast and complete within the time step of the simulation.

Input : c_i counts of molecules where $i = 1$ **to** $n_{molecules}$
Input : S matrix describing reaction stoichiometries where $S_{i,j}$ describes the coefficient for the i^{th} molecule in the j^{th} reaction
Input : `getPossibleReactions` function that takes c_i and S and returns all reactions that are possible
Input : `chooseRandomReaction` function that takes all possible reactions and returns one randomly chosen reaction
while *possible reactions remaining* **do**
  1. Get all possible reactions (r)
    $r =$ `getPossibleReactions`(S, c_i)
  2. Choose a random possible reaction (r_{choice}) to perform
    $r_{choice} =$ `chooseRandomReaction`(r)
  3. Perform r_{choice} by incrementing product counts and decrementing reactant counts
end
Result: Macromolecule complexes are formed from their subunits.

Algorithm 15: Algorithm for macromolecular complexation

Associated files

wcEcoli Path	File	Type
wcEcoli/models/ecoli/processes	complexation.py	process
wcEcoli/reconstruction/ecoli/dataclasses/process	complexation.py	data

Table C.14: Table of files for complexation.

Associated data

Stoichiometric coefficients that define 1,023 reactions to form protein complexes from EcoCyc [104].

Difference from *M. genitalium* model

This sub-model is implemented very similarly to the *M. genitalium* model of complexation. In the *M. genitalium* simulations, however, the selection of a complexation reaction was weighted by a multinomial distribution parameterized by substrate availability rather than a uniform distribution. We found that the choice of distribution had no major effect on behavior of the process. Additionally, the *M. genitalium* simulations describe 201 macromolecular complexes, whereas over 5 times as many are implemented in the *E. coli* model.

C.3.2 Metabolism

Metabolism

Model implementation

Flux balance analysis (FBA) is a common way to model large-scale metabolic network behavior with a low parameter requirement. However, traditional implementations of FBA are inappropriate for whole-cell modeling due to the dynamic nature of whole-cell simulation and fixed nature of the classic FBA objective function.

To alleviate this we use an alternative objective function that involves a multi-objective minimization for homeostatic metabolite composition and reaction kinetics that extends previous work by Birch *et al.*[21]. The effect of this multi-objective function is twofold: (1) to maintain cellular concentrations of small molecule metabolites and (2) to enforce constraints on metabolic fluxes calculated from Michaelis-Menten kinetics based on metabolite concentrations and curated kinetic parameters. A weighting factor is used to balance the contribution from the two objectives.

We used the metabolic network reconstruction from Karp *et al.* [96] because it was well-connected

to the rest of EcoCyc's resources and data which we relied on. This network reconstruction was based on the Orth model [139]. Different nutrient conditions (minimal M9, +amino acids, -oxygen, etc.) can be specified by changing bounds on metabolite import reactions, and shifts between these nutrient conditions can be programmatically varied.

Homeostatic objective The homeostatic objective attempts to maintain small molecule metabolite concentrations at a constant value. For example, if during a time-step the net effect of other `Process` execution transforms ATP to ADP, the concentration of ATP will be lower and ADP higher. The homeostatic objective ensures that the metabolic network will attempt to increase the ATP concentration and decrease the ADP concentration using chemical transformations available in the network.

A total of 140 metabolite set-point concentrations are specified in the objective ($C_{o,i}$ in Equation C.20). The homeostatic objective minimizes the deviation from these measured concentrations and can be specified as:

$$\text{minimize} \sum_i \left| 1 - \frac{C_i}{C_{o,i}} \right| \tag{C.20}$$

where C_i is the concentration of metabolite i and $C_{o,i}$ is the measured set-point concentration for metabolite i. Cytoplasmic concentrations were chosen based on data from Bennett *et al.* [19], and other components of biomass have set-point concentrations specified based on the overall composition of the cell (lipids, metal ions, etc.) [193] and can be dependent on the media environment of the simulation.

Kinetics objective The *E. coli* model simulates both metabolic enzyme expression via transcription and translation and dynamically maintains 140 metabolite concentrations. This enables the use of Michaelis-Menten kinetic equality constraints on metabolic fluxes using Equation C.21:

$$v_{o,j} = k_{cat} \cdot E \cdot \left(\frac{C_1}{C_1 + K_{m,1}} \right) \cdot \left(\frac{C_2}{C_2 + K_{m,2}} \right) \ldots \left(\frac{C_n}{C_n + K_{m,n}} \right) \tag{C.21}$$

where $v_{o,j}$ is the kinetic target for the flux through reaction j that has n substrates, k_{cat} is the catalytic turnover rate for enzyme E, $K_{m,n}$ is the saturation constant for substrate n, E is the concentration of metabolic enzyme, and C_n is the concentration of substrate n in reaction j.

Kinetics data was reviewed from over 12,000 papers identified from BRENDA [160]. We filtered out papers that did not have a k_{cat}, which did not use a lab strain, or which did not involve enzymes in our metabolic network. The result was roughly 1200 papers which we manually curated due

to our and others' observation that about 20% of the values in the BRENDA database are copied incorrectly from their primary source papers [16]. From this set, 181 constraints with a K_m and k_{cat} and 219 constraints with only a k_{cat} are used to constrain a total of 340 reactions (with some reactions having multiple constriants). Although some additional constraints were identified, they are not currently being used in the model. In particular, constraints were found for tRNA charging (18 reactions) but not used since tRNA charging is not explicitly included in the model. Additionally, contraints for two reactions involved in the Citric Acid cycle (succinate dehydrogenase and fumurate reductductase) were identified that, when enabled, caused a much higher glucose uptake rate than observed without kinetic constraints and higher than what has been experimentally measured. Based on this, we excluded these constraints from the model.

In cases where the enzyme parameters were recorded at non-physiological temperatures, we used the following scaling factor to adjust the k_{cat}:

$$2^{\frac{37-T}{10}} \tag{C.22}$$

where T is the reported temperature (in °C) for the experimental conditions—this increases the kinetic rate by a factor of 2 for every 10°C away from 37°C.

Similar to the homeostatic objective, the kinetics objective minimizes the deviation from a kinetic target that is calculated at each time step based on the enzyme and metabolite concentrations. Formally:

$$\text{minimize} \sum_j \left| 1 - \frac{v_j}{v_{o,j}} \right| \tag{C.23}$$

where v_j is the flux through reaction j and $v_{o,j}$ is the target flux for reaction j calculated from Equation C.21.

Difference from *M. genitalium* model

We have made a number of improvements to the metabolic sub-model compared to the *M. genitalium* implementation. In the *M. genitalium* simulations, metabolites were produced in a fixed ratio at every time step regardless of the behavior of the rest of the simulated cell—this could lead to pooling or depletion of metabolites. Furthermore, if one metabolite could not be produced, none of the metabolites could be produced. Our homeostatic objective fixes both of these shortcomings. In addition, we have quantitative data and a method to softly constrain 340 reactions.

Associated data

Parameter	Symbol	Units	Value	Reference
Metabolic network	S	-	Stoichiometric coefficients	[96]
Metabolic target fluxes	v_o	$\mu M/s$	[0, 87000]	See Table C.3.2
Metabolic fluxes (validation)	v_v	$\mu M/s$	[82, 1500]	[183]
Enzyme turnover number	k_{cat}	$1/s$	[0.00063, 38000]	Supp. Materials
Enzyme Michaelis constant	K_m	μM	[0.035, 550000]	Supp. Materials
Metabolite target concentration	C_o	μM	[0.063, 97000]	[19]

Table C.15: Table of parameters for metabolism process.

Input : C_i concentration for metabolite i

Input : $C_{o,i}$ concentration target for metabolite i

Input : $k_{cat,j}$ turnover number for enzyme j

Input : $K_{m,i,j}$ Michaelis constant for metabolite i for enzyme j

Input : E_j concentration for enzyme j

Input : S stoichiometric matrix for all reactions

1. Set physical constraints on reaction fluxes

For all reactions: $v_{min,j} = -\inf, v_{max,j} = +\inf$

For thermodynamically irreversible reactions: $v_{min,j} = 0$

If required enzyme not present: $v_{min,j} = v_{max,j} = 0$

2. Calculate kinetic target ($v_{o,j}$) for each reaction j based on the enzymes j and metabolites i associated with each reaction

$$v_{o,j} = k_{cat,j} \cdot E_j \cdot \prod_i \left(\frac{C_i}{K_{m,i,j} + C_i}\right)$$

3. Solve linear optimization problem

$$\begin{aligned} &\text{minimize} \sum_i \left|1 - \frac{C_i}{C_{o,i}}\right| + \lambda \sum_j \left|1 - \frac{v_j}{v_{o,j}}\right| \\ &\text{subject to} \quad S \cdot v = 0 \\ &\qquad\qquad\quad v_j \geq v_{min,j} \\ &\qquad\qquad\quad v_j \leq v_{max,j} \end{aligned}$$

4. Update concentrations of metabolites based on the solution to the linear optimization problem

Result: Metabolites are taken up from the environment and converted into other metabolites for use in other processes

Algorithm 16: Algorithm for Metabolism

Associated files

wcEcoli Path	File	Type
wcEcoli/models/ecoli/processes	metabolism.py	process
wcEcoli/reconstruction/ecoli/dataclasses/process	metabolism.py	data

Table C.16: Table of files for metabolism.

Energy requirements of cell maintenance

As was the case in our *M. genitalium* simulations, and in many flux balance analysis models, not all of the energy consumed by metabolic pathways, macromolecular polymerization, or other growth and non-growth associated processes is accounted for explicitly in our *E. coli* model. This is primarily due to a lack of experimental data and/or knowledge accounting for its usage. Similar to the *M. genitalium* model, we have incorporated reactions in the metabolic model with two parameters, Growth Associated Maintenance (GAM) and Non-Growth Associated Maintenance (NGAM), which modulate energy consumption associated with growth and cell maintenance.

Associated data

Parameter	Symbol	Units	Value	Reference
Growth associated maintenance	GAM	mmol/g	59.81	[71]
Non-growth associated maintenance	NGAM	mmol/g/h	8.39	[71]

Table C.17: Table of parameters for energy requirements of cell maintenance.

C.3.3 Balanced growth

Chromosome replication

Model implementation

Chromosome replication occurs through three steps that are implemented in the `ChromosomeFormation` and `ChromosomeElongation` processes. First, a round of replication is initiated at a fixed cell mass per origin of replication and generally occurs once per cell cycle (see Algorithm 17). Second, replication forks are elongated up to the maximal expected elongation rate, dNTP resource limitations, and template strand sequence (see Algorithm 18). Finally, replication forks terminate once they reach the end of their template strand and the chromosome immediately decatinates forming two

separate chromosome molecules (see Algorithm 19).

Input : m_{cell} cell mass
Input : $m_{critical}$ critical initiation mass
Input : n_{origin} number of origins of replication
Input : $n_{fork,f}$ number of replication forks on forward strand
Input : $n_{fork,r}$ number of replication forks on reverse strand
Input : $n_{chromosome}$ number of chromosome molecules
Input : C length of C period
Input : D length of D period
if $\frac{m_{cell}}{n_{origin}} > m_{critical}$ **then**
  if $n_{origin} > 1$ **then**
    $n_{origin} = n_{origin} + \frac{n_{fork,f}+n_{fork,r}}{2} \cdot n_{chromosome}$
  else
    $n_{origin} = n_{origin} + n_{chromosome}$
  end
  $n_{fork,f} = n_{fork,f} + n_{fork,f} \cdot n_{chromosome}$
  $n_{fork,r} = n_{fork,r} + n_{fork,r} \cdot n_{chromosome}$
end
Result: When cell mass is larger than critical initiation mass m_c another round of replication is initiated with correct number of replication forks

Algorithm 17: Algorithm for DNA replication initiation

Input : e maximal elongation rate of replication fork

Input : p_i position of forks on chromosome where $i = 1$ **to** n_{fork}

Input : δt length of current time step

Input : $c_{dNTP,j}$ counts of dNTP where $j = 1$ **to** 4 for dCTP, dGTP, dATP, dTTP

Input : L_k total length of each strand of chromosome from origin to terminus where $k = 1$ **to** 4 for forward/complement and reverse/complement.

for *each replication fork i on sequence k* **do**

1. Based on replication fork position p_i and maximal elongation rate e determine stop condition (s_i) for replication fork assuming no dNTP limitation.

$$s_i = \texttt{min}(p_i + e \cdot \delta t,\ L_k)$$

Stop condition is either maximal elongation rate scaled by the time step or the full length of sequence (i.e. the fork will terminate in this time step).

2. Derive sequence between replication fork position (p_i) and stop condition (s_i).

3. Based on derived sequence calculate the number of dNTPs required to polymerize sequence $c^{req}_{dNTP,i}$.

4. Elongate up to limits:

if $\texttt{all}(c^{req}_{dNTP,i} < c_{dNTP,j})$ **then**

Update the position of each replication fork to stop position

$$p_i = s_i$$

else

Attempt to equally elongate each replication fork update position of each fork to maximal position given the limitation of $c_{dNTP,j}$.

end

5. Update counts of $c_{dNTP,j}$ to reflect polymerization usage.

end

Result: Each replication fork is elongated up to the limit of available sequence, elongation rate, or dNTP limitation

Algorithm 18: Algorithm for DNA replication elongation

Input : p_i position of forks on chromosome where $i = 1$ **to** n_{fork}
Input : L_k total length of each strand of chromosome from origin to terminus where $k = 1$ **to** 4 for forward/complement and reverse/complement
Input : d_{queue} a double ended queue data structure that stores time(s) cell division should be triggered
Input : D D-period of cell cycle (time between completion of chromosome replication and cell division)
Input : t Current simulation time
for *each replication fork i on strand k* **do**
  if $p_i == L_k$ **then**
    1. Delete replication fork
    2. Divide remaining replication forks and origins of replication appropriately across the two new chromosome molecules
    3. Calculate time cell should trigger division based on current time of chromosome termination and push onto queue data structure
      d_{queue}.`push`$(t + D)$
  end
end
Result: Replication forks that have terminated are removed. A new chromosome molecule is created separating all remaining replication forks. Timer for D-period is started.

Algorithm 19: Algorithm for DNA replication termination

Associated files

`wcEcoli` Path	File	Type
`wcEcoli/models/ecoli/processes`	`chromosome_formation.py`	process
`wcEcoli/models/ecoli/processes`	`chromosome_elongation.py`	process
`wcEcoli/reconstruction/ecoli/dataclasses/process`	`replication.py`	data

Table C.18: Table of files for chromosome replication.

Difference from *M. genitalium* model

The physiology modeled is significantly different from what was implemented in the *M. genitalium* model. Initiation of DNA replication in *E. coli* no longer uses a DnaA based mechanistic model but instead uses a phenomenological model based on a constant mass per origin of replication triggering DNA replication initiation. The action of topoisomerases are not explicitly modeled. Replication forks no longer take into account all of the enzymes in the replisome but are point objects that traverse the chromosome sequence. Some differences exist because the *E. coli* model is not yet a

gene complete model. More importantly, certain changes enabled significant modeling advances in the *E. coli* model. These include modeling the DNA replication cycle over multiple growth rates, cell sizes, and conditions using a single unified framework, and enabling multiple rounds of replication to proceed simultaneously over multiple generations. Both advances were critical to the findings in this publication.

Associated data

Parameter	Symbol	Units	Value	Reference
Chromosome sequence	-	-	-	[22]
Replication fork elongation rate	e	nt/s	967	[32]
Mass per origin at DNA replication initiation	$m_{critical}$	origin/fgDW	[600,975]	Semi-quantitative fit [60]
C period	C	min	40	[137]
D period	D	min	20	[137]

Table C.19: Table of parameters for chromosome replication process.

Cell division

Model implementation

Cell division is modeled in the `ChromosomeElongation` process and `CellDivision` listener in the *E. coli* model. A Helmstetter-Cooper type model of chromosome replication initiation is coupled to cell division, inspired by work from Wallden *et al.* [190]. Chromosome replication initiation occurs at a fixed mass per origin of replication. Each initiation event is coupled to a cell division event after a constant period of time consisting of one round of chromosome replication and cytokinesis. Importantly, this constant period of time can span multiple cell division events.

Cell division itself is modeled as a binomial process where each daughter cell has an equal probability of inheriting the contents of the mother cell. The exception to this is if two chromosomes are present before cell division—each daughter is guaranteed to get one.

Algorithms 19 and 20 provide implementation details.

Input : d_{queue} a double ended queue data structure that stores time(s) cell division should be triggered

Input : c_i counts of all molecules in simulation at cell division where $i = 1$ **to** $n_{species}$

Input : p binomial partition coefficient

Input : n_{chrom} number of chromosome molecules

Input : `rand()` returns a random number from a uniform distribution between 0 and 1

Input : `randint()` returns a random integer either 0 or 1

if $t > d_{queue}$.`peek()` **then**

1. Trigger division and remove division time.

d_{queue}.`pop()`

2. Divide bulk contents of cell binomially. Number partitioned into daughter one is stored in $n_{daughter,1}$ and to daughter two in $n_{daughter,2}$.

for $i = 1$ **to** $n_{species}$ **do**

$n_{daughter,1} = 0$

for $j = 1$ **to** c_i **do**

if `rand()` $> p$ **then**

$n_{daughter,1} = n_{daughter,1} + 1$

end

end

$n_{daughter,2} = c_i - n_{daughter,1}$

end

3. Divide chromosome in binary manner. All replication forks and origins of replication associated with a chromosome molecule are partitioned as well. Number of chromosome molecules partitioned into daughter one is stored in $n_{chrom,daughter,1}$ and to daughter two in $n_{chrom,daughter,2}$.

if `mod(`n_{chrom}`,2)` **then**

$n_{chrom,daughter,1} = \frac{n_{chrom}}{2}$

else

$n_{chrom,daughter,1}$ =`floor(`$\frac{n_{chrom}}{2}$`)` + `randint()`

end

$n_{chrom,daughter,2} = n_{chrom} - n_{chrom,daughter,1}$

end

Result: Cell division is triggered at C+D time after DNA replication initiation. Contents of mother cell is divided between two daughter cells conserving mass.

Algorithm 20: Algorithm for cell division

Due to the interaction of the algorithms for Chromosome Replication and Cell Division, we occasionally obtain outliers when examining initial and added cell masses (clipped distributions are shown in Figure 4F in the main text). Figures C.3 and C.4 show histograms of these distributions (the

same data for Figure 4F, no clipping) and where the outliers fall relative to the x- and y-limits in Figure 4F—the dashed lines in each figure demarcate the x- and y- limits of the plots in Figure 4F. For example, in Figure C.3, the "Glucose minimal + 20 amino acids" histogram has dashed lines at 0.6 and 1.2, and in Figure C.4 has dashed lines at 0.45 and 1.5—these are the x- and y- limits, respectively, for that same plot in Figure 4F. We see that the outliers constitute at most 1-2% of the data.

Associated files

`wcEcoli` Path	File	Type
`wcEcoli/models/ecoli/processes`	`replication_elongation.py`	process
`wcEcoli/models/ecoli/listeners`	`cell_division.py`	listener

Table C.20: Table of files for transcription regulation.

Difference from *M. genitalium* model

The *E. coli* model is not yet a gene complete model and many of the mechanistic details of cell division are not implemented as they were in the *M. genitalium* model. In *E. coli*, cytokinesis, septation, and chromosome segregation are all not modeled explicitly. However, cell division in our *E. coli* model is consistent with growth at multiple growth rates, which was not the case in *M. genitalium*.

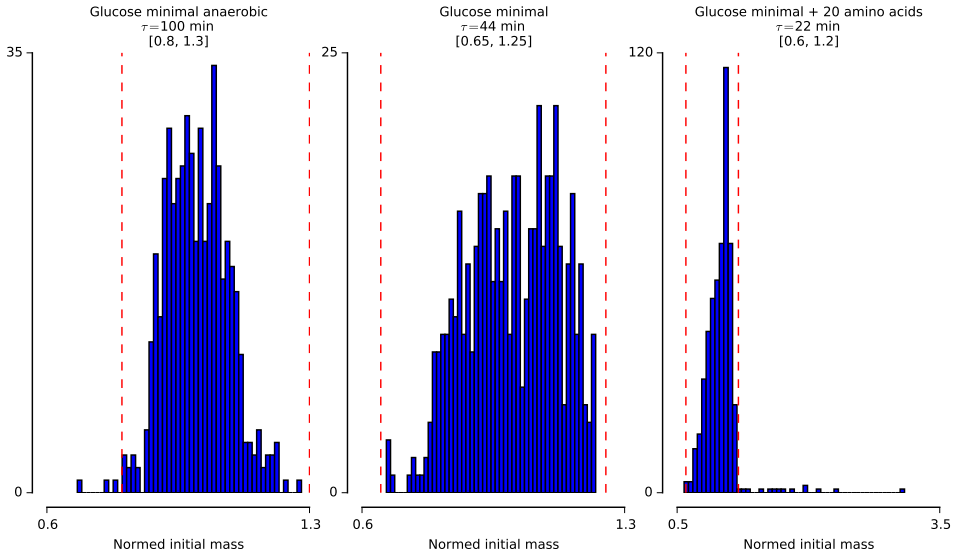


Figure C.3: Distributions of initial cell mass in 3 conditions. Red dashed lines indicate where the x-axis limits for Figure 4F (in the main text) fall relative to these distributions.

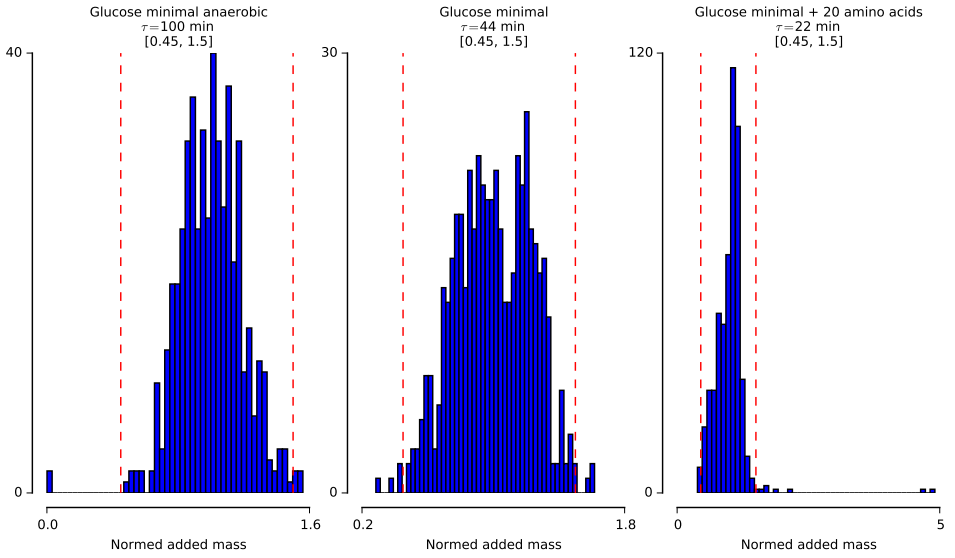


Figure C.4: Distributions of added cell mass in 3 conditions. Red dashed lines indicate where the y-axis limits for Figure 4F (in the main text) fall relative to these distributions.

C.4 Experimental procedures

Below are the materials and methods used to culture *E. coli* K-12 MG1655, perform RNA sequencing, and measure protein half-lives.

C.4.1 RNA sequencing

RNA sequencing was performed to characterize mRNA expression under a variety of environmental conditions. The contributions of this dataset is highlighted most in our evaluation of the Central Dogma model in the main text, but is integral to the function of all processes related to mRNAs.

Materials

Buffers

Stock Solution	**Volume (mL)**
alanine (free)	5
arginine (free)	65
asparagine (free)	40
aspartic acid (free)	5
cysteine (free)	50
glutamic acid (K monohydrate)	5
glutamine (free)	25
glycine (free)	5
histidine (free)	5
isoleucine (free)	5
leucine (free)	10
lysine (HCl)	80
methionine (free)	5
phenylalanine (free)	5
proline (free)	40
serine (free)	5
threonine (free)	40
tryptophan (free)	5
tyrosine (free)	125
valine (free)	5
VA Vitamin Solution	50
H_2O	350
TOTAL	1000

Table C.21: 5x Amino acids, Mix solutions together. Preparation of stock solutions is described in Table C.22
Filter sterilize with 0.2 uM filter. Aliquot and Freeze at -20° C.

Name	**FW**	**Weight (g)**	**Vol (mL)**
alanine (free)	89.09	1.78	25
arginine (free)	174.2	6.98	100
asparagine (free)	132.1	0.66	100
aspartic acid (free)	133.1	1.33	25
cysteine (free)	121.16	0.06	50
glutamic acid (K monohydrate)	203.23	3.05	25
glutamine (free)	146.20	1.8	100
glycine (free)	75.07	1.5	25
histidine (free)	155.15	0.78	25
isoleucine (free)	131.1	0.65	25
leucine (free)	131.2	0.66	100
lysine (HCl)	182.7	7.32	100
methionine (free)	149.2	0.75	25
phenylalanine (free)	165.2	0.83	100
proline (free)	115.10	1.15	25
serine (free)	105.1	8.4	100
threonine (free)	119.1	119	25
tryptophan (free)	204.2	0.26	25
tyrosine (free)	181.2	0.225	125
valine (free)	117.2	0.88	25

Table C.22: Amino Acid Stocks, Make each amino acid separately. Store at -20°C.

	Formula	**5x Salts**
Sodium Phosphate Dibasic (g)	Na_2HPO_4	8.475
Potassium Phosphate Monobasic (g)	KH_2PO_4	3.75
Sodium Chloride (g)	NaCl	0.625
Ammonium Chloride (g)	NH_4Cl	1.25

Table C.23: Salts, bring up to 250 mL each. Autoclave, aliquot into sterile 50 mL tubes.

	M9 Minimal Glucose	M9 Minimal Glucose + AAs
Volume (mL)	100	100
5x Salts (mL)	20	20
5x Amino acids (mL)	0	20
1 M $MgSO_4$ (μL)	200	200
1 M $CaCl_2$ (μL)	10	10
20% glucose (mL)	2	2
H_2O (mL)	77.8	57.8

Table C.24: Media Formulas

Methods

Cell Growth

Grow cells in requisite media overnight in 37°C incubator. In the morning, inoculate fresh cultures to an OD_{600} 0.02 in 10 mL of media contained in a 125 mL flask. Grow cultures on shaker in 37°C warm room.

RNA Extraction

Harvest cells at OD 0.4. Take 2 mL of cell culture and extract using Quiagen RNAeasy Protect Bacterial Mini Kit (Qiagen #74524) with RNAprotect Bacteria Reagent (# 76506) and RNase-Free DNase Set (# 79254) according to manufacturer's instructions. Performed extraction using lysozyme from ThermoScientific #90082. Measured RNA concentration using NanoDrop.

rRNA Cleanup

Remove rRNA from sample using RiboZero rRNA Removal Kit (Epicentre # MRZGN126), according to manufacturer's instructions.

RNA Quality

RNA quality was assessed using an Agilent 2100 BioAnalyzer (# G2938C), according to manufacturer's instructions by the SFGF facility at Stanford University.

cDNA prep and Sequencing

Library prep was performed by the SFGF facility at Stanford University according to the TruSeq Stranded Total RNA Sample Preparation Guide. Paired-end sequencing with read lengths of 75 bp was performed by the SFGF facility at Stanford University on an Illumina NextSeq 500. Approximately 20 million reads were obtained per sample.

Data availability

Sequencing data is available at GEO with accession number GSE85472.

Analysis

bbmap 34.33 [2] was used to pre-process sequencing data to trim reads, remove reads for common contaminants, and remove reads that map to non-coding RNA. RSEM 1.2.19 [118] was used for downstream processing and calculation of gene expression.

C.4.2 Protein half-life measurement

We tested the Central Dogma model under steady-state conditions, under which we expected that the rate of protein synthesis should equal the rate of decay. As shown in Figure 2G, this proved

to largely be the case in our simulations, with most of the production rates within an order of magnitude of the decay rate. We wondered whether some of the outliers in our comparison might be due to a more nuanced or specific value for the protein half-life. For that, we experimentally determined the half-lives of four well-characterized proteins (RpoH, RcsA, HelD, and PssA), and six protein outliers (DcuR, BioD, Rph, CarA, Pnp, and GshA). As illustrated in Figure 2H, we found that in all cases, the half-life predicted by our model was a better predictor of the data than the N-end rule.

Methods

Cell Growth

His-tagged gene plasmids (from the ASKA library without GFP) were transformed into MG1655. Duplicates of bacterial cultures were grown overnight at 37°C in M9 minimal media with 0.4% glucose and 20 μg/ml chloramphenicol for plasmid selection. In the morning, bacterial cultures were diluted to OD 0.03, and incubated at 37°C until they reached OD 0.3. At this point, bacterial cultures were diluted 1:2 in minimal media supplemented with IPTG (0.1mM) to induce protein over-expression. Cultures were grown on a shaker in a 37°C warm room for the requisite time of induction.

Time-course sampling

After the requisite time of IPTG induction (see Table C.25), a 9 mL sample was taken to measure the time 0 protein level. Then, 10 μg/mL tetracycline (the 10 mg/mL stock was made in 95% ethanol) was added to the rest of the culture to inhibit protein synthesis. Culture was then returned to 37°C. 9 mL samples were taken at indicated time points (10 min on ice followed by centrifugation for 10 min at 4000g, 4°C) to measure protein levels (see C.25. At each time point, culture OD was measured.

Cell lysates

To lyse cells we used BugBuster Master mix (Millipore, #71456-3) supplemented with Halt protease/phosphatase inhibitor cocktail (Thermo Scientific #78444), following the manufacturer's instructions.

Protein quantification

For protein quantification we used Pierce BCA protein assay kit (Thermo Scientific #23225), following the manufacturer's instructions.

Protein detection

For western blot detection of proteins we ran samples on a Simon machine (Protein Simple) using an

anti-His tag antibody (Novus Biologicals #NB100-64768) for protein detection and an anti-RNAP β antibody (BioLegend #663006) for capillary normalization (loading control).

Analysis

The measured amount of His-tagged protein (N_p) was normalized by the amount of β subunit RNAP loaded as a protein control ($RNAP$), i.e. $N = N_p/RNAP$; and then log-transformed. Linear regression was used to determine the first-order decay rate constant (k_d) as follows:

$$N = N_0 \exp(-k_d t),\ \log(N) = \log(N_0) - k_d t \tag{C.24}$$

Then, half-lives (see Figure C.5 and Figure C.6) were estimated by: $t_{1/2} = \frac{\log(2)}{k_d}$.

Gene	**ASKA Id**	$t_{1/2}$	**Ind.**[1] **(h)**	**Protein**[2] **(μg)**	**Ab. dilution**[3]	**Time points**[4]
RcsA	JW1935	25 min (c*)	2	10	1:100, 1:200	**0, 2, 5, 10, 30** min
RpoH	JW3426	25 min (c*)	2	10	1:100, 1:200	**0, 5, 10, 20**, 40 min
PssA	JW2569	10 h (c**)	2	1	1:100, 1:50	**0, 7, 18** h
HelD	JW0945	10 h (c**)	2	4	1:100, 1:100	**0, 1**, 3, 4, 7, 23 h
CarA	JW0030	2 min***	2	2	1:100, 1:100	**0, 0.17, 1, 4, 18** h
GshA	JW2663	2 min***	1	1	1:100, 1:100	**0, 0.17, 1, 4, 18** h
Pnp	JW5851	2 min***	1	4	1:100, 1:100	**0, 18, 24, 48** h
DcuR	JW4085	10 h	2	2	1:100, 1:100	**0, 0.5, 2.5**, 5.5, 18 h
BioD	JW0761	10 h	2	4	1:100, 1:100	**0, 0.67, 2**, 4, 18 h
Rph	JW3618	10 h	2	4	1:100, 1:100	**0, 0.5, 2.5**, 5.5, 18 h

Table C.25: Parameters used for half-life measurements. *Short half-lives that are well characterized in the literature. **Control proteins with minimal model discrepancies (half-life with highest confidence = 10 h). ***Short half-lives due to N-end rule, which dictates low stability if protein N-terminal is leucine. [1]IPTG induction used to over-express protein levels. [2]Amount of protein loaded for protein detection. [3]Antibody dilution in anti-His, anti-RNAP. [4]Note that time points highlighted in bold were used to calculate decay constant.

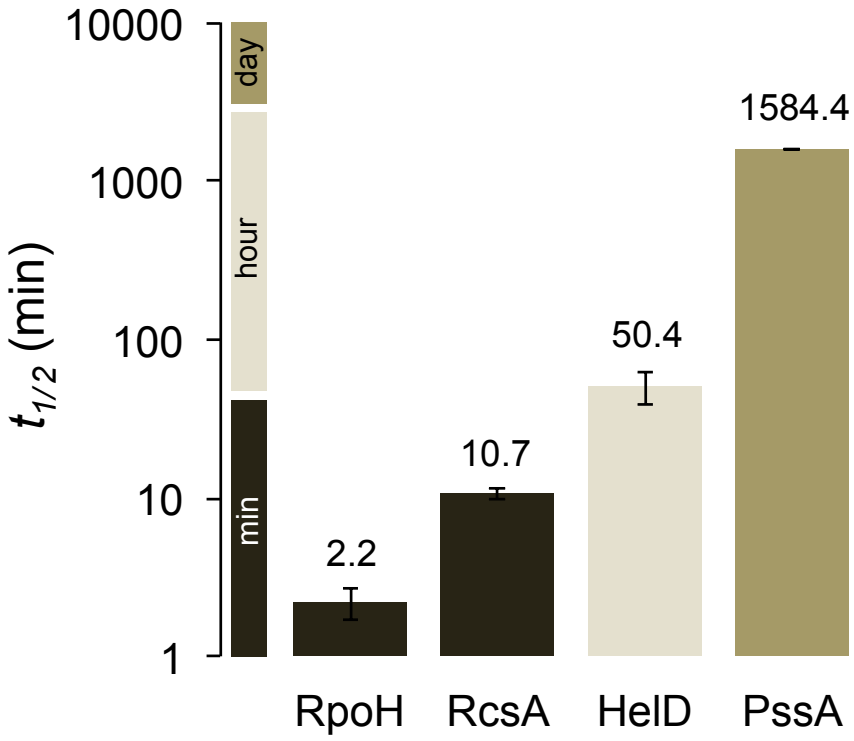


Figure C.5: **Protein half-lives measured for well characterized proteins (RpoH, RcsA), and control proteins with minimal model discrepancies (half-life with highest confidence = 10h).**

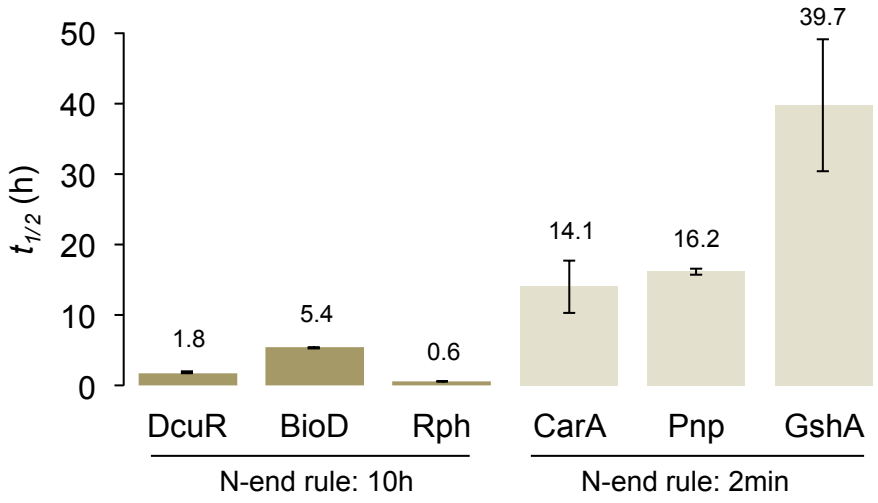


Figure C.6: **Protein half-lives measured.**

Bibliography

[1]

[2] Bbtools. `http://jgi.doe.gov/data-and-tools/bbtools/`. Accessed: 2017-02-16.

[3] Business Intelligence and Analytics Software.

[4] CUDA Toolkit Documentation.

[5] MPI Documents.

[6] Tableau Technology — Tableau Software.

[7] The MIT License (MIT) — Open Source Initiative.

[8] Whole-cell parameter estimation DREAM challenge - syn1876068.

[9] Announcement: Reducing our irreproducibility. *Nature*, 496(7446):398–398, April 2013.

[10] Kyle R Allison, Mark P Brynildsen, and James J Collins. Heterogeneous bacterial persisters and engineering approaches to eliminate them. *Current opinion in microbiology*, 14(5):593–598, 2011.

[11] Stephanie M Amato and Mark P Brynildsen. Persister heterogeneity arising from a single metabolic stress. *Current Biology*, 25(16):2090–2098, 2015.

[12] Steve Ashby, Pete Beckman, Jackie Chen, Phil Colella, Bill Collins, Dona Crawford, Jack Dongarra, Doug Kothe, Rusty Lusk, and Paul Messina. The opportunities and challenges of exascale computing—summary report of the advanced scientific computing advisory committee (ASCAC) subcommittee. US Department of Energy Office of Science. *US Department of Energy Office of Science*, 2010.

[13] A Bachmair. In vivo half-life of a protein is a function of its. *Science*, 3018930(179):234, 1986.

[14] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

[15] Nathalie Q Balaban, Jack Merrin, Remy Chait, Lukasz Kowalik, and Stanislas Leibler. Bacterial persistence as a phenotypic switch. *Science*, 305(5690):1622–1625, 2004.

[16] Arren Bar-Even, Elad Noor, Yonatan Savir, Wolfram Liebermeister, Dan Davidi, Dan S Tawfik, and Ron Milo. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–4410, 2011.

[17] T Barrett, S E Wilhite, P Ledoux, C Evangelista, I F Kim, M Tomashevsky, K A Marshall, K H Phillippy, P M Sherman, M Holko, A Yefanov, H Lee, N Zhang, C L Robertson, N Serova, S Davis, and A Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(D1):D991–D995, December 2012.

[18] Bryson D Bennett, Elizabeth H Kimball, Melissa Gao, Robin Osterhout, Stephen J Van Dien, and Joshua D Rabinowitz. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nature chemical biology*, 5(8):593–599, August 2009.

[19] Bryson D Bennett, Elizabeth H Kimball, Melissa Gao, Robin Osterhout, Stephen J Van Dien, and Joshua D Rabinowitz. Absolute metabolite concentrations and implied enzyme active site occupancy in escherichia coli. *Nature chemical biology*, 5(8):593–599, 2009.

[20] Jonathan A Bernstein, Arkady B Khodursky, Pei-Hsun Lin, Sue Lin-Chao, and Stanley N Cohen. Global analysis of mrna decay and abundance in escherichia coli at single-gene resolution using two-color fluorescent dna microarrays. *Proceedings of the National Academy of Sciences*, 99(15):9697–9702, 2002.

[21] Elsa W Birch, Madeleine Udell, and Markus W Covert. Incorporation of flexible objectives and time-linked simulation with flux balance analysis. *Journal of Theoretical Biology*, 345:12–21, March 2014.

[22] F R Blattner, G Plunkett, C A Bloch, N T Perna, V Burland, M Riley, J Collado-Vides, J D Glasner, C K Rode, G F Mayhew, J Gregor, N W Davis, H A Kirkpatrick, M A Goeden, D J Rose, B Mau, and Y Shao. The complete genome sequence of Escherichia coli K-12. *Science (New York, N.Y.)*, 277(5331):1453–1462, September 1997.

[23] Richard Bonneau, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteinn Thorsson. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*, 7(5):R36, 2006.

[24] Jason G Bragg and Sallie W Chisholm. Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PloS one*, 3(10):e3550, 2008.

[25] Abraham L Brass, Derek M Dykxhoorn, Yair Benita, Nan Yan, Alan Engelman, Ramnik J Xavier, Judy Lieberman, and Stephen J Elledge. Identification of host proteins required for

HIV infection through a functional genomic screen. *Science (New York, N.Y.)*, 319(5865):921–926, February 2008.

[26] Abraham L Brass, I-Chueh Huang, Yair Benita, Sinu P John, Manoj N Krishnan, Eric M Feeley, Bethany J Ryan, Jessica L Weyer, Louise van der Weyden, Erol Fikrig, David J Adams, Ramnik J Xavier, Michael Farzan, and Stephen J Elledge. The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell*, 139(7):1243–1254, December 2009.

[27] H. Bremer. Variation of generation times in Escherichia coli populations: its cause and implications. *Journal of general microbiology*, 128(12):2865–2876, December 1982.

[28] H. Bremer and G Churchward. An examination of the Cooper-Helmstetter theory of DNA replication in bacteria and its underlying assumptions. *Journal of Theoretical Biology*, 69(4):645–654, December 1977.

[29] H. Bremer and P Dennis. Feedback control of ribosome function in Escherichia coli. *Biochimie*, 90(3):493–499, March 2008.

[30] H. Bremer, P Dennis, and M Ehrenberg. Free RNA polymerase and modeling global transcription in Escherichia coli. *Biochimie*, 85(6):597–609, June 2003.

[31] H. Bremer and D Yuan. Chain growth rate of messenger RNA in Escherichia coli infected with bacteriophage T4. *Journal of Molecular Biology*, 34(3):527–540, June 1968.

[32] Hans Bremer and Patrick Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella: cellular and molecular biology,*, 2:1553–1569, 1996.

[33] Hans Bremer and Patrick P Dennis. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 3(1), 2008.

[34] Hans Bremer and Mns Ehrenberg. Guanosine tetraphosphate as a global regulator of bacterial rna synthesis: a model involving rna polymerase pausing and queuing. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1262(1):15 – 36, 1995.

[35] R Calendar. *The Bacteriophages*. Oxford University Press, USA, 2006.

[36] Manuel Campos, Ivan V Surovtsev, Setsu Kato, Ahmad Paintdakhi, Bruno Beltran, Sarah E Ebmeier, and Christine Jacobs-Wagner. A constant size extension drives bacterial cell size homeostasis. *Cell*, 159(6):1433–1446, December 2014.

[37] Javier Carrera, Raissa Estrela, Jing Luo, Navneet Rai, Athanasios Tsoukalas, and Ilias Tagkopoulos. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of escherichia coli. *Molecular systems biology*, 10(7):735, 2014.

[38] Javier Carrera, Guillermo Rodrigo, and Alfonso Jaramillo. Model-based redesign of global transcription regulation. *Nucleic acids research*, 37(5):e38, April 2009.

[39] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A Fulcher, Timothy A Holland, Ingrid M Keseler, Anamika Kothari, Aya Kubo, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 42(D1):D459–D471, 2014.

[40] Ron Caspi, Tomer Altman, Joseph M Dale, Kate Dreher, Carol A Fulcher, Fred Gilham, Pallavi Kaipa, Athikkattuvalasu S Karthikeyan, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Suzanne Paley, Liviu Popescu, Anuradha Pujar, Alexander G Shearer, Peifen Zhang, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 38(Database issue):D473–9, January 2010.

[41] Qian Chai, Bhupender Singh, Kristin Peisker, Nicole Metzendorf, Xueliang Ge, Santanu Dasgupta, and Suparna Sanyal. Organization of ribosomes and nucleoids in Escherichia coli cells during growth and in quiescence. *Journal of Biological Chemistry*, 289(16):11342–11352, April 2014.

[42] M Chamberlin and J Ring. Characterization of T7-specific ribonucleic acid polymerase. 1. General properties of the enzymatic reaction and the template specificity of the enzyme. *The Journal of biological chemistry*, 248(6):2235–2244, March 1973.

[43] Huiyi Chen, Katsuyuki Shiroguchi, Hao Ge, and Xiaoliang Sunney Xie. Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Molecular Systems Biology*, 11(1):781, 2015.

[44] Byung-Kwan Cho, Karsten Zengler, Yu Qiu, Young Seoub Park, Eric M Knight, Christian L Barrett, Yuan Gao, and Bernhard O Palsson. The transcription unit architecture of the Escherichia coli genome. *Nature Biotechnology*, 27(11):1043–1049, November 2009.

[45] Michele Clamp, Ben Fry, Mike Kamal, Xiaohui Xie, James Cuff, Michael F Lin, Manolis Kellis, Kerstin Lindblad-Toh, and Eric S Lander. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences*, 104(49):19428–19433, 2007.

[46] Martha R J Clokie, Jinyu Shan, Shaun Bailey, Ying Jia, Henry M Krisch, Stephen West, and Nicholas H Mann. Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environmental microbiology*, 8(5):827–835, May 2006.

[47] John A Cole and Zaida Luthey-Schulten. Whole cell modeling: from single cells to colonies. *Israel journal of chemistry*, 54(8-9):1219–1229, 2014.

[48] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[49] S Cooper. Cell division and DNA replication following a shift to a richer medium. *Journal of Molecular Biology*, 43(1):1–11, July 1969.

[50] S Cooper and C E Helmstetter. Chromosome replication and the division cycle of Escherichia coli B/r. *Journal of Molecular Biology*, 31(3):519–540, February 1968.

[51] Markus W Covert, Eric M Knight, Jennifer L Reed, Markus J Herrgard, and Bernhard O Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, May 2004.

[52] Markus W Covert and Bernhard O Palsson. Transcriptional regulation in constraints-based metabolic models of Escherichia coli. *The Journal of biological chemistry*, 277(31):28058–28064, August 2002.

[53] Markus W Covert, Nan Xiao, Tiffany J. Chen, and Jonathan R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics*, 24(18):2044–2050, September 2008.

[54] Markus W Covert, Nan Xiao, Tiffany J. Chen, and Jonathan R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics*, 24(18):2044–2050, 2008.

[55] FHC Crick. Project K:" The Complete Solution of E. Coli". *Perspectives in Biology and Medicine*, 1973.

[56] D G Dalbow and R Young. Synthesis time of beta-galactosidase in Escherichia coli B/r as a function of growth rate. *The Biochemical journal*, 150(1):13–20, July 1975.

[57] Eric H Davidson, Jonathan P Rast, Paola Oliveri, Andrew Ransick, Cristina Calestani, Chiou-Hwa Yuh, Takuya Minokawa, Gabriele Amore, Veronica Hinman, Cesar Arenas-Mena, Ochan Otim, C Titus Brown, Carolina B Livi, Pei Yun Lee, Roger Revilla, Alistair G Rust, Zheng jun Pan, Maria J Schilstra, Peter J C Clarke, Maria I Arnone, Lee Rowen, R Andrew Cameron, David R McClay, Leroy Hood, and Hamid Bolouri. A genomic regulatory network for development. *Science (New York, N.Y.)*, 295(5560):1669–1678, March 2002.

[58] PATRICK P DENNIS, Måns Ehrenberg, and Hans Bremer. Control of rRNA synthesis in Escherichia coli: a systems biology approach. *Microbiology and molecular biology reviews : MMBR*, 68(4):639–668, December 2004.

[59] M M Domach and M L Shuler. A finite representation model for an asynchronous culture of E. coli. *Biotechnology and Bioengineering*, 26(8):877–884, August 1984.

[60] W D Donachie. Relationship between cell size and time of initiation of DNA replication. *The American Journal of Gastroenterology*, 219(5158):1077–1079, September 1968.

[61] M J Donlin and K A Johnson. Mutants affecting nucleotide recognition by T7 DNA polymerase. *Biochemistry*, 33(49):14908–14917, December 1994.

[62] Ron O Dror, Robert M Dirks, J P Grossman, Huafeng Xu, and David E Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual Review of Biophysics*, 41:429–452, 2012.

[63] Arthur Eddington. *New Pathways in Science: Messenger Lectures (1934)*. Cambridge University Press, 2012.

[64] M Ehrenberg and C G Kurland. Costs of accuracy determined by a maximal growth rate constraint. *Quarterly reviews of biophysics*, 17(1):45–82, February 1984.

[65] Måns Ehrenberg, Hans Bremer, and PATRICK P DENNIS. Medium-dependent control of the bacterial growth rate. *Biochimie*, 95(4):643–658, April 2013.

[66] D. Endy. *Development and Application of a Genetically-structured Simulation for Bacteriophage*. Number v. 7. Dartmouth College, 1997.

[67] D. Endy, D. Kong, and J. Yin. Intracellular kinetics of a growing virus: A genetically structured simulation for bacteriophage T 7. *Biotechnology and Bioengineering*, 55(2):375–389, 1997.

[68] D. Endy, L You, J. Yin, and I J Molineux. Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10):5375–5380, May 2000.

[69] M J Engler and C.C. Richardson. Bacteriophage T7 DNA replication. Synthesis of lagging strands in a reconstituted system using purified proteins. *The Journal of biological chemistry*, 258(18):11197–11205, September 1983.

[70] Adam M Feist, Christopher S Henry, Jennifer L Reed, Markus Krummenacker, Andrew R Joyce, Peter D Karp, Linda J Broadbelt, Vassily Hatzimanikatis, and Bernhard O Palsson. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, 3, June 2007.

[71] Adam M Feist, Christopher S Henry, Jennifer L Reed, Markus Krummenacker, Andrew R Joyce, Peter D Karp, Linda J Broadbelt, Vassily Hatzimanikatis, and Bernhard Ø Palsson. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular systems biology*, 3(1):121, 2007.

[72] Jenny Finkel, Shipra Dingare, Christopher D Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. Exploring the boundaries: gene and protein identification in biomedical text. *BMC bioinformatics*, 6 Suppl 1:S5, 2005.

[73] Terrence S Furey. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews. Genetics*, 13(12):840–852, December 2012.

[74] Sean R Gallagher. One-dimensional SDS gel electrophoresis of proteins. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 10:Unit 10.2A, August 2006.

[75] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, et al. Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic acids research*, 44(D1):D133–D143, 2016.

[76] L R García and I J Molineux. Rate of translocation of bacteriophage T7 DNA across the membranes of Escherichia coli. *Journal of Bacteriology*, 177(14):4066–4076, July 1995.

[77] Marc Güell, Vera van Noort, Eva Yus, Wei-Hua Chen, Justine Leigh-Bell, Konstantinos Michalodimitrakis, Takuji Yamada, Manimozhiyan Arumugam, Tobias Doerks, Sebastian Kühner, Michaela Rode, Mikita Suyama, Sabine Schmidt, Anne-Claude Gavin, Peer Bork, and Luis Serrano. Transcriptome complexity in a genome-reduced bacterium. *Science (New York, N.Y.)*, 326(5957):1268–1271, November 2009.

[78] Jeremy Gunawardena. Silicon dreams of cells into symbols. *Nature Biotechnology*, 30(9):838–840, September 2012.

[79] H Hadas, M Einav, I Fishov, and A Zaritsky. Bacteriophage T4 development depends on the physiology of its host Escherichia coli. *Microbiology (Reading, England)*, 143 (Pt 1):179–185, January 1997.

[80] S. Hagens and M Loessner. Bacteriophage for biocontrol of foodborne pathogens: Calculations and considerations. *Current Pharmaceutical Biotechnology*, 11(1):58–68, 2010.

[81] Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, December 1999.

[82] C E Helmstetter. DNA synthesis during the division cycle of rapidly growing Escherichia coli B/r. *Journal of Molecular Biology*, 31(3):507–518, February 1968.

[83] Jörg D Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature reviews. Genetics*, 7(3):200–210, March 2006.

[84] J E Hopper, G Ko, and E T Young. Comparative analysis of the in vivo and in vitro expression of bacteriophage T7 messenger RNAs during infection of Escherichia coli. *Journal of Molecular Biology*, 94(4):539–554, June 1975.

[85] Dann Huh and Johan Paulsson. Non-genetic heterogeneity from stochastic partitioning at cell division. *Nature Genetics*, 43(2):95–100, February 2011.

[86] Alfredo J Ibáñez, Stephan R Fagerer, Anna Mareike Schmidt, Pawel L Urban, Konstantins Jefimovs, Philipp Geiger, Reinhard Dechant, Matthias Heinemann, and Renato Zenobi. Mass spectrometry-based metabolomics of single yeast cells. *Proceedings of the National Academy of Sciences*, 110(22):8790–8794, May 2013.

[87] Alina Ionel, Javier A Velázquez-Muriel, Daniel Luque, Ana Cuervo, José R Castón, José M Valpuesta, Jaime Martín-Benito, and José L Carrascosa. Molecular rearrangements involved in the capsid shell maturation of bacteriophage T7. *Journal of Biological Chemistry*, 286(1):234–242, January 2011.

[88] Nobuyoshi Ishii, Kenji Nakahigashi, Tomoya Baba, Martin Robert, Tomoyoshi Soga, Akio Kanai, Takashi Hirasawa, Miki Naba, Kenta Hirai, Aminul Hoque, Pei Yee Ho, Yuji Kakazu, Kaori Sugawara, Saori Igarashi, Satoshi Harada, Takeshi Masuda, Naoyuki Sugiyama, Takashi Togashi, Miki Hasegawa, Yuki Takai, Katsuyuki Yugi, Kazuharu Arakawa, Nayuta Iwata, Yoshihiro Toya, Yoichi Nakayama, Takaaki Nishioka, Kazuyuki Shimizu, Hirotada Mori, and Masaru Tomita. Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science (New York, N.Y.)*, 316(5824):593–597, April 2007.

[89] Anubhav Jain, Shyue Ping Ong, Wei Chen, Bharat Medasani, Xiaohui Qu, Michael Kocher, Miriam Brafman, Guido Petretto, Gian-Marco Rignanese, Geoffroy Hautier, Daniel Gunter, and Kristin A. Persson. Fireworks: a dynamic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience*, 27(17):5037–5059, 2015. CPE-14-0307.R2.

[90] Rishi Jain and Ranjan Srivastava. Metabolic investigation of host/pathogen interaction using MS2-infected Escherichia coli. *BMC Systems Biology*, 3:121, December 2009.

[91] Magnus Johansson, Elli Bouakaz, Martin Lovmar, and Måns Ehrenberg. The kinetics of ribosomal peptidyl transfer revisited. *Molecular cell*, 30(5):589–598, June 2008.

[92] Horace Freeland Judson. The eighth day of creation. *New York*, page 550, 1979.

[93] Suckjoon Jun and Sattar Taheri-Araghi. Cell-size maintenance: universal strategy revealed. *Trends in Microbiology*, 23(1):4–6, January 2015.

[94] Tomasz Kalwarczyk, Marcin Tabaka, and Robert Holyst. Biologistics–diffusion coefficients for complete proteome of Escherichia coli. *Bioinformatics*, 28(22):2971–2978, November 2012.

[95] Alexander Karlas, Nikolaus Machuy, Yujin Shin, Klaus-Peter Pleissner, Anita Artarini, Dagmar Heuer, Daniel Becker, Hany Khalil, Lesley A Ogilvie, Simone Hess, André P Mäurer, Elke Müller, Thorsten Wolff, Thomas Rudel, and Thomas F Meyer. Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*, 463(7282):818–822, February 2010.

[96] PD Karp, D Weaver, S Paley, C Fulcher, A Kubo, A Kothari, M Krummenacker, P Subhraveti, D Weerasinghe, S Gama-Castro, et al. The ecocyc database. ecosal plus 6. doi: 10.1128/ecosalplus. Technical report, ESP-0009-2013, 2014.

[97] Jonathan R. Karr, Nolan C Phillips, and Markus W Covert. WholeCellSimDB: a hybrid relational/HDF database for whole-cell model predictions. *Database : the journal of biological databases and curation*, 2014, 2014.

[98] Jonathan R. Karr, Jayodita C Sanghvi, Derek N Macklin, Abhishek Arora, and Markus W Covert. WholeCellKB: model organism databases for comprehensive whole-cell models. *Nucleic acids research*, 41(Database issue):D787–92, January 2013.

[99] Jonathan R. Karr, Jayodita C Sanghvi, Derek N Macklin, Abhishek Arora, and Markus W Covert. WholeCellKB: model organism databases for comprehensive whole-cell models. *Nucleic acids research*, 41(Database issue):D787–92, January 2013.

[100] Jonathan R. Karr, Jayodita C Sanghvi, Derek N Macklin, Abhishek Arora, and Markus W Covert. WholeCellKB: model organism databases for comprehensive whole-cell models. *Nucleic acids research*, 41(D1):D787–D792, 2013.

[101] Jonathan R. Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Jr., Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, 150(2):389–401, July 2012.

[102] Jonathan R. Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival Jr., Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, 150(2):389–401, July 2012.

[103] Priscilla Kemp, L René Garcia, and Ian J Molineux. Changes in bacteriophage T7 virion structure at the initiation of infection. *Virology*, 340(2):307–317, September 2005.

[104] Ingrid M Keseler, Amanda Mackie, Martin Peralta-Gil, Alberto Santos-Zavaleta, Socorro Gama-Castro, César Bonavides-Martínez, Carol Fulcher, Araceli M Huerta, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Luis Muñiz-Rascado, Quang Ong, Suzanne Paley, Imke Schröder, Alexander G Shearer, Pallavi Subhraveti, Mike Travers, Deepika Weerasinghe, Verena Weiss, Julio Collado-Vides, Robert P Gunsalus, Ian Paulsen, and Peter D Karp. EcoCyc: fusing model organism databases with systems biology. *Nucleic acids research*, 41(Database issue):D605–12, January 2013.

[105] Hiroaki Kitano. Computational systems biology. *The American Journal of Gastroenterology*, 420(6912):206–210, November 2002.

[106] Stefan Klumpp, Matthew Scott, Steen Pedersen, and Terence Hwa. Molecular crowding limits translation and cell growth. *Proceedings of the National Academy of Sciences*, 110(42):16754–16759, October 2013.

[107] Renate König, Silke Stertz, Yingyao Zhou, Atsushi Inoue, H-Heinrich Hoffmann, Suchita Bhattacharyya, Judith G Alamares, Donna M Tscherne, Mila B Ortigoza, Yuhong Liang, Qinshan Gao, Shane E Andrews, Sourav Bandyopadhyay, Paul De Jesus, Buu P Tu, Lars Pache, Crystal Shih, Anthony Orth, Ghislain Bonamy, Loren Miraglia, Trey Ideker, Adolfo García-Sastre, John A T Young, Peter Palese, Megan L Shaw, and Sumit K Chanda. Human host factors required for influenza virus replication. *Nature*, 463(7282):813–817, February 2010.

[108] Renate König, Yingyao Zhou, Daniel Elleder, Tracy L Diamond, Ghislain M C Bonamy, Jeffrey T Irelan, Chih-Yuan Chiang, Buu P Tu, Paul D De Jesus, Caroline E Lilley, Shannon Seidel, Amanda M Opaluch, Jeremy S Caldwell, Matthew D Weitzman, Kelli L Kuhen, Sourav Bandyopadhyay, Trey Ideker, Anthony P Orth, Loren J Miraglia, Frederic D Bushman, John A Young, and Sumit K Chanda. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, 135(1):49–60, October 2008.

[109] Sebastian Kühner, Vera van Noort, Matthew J Betts, Alejandra Leo-Macias, Claire Batisse, Michaela Rode, Takuji Yamada, Tobias Maier, Samuel Bader, Pedro Beltran-Alvarez, Daniel Castaño-Diez, Wei-Hua Chen, Damien Devos, Marc Güell, Tomas Norambuena, Ines Racke, Vladimir Rybin, Alexander Schmidt, Eva Yus, Ruedi Aebersold, Richard Herrmann, Bettina

Böttcher, Achilleas S Frangakis, Robert B Russell, Luis Serrano, Peer Bork, and Anne-Claude Gavin. Proteome organization in a genome-reduced bacterium. *Science (New York, N.Y.)*, 326(5957):1235–1240, November 2009.

[110] E Kutter. In kutter e., sulakvelidze a. *Bacteriophages: Biology and applications*, 2005.

[111] Piyush Labhsetwar, John Andrew Cole, Elijah Roberts, Nathan D Price, and Zaida A Luthey-Schulten. Heterogeneity in protein expression induces metabolic variability in a modeled Escherichia coli population. *Proceedings of the National Academy of Sciences*, 110(34):14006–14011, August 2013.

[112] Ruby Lee, Jonathan R. Karr, and Markus W Covert. WholeCellViz: data visualization for whole-cell models. *BMC bioinformatics*, 14(1):253, 2013.

[113] Ruby Lee, Jonathan R Karr, and Markus W Covert. Wholecellviz: data visualization for whole-cell models. *BMC bioinformatics*, 14(1):253, 2013.

[114] Seung-Joo Lee and Charles C Richardson. Molecular basis for recognition of nucleoside triphosphate by gene 4 helicase of bacteriophage T7. *Journal of Biological Chemistry*, 285(41):31462–31471, October 2010.

[115] Sun Bok Lee and James E Bailey. Analysis of growth rate effects on productivity of recombinant Escherichia coli populations using molecular mechanism models. *Biotechnology and Bioengineering*, 79(5):550–557, September 2002.

[116] Timothy K Lee, Elissa M Denny, Jayodita C Sanghvi, Jahlionais E Gaston, Nathaniel D Maynard, Jacob J Hughey, and Markus W Covert. A noisy paracrine signal determines the cellular NF-kappaB response to lipopolysaccharide. *Science Signaling*, 2(93):ra65, 2009.

[117] J.R. Leigh and Institution of Electrical Engineers. *Control Theory*. Control Series. Institution of Electrical Engineers, 2004.

[118] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[119] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, April 2014.

[120] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, 2014.

[121] Pawel Mackiewicz, Maria Kowalczuk, Dorota Mackiewicz, Aleksandra Nowicka, Malgorzata Dudkiewicz, Agnieszka Laszkiewicz, Miroslaw R Dudek, and Stanislaw Cebrat. How many protein-coding genes are there in the saccharomyces cerevisiae genome? *Yeast*, 19(7):619–629, 2002.

[122] Derek N Macklin, Nicholas A Ruggero, and Markus W Covert. The future of whole-cell modeling. *Current opinion in biotechnology*, 28:111–115, August 2014.

[123] Lisa U Magnusson, Anne Farewell, and Thomas Nyström. ppGpp: a global regulator in Escherichia coli. *Trends in Microbiology*, 13(5):236–242, May 2005.

[124] Tahrin Mahmood and Ping-Chang Yang. Western blot: technique, theory, and trouble shooting. *North American journal of medical sciences*, 4(9):429–434, September 2012.

[125] Tobias Maier, Alexander Schmidt, Marc Güell, Sebastian Kühner, Anne-Claude Gavin, Ruedi Aebersold, and Luis Serrano. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular Systems Biology*, 7(1), 2011.

[126] MR Maurizi. Proteases and protein degradation inescherichia coli. *Experientia*, 48(2):178–201, 1992.

[127] Nathaniel D Maynard, Elsa W Birch, Jayodita C Sanghvi, Lu Chen, Miriam V Gutschow, and Markus W Covert. A Forward-Genetic Screen and Dynamic Analysis of Lambda Phage Host-Dependencies Reveals an Extensive Interaction Network and a New Anti-Viral Strategy. *PLoS Genetics*, 6(7):e1001017, July 2010.

[128] Nathaniel D Maynard, Miriam V Gutschow, Elsa W Birch, and Markus W Covert. The virus as metabolic engineer. *Biotechnology Journal*, 5(7):686–694, July 2010.

[129] Nathaniel D Maynard, Derek N Macklin, Karla Kirkegaard, and Markus W Covert. Competing pathways control host resistance to virus via tRNA modification and programmed ribosomal frameshifting. *Molecular Systems Biology*, 8:567, 2012.

[130] M Meyer, B Wong, M Styczynski, T Munzner, and H Pfister. Pathline: A Tool For Comparative Functional Genomics. *Computer Graphics Forum*, 29(3):1043–1052, August 2010.

[131] Miriah Meyer, Tamara Munzner, Angela DePace, and Hanspeter Pfister. MulteeSum: a tool for comparative spatial and temporal gene expression data. *IEEE transactions on visualization and computer graphics*, 16(6):908–917, November 2010.

[132] Miriah Meyer, Tamara Munzner, and Hanspeter Pfister. MizBee: a multiscale synteny browser. *IEEE transactions on visualization and computer graphics*, 15(6):897–904, November 2009.

[133] Christian Miller, Björn Schwalb, Kerstin Maier, Daniel Schulz, Sebastian Dümcke, Benedikt Zacher, Andreas Mayer, Jasmin Sydow, Lisa Marcinowski, Lars Dölken, Dietmar E Martin, Achim Tresch, and Patrick Cramer. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular Systems Biology*, 7(1):–, January 2011.

[134] H J Morowitz. *The completeness of molecular biology.* Israel journal of medical sciences, 1984.

[135] D K Murray and H. Bremer. Control ofspoT-dependent ppGpp Synthesis and Degradation inEscherichia coli. *Journal of Molecular Biology*, 1996.

[136] Frederick Carl Neidhardt, John L Ingraham, and Moselio Schaechter. Physiology of the bacterial cell: a molecular approach. 1990.

[137] Frederick Carl Neidhardt, John L Ingraham, and Moselio Schaechter. *Physiology of the bacterial cell: a molecular approach.* Sinauer Associates Sunderland, MA, 1990.

[138] Jeffrey D Orth, Tom M Conrad, Jessica Na, Joshua A Lerman, Hojung Nam, Adam M Feist, and Bernhard O Palsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Molecular Systems Biology*, 7:535, 2011.

[139] Jeffrey D Orth, Tom M Conrad, Jessica Na, Joshua A Lerman, Hojung Nam, Adam M Feist, and Bernhard Ø Palsson. A comprehensive genome-scale reconstruction of escherichia coli metabolism2011. *Molecular systems biology*, 7(1):535, 2011.

[140] Jeffrey D Orth, Ines Thiele, and Bernhard O Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, March 2010.

[141] Paola Picotti, Mathieu Clément-Ziza, Henry Lam, David S Campbell, Alexander Schmidt, Eric W Deutsch, Hannes Röst, Zhi Sun, Oliver Rinner, Lukas Reiter, Qin Shen, Jacob J Michaelson, Andreas Frei, Simon Alberti, Ulrike Kusebauch, Bernd Wollscheid, Robert L Moritz, Andreas Beyer, and Ruedi Aebersold. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature*, 494(7436):266–270, January 2013.

[142] Oliver Purcell, Bonny Jain, Jonathan R. Karr, Markus W Covert, and Timothy K Lu. Towards a whole-cell modeling approach for synthetic biology. *Chaos (Woodbury, N.Y.)*, 23(2):025112, June 2013.

[143] Udi Qimron, Boriana Marintcheva, Stanley Tabor, and Charles C Richardson. Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America*, 103(50):19039–19044, December 2006.

[144] Udi Qimron, Stanley Tabor, and Charles C Richardson. New details about bacteriophage t7-host interactions-researchers are showing renewed interest in learning how phage interact with bacterial hosts, adapting to and overcoming their defenses. *Microbe*, 5(3):117, 2010.

[145] Karthik Raman and Nagasuma Chandra. Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics*, 10(4):435–449, July 2009.

[146] Harry P Rappaport. The fidelity of replication of the three-base-pair set adenine/thymine, hypoxanthine/cytosine and 6-thiopurine/5-methyl-2-pyrimidinone with T7 DNA polymerase. *The Biochemical journal*, 381(Pt 3):709–717, August 2004.

[147] Jennifer L Reed, Thuy D Vo, Christophe H Schilling, and Bernhard O Palsson. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biology*, 4(9):R54, 2003.

[148] Elijah Roberts, Andrew Magis, Julio O Ortiz, Wolfgang Baumeister, and Zaida Luthey-Schulten. Noise contributions in an inducible genetic switch: a whole-cell simulation study. *PLoS Computational Biology*, 7(3):e1002010, March 2011.

[149] Elijah Roberts, Andrew Magis, Julio O Ortiz, Wolfgang Baumeister, and Zaida Luthey-Schulten. Noise contributions in an inducible genetic switch: a whole-cell simulation study. *PLoS Computational Biology*, 7(3):e1002010, March 2011.

[150] J B Russell and G M Cook. Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological reviews*, 59(1):48–62, March 1995.

[151] P.D. Sadowski and C. Kerr. Degradation of Escherichia coli B deoxyribonucleic acid after infection with deoxyribonucleic acid-defective amber mutants of bacteriophage T7. *Journal of virology*, 6(2):149–155, August 1970.

[152] Jayodita C. Sanghvi, Sergi Regot, Silvia Carrasco, Jonathan R. Karr, Miriam Gutschow, Benjamin Jr. Bolival, and Markus W. Covert. Accelerated discovery via a whole-cell model. *Nature Methods*, 10(12):1192–1195, 2013.

[153] Jayodita C Sanghvi, Sergi Regot, Silvia Carrasco, Jonathan R. Karr, Miriam V Gutschow, Benjamin Bolival, and Markus W Covert. Accelerated discovery via a whole-cell model. *Nature methods*, 10(12):1192–1195, December 2013.

[154] John T Sauls, Dongyang Li, and Suckjoon Jun. Adder and a coarse-grained approach to cell size homeostasis in bacteria. *Current opinion in cell biology*, 38:38–44, February 2016.

[155] J.M. Savinell and B O Palsson. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *Journal of Theoretical Biology*, 154(4):421–454, 1992.

[156] M. Schaechter, O Maaløe, and N O KJELDGAARD. Dependency on medium and temperature of cell size and chemical composition during balanced grown of Salmonella typhimurium. *Journal of general microbiology*, 19(3):592–606, December 1958.

[157] Alexander Schmidt, Karl Kochanowski, Silke Vedelaar, Erik Ahrné, Benjamin Volkmer, Luciano Callipo, Kèvin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. The quantitative and condition-dependent escherichia coli proteome. *Nature biotechnology*, 2015.

[158] Alexander Schmidt, Karl Kochanowski, Silke Vedelaar, Erik Ahrné, Benjamin Volkmer, Luciano Callipo, Kèvin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. The quantitative and condition-dependent escherichia coli proteome. *Nature biotechnology*, 34(1):104–110, 2016.

[159] Ida Schomburg, Antje Chang, Sandra Placzek, Carola Söhngen, Michael Rother, Maren Lang, Cornelia Munaretto, Susanne Ulas, Michael Stelzer, Andreas Grote, Maurice Scheer, and Dietmar Schomburg. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic acids research*, 41(Database issue):D764–72, January 2013.

[160] Ida Schomburg, Antje Chang, Sandra Placzek, Carola Söhngen, Michael Rother, Maren Lang, Cornelia Munaretto, Susanne Ulas, Michael Stelzer, Andreas Grote, Maurice Scheer, and Dietmar Schomburg. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic acids research*, 41(Database issue):D764–72, January 2013.

[161] Matthew Scott, Stefan Klumpp, Eduard M Mateescu, and Terence Hwa. Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Molecular Systems Biology*, 10:747, 2014.

[162] Douglas W Selinger, Rini Mukherjee Saxena, Kevin J Cheung, George M Church, and Carsten Rosenow. Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome research*, 13(2):216–223, February 2003.

[163] N S Shepherd, G Churchward, and H. Bremer. Synthesis and activity of ribonucleic acid polymerase in Escherichia coli. *Journal of Bacteriology*, 141(3):1098–1108, March 1980.

[164] M L Shuler, S Leung, and C C Dick. A MATHEMATICAL MODEL FOR THE GROWTH OF A SINGLE BACTERIAL CELL*. *... of the New York Academy of ...*, 1979.

[165] Michael L Shuler, Patricia Foley, and Jordan Atlas. Modeling a minimal cell. *Methods in molecular biology (Clifton, N.J.)*, 881:573–610, 2012.

[166] Rae Silver, Kwabena Boahen, Sten Grillner, Nancy Kopell, and Kathie L Olsen. Neurotech for neuroscience: unifying concepts, organizing principles, and emerging tools. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(44):11807–11819, October 2007.

[167] M. Son, R.H. Watson, and P Serwer. The direction and rate of bacteriophage T7 DNA packaging in vitro. *Virology*, 196(1):282–289, September 1993.

[168] Daniel M Stoebel, Antony M Dean, and Daniel E Dykhuizen. The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products. *Genetics*, 178(3):1653–1660, March 2008.

[169] F.W. Studier. The genetics and physiology of bacteriophage T7. *Virology*, 39(3):562–574, November 1969.

[170] F.W. Studier and J.J. Dunn. Organization and expression of bacteriophage T7 DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, 47 Pt 2:999–1007, 1983.

[171] S. Tabor, H E Huber, and C.C. Richardson. Escherichia coli thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *The Journal of biological chemistry*, 262(33):16212–16223, November 1987.

[172] Sattar Taheri-Araghi. Self-Consistent Examination of Donachie's Constant Initiation Size at the Single-Cell Level. *Frontiers in microbiology*, 6:1349, 2015.

[173] Sattar Taheri-Araghi, Serena Bradde, John T Sauls, Norbert S Hill, Petra Anne Levin, Johan Paulsson, Massimo Vergassola, and Suckjoon Jun. Cell-size control and homeostasis in bacteria. *Current biology : CB*, 25(3):385–391, February 2015.

[174] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, May 2009.

[175] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science (New York, N.Y.)*, 329(5991):533–538, July 2010.

[176] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science (New York, N.Y.)*, 329(5991):533–538, July 2010.

[177] Yu Tanouchi, Anand Pai, Heungwon Park, Shuqiang Huang, Rumen Stamatov, Nicolas E Buchler, and Lingchong You. A noisy linear map underlies oscillations in cell size and gene expression in bacteria. *Nature*, June 2015.

[178] Ines Thiele and Bernhard O Palsson. Reconstruction annotation jamborees: a community approach to systems biology. *Molecular Systems Biology*, 6:361, April 2010.

[179] J W Tobias, T E Shrader, G Rocap, and A Varshavsky. The N-end rule in bacteria. *Science (New York, N.Y.)*, 254(5036):1374–1377, November 1991.

[180] JW Tobias, TE Shrader, G Rocap, and A Varshavsky. The n-end rule in bacteria. *Science*, 254(5036):1374–1377, 1991.

[181] M Tomita, K Hashimoto, K Takahashi, T S Shimizu, Y Matsuzaki, F Miyoshi, K Saito, S Tanida, K Yugi, J C Venter, and C A Hutchison. E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15(1):72–84, January 1999.

[182] M Tomita, K Hashimoto, K Takahashi, T S Shimizu, Y Matsuzaki, F Miyoshi, K Saito, S Tanida, K Yugi, J C Venter, and C A Hutchison. E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15(1):72–84, January 1999.

[183] Yoshihiro Toya, Nobuyoshi Ishii, Kenji Nakahigashi, Takashi Hirasawa, Tomoyoshi Soga, Masaru Tomita, and Kazuyuki Shimizu. 13c-metabolic flux analysis for batch culture of escherichia coli and its pyk and pgi gene knockout mutants based on mass isotopomer distribution of intracellular metabolites. *Biotechnology progress*, 26(4):975–992, 2010.

[184] David A Van Valen, Takamasa Kudo, Keara M Lane, Derek N Macklin, Nicolas T Quach, Mialy M DeFelice, Inbal Maayan, Yu Tanouchi, Euan A Ashley, and Markus W Covert. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLoS Computational Biology*, 12(11):e1005177, November 2016.

[185] A VARMA, B W Boesch, and B O Palsson. Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. *Applied and environmental Microbiology*, 59(8):2465–2473, August 1993.

[186] A Varma and B O Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and environmental Microbiology*, 60(10):3724, 1994.

[187] A VARMA and BO Palsson. Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use. *Bio-Technology*, 12(10):994–998, 1994.

[188] J Vind, M A Sørensen, M D Rasmussen, and S Pedersen. Synthesis of proteins in Escherichia coli is limited by the concentration of free ribosomes. Expression from reporter genes does not always reflect functional mRNA levels. *Journal of Molecular Biology*, 231(3):678–688, June 1993.

[189] N Wade. Life is pared to basics; complex issues arise. *New York Times*, 1999.

[190] Mats Wallden, David Fange, Ebba Gregorsson Lundius, Özden Baltekin, and Johan Elf. The synchronization of replication and division cycles in individual e. coli cells. *Cell*, 166(3):729–739, 2016.

[191] Mats Wallden, David Fange, Ebba Gregorsson Lundius, Özden Baltekin, and Johan Elf. The Synchronization of Replication and Division Cycles in Individual E. coli Cells. *Cell*, 166(3):729–739, July 2016.

[192] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, January 2009.

[193] Daniel S Weaver, Ingrid M Keseler, Amanda Mackie, Ian T Paulsen, and Peter D Karp. A genome-scale metabolic flux model of Escherichia coli K-12 derived from the EcoCyc database. *BMC Systems Biology*, 8:79, June 2014.

[194] U Wittig, R Kania, M Golebiewski, M Rey, L Shi, L Jong, E Algaa, A Weidemann, H Sauer-Danzwith, S Mir, O Krebs, M Bittkowski, E Wetsch, I Rojas, and W Muller. SABIO-RK–database for biochemical reaction kinetics. *Nucleic acids research*, 40(D1):D790–D796, December 2011.

[195] Pak Chung Wong, Han-Wei Shen, Christopher R Johnson, Chaomei Chen, and Robert B Ross. The Top 10 Challenges in Extreme-Scale Visual Analytics. *IEEE Computer Graphics and Applications*, 32(4):63–67, 2012.

[196] Y Yamada, P A Whitaker, and D NAKADA. Early to late switch in bacteriophage T7 development: functional decay of T7 early messenger RNA. *Journal of Molecular Biology*, 89(2):293–303, October 1974.

[197] Lingchong You, Patrick F Suthers, and John Yin. Effects of Escherichia coli physiology on growth of phage T7 in vivo and in silico. *Journal of Bacteriology*, 184(7):1888–1894, April 2002.

[198] Eva Yus, Tobias Maier, Konstantinos Michalodimitrakis, Vera van Noort, Takuji Yamada, Wei-Hua Chen, Judith A H Wodke, Marc Güell, Sira Martínez, Ronan Bourgeois, Sebastian Kühner, Emanuele Raineri, Ivica Letunic, Olga V Kalinina, Michaela Rode, Richard Herrmann, Ricardo Gutiérrez-Gallego, Robert B Russell, Anne-Claude Gavin, Peer Bork, and

Luis Serrano. Impact of genome reduction on bacterial metabolism and its regulation. *Science (New York, N.Y.)*, 326(5957):1263–1268, November 2009.

[199] X Zhang, P Dennis, M Ehrenberg, and H. Bremer. Kinetic properties of rrn promoters in Escherichia coli. *Biochimie*, 84(10):981–996, October 2002.

[200] Honglin Zhou, Min Xu, Qian Huang, Adam T Gates, Xiaohua D Zhang, John C Castle, Erica Stec, Marc Ferrer, Berta Strulovici, Daria J Hazuda, and Amy S Espeseth. Genome-scale RNAi screen for host factors required for HIV replication. *Cell host & microbe*, 4(5):495–504, November 2008.