

KEY ISSUES REVIEW

Simulating biological processes: stochastic physics from whole cells to colonies

To cite this article: Tyler M Earnest *et al* 2018 *Rep. Prog. Phys.* **81** 052601

View the [article online](#) for updates and enhancements.

Related content

- [Noise in biology](#)
Lev S Tsimring
- [Stochastic switching in biology: from genotype to phenotype](#)
Paul C Bressloff
- [Approximation and inference methods for stochastic biochemical kinetics—a tutorial review](#)
David Schnoerr, Guido Sanguinetti and Ramon Grima

Recent citations

- [Complex microbial systems across different levels of description](#)
Benedikt von Bronk *et al*





IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

Key Issues Review

Simulating biological processes: stochastic physics from whole cells to colonies

Tyler M Earnest^{2,3,7} , John A Cole^{1,7} and Zaida Luthey-Schulten^{1,2,3,4,5,6} 

¹ Department of Physics, University of Illinois, Urbana, IL, 61801, United States of America

² Department of Chemistry, University of Illinois, Urbana, IL, 61801, United States of America

³ National Center for Supercomputing Applications, University of Illinois, Urbana, IL, 61801, United States of America

⁴ Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana, IL, 61801, United States of America

⁵ Center for the Physics of Living Cells, University of Illinois, Urbana, IL, 61801, United States of America

⁶ Carl Woese Institute for Genomic Biology, University of Illinois, Urbana, IL, 61801, United States of America

E-mail: zan@illinois.edu

Received 25 August 2017, revised 4 December 2017

Accepted for publication 9 February 2018

Published 5 April 2018



Corresponding Editor Professor Robert H Austin

Abstract

The last few decades have revealed the living cell to be a crowded spatially heterogeneous space teeming with biomolecules whose concentrations and activities are governed by intrinsically random forces. It is from this randomness, however, that a vast array of precisely timed and intricately coordinated biological functions emerge that give rise to the complex forms and behaviors we see in the biosphere around us. This seemingly paradoxical nature of life has drawn the interest of an increasing number of physicists, and recent years have seen stochastic modeling grow into a major subdiscipline within biological physics. Here we review some of the major advances that have shaped our understanding of stochasticity in biology. We begin with some historical context, outlining a string of important experimental results that motivated the development of stochastic modeling. We then embark upon a fairly rigorous treatment of the simulation methods that are currently available for the treatment of stochastic biological models, with an eye toward comparing and contrasting their realms of applicability, and the care that must be taken when parameterizing them. Following that, we describe how stochasticity impacts several key biological functions, including transcription, translation, ribosome biogenesis, chromosome replication, and metabolism, before considering how the functions may be coupled into a comprehensive model of a ‘minimal cell’. Finally, we close with our expectation for the future of the field, focusing on how mesoscopic stochastic methods may be augmented with atomic-scale molecular modeling approaches in order to understand life across a range of length and time scales.

Keywords: chemical master equation, flux balance analysis, metabolism, reaction–diffusion master equation, stochastic chemical kinetics, gene expression, whole-cell modeling

(Some figures may appear in colour only in the online journal)

⁷Equal contribution

1. Introduction: beyond the mean

If so much of our modern understanding of biology is cast in the language of chemistry, and so much of our chemistry is built upon the bedrock of modern statistical and quantum physics, it is perhaps no surprise that so many physicists have wandered into the weeds of the biological sciences and made profound contributions there. These contributions began remarkably early, and have punctuated the history of modern biology. They include the investigations of researchers like Muller, Max Delbrück, and Erwin Schrödinger, who speculated about the physical properties necessary for a molecule to encode genetic information before DNA's role had even been implicated [1–3], and the work of Alan Turing who considered the types of reaction–diffusion systems that produce the spatial patterns seen in biological morphogenesis [4]. They also include experimental triumphs, like the crystallographic elucidation of the DNA double helix [5] by Watson, Crick and Franklin, the structure of proteins with the work of Kendrew *et al* [6] being the first, and Woese's sequence-based approach to phylogeny that would reveal the archaeal domain of life [7].

With the advent of modern computers, the predictive-ness of biophysical models has grown considerably. From the configurational space of proteins and other biomolecules elucidated through molecular dynamics (MD) and Brownian dynamics (BD) simulations, to the propagation of action potentials through neural networks, to the systems-level complexity of reaction networks involving thousands of biochemical species, theoretical biophysicists are diving deeper and uncovering more about the nature of life. Never far from their minds, however, is the knowledge that the apparent order they see in living systems emerges from the chaotic and inherently stochastic motions of a teeming soup of biomolecules. Understanding this 'order from disorder' remains an important part of understanding the biological world. After all, life only happens far from equilibrium, but we live in a universe of ever-increasing entropy; we can never faithfully describe the former without accounting for the randomness of the latter.

In this review we explore the current state of stochastic modeling of cellular processes at the level of whole cells. We will begin with some history, touching on a number of watershed theoretical and experimental results. We will then review the basic theory necessary to describe stochastic, spatially resolved chemical networks in section 2, then discuss simulation methodologies, available software to perform simulations, and how to design and parameterize such models. Section 3 will discuss how stochasticity plays a role in the fundamental molecular processes of biology: transcription, translation, DNA replication, ribosome biogenesis, and metabolism. In section 4, we will explore how these processes interact in the cell, and how to unite individual models of these processes to a model of a whole cell. We will then move from single cells to colonies in section 5 and describe how one builds colony-level models based on physical principles. Finally, we close with a discussion of the possibility of uniting mesoscopic stochastic modeling with atomic-scale molecular dynamics simulations in order to bridge the length- and time-scales of the molecular level to the level of an organism.

1.1. Stochastic physics in biology: a historical review

For much of the history of modern biology, researchers have largely been comfortable with descriptions of cellular processes that focus on mean behavior. Microbial growth, for example, is often described in terms of lag, log, and stationary phases, as though every cell in a population fits neatly into one of these categories at any given moment. Of course in reality, it's long been understood that substantial heterogeneity exists within populations of microbes. Persister cells, for example, were discovered in the 1940s and represent an essentially dormant fraction of log-phase populations that can survive antibiotic treatment [8]. This is not to mean that 'log-phase', cannot be a useful description, but rather that in many cases, cellular behavior should be thought of in terms of *distributions*. Some cells in a population may be growing exponentially, others may not be growing at all, and en masse the population has some effective growth rate.

The distributional nature of many cellular processes revealed itself slowly over many years. As early as the 1930s, researchers began to interpret variability in bacterial cell-cycle duration in terms of multi-step stochastic processes [9]. In 1945, Delbrück showed that the number of bacteriophages in infected *Escherichia coli* cells varied more than could be accounted for by the variability in the size of the cells themselves, and proposed stochasticity in the 'autocatalytic' process of viral reproduction as a possible explanation [10]. A few years later, Benzer showed that sub-optimally induced *E. coli* expressed variable numbers of β -D-galactosidase (part of the *lac* operon) [11]. Novick and Weiner [12] expanded on this work to show that at low methyl-1-thio- β -D-galactopyranoside concentrations (or TMG, a commonly used *lac* inducer) concentrations, induction of the *lac* operon was an 'all-or-none' process—some cells in a population expressed the operon at its full rate, while the others did not express it at all [12]. This finding was also attributed to the 'autocatalytic' (i.e. involving a positive feedback loop) nature of the induction. Maloney and Rotman [13] made these findings more concrete, measuring a bimodal distribution of β -D-galactosidase enzymes in *E. coli* under low levels of induction [13]. In 1976, [14] described variability in the durations of 'swimming' and 'tumbling' behaviors in individual bacteria in response chemical attractants, and attributed it to Poissonian fluctuations in the production of small numbers of regulator molecules [14].

By the 1990s, a number of groups were actively investigating the degree to which stochasticity in gene expression can lead to phenotypic variations within cell populations [15–19]. But it was still not until the first few years after the turn of the millennium that the study of gene expression 'noise' (often defined in terms of the ratio of the protein copy number variance to its squared mean, $\text{Var}[p]/E[p]^2$) would explode into the mainstream of biological physics research. It started in a now-classic article by [20]. Using *E. coli* engineered to express yellow and cyan fluorescent proteins, each from identical promoters located on opposite sides of the chromosome, the authors showed that 'intrinsic' and 'extrinsic' gene expression noise could be experimentally distinguished. Intrinsic noise is associated with the fluctuations due to discrete particle numbers

and random reaction times, whereas extrinsic noise describes the fluctuations arising from all other noise sources such as variability in the quantities of components in the gene expression machinery (e.g. RNA polymerase (RNAP), ribosomes, and ribonucleases). In the above example, extrinsic noise leads to correlated fluctuations between the cyan and yellow fluorescent protein copy numbers, however the fluctuations from the intrinsic noise sources are not correlated between the two reporter proteins. This work would be followed by a string of high-profile experimental and theoretical results that: implicated the role that single-molecule events can play in phenotypic switching [21]; determined the distributions of proteins and messenger RNA (mRNA) that should arise from stochastic gene expression [22, 23]; and measured—at the genome scale—the variability of gene products in *E. coli* [24] and later in yeast [25]. Among the most celebrated results to emerge during this time was that protein copy numbers should be expected to follow a gamma distribution [22–24] (see figures 1(a) and (b)), although countless other important results would come from many other luminaries in the field [26–41].

1.2. Fluctuations percolate through cellular networks

It should come as no surprise that noise associated with gene expression has a number of important implications for the fitness of a living cell. This is due in part to the fact that so many key cellular phenomena require the combined function of many different gene expression products. Consider, for example, the step-wise association of several proteins to form a macromolecule, like a ribosome (see section 3.4 for details). As the macromolecule forms, the production of each intermediate complex is limited by the number of its available precursors. But each precursor—be it a protein or the previous intermediate—is itself subject to stochastic variability in production, and as a result the noise in each protein gets passed on through the association network of the macromolecule. As another example, one can consider large biochemical reaction networks like bacterial metabolism. Often, multiple chemical reactions may be arranged in sequential steps in which one metabolite may be transformed into another and then another through the activity of multiple different enzymes (glycolysis is one such example of a roughly linear pathway, see section 4). Because the number of each enzyme will vary from cell to cell, the maximum reaction flux through each sequence of reactions will also vary. At the level of the whole network, this can have profound effects, giving rise to wide distributions of growth rates (see figures 1(c) and (d)), metabolic efficiencies, and metabolic byproduct formation rates [42].

1.3. Single cell super-resolution imaging reveals mRNA and protein distributions

The ongoing revolution in stochastic physics in biology has been facilitated in large part by a concurrent revolution in single cell super-resolution imaging. Starting in the late 1980s, a number of physicists and chemists began developing methods to detect single fluorescent molecules [45], and image them with resolutions below the 200 nm diffraction limit of traditional

optical microscopy. Among them, stimulated emission depletion (STED), which relies on the use of a secondary laser to inhibit the emission of fluorophores in a ring around the center of the primary excitation laser, and thereby narrow the point spread function (PSF) of the emitted fluorescence, was the first to be applied to the imaging of cells [46]. Over the following 10 years, a family of techniques—PALM [47], FPALM [48], and STORM [49]—emerged that exploit the stochastic nature of fluorescent emission. In essence, these approaches rely on emission from just a few fluorophores at a time; provided no two fluorescent molecules within the diffraction limit of each other emit simultaneously, the detected photons recorded over a series of time points can be collected and the positions of individual molecules can be reconstructed with errors on the order of nanometers [50]. These techniques have been further developed to capture the location of a fluorophore in the *z*-axis as well, leading to a three-dimensional map of the fluorescent molecules in the cell. Figure 2 shows an example of the type of data techniques like 3D STORM [50] can provide. Recent advances in super-resolution imaging include the MINFLUX methodology [51], which can achieve ~ 1 nm resolution.

The impact of super-resolution microscopy in biophysics is reflected in the 2014 Nobel prize in Chemistry awarded to Moerner, Betzig, and Hell ‘for the development of super-resolved fluorescence microscopy’. It has revealed the architecture of microtubule networks, and the dynamics of the molecular motors that traverse them in living cells [54, 55], the diffusive and subdiffusive motions of macromolecules like RNAP and ribosomes within and around the chromosome [56, 57], as well as the numbers and locations of individual genes, messengers, and proteins [24, 53, 58], and the intramolecular motions of biological machines and complexes [59].

1.4. Spatial heterogeneity and cellular architecture

Over the past few decades, the living cell has come to be viewed as a crowded, spatially heterogeneous space, the structure of which can have profound effects on a wide range of chemical and biological processes [60–62]. The details of this space are increasingly being revealed through both super-resolution optical microscopy, and groundbreaking advances in cryo-electron microscopy and tomography [63–65]. Obstructed diffusion within the cytoplasm, for example, has been implicated in both the slowed association times and enhanced rebinding times between transcription factors and their targets [66, 67]. Additionally, varying densities of DNA have been shown to lead to spatial dependencies in the distributions of key molecular machines, including RNAP and ribosomes [56, 68, 69], and by extension, the proteins and mRNA they create. As a result, understanding the details of many cellular processes requires not merely a stochastic description, but a spatially resolved description that accounts for the random motions of the particles involved and the local environment in which they reside. Stochastic reaction–diffusion simulations have been brought to bear on a number of important biological processes, including the assembly of cell division machinery in bacteria [70, 71], cell polarization in yeast [72], and ribosome assembly [69].

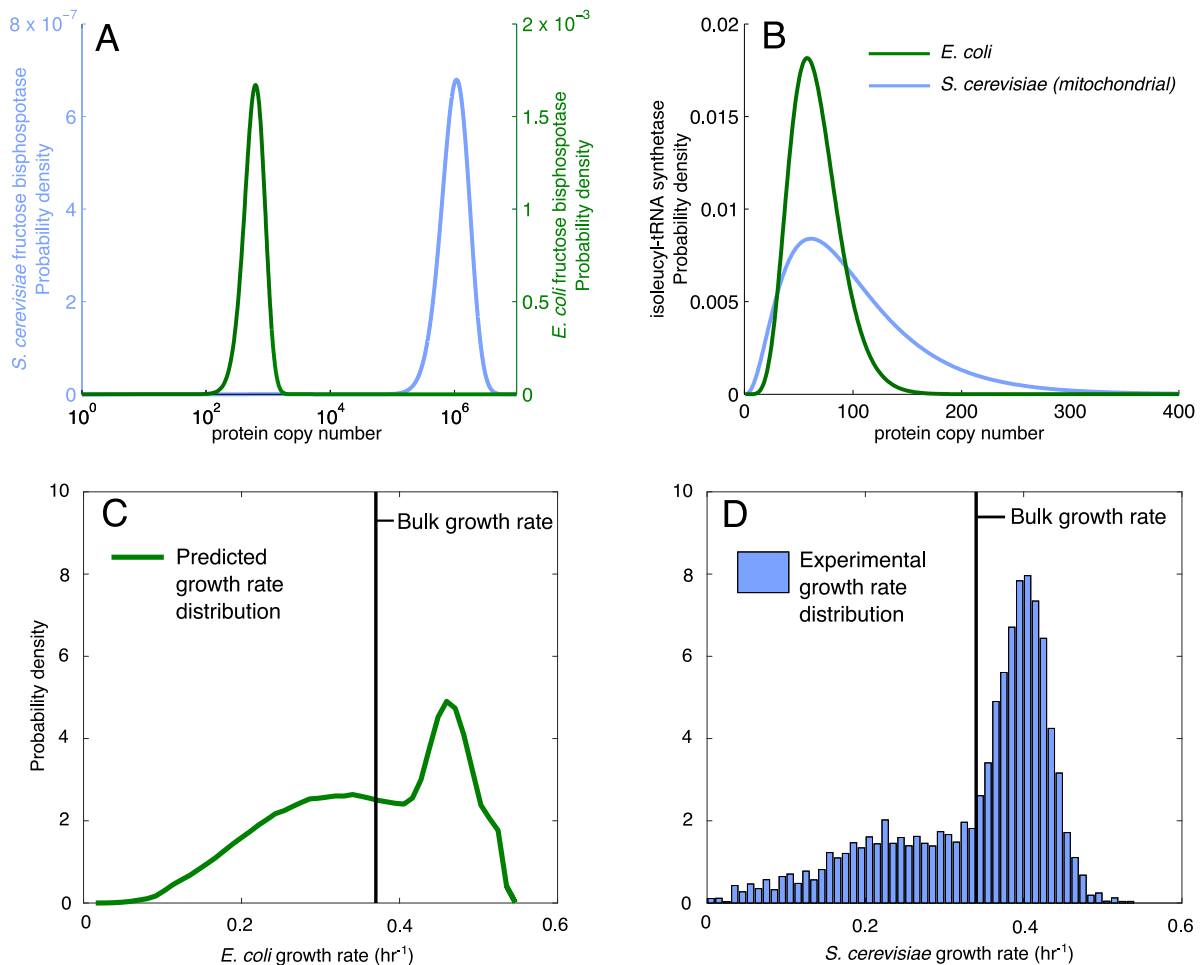


Figure 1. The distributional nature of life at the single cell level. Panels (A) and (B) show the distributions of copy numbers of fructose biphosphotase (part of the glycolytic pathway, see section 4.1 and figure 9) and isoleucyl-tRNA synthetase in the translation pathway of *E. coli* and the mitochondria of *S. cerevisiae*, respectively. Yeast cells are on the order of 100-fold larger than bacteria, and as a result, the fructose biphosphotase copy number, a cytosolic enzyme, appears in correspondingly larger numbers. The mitochondria within yeast are comparable to *E. coli* in size, and so the numbers of mitochondrial isoleucyl-tRNA synthetases appear in comparable numbers. Panels (C) and (D) show the similarity of predicted and experimentally measured growth rate distributions in *E. coli* and *S. cerevisiae*. The former was produced by sampling protein copy numbers from experimentally measured distributions, and using the resulting values as constraints in a genome-scale reconstruction of *E. coli* metabolism [42]. The latter shows experimental data adapted from Levy *et al* [43] (light blue), and a theoretical distribution also produced by sampling protein copy number distributions (dark blue, adapted from Labhsetwar *et al* [44]). Adapted from [44]. CC BY 3.0.

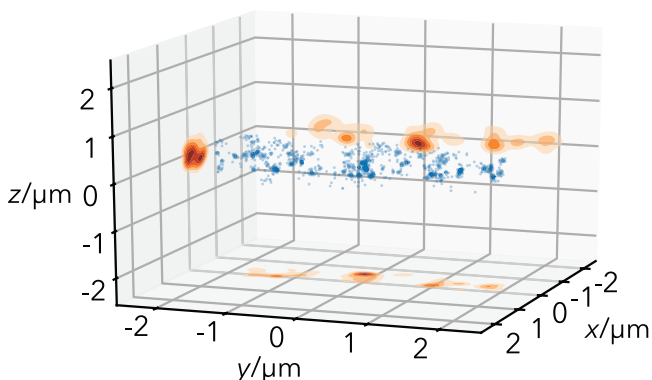


Figure 2. An example of 3D super-resolution data [52]. The small regulatory RNA, SgrS, in *E. coli* is labeled using single-molecule fluorescence *in situ* hybridization (smFISH) [53] and imaged using 3D STORM [50], revealing that the small RNA localize near the cell membrane. Blue points indicate individual molecule localization events, the orange density shows the localization event probability density projected onto the coordinate planes.

Due in part to their analytic intractability and significant computational complexity, a great body of work has thus far been devoted to developing efficient computational methods for simulating stochastic reaction diffusion systems. As a result, a major focus of what follows in this review will be devoted to these methods. Nevertheless, it is always the science that drives method development, and so the later portions of this review will focus on emerging questions in biology, and how the new computational methods can help to answer them.

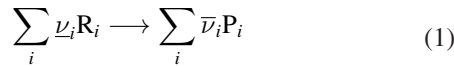
2. Spatially-resolved stochastic dynamics

In a stochastic, spatially resolved model of a cell, there exists the notion of the cellular state probability. This is a distribution over all spatial and chemical configurations of the cellular model. The time evolution of this distribution is governed by the reaction–diffusion master equation (RDME). For a general introduction to the physics of stochastic processes, the

reader is directed to the classic review by Chandrasekhar [73] in Reviews of Modern Physics and the textbooks by Gillespie [74], van Kampen [75], and Gardiner [76]. To define the RDME, we must first develop the basic notation and theory of stochastic chemical reaction networks.

2.1. Probability of the 'cellular state' and its equation of motion

We will denote an arbitrary chemical reaction of R_i reactants forming P_j products as



where ν_i is the stoichiometry for reactant i and $\bar{\nu}_j$ is the stoichiometry for product j . The rate of conversion through this reaction takes the form

$$\frac{dy}{dt} = k \prod_i c_i^{\nu_i}, \quad (2)$$

where i indexes reactants, the extent of reaction y is defined to be

$$y(t) = \frac{c_j(t) - c_j(0)}{\bar{\nu}_j - \nu_j}, \quad (3)$$

for any chemical species j , and the chemical concentration of species j is

$$c_j = \frac{n_j}{N_A \Omega}, \quad (4)$$

where n_i is the particle count, N_A is Avogadro's number, and Ω is the volume of the system. Equation (2) is not true in general. Only if the reaction is elementary, i.e. the reaction occurs in a single step with a single transition state, does this theory apply [77]. To apply this to non-elementary reactions, the reaction mechanism must be known so that each step can be represented as an elementary reaction.

The form of (2) can be derived from collision theory [77], however a simple justification follows from the fact that the reacting molecules must find each other within the reacting volume Ω in order to react. This means that the reaction rate must be proportional to the rate of particle encounters. The number of particle encounters per unit time depends on the number of ways that the reactant particles can come together to form a reacting complex. For example, the dimerization



requires two species to combine. There are n_A species, and the number of possible interactions is $n_A(n_A - 1)/2$ since in this reaction, a particle cannot interact with itself (the -1 term) and swapping the particles does not change the reaction (the $1/2$ term). This is clearly a binomial coefficient, and indeed the rate law for a single reaction can be written generally

$$\frac{dy}{dt} \propto \prod_i \binom{n_i}{\nu_i}. \quad (6)$$

In macroscopic systems, e.g. $n_i \sim N_A$, we can expand the binomial coefficient

$$\binom{n_i}{\nu_i} = \prod_{k=0}^{\nu_i} (n_i - k) = n_i^{\nu_i} \prod_{k=0}^{\nu_i} \left(1 - \frac{k}{n_i}\right) = n_i^{\nu_i} \left[1 + O\left(\frac{1}{n_i}\right)\right] \quad (7)$$

and truncate to zeroth order in $1/n_i$, from which (2) follows from the definition of the concentration, (4).

In order to maintain consistent units in (2), the dimensions of the chemical rate constant, k , depends on the reaction order,

$$\alpha = \sum_i \nu_i \quad (8)$$

as

$$[k] = \text{volume}^{\alpha-1} \text{time}^{-1}. \quad (9)$$

The reaction constant, k , encodes the details of the reaction kinetics, such as the diffusion rates of the reactants, the encounter geometry, temperature, and other microscopic details.

We will define a system of N_{rxn} reactions between N_{sp} chemical species as

$$X^T S = 0, \quad (10)$$

where

$$S = [\bar{\nu}_1 \ \bar{\nu}_2 \ \cdots \ \bar{\nu}_{N_{\text{rxn}}}] - [\nu_1 \ \nu_2 \ \cdots \ \nu_{N_{\text{rxn}}}] \quad (11)$$

is the $N_{\text{sp}} \times N_{\text{rxn}}$ stoichiometric matrix, constructed from the reactant (ν_r) and product ($\bar{\nu}_r$) stoichiometry vectors for all r reactions, such that matrix elements associated with products are positive and matrix elements associated with reactants are negative, and X symbolizes both the product and reactant chemical species. The system of chemical rate equations is then,

$$\frac{dc}{dt} = S \cdot J \quad (12)$$

where the flux vector is defined as,

$$J_r = k_r \prod_{i=1}^{N_{\text{sp}}} c_i^{S_{ir}}. \quad (13)$$

Equation (12), being a deterministic, continuous treatment, does not capture the true nature of the reactive dynamics of a chemical system at low particle numbers. The times and positions in which reactions occur are completely randomized due to Brownian motion of the molecules (the so-called Stoßzahlansatz, or molecular chaos hypothesis). Any memory of the prior state of the system is washed out after a time scale much shorter than the average time between reactions. The best that we can do is assign probabilities to the reactions and treat the system as a stochastic process. We assume that the system is 'well-stirred', meaning that the diffusion time scale is much shorter than the reaction time scale, which allows us to ignore spatial dependence. We also assume that the series of chemical reactions are described by a Poisson process with a rate which depends only on the current number of particles in the system, i.e. a Markov jump process. The defining equation of stochastic chemical kinetics in a 'well-stirred' environment is the chemical master equation (CME) [78, 79],

$$\frac{dP}{dt}(\mathbf{x}, t) = \sum_{r=1}^{N_{\text{rxn}}} a_r(\mathbf{x} - \mathbf{S}_r)P(\mathbf{x} - \mathbf{S}_r, t) - \sum_{r=1}^{N_{\text{rxn}}} a_r(\mathbf{x})P(\mathbf{x}, t) \quad (14)$$

where $a_r(\mathbf{x})$ is the propensity (i.e. transition rate) for reaction r while the system is in state \mathbf{x} , where

$$\mathbf{x}(t) = [x_1(t) \quad x_2(t) \quad \cdots \quad x_{N_{\text{sp}}}(t)]^T \quad (15)$$

which enumerates the particle counts for each species in the system, and \mathbf{S}_r is the column of the stoichiometric matrix corresponding to reaction r . Equation (14) describes the time evolution of the probability distribution function $P(\mathbf{x}, t)$ over the discrete space of particle number configurations of the system. The first summation in (14) is the rate of probability entering the state \mathbf{x} due to reactions from neighboring particle number states, while the second summation represents the rate of probability loss from \mathbf{x} due to reactions leaving the state. The CME performs bookkeeping on the states: probability lost from one state is immediately recovered in another. For an *in vivo* system, (14) can be considered as the equation of motion for the probability distribution function over cellular states. Here, the cellular state is the total number of biomolecules such as proteins, RNA, and metabolites.

The reaction propensities, $a_r(\mathbf{x})$, describe the transition rates between particle number states due to the action of reaction r . Again, we will only consider elementary reactions. The reaction propensity is

$$a_r(\mathbf{x}) = \kappa_r \prod_{i=1}^{N_{\text{sp}}} \binom{x_i}{S_{ir}}, \quad (16)$$

which follows from the same argument as (2), in that the overall rate of a reaction is proportional to the number of ways that the reactants can be grouped. However, here κ_r is the ‘stochastic rate constant’ not the deterministic rate constant k_r . They are related by

$$\kappa_r = (N_A \Omega)^{1-\alpha} k_r \quad (17)$$

since the deterministic rate equations are defined in terms of concentrations, whereas the CME is defined in terms of absolute numbers and the approximation (7) is no longer justified.

2.2. Stochastic simulations

It is difficult, if not outright impossible to solve the CME for many systems of interest, though a number of simplified models which capture the essential physics of important biological processes have analytic solutions, e.g. stochastic gene expression [22, 23, 80]. In general, for models including non-idealized descriptions of the chemical reaction network, the way around this difficulty is to generate trajectories sampled from the probability distribution the CME describes. The basic algorithm is the Gillespie direct method [81, 82], also known as the stochastic simulation algorithm (SSA). The algorithm was initially derived for gas phase kinetics and shown to be rigorous [79], then subsequently proven valid for the solution phase as well [78]. Starting out with the initial

species counts, \mathbf{x}_0 , the stoichiometric matrix, \mathbf{S} , and the propensity functions, $a_r(\mathbf{x})$, defined for each reaction r , the algorithm steps forward in time by randomly choosing the identity and time of the next reaction event. The time between subsequent reaction events is exponentially distributed, with a rate equal to the sum of all reaction propensities, a_{rxn} . The reason is transparent if you consider the CME for the current state and ignore incoming transitions,

$$\begin{aligned} \frac{dP_{\text{react}}}{dt} &= - \sum_{r=1}^{N_{\text{rxn}}} a_r(\mathbf{x})P_{\text{react}} \\ &= - \left(\sum_{r=1}^{N_{\text{rxn}}} a_r(\mathbf{x}) \right) P_{\text{react}} = -a_{\text{rxn}}P_{\text{react}}, \end{aligned} \quad (18)$$

whose solution is

$$P_{\text{react}}(t) = a_{\text{rxn}} e^{-a_{\text{rxn}} t}. \quad (19)$$

The probability that a reaction i fires is then simply

$$P_{\text{rxn}}(i) = \frac{a_i}{a_{\text{rxn}}}. \quad (20)$$

At each step of the SSA, a random reaction time $\tau \sim \text{Exp}(a_{\text{rxn}})$ is chosen, along with a random reaction index $i \sim P_{\text{rxn}}(\mathbf{a})$. The state is advanced by adding τ to the current time, and adding the net change of particles due to reaction i , i.e. the i th column of the stoichiometric matrix to the current particle counts.

Gillespie [82] presented an alternative algorithm, called the first reaction method. It differs from the direct algorithm in that a putative reaction time,

$$\tau_i \sim \text{Exp}(a_i(\mathbf{x})) \quad (21)$$

is computed for all reactions each time step. The minimum τ_i identifies both the time and index of the next reaction that fires. These two algorithms are mathematically equivalent [82], however the direct method is more computationally efficient since only two random variates are necessary per event as opposed to N_{rxn} as with the First Reaction Method.

Since Gillespie’s algorithms were published in [82], many improved algorithms have been published. The computational complexity of the direct method is $O(N_{\text{rxn}})$. The Next Reaction Method [83], which improves upon the First Reaction Method, is able to achieve $O(\log N_{\text{rxn}})$ complexity while only requiring a single random number per reaction event on average. The main feature of this method is that instead of recomputing all N_{rxn} tentative reaction times, $t + \tau_i$, after each reaction event, they are stored and used later in the simulation. They are only recomputed when the simulation advances past their scheduled time (and the system state is updated appropriately) or if reactions have occurred between t and $t + \tau_i$ which change the propensity for reaction i to fire. Techniques have been developed which improve upon the direct method such as partial propensity calculations [84–86] which are $O(N_{\text{sp}})$ instead of $O(N_{\text{rxn}})$, and methods which sort the reactions by propensity to decrease the number of iterations necessary to find a reaction [84, 87, 88], among others. An approximate method appropriate for large particle numbers is Tau leaping [89, 90]. This method evolves the dynamics forward in time using fixed time step τ ,

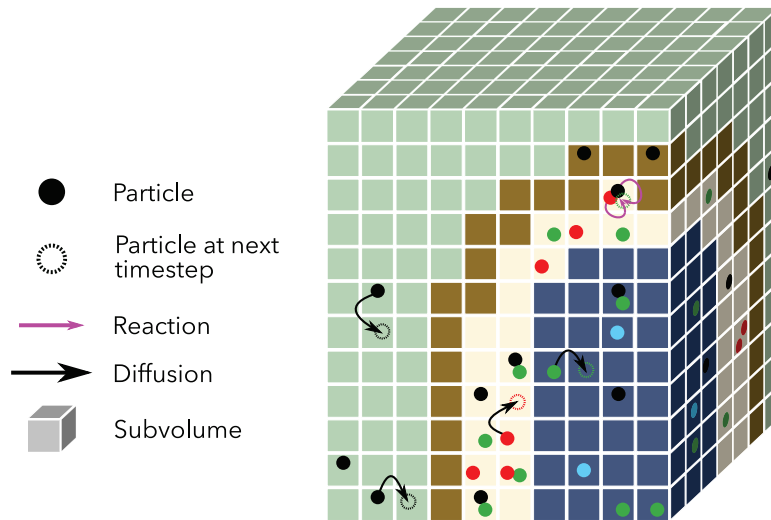


Figure 3. Schematic diagram of the RDME method using a cubic lattice. Circles indicate particles, their color indicates their species type. The cubes represent well mixed subvolumes of edge length λ , with their color indicating their site type. At each time step particles can diffuse to nearest neighbor subvolumes, or react with particles within the same subvolume. Site types allow for the behavior (e.g. varying diffusion coefficients or allowed chemical reactions) to vary with respect to position.

chosen to be larger than the expected time between reaction events yet short enough to ensure that the reaction propensities do not vary significantly, and updates the state to include the expected number of reactions which occur over the interval t to $t + \tau$. Other techniques for analyzing well-mixed stochastic systems beyond directly simulating the underlying stochastic process are reviewed in Schnoerr *et al* [91]. When applicable, these methods can provide analytic expressions for quantities such as mean and variance of particle abundances, and are generally much faster than direct simulation.

2.3. The reaction–diffusion master equation as a description of cellular processes

Cells can contain small copy numbers of chemical species which diffuse through a complex, crowded environment composed of other biomolecules and organelles. To accurately capture this behavior requires a stochastic, spatially resolved model. The two most common methods to simulate these systems at a biologically relevant scale are RDME and particle-based reaction–diffusion (PBRD) methods. The former method samples solutions from the probability distribution described by the RDME—an extension of the CME generalized to include spatial degrees of freedom. These RDME [92–95] methods do not track individual particles, but rather track the populations of chemical species within subvolumes of the simulation domain, as opposed to PBRD which accounts for the position of each particle in the simulation.

We assume that there is a combination of a length- and time-scale in which the system’s behavior can be considered well-mixed. Then for a particular choice of time scale, the reaction volume can be divided into subvolumes of a corresponding length scale such that both the chemical reactions and the diffusion between neighboring subvolumes can be treated as a Poisson process. The equation of motion governing this description is the RDME:

$$\begin{aligned} \frac{dP(\mathbf{x}, t)}{dt} = & \sum_{\nu}^V \sum_{r=1}^{N_{\text{rxn}}} [-a_r(\mathbf{x}_{\nu})P(\mathbf{x}_{\nu}, t) \\ & + a_r(\mathbf{x}_{\nu} - \mathbf{S}_r)P(\mathbf{x}_{\nu} - \mathbf{S}_r, t)] \\ & + \sum_{\mu}^V \sum_{\nu}^{\text{n.n.}(\mu)} \sum_{\alpha=1}^{N_{\text{sp}}} [-d_{\mu}^{\alpha} x_{\mu}^{\alpha} P(\mathbf{x}, t) \\ & + d_{\nu}^{\alpha} (x_{\nu}^{\alpha} + 1)P(\mathbf{x} + \mathbf{1}_{\nu}^{\alpha} - \mathbf{1}_{\mu}^{\alpha}, t)], \quad (22) \end{aligned}$$

where $P(\mathbf{x}, t)$ is the probability to find a configuration of particle counts and locations \mathbf{x} at time t . The first summation in (22) describes the flow of probability between different copy number states at every subvolume, and is simply a chemical master equation for that subvolume. The reaction propensities $a_r(\mathbf{x}_{\nu})$ give the transition probabilities due to reaction r firing at site ν , and are computed following (16). The vector \mathbf{S}_r is the r^{th} column of the stoichiometric matrix describing the change in species counts when reaction r fires. The second summation, where $\text{n.n.}(\mu)$ denotes all nearest neighbors of subvolume μ , describes the flow of probability due to diffusion between neighboring subvolumes, where each unique chemical species α is treated separately to allow for differing diffusion rates. The vector $\mathbf{1}_{\nu}^{\alpha}$ represents a single molecule of species α in volume ν , i.e. $(\mathbf{1}_{\nu}^{\alpha})_{\mu}^{\beta} = \delta_{\alpha\beta} \delta_{\mu\nu}$. Each subvolume is treated as well-stirred reaction volume, allowing for the reactions in each subvolume to be simulated independently. Figure 3 provides a schematic description of the dynamics simulated. These methods are generally less computationally expensive than particle-based methods, however excluded volume effects between reacting particles are neglected. Molecular crowding due to other molecules in the cell can be modeled through the introduction of obstacles in the lattice geometry [67, 96], however. The use of spatial discretization could lead to reduced accuracy compared to particle methods, but it has been shown that RDME methods approach the same

level of accuracy when the reaction radii are much smaller than the largest subvolume separation [97–100].

The spatial decomposition can take on any form provided that the separation between lattice sites is smaller than a particle would be expected to diffuse over a single time step. However it is convenient to decompose the system into a cubic lattice. In this case, the RDME takes the form

$$\begin{aligned} \frac{dP(\mathbf{x}, t)}{dt} = & \sum_{\nu} \sum_{r=1}^{N_{\text{rxn}}} [-a_r(\mathbf{x}_{\nu})P(\mathbf{x}_{\nu}, t) \\ & + a_r(\mathbf{x}_{\nu} - \mathbf{S}_r)P(\mathbf{x}_{\nu} - \mathbf{S}_r, t)] \\ & + \sum_{\nu} \sum_{\xi} \sum_{\alpha=1}^{N_{\text{sp}}} [-d_{\nu}^{\alpha} x_{\nu}^{\alpha} P(\mathbf{x}, t) \\ & + d_{\nu+\xi}^{\alpha} (x_{\nu+\xi}^{\alpha} + 1)P(\mathbf{x} + \mathbf{1}_{\nu+\xi}^{\alpha} - \mathbf{1}_{\nu}^{\alpha}, t)], \end{aligned} \quad (23)$$

where the configuration vector \mathbf{x} contains the number of species present at each individual lattice site,

$$\mathbf{x} = [\mathbf{x}_{1,1,1} \ \mathbf{x}_{1,1,2} \ \cdots \ \mathbf{x}_{1,1,N_z} \ \mathbf{x}_{1,2,1} \ \cdots \ \cdots \ \mathbf{x}_{1,N_y,N_z} \ \cdots \ \cdots \ \cdots \ \mathbf{x}_{N_x,N_y,N_z}]^T \quad (24a)$$

$$\mathbf{x}_{i,j,k} = [x_{i,j,k}^1 \ \cdots \ x_{i,j,k}^{N_{\text{sp}}}]^T, \quad (24b)$$

and the second term now describes particle diffusion in terms of the ξ transitions along the lattice axes: $\pm\hat{i}$, $\pm\hat{j}$, and $\pm\hat{k}$. For a cubic lattice with spacing λ , the diffusive propensity is

$$d_{\nu}^{\alpha} = \frac{D_{\nu}^{\alpha}}{\lambda^2}, \quad (25)$$

and is computed by treating diffusion as a discrete random walk of step size λ and associating the diffusion constant D_{ν}^{α} with the discrete step probability. The cubic lattice representation is simple to implement computationally, however it forces all subvolumes be the same size. This can be problematic where it is necessary to treat sections of the simulated volume at differing levels of detail. This arises frequently in the simulation of neurons since dendritic spines are much more narrow compared to the neuronal body.

A highly efficient method to sample trajectories from RDME on a cubic lattice is the multi-particle diffusion RDME (MPD-RDME) algorithm. It is based on an operator-splitting method [101] and the direct SSA [81], and is most similar to the Gillespie multiparticle (GMP) method developed by Rodriguez *et al* [102]. An implementation of this algorithm using the massively parallel architecture of modern graphics processing units (GPUs) hardware is available in the Lattice Microbes package [67, 92, 93, 96, 103, 104]. Lattice Microbes trajectories are capable of reaching hour long time scales—orders of magnitude longer than competing codes [94, 105–107]. By taking sufficiently short time steps such that particles are unlikely to take part in multiple reactions, the subvolumes are rendered independent, and can be simulated in parallel (implementation details can be found in Roberts *et al* [96], Roberts *et al* [92], and Hallock *et al* [93]).

The MPD-RDME algorithm represents the simulation volume as a cubic lattice, where each subvolume contains a

finite number of particles. The particles are represented by an array of integers, where the value of each integer greater than zero identifies both the presence of a particle and its species type. A value of zero indicates a vacancy. The simulation loop proceeds by executing the GPU-based procedures (called kernels) for diffusion in the x , y , and z directions sequentially, followed by the reaction kernel. The simulation time is updated as $t_{i+1} = t_i + \tau$, and once $t > t_{\text{final}}$ the loop exits and the simulation terminates. These kernels are executed in parallel on the GPU where each thread is responsible for a single subvolume. The simulation algorithm takes regular time steps, as opposed to the Gillespie direct algorithm which takes time steps of varying length sampled from an exponential distribution.

During a time step $[t, t + \tau]$, the probability of a reaction occurring is simply

$$P_{\text{react}} = \int_0^{\tau} dt a_{\text{rxn}} e^{-a_{\text{rxn}} t} = 1 - e^{-a_{\text{rxn}} \tau} \quad (26)$$

following (19). Each time step, a random number $\rho \sim \text{Uniform}(0, 1)$ is drawn and if $\rho < P_{\text{react}}$, then a reaction will occur at that time step. The specific reaction is then chosen according to (20), as in the exact stochastic simulation algorithm. The diffusion kernels proceed similarly. The probability that the particle leaves its site is

$$P_{\text{diffuse}} = 1 - e^{-a_{\text{dif}} \tau}, \quad (27)$$

where a_{dif} is the sum of the two diffusive propensities to transition along the diffusion kernel axis, e.g. $2d_{\nu}^{\alpha}$ in the case where the site types of the subvolumes at -1 , 0 , and $+1$ are all identical.

The nature of the MPD-RDME algorithm places constraints on the model parameters and the coarseness of the lattice. The largest diffusion constant in the system and the lattice spacing dictates the largest valid time step,

$$\tau < \frac{\lambda^2}{2 \max_{\alpha} D_{\alpha}}, \quad (28)$$

that can be taken. This relationship is a consequence of the fact that diffusion in the RDME is a discrete random walk. The decoupling of reactions from diffusion used in this method relies on a separation between diffusion and reaction time scales. We define the diffusion time scale to be

$$\tau_{\text{D}} = \frac{\lambda^2}{6D_{\text{max}}} \quad (29)$$

and the reaction time scale to be

$$\tau_{\text{R}} = \frac{1}{a_{\text{max}}}, \quad (30)$$

where a_{max} is the largest reaction propensity. Then

$$\tau_{\text{R}} \gg \tau_{\text{D}} \quad (31)$$

implies that

$$\lambda \ll \sqrt{\frac{6D_{\text{max}}}{a_{\text{max}}}}. \quad (32)$$

Substituting in the expressions for reaction propensities (16), we see that there are upper and lower bounds on the lattice size:

$$\lambda \ll \left(\frac{6D_{\max}}{J_{\max}^{(0)} N_A} \right)^{1/5} \quad (\text{zeroth-order}) \quad (33)$$

$$\lambda \ll \sqrt{\frac{6D_{\max}}{J_{\max}^{(1)}}} \quad (\text{first-order}) \quad (34)$$

$$\lambda \gg \frac{J_{\max}^{(2)}}{6D_{\max} N_A} \quad (\text{second-order}), \quad (35)$$

where $J_{\max}^{(i)}$ is the maximum i th-order flux (13) evaluated using typical lattice site concentrations.

In the implementation of the MPD-RDME algorithm, the simulation volume is represented by an $N_x \times N_y \times N_z \times N_p$ array of integers, where $N_{x,y,z}$ are the number of lattice sites in each dimension and N_p is the lattice occupancy. The finite lattice occupancy is a consequence of the GPU-oriented nature of the algorithm, allowing for the GPU to access the lattice memory in a regular pattern. This implies that the maximum particle concentrations that can be simulated are constrained by the maximum number of particles per subvolume. If a reaction or diffusion event causes any subvolume to exceed its capacity, the computation on the GPU must be placed on hold so that a process can run on the host to correct the overflow. Particles in the offending site are redistributed among the neighboring subvolumes which have empty particle slots available. Frequent overflows will cause the computational efficiency to plummet due to the repeated shuffling of the lattice data between the host and GPU.

The probability of an overflow occurring due to diffusion can be computed by considering particle placement as a series of Bernoulli trials [96]. Consider an empty lattice containing L_s subvolumes, each having a maximum occupancy of n_{\max} , to which we add N particles. The trial in this case is whether or not a particle is placed randomly at a particular lattice site. If all lattice sites are equally likely to receive a particle, then the probability of the success of a single trial is $p = 1/L_s$. The probability that a particular subvolume receives n particles then follows the binomial distribution

$$P(n) = \binom{N}{n} \left(\frac{1}{L_s} \right)^n \left(1 - \frac{1}{L_s} \right)^{N-n}. \quad (36)$$

The overflow probability of a single site is then

$$P_{\text{of}} = 1 - \sum_{n=0}^{n_{\max}} P(n), \quad (37)$$

from which it follows that the expected number of overflows, N_{of} , is

$$E[N_{\text{of}} | n_{\max}, L_s, N] = L_s \left[1 - \sum_{n=0}^{n_{\max}} \binom{N}{n} \left(\frac{1}{L_s} \right)^n \left(1 - \frac{1}{L_s} \right)^{N-n} \right]. \quad (38)$$

By considering the RDME simulation as a series of Bernoulli trials where success represents a time step with no particle overflows, it can be shown that the mean number of time steps between overflows is simply

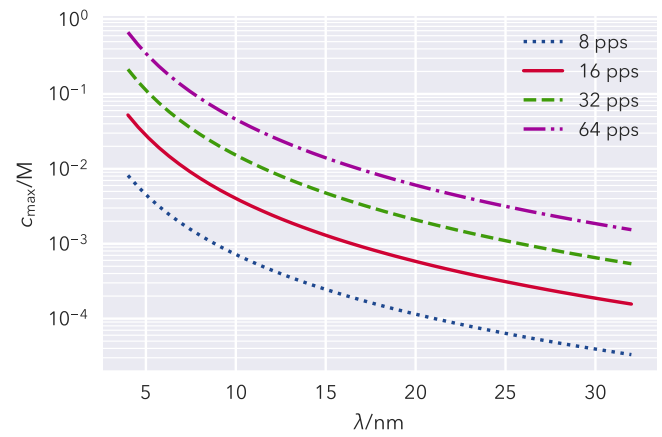


Figure 4. Dependence of maximum concentration on lattice spacing, λ , for a 1 fl simulation volume. The curves are plotted by solving (39) with a mean number of time steps between overflows, $t_{\text{of}} = 1000$, and using 8, 16, 32, or 64 particles per subvolume.

$$t_{\text{of}} = \frac{1}{1 - (1 - P_{\text{of}})^{L_s}}. \quad (39)$$

For Lattice Microbes, an acceptable number of time steps between overflows should be greater than 1000 in order to minimize the impact of host-GPU memory transfers on the simulations performance. Equation (39) can be solved numerically to find the appropriate lattice occupancy for a required particle density (figure 4). For an *E. coli* sized volume (1 fl), with a 32 nm lattice spacing at a maximum of 8 particles per subvolume, it follows from (39) that we can simulate systems with total particle concentrations of $\sim 33 \mu\text{M}$. At the highest concentrations seen in a model of ribosome biogenesis [69] ($42 \mu\text{M}$), this lattice configuration leads to overflows occurring once per 58 time steps on average, drastically decreasing the efficiency of the computation. Doubling the particles per subvolume to 16 effectively eliminates overflows, resulting in a mean of 2.6×10^{11} steps between overflows. To treat all proteins in an *E. coli* cell (5 mM), at a 32 nm lattice spacing would require a maximum particle occupancy of 156. Using the smallest possible lattice spacing of 20 nm (to fully contain a ribosome), we can reduce the maximum occupancy to 56.

2.4. RDME-based simulation software

Software to generate cell state trajectories following RDME has been available since at least [108] with the introduction of SmartCell [108]. SmartCell uses a version of the next reaction method which treats diffusion to neighboring subvolumes in the same way as reaction events. This technique was later employed in MesoRD [94, 109] in 2005 as the next subvolume method (NSM). Source code and binary distributions for these codes are available as of this writing, however the projects no longer appear to be in active development. GPGMP [110] is a GPU-accelerated code employing the GMP method [102]. GMP is a split-operator approach which treats the subvolumes as independent reaction volumes whose dynamics are simulated using the direct SSA. Diffusion of particles is simulated using a lattice gas automaton [101], a cellular

Table 1. Comparison of actively developed, freely available RDME-based spatially resolved stochastic simulation software.

| Software | Method | Geometry | Interface | Parallel | References |
|------------------|--------------|-------------|---------------|-----------|------------|
| Lattice Microbes | MPD-RDME | Lattice | Python 2 or 3 | MGPU, MPI | [92, 93] |
| MesoRD | NSM | Lattice | SBML | N | [94, 109] |
| NeuroRD | S_τ /GD | Mesh | XML | N | [111] |
| Spatioocyte | cRDME | hcp lattice | Python 2 | N | [123, 124] |
| STEPS | SSA/MND | Mesh | Python 2 | MPI | [120, 122] |
| URDME | NSM | Mesh | Python 2 | N | [95] |

automaton which allows multiple diffusion events to occur at each time step. Source code for GPGMP is available, however it is no longer in active development.

SmartCell, MesoRD, and GPGMP all require the cellular geometry to be discretized onto a regular cubic lattice, however many later codes do not have this restriction. NeuroRD [111] was the first code which represented the geometry as a tetrahedral mesh in order to accurately represent the geometry of dendritic spines. This code does not use a split-operator technique, but instead uses a spatial tau-leaping / gradientbased diffusion (S_τ /GD) method: both reactions and diffusion are simulated using tau-leaping and the gradient between neighboring subvolumes is used to sample the total number of particles migrating within a time step [112, 113]. URDME is another code which uses tetrahedral meshing, however it uses the NSM which exactly samples the underlying distribution [95]. This package has tight integration with the commercial finite-element analysis system Comsol Multiphysics, allowing for convenient generation and manipulation of mesh data. Due to the Python interface to URDME [114], this software is easily extended to other simulation modalities such as time-dependent geometry [115] and molecular crowding [116].

Lattice Microbes [92, 96, 103, 104], by nature of the highly parallelizable MPD-RDME algorithm, performs all RDME simulations on GPU hardware, allowing for simulation times on the order of hours to be completed within days. Implementations of Lattice Microbes are available which distribute the simulation domain over multiple GPUs attached to a single workstation (MGPU) or over multiple GPU-equipped compute nodes over MPI, which has allowed for hour-long simulations of yeast cells to be completed within 28 h [117]. Simulations are designed and controlled from a Python-based environment [103] similar to PyURDME [114], allowing the programmatic construction of complicated reaction models [69] and systems with time-varying geometry [118]. STAUCC [119] is another parallelized code which features both MPI and GPU implementations of the spatial tau-leaping algorithm. However neither source code nor binary distributions of the simulator are freely available. STEPS [120–122] is a GPU-based code using tetrahedral mesh geometry using a split operator approach. Reactions are performed using the SSA on a per subvolume basis, while diffusion is performed using a multinomial multiparticle diffusion (MND) method. NeuroRD, URDME, Lattice Microbes, and STEPS are all under active development. Table 1 provides a summary of the freely available voxel-based simulation software.

2.5. Particle-based simulation software

As an alternative to voxel-based methods, PBRD simulations allow for the study of stochastic reaction–diffusion systems without discretizing space. These particle-based methods track the position and identity of each particle in space, and evolve their positions in time using BD where the position of each particle i is updated as

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + \frac{1}{\zeta_i} \mathbf{f}_i(\{\mathbf{x}\}, t) \tau + \sqrt{2D_i} \boldsymbol{\eta}(t) \sqrt{\tau}, \quad (40)$$

where τ is the time step, D_i is the diffusion constant which is related to the drag coefficient ζ_i through the Einstein relation $D\zeta = k_B T$, $\mathbf{f}_i(\{\mathbf{x}\}, t)$ is the sum of forces acting on the particle, and $\boldsymbol{\eta}(t)$ is a Gaussian random variable with zero mean and unit variance. Reactions between particles are implemented through assigning reaction probabilities to interacting particles if their separation $\|\mathbf{x}_i - \mathbf{x}_j\|$ is less than the sum of their reaction radii. Depending on the implementation of the simulation software, attractive and repulsive interactions can be accounted for in these methods through the force term, allowing for the simulation of molecular crowding and aggregation.

MCell was the first PBRD simulation software, appearing in 1996 [125]. Particles are propagated in time using a Monte Carlo (MC) algorithm using a fixed time step which does not account for particle–particle interactions or excluded volume. In addition to 3D diffusion, particles can transition onto 2D structures such as membranes. Model specification with MCell can be done using text files, or a graphical interface (CellBlender) which takes advantage of the free open source 3D computer graphics package Blender. Smoldyn [126, 127], ChemCell [128], and Cell++ [129] use a similar diffusion methodology and have similar features. eGFRD [130] uses a novel Green function reaction dynamics (GFRD) algorithm which decomposes the many-body problem into one- and two-particle independent subproblems. These subproblems have analytic solutions to the Smoluchowski equation in the form of Green functions, which can be used to advance the simulation in forward in time. The major benefit to this technique is that larger time steps can be taken, accelerating the simulation time. Interestingly, the computational complexity of GFRD is $O(N_{sp}^{5/3})$ in contrast to the $O(N_{sp})$ scaling at constant volume that most PBRD algorithms follow. This leads to a crossover at high particle densities, where standard PBRD techniques are more efficient. Finally, ReaDDy [107] directly treats particle–particle interactions through forces, allowing for the highest level of computational detail and physical

Table 2. Comparison of actively developed freely available PBRD based spatially resolved stochastic simulation software. Packages which account for the excluded volume of reacting particles are indicated in the EV column.

| Software | Method | EV | Interface | Parallel | References |
|----------|--------|----|------------|----------------|------------|
| Smoldyn | MC | N | Text | N ^a | [126, 127] |
| MCell | MC | N | GUI & Text | N | [106, 125] |
| ReaDDy | BD | Y | Python 2 | N ^a | [107] |
| eGFRD | GFRD | N | Text | N | [130] |

^a Parallel implementation exist, yet are no longer maintained.

accuracy currently available. A parallelized implementation using CUDA has been released [131], however it is no longer in active development. For further information on this simulation methodology, the review by [132] provides an excellent overview. A summary of the actively developed simulation codes is provided in table 2.

2.6. RDME versus particle methods

In general, PBRD methods provide a more accurate description of the underlying phenomena, however they have the drawback of being significantly more computationally expensive. The serial RDME simulator MesoRD [94, 109] is approximately $3\times$ faster than MCell [106, 125] and Smoldyn [126, 127]. Lattice Microbes is approximately $100\text{--}250\times$ faster than MCell and Smoldyn on account of its GPU implementation and greater efficiency of the RDME [103]. However, PBRD can rigorously include particle interactions, while such interactions can only be approximated in RDME formalism [133]. For PBRD methods which account for excluded volume interactions directly, such as ReaDDy [107], the maximum allowable time step is on the order of nanoseconds since the average displacement per step should be small compared to the smallest particle radius. An indirect method of dealing with crowding in PBRD simulations has been proposed by [134], which allows for time steps bounded only by the maximum reaction timescale and the required spatial resolution. Here, the concentration and radii of non-reactive crowdors are used to compute the probability to reject a potential Brownian displacement or reaction event, otherwise the simulation algorithm is similar to other PBRD codes.

The RDME as an approximation does not converge to the expected behavior in the limit of $\lambda \rightarrow 0$, since the encounter probability of two particles vanishes at the limit of infinitesimal lattice spacing [97]. A modification to the standard methods, called the cRDME [135] or volume RDME (vRDME) [136], restricts the state of the system to a single particle per subvolume and allows second-order reactions between nearest-neighbor particles. This technique is exact in the limit of $\lambda \rightarrow 0$, but it requires a much finer lattice than standard RDME methods and neighboring lattice sites can no longer be computed independently, which renders the method computationally expensive and decreases its applicability to dense cellular systems. However the method recovers effects of excluded volume on the particle number statistics seen in

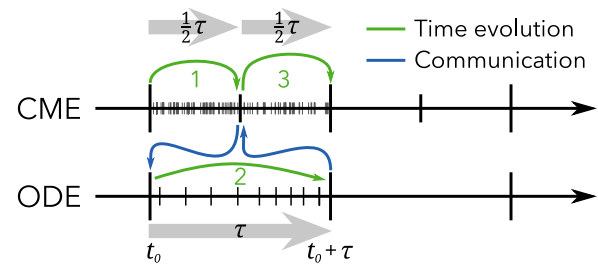


Figure 5. Communication times between the CME and ODE and the adaptive time steps within the ODE used in the hybrid scheme developed by Jahnke and Kreim [139]. First, the CME simulation is advanced in time by a half time step. Then the CME state at $t + \frac{1}{2}\tau$ is then used to update the ODE state. The ODE is then advanced in time by a full time step. Finally, the CME state is updated with the ODE state at $t + \tau$ and advanced to the end of the time step.

PBRD simulations with excluded volume [136], underscoring the importance of these non-reactive particle interactions.

2.7. Hybrid methods

Many processes within living cells, especially gene expression shown in figure 7, are characterized by low particle numbers and a high degree of randomness which brings about stochastic effects. So far we have written mostly about the RDME or CME descriptions of cellular processes where the system is assumed to follow a Markov jump process over the state space of particle numbers in time, to capture the discreteness of the population and the randomness of the dynamics (22). Unbiased realizations of the Markov processes are typically obtained with some variation of the widely used SSA; however, this algorithm is constrained by the fact that each single reaction event requires an update of the system, making the simulations for certain types of systems computationally costly. These problematic systems are characterized by large total reaction propensities; either due to large particle abundances or due to high reaction rates. In such a situation there may exist a partitioning of the species into slow species where their total reaction propensity is low and fast species whose total reaction propensity is high. A challenging and typical scenario is when species found in low abundance react with species found in high abundance, making the dynamics of the low-concentration species dependent on the dynamics of the high-concentration species. Numerical methods [137, 138] to improve the computational efficiency by reformulating the original scheme in a more economical way or by developing multi-scale stochastic approaches in which the high propensity segments of the system are described by ordinary differential equation (ODE) and the low propensity segments are treated stochastically are briefly reviewed in the 2012 article by Jahnke and Kreim [139]. Their review of the hybrid piecewise stochastic deterministic method (CME/ODE, see figure 5) includes a rigorous error analysis of their partitioning scheme which was validated against a pure SSA simulation of a rather small system. In these hybrid models, a Markov jump process describing the dynamics of the low-abundance species is coupled to deterministic rate equations modeling the high-abundance species. Such a partitioning works well

for cellular systems described by RDME or CME and typically improves the speed of the numerical simulations by a factor of 50–100, making it an indispensable tool for complex whole-cell simulations with a large number of species types, cellular components, and high concentrations of metabolites in the extracellular and intracellular regions. While it is intuitively tempting to assume a partial thermodynamic limit for the fast reactions involving a large number of species and just rescale the rate constants so the hybrid system so can be treated by a single stochastic model, this assumption cannot be made about the behavior of genetic switches in the early phases of the response to sugars, inducers, and metabolites. The optimal communication times between the CME and ODE descriptions as well as the time steps for each method must be accessed to verify that the hybrid description does not compromise the simulation accuracy or lose any of the stochastic effects which often have the greater impact on the cell's behavior.

In RDME simulations, which account for the spatial heterogeneity in the cell, additional factors need to be considered. In the case of highly abundant, slowly diffusing species, a hybrid simulation combining a deterministic reaction–diffusion partial differential equation (PDE) model is in order. The RDME/PDE hybrid has been studied previously [140–142], as well as the PBRD/PDE hybrid [143]. Depending on the resolution of the subvolumes compared to the dimensions of the largest moving particle, it may be again necessary to use a hybrid method in which the equations of motion for the large particles follow Brownian dynamics in order to correctly account for the effects of excluded volume. PBRD and RDME methodologies have been coupled previously using Smoldyn [144], however the implementation of BD did not account for excluded volume interactions nor other particle–particle interactions. The impetus for the coupling was to allow for the simulation of chemical systems over varying levels of length scale, e.g. a yeast cell treated with PBRD and its surrounding environment with RDME. A successful coupling of an on-lattice method with an off-lattice method which accounts for non-reactive interactions between particles would allow the effect of molecular crowding on the spatiotemporal behavior of cellular biochemistry to be comprehensively explored computationally.

2.8. Model development: how to design and parameterize kinetic models

The study of a stochastic, spatially resolved system begins with the description of the model. We must decide the geometry and architecture of the simulation domain, which chemical species to include and their initial distribution within the simulation volume, the diffusion coefficients of these species, and which reactions are significant and the rates in which they occur. Unfortunately, the reality is that biological models suffer from a lack of information. Either the only information available is from *in vitro* experiments whose results do not necessarily apply to the intracellular environment, the necessary experiments have not been performed, or there is no viable way to measure the quantity of interest. Thus, the

usual solution is to infer the parameters of interest through optimization of the model against the available experimental data. For modeling a system which behaves deterministically, this is straightforward yet non-trivial: fit the model solution to experimental measurements of the time course of the biological system. However for a stochastic system, the situation is far more complicated since the system is expected to behave differently from experiment to experiment—*distributions* must be considered. As a first step, an exploration of the parameter space of the chemical reaction network using deterministic rate equations can be helpful in acquainting one's self with the model.

With the complexity of chemical networks in biology, the resulting reaction models can have many parameters. Varying these parameters can have profound effects on the behavior of the model. For instance, the value of a parameter may determine whether a model of a genetic switch exhibits bistability [145]. However, it has been observed that many models have the surprising characteristic that some of the parameters, or functions of parameters, can be varied with little impact on the quality of fit to the experimental data. In fact, this has been claimed to be a *universal* property of dynamic models in biology [146], and has even been suggested to be exploited by Nature to provide biological robustness [147]. This feature is termed 'sloppiness' and is measured from the Hessian matrix of the model objective function. If the objective function,

$$C(\boldsymbol{\theta}) = \sum_{ij} \left(\frac{y_i(t_j; \boldsymbol{\theta}) - Y_{ij}}{\sigma_{ij}} \right)^2, \quad (41)$$

describes the deviation of the vector of experimental measurements Y_{ij} , with uncertainty σ_{ij} , at time point j , for experimental conditions i , from the measurement predicted by the model $y_i(t_j; \boldsymbol{\theta})$ for the set of model parameters $\boldsymbol{\theta}$, then the Hessian matrix is

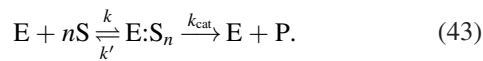
$$H_{ij}(\boldsymbol{\theta}) = \frac{\partial^2 C}{\partial \ln \theta_i \partial \ln \theta_j} \quad (42)$$

where the derivatives are taken with respect to the logarithm of the parameter value to account for the fact that the parameters will take on different units and scales of magnitude. When one inspects the eigenvalue spectrum of the Hessian, evaluated at a minimum of the objective function, it is common to see that the eigenvalues span many orders of magnitude. This is the manifestation of sloppiness: the parameter space is split into stiff dimensions where the eigenvalues are large and the model output is sensitive, and sloppy dimensions where the eigenvalues are small and the model is insensitive. A model is declared sloppy if the ratio of the largest eigenvalue to the smallest is greater than 10^3 [146]. Sloppiness should be not be confused with unidentifiability. Identifiability describes the ability to uniquely quantify all model parameters given sufficient experimental data, assuming that the model faithfully represents the phenomenon in question. Sloppy models are not necessarily unidentifiable [148].

Another problem with high dimensional models is that the fitness landscape of the objective function can be rugged, with many minima corresponding to acceptable fits to the

experimental data. Many of these minima in parameter space (including the global minimum), can take on values which are not realistic [149]. It is sometimes useful to break this symmetry using a regularization term in the objective function which penalizes unrealistic parameter values [150] while testing the viability of the model to accelerate the optimization process. A suggested way to deal with these problems is through optimal experiment design [151], which attempts to predict the experimental conditions whose data would minimize the uncertainty in the best-fit model parameters. However, White *et al* [152] warn that this solution can lead to an overall loss of predictive power of the model in spite of the lowered parameter uncertainty since the optimal experimental conditions can inadvertently magnify the effect of details which were left out of the model. Another way to quantify the uncertainty in the parameter estimates is to use bootstrapping [153]. The experimental data set is sampled with replacement to generate an ensemble of replicated data sets. The best fit parameters for each element of the ensemble allows us to construct a histogram of parameter values from which confidence intervals can be determined.

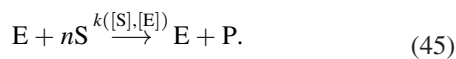
These chemical reaction networks are by their very nature coarse-grained, however further coarse-graining can be performed to reduce the number of reactions or to account for the lack of knowledge of the complete biochemical process. This sort of coarse-graining uses non-mass-action reaction propensities, the most common of which is Hill kinetics. Consider the scheme,



Assuming that $[E]$ is small compared to $[S]$, it can be shown that

$$-\frac{d[S]}{dt} = \frac{k_{\text{cat}}[E] \cdot [S]^n}{(k/k')^n + [S]^n}, \quad (44)$$

allowing us to replace (43) with



The right-hand-side of (44) can be written as

$$k(y) = \frac{V_{\text{max}} \cdot y^H}{K_{1/2}^H + y^H}, \quad (46)$$

for a concentration $y = [S]$, and can be interpreted as a saturating function of y , which reaches one half of its maximum rate,

$$V_{\text{max}} = k_{\text{cat}}[E], \quad (47)$$

when y reaches the concentration,

$$K_{1/2} = \frac{k}{k'}. \quad (48)$$

The steepness of the curve at $K_{1/2}$ is governed by the Hill coefficient,

$$H = n, \quad (49)$$

which can be understood as a measure of the ‘cooperativity’ of the reaction, i.e. the effect of binding a ligand affects the enzyme’s ability to bind subsequent molecules. Hill

coefficients greater than one describe cooperative binding where the binding rate of further substrates increases with the number of substrate molecules bound to the enzyme, and negative Hill coefficients describe the opposite situation. When $H = 1$, the reaction is non-cooperative and describes Michaelis–Menten (MM) kinetics. However, the value of H determined from fitting to (46) cannot be used to infer the number of binding substrates. The result is predicated on the assumption that reactions with order greater than two can occur, which is nonsense. Instead, ligands bind sequentially and it is possible that there can be intervening reactions involving the formation of an activated complex [154]. In general, the use of non-elementary propensity functions in RDME models is problematic. Lawson *et al* [155] reported a comparison between reactions with MM propensity functions and the equivalent mass-action reactions in an RDME model over varying lattice spacings. They showed that steady state ES abundance computed from the MM approximation diverges wildly from the mass-action kinetics as the lattice spacing decreases. Smith and Grima [156] later proposed that the reason for this discrepancy is that in the fast-diffusion limit, the RDME for the MM approximation does not converge to the expected MM approximated CME. They claim that because the use of non-elementary propensity functions implicitly assume that reactions represent the fastest time scale in the model, this is inconsistent with the assumptions underlying the RDME. However, in light of these complications the functional form of (46) fits remarkably well to the reaction rate of many systems. This means that many common reaction motifs seen in biology, such as MM reactions and reactions whose rates are described by Hill functions, must be decomposed into elementary mass-action kinetics prior to inclusion in an RDME model.

Stochastic models add another layer of complexity to the parameter estimation problem. Instead of being able to minimize a deterministic function (41), we must contend with the fact that a single parameter set can be associated with an infinite set of realizations of the system. Approximate Bayesian computation (ABC) [157] provides a simple way to estimate parameter values and their uncertainties which works well for stochastic simulations. Assuming a prior distribution of parameter values, informed by previous knowledge of the biology or just reasonable estimates of minimum and maximum values, ABC provides an estimate of the likelihood function of the model given experimental data. The idea is to comprehensively sample the prior parameter distribution and compute the deviation of the model from the experiment. A threshold ϵ is chosen as the acceptance criterion: all parameter values which lead to deviations from the experiment less than ϵ are accepted. Usually, the dimensionality of the experimental data is so high that the probability of sampling an acceptable parameter set is low. To remedy this, instead of directly comparing the model to the data, summary statistics are used. For example, the parameter space of a model describing a three-state genetic switch was explored using ABC [150]. A model of a two-state switch [67] was extended to include a third genetic state describing the situation when the *lac* repressor binds to two operators simultaneously, forming a DNA loop.

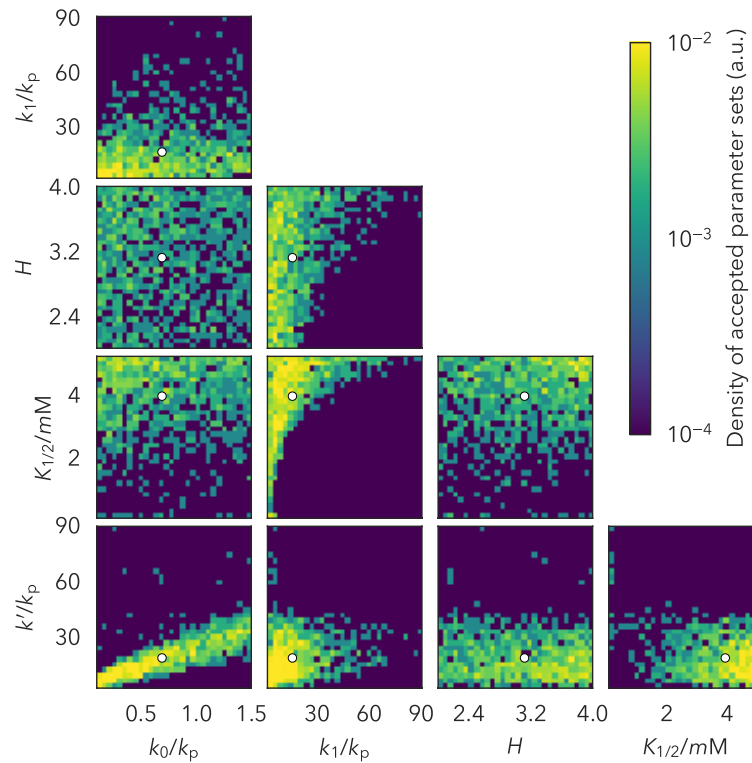


Figure 6. Example of ABC parameter estimation using a three-state model of the *lac* genetic switch in *E. coli* [150]. First order rate parameters are reported as multiples of the cell division rate (k_p), binding coefficients are reported in millimolar, and the Hill coefficients are dimensionless. The mean of the distribution is shown with a white marker. The parameter sets exhibiting bistability within a range of external conditions were accepted. Adapted from [150]. [CC BY 3.0](#).

The new state (loop) was connected to the repressed (off) state via the reactions,



Here the off-to-loop transition rate k' is constant, while the loop-to-off transition rate $k([\text{I}])$ is written as the Hill function

$$k([\text{I}]) = k_0 + (k_1 - k_0) \frac{[\text{I}]^H}{K_{1/2}^H + [\text{I}]^H}, \quad (51)$$

where k_0 is the loop-to-off transition rate without the presence of inducer ($[\text{I}] \rightarrow 0$) and k_1 is the loop-to-off transition rate at saturating inducer concentration ($[\text{I}] \rightarrow \infty$). The remaining model parameters were fixed, leading to a five-dimensional parameter space. The stochastic model was solved using the finite state projection (FSP) [158], which yields directly the probability distributions. The statistic of interest was the range of inducer concentration in which the system was bistable. Parameter sets which exhibited a range of bistability which included the experimentally observed bistability range were accepted. The prior distribution was a uniform distribution whose bounds were chosen using simple biological arguments. The resulting parameter distributions are shown in figure 6 as a series of two-dimensional histograms.

The joint histograms reveal the interdependence of the model parameters. The $k'-k_0$ histogram is an excellent example of model sloppiness: the ratio k'/k_0 clearly impacts the

bistability range, however the absolute values of the parameters are unimportant. This can be interpreted as the rate of switching between these transcriptional states is irrelevant: only the fraction of time spent in each state is important.

Using ABC was trivial in this example because the probability distributions were directly available. However in most cases the probability distribution is not accessible, instead we must infer the solution by simulating the stochastic process. In these cases, parameter estimation is significantly more difficult. The simulated distributions can be fit to the experimental data using a variety of objective functions such as the negative log-likelihood, the cumulative distribution function, or through comparing the distributions using a metric such as the Kullback–Leiber divergence [159]. Optimizing over these objective functions can be troublesome since evaluations are not deterministic and there is no objective function gradient available to provide to the optimizer. Srivastava and Rawlings [160] propose using the sample path method coupled with a derivative-free global optimization algorithm and bootstrapping for uncertainty analysis. For each simulated trajectory used to generate the histogram, the sample path method stores the value used to seed the random number generator. The random numbers are then replayed for each successive objective function evaluation, which effectively smooths the objective function in parameter space.

The previous examples were for well-mixed systems. Parameter optimization and sensitivity analysis for stochastic, spatially resolved systems has not been explored to the

same extent, unfortunately. The problem is that performing these simulations is generally many orders of magnitude more computationally expensive, making the direct exploration of parameter space difficult. Sensitivity analysis about a set of parameters in a one-dimensional RDME simulation has been described by [161]. They use finite differences to compute the derivative of a summary statistic with respect to a model parameter. They present a number of techniques based on the concept of variance reduction. Consider the summary statistic f evaluated over many realizations of the system at parameter sets θ and $\theta + \delta$, where δ is a small perturbation. The variance of the difference of the statistic between the two parameters is

$$\begin{aligned} \text{Var}[f(\theta + \delta) - f(\theta)] = & \text{Var}[f(\theta + \delta)] + \text{Var}[f(\theta)] \\ & - 2 \text{Cov}[f(\theta + \delta), f(\theta)]. \end{aligned} \quad (52)$$

By correlating the simulated time courses between the two parameter sets, $\text{Cov}[f(\theta + \delta), f(\theta)]$ is maximized which reduces the number of replicates necessary to estimate the statistic f . They present three methods: the split propensity method, the common Poisson process method, and the grouped sampling method. The split propensity method generates sample paths for two parameter sets simultaneously such that the total propensity for the next event considers whether the event occurs in the first, the second, or both systems. Since the perturbation is small, most events will occur in both systems such that the two sample paths are highly correlated. The common Poisson process method is similar to the sample path method in that the simulation of the two parameter sets share the same stream of random numbers. Finally the grouped sampling method extends the split propensity method to segment the space of possible events into $M + 2d$ groups, where M is the number of reactions, d is the dimension of the lattice, and the factor of 2 accounts for diffusion in both the positive and negative directions. The groups are then ordered by their voxel index. Events are selected by first choosing the group based on its total propensity, then the event from that group. The result is a sample path for both systems similar to the split propensity method with the added characteristic that shared events occur in similar positions in each system, further reducing the variance between the two statistics. In general, Lester *et al* [161] show that the grouped particle method requires the least number of simulations to achieve a given variance, followed by the coupled Poisson process method. For highly parallelized implementations of lattice-based stochastic simulators (i.e. Lattice Microbes), the coupled Poisson process method would be the most simple to apply with the least impact on simulation performance.

Recently, Schnoerr *et al* [162] have shown that an approximation to the likelihood function for stochastic, spatially resolved models can be derived using a connection between the RDME and spatial-temporal Cox point processes [163]. Spatial-temporal Cox processes are essentially Poisson processes over a given spatial domain and time interval where the Poissonian rate itself is a random variable. These processes are generally used to derive phenomenological models of stochastic systems evolving through time and space. Since

the resulting likelihood function is computed from the solution to a PDE or stochastic PDE, optimizing the likelihood function is computationally inexpensive. With the likelihood function readily available, it is possible to perform selection between sets of potential models using an information criterion. Schnoerr *et al* [162] report that the compute time necessary to optimize a four-parameter system was order of 10s, which suggests that this technique could be applied to parameter estimation in much larger, perhaps whole-cell models.

Other necessary data to construct whole-cell models include the diffusion coefficients and simulation geometry. As long as there is not a separation of diffusion scales between species in the model, RDME simulations are not in general sensitive to the exact value of the diffusion coefficients over time scales longer than the diffusion time [67, 69]. It is preferred to use *in vivo* measured diffusion rates when they are available, however this is rarely the case. Instead, estimations relating the radii of gyration to the diffusion coefficients in cytoplasm [164] are an appropriate replacement. Diffusion coefficients assigned to transitions between compartments with dissimilar diffusive properties can be computed as the geometric mean, $D_{a \leftrightarrow b} = \sqrt{D_a D_b}$.

The simulation volume is generally constructed manually based on measurements from microscopy [67, 69, 118, 120, 123]. However recently it has become possible to perform simulations using the actual 3D geometry data captured from tomographic methods [117, 165, 166]. Isaacson *et al* [166] used data from structured illumination microscopy of DAPI-stained DNA in mouse cells to study the diffusion of a transcription factor in the nucleus using an RDME simulation. Bartol *et al* [165] used serial electron microscopy sectioning of rat neurons to simulate Ca^{2+} transient formation using MCell. Earnest *et al* [117] presented two case studies in the use of cryoelectron tomograms as simulation geometry: a simulation of a hypothetical genetic switch in an *S. cerevisiae* cell, similar to the *lac* system in *E. coli*; and a simulation of an auto-repressing gene in a HeLa cell. The simulation geometry of the *S. cerevisiae* cell was inferred using measurements from a cryo-electron microscopy (cryo-EM), including the density of nuclear pores; whereas the HeLa simulation used the compartments created from the tomography data directly. The work focused on the construction of model geometry for use in lattice-based reaction–diffusion simulations, and suggested situations where experimentally derived geometry can be necessary. However, the importance of ‘real’ as opposed to idealized simulation geometries has yet to be studied in detail.

3. Stochasticity in universal cellular processes

If one were to consider the question ‘What is life?’ several essential properties should naturally spring to mind. Chief among them, living things have the ability to extract energy from their environment in order to grow, reproduce, and pass on their genes to the next generation. Of course, these properties alone are not strictly sufficient, and debate remains about what a formal definition of life could be (if one even exists), but they do give us a starting point to think about what behaviors

are truly important for understanding the living world. In this section, we review the current state of modeling several major cellular processes that are integral for the growth and replication of the cell. These processes include genetic information processing (transcription, translation, DNA replication, and ribosome assembly), and metabolism.

3.1. Transcription: modeling constitutive and regulated mRNA expression

The first step in the central dogma of molecular biology involves the transcription of a gene to form a messenger RNA. In its simplest form, this process can be modeled as the first order reaction



where D represents the gene, and m represents the mRNA, and k_t represents the transcription rate. Of course, this representation sweeps away a great deal of important biology (the association of transcription factors, the binding, open complex formation, and processivity of RNAP, etc), but it has nevertheless been used to great effect in countless high profile studies of gene expression (for just a few examples, see [22–24, 167, 168]).

The first order transcription and subsequent decay of mRNA,



which can be understood as a stochastic birth–death process, yields Poisson-distributed copy numbers with mean k_t/k_d . In many cases, accounting for the existence of regulatory machinery can be essential to accurately modeling the transcription of a gene. The binding of a transcription factor (TF) to either activate (transcriptional activator) or deactivate (transcriptional repressor) the gene can be straightforwardly included in a model by adding just a few reactions. For example, to account for the transcriptional activator T ,



where k_{tf} and k_{tf}' are the binding and dissociation rates of the TF to the gene, and D^* represents the gene in its transcriptionally active form. In the literature, the explicit dependence on the transcription factor is sometimes suppressed, and the gene is treated as though it undergoes first order transitions between inactive and active states with an effective rate $\tilde{k}_{tf} = k_{tf}\bar{n}_T$. This is often necessary when attempting analytical treatments, and it can be shown that (see [17]) the scheme



corresponding to the regulated production and decay of mRNA, yields a messenger mean and variance given by

$$\bar{m} = \frac{k_t}{k_d} \frac{\tilde{k}_{tf}}{\tilde{k}_{tf} + k_{tf}'} \quad (57a)$$

$$\sigma_m^2 = \bar{m} \left(1 + \frac{k_t k_{tf}'}{(\tilde{k}_{tf} + k_{tf}')(\tilde{k}_{tf} + k_{tf}' + k_d)} \right). \quad (57b)$$

Often, however, explicitly including the transcription factor, as well as the possible transitions the TF can undergo, can be necessary to capture the dynamics of the system.

Consider as an example the *lac* genetic switch—the ‘hydrogen atom’ of gene regulation. In this system, LacI, which represses transcription of the *lac* operon, can bind one or two lactose molecules. With lactose bound, LacI loses affinity for its DNA binding site, which in turn enables transcription of *lacZ*, *lacY*, and *lacA* to proceed. LacY is a lactose transporter, which creates a positive feedback loop—increased intracellular lactose enhances *lacY* expression, which in turn leads to enhanced lactose uptake. In general, feedback leads to effects such as the emergence of bistability due to up regulation (positive feedback) and noise reduction due to down regulation (negative feedback) of mRNA transcription. This is a ubiquitous feature of gene regulatory networks.

Due to their complexity, systems like this are amenable to analytical treatments only under certain simplifying assumptions (e.g. protein expression being modeled as a Poisson process [168], or occurring in bursts of geometrically-distributed sizes [169]). Nevertheless, these simplified descriptions—as well as more complete computational models [67, 150, 170]—have proven successful in describing a number of observed features of self-regulated genes, including the bimodality of LacY copy numbers at intermediate levels of inducer.

3.2. Translation: modeling protein distributions

Following transcription, the next step in gene expression is translation of the mRNA into its protein product. As was the case with transcription, translation is a complex process involving interactions of scores of different biomolecules (the association of the ribosomal subunits to the mRNA, interactions between the ribosome, mRNA, tRNAs, and elongation factors, etc) but it is often modeled simply as a first-order reaction. Because proteins are relatively long-lived, a common assumption is that they are lost primarily through dilution as the cell grows and divides. This too can be modeled as a first order reaction, giving a complete protein expression model,



where γ is the dilution rate, which can be related to the cell's doubling time, t_D , as $\gamma = (\ln 2)/t_D$. This model, and variations on it, has been analyzed on countless occasions. In particular, Friedman *et al* [22] in 2006 and Shahrezaei and Swain [23] in 2008 both showed that when the mRNA lifetime is assumed to be much shorter than that of the protein, the model gives rise to Poissonian 'bursts' of protein production of exponentially-distributed size, and leads to gamma-distributed protein counts [22, 23]:

$$P(p) = \frac{p^{a-1} e^{-p/b}}{\Gamma(a) b^a} \quad (59)$$

where the shape parameter, $a = k_i/\gamma$, represents the average number of protein production bursts in a cell cycle, and the scale parameter, $b = k_r/k_d$, represents the average number of proteins produced in a burst. This distribution has simple expressions for the mean and variance of the protein copy numbers:

$$E[p] = ab \quad (60a)$$

$$\text{Var}[p] = ab^2. \quad (60b)$$

The transition rates between the transcriptional states can divide the system into two regimes. The non-adiabatic regime is characterized by the transcriptional switching rates being much slower than the protein decay rates. Here, metastable populations with low and high average protein abundances develop corresponding to the inactive and active transcriptional states, respectively. The adiabatic regime is characterized by the transcriptional switching rates being much faster than the protein decay rates. The switching is fast enough that the metastable states blend into a single peak in the protein distribution and the two transcriptional states can then be considered as a single state with a mean transcription rate [33, 145].

In a seminal 2010 article, Taniguchi *et al* [24] constructed a library of over 1000 *E. coli* strains, each with a fluorescent protein fused to a different gene of interest. Then, by measuring the fluorescence of single cells of each strain, the authors estimated the copy number of each protein on the level of individual cells. They found that the experimental copy number distributions were fit extraordinarily well with a gamma distribution across a wide range of expression levels (see figures 1(a) and (b)) [24]. This work also provided genome-wide measurements of mRNA lifetimes and expression levels, and has become foundation for countless theoretical studies.

More general theories of gene regulation networks have been developed. Wang *et al* [36] describes a landscape and flux theory of gene regulatory networks. The time dependence of the various chemical abundances can be approximated continuously as

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}) + \boldsymbol{\eta}(t), \quad (61)$$

where $\mathbf{F}(\mathbf{x})$ is a generalized force describing the dynamics of the chemical abundances, \mathbf{x} , and $\boldsymbol{\eta}(t)$ is white noise with the autocorrelation function

$$\langle \boldsymbol{\eta}(t) \boldsymbol{\eta}(t') \rangle = 2\mathbf{D} \delta(t - t'), \quad (62)$$

with \mathbf{D} as the noise strength tensor. This leads a diffusion equation defined over concentrations which satisfies the conservation law

$$\frac{\partial P}{\partial t} + \nabla \cdot \mathbf{J}(\mathbf{x}, t) = 0. \quad (63)$$

Here $\mathbf{J}(\mathbf{x}, t)$ is the flux describing the evolution of the probability distribution, which at steady state $\nabla \cdot \mathbf{J}(\mathbf{x}, t) = 0$. It follows that steady state dynamics can be expressed in terms of a potential $U(\mathbf{x}) = -\log P_{s.s.}(\mathbf{x})$ and a non-equilibrium driving force $\mathbf{F}_{cf}(\mathbf{x}) = \mathbf{J}_{s.s.}/P_{s.s.}$ defined by the steady state probability distribution, $P_{s.s.}$, and steady state flux, $\mathbf{J}_{s.s.}$. This force is nonzero only when the curl of the flux is non-vanishing, which implies a loss of detailed balance. This admits an elegant interpretation of the network dynamics—the state of the system can be thought of as a charged particle under the influence of an electric potential U , and a magnetic force \mathbf{F}_{cf} . For a complete review of this theory, see [171]. This method was used to study the mammalian cell cycle [41]. Since protein abundances in such cells are generally large, the chemical Langevin description (61) is reasonable, however in bacterial systems this may not be appropriate.

3.3. DNA replication: modeling variations in initiation, termination, and replication duration

Cells can not reproduce without somehow copying their genetic material. But even in bacteria, where the process of DNA replication is comparatively straightforward, many details surrounding its control remain poorly understood. What is known is that bacterial cells initiate replication through the accumulation of DnaA to the chromosome near the origin of replication, *oriC*. This leads to DNA melting and strand separation, and the subsequent assembly of the replication machinery. Two replication forks form and progress in opposite directions along the circular chromosome simultaneously, meeting approximately opposite the origin of replication at one of several replication termination (*ter*) sites.

In slow-growing bacteria, the cell cycle can be neatly divided into three parts: the B-period represents the time between the start of the cell cycle (completion of the previous round of cell division) and the initiation of DNA replication, the C-period represents the time between DNA replication initiation and termination, and the D-period represents the time after replication but before division. Experiments have shown that the C- and D-periods remain relatively constant across large ranges of cellular growth rates [172]. This means that for cells with doubling times shorter than $C + D$ (approximately 60 min in *E. coli*), the cell cycle must begin with one set of replication forks already progressing, and fast-growing cells must maintain multiple sets of replication forks simultaneously [173]. As a result, individual cells will have different copy numbers of each gene at different times, and asynchronous populations will exhibit cell-to-cell variations in gene copy numbers (see figure 7).

Simulating the effects of DNA replication is fairly straightforward. Each gene in the computational model gets assigned

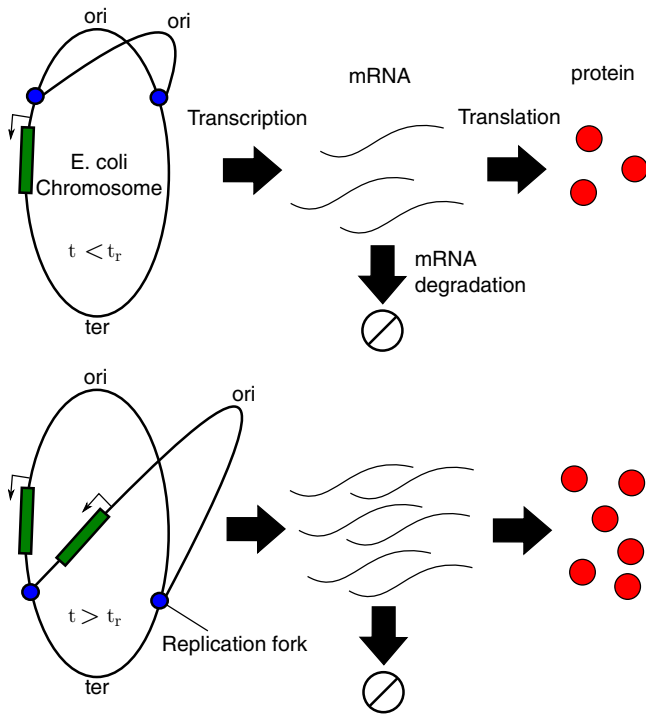


Figure 7. DNA replication induces variability in gene copy number. Early in the cell cycle ($t < t_r$), a given gene may exist in one copy, but later ($t > t_r$) it may exist in two copies. This gene copy number variability imparts a comparable level of protein noise as does variability in RNAP, ribosome, and ribonuclease copy numbers [80]. Reprinted figure with permission from [80], Copyright (2017) by the American Physical Society.

to it some replication time, t_r . For $t < t_r$, the gene exists in a low-copy number state (e.g. 1 for slow-growing cells, or 2, 4, etc for faster-growing cells); at $t = t_r$, the copy number is doubled, and remains so for the rest of the modeled cell cycle. Each t_r is a function of the gene's position on the chromosome, as well as the durations of the B- and C-periods, and the doubling time of the cell. Note that for slow-growing cells, $t_r = B + \chi C$ (where χ represents the position of the gene as a fraction from origin to terminus), meaning genes near the origin get replicated before genes closer to the terminus, but for faster growing cells, in which the cell cycle begins with a set of replication forks already progressing, genes close to the origin may start off in two copies and not double to four until late in the cell cycle.

The impact of variability in gene copy number on mRNA and protein statistics has been investigated experimentally and theoretically. In 2014, Jones *et al* [174] showed that DNA replication represents a major source of variability in mRNA concentrations [174]. Their results were refined in 2015 by Peterson *et al* [175], and then extended to protein variability by Cole and Luthey-Schulten [80] in 2017. All told, accounting for DNA replication can lead to a doubling or tripling (in slow- and fast-growing cells, respectively) of predicted mRNA Fano factors ($\text{Var}[m]/E[m]$) relative to analogous models that ignore the effect [175], and the contribution to protein noise associated with DNA replication is of comparable size to those associated with cell-to-cell variability in RNAP, ribosome, and ribonuclease E copy numbers [80].

3.4. Ribosome biogenesis: coupling transcription, translation, and ribosome assembly

With transcription and translation described, we are now in a position to discuss how ribosomes are formed. Ribosome biogenesis is a highly complex, tightly regulated process in the cell, where each 70S ribosome (composed of a 30S small subunit (SSU) and a 70S large subunit (LSU)) represents the end result of the coordinated transcription, translation, folding, and hierarchical assembly of three strands of ribosomal RNA (rRNA) and over four dozen ribosomal proteins (r-proteins). The SSU, tasked with binding and decoding mRNA, is composed of the 16S rRNA and 21 r-proteins, whereas the LSU is tasked with the synthesis of the nascent polypeptide chain and is composed of the 5S and 23S rRNA and 33 r-proteins.

Here we will differentiate assembly and biogenesis: assembly is the association of ribosomal proteins to rRNA, whereas biogenesis is the assembly process, as well as all necessary transcription and translation of the necessary components. Ribosomes appear to assemble *in vitro* [176–178] within hours, however in living cells the process concludes within minutes [179]. In bacteria, the ribosomes are not evenly distributed throughout the cell volume, but rather accumulate around the cell poles and inner membrane [56, 57, 67, 68, 180]. However, the disassociated large and small ribosomal subunits can be found anywhere in the cell with equal probability [68].

Ribosome biogenesis involves the cooperation of many molecular components. In bacteria the process involves the transcription of rRNA genes, the translation of r-protein, post-transcriptional processing and modification of both the rRNA and r-proteins, and assembly of the r-proteins and rRNA components to form the mature ribosomal subunits [181]. These processes all occur in parallel, perpetually throughout the cell cycle. The 54 r-proteins are classified by their order of binding. Primary proteins bind directly to the rRNA, secondary proteins require the presence of primary proteins in order to bind, and tertiary proteins require the presence of secondary proteins to bind. The r-proteins can compose 9–22% of the total protein counts in the cell [182, 183]. In addition, approximately 20 assembly cofactors are engaged to facilitate the process at various stages of the assembly process.

In 1966, Hosokawa *et al* [185] began the study of the assembly mechanism. Nomura and coworkers were the first to systematically study the order of r-protein assembly to the 16S rRNA to form the SSU [184]. By reconstituting the small subunit *in vitro* and measuring the equilibrium binding fractions of r-proteins, they found that the stability of binding an r-protein to an assembly intermediate can depend on which r-proteins are bound initially. They were able to infer an assembly hierarchy of binding dependencies, which is presented in figure 8(a). Later work in the field has shown through *in vitro* experiments that the assembly process can follow multiple pathways, which all follow a general binding order of 5'-domain r-proteins associating first, followed by the central-domain r-proteins, then finally the 3'-domain r-proteins [176–178, 186–191].

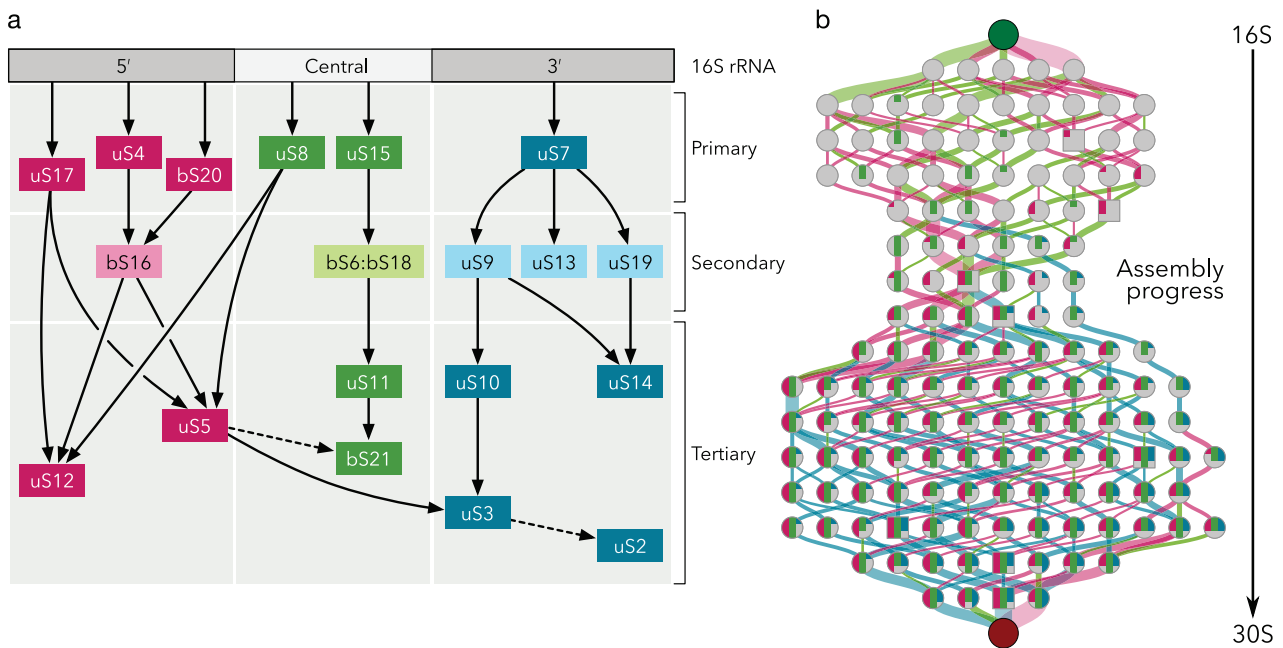


Figure 8. (a) The Nomura assembly map, which presents the thermodynamic protein binding dependencies to the 16S rRNA [184]. Arrows point from a protein to the protein that is dependent on it. (b) Assembly diagram describing the rRNA/protein association process at 40 °C [69]. Nodes represent specific rRNA/protein assembly intermediates, edges represent the binding of a ribosomal protein, where the color indicates the binding domain of that protein. The colored squares in each node indicate the proteins bound to that intermediate following the structure of panel (a). A filled square indicates that all proteins of the domain and binding order in the same position in the Nomura map are bound. Square nodes are intermediates which have been observed during *in vitro* assembly experiments [177].

Recently, Earnest *et al* [69] reported the first kinetic model of the assembly of the SSU, accounting for the association of 18 r-proteins to the 16S rRNA. A naïve approach to constructing such a kinetic model would consider the association of a protein to any possible protein/rRNA configuration—this would lead to $18!$ (6.4×10^{15}) reactions. Instead, the authors constructed the assembly reaction network by considering only r-protein association reactions which are consistent with the Nomura map. To parameterize the reaction network, rate parameters for two experimental conditions were inferred from pulse/chase mass spectrometry [176, 177]. The resulting kinetic models consisted of 1612 SSU assembly intermediates and 6997 protein/intermediate association reactions and were further reduced by removing the assembly intermediates from the network which contributed the least to the total assembly flux. This reduction resulted in simplified models with less than 150 assembly intermediates, which faithfully captured the topology of the r-protein/rRNA original (1612-intermediate) interaction network and reproduced the experimental protein binding kinetics described in [176] and [177] (figure 8(b)). Both models are consistent with an assembly mechanism inferred from cryo-EM of 30S assembly intermediates [177].

In order to model ribosome biogenesis, Earnest *et al* [69] integrated kinetic models of transcription and translation with the assembly model. The biogenesis model consisted of 251 species types: the SSU, LSU, rRNA, 18 ribosomal proteins, the ribosomal operons and associated mRNA, and over 140 SSU assembly intermediates. These species interact through approximately 1300 reactions: transcription, translation, assembly, and RNA degradation. Most reaction rates were taken directly from literature sources, with the exception of

the transcription rates which were chosen to produce 4500 ribosomes on average at steady state. Using cryo-electron tomography data describing slow-growing *E. coli* [67], they constructed a spatially resolved model using diffusion rates from the literature and simulated a full 120 min cell cycle using Lattice Microbes [92, 93]. Due to the relatively small number of 30S particles in the process of assembly and the large range of possible intermediates, the counts of individual 16S/r-protein configurations can be of the order of one per cell. Thus, to accurately describe the fluctuations due to small copy numbers, a spatially resolved representation accounting for the discreteness of chemical species was essential [192].

The model was further expanded in 2016 to account for DNA replication and cell division [118] in order to explore mRNA copy number variability associated with DNA replication. Cell cycle parameters, such as the delay between cell division and DNA replication initiation and the duration of replication, for the slow-growing *E. coli* were determined from a simple statistical model. This model described the probability to find a cell of a particular length with two copies of a gene, given the location of the gene in the genome. The model was fit to data from computed from images of a series of *E. coli* strains with fluorescently labeled genes distributed evenly throughout their genome. The cell geometry was allowed to vary over the cell cycle following the inferred cell cycle parameters. The ribosomal operons were duplicated and moved over the course of the simulation to include the effect of DNA replication. Though the cell growth and DNA replication were progressed deterministically by the simulation, not the underlying biochemistry, the work showed how ribosome number per unit volume remains constant over the

cell cycle, and paved the way for more complexly coupled whole cell models.

3.5. Metabolism: modeling how fluctuations in the microenvironment and gene expression affect metabolic behavior

Metabolism, the process by which chemical fuels are converted into energy, cellular building blocks, and other vital compounds (vitamins, cofactors, etc), represents something of a linchpin, tying essentially every other cellular process together. It produces the driving forces (by way of ATP, GTP, NADH, etc) and substrate pools (nucleic and amino acids) necessary to replicate the genome and express the plethora of RNA and proteins the cell needs to grow and divide. In so doing, metabolism involves the coordinated activity of hundreds or thousands of catalytic enzymes, each taking part in perhaps only a few related reactions whose reactants, products, and stoichiometry have been painstakingly elucidated through decades of biochemical research. Because of its complexity, modeling metabolism represents a major challenge in systems biology. In principle, any reaction network can be posed as a set of stochastic or deterministic differential equations; the problem with metabolism is that many of the enzymes involved are poorly characterized, and kinetic parameters can be scarce or vary wildly from study to study.

In the 1980s, researchers began to develop what would come to be known as flux balance analysis (FBA) as a way to partially circumvent the difficulty in parameterizing a kinetic description of metabolism. At its core, FBA is strikingly simple. Networks are described in terms of a stoichiometric matrix, \mathbf{S} , and a vector of (biomass-normalized) fluxes through each reaction, \mathbf{v} . It is assumed that there is no appreciable buildup or loss in the intracellular metabolite concentrations during growth, which leads to the ‘steady state’ requirement that:

$$\frac{d\mathbf{c}}{dt} = \mathbf{S} \cdot \mathbf{v} = 0. \quad (64)$$

Importantly, in addition to the metabolic reactions, the stoichiometric matrix also usually includes some sink reaction that consumes certain metabolites—amino acids, nucleic acids, lipids, etc—in their appropriate levels in order to produce a unit of biomass. The biomass reaction also includes a growth-associated ATP maintenance coefficient (GAM) which represents the ATP expenditure associated with peptide bond formation, RNA synthesis and recycling, and other uses of ATP required for growth. Non-growth associated ATP maintenance (NGAM) is also generally included as a $\text{ATP} \rightarrow \text{ADP} + \text{P}_i$ reaction in the stoichiometric matrix with a positive flux lower bound that can be fit to experimental data. By maximizing flux through this ‘biomass reaction’ subject to constraints placed on the fluxes through the metabolic reactions, a number of different environments and metabolic phenotypes can be modeled. This problem can be posed as the linear program,

$$\begin{aligned} &\text{maximize} && v_{\text{biomass}} \\ &\text{subject to} && \mathbf{S} \cdot \mathbf{v} = 0 \\ &&& \text{and } \mathbf{v}_{\text{l.b.}} \leq \mathbf{v} \leq \mathbf{v}_{\text{u.b.}}, \end{aligned} \quad (65)$$

where $\mathbf{v}_{\text{l.b.}}$ and $\mathbf{v}_{\text{u.b.}}$ represent vectors of lower- and upper-bounds for the reaction fluxes. For an outstanding introduction to FBA with an accompanying tutorial, see Orth *et al* [193].

What types of constraints are necessary to model metabolism in a realistic way? First, its important to describe the environment. If the model is intended to describe cells grown in aerobic glucose minimal media, then the fluxes that represent the uptake of oxygen, glucose, and salts should be limited to biologically reasonable values, while the fluxes that represent the uptake of other carbon sources not present in the media should be capped at zero. This type of ‘constraint-based modeling’ can be extremely powerful and extensible. Knock-out studies can be performed, for example, by constraining the flux through any reactions catalyzed by a given gene product to be zero. Other, more physically motivated constraints, can also be set. If a given reaction leads to a large change in Gibbs free energy, then it may be considered to be irreversible, and the lower-bound on the flux through it may be set to zero.

Extensions of this type of thermodynamic approach have been proposed and implemented on a number of occasions. Thermodynamics-based metabolic flux analysis (TMFA), first developed in 2007 [194], involved the detailed estimation of the free energy change in almost every reaction in the iJR904 model of *E. coli* (described in Reed *et al* [195]), and the requirement that it be negative for all reactions that carry non-zero flux. Research has continued along these lines, including the development of the ‘max-min driving force’ approach of Noor *et al* [196], which was used to predict specific metabolic reactions that pose a thermodynamic bottleneck, and what that means in terms of evolutionary pressures on the enzymes that catalyze those reactions. For example, the authors found the malate dehydrogenase was barely feasible at a pH of 7.5, and became thermodynamically infeasible below 7.0. They proposed that the high turnover rate (over 1000 reactions per second per enzyme molecule) may have evolved in order to compensate for the marginal ΔG_r of the reaction. Another particularly nice feature of these types of free energy-based approaches is that they naturally require some estimation of the metabolite concentrations [194, 196]—a detail missing in more traditional FBA approaches. Metabolome-wide Gibbs energy estimates have generally relied on experimental values and the group contribution method, but have recently been expanded to include values based on quantum chemistry calculations [197].

Constraints based on the structure and stability of enzymes have also been investigated. In 2013, Chang *et al* [198] investigated the thermotolerance of *E. coli* metabolism by using experimentally determined critical temperatures for the thermal denaturation of enzymes involved in an FBA model. The resulting simulations showed good agreement with experimental growth studies, and yielded new predictions regarding the specific enzymes that give rise to thermosensitivity [198].

Thermodynamics-based constraints are not the only physically motivated constraints that can be applied to FBA models. Limitations on reaction fluxes based on the volume or mass of the enzymes that catalyze the reactions have also been proposed. In general, these depend on physical and kinetic parameters of the enzymes, the latter of which can be estimated from literature values (which can often be found in resources like the BRENDA database [199]) and take the form:

$$\sum_{i=1}^{N_{\text{rxn}}} \frac{v_i}{k_i} n_i \leq N \quad (66)$$

where k_i represents the turnover rate of the enzyme that catalyzes reaction i (with units appropriately chosen such that v_i/k_i represents the number of enzymes per unit of biomass), n_i represents the volume or mass of each enzyme, and N represents the total volume or mass of catalytic proteins in a unit of biomass. These types of approaches have shown promise in understanding the origins of carbon catabolite repression, and the evolutionary forces that may have shaped the enzyme kinetics we observe today [200, 201].

The impact of stochastic gene expression on metabolic behavior has also been investigated using FBA. By sampling enzyme copy numbers from experimentally determined protein abundance distributions, Labhsetwar *et al* [42] simulated large numbers of individual *E. coli* cells, each with their own sets of metabolic constraints. This yielded a wide distribution of growth rates, and significant pathway level variability, including the emergence of subpopulations of cells that differentially used either of two glycolytic pathways (Emden–Meyerhof–Parnas or Entner–Doudoroff), and either the oxidative phosphorylation or acetate overflow pathways [42]. This approach has also recently been applied to the study of *S. cerevisiae* metabolism [44].

Finally, FBA has been leveraged to study the role of changing microenvironments in multicellular systems. By coupling an FBA description of metabolism and growth with a reaction–diffusion model of metabolite transport, Cole *et al* [202] modeled the spatial dependence of metabolic phenotypes in macroscopic (~ 1 mm) *E. coli* colonies. Their model recapitulated the experimentally measured penetration depth of oxygen into the colonies, the shape and radial expansion of the colonies, and predicted a previously unknown form of crossfeeding in which cells near the bottom of the colonies ferment glucose to acetate, and cells near the top consume the acetate [202]. This methodology will be described in greater detail in section 5.1.

4. Toward a whole cell model: uniting metabolism, transcription, translation, and DNA replication in the cell

4.1. Thermodynamics and free energies of the core metabolic network

Understanding the physical forces that determine the activity of a metabolic network requires a thermodynamic analysis of whether a given reaction or pathway is feasible. At or near equilibrium under physiological conditions (pH ~ 7 , ~ 300 K, and

~ 1 atm), the change in Gibbs free energy of a reaction is related to the reactant concentrations through the well-known equation,

$$\Delta G_r = \Delta G_r^\circ + RT \sum_{i=1}^{N_{\text{sp}}} S_i \ln c_i, \quad (67)$$

where c_i represents the concentration of species i , and S_i represents its stoichiometric coefficient in the given reaction. The free energy dissipated by a reaction can also be related to the fluxes and rate constants of the forward and backward reactions [203],

$$\Delta G_r = RT \ln \frac{J_r^-}{J_r^+}, \quad (68)$$

where the forward flux, $J_r^+ = k_r^+ \prod_i^{\text{reacts.}} c_i^{-S_{ir}}$, and the backward flux, $J_r^- = k_r^- \prod_j^{\text{prods.}} c_j^{S_{ij}}$ (see (13)), can be described in terms of their forward and backward rate constants, k_r^+ and k_r^- , respectively.

An enzyme catalyzing a reaction with a large negative ΔG_r has almost no backwards flux. In glycolysis, the process of converting glucose from the environment to ATP and pyruvate (an input to pathways generating further ATP), the reaction catalyzed by phosphofructokinase (PFK), which is coupled to ATP hydrolysis, dissipates some 20 kJ mol^{-1} , meaning less than one percent of the flux is in the reverse direction (thus it is considered an irreversible or ‘committed’ step; see figure 9(a)). Of course, not all glycolytic steps are so exergonic, but a few, including the reactions catalyzed by PFK and pyruvate kinase (PK), have rather large negative associated free energies. Because these reactions are so favorable, living cells have evolved mechanisms to tightly control the activity of the enzymes that catalyze them. In *E. coli*, for example, PFK I is allosterically activated by ADP and GDP, and inhibited by phosphoenolpyruvate (the penultimate metabolite in glycolysis) [204] such that the reaction rate is increased in conditions of low ATP, and decreased when the outputs of glycolysis cannot be consumed fast enough.

Glycolysis itself can be thought of as consisting of two parts—an upper part and a lower part. Upper glycolysis involves the stepwise conversion of glucose (a six carbon sugar) into two three carbon molecules of D-glyceraldehyde 3-phosphate (GAP). This process requires the consumption of two ATP molecules, the first being used to phosphorylate glucose (either directly through the use of hexokinase, or indirectly through the use of phosphoenolpyruvate via the phosphotransferase system in bacteria), and the second being used to phosphorylate β -D-fructose 6-phosphate to β -D-fructose 1,6-bisphosphate (carried out by PFK). The energetic cost of these ATP has led upper glycolysis to be thought of as the ‘investment’ phase of the pathway, with lower glycolysis making up the ‘pay-off’ phase.

Lower glycolysis converts GAP to pyruvate, but does so through steps that transform NAD^+ to NADH (via glyceraldehyde phosphate dehydrogenase (GAPDH)), and ADP to ATP (first via phosphoglycerate kinase (PGK) and again via PK). The net result of the upper and lower parts of the pathway is the formation of two ATP and two NADH for each glucose molecule.

concentrations of ATP, GTP, NADH, and NADPH (in *E. coli*, approximately 9.6, 4.9, 0.08, and 0.12 mM, respectively [207]).

Metabolic thermodynamics has recently been experimentally investigated by Park *et al* [207]. Using ^{13}C -labeled glucose, the reversibility, and in turn the free energy dissipated by many reactions in *E. coli*, *S. cerevisiae*, and mammalian central metabolism were inferred. The authors demonstrated their approach using triose phosphate isomerase (TPI); by labeling glucose carbons 1 and 2, the researchers showed that any unlabeled DHAP must have been generated via backward flux through TPI, rather than forward flux through fructose biphosphate aldolase (FBA) (see figure 9(b)). The ratio of labeled-to-unlabeled DHAP was then used to determine the ratio of forward-to-reverse TPI flux, and in turn, the ΔG of the reaction.

4.2. Kinetic model of core metabolic network

Based on the above considerations, a broadly applicable kinetic model of metabolism must be comprised not only of the known reactions, but also of many of the key regulatory connections between allosteric modulators and their targets. Without such regulation, a model whose parameters may be trained under one set of conditions (aerobic growth on glucose, for example), might have only limited predictive power under significantly different conditions (anaerobic culture, or growth on pyruvate or acetate etc).

A number of groups are actively developing kinetic models of metabolism [209–212] that account for regulation. Among them, Khodayari and Maranas [212] recently described a model of *E. coli* metabolism that encompasses 457 reactions among 337 metabolites and, importantly, includes almost 300 regulatory links. Their model was validated across a range of environmental conditions and genetic perturbations. Of particular interest, the Khodayari model breaks down complex reactions into multiple elementary reactions, fitting rate constants for each step. The enzyme-mediated reaction that transforms A–B, for example, might be written:



rather than a more compact approximate form, perhaps involving a Hill function (section 2.8). This can be extremely valuable when attempting stochastic simulations of the network in which the elementary reaction steps are the natural description of the system.

4.3. How cellular networks are linked

Marrying models of the universal cellular processes (transcription, translation, replication, ribosome biogenesis, and metabolism) into a comprehensive description of the living cell represents the holy grail of computational biophysics. From a practical standpoint, initial progress will likely require some simplification. Recent pioneering work in developing a synthetic ‘minimal’ cell—that is a cell with the smallest

number of genes necessary to sustain life—represents a unique opportunity in this respect. The current version of the minimal cell, based on *Mycoplasma mycoides* and dubbed JCVI-Syn3.0, has a genome comprised of only 473 genes (roughly a tenth of the *E. coli* genome) [213]. What makes the minimal cell so attractive from a modeling perspective is that many cellular networks are either greatly reduced or stripped out entirely. Metabolism, for example, involves only 165 enzymes (35% of the genome), and transcriptional regulation is almost nonexistent through the removal of most genes coding for transcription factors. Nevertheless, given an extremely rich growth medium, the minimal cell maintains a doubling time of around 3 h.

Biologically, the universal networks are linked in fairly obvious ways. Transcription and replication require pools of nucleic acids, translation requires pools of aminoacylated tRNAs, ribosome biogenesis requires both, and all of them require sufficient ATP and GTP to proceed. Estimates based on the chemical makeup of the cell indicate that roughly 70, 13, and 2% of available ATP is used in protein translation, maintaining mRNA pools, and chromosome replication, respectively [206, 214, 215]. These estimates generally involve the summation of ATP molecules used to carry out a given process. For example, in order to incorporate a single amino acid into a growing peptide chain, one ATP is hydrolyzed to AMP (roughly equivalent to hydrolyzing two ATP to ADP) in order to charge the amino acid, and two GTP (roughly equivalent to two ATP) are hydrolyzed during peptide bond formation, for a total of four ATP per amino acid. Importantly, this does not include the ATP expenditure associated with error checking and the rejection of incorrect aminoacylated tRNAs, for which estimates range up to around two ATP per peptide bond. Similarly, the glycolytic breakdown of glucose to lactate and acetate—the main ATP-generating pathway in the minimal cell—requires the transcription and translation of around a dozen enzymes. This means that the products of metabolism are directly used in transcription, translation, and replication, and vice versa. While these types of connections are easy to understand, they pose several challenges from a computational standpoint. One of the most important is that concentrations of different chemical species can span a wide range from just a few per cell (on the order of a nanomolar) in the case of mRNA and some proteins, all the way up to tens of millimolar in the case of some amino acids, sugars, and ATP [207]. Obviously the low-concentration species, which will be subject to appreciable variability, will require a stochastic description, but the high-concentration species might be better modeled using a deterministic approach in which the individual reaction events that consume or produce each molecule are averaged out. Coupled stochastic and deterministic modeling approaches have been implemented for a variety of systems, including the metabolism of clusters of cells [216], a model of *Mycoplasma genitalium* [217], and recently the mechanochemical dynamics of actin filaments [218]. Modeling the minimal cell will likely require a deterministic kinetic model of metabolism coupled with stochastic models of transcription, translation, ribosome biogenesis, and DNA replication.

5. Physics of 3D cellular communities

Cells do not exist in a vacuum. From isogenic colonies to complex biofilms and tissues, cells often live in dense consortia, interacting with each other in a multitude of ways. They can mechanically interact—like platelets adhering to each other during blood coagulation; they can interact through the directed production and dispersal of chemical messengers like toxins or signaling molecules—agents made at some cost to the individual cells but that serve the needs of the community; and they can interact passively, simply through the impact their consumption of resources and production of waste has on their microenvironment. These passive interactions can be surprisingly strong, driving different subpopulations of cells toward divergent modes of metabolism. While FBA alone can describe the average metabolic behavior of cells in homogeneous environments, and extensions like population FBA offer insight into how heterogeneity in the internal constituents of cells can affect their behavior, it is only recently that FBA has been applied to the study of how nonuniform external conditions can affect cellular metabolism and community dynamics.

While most of what follows will be devoted to the study of metabolism in multicellular systems, we note that this represents only one aspect of the broader field of multicellular modeling. Methods have been developed for simulating different aspects of the development of biofilms, tumors, and even embryos [219–222], and recently, stochastic models of the expression and intercellular exchange of proteins have begun to appear [223].

5.1. Spatially resolved FBA

In 2013, Cole *et al* [216] proposed the coupling of an FBA model of bacterial metabolism with an RDME model of metabolite transport within and around a clusters of cells [216]. Although the single-molecule resolution of the original implementation limited it to relatively low concentrations of metabolites and fairly small volumes (containing around 100 cells), these early simulations showed that FBA could be used to model the steep nutrient gradients that emerge within microbial colonies. Over the next two years, the general approach would be refined to the point where simulations of macroscopic colonies under realistic laboratory conditions could be accomplished [202, 224]. The basic method, called three-dimensional dynamic flux balance analysis (3DdFBA), can be represented as the solution to

$$\frac{\partial \phi_i(\mathbf{r}, t)}{\partial t} = \nabla \cdot [D_i(\mathbf{r}, \rho(\mathbf{r}, t)) \nabla \phi_i(\mathbf{r}, t)] + \rho(\mathbf{r}, t) M_i(\phi(\mathbf{r}, t)) \quad (70a)$$

$$\frac{\partial \rho(\mathbf{r}, t)}{\partial t} = \rho(\mathbf{r}, t) M_{\text{gr}}(\phi(\mathbf{r}, t)) + T(\rho(\mathbf{r}, t)), \quad (70b)$$

where $\phi_i(\mathbf{r}, t)$ represents the concentration of metabolite i (e.g. oxygen, glucose, and metabolic byproducts), and $\rho(\mathbf{r}, t)$ represents the density of cells. Obviously, the metabolites can diffuse as described by the space and density dependent diffusion coefficient $D_i(\mathbf{r}, \rho(\mathbf{r}, t))$, but they can also react through the local metabolic activity of the cells in each region

of space. The reaction term, $M_i(\phi(\mathbf{r}, t))$, is dependent on the local metabolite concentrations and is computed using FBA. The local cell density grows exponentially at rate $M_{\text{gr}}(\phi(\mathbf{r}, t))$, which corresponds to the biomass term in the FBA model, and can undergo some sort of transport, $T(\rho(\mathbf{r}, t))$, which expands the colony. In practice, this cell transport is usually assumed to be either diffusive [224], or result from cells ‘pushing’ outward as the local cell density grows beyond some maximal packing fraction [202] (see figure 10(a)). The majority of the parameters used in this model were determined from experimental data, however the maximum oxygen uptake rate was chosen to ensure that the carbon sources were utilized at their maximum uptake rates.

Using their own implementations of this spatially resolved dynamic FBA strategy, Harcombe *et al* [224] investigated two- and three-species codependent consortia in 2014, and Cole *et al* [202] showed that syntrophic metabolic behaviors can emerge even within isogenic populations in 2015 (see figure 10(b)).

5.2. Emergence of metabolic cooperativity

One of the strengths of using FBA in a spatiotemporal framework is that it enables modelers to investigate the interplay between neighboring cells in dense communities. It has long been known, for example, that oxygen is utilized so rapidly by *E. coli* that its penetration depth into a colony is only $\sim 50 \mu\text{m}$ [225]. This means that much of the bulk of the colony must be confined to an anaerobic metabolic state. What was unknown until recently was that the shallow oxygen penetration was also one of the driving forces behind an emergent form of cooperative crossfeeding [202].

Within an *E. coli* colony growing on a nutrient-rich substrate—like the glucose-infused agar commonly used in laboratories—cells near the bottom have access to food but little oxygen, while cells on the top have access to oxygen but little food. This drives the bottom of the colony into a fermentative mode of metabolism in which they produce formate, acetate, and other byproducts. Some of these byproducts, specifically acetate, can filter up to the top of the colony, where they can be aerobically catabolized by the cells near the top (see figure 10(b)). This crossfeeding was found to be extremely robust across variations in the shape and texture of the agar, its glucose concentrations, and across several common laboratory *E. coli* strains [226].

5.3. Future: metabolic reprogramming in tumor formation

One of the hallmarks of cancer is the reprogramming of metabolism toward increased proliferation rates [227]—analogous to the evolutionary pressure on microbial species to maximize their growth rate. The associated increase in metabolic rates carries with it another cancer hallmark: anoxia within the tumor and an associated enhancement in proangiogenic signaling [227]. The oxygen and nutrient gradients that form within a tumor have been shown to drive a form of lactate crossfeeding that bears a striking resemblance to the acetate crossfeeding present within *E. coli* colonies [202, 228–230]. Indeed, it has been noted by several authors that the dynamics

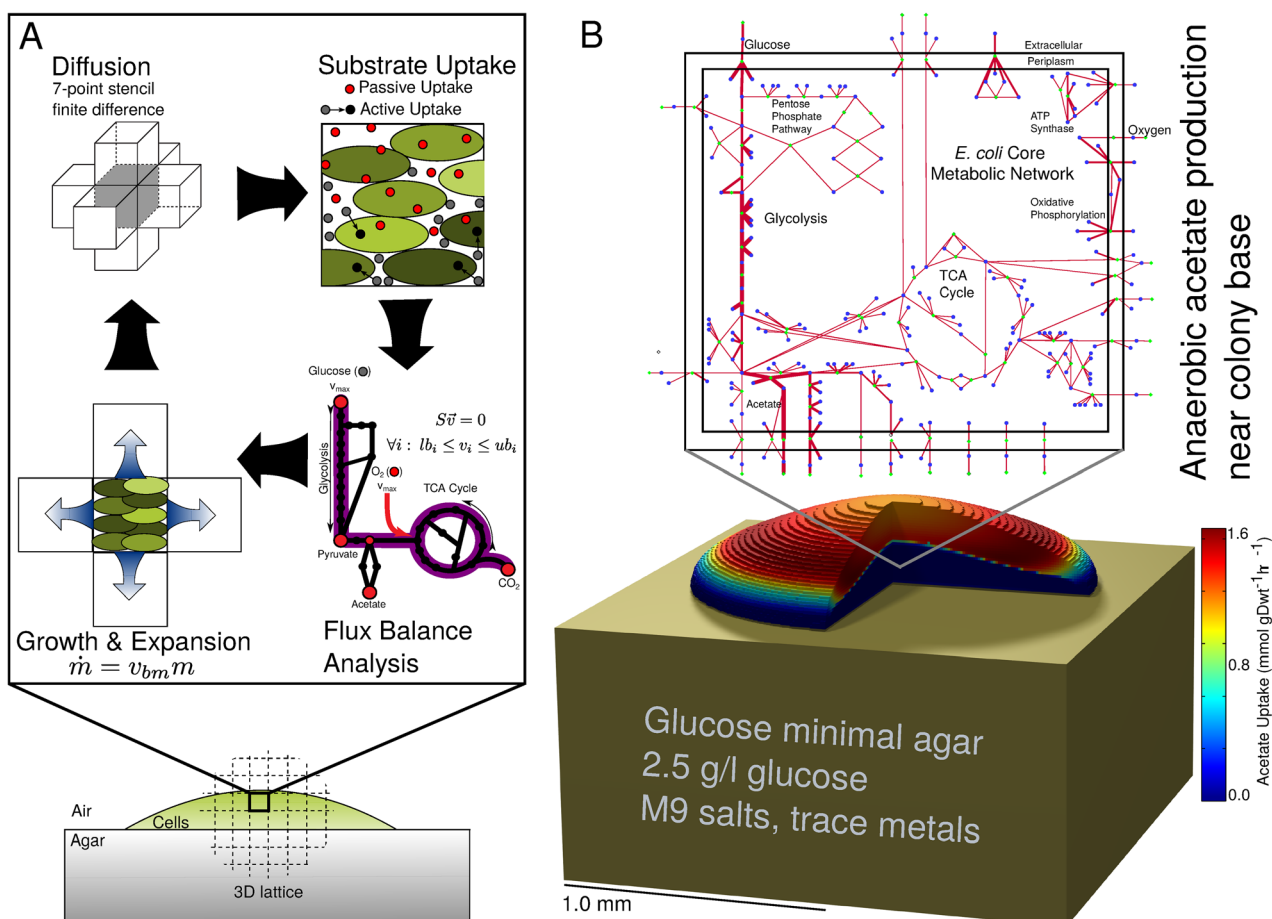


Figure 10. Simulating the metabolism and growth of a bacterial colony. (a) 3DdFBA methodology: the simulation volume is discretized into a 3D cubic lattice. In each voxel, the diffusion of metabolites is modeled using a seven-point finite difference scheme. Substrates can be either passively or actively imported. Metabolism is modeled using flux balance analysis. Growth and expansion of the colony is modeled assuming cells grow exponentially, and that they begin to spill out into neighboring lattice sites as their local packing fraction surpasses some critical value (~ 0.65) [202]. (b) Results of a simulation; *E. coli* colonies separate into anaerobic acetate producers near the base, and aerobic acetate consumers on top [202].

of microbial populations may offer some insight into the dynamics of a growing tumor [231–233]. This suggests that spatially resolved FBA methods like that of [202] may be leveraged to help understand tumor metabolism. Importantly, numerous human tissue- and cancer-specific FBA models are under active development by the community (see Nilsson and Nielsen [230] for an excellent review). These may be integrated in various combinations in order to construct models of growing tumors in the context of surrounding healthy tissues, potentially giving insight into emergent forms of crossfeeding both within the tumor and between it and nearby tissue.

6. Outlook

The ultimate goal of constructing whole-cell simulations is to describe life in the language of physics. The vocabulary is, of course, well known—we speak of thermodynamic potentials, molecular conformations, chemical transformations and reactions—but the grammar, the system of integrating methods that span vastly different time and length scales, and whose applicabilities are often mutually exclusive, into a cohesive

picture of the living cell remains elusive. This final section is intended to outline in broad terms our expectations for the future of the field.

A few things are clear. First, the types of coarse-graining that are necessary to describe living systems over cell-cycle time scales—namely RDME—are not by themselves sufficient. They leave too much unanswered; a coarse-grained simulation may indicate that two molecules interact, but it will take finer-grained methods like MD or BD to describe how they interact, and hybrid techniques like quantum mechanics/molecular mechanics (QM/MM) to describe their chemical reactivity. The problem is that these atomic-scale methods are considerably more costly computationally, describing motions on femtosecond time scales and Ångstrom length scales. The hope is that the two descriptions (fine and coarse) can be made to meet in the middle—to ‘overlap for a microsecond’. The coarse-grained models might capture the dynamics of a cell over long periods, and interesting intermediate states—DNA melting during replication initiation, polymerization of FtsZ prior to division, etc—may serve as starting points for further exploration using MD, BD, and QM/MM. The question is what would it take to get there?

All atom MD and coarse-grained BD simulations have contributed enormously to our understanding of the molecular mechanisms that underpin the living cell [234–239]. Both of these methods rely on force fields derived in large part from quantum chemical studies of bonded and non-bonded interactions within biomolecules and complexes. The importance of the MD simulations to structural biology is reflected in the 2013 Nobel Prize in Chemistry, awarded to Martin Karplus, Michael Levitt and Ariel Warshel for their development of multiscale models for complex chemical systems. MD and BD have facilitated a dynamic understanding of the structures of many large macromolecules, including the ribosome and ATP synthase (the structures of each of which also led to Nobel prizes). While there has been a steady refinement of the force fields and their associated dynamics, due in large part to protein folding studies that compare simulations and experiments, limitations on the number of atoms that can be simulated, and the duration of those simulations, remain a hurdle. What is required is faster scalable MD and BD codes that are capable of simulating billions of atoms for long times. NAMD [237], one of the most widely used supercomputer applications in the world, recently completed tests of a billion-atom MD simulation (a $10 \times 10 \times 10$ array of satellite tobacco mosaic viruses for a few picoseconds). For the first time, this puts short simulations of an entire minimal cell—on the order of a few billion atoms—within reach.

At the other end of the spectrum, RDME simulations of a minimal cell, which are in principle possible using today's codes, represent another type of challenge. To describe a cell is not just to describe the approximate locations of every particle, but to describe their reactions and interactions as well. Progress continues to be made in constructing and parameterizing the major cellular reaction networks—metabolism, transcription, translation, replication, and division—but combining them into a unified cohesive description continues to be a challenge. As noted in section 4.3, progress will likely come through the development of hybrid stochastic–deterministic methods, treating, as examples, the random transcription and translation of enzymes stochastically, and their role in catalyzing metabolic reactions involving high-concentration species deterministically.

Finally, it should not go without saying that at every scale, biological models are generally based (albeit sometimes loosely) on experimentally determined parameters. Development of these coarse-grained methods will place greater demands on experimentalists for the quantitative data necessary to parameterize the models. These include everything from vibrational spectra and crystallographic data for atomistic models, to reaction rates and -omics data for kinetic models. There can be no substitute for the systematic experimental determination and curation of these types of data. The current generation of scientists is incredibly fortunate in this regard. We have at our disposal robust and well-tested molecular force-fields [240, 241], and databases that include tens-of-thousands of atomic-resolution structures [242] and enzyme kinetic rates [199].

Data from -omics experiments are essential for the design and validation of whole-cell models, however generally they can only provide the mean abundances. Recent developments in microfluidics have made single-cell high-throughput

-omics experiments possible [243]. These experiments measure the distributions of species abundances, which will be necessary to parameterize or validate these stochastic models with inherent cell-to-cell variability. Most experimentally derived kinetic parameters used in current models are derived from *in vitro* conditions, however there can be significant differences in enzyme activity compared to *in vivo*. For example, a deterministic, well-mixed model of glycolysis in *S. cerevisiae* originally parameterized with *in vitro* derived parameters exhibited qualitatively wrong steady-state behavior under glucose-limited conditions [244], yet kinetic parameters measured under conditions closer to *in vivo* yielded a model having the right behavior, and was predictive over the five experimental conditions tested [245]. The development of high-throughput measurement techniques to quantify the kinetic parameters of metabolic enzymes would be especially helpful in the development of whole-cell models.

Obviously, there remains a great deal of work ahead before even a minimal cell can be accurately described on all scales. But with ongoing technical advances making molecular simulations larger and faster, and making hybrid RDME simulations more comprehensive, a true *in silico* cell may soon be within our grasp.

Acknowledgments

This work is supported by the National Science Foundation (NSF) grants MCB-1244570 (TME, ZLS), Center for the Physics of Living Cells: PHY-1430124 (TME, ZLS), Physics of Living Systems Student Research Network: PHY-1505008 (JAC, ZLS), the National Institutes of Health grants 4 P41 GM104601-27 (JAC, ZLS) and 5 R01 GM112659-03 (ZLS), the U.S. Department of Energy, Office of Science, Biological and Environmental Research as part of the Adaptive Biosystems Imaging Scientific Focus Area (TME). This work was made possible, in part, by resources from the National Center for Supercomputing Applications (TME).

ORCID iDs

Tyler M Earnest  <https://orcid.org/0000-0002-1630-0791>

Zaida Luthey-Schulten  <https://orcid.org/0000-0001-9749-8367>

References

- [1] Muller H J 1922 Variation due to change in the individual gene *Am. Naturalist* **56** 32–50
- [2] Zimmer K, Timofeev-Ressovsky N and Delbrück M 1935 *Über die Natur der Genmutation und der Genstruktur* (Berlin: Weidmannsch Buchhandlung)
- [3] Schrödinger E and Roger P 1992 *What Is Life?* (Cambridge: Cambridge University Press) (<https://doi.org/10.1017/CBO9781139644129>)
- [4] Turing A M 1990 The chemical basis of morphogenesis *Bull. Math. Biol.* **52** 153–97
- [5] Watson J D and Crick F H 1953 Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid *Nature* **171** 737–8

- [6] Kendrew J C, Bodo G, Dintzis H M, Parrish R, Wyckoff H and Phillips D C 1958 A three-dimensional model of the myoglobin molecule obtained by x-ray analysis *Nature* **181** 662–6
- [7] Woese C R and Fox G E 1977 Phylogenetic structure of the prokaryotic domain: The primary kingdoms *Proc. Natl Acad. Sci.* **74** 5088–90
- [8] Bigger J W 1944 Treatment of staphylococcal infections with penicillin by intermittent sterilisation *Lancet* **244** 497–500
- [9] Rahn O 1932 A chemical explanation of the variability of the growth rate *J. Gen. Physiol.* **15** 257–77
- [10] Delbrück M 1945 The burst size distribution in the growth of bacterial viruses (bacteriophages) *J. Bacteriol.* **50** 131
- [11] Benzer S 1953 Induced synthesis of enzymes in bacteria analyzed at the cellular level *Biochim. Biophys. Acta* **11** 383–95
- [12] Novick A and Weiner M 1957 Enzyme induction as an all-or-none phenomenon *Proc. Natl Acad. Sci.* **43** 553–66
- [13] Maloney P C and Rotman B 1973 Distribution of suboptimally induced β -d-galactosidase in *Escherichia coli*: the enzyme content of individual cells *J. Mol. Biol.* **73** 77–91
- [14] Spudich J L and Koshland D E 1976 Non-genetic individuality: chance in the single cell *Nature* **262** 467–71
- [15] Ko M S 1992 Problems and paradigms: induction mechanism of a single gene molecule: Stochastic or deterministic? *Bioessays* **14** 341–6
- [16] Heitzler P and Simpson P 1991 The choice of cell fate in the epidermis of *Drosophila* *Cell* **64** 1083–92
- [17] Peccoud J and Ycart B 1995 Markovian modeling of gene-product synthesis *Theor. Population Biol.* **48** 222–34
- [18] McAdams H H and Arkin A 1997 Stochastic mechanisms in gene expression *Proc. Natl Acad. Sci.* **94** 814–9
- [19] Fiering S, Whitelaw E and Martin D I 2000 To be or not to be active: the stochastic nature of enhancer action *Bioessays* **22** 381–7
- [20] Elowitz M B, Levine A J, Siggia E D and Swain P S 2002 Stochastic gene expression in a single cell *Science* **297** 1183–6
- [21] Choi P J, Cai L, Frieda K and Xie X S 2008 A stochastic single-molecule event triggers phenotype switching of a bacterial cell *Science* **322** 442–6
- [22] Friedman N, Cai L and Xie X S 2006 Linking stochastic dynamics to population distribution: an analytical framework of gene expression *Phys. Rev. Lett.* **97** 168302
- [23] Shahrezaei V and Swain P S 2008 Analytical distributions for stochastic gene expression *Proc. Natl Acad. Sci.* **105** 17256–61
- [24] Taniguchi Y, Choi P J, Li G-W, Chen H, Babu M, Hearn J, Emili A and Xie X S 2010 Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells *Science* **329** 533–8
- [25] Dénervaud N, Becker J, Delgado-Gonzalo R, Damay P, Rajkumar A S, Unser M, Shore D, Naef F and Maerkl S J 2013 A chemostat array enables the spatio-temporal analysis of the yeast proteome *Proc. Natl Acad. Sci.* **110** 15842–7
- [26] Thattai M and van Oudenaarden A 2001 Intrinsic noise in gene regulatory networks *Proc. Natl Acad. Sci.* **98** 8614–9
- [27] Ozbudak E M, Thattai M, Kurtser I, Grossman A D and van Oudenaarden A 2002 Regulation of noise in the expression of a single gene *Nat. Genet.* **31** 69
- [28] Swain P S, Elowitz M B and Siggia E D 2002 Intrinsic and extrinsic contributions to stochasticity in gene expression *Proc. Natl Acad. Sci.* **99** 12795–800
- [29] Golding I, Paulsson J, Zawilski S M and Cox E C 2005 Real-time kinetics of gene activity in individual bacteria *Cell* **123** 1025–36
- [30] Kærn M, Elston T C, Blake W J and Collins J J 2005 Stochasticity in gene expression: from theories to phenotypes *Nat. Rev. Genet.* **6** 451
- [31] Raj A, Peskin C S, Tranchina D, Vargas D Y and Tyagi S 2006 Stochastic mRNA synthesis in mammalian cells *PLoS Biol.* **4** e309
- [32] Schultz D, Jacob E B, Onuchic J N and Wolynes P G 2007 Molecular level stochastic model for competence cycles in *Bacillus subtilis* *Proc. Natl Acad. Sci.* **104** 17582–7
- [33] Schultz D, Onuchic J N and Wolynes P G 2007 Understanding stochastic simulations of the smallest genetic networks *J. Chem. Phys.* **126** 245102
- [34] Acar M, Mettetal J T and van Oudenaarden A 2008 Stochastic switching as a survival strategy in fluctuating environments *Nat. Genet.* **40** 471–5
- [35] Boeger H, Griesenbeck J and Kornberg R D 2008 Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription *Cell* **133** 716–26
- [36] Wang J, Xu L and Wang E 2008 Potential landscape and flux framework of nonequilibrium networks: robustness, dissipation and coherence of biochemical oscillations *Proc. Natl Acad. Sci.* **105** 12271–6
- [37] MacNeil L T and Walhout A J 2011 Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression *Genome Res.* **21** 645–57
- [38] So L-H, Ghosh A, Zong C, Sepúlveda L A, Segev R and Golding I 2011 General properties of transcriptional time series in *Escherichia coli* *Nat. Genet.* **43** 554–60
- [39] Hensel Z, Feng H, Han B, Hatem C, Wang J and Xiao J 2012 Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis *Nat. Struct. Mol. Biol.* **19** 797–802
- [40] Lu M, Onuchic J and Ben-Jacob E 2014 Construction of an effective landscape for multistate genetic switches *Phys. Rev. Lett.* **113** 078102
- [41] Li C and Wang J 2014 Landscape and flux reveal a new global view and physical quantification of mammalian cell cycle *Proc. Natl Acad. Sci.* **111** 14130–5
- [42] Labhsetwar P, Cole J A, Roberts E, Price N D and Luthey-Schulten Z A 2013 Heterogeneity in protein expression induces metabolic variability in a modeled *Escherichia coli* population *Proc. Natl Acad. Sci.* **110** 14006–11
- [43] Levy S F, Ziv N and Siegal M L 2012 Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant *PLoS Biol.* **10** e1001325
- [44] Labhsetwar P, Melo M C R, Cole J and Luthey-Schulten Z 2017 Population FBA predicts metabolic phenotypes in yeast *PLoS Comput. Biol.* **13** e1005728
- [45] Moerner W E and Kador L 1989 Optical detection and spectroscopy of single molecules in a solid *Phys. Rev. Lett.* **62** 2535
- [46] Klar T A, Jakobs S, Dyba M, Egner A and Hell S W 2000 Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission *Proc. Natl Acad. Sci.* **97** 8206–10
- [47] Betzig E, Patterson G H, Sougrat R, Lindwasser O W, Olenych S, Bonifacino J S, Davidson M W, Lippincott-Schwartz J and Hess H F 2006 Imaging intracellular fluorescent proteins at nanometer resolution *Science* **313** 1642–5
- [48] Hess S T, Girirajan T P and Mason M D 2006 Ultra-high resolution imaging by fluorescence photoactivation localization microscopy *Biophys. J.* **91** 4258–72
- [49] Rust M J, Bates M and Zhuang X 2006 Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM) *Nat. Methods* **3** 793–6
- [50] Huang B, Bates M and Zhuang X 2009 Super-resolution fluorescence microscopy *Ann. Rev. Biochem.* **78** 993–1016
- [51] Balzarotti F, Eilers Y, Gwosch K C, Gynnå A H, Westphal V, Stefani F D, Elf J and Hell S W 2016 Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes *Science* **355** 606–12

- [52] Park S, Zhang J, Reyer M and Fei J 2017 Theme 5: Quantitative imaging & cell simulation of small regulatory RNA *Center for the Physics of Living Cells, Summer school (University of Illinois Urbana-Champaign, Urbana, IL, USA, 2017)*
- [53] Fei J, Singh D, Zhang Q, Park S, Balasubramanian D, Golding I, Vanderpool C K and Ha T 2015 Determination of *in vivo* target search kinetics of regulatory noncoding RNA *Science* **347** 1371–4
- [54] Yildiz A, Tomishige M, Vale R D and Selvin P R 2004 Kinesin walks hand-over-hand *Science* **303** 676–8
- [55] Kural C, Kim H, Syed S, Goshima G, Gelfand V I and Selvin P R 2005 Kinesin and dynein move a peroxisome *in vivo*: a tug-of-war or coordinated movement? *Science* **308** 1469–72
- [56] Bakshi S, Siryaporn A, Goulian M and Weisshaar J C 2012 Superresolution imaging of ribosomes and RNA polymerase in live *Escherichia coli* cells *Mol. Microbiol.* **85** 21–38
- [57] Wang W, Li G-W, Chen C, Xie X S and Zhuang X 2011 Chromosome organization by a nucleoid-associated protein in live bacteria *Science* **333** 1445–9
- [58] Moffitt J R, Pandey S, Boettiger A N, Wang S and Zhuang X 2016 Spatial organization shapes the turnover of a bacterial transcriptome *eLife* **5** e13065
- [59] Roy R, Hohng S and Ha T 2008 A practical guide to single-molecule FRET *Nat. Methods* **5** 507–16
- [60] Zimmerman S B and Minton A P 1993 Macromolecular crowding: biochemical, biophysical and physiological consequences *Ann. Rev. Biophys. Biomol. Struct.* **22** 27–65
- [61] Luby-Phelps K 1994 Physical properties of cytoplasm *Curr. Opin. Cell Biol.* **6** 3–9
- [62] Luby-Phelps K 1999 Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area *Microcompartmentation and Phase Separation in Cytoplasm (Int. Review of Cytology)* (San Diego, CA: Academic) pp 189–221
- [63] Lučić V, Rigort A and Baumeister W 2013 Cryo-electron tomography: the challenge of doing structural biology *in situ* *J. Cell Biol.* **202** 407–19
- [64] Mahamid J, Pfeffer S, Schaffer M, Villa E, Danev R, Cuellar L K, Forster F, Hyman A A, Plitzko J M and Baumeister W 2016 Visualizing the molecular sociology at the HeLa cell nuclear periphery *Science* **351** 969–72
- [65] Beck F *et al* 2012 Near-atomic resolution structural model of the yeast 26S proteasome *Proc. Natl Acad. Sci.* **109** 14870–5
- [66] Hammar P, Leroy P, Mahmutovic A, Marklund E G, Berg O G and Elf J 2012 The lac repressor displays facilitated diffusion in living cells *Science* **336** 1595–8
- [67] Roberts E, Magis A, Ortiz J O, Baumeister W and Luthey-Schulten Z 2011 Noise contributions in an inducible genetic switch: a whole-cell simulation study *PLoS Comput. Biol.* **7** e1002010
- [68] Sanamrad A, Persson F, Lundius E G, Fange D, Gynna A H and Elf J 2014 Single-particle tracking reveals that free ribosomal subunits are not excluded from the *Escherichia coli* nucleoid *Proc. Natl Acad. Sci.* **111** 11413–8
- [69] Earnest T M, Lai J, Chen K, Hallock M J, Williamson J R and Luthey-Schulten Z 2015 Toward a whole-cell model of ribosome biogenesis: kinetic modeling of SSU assembly *Biophys. J.* **109** 1117–35
- [70] Fange D and Elf J 2006 Noise-induced Min phenotypes in *E. coli* *PLoS Comput. Biol.* **2** e80
- [71] Halatek J and Frey E 2012 Highly canalized MinD transfer and MinE sequestration explain the origin of robust MinCDE-protein dynamics *Cell Rep.* **1** 741–52
- [72] Lawson M J, Drawert B, Khammash M, Petzold L and Yi T-M 2013 Spatial stochastic dynamics enable robust cell polarization *PLoS Comput. Biol.* **9** e1003139
- [73] Chandrasekhar S 1943 Stochastic problems in physics and astronomy *Rev. Mod. Phys.* **15** 1–89
- [74] Gillespie D 1992 *Markov Processes: an Introduction for Physical Scientists* (Boston, MA: Academic)
- [75] van Kampen N G 2007 *Stochastic Processes in Physics and Chemistry* 3rd edn (Amsterdam: North-Holland)
- [76] Gardiner C 2009 *Stochastic Methods: a Handbook for the Natural and Social Sciences (Springer Series in Synergetics vol 13)* 4th edn (Berlin: Springer)
- [77] Arnaut L G, Formosinho S J and Burrows H 2006 *Chemical Kinetics: from Molecular Structure to Chemical Reactivity* (Amsterdam: Elsevier)
- [78] Gillespie D T 2009 A diffusional bimolecular propensity function *J. Chem. Phys.* **131** 164109
- [79] Gillespie D T 1992 A rigorous derivation of the chemical master equation *Physica A* **188** 404–25
- [80] Cole J A and Luthey-Schulten Z 2017 Careful accounting of extrinsic noise in protein expression reveals correlations among its sources *Phys. Rev. E* **95** 062418
- [81] Gillespie D T 1977 Exact stochastic simulation of coupled chemical reactions *J. Phys. Chem.* **81** 2340–61
- [82] Gillespie D T 1976 A general method for numerically simulating the stochastic time evolution of coupled chemical reactions *J. Comput. Phys.* **22** 403–34
- [83] Gibson M A and Bruck J 2000 Efficient exact stochastic simulation of chemical systems with many species and many channels *J. Phys. Chem. A* **104** 1876–89
- [84] Ramaswamy R, Gonzalez-Segredo N and Sbalzarini I F 2009 A new class of highly efficient exact stochastic simulation algorithms for chemical reaction networks *J. Chem. Phys.* **130** 244104
- [85] Indurkha S and Beal J 2010 Reaction factoring and bipartite update graphs accelerate the Gillespie algorithm for large-scale biochemical systems *PLoS One* **5** e8125
- [86] Ramaswamy R and Sbalzarini I F 2010 A partial-propensity variant of the composition-rejection stochastic simulation algorithm for chemical reaction networks *J. Chem. Phys.* **132** 044102
- [87] Cao Y, Li H and Petzold L 2004 Efficient formulation of the stochastic simulation algorithm for chemically reacting systems *J. Chem. Phys.* **121** 4059
- [88] McCollum J M, Peterson G D, Cox C D, Simpson M L and Samatova N F 2006 The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior *Comput. Biol. Chem.* **30** 39–49
- [89] Gillespie D T 2001 Approximate accelerated stochastic simulation of chemically reacting systems *J. Chem. Phys.* **115** 1716
- [90] Rathinam M, Petzold L R, Cao Y and Gillespie D T 2003 Stiffness in stochastic chemically reacting systems: the implicit tau-leaping method *J. Chem. Phys.* **119** 12784
- [91] Schnoerr D, Sanguinetti G and Grima R 2017 Approximation and inference methods for stochastic biochemical kinetics—a tutorial review *J. Phys. A: Math. Theor.* **50** 093001
- [92] Roberts E, Stone J E and Luthey-Schulten Z 2013 Lattice Microbes: high-performance stochastic simulation method for the reaction–diffusion master equation *J. Comput. Chem.* **3** 245–55
- [93] Hallock M J, Stone J E, Roberts E, Fry C and Luthey-Schulten Z 2014 Simulations of reaction diffusion processes over biologically-relevant size and time scales using multi-GPU workstations *Parallel Comput.* **40** 86–99
- [94] Hattne J, Fange D and Elf J 2005 Stochastic reaction–diffusion simulation with MesoRD *Bioinformatics* **21** 2923–4
- [95] Drawert B, Engblom S and Hellander A 2012 URDME: a modular framework for stochastic simulation of reaction-transport processes in complex geometries *BMC Syst. Biol.* **6** 76

- [96] Roberts E, Stone J E, Sepulveda L, Hwu W-M W and Luthey-Schulten Z 2009 Long time-scale simulations of *in vivo* diffusion using GPU hardware *IEEE Int. Symp. on Parallel & Distributed Processing* (<https://doi.org/10.1109/ipdps.2009.5160930>)
- [97] Isaacson S A 2009 The reaction–diffusion master equation as an asymptotic approximation of diffusion to a small target *SIAM J. Appl. Math.* **70** 77–111
- [98] Isaacson S A and Isaacson D 2009 Reaction–diffusion master equation, diffusion-limited reactions and singular potentials *Phys. Rev. E* **80** 066106
- [99] Erban R and Chapman S J 2009 Stochastic modelling of reaction–diffusion processes: algorithms for bimolecular reactions *Phys. Biol.* **6** 046001
- [100] Elf J and Ehrenberg M 2004 Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases *Syst. Biol.* **1** 230–6
- [101] Chopard B, Frachebourg L and Droz M 1994 Multiparticle lattice gas automata for reaction diffusion systems *Int. J. Mod. Phys. C* **05** 47–63
- [102] Rodriguez J V, Kaandorp J A, Dobrzynski M and Blom J G 2006 Spatial stochastic modelling of the phosphoenolpyruvate-dependent phosphotransferase (PTS) pathway in *Escherichia coli* *Bioinformatics* **22** 1895–901
- [103] Peterson J R, Hallock M J, Cole J A and Luthey-Schulten Z A 2013 A problem solving environment for stochastic biological simulations *PyHPC 2013* (Supercomputing)
- [104] Hallock M J and Luthey-Schulten Z 2016 Improving reaction kernel performance in Lattice Microbes: Particle-wise propensities and run-time generated code *IPDPS Workshops* (IEEE Computer Society) pp 428–34
- [105] Andrews S S, Addy N J, Brent R and Arkin A P 2010 Detailed simulations of cell biology with Smoldyn 2.1 *PLoS Comput. Biol.* **6** e1000705
- [106] Kerr R A, Bartol T M, Kaminsky B, Dittrich M, Chang J-C J, Baden S B, Sejnowski T J and Stiles J R 2008 Fast Monte Carlo simulation methods for biological reaction–diffusion systems in solution and on surfaces *SIAM J. Sci. Comput.* **30** 3126–49
- [107] Schöneberg J and Noé F 2013 ReaDDy—a software for particle-based reaction–diffusion dynamics in crowded cellular environments *PLoS One* **8** e74261
- [108] Ander M, Tomás-Oliveira I, Ferkinghoff-Borg J, Beltrao P, Foglierini M, Ventura B D, Serrano L and Lemerle C 2004 SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks *Syst. Biol.* **1** 129–38
- [109] Fange D, Mahmutovic A and Elf J 2012 MesoRD 1.0: Stochastic reaction–diffusion simulations in the microscopic limit *Bioinformatics* **28** 3155–7
- [110] Vigelius M, Lane A and Meyer B 2010 Accelerating reaction–diffusion simulations with general-purpose graphics processing units *Bioinformatics* **27** 288–90
- [111] Oliveira R F, Terrin A, Benedetto G D, Cannon R C, Koh W, Kim M, Zaccolo M and Blackwell K T 2010 The role of type 4 phosphodiesterases in generating microdomains of cAMP: Large scale stochastic simulations *PLoS One* **5** e11725
- [112] Koh W and Blackwell K T 2011 An accelerated algorithm for discrete stochastic simulation of reaction–diffusion systems using gradient-based diffusion and tau-leaping *J. Chem. Phys.* **134** 154103
- [113] Koh W and Blackwell K T 2012 Improved spatial direct method with gradient-based diffusion to retain full diffusive fluctuations *J. Chem. Phys.* **137** 154111
- [114] Drawert B, Trogdon M, Toor S, Petzold L and Hellander A 2016 MOLNs: a cloud platform for interactive, reproducible and scalable spatial stochastic computational experiments in systems biology using PyURDME *SIAM J. Sci. Comput.* **38** C179–202
- [115] Drawert B, Hellander S, Trogdon M, Yi T-M and Petzold L 2016 A framework for discrete stochastic simulation on 3D moving boundary domains *J. Chem. Phys.* **145** 184113
- [116] Golkaram M, Hellander S, Drawert B and Petzold L R 2016 Macromolecular crowding regulates the gene expression profile by limiting diffusion *PLoS Comput. Biol.* **12** e1005122
- [117] Earnest T M, Watanabe R, Stone J E, Mahamid J, Baumeister W, Villa E and Luthey-Schulten Z 2017 Challenges of integrating stochastic dynamics and cryo-electron tomograms in whole-cell simulations *J. Phys. Chem. B* **121** 3871–81
- [118] Earnest T M, Cole J A, Peterson J R, Hallock M J, Kuhlman T E and Luthey-Schulten Z 2016 Ribosome biogenesis in replicating cells: Integration of experiment and theory *Biopolymers* **105** 735–51
- [119] D’Agostino D, Pasquale G, Clematis A, Maj C, Mosca E, Milanesi L and Merelli I 2014 Parallel solutions for voxel-based simulations of reaction–diffusion systems *BioMed Res. Int.* **2014** 1–10
- [120] Chen W and Schutter E D 2017 Parallel STEPS: large scale stochastic spatial reaction–diffusion simulation with high performance computers *Frontiers Neuroinf.* **11**
- [121] Hepburn I, Chen W and Schutter E D 2016 Accurate reaction–diffusion operator splitting on tetrahedral meshes for parallel stochastic molecular simulations *J. Chem. Phys.* **145** 054118
- [122] Hepburn I, Chen W, Wils S and Schutter E D 2012 STEPS: efficient simulation of stochastic reaction–diffusion models in realistic morphologies *BMC Syst. Biol.* **6** 36
- [123] Arjunan S N V and Tomita M 2009 A new multicompartmental reaction–diffusion modeling method links transient membrane attachment of *E. coli* MinE to E-ring formation *Syst. Synth. Biol.* **4** 35–53
- [124] Arjunan S N V and Takahashi K 2017 Multi-algorithm particle simulations with Spatiocyte *Methods in Molecular Biology* (New York: Springer) pp 219–36
- [125] Stiles J R, van Helden D, Bartol T M, Salpeter E E and Salpeter M M 1996 Miniature endplate current rise times less than 100 microseconds from improved dual recordings can be modeled with passive acetylcholine diffusion from a synaptic vesicle *Proc. Natl Acad. Sci.* **93** 5747–52
- [126] Andrews S S 2016 Smoldyn: particle-based simulation with rule-based modeling, improved molecular interaction and a library interface *Bioinformatics* **33** 710–7
- [127] Dematte L 2012 Smoldyn on graphics processing units: massively parallel Brownian dynamics simulations *IEEE/ACM Trans. Comput. Biol. Bioinf.* **9** 655–67
- [128] Plimpton S J and Slepoy A 2005 Microbial cell modeling via reacting diffusive particles *J. Phys.: Conf. Ser.* **16** 305–9
- [129] Sanford C, Yip M L, White C and Parkinson J 2006 Cell++—simulating biochemical pathways *Bioinformatics* **22** 2918–25
- [130] van Zon J S and Wolde P R 2005 Green’s-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space *J. Chem. Phys.* **123** 234910
- [131] Biedermann J, Ullrich A, Schöneberg J and Noé F 2015 ReaDDyMM: fast interacting particle reaction–diffusion simulations using graphical processing units *Biophys. J.* **108** 457–61
- [132] Schöneberg J, Ullrich A and Noé F 2014 Simulation tools for particle-based reaction–diffusion dynamics in continuous space *BMC Biophys.* **7**

- [133] Cianci C, Smith S and Grima R 2017 Capturing Brownian dynamics with an on-lattice model of hard-sphere diffusion *Phys. Rev. E* **95** 052118
- [134] Smith S and Grima R 2017 Fast simulation of brownian dynamics in a crowded environment *The J. Chem. Phys.* **146** 024105
- [135] Isaacson S A 2013 A convergent reaction–diffusion master equation *J. Chem. Phys.* **139** 054101
- [136] Cianci C, Smith S and Grima R 2016 Molecular finite-size effects in stochastic models of equilibrium chemical systems *J. Chem. Phys.* **144** 084101
- [137] Alfonsi A, Cancès E, Turinici G, Ventura B D and Huisinga W 2005 Adaptive simulation of hybrid stochastic and deterministic models for biochemical systems *ESAIM: Proc.* **14** 1–13
- [138] Cao Y, Gillespie D T and Petzold L R 2005 The slow-scale stochastic simulation algorithm *J. Chem. Phys.* **122** 014116
- [139] Jahnke T and Kreim M 2012 Error bound for piecewise deterministic processes modeling stochastic reaction systems *Multiscale Model. Simul.* **10** 1119–47
- [140] Yates C A and Flegg M B 2015 The pseudo-compartment method for coupling partial differential equation and compartment-based models of diffusion *J. R. Soc. Interface* **12** 20150141
- [141] Taylor P R, Baker R E, Simpson M J and Yates C A 2016 Coupling volume-excluding compartment-based models of diffusion at different scales: voronoi and pseudo-compartment approaches *J. R. Soc. Interface* **13** 20160336
- [142] Harrison J U and Yates C A 2016 A hybrid algorithm for coupling partial differential equation and compartment-based dynamics *J. R. Soc. Interface* **13** 20160335
- [143] Schaff J C, Gao F, Li Y, Novak I L and Slepchenko B M 2016 Numerical approach to spatial deterministic-stochastic models arising in cell biology *PLoS Comput. Biol.* **12** e1005236
- [144] Robinson M, Andrews S S and Erban R 2015 Multiscale reaction–diffusion simulations with Smoldyn *Bioinformatics* **31** 2406–8
- [145] Feng H, Han B and Wang J 2011 Adiabatic and non-adiabatic non-equilibrium stochastic dynamics of single regulating genes *The J. Phys. Chem. B* **115** 1254–61
- [146] Gutenkunst R N, Waterfall J J, Casey F P, Brown K S, Myers C R and Sethna J P 2007 Universally sloppy parameter sensitivities in systems biology models *PLoS Comput. Biol.* **3** e189
- [147] Daniels B C, Chen Y-J, Sethna J P, Gutenkunst R N and Myers C R 2008 Sloppiness, robustness and evolvability in systems biology *Curr. Opin. Biotechnol.* **19** 389–95
- [148] Chis O-T, Villaverde A F, Banga J R and Balsa-Canto E 2016 On the relationship between sloppiness and identifiability *Math. Biosci.* **282** 147–61
- [149] Slezak D F, Suárez C, Cecchi G A, Marshall G and Stolovitzky G 2010 When the optimal is not the best: parameter estimation in complex biological models *PLoS One* **5** e13283
- [150] Earnest T M, Roberts E, Assaf M, Dahmen K and Luthey-Schulten Z 2013 DNA looping increases the range of bistability in a stochastic model of the lac genetic switch *Phys. Biol.* **10** 026002
- [151] Silk D, Kirk P D W, Barnes C P, Toni T and Stumpf M P H 2014 Model selection in systems biology depends on experimental design *PLoS Comput. Biol.* **10** e1003650
- [152] White A, Tolman M, Thames H D, Withers H R, Mason K A and Transtrum M K 2016 The limitations of model-based experimental design and parameter estimation in sloppy systems *PLoS One* **12** e1005227
- [153] Joshi M, Seidel-Morgenstern A and Kremling A 2006 Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems *Metabolic Eng.* **8** 447–55
- [154] Weiss J N 1997 The Hill equation revisited: Uses and misuses *FASEB J.* **11** 835–41
- [155] Lawson M J, Petzold L and Hellander A 2015 Accuracy of the Michaelis–Menten approximation when analysing effects of molecular noise *J. R. Soc. Interface* **12** 20150054
- [156] Smith S and Grima R 2016 Breakdown of the reaction–diffusion master equation with nonelementary rates *Phys. Rev. E* **93** 052135
- [157] Sunnåker M, Busetto A G, Numminen E, Corander J, Foll M and Dessimoz C 2013 Approximate bayesian computation *PLoS Comput. Biol.* **9** e1002803
- [158] Munsky B and Khammash M 2006 The finite state projection algorithm for the solution of the chemical master equation *J. Chem. Phys.* **124** 044104
- [159] Poovathingal S and Gunawan R 2010 Global parameter estimation methods for stochastic biochemical systems *BMC Bioinf.* **11** 414
- [160] Srivastava R and Rawlings J B 2014 Parameter estimation in stochastic chemical kinetic models using derivative free optimization and bootstrapping *Comput. Chem. Eng.* **63** 152–8
- [161] Lester C, Yates C A and Baker R E 2017 Efficient parameter sensitivity computation for spatially extended reaction networks *J. Chem. Phys.* **146** 044106
- [162] Schnoerr D, Grima R and Sanguinetti G 2016 Cox process representation and inference for stochastic reaction–diffusion processes *Nat. Commun.* **7** 11729
- [163] Cressie N and Wikle C K 2011 *Statistics for Spatio-Temporal Data* (New York: Wiley)
- [164] Kalwarczyk T, Tabaka M and Holyst R 2012 Biologistics—diffusion coefficients for complete proteome of *Escherichia coli* *Bioinformatics* **28** 2971–8
- [165] Bartol T M, Keller D X, Kinney J P, Bajaj C L, Harris K M, Sejnowski T J and Kennedy M B 2015 Computational reconstitution of spine calcium transients from individual proteins *Frontiers Synaptic Neurosci.* **7**
- [166] Isaacson S A, McQueen D M and Peskin C S 2011 The influence of volume exclusion by chromatin on the time required to find specific DNA binding sites by diffusion *Proc. Natl Acad. Sci.* **108** 3815–20
- [167] Hornos J E M, Schultz D, Innocentini G C P, Wang J, Walczak A M, Onuchic J N and Wolynes P G 2005 Self-regulating gene: an exact solution *Phys. Rev. E* **72** 051907
- [168] Grima R, Schmidt D R and Newman T J 2012 Steady-state fluctuations of a genetic feedback loop: an exact solution *The J. Chem. Phys.* **137** 035104
- [169] Kumar N, Platini T and Kulkarni R V 2014 Exact distributions for stochastic gene expression models with bursting and feedback *Phys. Rev. Lett.* **113** 268105
- [170] Choudhary K, Oehler S and Narang A 2014 Protein distributions from a stochastic model of the lac operon of *E. coli* with DNA looping: analytical solution and comparison with experiments *PLoS One* **9** e102580
- [171] Wang J 2015 Landscape and flux theory of non-equilibrium dynamical systems with application to biology *Adv. Phys.* **64** 1–137
- [172] Michelsen O, De Mattos M J T, Jensen P R and Hansen F G 2003 Precise determinations of C and D periods by flow cytometry in *Escherichia coli* K-12 and B/r *Microbiology* **149** 1001–10
- [173] Cooper S and Helmstetter C E 1968 Chromosome replication and the division cycle of *Escherichia coli* B/r *J. Mol. Biol.* **31** 519–40
- [174] Jones D L, Brewster R C and Phillips R 2014 Promoter architecture dictates cell-to-cell variability in gene expression *Science* **346** 1533–6

- [175] Peterson J R, Cole J A, Fei J, Ha T and Luthey-Schulten Z A 2015 Effects of DNA replication on mRNA noise *Proc. Natl Acad. Sci.* **112** 15886–91
- [176] Bunner A E, Beck A H and Williamson J R 2010 Kinetic cooperativity in *Escherichia coli* 30S ribosomal subunit reconstitution reveals additional complexity in the assembly landscape *Proc. Natl Acad. Sci.* **107** 5417–22
- [177] Mulder A M, Yoshioka C, Beck A H, Bunner A E, Milligan R A, Potter C S, Carragher B and Williamson J R 2010 Visualizing ribosome biogenesis: parallel assembly pathways for the 30S subunit *Science* **330** 673–7
- [178] Sashital D G, Greeman C A, Lyumkis D, Potter C S, Carragher B and Williamson J R 2014 A combined quantitative mass spectrometry and electron microscopy analysis of ribosomal 30S subunit assembly in *E. coli* *eLife* **3** e04491
- [179] Lindahl L 1975 Intermediates and time kinetics of the *in vivo* assembly of *Escherichia coli* ribosomes *J. Mol. Biol.* **15**–37
- [180] Chai Q, Singh B, Peisker K, Metzendorf N, Ge X, Dasgupta S and Sanyal S 2014 Organization of ribosomes and nucleoids in *Escherichia coli* cells during growth and in quiescence *J. Biol. Chem.* **289** 11342–52
- [181] Kaczanowska M and Rydén-Aulin M 2007 Ribosome biogenesis and the translation process in *Escherichia coli* *Microbiol. Mol. Biol. Rev.* **71** 477–94
- [182] Bremer H and Dennis P P 1996 Modulation of chemical composition and other parameters of the cell by growth rate *Escherichia coli* and *Salmonella Typhimurium: Cellular and Molecular Biology* 2nd edn, ed F C Neidhardt *et al* (Washington, DC: ASM) pp 1553–69
- [183] Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J and Milo R 2014 Visual account of protein investment in cellular functions *Proc. Natl Acad. Sci.* **111** 8488–93
- [184] Held W A, Ballou B, Mizushima S and Nomura M 1974 Assembly mapping of 30S ribosomal proteins from *Escherichia coli*: further studies *J. Biol. Chem.* **249** 3103–11
- [185] Hosokawa K, Fujimura R K and Nomura M 1966 Reconstitution of functionally active ribosomes from inactive subparticles and proteins *Proc. Natl Acad. Sci.* **55** 198–204
- [186] Adilakshmi T, Ramaswamy P and Woodson S A 2005 Protein-independent folding pathway of the 16S rRNA 5' domain *J. Mol. Biol.* **351** 508–19
- [187] Adilakshmi T, Bellur D L and Woodson S A 2008 Concurrent nucleation of 16S folding and induced fit in 30S ribosome assembly *Nature* **455** 1268–72
- [188] Kim H, Abeysirigunawardena S C, Chen K, Mayerle M, Ragunathan K, Luthey-Schulten Z, Ha T and Woodson S A 2014 Protein-guided RNA dynamics during early ribosome assembly *Nature* **506** 334–8
- [189] Talkington M W T, Siuzdak G and Williamson J R 2005 An assembly landscape for the 30S ribosomal subunit *Nature* **438** 628–32
- [190] Sykes M T and Williamson J R 2009 A complex assembly landscape for the 30S ribosomal subunit *Ann. Rev. Biochem.* **38** 197–215
- [191] Davis J H and Williamson J R 2017 Structure and dynamics of bacterial ribosome biogenesis *Phil. Trans. R. Soc. B: Biol. Sci.* **372** 20160181
- [192] Mahmutovic A, Fange D, Berg O G and Elf J 2012 Lost in presumption: stochastic reactions in spatial models *Nat. Methods* **9** 1163–6
- [193] Orth J D, Thiele I and Palsson B Ø 2010 What is flux balance analysis? *Nat. Biotechnol.* **28** 245–8
- [194] Henry C S, Broadbelt L J and Hatzimanikatis V 2007 Thermodynamics-based metabolic flux analysis *Biophys. J.* **92** 1792–805
- [195] Reed J L, Vo T D, Schilling C H and Palsson B Ø 2003 An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR) *Genome Biol.* **4** R54
- [196] Noor E, Bar-Even A, Flamholz A, Reznik E, Liebermeister W and Milo R 2014 Pathway thermodynamics highlights kinetic obstacles in central metabolism *PLoS Comput. Biol.* **10** e1003483
- [197] Jinich A, Rappoport D, Dunn I, Sanchez-Lengeling B, Olivares-Amaya R, Noor E, Even A B and Aspuru-Guzik A 2014 Quantum chemical approach to estimating the thermodynamics of metabolic reactions *Sci. Rep.* **4** 7022
- [198] Chang R L, Andrews K, Kim D, Li Z, Godzik A and Palsson B Ø 2013 Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli* *Science* **340** 1220–3
- [199] Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J and Schomburg D 2017 BRENDA in 2017: new perspectives and new tools in BRENDA *Nucl. Acids Res.* **45** D380–8
- [200] Beg Q K, Vazquez A, Ernst J, de Menezes M A, Bar-Joseph Z, Barabási A-L and Oltvai Z N 2007 Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity *Proc. Natl Acad. Sci.* **104** 12663–8
- [201] Adadi R, Volkmer B, Milo R, Heinemann M and Shlomi T 2012 Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters *PLoS Comput. Biol.* **8** e1002575
- [202] Cole J A, Kohler L, Hedhli J and Luthey-Schulten Z 2015 Spatially-resolved metabolic cooperativity within dense bacterial colonies *BMC Syst. Biol.* **9**
- [203] Beard D A and Qian H 2007 Relationship between thermodynamic driving force and one-way fluxes in reversible processes *PLoS One* **2** e144
- [204] Keseler I M *et al* 2017 The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12 *Nucl. Acids Res.* **45** D543–50
- [205] Rich P 2003 The molecular machinery of keilin's respiratory chain *Biochem. Soc. Trans.* **31** 1095–105
- [206] Milo R, Jorgensen P, Moran U, Weber G and Springer M 2009 BioNumbers—the database of key numbers in molecular and cell biology *Nucl. Acids Res.* **38** D750–3
- [207] Park J O, Rubin S A, Xu Y-F, Amador-Noguez D, Fan J, Shlomi T and Rabinowitz J D 2016 Metabolite concentrations, fluxes and free energies imply efficient enzyme usage *Nat. Chem. Biol.* **12** 482–9
- [208] Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R and Heinemann M 2015 The quantitative and condition-dependent *Escherichia coli* proteome *Nat. Biotechnol.* **34** 104–10
- [209] Millard P, Smallbone K and Mendes P 2017 Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in *Escherichia coli* *PLOS Comput. Biol.* **13** e1005396
- [210] Peskov K, Mogilevskaia E and Demin O 2012 Kinetic modelling of central carbon metabolism in *Escherichia coli* *FEBS J.* **279** 3374–85
- [211] Du B, Zielinski D C, Kavvas E S, Dräger A, Tan J, Zhang Z, Ruggiero K E, Arzumanyan G A and Palsson B Ø 2016 Evaluation of rate law approximations in bottom-up kinetic models of metabolism *BMC Syst. Biol.* **10** 40
- [212] Khodayari A and Maranas C D 2016 A genome-scale *Escherichia coli* kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains *Nat. Commun.* **7** 13806

- [213] Hutchison C A *et al* 2016 Design and synthesis of a minimal bacterial genome *Science* **351** aad6253
- [214] Stouthamer A 1973 A theoretical study on the amount of ATP required for synthesis of microbial cell material *Antonie van Leeuwenhoek* **39** 545–65
- [215] Pontes M H, Sevostyanova A and Groisman E A 2015 When too much ATP is bad for protein synthesis *J. Mol. Biol.* **427** 2586–94
- [216] Cole J, Hallock M J, Labhsetwar P, Peterson J R, Stone J E, Luthey-Schulten Z 2014 Stochastic simulations of cellular processes: from single cells to colonies *Computational Systems Biology* 2nd edn, ed A Kriete and R Eils (Amsterdam: Elsevier) ch 13
- [217] Karr J R, Sanghvi J C, Macklin D N, Gutschow M V, Jacobs J M, Bolival B, Assad-Garcia N, Glass J I and Covert M W 2012 A whole-cell computational model predicts phenotype from genotype *Cell* **150** 389–401
- [218] Popov K, Komianos J and Papoian G A 2016 MEDYAN: Mechanochemical simulations of contraction and polarity alignment in actomyosin networks *PLoS Comput. Biol.* **12** e1004877
- [219] Horn H and Lackner S 2014 Modeling of biofilm systems: a review *Productive Biofilms* (Berlin: Springer) pp 53–76
- [220] Deisboeck T S, Wang Z, Macklin P and Cristini V 2011 Multiscale cancer modeling *Ann. Rev. Biomed. Eng.* **13** 127–55
- [221] Yamada K M and Cukierman E 2007 Modeling tissue morphogenesis and cancer in 3D *Cell* **130** 601–10
- [222] Pouille P-A and Farge E 2008 Hydrodynamic simulation of multicellular embryo invagination *Phys. Biol.* **5** 015005
- [223] Smith S, Cianci C and Grima R 2016 Analytical approximations for spatial stochastic gene expression in single cells and tissues *J. R. Soc. Interface* **13** 20151051
- [224] Harcombe W R *et al* 2014 Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics *Cell Rep.* **7** 1104–15
- [225] Peters A, Wimpenny J and Coombs J 1987 Oxygen profiles in and in the agar beneath, colonies of *Bacillus cereus*, *Ataphylococcus albus* and *Escherichia coli* *Microbiology* **133** 1257–63
- [226] Peterson J R, Cole J A and Luthey-Schulten Z 2017 Parametric studies of metabolic cooperativity in *Escherichia coli* colonies: Strain and geometric confinement effects *PLoS One* **12** e0182570
- [227] Hanahan D and Weinberg R A 2000 The hallmarks of cancer *Cell* **100** 57–70
- [228] Guillaumond F *et al* 2013 Strengthened glycolysis under hypoxia supports tumor symbiosis and hexosamine biosynthesis in pancreatic adenocarcinoma *Proc. Natl Acad. Sci.* **110** 3919–24
- [229] Sonveaux P *et al* 2008 Targeting lactate-fueled respiration selectively kills hypoxic tumor cells in mice *J. Clin. Invest.* **118** 3930–42
- [230] Nilsson A and Nielsen J 2017 Genome scale metabolic modeling of cancer *Metabolic Eng.* **43** 103–12
- [231] Lambert G, Estévez-Salmeron L, Oh S, Liao D, Emerson B M, Tlsty T D and Austin R H 2011 An analogy between the evolution of drug resistance in bacterial communities and malignant tissues *Nat. Rev. Cancer* **11** 375–82
- [232] de Bruin E C, Taylor T B and Swanton C 2013 Intra-tumor heterogeneity: lessons from microbial evolution and clinical implications *Genome Med.* **5** 101
- [233] Ben-Jacob E, Coffey D S and Levine H 2012 Bacterial survival strategies suggest rethinking cancer cooperativity *Trends in Microbiology* **20** 403–10
- [234] Whitford P C, Noel J K, Gosavi S, Schug A, Sanbonmatsu K Y and Onuchic J N 2009 An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields *Proteins: Struct. Funct. Bioinf.* **75** 430–41
- [235] Leopold P E, Montal M and Onuchic J N 1992 Protein folding funnels: a kinetic approach to the sequence-structure relationship *Proc. Natl Acad. Sci.* **89** 8721–5
- [236] Gumbart J, Wang Y, Aksimentiev A, Tajkhorshid E and Schulten K 2005 Molecular dynamics simulations of proteins in lipid bilayers *Curr. Opin. Struct. Biol.* **15** 423–31
- [237] Phillips J C, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel R D, Kale L and Schulten K 2005 Scalable molecular dynamics with NAMD *J. Comput. Chem.* **26** 1781–802
- [238] McGuffee S R and Elcock A H 2010 Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm *PLoS Comput. Biol.* **6** e1000694
- [239] Yu I, Mori T, Ando T, Harada R, Jung J, Sugita Y and Feig M 2016 Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm *eLife* **5** e19274
- [240] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I and Mackerell A D 2010 CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields *J. Comput. Chem.* **31** 671–90
- [241] Wang J, Wolf R M, Caldwell J W, Kollman P A and Case D A 2004 Development and testing of a general amber force field *J. Comput. Chem.* **25** 1157–74
- [242] Burley S K, Berman H M, Kleywegt G J, Markley J L, Nakamura H and Velankar S 2017 Protein data bank (PDB): the single global macromolecular structure archive *Methods in Molecular Biology* (New York: Springer) pp 627–41
- [243] Prakadan S M, Shalek A K and Weitz D A 2017 Scaling by shrinking: empowering single-cell ‘omics’ with microfluidic devices *Nat. Rev. Genet.* **18** 345–61
- [244] Teusink B *et al* 2000 Can yeast glycolysis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing biochemistry *Eur. J. Biochem.* **267** 5313–29
- [245] van Eunen K, Kiewiet J A L, Westerhoff H V and Bakker B M 2012 Testing biochemistry revisited: How *in vivo* metabolism can be understood from *in vitro* enzyme kinetics *PLoS Comput. Biol.* **8** e1002483



Tyler M. Earnest is a postdoctoral researcher in the laboratory of Professor Luthey-Schulten. Dr. Earnest received a B.S. in Chemistry and a B.S. in Physics from the South Dakota School of Mines and Technology in 2008, a M.S in Physics in 2012 and a Ph.D. in Physics in 2016 from the University of Illinois at Urbana-Champaign.



John A. Cole received a B. S. in Physics from Rutgers University in 2005, and an M. S. in Computational Biology from the New Jersey Institute of Technology in 2008 and his Ph.D. from the University of Illinois at Urbana-Champaign in 2017.



Professor Schulten received a B.S. in Chemistry from the University of Southern California in 1969, an M.S. in Chemistry from Harvard University in 1972, and a Ph.D. in Applied Mathematics from Harvard University in 1975. From 1975 to 1980, she was a Research Fellow at the Max-Planck Institute for Biophysical Chemistry in Göttingen, and from 1980 to 1985, she was a Research Fellow in the Department of Theoretical Physics at the Technical University of Munich. Professor Schulten has been at the University of Illinois since 1987, where she is currently the William and Janet Lycan Professor of Chemistry, co-director of the NSF Center for the Physics of Living Cells, and co-investigator at the NIH Resource of Macromolecular Modeling and Bioinformatics at the Beckman Institute.