frontiers | Frontiers in Cell and Developmental Biology

TYPE Review
PUBLISHED 07 November 2023
DOI 10.3389/fcell.2023.1260507

OPEN ACCESS

EDITED BY
Michael Blinov,
UCONN Health, United States

REVIEWED BY
Zaida Ann Luthey-Schulten,
University of Illinois at Urbana-Champaign, United States
Markus Covert,
Stanford University, United States

*CORRESPONDENCE
Ali Navid,
navid1@llnl.gov

RECEIVED 18 July 2023
ACCEPTED 19 October 2023
PUBLISHED 07 November 2023

CITATION
Georgouli K, Yeom J, Blake RC and Navid A (2023), Multi-scale models of whole cells: progress and challenges. *Front. Cell Dev. Biol.* 11:1260507. doi: 10.3389/fcell.2023.1260507

Multi-scale models of whole cells: progress and challenges

Konstantia Georgouli[1], Jae-Seung Yeom[2], Robert C. Blake[2] and Ali Navid[1]*

[1]Biosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, United States, [2]Center for Applied Scientific Computing, Computing Directorate, Lawrence Livermore National Laboratory, Livermore, CA, United States

Whole-cell modeling is "the ultimate goal" of computational systems biology and "a grand challenge for 21st century" (Tomita, Trends in Biotechnology, 2001, 19(6), 205–10). These complex, highly detailed models account for the activity of every molecule in a cell and serve as comprehensive knowledgebases for the modeled system. Their scope and utility far surpass those of other systems models. In fact, whole-cell models (WCMs) are an amalgam of several types of "system" models. The models are simulated using a hybrid modeling method where the appropriate mathematical methods for each biological process are used to simulate their behavior. Given the complexity of the models, the process of developing and curating these models is labor-intensive and to date only a handful of these models have been developed. While whole-cell models provide valuable and novel biological insights, and to date have identified some novel biological phenomena, their most important contribution has been to highlight the discrepancy between available data and observations that are used for the parametrization and validation of complex biological models. Another realization has been that current whole-cell modeling simulators are slow and to run models that mimic more complex (e.g., multi-cellular) biosystems, those need to be executed in an accelerated fashion on high-performance computing platforms. In this manuscript, we review the progress of whole-cell modeling to date and discuss some of the ways that they can be improved.

KEYWORDS

whole-cell modeling, systems biology, multi-scale models, data integration, high performance computing

1 Introduction

Biology once was considered a data poor science. That era has long passed. Today, thanks to revolutionary advances in sequencing and other high-throughput analytical techniques, staggering amount of biological data is being collected (Marx, 2013). Soon the cost of storing and analyzing the biological data could be more concerning than the cost of generating it (Fritz et al., 2011; Berger et al., 2013; Jagadish et al., 2014; Stephens et al., 2015). Further complicating the challenge, the data that is being generated is highly heterogeneous. The data is also variable. At times, measurements from the same biosystem but from different groups, or even the same group but on different days or on different instruments could disagree with one another. Therefore, data processing and integration from widely diverse databases have become important tasks during *in silico* systematic analyses (Bajcsy et al., 2005; Shamim et al., 2010).

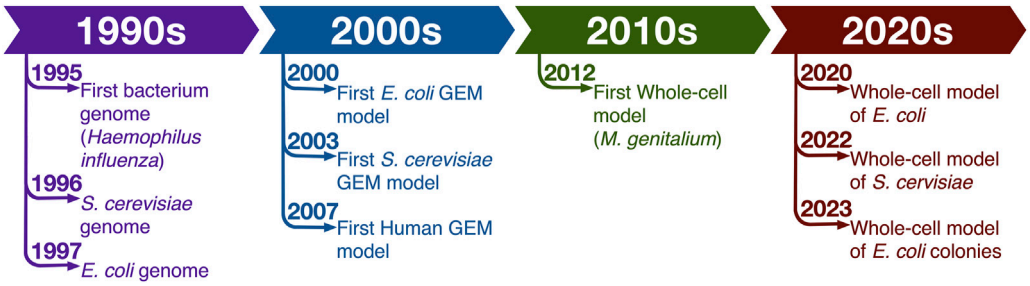

FIGURE 1
Timeline of some of the important milestones in development of whole-cell models.

2 Whole-cell models

French polymath René Descartes in his Discourses put forth the idea that the world behaves like a clockwork machine and therefore it can be understood by dividing it into smaller pieces and studying the individual components (Descartes, 1984). Molecular biology investigations followed this idea for most of 20th century. But while reductionist studies dominated the field and provided invaluable insights into workings of specific processes in various model organisms, the Aristotelian view that "the totality is not, as it were, a mere heap, but the whole is something besides the parts" (Cohen and Reeve, 2000) always had advocates among biologists. These detractors observed the emergent behavior of whole systems and argued that the observations that structures of systems organized and controlled the performance of the component parts refuted the reductionist basis of many studies since they failed to account for critical system-level orchestrations. For a long time, holistic analyses were impossible due to absence of system-level data. That shortcoming has now been overcome and the ready availability of various types of omics data have led to a renaissance in the field of systems biology (Figure 1).

Soon after first genomes became available, computational system-level models were developed. Genome-scale models of metabolism (GEMs) are among the most widely used system-level models. Metabolism was chosen as one of the first bioprocesses to be examined on a system-level thanks to tireless efforts of biochemists and microbiologists who for generations conducted extensive targeted mechanistic analyses of enzymes and pathways (Hill, 1970; Schilling et al., 1999; Papin et al., 2003; Cornish-Bowden, 2013; Johnson, 2013) and bioinformaticians who processed and deposited this information in numerous databases.

Coupling of GEMs with constraint-based reconstruction and analysis (COBRA) methods such as popular Flux Balance Analysis (FBA) has provided a wealth of general information regarding fundamental organization and function of metabolic pathways (e.g., (Almaas et al., 2004; Almaas et al., 2005)) while on a biosystem specific level it has shed light on the metabolic capabilities of the modeled organisms, their environmental niches and the robustness of their metabolism to environmental and genetic perturbations.

The popularity of these constraint-based modeling approaches stems from the fact that they utilize the data that is readily available (annotated genomes, empirical measurements of growth, nutrient uptake, and byproduct excretion) and circumvent the issue of dearth of kinetic data that plague generation of system-level kinetic models. Some system-level kinetic models have been developed e.g., (Klipp, 2007; Bordbar et al., 2015; Jamei, 2016), but they usually tend to account for the activity of significantly fewer genes than COBRA models due to a lack of detailed kinetic data for all cellular processes. There have been many methods developed that use Bayesian parameter estimation to predict reasonable thermodynamic and kinetic values to constrain COBRA models e.g., (Liebermeister and Klipp, 2006a; Liebermeister and Klipp, 2006b; Stanford et al., 2013) and subsequently there have been a number of attempts to add kinetic information to FBA models (e.g., (Jamshidi and Palsson, 2008; Adadi et al., 2012; Stanford et al., 2013; Chowdhury et al., 2015; Pozo et al., 2015; Khodayari and Maranas, 2016; Sánchez et al., 2017; Shameer et al., 2022)). Despite this progress, currently the vast majority of FBA models do not contain kinetic information.

Given their wide range of uses many upgrades to FBA methods have been made to incorporate heterogenous omics data into them. Many methods have been developed that constrain COBRA models with omics data other than genome (e.g., (Becker and Palsson, 2008; Chandrasekaran and Price, 2010; Zur et al., 2010; Jensen and Papin, 2011; Fang et al., 2012; Navid and Almaas, 2012; Sánchez et al., 2017; Bekiaris and Klamt, 2020; Hadadi et al., 2020; Di Filippo et al., 2022)). Several methods have also been developed that analyze multi-omics data using machine learning models prior to their incorporation into FBA models (Kim et al., 2016; Zampieri et al., 2019; Lewis and Kemp, 2021; Sahu et al., 2021). In one case, FBA was embedded into artificial neural networks resulting in a hybrid mechanistic-machine learning model that allows quantitative predictions of medium uptake fluxes based solely on medium composition (Faure et al., 2023). This development could greatly improve our ability to develop condition- and species-specific GEMs using data that are more readily available and easier to access.

There are also models available that account for the sequence-specific synthesis of gene products, their function and all catalyzed biochemical processes (Thiele et al., 2012; Ma et al., 2017). However, despite all these advances in COBRA modeling, all GEM models and upgraded variants do not fully account for activity of every known biological molecule and process. It is also important to account for the structure of the cell since most molecular processes use it to collocate into interacting modules at multiple scales (Betts and Russell, 2007). While GEMs for eukaryotes bin the reactions of metabolic reconstructions into different cellular compartments, they

do not explicitly account for clustering of molecules and proteins within prokaryotes or organelles in a manner that could explain observed interacting units. Additionally, most GEMs contain many sources or sinks of energy and metabolites which hinder accurate and detailed description of mechanisms associated with homeostasis in a system (Roberts, 2014). Whole-cell models aim to overcome these limitations.

Whole-cell models, as with other "system-level" models aim to predict cellular phenotypes from genotype and biochemical and biophysical characteristics of the environment. Where WCM supersedes the other modeling efforts is the ambitious goal of incorporating the function of each gene, gene product, and metabolite in the modeled system (Karr et al., 2015). Thus, WCMs serve as nearly comprehensive knowledgebases for the modeled system. They allow *in silico* experiments that can lead to prediction of novel biological phenomena, identification of gaps in our knowledge, generation of new hypotheses and design of new studies (Tomita, 2001). The models can be easily updated with new information which can be a quick way of ascertaining the significance of new discoveries. Also, in this golden age of machine learning, regression techniques can be used to examine large heterogenous biological datasets and with a relatively high degree of accuracy predict phenotypes (Guzzetta et al., 2010; Smith et al., 2020; Guo and Li, 2023); in fact WCMs are the ideal complementary models to the black box nature of machine learning models and can provide a mechanistic underpinning to the predicted phenotypes.

2.1 Whole-cell model of *Mycoplasma genitalium*

The first whole-cell model, one that can reasonably claim to incorporate the activity of nearly all molecules in a system, was developed for the small bacterium *M. genitalium* (Karr et al., 2012). *M. genitalium* is a facultative anaerobic pathogen that can cause sexually transmitted diseases. In men it causes nongonococcal urethritis and in women it could cause a variety of ailments including cervicitis, endometritis, pelvic inflammation, infertility, and even unfavorable birth outcomes.

Although *M. genitalium* (MG) does have some medical significance, the main reason why it was chosen as the first organism for development of a WCM was that it has one of the smallest known genomes (~580 kb and 480 coded proteins) (Fraser et al., 1995). Also, compared to other genomes, including well studied model organisms like *E. coli*, MG's genome contains significantly fewer genes of unknown function. Despite its small size and complexity, the development of the MG model was still a monumental undertaking and was a very labor-intensive process. The model contains 1900 parameters from over 900 publications and is nearly 3000 pages of Matlab code. It divides the activity of all annotated MG gene products into 28 subcellular processes. To ensure the most accurate representation and simulation, the most appropriate mathematical modeling method was used for each subcellular process. To link all these disparate models together, the developers devised a hybrid modeling approach where all 28 mathematical modules are linked to a subset of other modules via 16 cell variables. Metabolism in the MG WCM uses similar metabolic reconstructions as GEMs; however, the internal fluxes of the reactions are dynamically constrained by multiplying the amount of catalyzing enzyme present in the system (a variable in the WCM) by its catalytic constant (k_{cat}).

The simulation starts with an initial set of values for these variables. All the modules then run for a set period (e.g., 1 s) and afterwards the value of each cell variable is updated based on input from all the modules that link to it (Figure 2). Once the variables have been updated, the modules are run again using the new values. The process continues until a preset biological objective has been accomplished. Given the complexity of the problem, the amount of data that needed to be transferred back and forth between variables and modules, and the inefficiency of the solver, the simulation time for the original MG model was slow (~1 day for 1 cell cycle). The model provided some interesting insights into working of MG and predicted some novel phenotypes.

In cases where experimental results and model predictions disagreed, gaps in our knowledge were identified and some parameter values were corrected (Karr et al., 2012). This type of model-driven knowledge gap filling and correction is a strong suit of WCMs. For example, the MG WCM was used in a follow up work by Sanghvi and coworkers (Sanghvi et al., 2013) to compare the WCM predicted growth rates for all non-lethal single-gene deletions with experimental data. In cases of quantitative disagreement between model predictions and experimental measurements, the authors examined the "molecular pathology" of each gene-deleted strain and identified gene targets which during the genome annotation process had been wrongly assigned a function or had a missing function that was not included in the model. In some other cases they identified alternate metabolic pathways that could compensate for loss of a gene product. Finally, given the more quantitative nature of WCM (in comparison to FBA models) due to their incorporation of kinetic data into their metabolic simulations; the authors were able to use the quantitative differences between model predictions and experiments to predict appropriate kinetic parameters for several critical enzymes. The predicted values were experimentally validated. Comparing the new measured values with the literature data that originally was used to train the MG WCM showed significant differences, in some cases up to four orders of magnitude.

The ability of WCMs to reliably predict in a quantitative manner the *in vivo* dynamics of a system; information that cannot easily be measured but is invaluable for assessing the state of a system and guiding efforts to alter it, makes WCMs critical tools for biological engineering projects. For example, WCMs can provide invaluable information about how incorporating synthetic gene circuits in an organism could alter the working of the system and how internal processes that are almost always unaccounted for *in silico* models can divert the system behavior away from desired outcome. In this vein, Purcell et al. (2013) used the MG WCM to examine the effects of adding genes into MG. They also examined how codon usage affects gene expression and in agreement with results from *E. coli* (Kudla et al., 2009). They found no difference in expression rates. Recently (Rees-Garbutt et al., 2020) have used the MG WCM within a design-simulate-test framework to predict a minimal genome that (if biologically correct) could be smaller than *JCVI-Syn3.0* minimal genome bacterium.

3 Progress

3.1 Whole-cell model of *Escherichia coli*

While the development of MG whole-cell model (WC-MG) was a monumental achievement and has been used to highlight the

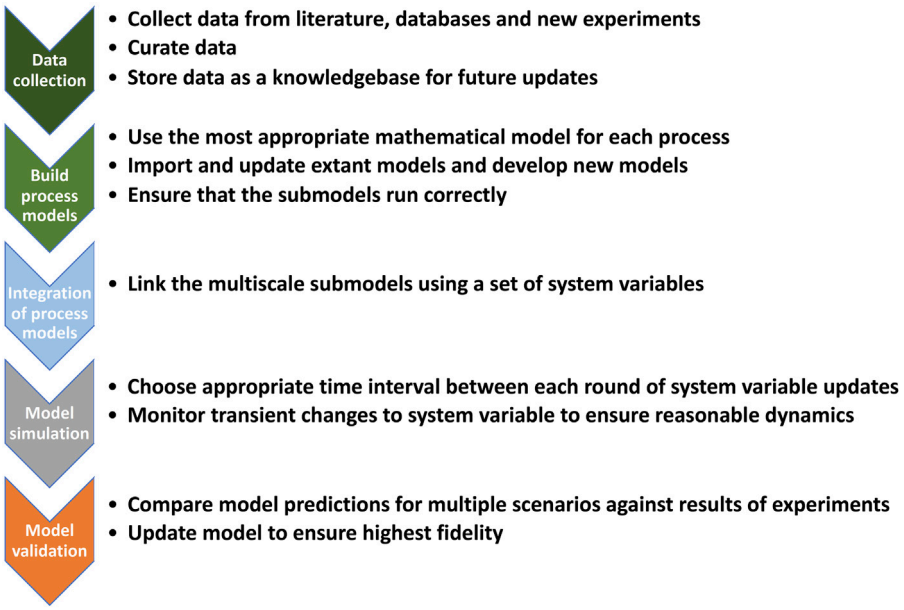

FIGURE 2
Assembly process for whole-cell models.

immense potential of WCM for a variety of important uses, WC-MG has limited utility for common uses of *in silico* models such as predicting targets or outcomes for bioengineering. To have that ability, the logical next organism to be modeled needed to be the best studied bioengineering chassis organism, namely, *E. coli*. To that end, a hybrid multi-math, multi-scale model for *E. coli* has been developed (WC-EC) (Macklin et al., 2020). It incorporates the function of over 40% of the well-annotated genes in *E. coli* genome (1,214 genes). Although the model does not account for activity of every gene product in *E. coli*, the model is significantly larger than the WC-MG (>10,000 mathematical equations and >19,000 parameters). This is not surprising given that *E. coli*'s genome is an order of magnitude larger than MG's and *E. coli* has nearly 50 times more molecules. *E. coli*'s metabolism and regulatory mechanisms are also significantly more sophisticated than those for MG. Another advantage of WC-EC over WC-MG is that 100% of former's parameters are derived from experimental measurements compared to less than 30% of the WC-MG parameters. The WC-EC, in addition to omics data, is informed by a large amount of kinetic data. This data was collected from 1,200 hand-curated papers after reviewing 12,000 papers in the BRENDA (Schomburg et al., 2002; Chang et al., 2009) database. The fact that all the parameters in WC-EC are empirically measured allowed its use for examining the cross-consistency between the disparate data sources that were used for its parameterization. The results of analyses showed that most of the data used for the development of WC-EC were consistent with predicted behaviors. However, parameter sets that were not consistent resulted in discrepancies that were alarming. For example, the incorporated data for rate of activity by ribosomes and RNA polymerases were too low to result in measured growth rates. Another interesting finding was that some essential genes are not transcribed during division cycles and yet cells proliferate. This latter finding is a strong reminder that besides the catalytic capability and concentration of an enzyme, the time course of its production and eventual degradation can also have a significant effect on the robustness of a system to environmental and genetic perturbations.

After the publication of WC-EC, its creators have initiated the *E. coli* whole-cell modeling project (Sun et al., 2021). The project aims to expand on the published WC-EC model and ultimately develop the most detailed model *E. coli* ever. The project invites input and collaboration from the scientific community to accelerate the development process. As part of this effort, updated versions of WC-EC have been developed. One update (Ahn-Horst et al., 2022) incorporates additional growth rate control regulations such as global regulator guanosine tetraphosphate, as well as dynamics of amino acid biosynthesis and translation. The additions significantly improve the WC-EC's ability to simulate dynamics of cellular responses as a response to environmental perturbations. Another update (Choi and Covert, 2023) added accurate tRNA aminoacylation, codon-based polypeptide elongation, and N-terminal methionine cleavage mechanisms to WC-EC which permits better examination of inconsistencies between different types of measurements. The updated model was used to verify that *in vitro* tRNA aminoacylation measurements are insufficient for cellular proteome maintenance. The model predicted a positive feedback mechanism that regulates arginine synthesis.

3.2 Whole-cell model of Saccharomyces cerevisiae

Saccharomyces cerevisiae's (SC, Brewer's yeast) genome was the first eukaryotic genome to be sequenced (Goffeau et al., 1996). SC is

an extremely important organism economically. It is genetically tractable and has been engineered through a plethora of homologous recombination techniques. Overall, SC is the best studied single cell eukaryotic organism. Given this distinction, SC was the obvious best choice for developing the first whole-cell model of a multi-compartmented organism. The yeast whole-cell model (WM_S288C) (Ye et al., 2020) was developed by expanding upon an earlier FBA model of the organism (Österlund et al., 2013). It incorporates products of 6,447 genes (100% of genome), 975 metabolites and 6,156 reactions. Overall, it includes 26 cellular processes. Unlike WC-EC, not all incorporated parameters were available from yeast experiments. So instead, measurements from other organisms were used. The WM_S288C's predictions were validated against experimental results and when compared against predictions from its progenitor FBA model they showed significant improvement (e.g., precision of accurately predicting essential genes WM_S288C 70%, FBA model 28%). The developers used the model to conduct an extensive study of roles of various molecules in the system. They ascertained the function of 1,140 essential genes, thus providing a mechanistic understanding of vulnerable processes under different conditions. They also gained new insights into function of non-essential genes, namely, that these genes can regulate nucleotide concentrations and thus affect cellular growth rates.

3.3 Vivarium

As noted earlier, whole-cell models integrate a diverse set of intracellular processes using numerous simulation methods. When developing the first whole-cell model, accuracy and completeness were primary considerations. Speed of simulation was a secondary consideration. However, (Karr et al., 2012), did attempt to speed up the whole-cell simulation by executing multiple pathway sub-models simultaneously for the agreed simulation time interval using multiple CPU cores with one per pathway in Matlab (Gunawardena, 2012). This attempt exposed a few significant challenges to speeding up simulations of hybrid models. Firstly, the time interval for all pathways is restricted by the smallest time interval needed by any individual pathway. Secondly, the level of parallelism is limited by the number of pathways. Thirdly, the pathways tend to be extremely heterogeneous in terms of the computational work needed to advance within the selected time interval. Consequently, simulating the same interval for different pathways may require vastly different computing times, making the parallelization essentially ineffective.

To answer some of these problems, Vivarium (Agmon et al., 2022), a platform for integrative multi-scale modeling, has been developed. It provides an interface for combining existing models in the nested hierarchies of multiple scales via a discrete event simulation engine. This eases the software engineering task of combining smaller pathways into a larger whole-cell model. Vivarium makes it easier to combine multiple pathways together and thus allows larger models and more computational parallelism. Vivarium offers utilities to partition molecular species shared between pathways based on expected demand in such a way that mass is conserved. In this way, individual pathways can run independently from each other within a time interval. Vivarium can also leverage the message-passing of the Python multiprocessing module to exploit the inherent parallelism in the model across multiple cores and multiple processors. While the original version of Vivarium faced some of the same limitations as the original WCM models—linked timesteps, parallelism by pathways, and uneven computational load between pathways but updates have been made and are on the way that answer some of these issues (Skalnik et al., 2023).

3.4 Unbalanced growth and non-steady-state metabolism

In all WCMs developed so far, metabolism is solved using updated variants of FBA method that account for each enzyme's abundance and catalytic rate constant. Typical FBA models use a rigid biomass reaction where a single set of stoichiometric coefficients define the ratio of reactants that are used for production of a set amount of biomass and a fixed set of coefficients to define the other byproducts of cell maintenance and replication (Orth et al., 2010). This balance growth assumption is valid for most conditions, particularly if one must assume a long-term analysis. However, for the development of WCMs where FBA models are integrated in a hybrid format to interact with dynamic simulations of bioprocesses with significantly shorter timescales, this assumption is problematic. To overcome this flaw, (Birch et al., 2014), developed two variations of FBA called flexible FBA (flexFBA) and time-linked FBA (tFBA) that when run simultaneously within WCMs improve the accuracy of model predictions. In flexFBA, the fixed ratios of biomass reactants have been removed in the objective function. This eliminates the classical assumption of balanced growth. In tFBA the ratios between the reactants and byproducts in the biomass equation are no longer fixed and thus the common steady-state growth constraint of classical FBA is eased. Using these methods for WCM allows for "short time" FBA which allows integration of output from different types of mathematical models.

3.5 Colony-scale whole system modeling

Phenotypic heterogeneity in a microbial community, particularly those that persist for more than one generation can have a significant impact resilience of a system to environmental changes and threats. Bacterial persistence, the phenomenon where genetically identical bacterial colonies behave heterogeneously to introduction of antibiotics is known to play a key role in development of antibiotic resistance in bacteria (Gefen and Balaban, 2009). The heterogenous differences could stem molecular processes, such as stochastic expression of antibiotic resistance genes (Akiyama and Kim, 2021). Mechanistic WCMs are ideal tools for gaining a system level understanding of these phenomena. But to gain a colony level perspective requires simulating many cells interacting with one another via a shared environment. Vivarium allows such multi-scale simulations and Skalnik et al. (2023) have used it to alter WC-EC model and develop the first colony level holistic model. The model was then run in parallel using cloud computing to study the emergence of antibiotic

resistance in *E. coli* when treated with two antibiotics with different modes of action.

4 Challenges

Despite all the advances and progress in the development of WCMs over the last decades, there are still persistent fundamental challenges that hinder not only the development of new models but also any efforts to develop computational tools for accelerating model simulation. In this section, we will discuss these challenges and propose possible solutions.

4.1 Data collection

As the aim of WCMs is to accurately and comprehensively predict the cell behavior, a huge amount of biological data is needed for model parameterization and validation. This need increases with the complexity and size of the cell (Babtie and Stumpf, 2017). The main challenge with efforts at gathering the needed data is ensuring that the publicly available data is in a useable format. This will allow easy identification, extraction, and aggregation of high-quality data. Unfortunately, the high dimensionality, the heterogeneity, and the lack of sufficient annotation of the data pose important challenges regarding their interpretation, and reusability. These challenges have led to calls for standardization of databases, simulation softwares and overall modeling standards (Waltemath et al., 2016).

Fortunately, a variety of tools and databases have been developed to facilitate the data collection and aggregation process. These tools also ease the burden of additional curation of data. For example, there are many repositories providing pathway/genome information such as BioCyc (Karp et al., 2017), BiGG (Schellenberger et al., 2010; King et al., 2015a), WholeCellKB (Karr et al., 2013), KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2004; Kanehisa et al., 2016) and BRENDA (Schomburg et al., 2002; Chang et al., 2009). In addition, there are databases that include experimental data for a specific organism, such as EcoCyc (Keseler et al., 2011; Keseler et al., 2017) where interestingly in its latest version (Karp et al., 2023) there is a bidirectional connection with the *E. coli* whole-cell modeling project that can be used for importing data from EcoCyc to parametrize the WCM and updating the WCM with EcoCyc's latest mechanistic information. Human curation of data collected on bioprocesses is key to developing accurate WCMs and to this end visualization of metabolic maps can provide extremely valuable insights for data integration. Network visualization tools such as Escher (King et al., 2015b; Rowe et al., 2018) and Pathview (Luo and Brouwer, 2013; Luo et al., 2017) can be used for this task. However, these tools rely on pre-drawn maps and cannot support inputs of large networks with multi-type models.

In cases when data have not been deposited in any database, literature text mining tools for extracting biological data like Integrated Network and Dynamical Reasoning Assembler (INDRA) (Gyori et al., 2017; Bachman et al., 2023), BioQRator (Kwon et al., 2014) and PubTator (Wei et al., 2013) can help with data collection and curation efforts. However, despite these resources, there are still a few problems that need to be addressed.

Some parameters still remain unknown or of poor quality. This is because while we have been generating massive amounts of omics data, we have badly neglected measuring data needed for building kinetic models. While there are databases such as BRENDA (Chang et al., 2009) that contain some kinetic parameters such as catalytic turnover rates and substrate-protein affinity coefficients, there is wide variability between measured values even for the same organisms. Sometimes, the only available data is from an organism that might be in a different phyla or even biological kingdom.

Another problem that is a major issue with all system-level biological modeling efforts is inaccurate assignment of function to gene products. It has been shown that different annotation tools can assign widely different functions for the same proteins, particularly for proteins of non-model organism (Griesemer et al., 2018). WCMs' ability to reconcile kinetic parameters is another significant means in our toolbox for overcoming the errors prevalent in the data we use for model parameterization. Given that WCMs integrate large heterogenous sets of data, they can be used to examine the incorporated data and through cross-validation improve the accuracy of model parameters. These types of data cross-validation and correction have already been shown to be a strength of WCMs (Sanghvi et al., 2013; Macklin et al., 2020).

Finally, we have been mostly overlooking the activities of "underground" metabolic processes in our models. Underground metabolic processes are biochemical reactions that occur due to promiscuity of enzymes. In our biological network reconstructions, we usually only include the canonical function for a protein and associated reactions if the proteins are enzymes. We typically ignore low flux reactions that occur when proteins interact with alternate metabolites. While the activity of underground metabolism under most conditions is very low, under extraordinary conditions their reaction rates can significantly increase and lead to evolution of new pathways and adaptation to new environments (Notebaart et al., 2018). Omission of underground metabolic processes from WCMs could affect the accuracy of model predictions, particularly when examining the behavior of a system under stress.

A promising solution to the problem of poor quality or missing parameters can be use of sophisticated machine learning techniques. Using big biological datasets with state-of-the-art methods like deep learning approach for symbolic regression (Petersen et al., 2019), where interpretable models can be generated by inferencing the optimal format of equations and parameters from given data, could predict some of these values.

4.2 Data and model integration

Combining heterogenous data together is a labor-intensive process, though advances are being made that make it easier to use disparate data and assemble it into a large model. The biomodels database (Juty et al., 2015; Malik-Sheriff et al., 2020) is one such database that captures reaction and metabolic pathways for many different cellular models. The model physiome project (Hunter et al., 2006) offers another. An ideal way of accelerating the process of WCM development is to import extant models and use them as submodels in WCMs. Chelliah et al. (2015) and Pan et al. (2021) have offered means to automatically and programmatically link

disparate submodels together into one cohesive whole. Bouhaddou et al. (2018) make the case that it is important to distribute the tools and thus conditions needed for a study can be "unit-tested" like software subroutines. In this way each individual model can be checked for errors and results can be reproduced in isolation before assembled into a larger whole. Other groups agree about the need for greater reproducibility for computational models (Papin et al., 2020; Niarakis et al., 2022). Developments of tools like Memote (Lieven et al., 2020) for standardizing the GEMs and FROG ensemble of analyses for ensuring reproducibility of published models (Tatka et al., 2023) have significantly increased confidence in the quality of models that will be incorporated in future WCMs.

Though advances are being made in automatically assembling disparate data together, researchers must take care to make sure each data source is appropriate for the task at hand. This requires an extensive literature search with proper data provenance to ensure each pathway and parameter is appropriately sourced and justified.

Once this data is assembled, deciding how best to simulate the model is no small task. From a software engineering standpoint, reference code implementations from different research teams are usually completely incompatible with each other. This requires recoding and translating, which is why having reproducible results are so important. Model definition languages like SBML (Hucka et al., 2018), CellML (Lloyd et al., 2004), and Modelica (Fritzson and Engelson, 1998) offer an advantage here because they separate the model definition from its numerical implementation, which simplifies composing different cellular models from different sources.

From a mathematical/numerical analysis standpoint, it can be difficult to decide how to integrate the different models into one cohesive whole that can offer numerically sound predictions. How the hybrid modeling process deals with the different time scales for the various types of mathematical models is a major challenge. For example, FBA models do not follow a time-varying process at all—they assume that the system operates at steady state and instantaneously adjusts to changes in order to optimize some biological objective. Ordinary differential equations (ODEs) and stochastic differential equations (SDEs) give continuous approximations of the evolution of high-concentration chemical concentrations within a component. There are well-established best practices on how to simulate ODEs/SDEs accurately, but best practices like simulating all the equations together with a global adaptive timestep fall at odds with WCM's practical need to modularize and separate different subcomponents from each other. For low-concentration chemical pathways, simulation methods like discrete chemical kinetics are preferred (Gillespie et al., 2007; Gillespie et al., 2013). Putting these disparate mathematical models together is hard, and care must be taken to ensure that artificial numerical artifacts are not introduced in the process. Here are some examples of difficulties that can arise when combining multiple different mathematical models.

- Each numerical method has different time stepping requirements. It is unclear how one determines which method controls the global timestep.
- The frequency of synchronization between different numerical mathematical models is unknown.
- In cases when ODE method is extremely stiff and requires miniscule timesteps the simulation can grind to a halt.
- The method for synchronizing continuous models like ODE/SDE with discrete chemical kinetics is unknown.
- When the concentration of a molecule gets too low in an ODE model there is a need to switch to discrete chemical kinetics. Current hybrid modeling method cannot handle this switch.
- At times it will be necessary for models to evolve independently from each other while at other times they need to be tightly coupled and must be solved together. This requires an evolving architecture of links between submodels and system variables which currently is unavailable.

None of these problems have simple solutions. It is up to the individual research teams to find the modeling format that provides the most accurate predictions and useable models. However, this level of variance could drastically lower the reusability of the models for other studies.

Aside from physical and mathematical scaling problems, from a computational viewpoint, solving the different types of models can be quite intense. FBA simulators require linear programming solvers, which have $O(n^3)$ computational requirements (i.e., every time the size of the model doubles, you need eight times the computational resources). As models get larger, it is unclear how one can spread this work across many processors to speed up the simulation. ODE/SDE solvers are usually extremely efficient, but whole-cell modeling is an inherently multi-physics and multiscale problem, with stiff processes that evolve/oscillate on a microscale timescale interacting with processes that evolve on a timescale of days. How do you synchronize these disparate timescales efficiently, and how do you separate the workflow onto multiple processors without incurring too much communication overhead? Discrete chemical kinetics require timing and tracking every chemical reaction in a cell. As concentration increases, your timestep becomes prohibitively small. How do you keep these systems from dominating the computational running time as they interact with high-concentration ODE models? How do you split these discrete chemical reactions onto multiple processors to help distribute the computational load?

4.3 Slow simulators

Although development of Vivarium (Agmon et al., 2022) has helped with some of the issues that plague simulation speed of complex whole-cell models, it is still limited to running on a single CPU with multiple cores although in principle it can extend to support distributed memory systems. Nevertheless, load balancing remains challenging while limiting the speedup.

While it might be possible to answer some of the problems associated with simulation of complex systems by building accurate reduced models (e.g., (Gates et al., 2021; Avanzini et al., 2023)), alternative solutions have been proposed. Goldberg et al. (2016) envision highly parallel whole-cell simulations by clustering species and reactions into groups that interact infrequently with each other and by simulating them in the parallel discrete event simulation (PDES) paradigm

(Jefferson et al., 1987). PDES enables further parallelism otherwise difficult to leverage via speculative execution and rollback management (Jefferson et al., 1987). This requires elaborate implementation and is currently under development.

Other potential remedies include parallelization of individual sub-models, especially the computationally demanding ones. Among the modeling approaches used in whole-cell models, stochastic simulation algorithm (SSA) (Gillespie, 1976; Gillespie, 1977) implements the most detailed model of discrete biochemical reaction events. SSA is necessary for accurately simulating statistically correct trajectories of species especially with low constituent counts. As more and more kinetic data become available for developing more accurate models, SSA can be used to simulate larger reaction networks. However, its computational cost is prohibitive for the scale of whole-cell models, even for the smallest organisms.

A popular approach to speed up an SSA simulations is to simultaneously execute multiple independent realizations of a simulation (Klingbeil et al., 2011; Sanft et al., 2011). Unfortunately, this approach is not directly beneficial to whole-cell modeling as it couples SSA-based models with other types of models for a simulation run.

However, there exist a variety of SSA methods (Gillespie, 1977). Especially, the next reaction method (NRM) (Gibson and Bruck, 2000) exposes opportunities for parallel processing. It employs a dependency graph to identify the coupling between reactions via their commonly referenced species (biomolecules in WCMs), and to selectively update the propensity and the time of the next occurrence of each reaction impacted by the fired one (Gibson and Bruck, 2000). Such updates can be processed independently of each other (Yeom et al., 2021). The degree of parallelism here is bounded by the number of system updates, i.e., the number of reactions involving the species consumed or produced by the reaction fired as well as the cost reduction in updating the priority queue. Some species may be shared by many reactions. This will result in a non-trivial number of updates, exposing the performance optimization opportunity. Goldberg et al. theorizes a PDES-based approach to parallelize SSA for distributed memory systems (Goldberg et al., 2020).

The cost of a single update itself may not be significant and dedicating a processor to that may not be beneficial. Therefore, an existing approach partitions the reaction network into multiple subnetworks and updates them simultaneously with one processor per group of reactions of each subnetwork via OpenMP (Yeom et al., 2021). Partitioning a network of highly skewed degree distribution for load balancing is known to be challenging (Gonzalez et al., 2012; Yeom et al., 2014). In the bipartite-graph abstraction of biochemical networks, a reaction node represents a computation, and a species node does a state. The edge indicates the dependency of the computation on the states. If a state is referenced by different reactions across multiple subnetworks over distributed memory systems, state replication, maintained by a means of coherent updates, may help mitigate the message passing cost. When parallelized for shared memory systems, the state must be accessed in a coordinated fashion among different processors to maintain consistency (Yeom et al., 2021). For balancing compute loads across processors, partitioning must consider the distribution of aggregate reaction update rates of subnetworks, which dynamically evolve through the course of simulation. This presents another challenge for load balancing and may require re-partitioning.

There exist works that parallelize SSA using accelerator hardware (Indurkhya and Beal, 2010; Komarov and D'Souza, 2012; Manolakos and Kouskoumvekakis, 2017). However, these approaches assume only the mass-action type reactions (van der Schaft et al., 2013) and leverage it for parallelization. These do not support general forms of reaction rate formula to accommodate diverse modeling practices in the field, or do not support the community standard model description, such as SBML, to its full reaction expression capacity (Bornstein et al., 2008; Sayikli and Bagci, 2011; Erdem et al., 2022).

ODE is another common simulation method used in WCM, and there exist solver packages that speed up by distributed memory parallelism using MPI along with node-level acceleration using GPU or OpenMP (Fidler et al., 2019; Balos et al., 2021; Städter et al., 2021; Elrod et al., 2022).

5 Conclusion

The field of whole-cell modeling is growing. Since the publication of the first WCM a decade ago a handful of models for important research, industrial, and medicinal model systems have been developed. Other than the ones mentioned above earlier, WCMs have been developed for JCVI-syn3A (Thornburg et al., 2022) and human epithelial cells (Ghaemi et al., 2020). Given the difficult and very labor-intensive process of developing WCMs, this is a remarkable achievement and a testament to how scientists view the potential of these models. The creation of these models has led to the development of whole-cell structural models (Maritan et al., 2022; Stevens et al., 2023) and even multicellular whole community models (Skalnik et al., 2023).

There are still several problems that need to be addressed before the use of these models becomes as common as usage of genome-scale models of metabolism. These include problems with data collection, model integration and parallel simulation of hybrid models. However, advances thus far are a good indication that these obstacles will soon be overcome.

Author contributions

KG: Writing–original draft, Writing–review and editing. JY: Writing–original draft, Writing–review and editing. RCB: Writing–original draft, Writing–review and editing. AN: Funding acquisition, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the Laboratory Research and Development program

(19-ERD-030) at LLNL and partially by the LLNL μBiospheres Scientific Focus Area, funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science program under FWP SCW1039.

Acknowledgments

The authors would like to thank Drs. Arthur Goldberg, Jonathan Karr, Marc Birtwistle, and Eran Agmon for sharing their experiences in developing large multi-scale systems models and insights into challenges associated with whole-cell modeling. Work at LLNL was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-JRNL-851344.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

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 (7), e1002575. doi:10.1371/journal.pcbi.1002575

Agmon, E., Spangler, R. K., Skalnik, C. J., Poole, W., Peirce, S. M., Morrison, J. H., et al. (2022). Vivarium: an interface and engine for integrative multiscale modeling in computational biology. *Bioinformatics* 38 (7), 1972–1979. doi:10.1093/bioinformatics/btac049

Ahn-Horst, T. A., Mille, L. S., Sun, G., Morrison, J. H., and Covert, M. W. (2022). An expanded whole-cell model of *E. coli* links cellular physiology with mechanisms of growth rate control. *npj Syst. Biol. Appl.* 8 (1), 30. doi:10.1038/s41540-022-00242-9

Akiyama, T., and Kim, M. (2021). Stochastic response of bacterial cells to antibiotics: its mechanisms and implications for population and evolutionary dynamics. *Curr. Opin. Microbiol.* 63, 104–108. doi:10.1016/j.mib.2021.07.002

Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z. N., and Barabasi, A. L. (2004). Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 427 (6977), 839–843. doi:10.1038/nature02289

Almaas, E., Oltvai, Z. N., and Barabasi, A. L. (2005). The activity reaction core and plasticity of metabolic networks. *PLoS Comput. Biol.* 1 (7), e68. doi:10.1371/journal.pcbi.0010068

Avanzini, F., Freitas, N., and Esposito, M. (2023). Circuit theory for chemical reaction networks. *Phys. Rev. X* 13 (2), 021041. doi:10.1103/physrevx.13.021041

Babtie, A. C., and Stumpf, M. P. H. (2017). How to deal with parameters for whole-cell modelling. *J. R. Soc. Interface* 14 (133), 20170237. doi:10.1098/rsif.2017.0237

Bachman, J. A., Gyori, B. M., and Sorger, P. K. (2023). Automated assembly of molecular mechanisms at scale from text mining and curated databases. *Mol. Syst. Biol.* 19 (5), e11325. doi:10.15252/msb.202211325

Bajcsy, P., Han, J., Liu, L., and Yang, J. (2005). Survey of biodata analysis from a data mining perspective. *Data Min. Bioinforma.* 2005, 9–39. doi:10.1007/1-84628-059-1_2

Balos, C. J., Gardner, D. J., Woodward, C. S., and Reynolds, D. R. (2021). Enabling GPU accelerated computing in the SUNDIALS time integration library. *Parallel Comput.* 108, 102836. doi:10.1016/j.parco.2021.102836

Becker, S. A., and Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* 4 (5), e1000082. doi:10.1371/journal.pcbi.1000082

Bekiaris, P. S., and Klamt, S. (2020). Automatic construction of metabolic models with enzyme constraints. *BMC Bioinforma.* 21 (1), 19. doi:10.1186/s12859-019-3329-9

Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. *Nat. Rev. Genet.* 14 (5), 333–346. doi:10.1038/nrg3433

Betts, M. J., and Russell, R. B. (2007). The hard cell: from proteomics to a whole cell model. *FEBS Lett.* 581 (15), 2870–2876. doi:10.1016/j.febslet.2007.05.062

Birch, E. W., Udell, M., and Covert, M. W. (2014). Incorporation of flexible objectives and time-linked simulation with flux balance analysis. *J. Theor. Biol.* 345, 12–21. doi:10.1016/j.jtbi.2013.12.009

Bordbar, A., McCloskey, D., Zielinski, D. C., Sonnenschein, N., Jamshidi, N., and Palsson, B. O. (2015). Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics. *Cell Syst.* 1 (4), 283–292. doi:10.1016/j.cels.2015.10.003

Bornstein, B. J., Keating, S. M., Jouraku, A., and Hucka, M. (2008). LibSBML: an API library for SBML. *Bioinformatics* 24 (6), 880–881. doi:10.1093/bioinformatics/btn051

Bouhaddou, M., Barrette, A. M., Stern, A. D., Koch, R. J., DiStefano, M. S., Riesel, E. A., et al. (2018). A mechanistic pan-cancer pathway model informed by multi-omics data interprets stochastic cell fate responses to drugs and mitogens. *PLoS Comput. Biol.* 14 (3), e1005985. doi:10.1371/journal.pcbi.1005985

Chandrasekaran, S., and Price, N. D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci.* 107 (41), 17845–17850. doi:10.1073/pnas.1005139107

Chang, A., Scheer, M., Grote, A., Schomburg, I., and Schomburg, D. (2009). BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic acids Res.* 37 (1), D588–D592. doi:10.1093/nar/gkn820

Chelliah, V., Juty, N., Ajmera, I., Ali, R., Dumousseau, M., Glont, M., et al. (2015). BioModels: ten-year anniversary. *Nucleic Acids Res.* 43 (D1), D542–D548. doi:10.1093/nar/gku1181

Choi, H., and Covert, M. W. (2023). Whole-cell modeling of *E. coli* confirms that *in vitro* tRNA aminoacylation measurements are insufficient to support cell growth and predicts a positive feedback mechanism regulating arginine biosynthesis. *Nucleic Acids Res.* 51 (12), 5911–5930. doi:10.1093/nar/gkad435

Chowdhury, A., Khodayari, A., and Maranas, C. D. (2015). Improving prediction fidelity of cellular metabolism with kinetic descriptions. *Curr. Opin. Biotechnol.* 36, 57–64. doi:10.1016/j.copbio.2015.08.011

Cohen, S. M., and Reeve, C. D. C. (2000). *Aristotle's metaphysics*.

Cornish-Bowden, A. (2013). The origins of enzyme kinetics. *FEBS Lett.* 587 (17), 2725–2730. doi:10.1016/j.febslet.2013.06.009

Descartes, R. (1984). *The philosophical writings of Descartes*. Cambridge: Cambridge University Press.

Di Filippo, M., Pescini, D., Galuzzi, B. G., Bonanomi, M., Gaglio, D., Mangano, E., et al. (2022). INTEGRATE: model-based multi-omics data integration to characterize multi-level metabolic regulation. *PLoS Comput. Biol.* 18 (2), e1009337. doi:10.1371/journal.pcbi.1009337

Elrod, C., Ma, Y., Althaus, K., and Rackauckas, C. (2022). *Parallelizing explicit and implicit extrapolation methods for ordinary differential equations* (United States: IEEE).

Erdem, C., Mutsuddy, A., Bensman, E. M., Dodd, W. B., Saint-Antoine, M. M., Bouhaddou, M., et al. (2022). A scalable, open-source implementation of a large-scale mechanistic model for single cell proliferation and death signaling. *Nat. Commun.* 13 (1), 3555. doi:10.1038/s41467-022-31138-1

Fang, X., Wallqvist, A., and Reifman, J. (2012). Modeling phenotypic metabolic adaptations of *Mycobacterium tuberculosis* H37Rv under hypoxia. *PLoS Comput. Biol.* 8 (9), e1002688. doi:10.1371/journal.pcbi.1002688

Faure, L., Mollet, B., Liebermeister, W., and Faulon, J.-L. (2023). A neural-mechanistic hybrid approach improving the predictive power of genome-scale metabolic models. *Nat. Commun.* 14 (1), 4669. doi:10.1038/s41467-023-40380-0

Fidler, M., Hallow, M., Wilkins, J., and Wang, W. (2019). RxODE: facilities for simulating from ODE-based models. *R. package version* 1 (9).

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., et al. (1995). The minimal gene complement of Mycoplasma genitalium. *Science* 270 (5235), 397–403. doi:10.1126/science.270.5235.397

Fritz, M. H.-Y., Leinonen, R., Cochrane, G., and Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21 (5), 734–740. doi:10.1101/gr.114819.110

Fritzson, P., and Engelson, V. (1998). *Modelica—a unified object-oriented language for system modeling and simulation1998* (Berlin, Germany: Springer).

Gates, A. J., Brattig Correia, R., Wang, X., and Rocha, L. M. (2021). The effective graph reveals redundancy, canalization, and control pathways in biochemical regulation and signaling. *Proc. Natl. Acad. Sci.* 118 (12), e2022598118. doi:10.1073/pnas.2022598118

Gefen, O., and Balaban, N. Q. (2009). The importance of being persistent: heterogeneity of bacterial populations under antibiotic stress. *FEMS Microbiol. Rev.* 33 (4), 704–717. doi:10.1111/j.1574-6976.2008.00156.x

Ghaemi, Z., Peterson, J. R., Gruebele, M., and Luthey-Schulten, Z. (2020). An in-silico human cell model reveals the influence of spatial organization on RNA splicing. *PLoS Comput. Biol.* 16 (3), e1007717. doi:10.1371/journal.pcbi.1007717

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 (9), 1876–1889. doi:10.1021/jp993732q

Gillespie, D. T. (1976). A General method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* 22, 403–434. doi:10.1016/0021-9991(76)90041-3

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81, 2340–2361. doi:10.1021/j100540a008

Gillespie, D. T., Hellander, A., and Petzold, L. R. (2013). Perspective: stochastic algorithms for chemical kinetics. *J. Chem. Phys.* 138 (17), 170901. doi:10.1063/1.4801941

Gillespie, D. T., Lampoudi, S., and Petzold, L. R. (2007). Effect of reactant size on discrete stochastic chemical kinetics. *J. Chem. Phys.* 126 (3), 034302. doi:10.1063/1.2424461

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274 (5287), 563–567. doi:10.1126/science.274.5287.546

Goldberg, A. P., Chew, Y. H., and Karr, J. R. (2016). *Toward scalable whole-cell modeling of human cells* (United States: ACM).

Goldberg, A. P., Jefferson, D. R., Sekar, J. A. P., and Karr, J. R. (2020). *Exact parallelization of the stochastic simulation algorithm for scalable simulation of large biochemical networks*. arXiv preprint arXiv:200505295.

Gonzalez J. E., Low Y., Gu H., Bickson D., and Guestrin C. (2012). *{PowerGraph}: distributed {Graph-Parallel} computation on natural graphs* (United States: USENIX Association).

Griesemer, M., Kimbrel, J. A., Zhou, C. E., Navid, A., and D'haeseleer, P. (2018). Combining multiple functional annotation tools increases coverage of metabolic annotation. *BMC genomics* 19 (1), 948. doi:10.1186/s12864-018-5221-9

Gunawardena, J. (2012). Silicon dreams of cells into symbols. *Nat. Biotechnol.* 30 (9), 838–840. doi:10.1038/nbt.2358

Guo, T., and Li, X. (2023). Machine learning for predicting phenotype from genotype and environment. *Curr. Opin. Biotechnol.* 79, 102853. doi:10.1016/j.copbio.2022.102853

Guzzetta, G., Jurman, G., and Furlanello, C. (2010). A machine learning pipeline for quantitative phenotype prediction from genotype data. *BMC Bioinforma.* 11 (8), S3–S9. doi:10.1186/1471-2105-11-S8-S3

Gyori, B. M., Bachman, J. A., Subramanian, K., Muhlich, J. L., Galescu, L., and Sorger, P. K. (2017). From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol.* 13 (11), 954. doi:10.15252/msb.20177651

Hadadi, N., Pandey, V., Chiappino-Pepe, A., Morales, M., Gallart-Ayala, H., Mehl, F., et al. (2020). Mechanistic insights into bacterial metabolic reprogramming from omics-integrated genome-scale models. *NPJ Syst. Biol. Appl.* 6 (1), 1. doi:10.1038/s41540-019-0121-4

Hill, R. (1970). *The chemistry of life: eight lectures on the history of biochemistry*. Cambridge: CUP Archive.

Hucka, M., Bergmann, F. T., Dräger, A., Hoops, S., Keating, S. M., Le Novère, N., et al. (2018). The Systems Biology Markup Language (SBML): language specification for level 3 version 2 core. *J. Integr. Bioinforma.* 15 (1), 20170081. doi:10.1515/jib-2017-0081

Hunter, P. J., Li, W. W., McCulloch, A. D., and Noble, D. (2006). Multiscale modeling: physiome project standards, tools, and databases. *Computer* 39 (11), 48–54. doi:10.1109/mc.2006.392

Indurkhya, S., and Beal, J. (2010). Reaction factoring and bipartite update graphs accelerate the Gillespie algorithm for large-scale biochemical systems. *PloS one* 5 (1), e8125. doi:10.1371/journal.pone.0008125

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., et al. (2014). Big data and its technical challenges. *Commun. ACM* 57 (7), 86–94. doi:10.1145/2611567

Jamei, M. (2016). Recent advances in development and application of physiologically-based pharmacokinetic (PBPK) models: a transition from academic curiosity to regulatory acceptance. *Curr. Pharmacol. Rep.* 2, 161–169. doi:10.1007/s40495-016-0059-9

Jamshidi, N., and Palsson, B. Ø. (2008). Formulating genome-scale kinetic models in the post-genome era. *Mol. Syst. Biol.* 4 (1), 171. doi:10.1038/msb.2008.8

Jefferson, D., Beckman, B., Wieland, F., Blume, L., and DiLoreto, M. (1987). *Time warp operating system* (United States: ACM).

Jensen, P. A., and Papin, J. A. (2011). Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* 27 (4), 541–547. doi:10.1093/bioinformatics/btq702

Johnson, K. A. (2013). A century of enzyme kinetic analysis, 1913 to 2013. *FEBS Lett.* 587 (17), 2753–2766. doi:10.1016/j.febslet.2013.07.012

Juty, N., Ali, R., Glont, M., Keating, S., Rodriguez, N., Swat, M. J., et al. (2015). BioModels: content, features, functionality, and use. *CPT pharmacometrics Syst. Pharmacol.* 4 (2), e3–e68. doi:10.1002/psp4.3

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic acids Res.* 32 (1), D277–D280. doi:10.1093/nar/gkh063

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids Res.* 44 (D1), D457–D462. doi:10.1093/nar/gkv1070

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., et al. (2017). The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinforma.* 20, 1085–1093. doi:10.1093/bib/bbx085

Karp, P. D., Paley, S., Caspi, R., Kothari, A., Krummenacker, M., Midford, P. E., et al. (2023). The EcoCyc database. *EcoSal Plus* 2023, eesp0002. eesp-0002. doi:10.1128/ecosalplus.esp-0002-2023

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Arora, A., and Covert, M. W. (2013). WholeCellKB: model organism databases for comprehensive whole-cell models. *Nucleic Acids Res.* 41, D787–D792. doi:10.1093/nar/gks1108

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., et al. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150 (2), 389–401. doi:10.1016/j.cell.2012.05.044

Karr, J. R., Takahashi, K., and Funahashi, A. (2015). The principles of whole-cell modeling. *Curr. Opin. Microbiol.* 27, 18–24. doi:10.1016/j.mib.2015.06.004

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muniz-Rascado, L., et al. (2011). EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.* 39, D583–D590. doi:10.1093/nar/gkq1143

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., et al. (2017). The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* 45 (1), D543–D50. doi:10.1093/nar/gkw1003

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 (1), 13806. doi:10.1038/ncomms13806

Kim, M., Rai, N., Zorraquino, V., and Tagkopoulos, I. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nat. Commun.* 7 (1), 13090. doi:10.1038/ncomms13090

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015b). Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS Comput. Biol.* 11 (8), e1004321. doi:10.1371/journal.pcbi.1004321

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., et al. (2015a). BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids Res.* 44 (D1), D515–D522. doi:10.1093/nar/gkv1049

Klingbeil, G., Erban, R., Giles, M., and Maini, P. K. (2011). STOCHSIMGPU: parallel stochastic simulation for the Systems Biology Toolbox 2 for MATLAB. *Bioinformatics* 27 (8), 1170–1171. doi:10.1093/bioinformatics/btr068

Klipp, E. (2007). Modelling dynamic processes in yeast. *Yeast* 24 (11), 943–959. doi:10.1002/yea.1544

Komarov, I., and D'Souza, R. M. (2012). Accelerating the Gillespie exact stochastic simulation algorithm using hybrid parallel execution on graphics processing units. *PLoS One* 7 (11), e46693. doi:10.1371/journal.pone.0046693

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *science* 324 (5924), 255–258. doi:10.1126/science.1170160

Kwon, D., Kim, S., Shin, S.-Y., Chatr-aryamontri, A., and Wilbur, W. J. (2014). Assisting manual literature curation for protein–protein interactions using BioQRator. *Database* 2014, bau067. doi:10.1093/database/bau067

Lewis, J. E., and Kemp, M. L. (2021). Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. *Nat. Commun.* 12 (1), 2700. doi:10.1038/s41467-021-22989-1

Liebermeister, W., and Klipp, E. (2006a). Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Model.* 3, 41–13. doi:10.1186/1742-4682-3-41

Liebermeister, W., and Klipp, E. (2006b). Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theor. Biol. Med. Model.* 3 (1), 42–11. doi:10.1186/1742-4682-3-42

Lieven, C., Beber, M. E., Olivier, B. G., Bergmann, F. T., Ataman, M., Babaei, P., et al. (2020). MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.* 38 (3), 272–276. doi:10.1038/s41587-020-0446-y

Lloyd, C. M., Halstead, M. D. B., and Nielsen, P. F. (2004). CellML: its future, present and past. *Prog. biophysics Mol. Biol.* 85 (2), 433–450. doi:10.1016/j.pbiomolbio.2004.01.004

Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29 (14), 1830–1831. doi:10.1093/bioinformatics/btt285

Luo, W., Pant, G., Bhavnasi, Y. K., Blanchard, S. G., Jr, and Brouwer, C. (2017). Pathview Web: user friendly pathway visualization and data integration. *Nucleic acids Res.* 45 (W1), W501–W8. doi:10.1093/nar/gkx372

Ma, D., Yang, L., Fleming, R. M. T., Thiele, I., Palsson, B. O., and Saunders, M. A. (2017). Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. *Sci. Rep.* 7 (1), 40863. doi:10.1038/srep40863

Macklin, D. N., Ahn-Horst, T. A., Choi, H., Ruggero, N. A., Carrera, J., Mason, J. C., et al. (2020). Simultaneous cross-evaluation of heterogeneous *E. coli* datasets via mechanistic simulation. *Science* 369 (6502), eaav3751. doi:10.1126/science.aav3751

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., et al. (2020). BioModels—15 years of sharing computational models in life science. *Nucleic Acids Res.* 48 (D1), D407–D15. doi:10.1093/nar/gkz1055

Manolakos, E. S., and Kouskoumvekakis, E. (2017). *StochSoCs: high performance biocomputing simulations for large scale Systems Biology* (United States: IEEE).

Maritan, M., Autin, L., Karr, J., Covert, M. W., Olson, A. J., and Goodsell, D. S. (2022). Building structural models of a whole Mycoplasma cell. *J. Mol. Biol.* 434 (2), 167351. doi:10.1016/j.jmb.2021.167351

Marx, V. (2013). Biology: the big challenges of big data. *Nature* 498 (7453), 255–260. doi:10.1038/498255a

Navid, A., and Almaas, E. (2012). Genome-level transcription data of *Yersinia pestis* analyzed with a New metabolic constraint-based approach. *BMC Syst. Biol.* 6 (1), 150. doi:10.1186/1752-0509-6-150

Niarakis, A., Waltemath, D., Glazier, J., Schreiber, F., Keating, S. M., Nickerson, D., et al. (2022). Addressing barriers in comprehensiveness, accessibility, reusability, interoperability and reproducibility of computational models in systems biology. *Briefings Bioinforma.* 23 (4), bbac212. doi:10.1093/bib/bbac212

Notebaart, R. A., Kintses, B., Feist, A. M., and Papp, B. (2018). Underground metabolism: network-level perspective and biotechnological potential. *Curr. Opin. Biotechnol.* 49, 108–114. doi:10.1016/j.copbio.2017.07.015

Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28 (3), 245–248. doi:10.1038/nbt.1614

Österlund, T., Nookaew, I., Bordel, S., and Nielsen, J. (2013). Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. *BMC Syst. Biol.* 7 (1), 36. doi:10.1186/1752-0509-7-36

Pan, M., Gawthrop, P. J., Cursons, J., and Crampin, E. J. (2021). Modular assembly of dynamic models in systems biology. *PLoS Comput. Biol.* 17 (10), e1009513. doi:10.1371/journal.pcbi.1009513

Papin, J. A., Mac Gabhann, F., Sauro, H. M., Nickerson, D., and Rampadarath, A. (2020). *Improving reproducibility in computational biology research*. San Francisco, CA USA: Public Library of Science, e1007881.

Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A., and Palsson, B. O. (2003). Metabolic pathways in the post-genome era. *Trends Biochem. Sci.* 28 (5), 250–258. doi:10.1016/S0968-0004(03)00064-1

Petersen, B. K., Landajuela, M., Mundhenk, T. N., Santiago, C. P., Kim, S. K., and Kim, J. T. (2019). *Deep symbolic regression: recovering mathematical expressions from data via risk-seeking policy gradients*. arXiv preprint arXiv:191204871. 2019.

Pozo, C., Miró, A., Guillén-Gosálbez, G., Sorribas, A., Alves, R., and Jiménez, L. (2015). Gobal optimization of hybrid kinetic/FBA models via outer-approximation. *Comput. Chem. Eng.* 72, 325–333. doi:10.1016/j.compchemeng.2014.06.011

Purcell, O., Jain, B., Karr, J. R., Covert, M. W., and Lu, T. K. (2013). Towards a whole-cell modeling approach for synthetic biology. *Chaos* 23 (2), 025112. doi:10.1063/1.4811182

Rees-Garbutt, J., Chalkley, O., Landon, S., Purcell, O., Marucci, L., and Grierson, C. (2020). Designing minimal genomes using whole-cell models. *Nat. Commun.* 11 (1), 836. doi:10.1038/s41467-020-14545-0

Roberts, E. (2014). Cellular and molecular structure as a unifying framework for whole-cell modeling. *Curr. Opin. Struct. Biol.* 25, 86–91. doi:10.1016/j.sbi.2014.01.005

Rowe, E., Palsson, B. O., and King, Z. A. (2018). Escher-FBA: a web application for interactive flux balance analysis. *BMC Syst. Biol.* 12, 84–87. doi:10.1186/s12918-018-0607-5

Sahu, A., Blätke, M.-A., Szymański, J. J., and Töpfer, N. (2021). Advances in flux balance analysis by integrating machine learning and mechanism-based models. *Comput. Struct. Biotechnol. J.* 19, 4626–4640. doi:10.1016/j.csbj.2021.08.004

Sánchez, B. J., Zhang, C., Nilsson, A., Lahtvee, P. J., Kerkhoven, E. J., and Nielsen, J. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* 13 (8), 935. doi:10.15252/msb.20167411

Sanft, K. R., Wu, S., Roh, M., Fu, J., Lim, R. K., and Petzold, L. R. (2011). StochKit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics* 27 (17), 2457–2458. doi:10.1093/bioinformatics/btr401

Sanghvi, J. C., Regot, S., Carrasco, S., Karr, J. R., Gutschow, M. V., Bolival, B., et al. (2013). Accelerated discovery via a whole-cell model. *Nat. Methods* 10 (12), 1192–1195. doi:10.1038/nmeth.2724

Sayikli, C., and Bagci, E. Z. (2011). *Limitations of using mass action kinetics method in modeling biochemical systems: illustration for a second order reaction* (Berlin, Germany: Springer).

Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. O. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinforma.* 11, 213. doi:10.1186/1471-2105-11-213

Schilling, C. H., Schuster, S., Palsson, B. O., and Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.* 15 (3), 296–303. doi:10.1021/bp990048k

Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F., and Schomburg, D. (2002). BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* 27 (1), 54–56. doi:10.1016/s0968-0004(01)02027-8

Shameer, S., Wang, Y., Bota, P., Ratcliffe, R. G., Long, S. P., and Sweetlove, L. J. (2022). A hybrid kinetic and constraint-based model of leaf metabolism allows predictions of metabolic fluxes in different environments. *Plant J.* 109 (1), 295–313. doi:10.1111/tpj.15551

Shamim, A., Shaikh, M. U., and Malik, S. U. R. (2010). "Intelligent data mining in autonomous heterogeneous distributed bio databases," in 2010 Second International Conference on Computer Engineering and Applications, Bali, Indonesia, 2010 19-21 March.

Skalnik, C. J., Cheah, S. Y., Yang, M. Y., Wolff, M. B., Spangler, R. K., Talman, L., et al. (2023). Whole-cell modeling of *E. coli* colonies enables quantification of single-cell heterogeneity in antibiotic responses. *PLOS Comput. Biol.* 19 (6), e1011232. doi:10.1371/journal.pcbi.1011232

Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., et al. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinforma.* 21 (1), 119–218. doi:10.1186/s12859-020-3427-8

Städter, P., Schälte, Y., Schmiester, L., Hasenauer, J., and Stapor, P. L. (2021). Benchmarking of numerical integration methods for ODE models of biological systems. *Sci. Rep.* 11 (1), 2696. doi:10.1038/s41598-021-82196-2

Stanford, N. J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P., and Liebermeister, W. (2013). Systematic construction of kinetic models from genome-scale metabolic networks. *PloS one* 8 (11), e79195. doi:10.1371/journal.pone.0079195

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLoS Biol.* 13 (7), e1002195. doi:10.1371/journal.pbio.1002195

Stevens, J. A., Grünewald, F., van Tilburg, P. A. M., König, M., Gilbert, B. R., Brier, T. A., et al. (2023). Molecular dynamics simulation of an entire cell. *Front. Chem.* 11, 1106495. doi:10.3389/fchem.2023.1106495

Sun, G., Ahn-Horst, T. A., and Covert, M. W. (2021). The *E. coli* whole-cell modeling project. *EcoSal plus* 9 (2), eESP00012020. eESP-0001. doi:10.1128/ecosalplus.ESP-0001-2020

Tatka, L. T., Smith, L. P., Hellerstein, J. L., and Sauro, H. M. (2023). Adapting modeling and simulation credibility standards to computational systems biology. *J. Transl. Med.* 21 (1), 501. doi:10.1186/s12967-023-04290-5

Thiele, I., Fleming, R. M. T., Que, R., Bordbar, A., Diep, D., and Palsson, B. O. (2012). Multiscale modeling of metabolism and macromolecular synthesis in *E. coli* and its application to the evolution of codon usage. *PLoS One* 7, e45635. doi:10.1371/journal.pone.0045635

Thornburg, Z. R., Bianchi, D. M., Brier, T. A., Gilbert, B. R., Earnest, T. M., Melo, M. C. R., et al. (2022). Fundamental behaviors emerge from simulations of a living minimal cell. *Cell* 185 (2), 345–360.e28. doi:10.1016/j.cell.2021.12.025

Tomita, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol.* 19 (6), 205–210. doi:10.1016/s0167-7799(01)01636-5

van der Schaft, A., Rao, S., and Jayawardhana, B. (2013). On the mathematical structure of balanced chemical reaction networks governed by mass action kinetics. *SIAM J. Appl. Math.* 73 (2), 953–973. doi:10.1137/11085431x

Waltemath, D., Karr, J. R., Bergmann, F. T., Chelliah, V., Hucka, M., Krantz, M., et al. (2016). Toward community standards and software for whole-cell modeling. *IEEE Trans. Biomed. Eng.* 63 (10), 2007–2014. doi:10.1109/TBME.2016.2560762

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids Res.* 41 (W1), W518–W522. doi:10.1093/nar/gkt441

Ye, C., Xu, N., Gao, C., Liu, G., Xu, J., Zhang, W., et al. (2020). Comprehensive understanding of *Saccharomyces cerevisiae* phenotypes with whole-cell model WM_S288C. *Biotechnol. Bioeng.* 117 (5), 1562–1574. doi:10.1002/bit.27298

Yeom J., Bhatele, A., Bisset, K., Bohm, E., Gupta, A., and Kale, L. V. (2014). *Overcoming the scalability challenges of epidemic simulations on blue waters* (United States: IEEE).

Yeom, J., Georgouli, K., Blake, R., and Navid, A. (2021). Towards dynamic simulation of a whole cell model. Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics.

Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput. Biol.* 15 (7), e1007084. doi:10.1371/journal.pcbi.1007084

Zur, H., Ruppin, E., and Shlomi, T. (2010). iMAT: an integrative metabolic analysis tool. *Bioinformatics* 26 (24), 3140–3142. doi:10.1093/bioinformatics/btq602