Simulating single-cell metabolism using a stochastic flux-balance analysis algorithm

David S. Tourigny[1*], Arthur P. Goldberg[2], and Jonathan R. Karr[2]

[1]*Irving Institute for Cancer Dynamics, Columbia University, Schermerhorn Hall, 1190 Amsterdam Ave, New York, NY 10027, USA*

[2]*Icahn Institute for Data Science and Genomic Technology, and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA*

[*]*Correspondence: dst2156@columbia.edu*

Abstract

Stochasticity from gene expression in single cells is known to drive metabolic heterogeneity at the population-level, which is understood to have important consequences for issues such as microbial drug tolerance and treatment of human diseases like cancer. Despite considerable advancements in profiling the genomes, transcriptomes, and proteomes of single cells, it remains difficult to experimentally characterise their metabolism at genome-scale. Computational methods could bridge this gap toward a systems understanding of single-cell biology. To address this challenge, we developed stochastic simulation algorithm with flux-balance analysis embedded (SSA-FBA), a genome-scale modelling framework for simulating the stochastic dynamics of the metabolism of individual cells. SSA-FBA extends the constraint-based modelling formalism of metabolic network modelling to the single-cell regime, enabling simulation when experimentation is intractable. We also developed an efficient implementation of SSA-FBA that leverages the topology of embedded FBA models to significantly reduce the computational cost of simulation. As a preliminary case study, we built a single-cell model of *Mycoplasma pneumoniae*, and used SSA-FBA to explore the role of stochasticity on the dynamics of ATP production and consumption at the single-cell level.

Introduction

Describing the phenotypic behaviour of single cells is critical for a deeper understanding of biological tissues, organisms and populations. Recent experimental advances are driving a data explosion in systems biology by enabling researchers to profile multiple dimensions of single cells, including their genome, transcriptome, and proteome [1, 2, 3]. Such single-cell measurements can yield information about thousands of individual cells in a single experiment. This can provide insights on intracellular function and the role of intercellular heterogeneity in a variety of biological systems ranging from microbial populations [4, 5] to human diseases such as cancer [6, 7].

Although metabolism is a key aspect of cellular physiology, methodologies for probing the metabolism of single cells are comparatively immature [8, 9, 10, 12, 13]. This presents a barrier to studying a variety of phenomena such as metabolic reprogramming in tumours, which is now understood to be a central hallmark of cancer [14, 15]. It is more challenging to experimentally probe the metabolism of single cells due to low abundances of many metabolites, the compartmentalisation of eukaryotic cells, and the wide structural diversity of metabolites [10]. Moreover, faster timescales of enzymatic reactions makes measuring the dynamics of metabolism more difficult than measuring that of DNA replication or gene expression. These experimental obstacles to studying single-cell metabolism necessitate the development of computational techniques that infer the metabolism of single cells from other information, such as single-cell transcriptomic or proteomic data and population-level metabolic data [11].

While our current capabilities to probe and model the metabolism of single-cells are limited, considerable attention has been devoted to the metabolism of cellular populations. For example, metabolic network modelling has achieved a great deal of success combining limited experimental data and computational simulation to study gene essentiality and guide metabolic engineering [16, 17, 18]. Extensions of these population-based metabolic network modelling frameworks, such as dynamic metabolism expression models [19] and dynamic enzyme-cost flux-balance analysis [20], have also been developed to incorporate the dynamics of gene expression. However, these deterministic approaches cannot capture effects that are relevant to the metabolism of single cells,

where stochasticity is understood to play a pivotal role [21, 22]. To date, there have only been a handful of attempts (see [23] and references therein) to extend metabolic network modelling to the single-cell regime. These efforts have mainly focused on integrating single-cell transcriptomics data with flux-balance analysis (FBA) in the context of cancer. Critically, these studies have not addressed the temporal dynamics of single cells, which is a key feature of their behaviour.

The existing approaches to stochastic modelling of single-cell metabolism can be organised into two categories: (a) (semi-)analytical treatment of rigorous models of individual metabolic pathways involving the expression of one [24, 25] or several [26, 27] enzymes catalysing a handful of metabolic reactions, or (b) ad hoc simulation of whole-cell models [28, 29] involving hybrid methods that combine ordinary differential equations (ODEs), particle-based stochastic simulation algorithms (SSA, [30, 31]), and dynamic FBA (DFBA, [32]). Although the shared aim of these approaches is to relate single-cell behaviour to genotype, the two categories fall at opposite ends of a wide spectrum: whole-cell modelling attempts to accommodate as much detail as possible, but no framework is yet available to rigorously simulate entire cells. On the other hand, analytic methods that can provide a mechanistic understanding of small pathways are intractable for elucidating the complexity of entire cells. Here, we introduce a stochastic extension of FBA that we call stochastic simulation algorithm with flux-balance analysis embedded (SSA-FBA), which enables systematic systems-scale simulations of the metabolism of single cells. We believe SSA-FBA is a powerful computational tool for simulating stochastic dynamics of metabolic network models at the single-cell level.

The remainder of this paper is organised as follows. We first outline the concepts of SSA-FBA and relate it to a formal description of single-cell metabolism based on the chemical master equation. Next, we compare exact and approximate implementations of SSA-FBA, and then introduce an advanced algorithm that significantly improves the efficiency of exact simulations. As a case study, we subsequently present simulation results of an SSA-FBA model of a single *Mycoplasma pneumoniae* cell representing 505 reactions where we also explore the consequences of stochasticity for the dynamics of adenosine triphosphate (ATP) production and consumption at the single-cell

level. We conclude the paper with a discussion of the main results and directions for future work.

Results

Stochastic simulation algorithm with flux-balance analysis embedded

Stochasticity in the metabolism and growth of single cells is generally believed to emerge primarily from fluctuations in enzyme expression levels [33, 21, 22, 24, 25] (Figure 1). Due to the relatively high copy numbers of most metabolites, metabolism is thought to have little intrinsic stochastic variation compared to gene expression [21]. This observation motivates SSA-FBA as an appropriate framework for modelling single-cell metabolism, where the dynamics of reaction fluxes internal to a metabolic network are captured deterministically by FBA, while SSA is used to model changes in the copy numbers of enzyme molecules and metabolites that are produced or consumed on the periphery of the metabolic network.

The combined single-cell network of M metabolic and enzyme expression reactions can be captured by the chemical master equation (CME) [34]

$$\frac{dP(\mathbf{n})}{dt} = \sum_{j=1}^{M} a_j(\mathbf{n}-\mathbf{S}_j)P(\mathbf{n}-\mathbf{S}_j) - a_j(\mathbf{n})P(\mathbf{n}) \tag{1}$$

where $\mathbf{n}$ is an N-dimensional vector for counts of N chemical species, $a_j(\mathbf{n})$ is the propensity value of reaction j given $\mathbf{n}$, and $\mathbf{S}_j$ is the stoichiometry of reaction j. The probability density function $P(\mathbf{n})$ describes the probability of the system to occupy state $\mathbf{n}(t)$ at time t. Because CMEs are usually too complex to be solved directly, investigators often use SSA to sample trajectories through the state space of the CME [30, 31]. One limitation of SSA is its computational cost, which is particularly acute for large chemical reaction networks. Various performance enhancements have therefore been developed to improve the run time of SSA. Since the average counts of enzymes are many times lower than those of metabolites, and metabolic reactions operate on a much faster time-scale than that of gene synthesis and degradation, several researchers have used *stochastic*

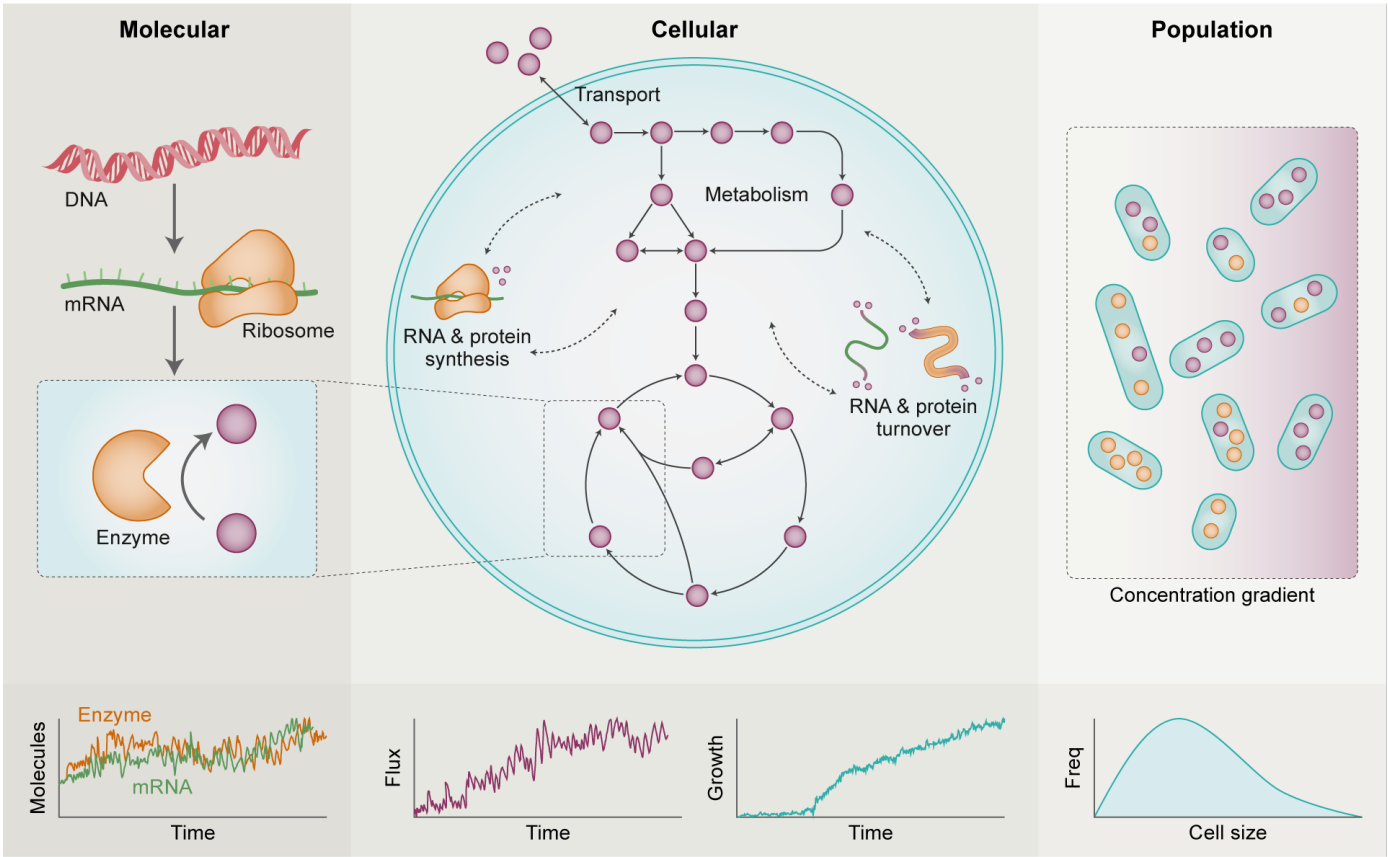


Figure 1: The metabolic behaviour of single cells is influenced by the spatiotemporal dynamics of metabolites and enzymes at multiple scales. Temporal variation in the metabolic behaviour of single cells is primarily driven by fluctuations in gene expression and amplified or attenuated by their metabolic networks. For example, the structure of the reaction network combined with fluctuating enzyme and metabolite levels determine intracellular metabolic fluxes that, in turn, govern the energy supply and growth dynamics of individual cells. The resulting stochastic nature of these properties at the single-cell level drives phenotypic heterogeneity at the population level, which can also be influenced by external factors such as environmental conditions.

quasi-steady state assumptions (stochastic generalisations of the quasi-steady state assumption in DFBA [32]) to approximate SSA trajectories at a lower computational cost. These methods correspond to a reduction of the CME (1) on the basis of time-scale separation (see [35, 36, 37] and Supplementary Appendix S1).

While the above time-scale separation methods could in principle be used to simulate single-cell metabolism, investigators rarely have sufficient data to construct detailed dynamical models of metabolism in single cells due to the experimental challenges outlined in the Introduction. Instead, we realised that the constraint-based formalism of (D)FBA [16, 17, 32] could be used to identify a numerical solution to the deterministic quasi-steady state conditions of the metabolic reaction network with a small amount of data at a modest computational cost. In short, SSA-FBA embeds an FBA model into SSA, analogous to the way that an FBA model is embedded into a system of ODEs in DFBA. This embedding enables SSA-FBA to simulate the stochastic behaviour of the metabolism of single cells analogous to how DFBA enables deterministic simulation of the dynamic behaviour of the metabolism of populations.

First, SSA-FBA separates reactions into three mutually disjoint subsets based on whether the species that participate in the reaction are defined to be internal or external to the metabolic reaction network (Figure 2a). The three subsets of reactions in an SSA-FBA model are as follows:

- **FBA only reactions**: reactions that interconvert among the internal species,
- **SSA only reactions**: reactions that interconvert among the external species,
- **SSA-FBA reactions**: reactions that convert between the internal and external species.

For example, the SSA only reactions may correspond to the synthesis and degradation of gene products, while the FBA only and SSA-FBA reactions may include reactions involved in trans-membrane transport and central metabolism.

Since external species, such as enzymes and transporters, often influence the rates or propensity values of metabolic reactions, the FBA only and SSA-FBA reactions can be constrained or bounded by the dynamic counts of external species (Figure 2a). Mathematically, this can be expressed as

the following linear programming (LP) problem

$$\begin{aligned} &\text{maximise}: \quad z = \mathbf{c} \cdot \mathbf{a}_{FBA} \\ &\text{subject to}: \quad \mathbf{S} \cdot \mathbf{a}_{FBA} = 0, \quad \mathbf{l}(\mathbf{n}) \leq \mathbf{a}_{FBA} \leq \mathbf{u}(\mathbf{n}) \end{aligned} \tag{2}$$

where $\mathbf{a}_{FBA}$ is a vector containing the propensity values of FBA only and SSA-FBA reactions, $\mathbf{S}$ is the stoichiometry matrix of the metabolic reaction network, and $\mathbf{l}(\mathbf{n}), \mathbf{u}(\mathbf{n})$ are bounds that depend on the counts of external species (see Supplementary Appendix S1 for extended discussion). The coefficient vector $\mathbf{c}$ is chosen to reflect a biologically-relevant objective such as maximising the rate of production of the metabolites required for growth. Solving the LP problem (2) for a given species counts $\mathbf{n}$ returns an optimal set of propensity values for FBA only and SSA-FBA reactions, while propensity values for the SSA only reactions are calculated from $\mathbf{n}$ by direct evaluation of a reaction rate function as in SSA [30, 31].

Next, as outlined in Figures 2b and 2c, SSA-FBA combines propensity values for SSA-FBA reactions obtained from an optimal solution to the LP problem (2) with the propensity values of SSA only reactions to determine the next reaction event according to Gillespie's original algorithm [30, 31] (although this step can also be replaced by more advanced methods such as the Next Reaction Method [38]). Since FBA only reactions do not affect the counts of external species, their propensity values are not required at this stage. Finally, execution of either an SSA only or SSA-FBA reaction updates the counts of external species, which in turn update the bounds of both FBA only and SSA-FBA propensity values in (2).

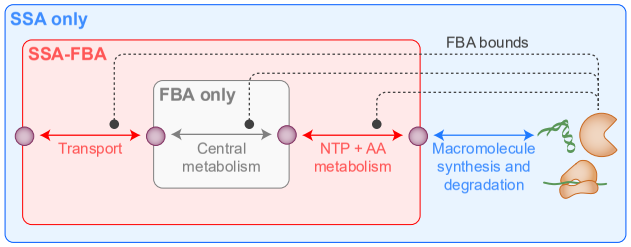
In summary, SSA-FBA embeds FBA within SSA by using an LP problem to predict an optimal propensity value for each metabolic reaction, using the counts of species predicted by SSA. This bidirectional coupling of SSA and FBA is analogous to the coupling of FBA and ODE integration in DFBA [32]. One challenge of both DFBA and SSA-FBA is that solutions to (2)– and therefore the optimal propensity values of SSA-FBA reactions– are not necessarily unique. However, it turns out to always be possible to formulate a lexicographic version of the LP problem by including multiple biological objective functions to be optimised sequentially in order to guarantee

uniqueness and retain compatibility with the implementation described in the next section (see also [39]). It is important to highlight that, unlike variants of SSA that approximate trajectories of the CME (1) in order to reduce the computational overhead [35, 36, 37], the goal of SSA-FBA is to approximately predict the metabolic dynamics of single cells in the absence of detailed kinetic information about each species and reaction [16, 17]. The modeller can control the degree of this approximation by how they choose to partition reactions into the three subsets (FBA only, SSA only, or SSA-FBA). As a rule of thumb, we recommend that modellers distinguish between internal and external species (and hence partitioning of reactions) based loosely on the time-scales of the species. Further guidelines for encoding SSA-FBA models and additional information about the relationship between SSA-FBA and the CME are outlined in Supplementary Appendix S1.

Efficient implementation of SSA-FBA

As described above, SSA-FBA iteratively (i) uses the counts of the species and reaction rate laws to calculate the propensity values for each SSA only reaction, (ii) uses the counts of the species in FBA to calculate the propensity values of each SSA-FBA reaction, (iii) uses SSA to select the next reaction and (iv) updates the species counts based on stoichiometry of the selected SSA only or SSA-FBA reaction. The direct method for performing an SSA-FBA simulation in this way is summarised in the left column of Figure 2b. At first glance, the direct method appears to be computationally expensive for larger models because it appears to require solving one FBA problem per SSA execution event.

One way to reduce the computational cost associated with the direct SSA-FBA simulation method is to update the SSA-FBA propensity values less frequently, assuming a time interval (Δ_{event}) across which an optimal solution to the embedded FBA problem does not change appreciably. When the end of the interval is reached, the constraints of the embedded FBA problem can be updated using the current counts of species and re-solved for a new set of SSA-FBA propensity values used to parameterise integration over the next interval. This approximate SSA-FBA simulation method (centre column in Figure 2b) is analogous to the simplest implementation of DFBA



(a) SSA-FBA separates reactions into SSA only (blue), SSA-FBA (red) and FBA only (grey) subsets based on the species that participate in each reaction. In this example, which models the coupling of transmembrane transport and central metabolism to macromolecule synthesis and degradation via amino acid (AA) and nucleotide triphosphate (NTP) metabolism, metabolites in central carbon metabolism are internal to the metabolic reaction network. Gene products such as enzymes and transporters contribute to the rates or propensity values of FBA and SSA-FBA reactions, and this is modelled through FBA bounds on these reactions (dotted lines).

Direct method

Set *event* to zero

While *event* less than *maximum events* **do**

- Calculate SSA-only propensities
- Calculate SSA-FBA propensities using FBA
- Execute SSA event
- **Increment** *event*

Approximate method

Set *event*, *counter* to zero

While *event* less than *maximum events* **do**

- Calculate SSA-only propensities
- **If** *counter* equals *time interval* **do**
  - Calculate SSA-FBA propensities using FBA
  - **Set** *counter* to zero
- Execute SSA event
- **Increment** *event*

Optimal basis method

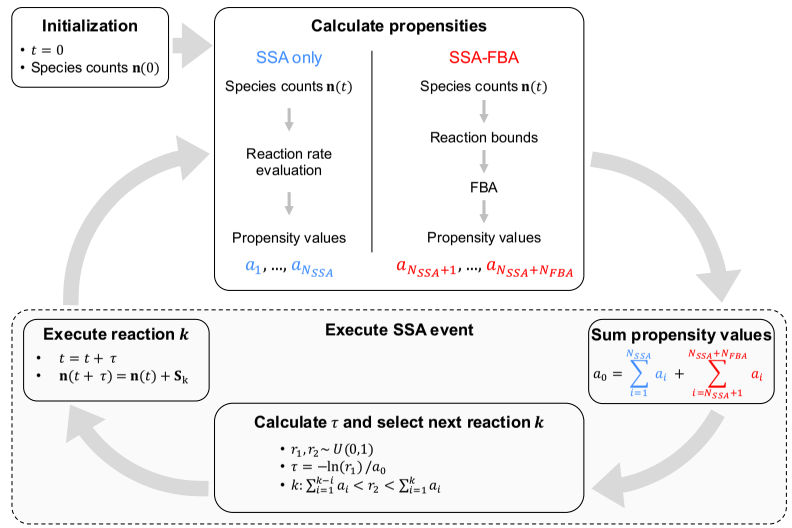
Set *event* to zero

- Calculate *optimal basis* using FBA

While *event* less than *maximum events* **do**

- Calculate SSA-only propensities
- **If** *optimal basis* is invalid **do**
  - Calculate *optimal basis* using FBA
- Calculate SSA-FBA propensities using *optimal basis*
- Execute SSA event
- **Increment** *event*

(b) Pseudo code for the direct, approximate and optimal basis SSA-FBA simulation methods. The approximate method speeds up SSA-FBA simulation by only calculating SSA-FBA propensity values at a fixed time interval (represented by Δ_{event} in main text). The optimal basis method speeds up SSA-FBA without approximation by only calculating SSA-FBA propensity values when the optimal basis has changed.



(c) Following initialisation, each iteration of an SSA-FBA simulation first calculates the propensity values for SSA only reactions (blue) by direct evaluation of reaction rate functions (as in Gillespie's original algorithm, [30, 31]) and calculates optimal propensity values for SSA-FBA reactions (red) using an optimal solution of the FBA model, where current species counts are used in reaction bounds. Next, both types of propensity values are combined to select the next reaction and its execution time. Finally, species counts are updated according to the stoichiometry of the selected reaction.

Figure 2: Overview of SSA-FBA model structure and simulation algorithms.

[32] used in the simulation of whole-cell models [28, 29]. To evaluate the performance of the direct and approximate simulation methods, we designed a toy model that contained only two SSA only reactions (R0 and R1), one SSA-FBA reaction (R2, growth rate of *Mycoplasma genitalium*) and one variable FBA bound (on the oxygen transport reaction). The toy model contained three species S_0, S_1, S_2 and is represented by the reaction schema

$$S_0 \underset{\text{R1}}{\overset{\text{R0}}{\rightleftarrows}} S_1 \overset{\text{R2}}{\rightarrow} S_2. \tag{3}$$

In this schema, the propensity value of reaction R2 depends indirectly on the count of species S_1, which bounds the maximal rate of the oxygen uptake reaction in the metabolic network model of *M. genitalium* [28] based on the functional form $u(S_1) = S_1/10$.

We began simulations with initial species counts $S_0 = 1000$, $S_1 = 0$, $S_2 = 0$, and simulated the toy model across 3000 reaction execution events using either the direct or approximate SSA-FBA simulation method. In particular, we used the toy model to explore how the choice of Δ_{event} affects results of SSA-FBA simulations using the approximate method. When Δ_{event} is small, the variances of approximate SSA-FBA simulations are comparable to those obtained using the direct method (Supplementary Figures S1a and S1b) and increasing Δ_{event} increases this variance (Supplementary Figure S1c). The toy model highlighted a shortcoming of the approximate SSA-FBA simulation method that trajectories occasionally (particularly for larger Δ_{event} as visible in Supplementary Figure S1c, but observed for all Δ_{event} tested) display numerical instability and diverge from realistic values towards the latter end of a simulation. This type of numerical instability also arises during direct integration of many DFBA models [39], but is particularly problematic for SSA-FBA because its stochastic nature means the pathology cannot be conclusively ruled-out on the basis of trial simulations (for example, to establish optimal size of Δ_{event}). The origin of the problem is that the embedded FBA problem can become infeasible within an integration interval, which induces a closed domain of definition for the dynamic system [39].

We therefore searched for a strategy to reduce the computational cost of SSA-FBA simulation

without the loss of accuracy and numerical stability associated with approximation. Motivated by a recent method for numerical integration of ODEs with embedded LP problems [39], we developed a more efficient algorithm that leverages the fact that FBA problems only need to be resolved when their optimal basis changes (right column in Figure 2b and Supplementary Appendix S2), which is typically much less frequently that the firing of each SSA reaction. The optimal basis simulation method is mathematically equivalent to the direct SSA-FBA simulation method, but substantially faster because it requires solving many fewer LP problems. Supplementary Figure S1d displays representative results of the comparisons between the fast and direct SSA-FBA simulation methods, showing that the former improves run time by an order of magnitude over the latter. The fast optimal basis algorithm also scales better in the number of reaction execution events than the direct implementation.

Case study: dynamics of ATP in *M. pneumoniae*

In this section, we present results from a case study using SSA-FBA to simulate the metabolism of a single *M. pneumoniae* cell. *Mycoplasma* have the smallest genomes among known freely living cells, and *M. pneumoniae* is one of the best-characterised members of this family [40], making it an excellent case study for understanding the metabolism of a single cell. Its close relative *M. genitalium* was previously used to build a whole-cell model [28].

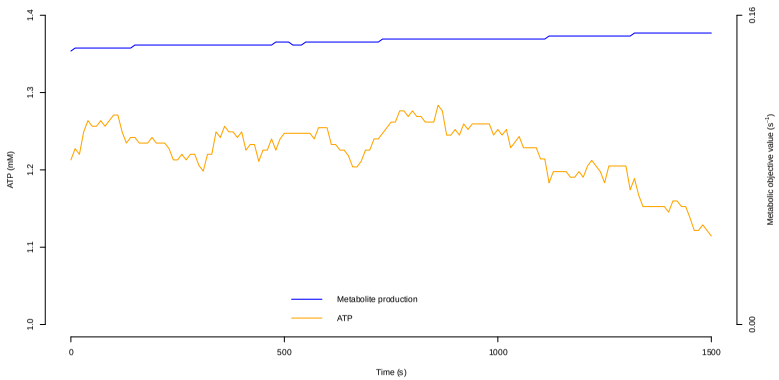
We constructed a model of *M. pneumoniae* consisting of 505 biochemical reactions. Most parameter values were derived from metabolomics, transcriptomics, and proteomics data about *M. pneumoniae* [41, 42, 43, 44], with the exception of rate constants for metabolic reactions (for which many fewer experimental observations are available) that are drawn from a variety of other bacteria [45]. The model contains a metabolic reaction network with 86 metabolic reactions, including a metabolite production pseudo- reaction that constitutes the objective function of the embedded FBA problem. There are 81 protein species that either function individually or in complex (27 complexes in total) as enzymes or transporters regulating flux through the FBA reactions, or serve a direct role in gene expression (ribosomal proteins, RNA polymerase or RNAse subunits). Fur-

thermore, each protein species is associated with an mRNA molecule (in addition to the three ribosomal RNAs) and four reactions corresponding to RNA transcription, RNA degradation, protein synthesis and protein degradation, which together with complexation reactions makes 420 SSA only reactions. We designated species on the periphery of the metabolic reaction network that appear as reactants in the Michaelis-Menten-like rate laws for SSA only reactions (nucleoside triphosphates and amino acids) as external species. Extracellular substrates were also considered external species. Therefore, reactions of the embedded FBA model involved in the production or consumption of nucleoside triphosphates and amino acids or extracellular transport were assigned to the SSA-FBA reaction subset. The remaining reactions of the embedded FBA model made up the FBA only subset. Supplementary File S1 contains all of the compartments, species, reactions, rate laws, and parameters; Supplementary Appendix S3 provides an extended description of the model.

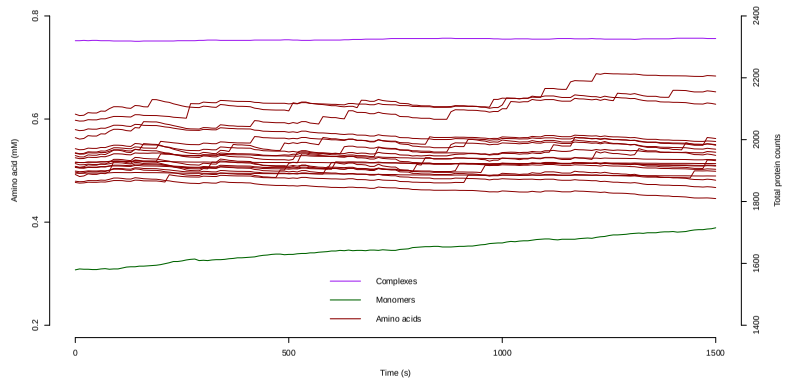
The single-cell model is based partly on metabolic rate constants measured for bacteria distantly related to *M. pneumoniae*. The average fold variation for wild-type bacterial enzymatic reaction rate constants (k_{cat} values) in BRENDA [45] is 3951.6, as measured by $(k_{cat}^{max} - k_{cat}^{min})/k_{cat}^{min}$, suggesting one or more of these model parameters require calibration in order to generate predictions consistent with the observed growth physiology of *M. pneumoniae*. We therefore parsimoniously searched metabolic parameter space by individually varying enzymatic reaction rate constants, identifying that growth rate is sensitive to the k_{cat} of guanylate kinase (GMK, gene MPN246). To generate each set of simulation results we ran a thousand instances of the single-cell *M. pneumoniae* model using the fast SSA-FBA method initialised with counts of species randomly sampled from a Poisson distribution parametrised by their mean values across the cell cycle. In each case we analysed simulations over a 25 min interval of biological time, across which cell volume is assumed to be approximately constant given the relatively long (6-8 hrs) doubling time of *M. pneumoniae* [40], excluding the first 25 mins of each 50 min simulation to prevent initial transients from contaminating the analysis. Figures 3a and 3b display results from an example simulation.

The k_{cat} value of GMK was varied from its lowest measured value (equal to the k_{cat} of adenylate kinase (ADK) in *Bacillus subtilis*, a much faster growing gram-positive bacteria), by increasing in 10-fold increments towards its highest estimated value (ADK in *Vibrio natriegens*). Notably, we found that the distribution of SSA execution events per simulation becomes bimodal as the GMK reaction rate constant increases (Supplementary Figure S2). This bimodal distribution computationally separates simulations into two distinct groups with a low (less than 10,000 SSA execution events) or high (more than 10,000 SSA execution events) number of SSA execution events per simulation, which are biologically associated with low or high time-averaged GTP concentrations: 0.04 ± 0.06 mM vs 0.47 ± 0.29 mM (mean $\pm$ SD) in low vs high group, respectively. For the largest value of the GMK reaction rate constant tested, simulations in the low GTP group (146 out of 1000 simulations) were associated with a relatively lower average proportion of SSA execution events corresponding to metabolic reactions compared to simulations in the high GTP group: 79.36 % vs 99.16 % in low GTP group vs high GTP group, respectively, implying that higher time-averaged GTP concentrations are associated with increased metabolic activity. Conversely, a relatively higher average proportion of SSA execution events correspond to translation reactions in the low GTP group compared to the high GTP group: 13.46 % compared with 0.50 % in low GTP group vs high GTP group, respectively, consistent with the fact that translation is a major consumer of intracellular GTP. No clear difference between randomised initial conditions between groups appears to govern the separation of simulations into low GTP (metabolically less active) and high GTP (metabolically more active) groups (Supplementary Figure S3), indicating that the bimodal nature of this distribution is possibly a feature of intrinsic network stochasticity coupled to FBA in a way that depends on the value of the GMK reaction rate constant.

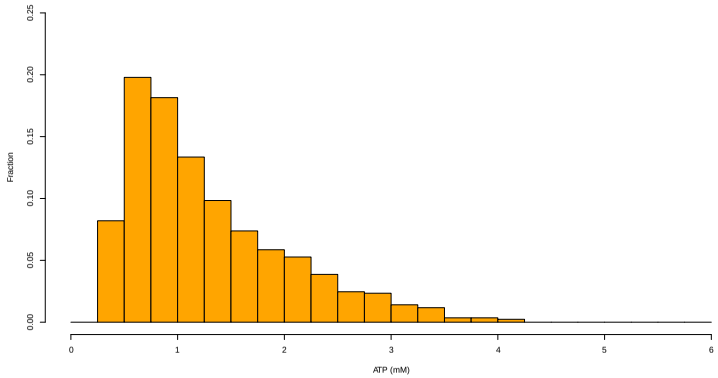
Given the critical importance of ATP as source of energy for intracellular reactions supporting growth, we studied the distribution of time-averaged ATP concentrations across the population of simulated cells. We found that the time-averaged concentrations ATP across the group of metabolically more active cells is positively skewed (Figure 3c, compare with Figure 4d in [46]) in agreement with experimental measurements of ATP concentrations inside individual bac-



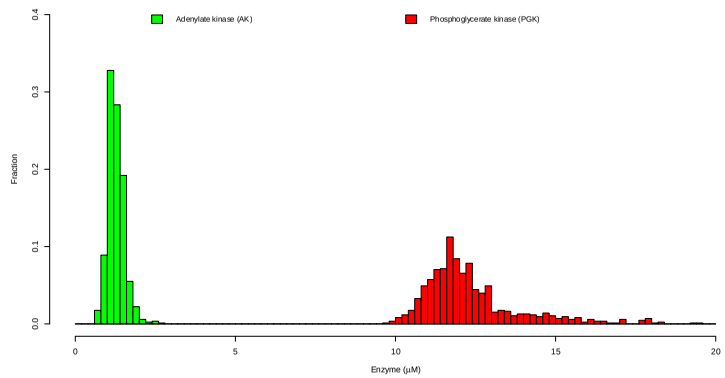
(a) Dynamics of intracellular ATP concentration and the flux through the metabolite production pseudo-reaction in representative simulation instance.



(b) Dynamics of intracellular concentrations of all twenty amino acids (average amino acid concretion in bold line) and total counts of all proteins (split by complex and monomers) in representative simulation instance.



(c) Distribution of time-averaged ATP concentrations.



(d) Distribution of time-averaged ADK and PGK concentrations.

Figure 3: Results from SSA-FBA simulations of *M. pneumoniae* single-cell model. In generating these plots, the first 1500 secs of each simulation were excluded from analysis to allow for initial transients to die out.

teria [46]: 1.28 $\pm$ 0.75 mM vs 1.54 $\pm$ 1.22 mM (mean $\pm$ SD) and skew values of 1.16 vs 2.20 for simulation vs experiment, respectively. Conversely, the distribution of ATP concentrations in metabolically less active cells was slightly negatively skewed: 0.81 $\pm$ 0.22 mM (mean $\pm$ SD) and skew value of -0.35. To explain the diversity of ATP concentrations across single cells, we focused on the enzymes that account for the majority of metabolic reaction execution events and found the most represented are phosphofructokinase (PFK; GTP consumption), phosphoglycerate kinase (PGK; ATP and GTP production), pyruvate kinase (PYK; GTP production), ADK (ATP production), GMK (GTP consumption) and the glucose import reaction. This confirms that central glycolysis and nucleoside triphosphate homeostasis are the major metabolic pathways active during simulation. We found a significant positive correlation between time-averaged ATP concentrations and ADK but not PGK (used as a representative for glycolysis) enzyme counts in the metabolically more active cells ($R = 0.18$, $p < 10^{-7}$ and $R = 0.04$, $p = 0.24$, respectively, based on Pearson's product moment correlation coefficient), but no correlations in metabolically less active cells. Moreover, time-averaged ADK and PGK concentrations were positively skewed across the metabolically more active group (1.24 $\pm$ 0.26 μM and 12.35 $\pm$ 1.52 μM (mean $\pm$ SD) and skew values 1.00 and 1.67, respectively, see also Figure 3d), suggesting that the asymmetric metabolite distributions observed both here and experimentally [46] are perhaps the result of fluctuating enzyme levels combined with the effect of reaction network coupling that serves to integrate and propagate multiple sources of stochasticity.

Discussion

In this paper we have presented SSA-FBA, the first framework for simulating the complete metabolic networks of single cells at the resolution of individual species and reactions. SSA-FBA is a hybrid method for embedding FBA in SSA that is well-suited to simulating metabolic networks for which detailed kinetic information is often lacking. We also have developed an advanced optimal basis algorithm for efficiently executing SSA-FBA without approximation. A case study using our

algorithm to simulate the metabolism of an individual *M. pneumoniae* cell demonstrates that SSA-FBA has the potential to reveal how stochasticity at the single-cell level contributes to metabolic heterogeneity at the population level. Our results have helped to identify the sources of variation observed in experimental measurements of important metabolites in single cells, indicating that reaction network coupling contributes to population heterogeneity by amplifying or attenuating the sources noise in an integrative fashion. SSA-FBA can therefore be used to probe metabolic heterogeneity in a manner complementary to alternative existing experimental and computational methods [11, 22, 25, 27].

Future extensions of our work will involve applying SSA-FBA to larger, more realistic models of entire cells, in addition to reconsidering how standard constraint-based formulations of metabolite pools and an objective function should be adapted to suit the biological nature of single-cell biology. More complex models of single-cell metabolism could include mechanisms that impart regulatory control of stochasticity [47], and pave the way for combining insights from simulation with experimental advances in microfluidics [48] or real-time quantification of RNA translation events within individual cells [49]. From an algorithmic perspective, SSA-FBA could be extended to incorporate additional time-scales governed by continuous stochastic or deterministic processes (e.g., [50]), and its computational efficiency perhaps further enhanced through parallelisation methods [51].

Materials and Methods

We implemented an SSA-FBA simulation package in C++ and Python. Further details on SSA-FBA, its implementation and the case study can be found in the Supplementary Appendix, and all code and data are freely available open-source at https://gitlab.com/davidtourigny/single-cell-fba.

Acknowledgements

Part of this work was completed while DST was a Simons Foundation Fellow of the Life Sciences Research Foundation hosted by Columbia University Irving Medical Center. The work of JRK was supported by National Science Foundation award 1649014 and National Institutes of Health award R35GM119771. We thank ME Beber and J Carrasco Muriel for their close collaboration with the first author on a related project that provided some of the motivation for the current study, and Maria Lluch-Senar, Veronica Llorens, and Luis Serrano for discussions on the physiology of Mycoplasmas.

References

[1] Schier AF (2020) Single-cell biology: beyond the sum of its parts. *Nat. Methods* 17: 17-20.

[2] Angerer P, Simon L, Tritschler S, Wolf FA, Fischer D, Theis FJ (2017) Single cells make big data: new challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* 4: 85-91.

[3] Stuart T, Satija R (2019) Integrative single-cell analysis. *Nat. Rev. Genet.* 20: 257-272.

[4] Ackermann M (2015) A functional perspective on phenotypic heterogeneity in microorganisms. *Nat. Rev. Microbiol.* 13: 497-508.

[5] Schreiber F, Ackermann M (2020) Environmental drivers of metabolic heterogeneity in clonal microbial populations. *Curr. Opin. Biotechnol.* 62: 202-211.

[6] Dagogo-Jack I, Shaw AT (2018) Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* 15: 81-94.

[7] Xiao Z, Locasale JW, Dai Z (2020) Metabolism in the tumor microenvironment: insights from single-cell analysis. *Oncoimmunology* 9: 1726556.

[8] Rubakhin SS, Lanni EJ, Sweedler JV (2013) Progress toward single cell metabolomics. *Curr. Opin. Biotechnol.* 24: 95-104.

[9] Fessenden M (2016) Metabolomics: small molecules, single cells. *Nature* 540: 153-155.

[10] Emara S, Amer S, Ali A, Abouleila Y, Oga A, Masujima T (2017) Single-Cell Metabolomics. In: Sussulini A. (eds) *Metabolomics: From Fundamentals to Clinical Applications. Advances in Experimental Medicine and Biology*, vol 965. Springer, Cham.

[11] Evers TMJ, Hochane M, Tans SJ, Heeren RMA, Semrau S, Nemes P, Mashaghi A (2019) Deciphering metabolic heterogeneity by single-cell analysis. *Anal. Chem.* 91: 13314-13323.

[12] Duncan KD, Fyrestam J, Lanekoff I (2019) Advances in mass spectrometry based single-cell metabolomics. *Analyst.* 144: 782-793.

[13] Damiani C, Gaglio D, Sacco E, Alberghina L, Vanoni M (2020) Systems metabolomics: from metabolomic snapshots to design principles. *Curr. Opin. Biotechnol.* 63: 190-199.

[14] Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell.* 144: 646-674.

[15] Pavlova NN, Thompson CB (2016) The emerging hallmarks of cancer metabolism. *Cell Metab.* 12: 27-47.

[16] Edwards JS, Covert M, Palsson B (2002) Metabolic modelling of microbes: the flux-balance approach. Environ Microbiol. 4: 133-140.

[17] O'Brien EJ, Monk JM, Palsson BO (2015) Using genome-scale models to predict biological capabilities. *Cell* 161: 971-987.

[18] Angione C (2019) Human systems biology and metabolic modelling: a review- from disease metabolism to precision medicine. *Biomed Res. Int.* 2019: 8304260.

[19] Yang L, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO (2019) DynamicME: dynamic simulation and refinement of integrated models of metabolism and protein expression. *BMC Syst Biol.* 13: 2.

[20] Waldherr S, Oyarzún DA, Bockmayr A (2015) Dynamic optimization of metabolic networks coupled with gene expression. *J. Theor. Biol.* 365: 469-485.

[21] Kiviet DJ, Nghe P, Walker N, Boulineau S, Sunderlikova V, Tans SJ (2014) Stochasticity of metabolism and growth at the single-cell level. *Nature* 514: 376-379.

[22] Wehrens M, Büke F, Nghe P, Tans SJ (2018) Stochasticity in cellular metabolism and growth: approaches and consequences. *Curr. Opin. Syst. Biol.* 8: 131-136.

[23] Damiani C, Maspero D, Di Filippo M, Colombo R, Pescini D, Graudenzi A, Westerhoff HV, Alberghina L, Vanoni M, Mauri G (2019) Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *PLoS Comput. Biol.* 15: e1006733.

[24] Levine E, Hwa T (2007) Stochastic fluctuations in metabolic pathways. *Proc. Natl. Acad. Sci. USA.* 104: 9224-9229.

[25] Tonn MK, Thomas P, Barahona M, Oyarzún DA (2019) Stochastic modelling reveals mechanisms of metabolic heterogeneity. *Commun. Biol.* 2: 108.

[26] Thomas P, Terradot G, Danos V, Weiße AY (2018) Sources, propagation and consequences of stochasticity in cellular growth. *Nat. Commun.* 9: 4528.

[27] Tonn MK, Thomas P, Barahona M, Oyarzún DA (2020) Computation of single-cell metabolite distributions using mixture models. bioRxiv 2020.10.07.329342

[28] Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B Jr., Assad-Garcia N, Glass JI, Covert MW (2012) A whole-cell computational model predicts phenotype from genotype. *Cell* 150: 389-401.

[29] Karr JR, Takahashi K, Funahashi A (2015) The principles of whole-cell modeling. *Curr. Opin. Microbiol.* 27: 18-24.

[30] Gillespie DT (1976) General method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* 22: 403-434.

[31] Gillespie DT (1977) Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* 81: 2340-2361.

[32] Mahadevan R, Edwards JS, Doyle FJ 3rd (2002) Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys. J.* 83: 1331-1340.

[33] Elowitz MB, Levine AJ, Siggia ED (2002) Stochastic gene expression in a single cell. *Science* 279: 1183-1186.

[34] Gillespie DT (1992) A rigorous derivation of the chemical master equation. *Physica A* 188: 404-425.

[35] Rao CV, Arkin AP (2003) Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *J. Chem. Phys.* 118: 4999-5010.

[36] Cao Y, Gillespie DT, Petzold LR (2005) The slow-scale stochastic simulation algorithm. *J. Chem. Phys.* 122: 014116.

[37] Cao Y, Gillespie DT, Petzold LR (2005) Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *J. Comput. Phys.* 206: 395-411.

[38] Gibson, MA, Bruck, J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A* 104: 1876-1889.

[39] Harwood SM, Höffner K, Barton PI (2016) Efficient solution of ordinary differential equations with a parametric lexicographic linear program embedded. *Numer. Math.* 133: 623-653.

[40] Waites KB, Talkington DF (2004) *Mycoplasma pneumoniae* and its role as a human pathogen. *Clin. Microbiol. Rev.* 17: 697-728.

[41] Güell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kühner S, Rode M, Suyama M, Schmidt S, Gavin A-C, Bork P, Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. *Science* 326: 1268-1271.

[42] Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castaño-Diez D, Chen W-H, Devos D, Güell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Böttcher B, Frangakis AS, Russell RB, Serrano L, Bork P, Gavin A-C (2009) Proteome organization in a genome-reduced bacterium. *Science* 326: 1235-1240.

[43] Maier T, Schmidt A, Güell M, Kühner S, Gavin AC, Aebersold R, & Serrano L (2011) Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol. Syst. Biol.* 7: 511.

[44] Yus E, Lloréns-Rico V, Martínez S, Gallo C, Eilers H, Blötz C, Stülke J, Lluch-Senar M, Serrano L (2019) Determination of the gene regulatory network of a genome-reduced bacterium highlights alternative regulation independent of transcription factors. *Cell Syst.* 9: 143-158.

[45] Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Söhngen C, Stelzer M, Thiele J, Schomburg D (2011) BRENDA, the enzyme information system in 2011. Nucleic. Acids Res. 39: D670-D676.

[46] Yaginuma H, Kawai S, Tabata K, Tomiyama K, Kakizuka A, Komatsuzaki T, Noji H, Imamura H (2014) Diversity in ATP concentrations in a single bacterial cell population revealed by quantitative single-cell imaging. *Sci. Rep.* 4: 6522.

[47] Borri A, Palumbo P, Singh A (2016) Impact of negative feedback in metabolic noise propagation. *IET Syst. Biol.* 10: 179-186.

[48] Leygeber M, Lindemann D, Sachs CC, Kaganovitch E, Wiechert W, Nöh K, Kohlheyer D (2019) Analyzing microbial population heterogeneity-expanding the toolbox of microfluidic single-sell cultivations. *J. Mol. Biol.* 431: 4569-4588.

[49] Morisaki T, Lyon K, DeLuca KF, DeLuca JG, English BP, Zhang Z, Lavis LD, Grimm JB, Viswanathan S, Looger LL, Lionnet T, Stasevich TJ (2016) Real-time quantification of single RNA translation dynamics in living cells. *Science* 352: 1425-1429.

[50] Haseltine EL, Rawlings JB (2002) Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.* 117: 6959.

[51] Goldberg AP, Jefferson DR, Sekar JAP, Karr JR (2020) Exact parallelization of the stochastic simulation algorithm for scalable simulation of large biochemical networks. *arXiv preprint* https://arxiv.org/abs/2005.05295