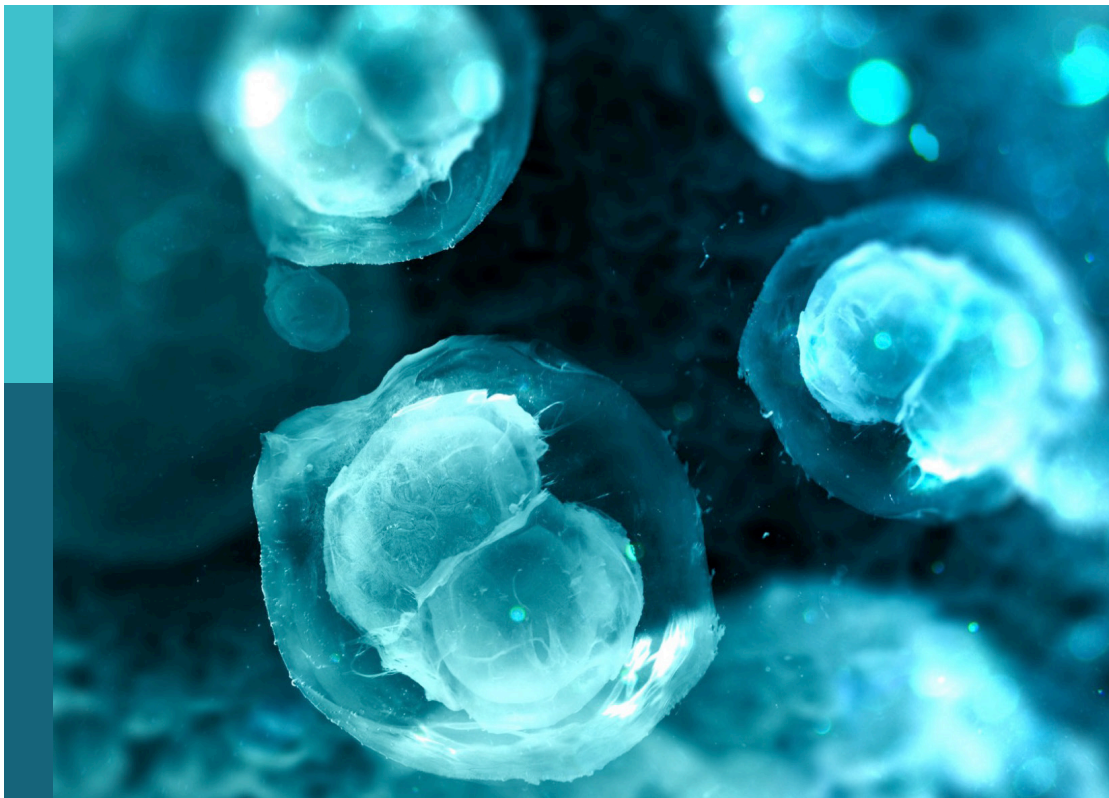
frontiers | Research Topics

Network-based mathematical modeling in cell and developmental biology

Edited by
Susan Mertins and Michael Blinov

Published in
Frontiers in Cell and Developmental Biology



ISSN 1664-8714
ISBN 978-2-8325-5346-6
DOI 10.3389/978-2-8325-5346-6

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Network-based mathematical modeling in cell and developmental biology

Topic editors

Susan Mertins — Leidos Biomedical Research, Inc., United States

Michael Blinov — UCONN Health, United States

Citation

Mertins, S., Blinov, M., eds. (2024). *Network-based mathematical modeling in cell and developmental biology*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-5346-6

Table of contents

frontiers | Frontiers in Cell and Developmental Biology

TYPE Editorial
PUBLISHED 12 August 2024
DOI 10.3389/fcell.2024.1475005

OPEN ACCESS

EDITED AND REVIEWED BY
Andrew B. Goryachev,
University of Edinburgh, United Kingdom

*CORRESPONDENCE
Michael L. Blinov,
blinov@uchc.edu
Susan D. Mertins,
smertins@biosystemsstrategies.com

RECEIVED 02 August 2024
ACCEPTED 05 August 2024
PUBLISHED 12 August 2024

CITATION
Blinov ML and Mertins SD (2024) Editorial: Network-based mathematical modeling in cell and developmental biology.
Front. Cell Dev. Biol. 12:1475005.
doi: 10.3389/fcell.2024.1475005

Editorial: Network-based mathematical modeling in cell and developmental biology

Michael L. Blinov[1]* and Susan D. Mertins[2]*

[1]Center for Cell Analysis and Modeling, University of Connecticut School of Medicine, Farmington, CT, United States, [2]BioSystems Strategies, LLC, Frederick, MD, United States

KEYWORDS

cellular reaction networks, mathematical modelling, math biology, differential equations, graph theory, boolean models, whole cell modelling

Editorial on the Research Topic
Network-based mathematical modeling in cell and developmental biology

The Research Topic on Network-based Mathematical Modeling in Cell and Developmental Biology raised various subjects of interest for mathematical modelers that also provide insight for future studies by bench scientists. In essence, three themes emerged and will be discussed below. First, the crucial and essential Research Topic of reproducibility of published models was addressed. Secondly, novel mathematical and computational approaches enabled a deeper understanding of biological systems. A third theme arose through the significant efforts aimed at creating whole cell models and downstream applications.

Enhancing reproducibility through rigor and transparency is a long-term goal of NIH (Collins and Tabak, 2014) and other agencies such as NSF. Wet-lab experiments can be difficult to reproduce because of variations in the conditions. Are models, essentially computational experiments, easily reproducible? Guided by FAIR principles (Findable, Accessible, Interoperable, and Reusable), Pedro Mendes examined a highly cited mathematical model that described segment polarity in *Drosophilia* published by Von Dassov et al. (2000). The unavailability of the original software forced the author to recode the model, which was a labor-intensive process that required *de novo* model implementation. The major take-home message from the report is that publication of mathematical models in a widely used standard format is essential, as only this will ensure the model is reproducible in the future.

Several novel mathematical approaches were taken by investigators to better understand cellular reaction networks. Marrone et al. described the use of nullclines, curves on a plane that are solutions to the differential equations, for analysis of systems with more than two variables. The authors followed Zhang et al. (2011) in considering pseudo-nullclines (an analog of nullclines for a system that can be decomposed into two modules) and used them to reproduce the dynamics of several well-known systems such as the embryonic cell cycle and MAPK cascade. Glazer et al. developed a new Monte Carlo Boolean Modeler (MC-Boomer) method to generate large (hundreds of thousands) collections of Boolean models whose simulations agree with observed data. A pipeline for analyzing these models and discovering novel regulatory interactions was developed and applied to a well-known model of the *Drosophila* segment polarity network (Albert and Othmer, 2003). Analysis of the models generated by MC-Boomer can be used to identify alternate hypotheses for the gene regulatory mechanism that could be then experimentally validated. Eidi et al. used stochastic modeling to investigate

the spatial arrangement of cell types during stem cell division, governed by two Turing signaling patterns. Their model predicts the pattern of the differentiated cells and identifies the signaling patterns that influenced the formation of the cellular structure.

Biological networks are usually described as graphs with nodes representing biological entities (e.g., genes, proteins, functional complexes) and connecting edges representing influences on their behavior. Thus, graph-theoretical methods are under constant development, as illustrated by the two manuscripts in this Research Topic. Nam and Gunawardena introduced the linear framework—a graph-theoretic approach to analyzing biomolecular systems described by continuous time Markov processes, such as post-translational modification and gene regulation. The nodes represented individual molecular states and edges represented the probabilities of transitions between molecular states. This report described the application of linear frameworks before the steady state was reached. Specifically, the authors showed that the properties of the First Passage Time (FPT) were functions of the edge labels. The FPT defined a timescale of single-molecule kinetics, such as the enzyme's completion time, and the approach described by the authors can be used for the analysis of real-time single-molecule data. Lecca et al. represented a biological network as a system of springs, in which the nodes constituted the masses and the edges were springs that connected these masses. Further, they defined latent geometry through the embedding of the spring network model into the metric space (Euclidean, hyperbolic, and spherical). Geometric properties of the embedded network (such as nodes clustering according to their radial coordinates) can be used for the analysis of the original biological network. The authors analyzed the transcriptome network of chronic myeloid leukemia and identified a set of candidate driver genes for network dynamics.

The third theme that emerged under the Research Topic addressed mathematical modeling beyond biochemical signaling networks. Georgouli et al. provided a review of existing multi-scale models of whole cells, starting from genome-scale models of metabolism developed in the nineties to the first whole-cell model incorporating the activity of nearly all molecules in *M. genitalium* to the recent efforts to develop the whole-cell model of *E. coli*. Gilbert et al. continued the theme of whole cell modeling by building a computational model of chromosome replication in a synthetic minimal bacterial cell. The authors used Langevin simulations to analyze chromosome organization. The authors noted that the polymer model of the chromosome can be used to prepare molecular dynamics models of entire Syn3A cells, validating cell states predicted by the whole-cell models. Hansen et al. discussed a topic closely related to whole-cell modeling–digital twins, multi-scale computational models of tissue and organs that work as a substitute for real human organ systems and can predict physiological events from genomic and molecular data. The authors presented opportunities and challenges for building digital twins, including parameter uncertainty, the use of artificial intelligence (AI) and machine learning (ML) methods to speed up model building, simulation and validation, and assessing the quality of predictions.

In summary, the Research Topic, Network-based Mathematical Modeling in Cell and Developmental Biology, gathered many state-of-the-art studies that will guide future directions. The melding of dry and wet laboratory studies is anticipated to advance our understanding of Systems Biology in new and exciting ways.

Author contributions

MB: Writing–original draft, Writing–review and editing. SM: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. MB wishes to acknowledge NIH funding R24 GM137787 and P41 EB023912.

Acknowledgments

We wish to thank all the contributors to this Research Topic. Through advancing network-based modeling mathematically, a deeper understanding of biological complexity will be gained. The co-authors are grateful to the staff at Frontiers Cell and Developmental Biology for their editorial efforts that allowed us to complete the Research Topic. SM wishes to thank her colleagues at Frederick National Laboratory for Cancer Research and the ATOM Research Alliance for their thoughts and input.

Conflict of interest

Author SM was employed by BioSystems Strategies, LLC.

The remaining author declares 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

Albert, R., and Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J. Theor. Biol.*, 223(1), 1–18. doi:10.1016/s0022-5193(03)00035-3

Collins, F. S., and Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, 505(7485), 612–613. doi:10.1038/505612a

Von Dassow, G., Meir, E., Munro, E. M., and Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature*, 406(6792), 188–192. doi:10.1038/35018085

Zhang, T., Schmierer, B., and Novák, B. (2011). Cell cycle commitment in budding yeast emerges from the cooperation of multiple bistable switches. *Open Biol.*, 1(3), 110009. doi:10.1098/rsob.110009

frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 06 June 2023
DOI 10.3389/fcell.2023.1201673

OPEN ACCESS

EDITED BY
Susan Mertins,
Leidos Biomedical Research, Inc.,
United States

REVIEWED BY
Andreas Dräger,
University of Tübingen, Germany
David Phillip Nickerson,
The University of Auckland, New Zealand

*CORRESPONDENCE
Pedro Mendes,
pmendes@uchc.edu

RECEIVED 06 April 2023
ACCEPTED 30 May 2023
PUBLISHED 06 June 2023

CITATION
Mendes P (2023), Reproducibility and FAIR principles: the case of a segment polarity network model.
Front. Cell Dev. Biol. 11:1201673.
doi: 10.3389/fcell.2023.1201673

Reproducibility and FAIR principles: the case of a segment polarity network model

Pedro Mendes[1,2]*

[1]Center for Cell Analysis and Modeling, University of Connecticut School of Medicine, Farmington, CT, United States, [2]Department of Cell Biology, University of Connecticut School of Medicine, Farmington, CT, United States

The issue of reproducibility of computational models and the related FAIR principles (findable, accessible, interoperable, and reusable) are examined in a specific test case. I analyze a computational model of the segment polarity network in *Drosophila* embryos published in 2000. Despite the high number of citations to this publication, 23 years later the model is barely accessible, and consequently not interoperable. Following the text of the original publication allowed successfully encoding the model for the open source software COPASI. Subsequently saving the model in the SBML format allowed it to be *reused* in other open source software packages. Submission of this SBML encoding of the model to the BioModels database enables its *findability* and *accessibility*. This demonstrates how the FAIR principles can be successfully enabled by using open source software, widely adopted standards, and public repositories, facilitating reproducibility and reuse of computational cell biology models that will outlive the specific software used.

KEYWORDS

reproducibility, model reuse, computational modeling, ODE modeling, systems biology, SBML, segment polarity network

1 Introduction

Embryonic development is characterized by frequent dynamic changes in gene expression that lead to the formation of different tissues and organs. Several patterns form during development caused by the interaction of biochemical reactions and diffusion, which was first suggested by the pioneering work of Turing (1952). Since then computational models have been used to attempt to rationalize the formation of various patterns that are crucial in development. One of these is the formation of segments in the body of insects, studied intensively in the *Drosophila* embryo (Jaeger, 2009). Insects, and other arthropods, have segmented bodies with each segment being a unit bearing a pair of appendages (such as legs). The formation of these segments during embryogenesis originates from periodic patterns of gene expression that occur in various stages. First, genes maternally expressed determine the broad regions of the body (anterior, posterior and terminal), followed by the expression of "gap genes" and then by "pair-rule genes". Mutations on the gap genes delete contiguous segments, while mutations on pair-rule genes affect every other segment. These stages happen when cell boundaries (membranes) have not yet formed and thus multiple nuclei share a common cytoplasm (a syncytium). After separation of nuclei into separate cells, by formation of plasma membranes, the "segment polarity genes" are expressed at different levels in each cell forming a pattern that will ensure the persistent polarity of the segments throughout the rest of embryonic development.

TABLE 1 Publications that reproduced or re-used the von Dassow et al. (2000) SPN model.

References	Description	Approach	Software
von Dassow and Odell (2002)	re-used original SPN model	ODE	Ingeneue[a]
Albert and Othmer (2003)	Boolean network similar but not equal to original SPN	Boolean	unknown
Tegner et al. (2003)	Single-cell version of original SPN, without diffusive transitions	ODE	unknown
Ingolia (2004)	re-coded original SPN model	ODE	C program[b]
Ma et al. (2006)	re-coded original SPN model	ODE	C program[b]
Gutenkunst et al. (2007)	re-coded original SPN model	ODE	SloppyCell[c]
Daniels et al. (2008)	re-used code from Gutenkunst et al. (2007)[d]	ODE	SloppyCell[c]
Chaves et al. (2009)	simplification of SPN model ODEs[e]	algebraic	N/A
Dayarian et al. (2009)	simplification of SPN model ODEs[e]	algebraic	unknown
Kim and Fernandes (2009)	re-coded diploid version of SPN model	ODE	Mathematica[b]
Mallavarapu et al. (2009)	re-coded original SPN model	ODE	Little b[a]
Albert et al. (2011)	re-coded original SPN model	algebraic	MATLAB[b]
Zañudo et al. (2017)	re-used original SPN model	ODE	Python[b]
Rozum and Albert (2018)	re-coded single-cell version of original SPN model	algebraic	Python
Marazzi et al. (2022)	re-used SBML model from Daniels et al. (2008)[d]	ODE	COPASI

[a]Software no longer available.
[b]Code not publicly available.
[c]Software available from https://sloppycell.sourceforge.net/.
[d]SBML version available from https://sethna.lassp.cornell.edu/Sloppy/vonDassow/model.html.
[e]Used a square grid of cells.

The year 2000 is often considered to mark the beginning of the modern systems biology era. This derives from several events that happened in that year, such as the founding of the Institute for Systems Biology, the first International Conference for Systems Biology, and the publication of various articles that are now considered "classics". One of those publications, by von Dassow et al. (2000), describes a model of the *Drosophila* segment polarity network, where a gene regulatory network operates in each one of a series of neighboring cells, with their protein products also interacting across cells (hereafter, the "SPN model"). The main conclusion, from a set of computer simulations sampling the SPN model's parameter space, was that it is "remarkably" robust as many more random combinations of parameter values than expected give rise to the characteristic spatial gene expression pattern required for segmentation. The inference that the network structure, rather than a narrow set of parameter values, is determinant to the phenotype has been cited as a general property of systems by more than one thousand publications to this date. Another conclusion derived from those results is that the phenotype is therefore robust against perturbation of the parameters—and this has also frequently been assumed to be a general property of biological systems.

An important activity in computational systems biology is the deposition of models in public repositories using standard formats like SBML (Hucka et al., 2003) or CellML (Hedley et al., 2001). This allows any scientist to easily find and access those models and use them to run simulations or derive new ones using several compatible software applications. Through the last couple decades most classic models have been added to model repositories.

Surprisingly, being described in such a highly cited publication, the SPN model is not available in any of the four major systems biology model repositories: BioModels database (Le Novère et al., 2006; Malik-Sheriff et al., 2020), the Physiome model repository (Yu et al., 2011), JWS online (Olivier and Snoep, 2004), or the database of Virtual Cell published models (Moraru et al., 2008). To make matters worse, the software Ingeneue (Meir et al., 2002; Kim, 2009), used to create this model, is no longer available, not even through the Wayback Machine (Internet Archive, 1996). Web searches revealed an SBML implementation (Sethna, 2008) which encodes the mathematics of the model in a 4 × 6 grid of cells, but not the biochemical network.

Given the importance that the results obtained from the SPN model have had in systems biology it seems important that they be available in a well-supported software simulator and distributed in a standard format by one of the model repositories. I therefore set to encode this model with COPASI (Hoops et al., 2006; Bergmann et al., 2017) and to make sure that it was correctly implemented, use it to reproduce the simulation results of von Dassow et al. (2000), at least partially. It has been noted that reproducing results from computational studies in general (Mesirov, 2010; Peng, 2011; Stodden et al., 2016), and also computatational systems biology (Waltemath and Wolkenhauer, 2016; Mendes, 2018; Tiwari et al., 2021), is as hard as with laboratory experiments. This has also been the case here and the obstacles encountered are described below.

Through a careful examination of the publications that cite von Dassow et al. (2000), I was able to identify 15 cases where the SPN model was reused (Table 1). Only two actually reproduced their

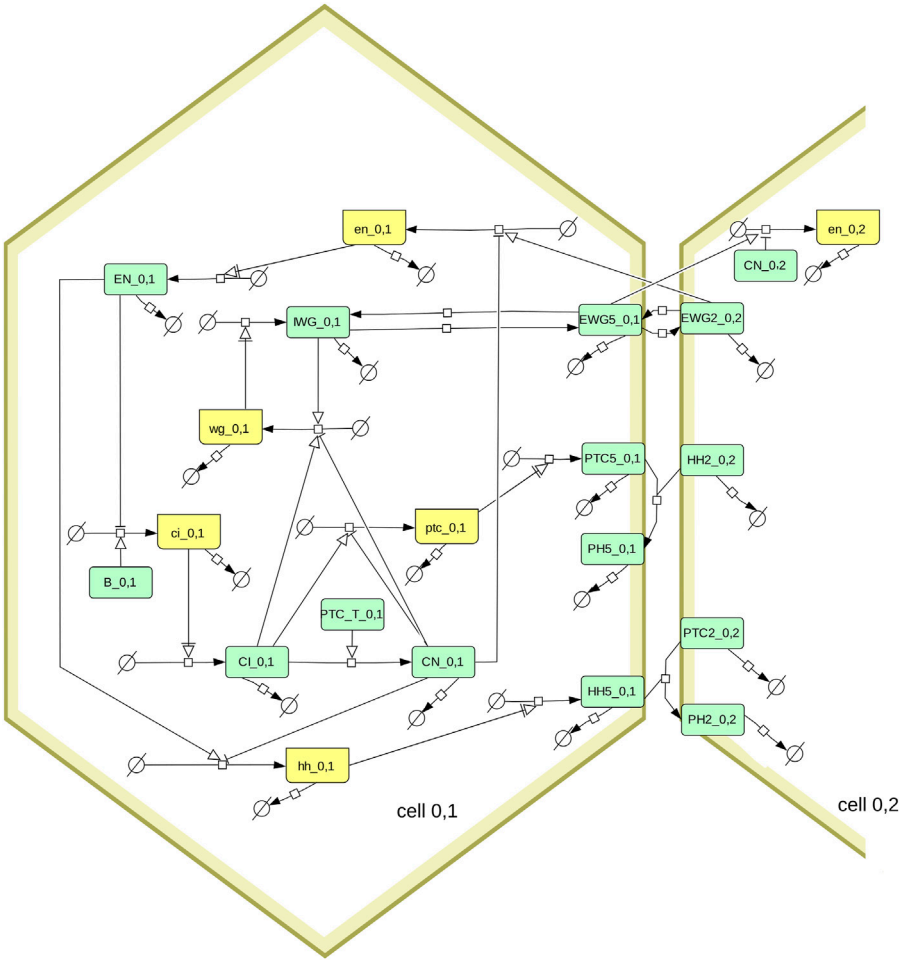


FIGURE 1
Diagram of the segment polarity network following the SBGN standard (Le Novère et al., 2009; Touré et al., 2021). Boxes in light green represent proteins, boxes in yellow represent mRNA. The full model includes several hexagonal cells, this diagram shows only one (cell_0,1) and its interactions with one of its neighbors (cell_0,2). Note that the membrane proteins (EWG, PTC, HH, and PH) exist in six pools, one for each side of the hexagonal cell. Only the proteins in side 5 are shown on the diagram, as well as the proteins on side 2 of the neighboring cell. The membrane proteins are allowed to diffuse between sides of the hexagon, which is also not shown here (*eg.* EGW5_0,1 can transfer reversibly to EGW4_0,1 and EGW6_0,1). The box labeled PTC_T_0,1 represents the sum of all PTC species (from the six sides of the membrane of cell_0,1).

results (Ingolia, 2004; Ma et al., 2006), and another expanded the analysis to diploidy (Kim and Fernandes, 2009). Several authors used the SPN model to illustrate other issues, such as robustness (Chaves et al., 2009; Dayarian et al., 2009; Albert et al., 2011), "sloppyness" (Gutenkunst et al., 2007; Daniels et al., 2008), or new methodologies (Tegner et al., 2003; Zañudo et al., 2017; Rozum and Albert, 2018; Marazzi et al., 2022). Several software applications were used, such as the original Ingeneue (Meir et al., 2002; Kim, 2009) and Little b (Mallavarapu et al., 2009), both now unavailable, and bespoke C programs that were never distributed (Ingolia, 2004; Ma et al., 2006)—all those results are now difficult to reproduce. Only the Sethna group publications (Gutenkunst et al., 2007; Daniels et al., 2008) resulted in a version of the model that is runnable in several simulators; Marazzi et al. (2022) re-used that model and also provided a COPASI version in their GitHub repository.

This exercise identifies issues that hinder reproducibility and reuse of biomodels, and illustrates how they can be overcome with modern open science practices addressing the FAIR principles (Wilkinson et al., 2016). Reproducing it required a certain level of "archeological" craft to find missing parts. I hope that this also serves as a demonstration of procedures that make models usable beyond the lifetime of the software that created them. Of course, the SPN model was an important and early application of computational systems biology to developmental biology, and reproducing its results is also not irrelevant.

2 Methods

2.1 Software

Model simulations and parameter sampling were carried out with COPASI version 4.39 (Hoops et al., 2006; Bergmann et al., 2017, RRID:SCR_014260), Virtual Cell version 7.5.0 (Schaff et al., 1997; Moraru et al., 2008, RRID:SCR_007421), Tellurium version 2.2.7 (Choi et al., 2018) that uses libRoadRunner version 2.3.2 (Welsh

et al., 2023, RRID:SCR_014763), and AMICI version 0.11.25 (Fröhlich et al., 2021), which was accessed through runBioSimulations (Shaikh et al., 2021, RRID:SCR_019110). The model file was constructed with python scripts using the BasiCO package that interfaces with COPASI (Bergmann, 2023). Simulations were run at the local high-performance computing cluster using the Cloud-COPASI web interface (Kent et al., 2012). Results were visualized with COPASI, with Gnuplot version 5.4.3 (Williams and Kelley, 2022, RRID:SCR_008619), or with the Python libraries Seaborn (Waskom, 2021) (RRID:SCR_018132) and Matplotlib (Hunter, 2007, RRID:SCR_008624). The SBGN diagram of Figure 1 was created using Cell Designer version 4.4 (Funahashi et al., 2003, RRID:SCR_007263) and then edited with Inkscape version 1.1 (RRID:SCR_014479).

2.2 Model

The model used here is the segment polarity network model described by von Dassow et al. (2000). Briefly it represents a hexagonal array of cells, where each cell can express various genes (*wingless*, *engrailed*, *hedgehog*, *cubitus interruptus*, and *patched*) and where their protein products interact within a cell, and across neighboring cells. Figure 1 depicts the interaction network using the SBGN standard (Le Novère et al., 2009; Touré et al., 2021). Note that von Dassow et al. (2000) analyze two versions of this model, one having less interactions than the other. Here we only look at their full model (i.e., including the dashed arrows in the diagram of their Box 1). Since a 1 × 4 grid of cells is enough to replicate the results (von Dassow et al., 2000), that was used here to obtain all results.

My implementation of the model was first created for the widely used software COPASI (Hoops et al., 2006; Bergmann et al., 2017) through a Python script that creates a model with arbitrary number of cells at the user's desire. A second script was created to generate the same model with only one cell, where the interacting species from neighboring cells are included as fixed concentrations. COPASI generates the full set of ODEs automatically based on the network and reaction kinetic rate laws. Unlike the SBML version from Sethna (2008), here we have the full reaction network, not just the differential equations. A small formal difference between this version and the original SPN model, is that COPASI expresses ODEs in terms of the species amounts rather than concentrations, but since the cell volumes are not variable this makes no difference and both sets of equations are equivalent.

The model makes extensive use of Hill-type functions where various terms appear in the form $base^{exponent}$. This is often problematic in IEEE floating point since, for non-integer exponents, those operations are carried out based on the equivalence:

$$base^{exponent} = e^{exponent \times \log(base)}. \tag{1}$$

Therefore, calculations fail when *base* is negative, even if infinitesimally small (generates a NaN, which in COPASI is translated to an error "Invalid state"). Unfortunately, due to the nature of predictor-corrector integration algorithms, this can easily happen during a time course integration if one species concentration becomes very close to zero. In order to avoid this problem one can use a kind of "guarded" exponentiation:

$$base^{exponent} \simeq \max(\epsilon, base)^{exponent}, \ \epsilon > 0. \tag{2}$$

Applying this protection to the model changes the rate laws. For example, the rate law for transcription with inducer-repressor pair changes from the original:

$$V \cdot \frac{I \cdot \left(1 - \frac{R^{h_2}}{k_2^{h_2} + R^{h_2}}\right)^{h_1}}{k_1^{h_1} + I \cdot \left(1 - \frac{R^{h_2}}{k_2^{h_2} + R^{h_2}}\right)^{h_1}} \tag{3}$$

to the alternative:

$$V \cdot \frac{I \cdot \max\left(\epsilon, 1 - \frac{\max(\epsilon,R)^{h_2}}{k_2^{h_2} + \max(\epsilon,R)^{h_2}}\right)^{h_1}}{k_1^{h_1} + I \cdot \max\left(\epsilon, 1 - \frac{\max(\epsilon,R)^{h_2}}{k_2^{h_2} + \max(\epsilon,R)^{h_2}}\right)^{h_1}}. \tag{4}$$

The terms $k_1^{h_1}$ and $k_2^{h_2}$ are not protected by a "guard" because k_1 and k_2 are constants that are always positive. In the results presented here I have used $\epsilon = 10^{-80}$, which reduced the incidence of simulations with NaNs from around 10%–0.1%. von Dassow et al. (2000) did not describe how they avoided this problem within the software Ingeneue. Use of these alternative rate laws was necessary for the random parameter sampling, but for specific time course simulations one can almost always use the original rate laws as described in von Dassow et al. (2000).

Several aspects of the original SPN model were not fully described by von Dassow et al. (2000) and I have had to resort to later publications to infer what they could be. For the sake of complete transparency, here are all the details that had to be inferred from sources other than the original article.

- Parameter H_{EWG} does not feature in the differential equations of the Supplementary Material S1 or in von Dassow and Odell (2002), instead there the proteins *EWG* and *IWG* have the same half-life (H_{IWG}). However the parameter is clearly described as one of the 48 parameters sampled in Meir et al. (2002), from the same group. Thus in my implementation *EWG* has its own half-life H_{EWG}.
- The identity of the 48 parameters that are sampled was not described unequivocally. There are in fact 53 parameters in the model (when considering 4 cells), so while 46 were obvious from their Supplementary Table S1, the other 2 could have been any of the remaining 7 . . . Again, a Figure in Meir et al. (2002) provided the identity of the 48 parameters (which include the one mentioned in the previous bullet).
- The ranges for parameter samplings are provided in Supplementary Table S1, however it missed including the ranges for parameters PTC_0 and HH_0. Kim (2009) mentions this range as 1–1000 (their table 3, parameters "max"), while an Ingeneue network file (named spg1_01_4cell.net), recovered from the Internet Archive (Kim, 2010), suggests it could be 10^3–10^6. I ran simulations with both ranges, and the range 1–1000 produces results closer to those reported by von Dassow et al. (2000).
- The score function used to identify parameter sets that result in the desired properties was described without sufficient

detail. This scoring function is a composite of a function to identify the gene expression pattern (Eq. 15 of their Supplementary Material S1), and another to detect stable stripes (Eq. 16 of their Supplementary Material S1); the final score being the largest of these two. The text does not specify clearly what the symbols of Eq. (16) mean, particularly the *StripeScore*. Thus I only used Eq. (15) for scoring. By definition my results should identify more parameter sets than the full scoring criterion (since we are looking for scores below a threshold of 0.2).

- The initial conditions probed in each line of Table 1 of the original paper are not specified exactly, instead they provide ranges, such as < 20% value, or 20%–60%, not saying whether the values used were random within that range or some actual specific values. I used 0.15 for when they indicate < 20%, 0.4 for when they specify 20%–60%, and 0.9 when they specify 60%–100%. For the "degraded" initial condition this is even more problematic as they only provided a bar chart without axes, rather than actual values. The values I used here are specified in the Python code and in the COPASI and SBML files for the time course described below.

As described in the Interoperability section of Results, below, the model can be exported from COPASI in standard formats, particularly the systems biology markup language (SBML, Hucka et al., 2003; Keating et al., 2020) and the OMEX format (Bergmann et al., 2014) containing a SBML file for the model and a SED-ML (Waltemath et al., 2011b) file with the simulation specification.

3 Results

3.1 Reproducibility

It is rather unfortunate that the term "reproducibility" has itself been used with various different meanings. This confusion in terminology was discussed in detail by Goodman et al. (2016), Plesser (2018), Miłkowski et al. (2018), and especially Barba (2018). As previously (Mendes, 2018), I will follow the definitions of Goodman et al. (2016), which specifies three distinct types of reproducibility.

- *reproducibility of methods* requires one to be able to exactly reproduce the results using the same methods on the same data;
- *reproducibility of results* requires one to obtain similar results in an independent study applying similar procedures;
- *reproducibility of inferences* requires the same conclusions to be reached in an independent replication potentially following a different methodology.

Because the software Ingeneue, originally used to build and simulate the SPN model, has now disappeared from circulation, reproducibility of methods can no longer be effectively carried out. In a later publication von Dassow and Odell (2002) appear to have reproduced the results with the same software (see Table 1), however since these are the original authors, that can hardly be seen as independent verification. Of all the works listed in Table 1, only Ingolia (2004) and Ma et al. (2006) can be seen as independent reproductions of the original results. Unfortunately those two publications used their own C programs but did not publish them. It was work in Sethna's lab (Gutenkunst et al., 2007; Daniels et al., 2008) that resulted in an electronic version of the model being created in the SBML format that is still available (see notes to Table 1), and which was re-used by Marazzi et al. (2022). However this SBML implementation coded the ODEs directly without representing the reaction network, an important limitation.

I attempted to reproduce the results of Table 1 in von Dassow et al. (2000), displayed in our Table 2. Overall these results match the original ones fairly well. There are some discrepancies in two samplings, but these are likely due to the uncertainty on the actual initial values, as pointed out in Methods. Bear in mind that these are very small samples of a 48-dimensional parameter space and the differences may just be due to random sampling. Figure 2 displays the succesful parameter sets in the sampling with crisp initial conditions, corresponding to Figure 2A in von Dassow et al. (2000). Careful comparison between the Figure and the original one reveals similar distributions. For example, in both cases κ_{CNen} rarely takes large values. The conclusions taken by von Dassow et al. (2000) would not change if their Figure 2A was substituted by this Figure 2. Taking these results together, I propose that the current implementation of the SPN model matches the results of the original—*reproducibility of results*.

3.2 Interoperability

To demonstrate that this implementation of the SPN model is interoperable across different software, a specific time course was chosen to be run by several simulators (herafter named *timecourse1*). One of the successful parameter sets generated in the random sampling with the "degraded" initial condition was chosen and saved as a native COPASI file, an SBML Level 3 Version 1 file (Hucka et al., 2018), and an OMEX file (Bergmann et al., 2014). Both the COPASI and OMEX files include the specification of the time course (end time of 1100 time units, sampled every 5 time units), though the SBML file requires that time course to be specified separately in the destination simulator.

Timecourse1 was simulated in four different software tools: COPASI, Virtual Cell (Schaff et al., 1997; Moraru et al., 2008), Tellurium (Choi et al., 2018), and AMICI (Fröhlich et al., 2021). It was run locally with COPASI, Virtual Cell, and Tellurium, and through the web service runBioSimulations (Shaikh et al., 2021) with AMICI. COPASI used the native file format, Tellurium used the SBML (through a small Python script runTellurium.py), while Virtual Cell and AMICI used the OMEX file.

Figures 3, 4 display the time course simulations obtained with four different software. There are no visible differences in the trajectories displayed confirming that these packages are all equally able to reproduce the results. Note that different ODE solvers were used by each one: COPASI used LSODA (Petzold, 1983), Virtual Cell used a fixed-step size Adams-Moulton method (Han and Han, 2002), Tellurium used CVODE (using the Adams-Moulton variable order, variable step size method) and AMICI used

TABLE 2 Frequency of solutions as a function of initial conditions.

	Von Dassow et al. (2000)			This work		
Initial conditions	Hits	Tries	Hit rate	Hits	Tries	Hit rate
Crisp	1,192	240,000	1/201	1,015	239,272	1/236
Degraded	149	750,000	1/5,000	22	749,988	1/34,090
Crisp, plus ubiquitous low-level *ci* and *ptc*	110	41,258	1/375	91	41,941	1/461
3-cell band of *ci*, *wg* stripe on posterior margin	69	40,338	1/585	97	41,994	1/433
3-cell band of *ptc*, *en* stripe on anterior margin	127	36,196	1/285	102	37,994	1/372
3-cell band of *ptc*, out-of-phase 3-cell band of *ci*	16	226,084	1/14,130	168	229,996	1/1,369
10.5281/zenodo.7772570 Close to target pattern	464	21,526	1/46	556	21,992	1/39

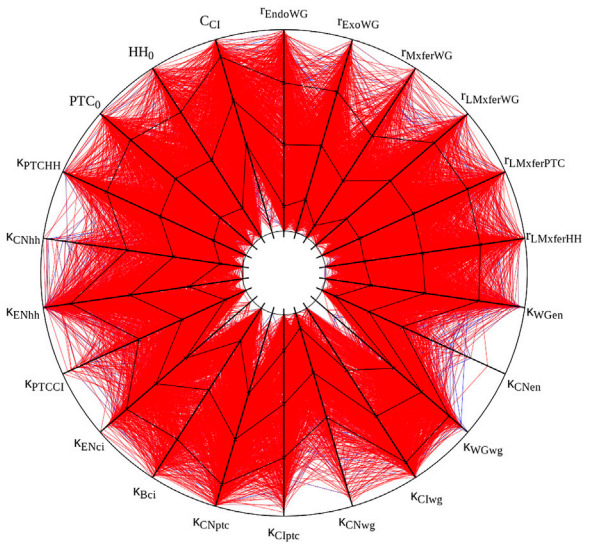


FIGURE 2
Graphic representation of "solutions" obtained with crisp initial conditions. All 1,015 parameter sets with a score below 0.2 are displayed. Black lines plot mean and standard deviation. Each spoke represents the log-scale range of one parameter. Half-lives and cooperativity coefficients are omitted, as in Panel 2A of von Dassow et al. (2000). This figure was created with the open source software Gnuplot and its source is included with the available data sets (see Data Availability Statement).

CVODES, both part of the SUNDIALS suite (Hindmarsh et al., 2005).

3.3 Findability and accessability

To promote findability and accessibility, the model files and associated scripts are made available through the following channels: a) a GitHub repository https://github.com/pmendes/models/tree/main/vonDassow2000, b) a Zenodo accession DOI (doi:10.5281/zenodo.7772570), c) a submission to the Biomodels database (MODEL2304060001), and d) model files deposited in the database of public Virtual Cell models. Note that the complete result files are only accessible through Zenodo since several files were larger than the limit at GitHub.

3.4 Reuse

To demonstrate how the model can be reused for different purposes, I decided to ask the question "how often do parameter sets of the SPN model have multiple steady states?" Earlier von Dassow and Odell (2002) and especially Ingolia (2004) proposed that the robustness of pattern formation in the SPN model is due to multi-stability of steady states. Ingolia (2004) showed this in SPN models of a single cell (where the interacting species from the neighboring cells are kept constant). Here I investigate the answer to this question in a 1 × 4 array of cells. The strategy I used is as follows.

1. Generate p random sets of parameter values;
2. For each set of parameter values generate i random sets of initial conditions and calculate their steady state by integration;
3. Determine how many sets of parameter values produced more than one steady state.

COPASI can easily to carry out such a study directly with the *Parameter scan* and *Steady state* tasks. The steady state task was applied here disabling the Newton method and therefore only using ODE integration to find the steady state reachable from the initial conditions (the steady state resolution was set to 10^{-4} and the criterion used was "distance and time"). With the parameter scan task, 5,000 random parameter sets were sampled, using the same rules as in Section 3.1 above. Then, for each parameter set, it sampled 15 random initial conditions. Since we use a model of 1 × 4 array of cells, the initial conditions are composed of 132 species concentrations that were sampled in the interval [0,1].

From the 5,000 random parameter sets generated, 3,387 had at least one steady state (the remainder are likely to contain limit cycles, but this was not investigated). Of those 3,387 parameter sets with steady states, 498 contained more than one steady state. This rate of 1/10 parameter sets displaying multistability is not entirely surprising given the study by Ingolia (2004) which highlighted the positive feedbacks contained in the SPN model. Nevertheless it is interesting to investigate if these

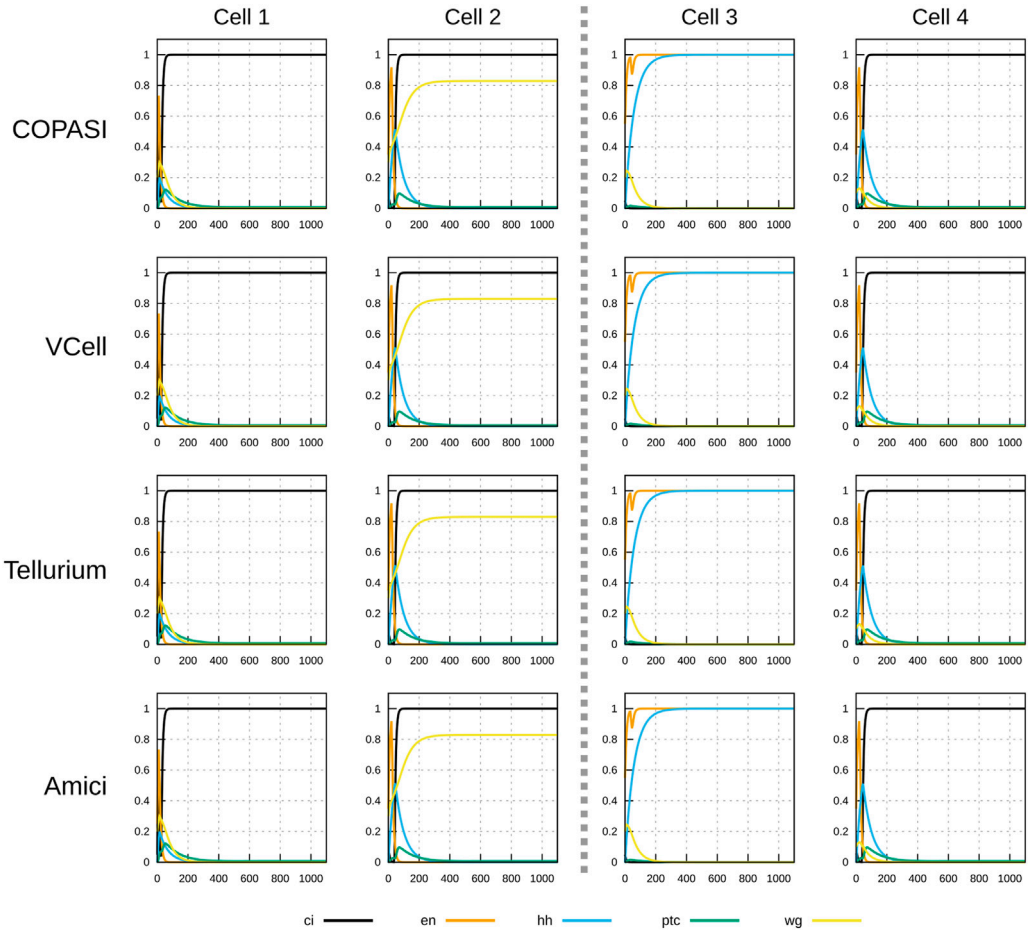


FIGURE 3
Time course simulation of mRNA species in a 1×4 arrangement of cells using a parameter set obtained by random sampling from the "degraded" initial condition (see Table 2). Columns represent the different cells; the middle dashed line separating cell 2 and cell 3 represents a parasegmental boundary. Displayed in each plot are the time evolution of all mRNA species in that cell. Note the formation of the expected segment polarity pattern around the parasegmental boundary, with high levels of *wingless* and *patched* in cell 2, and high levels of *engrailed* and *hedgehog* in cell 3. Each row corresponds to simulations carried out by different software. COPASI used the LSODA algorithm with absolute tolerance 10^{-13} and relative tolerance 10^{-8}. Virtual Cell used a fixed step size Adams-Moulton algorithm (step size 0.1). Tellurium used CVODE non-stiff algorithm (variable step size, variable order Adams-Moulton) with absolute tolerance of 10^{-12} and relative tolerance of 10^{-6}. AMICI used CVODES with absolute tolerance of 10^{-16} and relative tolerance of 10^{-8}. Results from the four simulators are visibly the same.

498 parameter sets have special characteristics *versus* the other 2,889 that have only one steady state.

The distributions of parameter values that support multiple steady states was compared with those that appear to only support a single steady state. Calculation of the relative change in the median values for each parameter in the single steady state set *versus* the multiple steady state set revealed that only κ_{CNptc} shows a large difference, with a median 5-fold larger in the multiple steady state set than in the single steady state set. Three others have much lower differences: κ_{CNen} 0.7-fold smaller, κ_{CIptc} 0.46-fold smaller, and HH_0 0.45-fold smaller. The other 44 parameters have smaller differences. Figure 5 depicts the distributions of values of κ_{CNptc} and κ_{CNen} for the two data sets. Supplementary Figures S1–S3 depict histograms for all of the 48 parameters. There seems to be very few parameter sets that lead to multiple steady states with low values of κ_{CNptc}, while many more have high values for this parameter. This suggests that in order to achieve multiple stability the repression of *patched* (*ptc*) transcription by the truncated protein product of *cubitus interruptus* (*CN*) should be weak. Note that there is another negative feedback loop between these two genes, through induction of *ptc* transcription by the full length *cubitus interruptus* protein (*CI*).

4 Discussion

It is widely recognized that there is a "reproducibility crisis" in science (Baker, 2016) that includes computational science (Mesirov, 2010; Peng, 2011; Stodden et al., 2016) and indeed computational modeling of biological systems (Waltemath and Wolkenhauer, 2016; Mendes, 2018; Tiwari et al., 2021). I and others argue that reproducibility of results obtained from computer simulations of biological models (biomodels) could be enhanced by using open source software (Ince et al., 2012; Mendes, 2018) that implement

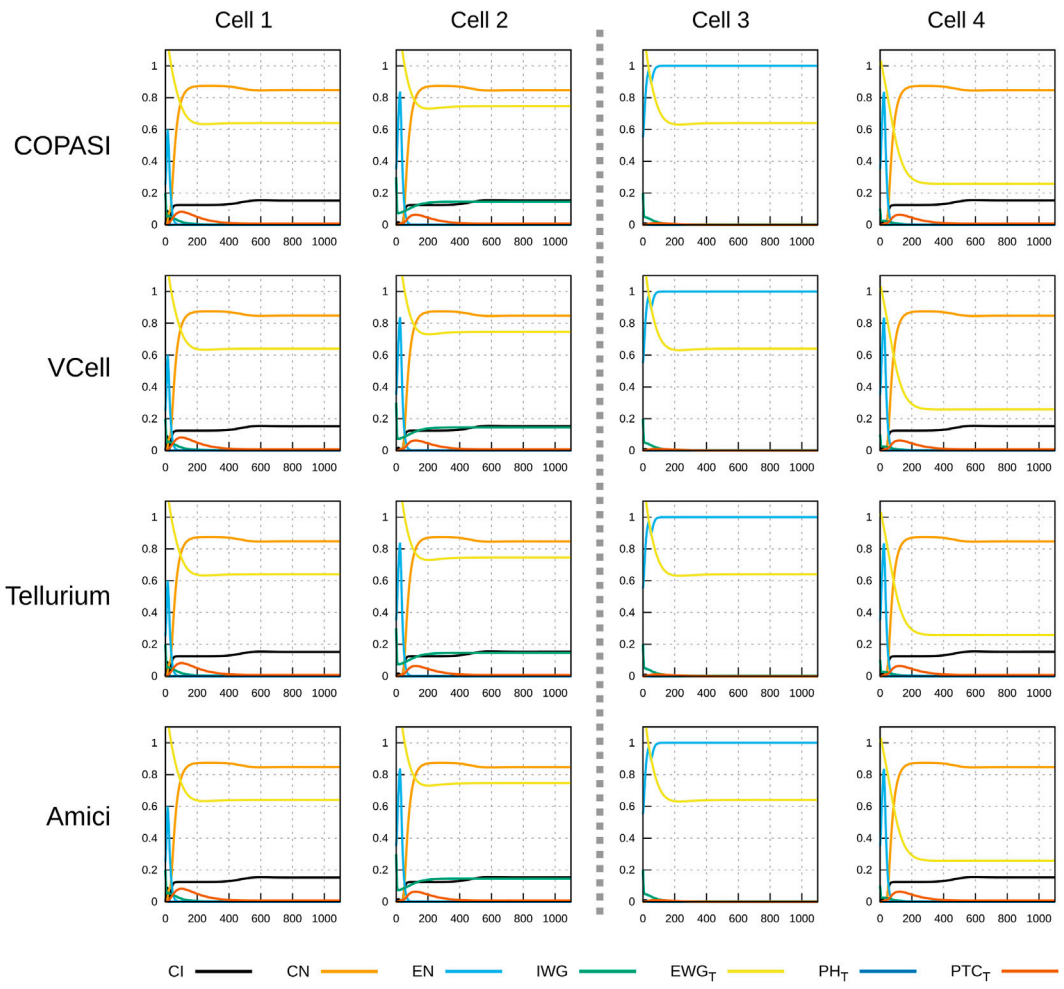


FIGURE 4
Time course simulation of protein species as in Figure 3. Displayed in each plot are the time evolution of some of the protein species in that cell. Species EWG_T represents the total amount of *EWG* protein (product of *wingless*) located in the membranes of the six neighboring cells to the one displayed; PH_T is the sum of all patched–hedgehog complexes located in the six sides of that cell's membrane, and PTC_T is the sum of all free patched receptor located in the six sides of that cell's membrane. Each row corresponds to simulations carried out by different software with different algorithms. As in Figure 3, there are no visible differences in the results of the four simulators.

widely adopted standards (Waltemath and Wolkenhauer, 2016; Blinov et al., 2021; Porubsky et al., 2021), which are part of various sets of rules proposed in the last 2 decades (Le Novère et al., 2005; Waltemath et al., 2011a; Lewis et al., 2016; Porubsky et al., 2020). Adoption of such practices, though, will only become widespread when enforced by publishers (Schnell, 2018; Stodden et al., 2018) and funding agencies (Yale Law School Roundtable Participants, 2010). A recent move by the US National Institutes of Health to enforce standards for data management (National Institutes of Health, 2020) is an encouraging move in that direction.

While reproducibility is a fundamental part of the scientific process (Popper, 1959), another important aspect is that new discoveries are almost always dependent on previous results, methodologies, and theories. To facilitate reuse of scientific data the community is increasingly adopting the so-called FAIR data principles (Wilkinson et al., 2016) which promote *Findability*, *Accessibility*, *Interoperability*, and *Reuse* of data. While biomodels are usually seen as mathematics or software, they are operationally complex data objects and these principles ought to apply to them as well. Here I reproduced the reaction network, ODE model and associated simulations described in the classic systems biology paper by von Dassow et al. (2000) with the software COPASI. I then exported the model and simulation specifications in community-derived standard formats that are supported by many software applications. Finally these files were contributed to model and data repositories. This essentially makes the model available to be manipulated by a large number of software applications, not only extant but likely future ones. Even if the standards used here will be abandoned in the future, it is most likely that converters would be developed to upgrade models to the new standards. Model and data repositories are also expected to last a long time. Thus this classic systems biology and development model is now available to a wide community, enabling its re-use for many decades.

As in previous case studies (e.g., Jablonsky et al., 2011; Tiwari et al., 2021), not all required information to reproduce the model and simulations were available in the original publication.

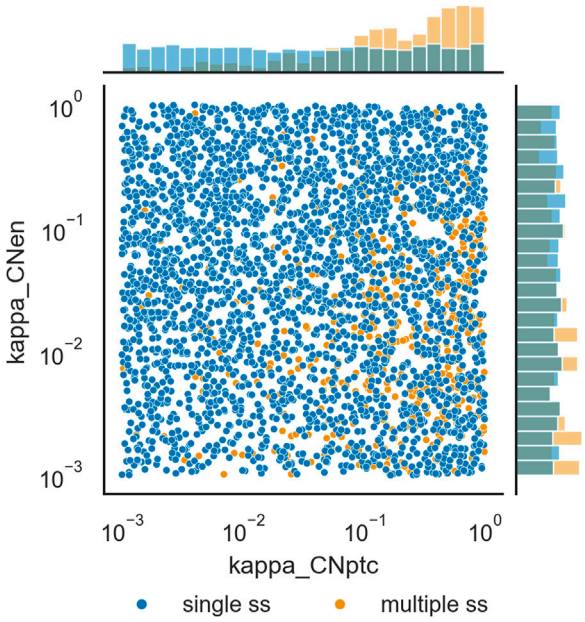


FIGURE 5
Distribution of values of κ_{CNptc} and κ_{CNen} originating single steady states, *versus* those originating multiple steady states. Scatterplot of values of the two parameters and histograms of their distribution. Darker blue circles represent parameter sets for which only one steady state was identified, lighter orange circles represent parameter sets for which more than one steady state could be identified.

Fortunately, there were subsequent publications by the authors and other members of their teams that hinted at the missing pieces. In some cases there is still uncertainty whether I made the correct choices, however the results obtained (Figure 2) are sufficiently close to the original that these choices are at least validated to be highly plausible. This supports previous suggestions (Claerbout and Karrenbach, 1992; Hothorn and Leisch, 2011; Stodden et al., 2016) that true computational reproducibility requires availability of electronic executable versions. Unfortunately textual descriptions are almost always deficient in details, as it is only too easy to miss something.

While the missing information in von Dassow et al. (2000) could be seen as a negative, I note that at the time the software Ingeneue was distributed together with files that allowed reproduction of the results. Additionally the model was actually described in great detail, so much that I was able to re-implement it. It is not uncommon to come across cases where even the model equations are not listed (see, e.g., Hübner et al., 2011, for a survey). However, this also highlights that publishing an electronic version alone is not guarantee that others in the future will be able to use it. In this case the software Ingeneue is no longer distributed and thus the electronic version is essentially lost (I could have tried to seek a copy from the original authors but I decided not to do so in order to test whether I could reproduce it with the information available). Publication of models in a widely used standard format is essential, as only this will assure the model to be interoperable by future software. Again, this is not a criticism of this 23 year-old publication, since at that time the relevant standards were nonexistent.

In conclusion: we have all the tools needed to make computational systems biology models FAIR. They should be encoded in standard formats with relevant metadata and deposited in widely used repositories. Only this will assure that future researchers will be able to study and re-use these models. Any other option, such as only describing model equations, making the model available "upon request", or non-standard electronic encodings of the model will likely be lost within a decade or less.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: GitHub: https://github.com/pmendes/models/tree/main/vonDassow2000 Zenodo: https://zenodo.org/record/7772570 Biomodels: https://www.ebi.ac.uk/biomodels/MODEL2304060001.

Author contributions

PM created the concept and design of the study, run all computations, wrote the entire manuscript. PM revised, read, and approved the submitted version.

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R24 GM137787. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgments

I am grateful to Lauren Marazzi who drew my attention to the issues of findability and accessibility of this model; to Frank Bergmann who improved the BasiCO package at my request with incredible speed; to Ion Moraru and Lucian P. Smith for help with appropriately running Virtual Cell and Tellurium, respectively. I am also grateful to Eran Agmon for many discussions about interoperability of biomodels.

Conflict of interest

The author declares 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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their

affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1201673/full#supplementary-material

References

Albert, R., and Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. *J. Theor. Biol.* 223, 1–18. doi:10.1016/s0022-5193(03)00035-3

Albert, R., DasGupta, B., Hegde, R., Sivanathan, G. S., Gitter, A., Gürsoy, G., et al. (2011). Computationally efficient measure of topological redundancy of biological and social networks. *Phys. Rev. E, Stat. Nonlinear, Soft Matter Phys.* 84, 036117. doi:10.1103/PhysRevE.84.036117

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454. doi:10.1038/533452a

Barba, L. A. (2018). *Terminologies for reproducible research. arXiv preprint*. doi:10.48550/arXiv.1802.03311

Bergmann, F. T., Adams, R., Moodie, S., Cooper, J., Glont, M., Golebiewski, M., et al. (2014). COMBINE archive and OMEX format: One file to share all information to reproduce a modeling project. *BMC Bioinforma.* 15, 369. doi:10.1186/s12859-014-0369-z

Bergmann, F. T., Hoops, S., Klahn, B., Kummer, U., Mendes, P., Pahle, J., et al. (2017). COPASI and its applications in biotechnology. *J. Biotechnol.* 261, 215–220. doi:10.1016/j.jbiotec.2017.06.1200

Bergmann, F. T. (2023). *basico: a simplified python interface to COPASI*. doi:10.5281/zenodo.7665294

Blinov, M. L., Gennari, J. H., Karr, J. R., Moraru, I. I., Nickerson, D. P., and Sauro, H. M. (2021). Practical resources for enhancing the reproducibility of mechanistic modeling in systems biology. *Curr. Opin. Syst. Biol.* 27, 100350. doi:10.1016/j.coisb.2021.06.001

Chaves, M., Sengupta, A., and Sontag, E. D. (2009). Geometry and topology of parameter space: Investigating measures of robustness in regulatory networks. *J. Math. Biol.* 59, 315–358. doi:10.1007/s00285-008-0230-y

Choi, K., Medley, J. K., König, M., Stocking, K., Smith, L., Gu, S., et al. (2018). Tellurium: An extensible python-based modeling environment for systems and synthetic biology. *Bio Syst.* 171, 74–79. doi:10.1016/j.biosystems.2018.07.006

Claerbout, J. F., and Karrenbach, M. (1992). "Electronic documents give reproducible research a new meaning," in *SEG technical program expanded abstracts 1992* (Houston, TX: Society of Exploration Geophysicists), 601–604. doi:10.1190/1.1822162

Daniels, B. C., Chen, Y.-J., Sethna, J. P., Gutenkunst, R. N., and Myers, C. R. (2008). Sloppiness, robustness, and evolvability in systems biology. *Curr. Opin. Biotechnol.* 19, 389–395. doi:10.1016/j.copbio.2008.06.008

Dayarian, A., Chaves, M., Sontag, E. D., and Sengupta, A. M. (2009). Shape, size, and robustness: Feasible regions in the parameter space of biochemical networks. *PLoS Comput. Biol.* 5, e1000256. doi:10.1371/journal.pcbi.1000256

Fröhlich, F., Weindl, D., Schälte, Y., Pathirana, D., Paszkowski, L., Lines, G. T., et al. (2021). Amici: High-performance sensitivity analysis for large ordinary differential equation models. *Bioinformatics* 37, 3676–3677. doi:10.1093/bioinformatics/btab227

Funahashi, A., Morohashi, M., Kitano, H., and Tanimura, N. (2003). CellDesigner: A process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO* 1, 159–162. doi:10.1016/S1478-5382(03)02370-9

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Sci. Transl. Med.* 8, 341ps12. doi:10.1126/scitranslmed.aaf5027

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* 3, 1871–1878. doi:10.1371/journal.pcbi.0030189

Han, T. M., and Han, Y. (2002). Solving implicit equations arising from Adams-Moulton methods. *BIT Numer. Math.* 42, 336–350. doi:10.1023/A:1021951025649

Hedley, W. J., Nelson, M. R., Bellivant, D. P., and Nielsen, P. F. (2001). A short introduction to CellML. *Philosophical Trans. R. Soc. Lond. Ser. A* 359, 1073–1089. doi:10.1098/rsta.2001.0817

Hindmarsh, A. C., Brown, P. N., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E., et al. (2005). Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Trans. Math. Softw.* 31, 363–396. doi:10.1145/1089014.1089020

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., et al. (2006). COPASI—A COmplex PAthway SImulator. *Bioinformatics* 22, 3067–3074. doi:10.1093/bioinformatics/btl485

Hothorn, T., and Leisch, F. (2011). Case studies in reproducibility. *Briefings Bioinforma.* 12, 288–300. doi:10.1093/bib/bbq084

Hübner, K., Sahle, S., and Kummer, U. (2011). Applications and trends in systems biology in biochemistry. *FEBS J.* 278, 2767–2857. doi:10.1111/j.1742-4658.2011.08217.x

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003). The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531. doi:10.1093/bioinformatics/btg015

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

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* 9, 90–95. doi:10.1109/MCSE.2007.55

Ince, D. C., Hatton, L., and Graham-Cumming, J. (2012). The case for open computer programs. *Nature* 482, 485–488. doi:10.1038/nature10836

Ingolia, N. T. (2004). Topology and robustness in the Drosophila segment polarity network. *PLoS Biol.* 2, e123. doi:10.1371/journal.pbio.0020123

Internet Archive (1996). *Wayback machine*. Available at: https://web.archive.org/.

Jablonsky, J., Bauwe, H., and Wolkenhauer, O. (2011). Modeling the calvin-benson cycle. *BMC Syst. Biol.* 5, 185. doi:10.1186/1752-0509-5-185

Jaeger, J. (2009). Modelling the *Drosophila* embryo. *Mol. Biosyst.* 5, 1549–1568. doi:10.1039/b904722k

Keating, S. M., Waltemath, D., König, M., Zhang, F., Dräger, A., Chaouiya, C., et al. (2020). SBML level 3: An extensible format for the exchange and reuse of biological models. *Mol. Syst. Biol.* 16, e9110. doi:10.15252/msb.20199110

Kent, E., Hoops, S., and Mendes, P. (2012). Condor-COPASI: High-throughput computing for biochemical networks. *BMC Syst. Biol.* 6, 91. doi:10.1186/1752-0509-6-91

Kim, K. J., and Fernandes, V. M. (2009). Effects of ploidy and recombination on evolution of robustness in a model of the segment polarity network. *PLoS Comput. Biol.* 5, e1000296. doi:10.1371/journal.pcbi.1000296

Kim, K. J. (2009). Ingeneue: A software tool to simulate and explore genetic regulatory networks. *Methods Mol. Biol.* 500, 169–200. doi:10.1007/978-1-59745-525-1_6

Kim, K. J. (2010). *IngeneueInMathematica*. Available at: https://web.archive.org/web/20100813195616/http://rusty.fhl.washington.edu/ingeneue/mathematica.html.

Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* 23, 1509–1515. doi:10.1038/nbt1156

Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., et al. (2006). BioModels database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* 34, D689–D691. doi:10.1093/nar/gkj092

Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., et al. (2009). The systems biology graphical notation. *Nat. Biotechnol.* 27, 735–741. doi:10.1038/nbt.1558

Lewis, J., Breeze, C. E., Charlesworth, J., Maclaren, O. J., and Cooper, J. (2016). Where next for the reproducibility agenda in computational biology? *BMC Syst. Biol.* 10, 52. doi:10.1186/s12918-016-0288-x

Ma, W., Lai, L., Ouyang, Q., and Tang, C. (2006). Robustness and modular design of the Drosophila segment polarity network. *Mol. Syst. Biol.* 2, 70. doi:10.1038/msb4100111

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, D407–D415. doi:10.1093/nar/gkz1055

Mallavarapu, A., Thomson, M., Ullian, B., and Gunawardena, J. (2009). Programming with models: Modularity and abstraction provide powerful capabilities for systems biology. *J. R. Soc. Interface* 6, 257–270. doi:10.1098/rsif.2008.0205

Marazzi, L., Shah, M., Balakrishnan, S., Patil, A., and Vera-Licona, P. (2022). Netisce: A network-based tool for cell fate reprogramming. *npj Syst. Biol. Appl.* 8, 21. doi:10.1038/s41540-022-00231-y

Meir, E., Munro, E. M., Odell, G. M., and Von Dassow, G. (2002). Ingeneue: A versatile tool for reconstituting genetic networks, with examples from the segment polarity network. *J. Exp. Zoology* 294, 216–251. doi:10.1002/jez.10187

Mendes, P. (2018). Reproducible research using biomodels. *Bull. Math. Biol.* 80, 3081–3087. doi:10.1007/s11538-018-0498-z

Mesirov, J. P. (2010). Computer science. Accessible reproducible research. *Science* 327, 415–416. doi:10.1126/science.1179653

Miłkowski, M., Hensel, W. M., and Hohol, M. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *J. Comput. Neurosci.* 45, 163–172. doi:10.1007/s10827-018-0702-z

Moraru, I., Morgan, F., Li, Y., Loew, L., Schaff, J., Lakshminarayana, A., et al. (2008). Virtual Cell modelling and simulation software environment. *IET Syst. Biol.* 2, 352–362. doi:10.1049/iet-syb:20080102

National Institutes of Health (2020). *NOT-OD-21-013: Final NIH policy for data management and sharing*. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.

Olivier, B. G., and Snoep, J. L. (2004). Web-based kinetic modelling using JWS Online. *Bioinformatics* 20, 2143–2144. doi:10.1093/bioinformatics/bth200

Peng, R. D. (2011). Reproducible research in computational science. *Science* 334, 1226–1227. doi:10.1126/science.1213847

Petzold, L. (1983). Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM J. Sci. Stat. Comput.* 4, 136–148. doi:10.1137/0904010

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Front. Neuroinformatics* 11, 76. doi:10.3389/fninf.2017.00076

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Porubsky, V. L., Goldberg, A. P., Rampadarath, A. K., Nickerson, D. P., Karr, J. R., and Sauro, H. M. (2020). Best practices for making reproducible biochemical models. *Cell. Syst.* 11, 109–120. doi:10.1016/j.cels.2020.06.012

Porubsky, V., Smith, L., and Sauro, H. M. (2021). Publishing reproducible dynamic kinetic models. *Briefings Bioinforma.* 22, bbaa152. doi:10.1093/bib/bbaa152

Rozum, J. C., and Albert, R. (2018). Identifying (un)controllable dynamical behavior in complex networks. *PLoS Comput. Biol.* 14, e1006630. doi:10.1371/journal.pcbi.1006630

Schaff, J., Fink, C. C., Slepchenko, B., Carson, J. H., and Loew, L. M. (1997). A general computational framework for modeling cellular structure and function. *Biophysical J.* 73, 1135–1146. doi:10.1016/S0006-3495(97)78146-3

Schnell, S. (2018). "Reproducible" research in mathematical sciences requires changes in our peer review culture and modernization of our current publication approach. *Bull. Math. Biol.* 80, 3095–3105. doi:10.1007/s11538-018-0500-9

Sethna, J. P. (2008). Segment polarity model. Available at: https://sethna.lassp.cornell.edu/Sloppy/vonDassow/model.html.

Shaikh, B., Marupilla, G., Wilson, M., Blinov, M. L., Moraru, I. I., and Karr, J. R. (2021). RunBioSimulations: An extensible web application that simulates a wide range of computational modeling frameworks, algorithms, and formats. *Nucleic Acids Res.* 49, W597–W602. doi:10.1093/nar/gkab411

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., et al. (2016). Enhancing reproducibility for computational methods. *Science* 354, 1240–1241. doi:10.1126/science.aah6168

Stodden, V., Seiler, J., and Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci. U. S. A.* 115, 2584–2589. doi:10.1073/pnas.1708290115

Tegner, J., Yeung, M. K. S., Hasty, J., and Collins, J. J. (2003). Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5944–5949. doi:10.1073/pnas.0933416100

Tiwari, K., Kananathan, S., Roberts, M. G., Meyer, J. P., Sharif Shohan, M. U., Xavier, A., et al. (2021). Reproducibility in systems biology modelling. *Mol. Syst. Biol.* 17, e9982. doi:10.15252/msb.20209982

Touré, V., Dräger, A., Luna, A., Dogrusoz, U., and Rougny, A. (2021). *The systems biology graphical notation: Current status and applications in systems medicine*. Oxford: Academic Press, 372–381. doi:10.1016/B978-0-12-801238-3.11515-6

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Trans. R. Soc. Lond. Ser. B, Biol. Sci.* 237, 37–72. doi:10.1098/rstb.1952.0012

von Dassow, G., and Odell, G. M. (2002). Design and constraints of the Drosophila segment polarity module: Robust spatial patterning emerges from intertwined cell state switches. *J. Exp. Zoology* 294, 179–215. doi:10.1002/jez.10144

von Dassow, G., Meir, E., Munro, E. M., and Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature* 406, 188–192. doi:10.1038/35018085

Waltemath, D., and Wolkenhauer, O. (2016). How modeling standards, software, and initiatives support reproducibility in systems biology and systems medicine. *IEEE Trans. Bio-Medical Eng.* 63, 1999–2006. doi:10.1109/TBME.2016.2555481

Waltemath, D., Adams, R., Beard, D. A., Bergmann, F. T., Bhalla, U. S., Britten, R., et al. (2011a). Minimum information about a simulation experiment (MIASE). *PLoS Comput. Biol.* 7, e1001122. doi:10.1371/journal.pcbi.1001122

Waltemath, D., Adams, R., Bergmann, F. T., Hucka, M., Kolpakov, F., Miller, A. K., et al. (2011b). Reproducible computational biology experiments with SED-ML—The simulation experiment description markup language. *BMC Syst. Biol.* 5, 198. doi:10.1186/1752-0509-5-198

Waskom, M. L. (2021). seaborn: statistical data visualization. *J. Open Source Softw.* 6, 3021. doi:10.21105/joss.03021

Welsh, C., Xu, J., Smith, L., König, M., Choi, K., and Sauro, H. M. (2023). libRoadRunner 2.0: a high performance SBML simulation and analysis library. *Bioinformatics* 39, btac770. doi:10.1093/bioinformatics/btac770

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

Williams, T., and Kelley, C. (2022). *Gnuplot 5.4.3: An interactive plotting program*. Available at: http://www.gnuplot.info/.

Yale Law School Roundtable Participants (2010). Reproducible research. *Comput. Sci. Eng.* 12, 8–13. doi:10.1109/MCSE.2010.113

Yu, T., Lloyd, C. M., Nickerson, D. P., Cooling, M. T., Miller, A. K., Garny, A., et al. (2011). The physiome model repository 2. *Bioinformatics* 27, 743–744. doi:10.1093/bioinformatics/btq723

Zañudo, J. G. T., Yang, G., and Albert, R. (2017). Structure-based control of complex networks with nonlinear dynamics. *Proc. Natl. Acad. U. S. A.* 114, 7234–7239. doi:10.1073/pnas.1617387114

frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 09 August 2023
DOI 10.3389/fcell.2023.1214962

OPEN ACCESS

EDITED BY
Susan Mertins,
Leidos Biomedical Research, Inc.,
United States

REVIEWED BY
Cemal Erdem,
Clemson University, United States
Lexy Von Diezmann,
University of Minnesota Twin Cities,
United States

*CORRESPONDENCE
Zaida Luthey-Schulten,
zan@illinois.edu

RECEIVED 30 April 2023
ACCEPTED 10 July 2023
PUBLISHED 09 August 2023

CITATION
Gilbert BR, Thornburg ZR, Brier TA, Stevens JA, Grünewald F, Stone JE, Marrink SJ and Luthey-Schulten Z (2023), Dynamics of chromosome organization in a minimal bacterial cell.
Front. Cell Dev. Biol. 11:1214962.
doi: 10.3389/fcell.2023.1214962

Dynamics of chromosome organization in a minimal bacterial cell

Benjamin R. Gilbert[1], Zane R. Thornburg[1], Troy A. Brier[1], Jan A. Stevens[2], Fabian Grünewald[2], John E. Stone[3,4], Siewert J. Marrink[2] and Zaida Luthey-Schulten[1,4,5]*

[1]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [2]Molecular Dynamics Group, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, Netherlands, [3]NVIDIA Corporation, Santa Clara, CA, United States, [4]NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [5]NSF Center for the Physics of Living Cells, Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, United States

Computational models of cells cannot be considered complete unless they include the most fundamental process of life, the replication and inheritance of genetic material. By creating a computational framework to model systems of replicating bacterial chromosomes as polymers at 10 bp resolution with Brownian dynamics, we investigate changes in chromosome organization during replication and extend the applicability of an existing whole-cell model (WCM) for a genetically minimal bacterium, JCVI-syn3A, to the entire cell-cycle. To achieve cell-scale chromosome structures that are realistic, we model the chromosome as a self-avoiding homopolymer with bending and torsional stiffnesses that capture the essential mechanical properties of dsDNA in Syn3A. In addition, the conformations of the circular DNA must avoid overlapping with ribosomes identitied in cryo-electron tomograms. While Syn3A lacks the complex regulatory systems known to orchestrate chromosome segregation in other bacteria, its minimized genome retains essential loop-extruding structural maintenance of chromosomes (SMC) protein complexes (SMC-scpAB) and topoisomerases. Through implementing the effects of these proteins in our simulations of replicating chromosomes, we find that they alone are sufficient for simultaneous chromosome segregation across all generations within nested theta structures. This supports previous studies suggesting loop-extrusion serves as a near-universal mechanism for chromosome organization within bacterial and eukaryotic cells. Furthermore, we analyze ribosome diffusion under the influence of the chromosome and calculate *in silico* chromosome contact maps that capture inter-daughter interactions. Finally, we present a methodology to map the polymer model of the chromosome to a Martini coarse-grained representation to prepare molecular dynamics models of entire Syn3A cells, which serves as an ultimate means of validation for cell states predicted by the WCM.

KEYWORDS

whole-cell modeling, chromosome replication, chromosome segregation, brownian dynamics, smc proteins, topoisomerase, Martini model

1 Introduction

The goal of computational modeling of a single cell is to create whole-cell models (WCMs) that propagate the state of an entire cell through time, where the propagation is governed by the chemical and physical interactions within the cell and between the cell and its environment (Karr et al., 2012; Goldberg et al., 2018; Macklin et al., 2020; Marucci et al., 2020; Luthey-Schulten et al., 2022; Maritan et al., 2022; Thornburg et al., 2022). To model any cell in 3D, configurations and dynamics of the chromosome(s) are critical in defining the spatial heterogeneity of gene expression over the course of a cell-cycle (Llopis et al., 2010). While there are several existing models that can simulate entire bacterial chromosomes (Buenemann and Lenz, 2010; Messelink et al., 2021; Wasim et al., 2021), relatively few are at spatial resolutions less than hundreds to thousands of base pairs (bp) per particle (Hacker et al., 2017; Goodsell et al., 2018; Gilbert et al., 2021). Here, we introduce a computational model to simulate the 3D dynamics of the chromosome of a genetically minimal bacterium, JCVI-syn3A, at 10-bp resolution including replicating chromosome states (Cooper and Helmstetter, 1968; Bremer and Dennis, 2008; Youngren et al., 2014) and loop-extrusion by structural maintenance of chromosomes (SMC) protein complexes (Hirano, 2006; Alipour and Marko, 2012; Ganji et al., 2018; Lioy et al., 2020; Davidson and Peters, 2021; Lee et al., 2021).

JCVI-syn3A is a minimal bacterial cell with a chemically synthesized 543 kbp genome composed of 493 genes (Breuer et al., 2019). The SynX-series of organisms began with JCVI-syn1.0, which was created by transplanting a chemically synthesized *Mycoplasma mycoides* genome into living *Mycoplasma* cells (Gibson et al., 2010). JCVI-syn3.0 was subsequently created by synthetically reducing the 1,079 kbp genome of Syn1.0 until what was considered a genetically minimal bacterium with a 531 kbp genome, stripped of all but the necessary components to continue proliferating, was achieved (Hutchison et al., 2016). Finally, Syn3A was created by re-introducing 19 genes from Syn1.0 back into Syn3.0's genome. While this produced an arguably less-minimal bacterium, it increased the growth rate (180 min doubling-time in Syn3.0 to 110 min doubling-time in Syn3A) (Breuer et al., 2019) and restored a regular spherical morphology to the cells (Pelletier et al., 2021).

With a genome and physical size approximately one-tenth the size of the model bacterium *Escherichia coli*, Syn3A is ideally suited for whole-cell modeling due to the corresponding reduction in complexity. Syn3A′s initial cell state was defined through experimental charactizations of its biochemical components — genome-wide gene-essentiality and proteomics (Breuer et al., 2019), metabolomics (Haas et al., 2022), lipidomics (Thornburg et al., 2022), and cellular architecture from cryo-electron tomography (cryo-ET) (Gilbert et al., 2021). Systematic investigations of the interactions amongst Syn3A′s biochemical components were undertaken — defining the metabolic map (Breuer et al., 2019), genetic information processes (Thornburg et al., 2019), and reaction kinetics of coupled metabolic/genetic information processes (Thornburg et al., 2022). By combining these with hybrid stochastic-deterministic methods leveraging GPU-accelerated simulation software (Roberts et al., 2012; Hallock et al., 2014; Bianchi et al., 2018), a well-stirred WCM (WS-WCM) and 3D spatially resolved WCM (4D-WCM) that predict time-dependent Syn3A cell states were created (Thornburg et al., 2022).

However, due to the methodology used to model the chromosome (Gilbert et al., 2021), the existing 4D-WCM was limited to the part of the cell-cycle prior to the onset of DNA replication (Thornburg et al., 2022). This study resolves that issue by transitioning from a lattice polymer model to a continuum polymer model (Figure 1A) of the chromosome, while retaining the previous model's strengths; namely, the ability to fold chromosomes within cellular architectures dictated by cryo-ET and a high spatial resolution (10 bp per monomer) that enables modeling of the hetergeneous diffusion of macromolecular complexes due to excluded-volume interactions with the chromosome. Furthermore, the new method allows for progressive DNA replication of the chromosome to reach nontrivial replication states (Cooper and Helmstetter, 1968; Bremer and Dennis, 2008; Youngren et al., 2014; Khan et al., 2016; Wasim et al., 2021; Pountain et al., 2022) and for the segregation of daughter chromosomes (Goloborodko et al., 2016a; Gogou et al., 2021) under the influence of known essential components (Breuer et al., 2019), SMC-complexes (Ganji et al., 2018; Lee et al., 2021) and topoisomerases (Wang, 1991; McKie et al., 2021; Sutormin et al., 2021; Conin et al., 2022). These nontrivial replication states have *Ori*:*Ter* ratios greater than 2:1 (Figure 2), where *Ori* is the origin of replication and *Ter* is the terminus of replication, and were predicted in Syn3A by WS-WCM simulations and measured by experimental quantitative-PCR (qPCR) (Thornburg et al., 2022). These new capabilities lay the groundwork for the extension of the 4D-WCM to the full cell-cycle. Additionally, by using a binary tree model (Figure 2A) the full spectrum of replication of states for a circular chromosome can be explored and *in silico* chromosome contact maps resolving inter-daughter interactions can be calculated (Figure 3A).

Beyond the information stored in the sequence of the genome, the 3D organization of eukaryotic (Kempfer and Pombo, 2019) and bacterial (Dame et al., 2019; Lioy et al., 2021) genomes plays a role in cellular behavior (Dekker and Mirny, 2016). While imaging techniques such as DNA-FISH (Giorgetti and Heard, 2016) provide insights about targeted interactions, the wide-spread accessibility of next-generation sequencing (Goodwin et al., 2016) catalyzed the proliferation of sequence-based techniques that assess genome-wide interactions such as DNA-protein binding using CHIP-seq (Park, 2009) and DNA-DNA proximity using chromosome conformation capture (3C) (Dekker et al., 2002). Following the creation of 3C, many variations have been developed (Denker and de Laat, 2016; Goel and Hansen, 2020), the most well-known of which is perhaps Hi-C (Lieberman-Aiden et al., 2009). Although researchers have a stunning breadth of experimental data characterizing interactions throughout the genome, computational models (Rosa and Zimmer, 2014; Tiana and Giorgetti, 2019) are required to solve the inverse problem of determining 3D genome organization (Di Pierro et al., 2017; Messelink et al., 2021; Shi and Thirumalai, 2023) and provide mechanistic insights (Sanborn et al., 2015; Fudenberg et al., 2016; Banigan et al., 2020; Fiorillo et al., 2021).

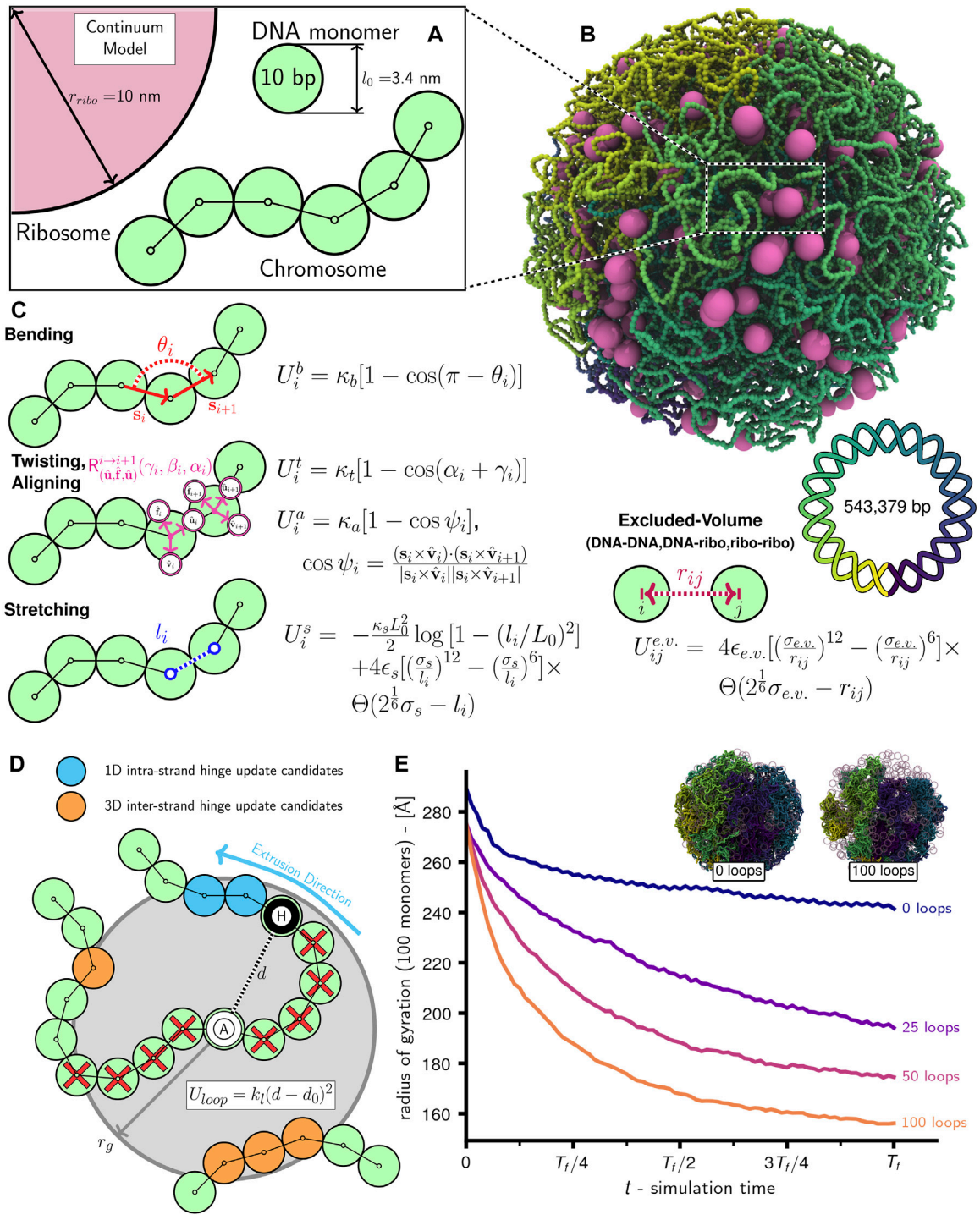


FIGURE 1
Schematic of chromosome modeling protocol: **(A)** Diagram of system with ribosomes and chromosome comprised of 10 bp DNA monomers. **(B)** Snapshot of system with an unreplicated 54,338 monomer Syn3A chromosome and 500 ribosomes in a 200 nm radius cell. **(C)** Bending (U_i^b), twisting (U_i^t), aligning (U_i^a), and stretching (U_i^s) potential energy functions for intramonomer interactions, and potential energy functions for excluded-volume interactions ($U_{ij}^{e.v.}$) between DNA monomers and ribosomes. **(D)** DNA loops are created by applying harmonic bonds between pairs of "anchor" (A, white) and "hinge" (H, black) monomers. Loop-extrusion is simulated by periodically updating the hinge monomer from the set of candidates within the grab radius, r_g. Monomers with a red cross are excluded from hinge updates due to not satisfying the minimal loop length requirement, $L_{\min}$. **(E)** Average windowed radius of gyration as a function of time for simulations of a single unreplicated chromosome with varying numbers of loops. Simulations were run for 4.0E + 6 timesteps with parameters given in Section 2.4.4. Inset are snapshots of the simulations with 0 loops and 100 loops at $t = T_f$.

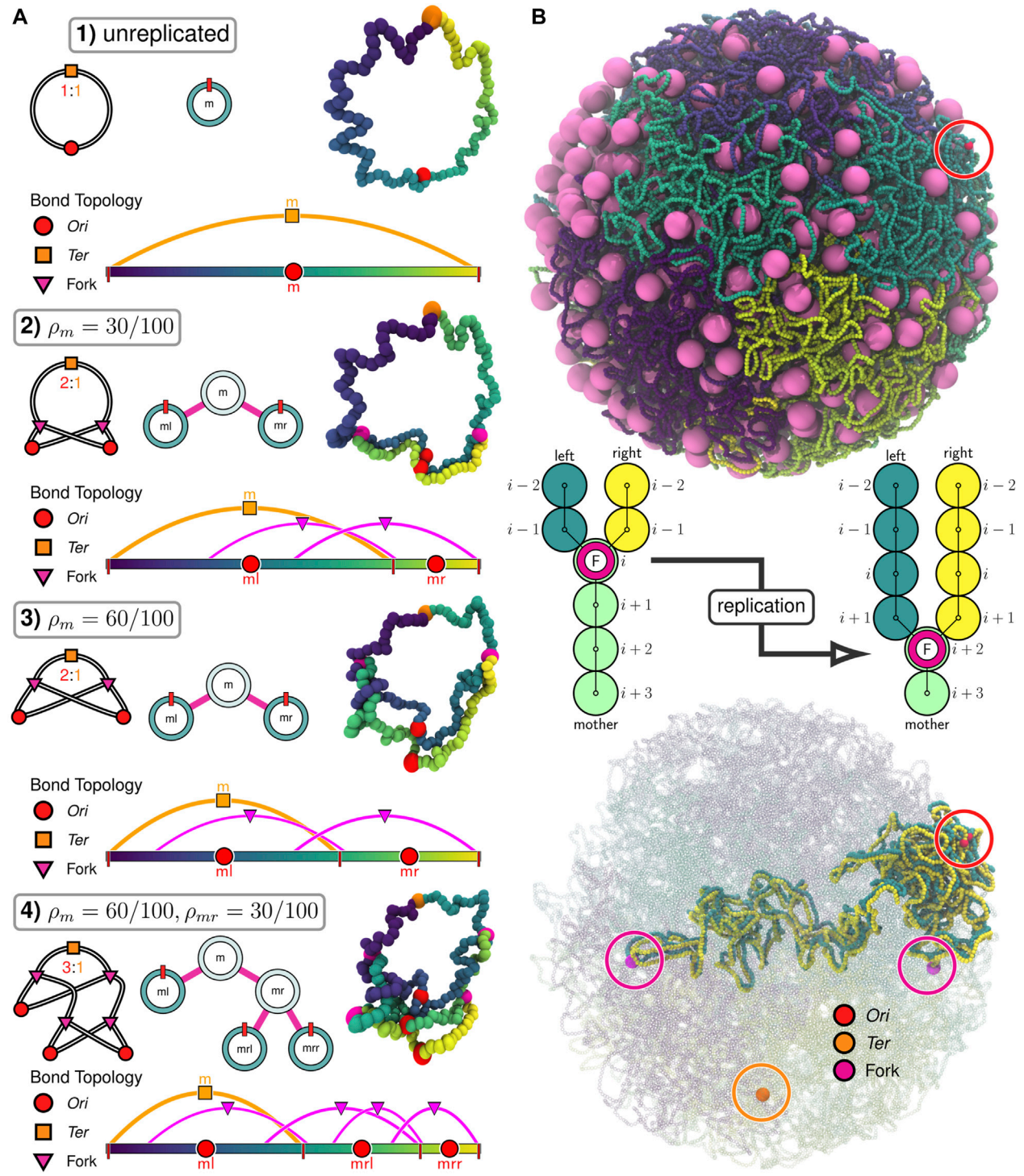


FIGURE 2
Replication of chromosomes in polymer model: **(A)** Progressive replication ($\rho_i = \rho_i^{CW} + \rho_i^{CCW}$) of 100 monomer circular DNA using binary tree model. For each of the four stages of replication, we show the theta structure in the top-left, the binary tree representation in the top-middle, the physical model in the top-right, and the bond topology of the physical model in the bottom. The bond topology displays all monomers using the colorbar at the bottom. Adjacent monomers in regions of the colorbar partitioned by red lines (chromosome boundaries) are bonded. All other bonds in the system (*Ters* creating circular chromosomes and forks creating theta structures) are depicted using arcs between the bonded monomers. **(B)** Beginning with an unreplicated Syn3A chromosome (543,379 bp) within a 200 nm radius cell containing 500 ribosomes, 20,000 bp (2,000 monomers) were replicated using the train-track model (see schematic). The *Oris*, *Ter*, and forks in the replicated system are highlighted with circles.

TABLE 1 Comparative proteomics of proteins in Syn3A that are known to interact with bacterial chromosomes (SMC-scpAB, DNA-gyrase, topoisomerase-IV, and HU). Values were taken from Supplementary Table S1 of (Thornburg et al., 2022).

Protein (stoichiometry)	Locus	# Syn3A	#/Genome size [#/bp]		
			Syn3A	*E. coli*[a]	*B. subtilis*[b]
SMC (2)	0415	202	3.72E-4	2.18E-3[c]	1.07E-4
ScpA (1)	0327	1 (10)[d]	1.84E-5	-	-
ScpB (2)	0328	31	5.71E-5	-	1.88E-5
gyrase-A (2)	0007	298	5.48E-4	1.86E-3	3.05E-4
gyrase-B (2)	0006	244	4.49E-4	1.32E-3	1.67E-4
topoIV-A (2)	0453	156	2.87E-4	2.77E-4	4.83E-5
topoIV-B (2)	0452	157	2.89E-4	1.38E-4	4.71E-5
HU (2)[e]	0350	28	5.15E-5	2.69E-3	2.01E-3

[a]4.6 Mbp genome.
[b]4.4 Mbp genome.
[c]SMC complex in *E. coli* is MukBEF, with stoichiometry of 4:2:4.
[d]Proteins with counts less than 10 were assumed to be a minimum of 10 in whole-cell simulations.
[e]Greatest %-identity with α-subunit (HUα), HU, forms homo- (αα or ββ) and heterodimers (αβ) in *E. coli*.

Syn3A is a compelling system for a systematic study of bacterial chromosomes (Birnie and Dekker, 2020), including but not limited to their replication and segregation, and the bacterial cell cycle (Olivi et al., 2021) because the protein functions encoded by its remaining essential genes hypothetically represent the minimal ingredients necessary for successful proliferation of bacterial cells. Based on what is known of chromosome organizing elements, key among these minimal ingredients should be at least one creating DNA loops (DNA regions distant in sequence but constrained in close spatial proximity) (Davidson and Peters, 2021) and one resolving DNA knots and catenanes (McKie et al., 2021). Syn3A′s genome encodes the prokaryotic condensin complex, SMC-scpAB (Table 1), whose *Saccharomyces cerevisiae* homolog extrudes DNA loops at rates of hundreds of bp per second (Ryu et al., 2021), along with two type-II topoisomerases that allow strand-passage of dsDNA (Liu et al., 1980), DNA-gyrase (gyrase) and topoisomerase-IV (topo-IV) (Table 1) — all of these genes were found to be essential by transposon mutagenesis (Breuer et al., 2019). We compare Syn3A′s proteomics counts of SMC-scpAB and type-II topoisomerases with respect to the model bacteria *E. coli* and *B. subtilis* on the basis of their counts relative to the total DNA content of the genome, as the DNA is what these proteins manipulate. After accounting for the 4:2:4 stoichiometry (Lee et al., 2021) of *E. coli*′s SMC complex, MukBEF, we find that the densities of SMC core proteins per bp of genome are ranked in decreasing order as 1) *E. coli*, 2) Syn3A, 3) *B. subtilis* (Table 1). However, the difference between *E. coli* and *B. subtilis* is only one order-of-magnitude and we can conjecture that this might be due to Syn3A and *E. coli* compensating for their lack of a parABS system (Livny et al., 2007; Badrinarayanan et al., 2015) that preferentially loads SMC complexes onto the chromosome (Marbouty et al., 2015; Tran et al., 2017). We find similar trends among the densities of the two type-II topoisomerases (Table 1). Given the comparable densities of these chromosome-manipulating proteins between Gram-positive (Syn3A and *B. subtilis*) and -negative (*E. coli*) bacteria, we feel that Syn3A is a suitable system in which to study the dynamics of bacterial chromosome organization.

As was noted in a previous study (Gilbert et al., 2021), unlike many bacteria Syn3A codes for a single nucleoid-associated protein (NAP) (Dame, 2005; Lioy et al., 2018; Verma et al., 2019; Lioy et al., 2021), HU (JCVISYN3A_0350), which is known to have two binding modes: 1) low-affinity binding to linear DNA and 2) high-affinity binding to structurally deformed DNA (Kamashev, 2000; Verma et al., 2023). One result of HU binding is the stabilization of supercoiling (Le et al., 2013; Lioy et al., 2018; Strzałka et al., 2022). Curiously, while the HU gene was found to be essential by transposon mutagenesis, the proteomics count is so vastly lower than that of *E. coli*, *B. subtilis*, and related-organism *Mesoplasma florum* (Gilbert et al., 2021) that its genome-wide influence (Pelletier et al., 2012) is likely to be negligible. Furthermore, chromosome contact maps from 3C-seq libraries of Syn3A do not exhibit chromosome interaction domains (CIDs) (Gilbert et al., 2021), which are known to be a result of persistent supercoiling (Le et al., 2013; Trussart et al., 2017; Lioy et al., 2018). Given these considerations, we hypothesize that HU's lingering essentiality in Syn3A is a reflection of it only acting through an interaction specific to the high-affinity binding mode. In *E. coli*, HU is known to interact with replication initiator protein DnaA (Chodavarapu et al., 2007), HUα/DNA stoichiometry has been shown to increase for faster-growing *E. coli* cells (Abebe et al., 2017), and experimental evidence suggests a mechanism of HU promoting duplex unwinding at the *oriC* replication origin (Yoshida et al., 2023). Additionally, HU is essential for replication initiation in Gram-positive *B. subtilis* (Karaboja and Wang, 2022; Schramm and Murray, 2022), whose replication origin is similarly a DnaA-based *oriC*. Based on these results in other bacteria and HU's enhanced binding to dsDNA repair and recombination intermediates (Kamashev, 2000), we believe a small number of HU was retained to fulfill an essential role during replication initiation using a DnaA-based *oriC* in Syn3A (Thornburg et al., 2019), but do not expect it to influence

chromosome-scale organization with its reduced proteomics count, and therefore exclude HU from our model of the chromosome.

Given the absence of NAPs structuring Syn3A′s chromosome into bacterial chromatin (Dame and Tark-Dame, 2016), we chose to model the chromosome of essentially naked dsDNA as a twistable and elastic worm-like chain (Klenin et al., 1998; Brackley et al., 2014; Maffeo and Aksimentiev, 2020). The polymer is comprised of 3.4 nm diameter spherical monomers containing 10 bp of chromosomal DNA (Figure 1A), 54,338 such monomers bonded in a circle are used to create Syn3A′s 543,379 bp circular chromosome (Figure 1B). Adjacent monomers interact through stretching, bending, and twisting potentials (Figure 1C) that reproduce the tensile, bending, and torsional stiffness of dsDNA (Cocco et al., 2002; Brackley et al., 2014), and are parameterized by linear (45 nm) (Manning, 2006; Geggier et al., 2010; Mantelli et al., 2011) and twist (85 nm) (Mosconi et al., 2009) persistence lengths. The monomers are subject to non-bonded interactions that prevent strand-crossings and cause them to avoid ribosomes modeled as 20 nm diameter spherical particles (Figure 1C). We chose to neglect electrostatics and hydrodynamic interactions in this current model. The complete system of chromosomes and ribosomes is simulated using a Brownian dynamics (Snook, 2007) integrator for aspherical particles in LAMMPS (Thompson et al., 2022). To explore the influence of loop-extruding SMC complexes and strand-crossing type-II topoisomerase in this framework, we have developed algorithms to selectively introduce and remove additional terms in the energy function that emulate their effects.

While the computational methodologies described in this paper are tailored to reaching the longer-timescales necessary for WCMs that include fundamental processes of bacterial life such as chromosome replication and segregation, returning to the near-atomic scale provides the ultimate means of validation and reveals additional insights. Researchers have previously completed molecular dynamics (MD) simulations of representative volumes of bacterial cytoplasm (Yu et al., 2016; Rickard et al., 2019; Heo et al., 2022), but only recently has it become possible to prepare a MD simulation of an entire bacterium (Stevens et al., 2023). We will describe how our polymer model for the chromosome can be directly mapped to a coarse-grained Martini model (Marrink et al., 2022) of dsDNA that is ready to be simulated using Gromacs-2023 (Páll et al., 2020; Abraham et al., 2023).

2 Materials and methods

2.1 Twistable polymer model

The chromosome is modeled under the assumption that due to the low density of NAPs in Syn3A, the vast majority of the chromosome is essentially naked dsDNA in a good solvent (Breuer et al., 2019; Szatmári et al., 2020; Thornburg et al., 2022). The naked dsDNA is represented as a twistable and elastic worm-like chain of spherical monomers, each of which contain 10 bp of DNA and have a radius (r_{DNA}) of 1.7 nm. We model the 10 bp monomers as spheres rather than 3.4 nm cylindrical segments with diameters equal to that of a dsDNA helix (2 nm) because using isotropic pair potentials for spherical particles is less computationally intensive. We consider the spherical monomer approximation acceptable for our chromosome-scale model because relative to a cylindrical segment the excluded volume is overestimated by less than a factor of two and the translational damping (Section 2.2) is overestimated by only 15% (Supplementary Analyses). Monomers interact through the energy function from Brackley et al. (Brackley et al., 2014) — the monomers are bonded using finitely extensible nonlinear elastic (FENE) potentials (l_i in Figure 1C), the bending stiffness of dsDNA is implemented using a cosine potential whose argument is the angle (θ_i in Figure 1C) between $(i-1)$-, i-, and $(i+1)$-th monomers, and the torsional stiffness of dsDNA is implemented using a cosine potential whose argument is the sum of Euler angles parameterizing the rotation matrix describing the transformation between the local coordinate systems, $(\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{f}}_i, \hat{\boldsymbol{v}}_i)$, of i- and $(i+1)$-th monomers (α_i and γ_i in Figure 1C). The linear and torsional stiffness parameters, κ_b and κ_t, are determined based on the assumed linear persistence length, l_p, of 45 nm (Manning, 2006; Geggier et al., 2010; Mantelli et al., 2011) and twist persistence length, l_t, of 85 nm (Mosconi et al., 2009), respectively. The alignment term in the potential serves to align the $\hat{\boldsymbol{u}}_i$ basis vector of the i-th monomer's local coordinate system with the displacement vector between the i- and $(i+1)$-th monomers, $\boldsymbol{s}_i$. While the monomer orientations and torsional interactions play a limited role in the current simulations due the assumption of a relaxed supercoiling state, we elected to include them for a few reasons. First, the train-track model of replication (Section 2.6) uses the monomer orientations to specify the coordinates of the daughter chromosomes (Figure 2B). Second, in the future we intend to use the model to investigate chromosome organization due to DNA-binding HU (Lioy et al., 2018) when its expression is restored and to mechanochemically couple transcriptional activity in the 4D-WCM to the torsional state of the chromosome (Liu and Wang, 1987; Chong et al., 2014; Dorman, 2019; Kim et al., 2019; Chatterjee et al., 2021; Guo et al., 2021; Geng et al., 2022).

The chromosome as a whole is modeled as a homopolymer and all monomers, including those representing the *Ori*s, *Ter*s, and forks, have an identical radius of r_{DNA}. The ribosomes are modeled as spheres with a radius (r_{ribo}) of 10.0 nm. Not pictured in Figure 1 are boundary particles with a radius (r_{bdry}) of $5r_{\mathrm{DNA}}$ that create the closed membrane shape. All non-bonded particles interact through purely repulsive Weeks-Chandler-Andersen (WCA) pair potentials (Figure 1C), which serve to prevent dsDNA strand-crossings (Supplementary Video SV1), create the excluded-volume interactions between the chromosome and ribosomes, and confine all DNA monomers and ribosomes within the surface comprised of boundary particles.

The total potential energy function for the chromosome/ribosome system is

$$\begin{aligned} U = &\sum_{i=1}^{N_{\mathrm{DNA}}} \left[U_i^b + U_i^t + U_i^a + U_i^s\right] + \\ &+ \sum_{i=1}^{N_{\mathrm{DNA}}-1} \sum_{j=i+1}^{N_{\mathrm{DNA}}} U_{ij}^{\mathrm{DNA-DNA}} + \sum_{i=1}^{N_{\mathrm{DNA}}} \sum_{j}^{N_{\mathrm{ribo}}} U_{ij}^{\mathrm{DNA-ribo}} \\ &+ \sum_{i=1}^{N_{\mathrm{ribo}}-1} \sum_{j=i+1}^{N_{\mathrm{ribo}}} U_{ij}^{\mathrm{ribo-ribo}} \\ &+ \sum_{i=1}^{N_{\mathrm{bdry}}} \sum_{j}^{N_{\mathrm{DNA}}} U_{ij}^{\mathrm{bdry-DNA}} + \sum_{i=1}^{N_{\mathrm{bdry}}} \sum_{j}^{N_{\mathrm{ribo}}} U_{ij}^{\mathrm{bdry-ribo}}, \end{aligned} \tag{2.1}$$

where the details of the energy functions may be found in Figure 1A. Soft pair potentials of the form

$$U_{ij}^{\text{soft/topo}} = \epsilon_{\text{soft/topo}}\left[1 + \cos\left(\frac{\pi r_{ij}}{\sigma_{ij}}\right)\right], \quad \text{where} \quad r_{ij} < \sigma_{ij} \tag{2.2}$$

are used to reduce overlaps during energy minimizations (replacing $U_{ij}^{\text{DNA–DNA}}$ and $U_{ij}^{\text{DNA–ribo}}$) and permit strand-crossings of DNA (Supplementary Video SV1) under the assumed action of topoisomerases (replacing $U_{ij}^{\text{DNA–DNA}}$). Additionally, the FENE bonds between monomers are replaced with harmonic bonds of the form

$$U_i^s = k_{\min}(l_i - l_0)^2 \tag{2.3}$$

during the initial energy minimizations to prevent over-stretching. Excluding the SMC looping interactions, which are described in greater detail in Section 2.3, all energetic parameters for the potential energy function are listed in Table 2.

2.2 Brownian dynamics

The time-integration was carried out using an OpenMP-accelerated version of the Brownian dynamics integrator for aspherical particles (Delong et al., 2015; Ilie et al., 2015) in LAMMPS (Thompson et al., 2022). The Brownian equation of motion

$$\frac{d\boldsymbol{x}_i}{dt} = \frac{\boldsymbol{F}_{\text{system}} + \boldsymbol{F}_{\text{random}}}{\gamma_i} \tag{2.4}$$

approximates the overdamped limit of the Langevin equation

$$\frac{m_i}{\gamma_i}\frac{d^2\boldsymbol{x}_i}{dt^2} = -\frac{d\boldsymbol{x}_i}{dt} + \frac{\boldsymbol{F}_{\text{system}} + \boldsymbol{F}_{\text{random}}}{\gamma_i}, \tag{2.5}$$

and is only an accurate approximation if the inertial forces are insignificant compared to the viscous forces (Snook, 2007). The mass of the 10 bp monomers is sequence-independent and was calculated as the molar mass of an average 10 bp sequence from Syn3A′s genome (Breuer et al., 2019). We model only complete 70S ribosomes with an assumed mass of 2,700 KDa (Yamamoto et al., 2006). Both ribosomes and DNA monomers are assumed to behave as spherical particles undergoing normal Brownian motion in a Newtonian fluid. In the case of the ribosomes, their characteristic size is 20 nm when we do not include polysomes (multiple ribosomes translating a single mRNA) (Xue et al., 2022), and their motion should be decoupled from metabolic activity due to falling below a 30 nm size threshold (Parry et al., 2014). Although the chromosome is a cytoplasmic component with size well in excess of this threshold, we model the DNA monomers under the same simplifying assumption of normal Brownian motion. In reality, bacterial chromosome dynamics are a result of ATP-dependent motion (Weber et al., 2012), and part of this motion originates from loop-extrusion by SMC (Hirano, 2006), which is addressed by another part of our computational model (Section 2.3). The damping coefficients for the translational and rotational motion of DNA monomers and ribosomes are listed in Table 3. Translational damping constants, γ_i^{T}, were calculated using the Stokes-Einstein equation for spherical particles (Snook, 2007)

$$\gamma_i^{\text{T}} = 6\pi\eta r_i \tag{2.6}$$

TABLE 2 Potential energy parameters for the chromosome and ribosome system. All simulation units are using "units real" in LAMMPS (Thompson et al., 2022).

Parameter	Symbol	Simulation units	
		Quantity	Unit
DNA monomer radius	r_{DNA}	1.7E+1	Å
ribosome radius	r_{ribo}	1.0E+2	Å
boundary particle radius	r_{bdry}	$2.5r_{\text{DNA}}$	Å
eq. monomer spacing	l_0	$2r_{\text{DNA}}$	Å
linear persistence length	l_p	4.5E+2	Å
twist persistence length	l_t	8.5E+2	Å
bending energy	κ_b/k_BT	$l_p/(2r_{\text{DNA}})$	n.d.
twisting energy	κ_t/k_BT	$l_t/(2 \times (2r_{\text{DNA}}))$	n.d.
aligning energy	κ_a	$2\kappa_t$	Kcal/mol
FENE rep. energy	ϵ_s/k_BT	1.0	n.d.
FENE rep. length	σ_s	$2r_{\text{DNA}}$	Å
FENE att. energy	$\kappa_s\sigma_s^2/k_BT$	1.0E+2	n.d.
FENE finite-length	L_0	$1.5\sigma_s$	Å
DNA-DNA WCA energy	$\epsilon_{\text{DNA-DNA}}/k_BT$	1.0	n.d.
DNA-DNA WCA length	$\sigma_{\text{DNA-DNA}}$	$2r_{\text{DNA}}$	Å
DNA-ribo WCA energy	$\epsilon_{\text{DNA-ribo}}/k_BT$	1.0	n.d.
DNA-ribo WCA length	$\sigma_{\text{DNA-ribo}}$	$r_{\text{DNA}} + r_{\text{ribo}}$	Å
ribo-ribo WCA energy	$\epsilon_{\text{ribo-ribo}}/k_BT$	1.0	n.d.
ribo-ribo WCA length	$\sigma_{\text{ribo-ribo}}$	$2r_{\text{ribo}}$	Å
bdry-DNA WCA energy	$\epsilon_{\text{bdry-DNA}}/k_BT$	1.0	n.d.
bdry-DNA WCA length	$\sigma_{\text{bdry-DNA}}$	$r_{\text{bdry}} + r_{\text{DNA}}$	Å
bdry-ribo WCA energy	$\epsilon_{\text{bdry-ribo}}/k_BT$	1.0	n.d.
bdry-ribo WCA length	$\sigma_{\text{bdry-ribo}}$	$r_{\text{bdry}} + r_{\text{ribo}}$	Å
soft pairs	$\epsilon_{\text{soft}}/k_BT$	1.0	n.d.
topoisomerase pairs	$\epsilon_{\text{topo}}/k_BT$	1.0E-1	n.d.
minimization bond energy	$k_{\min}l_0^2/k_BT$	1.0E+3	n.d.

with the dynamic viscosity used previously in the 4D-WCM (Thornburg et al., 2022). Rotational damping constants, γ_i^{R}, were calculated assuming no-slip boundary conditions between the spherical solute particles and surrounding solvent

$$\gamma_i^{\text{R}} = \frac{\gamma_i^{\text{T}} r_i^2}{3}. \tag{2.7}$$

the timestep, Δt = 0.1 ns, was selected such that it satisfies the conditions of the overdamped limit of the Langevin equation, $\Delta t \gg m_i/\gamma_i^{\text{T}}$ and $\Delta t \gg I_i/\gamma_i^{\text{R}}$ (where $I_i = 2m_i r_i^2/5$), while remaining small enough to prevent unphysical strand crossings (Supplementary Video SV1). The boundary particles are held fixed at their initial coordinates and are not subject to coordinate updates due to energy minimizations nor time-integrations.

2.3 SMC-induced DNA loops

The 3D loop-extruding action of SMC protein complexes are simulated using the methodology of Bonato and Michieletto (Bonato and Michieletto, 2021; Ryu et al., 2021), which simulates

TABLE 3 Time-integration parameters for the Brownian dynamics simulations. All simulation units are using "units real" in LAMMPS (Thompson et al., 2022).

Parameter	Symbol	Simulation units	
		Quantity	Unit
thermal energy	k_BT	6.16	Kcal/mol
DNA monomer mass	m_{DNA}	6.18E+3	g/mol
ribosome mass	m_{ribo}	2.11E+6	g/mol
DNA monomer rotational inertia	I_{DNA}	7.14E+5	(g/mol)·Å^2
ribosome rotational inertia	I_{ribo}	8.45E+9	(g/mol)·Å^2
dynamic viscosity	η	7.04E+1	(g/mol)/(fs·Å)
monomer translational damping	$\gamma^{\text{T}}_{\text{DNA}}$	2.39E+4	(g/mol)/fs
ribosome translational damping	$\gamma^{\text{T}}_{\text{ribo}}$	2.81E+5	(g/mol)/fs
monomer rotational damping	$\gamma^{\text{R}}_{\text{DNA}}$	9.21E+6	(g/mol)·Å^2/fs
ribosome rotational damping	$\gamma^{\text{R}}_{\text{ribo}}$	1.50E+10	(g/mol)·Å^2/fs
monomer translational time-scale	$\tau^{\text{T}}_{\text{DNA}}$	2.74E-1	fs
ribosome translational time-scale	$\tau^{\text{T}}_{\text{ribo}}$	1.59E+1	fs
monomer rotational time-scale	$\tau^{\text{R}}_{\text{DNA}}$	8.21E-2	fs
ribosome rotational time-scale	$\tau^{\text{R}}_{\text{ribo}}$	4.77E+0	fs
simulation timestep	Δt	1.0E+5	fs

the action of SMC heads associating with the DNA and then translocating the DNA between the head and the hinge (Nunez et al., 2019). DNA loops are created by adding harmonic bonds between "anchor" and "hinge" monomers (Figure 1D)

$$U_{\text{loop}} = k_l\,(d - d_0)^2, \tag{2.8}$$

rather than explicitly simulating the conformational changes of SMC protein complexes (Higashi et al., 2021; Nomidis et al., 2022). Due to physical considerations of the bending stiffness of dsDNA, the anchor and hinge monomers of all loops are required to be separated by a minimal loop length, L_{min}, in units of bonded monomer distance (Figure 1D). Loops are initialized by first identifying regions of the chromosome accessible to loops by determining contiguous series of bonded monomers that are partitioned by replication forks at either end. Anchors are then randomly assigned to each of the regions with a probability proportional to the number of monomers in the region relative to the total number of looping accessible monomers across all regions. The region-assigned anchors are distributed uniformly within their respective regions. Finally, for each anchor a matching hinge is selected in a random direction along the polymer, and at a distance of bonded monomers that is equal to the minimal loop length.

Loop extrusion is simulated by periodically pausing the time-integration and updating the positions of the hinges while leaving the anchors fixed. There are two possible events during these hinge-update steps (Bonato and Michieletto, 2021) — 1) intra-strand motion in which the hinge advances in 1D along the current strand in the previously assigned direction or 2) inter-strand motion in which the hinge unbinds from the current strand with probability p_{unbind} and rebinds to a new strand within a 3D spherical volume centered about the anchor (Figure 1D). For this study we made the simplifying assumption that only intra-strand motion is permitted ($p_{\text{unbind}} = 0$), which has been used in other studies (Ryu et al., 2021), but the software is capable of simulating inter-strand motion. For both types of updates, only monomers whose distance from the anchor monomer is less than the grab radius, r_g, and in the case of intra-strand updates, whose bonded monomer distance on the current strand is greater than the minimal loop length, are considered as viable update candidates (Figure 1D). The grab radius is chosen to be 50 nm based on the coiled-coil structure of SMC protein complexes (Diebold-Durand et al., 2017). Based on results showing that eukaryotic SMC complexes can traverse one-another to form Z-loops (Kim et al., 2020), we do not include any interactions between hinge and anchors that are not paired.

If the first case of intra-strand motion is selected with probability $1 - p_{\text{unbind}}$, the update monomer is selected from the set of intra-strand candidates by sampling a Poisson distribution with mean $L_{\text{ext-avg}}$ and truncated at $L_{\text{ext-max}}$. Based on step-size distributions measured with magnetic tweezers (Ryu et al., 2021) and analytical calculations (Takaki et al., 2021), we chose these to be $L_{\text{ext-avg}} = 20$ monomers (68 nm) and $L_{\text{ext-max}} = 30$ monomers (102 nm). Should there be no intra-strand candidates, the hinge will remain at its current monomer. If the second case of inter-strand motion is selected with p_{unbind}, the update monomer is selected from the set of inter-strand candidates with equal probability. Should there be no inter-strand candidates following an unbinding, the hinge will remain unbound until there are inter-strand candidates in a subsequent hinge update step. The pseudocode for this process is presented in Supplementary Algorithm S1.

The length-scale of the grab radius is much greater than that of pairwise interactions between non-bonded DNA monomers, we therefore make the simplifying assumption that the DNA monomers available as hinge update candidates have a nearly uniform distribution within the spherical volume of radius r_g centered about any anchor. Under such conditions, the average separation distance, $\bar{d}$, between the anchor and hinge following a hinge update may then be calculated as

$$\bar{d} = \frac{\int_0^{r_g} \mathrm{d}r\,(r \times 4\pi r^2)}{\int_0^{r_g} \mathrm{d}r 4\pi r^2} = \frac{3}{4} r_g, \tag{2.9}$$

the loop will then perform on average the amount of work, $\bar{W}_{\text{loop}}$, necessary to pull the hinge and anchor to their equilibrium separation distance

$$\bar{W}_{\text{loop}} = -\left[U_{\text{loop}}(d_0) - U_{\text{loop}}(\bar{d})\right] = k_l\,(\bar{d} - d_0)^2. \tag{2.10}$$

given that each extrusion event (emulated by hinge updates and subsequent pulling in this case) was measured to complete approximately $4k_BT$ of work (Ryu et al., 2021) and ATP hydolysis is sufficient to provide this, we estimate the spring constant in our model to be

$$k_l = \frac{4k_BT}{(\bar{d} - d_0)^2}. \tag{2.11}$$

all spatial, energetic, and probabilistic parameters for the loop-extrusion model are listed presented in Table 4.

2.4 Polymer model simulation protocols

2.4.1 Simulation software

All polymer model simulations were performed using the C++ program btree_chromo (Supplementary Table S1), which implements the binary tree model of replication states, replication within the chromosome system using the train-track model, and Brownian dynamics simulations that include the effects of SMC complexes and topoisomerases by calling LAMMPS as a library (Thompson et al., 2022). This program is executed from the command-line and takes a single input script of program directives that it then parses into commands and parameters before executing in sequence. Additionally, a number of metacommands are included that allow for sections of the script to be repeated within loops and other similar functions that aid in defining simulation protocols. All directives are documented within the project's repository (Supplementary Table S1). Spatial, energetic, and temporal parameters for the model and subroutines that are regularly performed during the course of simulations are stored within a separate directory as a set of LAMMPS input scripts that are fed into the LAMMPS simulation object. The directory containing LAMMPS input scripts can be redefined, allowing the user to systematically test alternate chromosome models or change models on-the-fly within a simulation. Walltimes for a representative selection of the simulations presented in this study are included in Supplementary Table S2.

2.4.2 Generating initial conditions

Initial configurations of the chromosome are generated using an algorithm based on a midpoint-displacement approach (Fournier et al., 1982) that builds three-dimensional, closed curves resembling Koch curves (von Koch, 1904) out of spherocylinder segments (i.e., cylinders with hemispherical caps) that overlap about the centerpoint of the caps (Supplementary Figure S1A). Given a spherical cell containing a known spatial distribution of ribosomes, the initially unrelaxed configuration of the continuum model is placed within the confines of the spherical cell by growing a circular and self-avoiding chain of spherocylinders. The freely-jointed chain of spherocylinders uses a series of decreasing cylinder lengths during the growth process to generate a chromosome configuration organized as a fractal globule (Lua et al., 2004) with clearly-defined chromosomal territories (Lieberman-Aiden et al., 2009; Sanborn et al., 2015), which is consistent with our previous lattice methodology (Gilbert et al., 2021). This is accomplished using an iterative procedure in which a specified number of spherocylinder segments are added. Self- and ribosome-avoidance are imposed at every stage between the spherocylinder segments and the spherical ribosomes. Furthermore, tracking the crossing of the spherocylinders during segment addition steps was used to prevent the introduction of knots. In the final step, spherical monomers with radii equal to the spherocylinder radii (17.0 Å) are then interpolated along the spherocylinders and any remaining monomers are inserted using an equivalent midpoint-displacement method. The model of an unreplicated Syn3A chromosome is comprised of 54,338 monomers, each containing 10 bp. This method creates suitable chromosome configurations for both the small and large Syn3A cell geometries and ribosome distributions reconstructed from cryo-ET (Gilbert et al., 2021) (Supplementary Figures S1B–C) and has been further extended to fill cell geometries with multiple circular chromosomes simultaneously (Supplementary Figure S2).

TABLE 4 Energetic, spatial, and probabilistic parameters for SMC loops. All simulation units are using "units real" in LAMMPS (Thompson et al., 2022).

Parameter	Symbol	Simulation units	
		Quantity	Unit
equilibrium bond distance	d_0	$4r_{\mathrm{DNA}}$	Å
grab radius	r_g	500.0	Å
average grab distance	$\bar{d}$	$3r_g/4$	Å
spring constant	k_l	2.61E-2	Kcal/(mol·Å^2)
minimum loop length	L_{min}	5	# monomers
average 1D extrusion length	$L_{\mathrm{ext\text{-}avg}}$	20	# monomers
maximum 1D extrusion length	$L_{\mathrm{ext\text{-}max}}$	30	# monomers
unbinding probability	p_{unbind}	0.0	n.d.

2.4.3 Standard polymer model simulations

At the start of any polymer model simulation and before any Brownian dynamics steps are taken, potential particle overlaps are relaxed by running the following sequence of minimizations and short runs (Table 5): 1) `minimize_soft_harmonic`, 2) `run_soft_harmonic`, 3) `minimize_hard_harmonic`, 4) `run_hard_harmonic`, and 5) `minimize_hard_FENE`. The stopping criteria and maximum number of iterations for each of these are defined within the directory of input scripts. This is sufficient to relax the initial structure without significantly altering it, while remaining tolerant to the insertion of new monomers, ribosomes, or reshaping of the boundary. Brownian dynamics integration then proceeds using `run_hard_FENE` to simulate the system with stretching, bending, and twisting of the dsDNA polymer while preventing strand-crossings. Following replication using the train-track model (Figure 2B), the system is relaxed using the previously mentioned protocol to resolve particle overlaps that may have resulted from the addition of new monomers.

2.4.4 Simulations with SMC-looping and topoisomerases

Given that SMC complexes and topoisomerases were identified to be essential in Syn3A by transposon mutagenesis, we developed a simulation method to describe their interaction with the DNA at the scale of the full chromosome. Simulations of systems that include SMC-looping and the action of topoisomerases are performed using an algorithm that iteratively alternates between updating loop locations, minimizing the now non-equilibrium system's energy, and performing Brownian dynamics steps (Supplementary Algorithm S2). We chose to use this approach because the small timesteps (Δt = 0.1 ns) used to prevent strand-crossings of the 10 bp monomers would otherwise prevent us from running Brownian dynamics over timescales required for multiple loop-extrusion steps that occur on the order of seconds (Ryu et al., 2021). Intermittently, this process is stopped to run a set of Brownian dynamics steps with DNA-DNA pair interactions replaced by soft potentials permitting strand-crossings, `run_topoDNA_FENE` (Table 5). This and similar approaches have been used in previous studies to model the net effect of

TABLE 5 Models used during energy minimizations (`minimize_"bonds_pairs"`) and Brownian dynamics time-integrations (`run_"bonds_pairs"`) of the system. Hard-pair interactions are used between boundary particles and all other particles for every model.

Model	DNA bonds	Pair interactions		
		DNA-DNA	DNA-ribo	ribo-ribo
`soft_harmonic`	harmonic	soft	soft	soft
`soft_FENE`	FENE	soft	soft	soft
`hard_harmonic`	harmonic	hard	hard	hard
`hard_FENE`	FENE	hard	hard	hard
`topoDNA_harmonic`	harmonic	topo	hard	hard
`topoDNA_FENE`	FENE	topo	hard	hard

topoisomerases (Goloborodko et al., 2016a; Mitra et al., 2022b). We note that this better emulates topo-IV rather than gyrase, but we feel this is appropriate given that topo-IV is known to primarily decatenane replication products (Zechiedrich et al., 1997; Cebrián et al., 2015). The number of loops, duration of loop simulations before updates (Δt_{loops}), frequency of topoisomerase runs (T_{topo}), and duration of topoisomerase runs (ΔT_{loops}) are specified by the user. For the simulations in this study we used the following values in units of timesteps: Δt_{loops} = 10,000, T_{topo} = 50,000, Δt_{loops} = 50,000. Additionally, this algorithm was restarted every 100,000 timesteps to sample new locations for the loop anchors. Simulations show that increased loop numbers lead to greater chromosome compaction (Figure 1E), with 100 loops reducing the windowed radius of gyration by approximately 35% relative to the case with 0 loops.

2.5 Replication states

Beyond the configurational state of the chromosome, we wish to consider the replication state of the chromosome system. We will use a binary tree model (Taylor and Garnier, 2009) (Figure 2A), where the replication state is described by the extent of replication for each of the possible *Oris*. The *Oris* are labeled by their lineage relative to the mother chromosome (*m*), i.e., the root of the tree. For example, replication of the mother chromosome produces two new daughter *Oris*, a left daughter (*ml*) and a right daughter (*mr*). This pattern continues for subsequent generations, i.e., the mother's right daughter (*mr*) will create the daughter *Oris* labeled *mrl* and *mrr* when it undergoes replication (Figure 2A). Aside from the initial mother chromosome, we uniformly refer to *Oris* represented as leaves in the binary tree (Figure 2A) as "daughters" and use the label to describe the generation, i.e., a daughter (*ml*) vs. a granddaughter (*mrl*).

If we assume the mother is the zero-th generation, we can write the space of labels for the q-th generation as $\boldsymbol{I}_q = \{I_0, I_1, \ldots, I_{q-1}, I_q\}$, where $I_0 = m$ and $I_j \in (l, r)$ for all $j > 0$. This is essentially a q-dimensional vector of binary values (the zero-th element is trivially constant), but for clarity we will write it as a list of labels selecting the left/right daughters at each generation. If we have a chromosome in the q-th generation with the label $\boldsymbol{i}_q$, then we denote the labels of its daughters in the $(q + 1)$-th generation as $\boldsymbol{i}_{ql}$ and $\boldsymbol{i}_{qr}$. Conversely, if we have a chromosome in the q-th generation with label $\boldsymbol{i}_q$, then we denote the label of its mother in the $(q - 1)$-th generation as $\boldsymbol{i}_{q(-)}$.

The genomic content of any daughter chromosome is determined by the extent of replication of its mother. i.e., the genomic content of the chromosome labeled $\boldsymbol{i}_q$ is determined by the extent of replication, ρ, of the chromosome labeled $\boldsymbol{i}_{q(-)}$. Given this, the replication microstate of some general chromosome system with a maximum generation of q is given by the vector

$$\begin{aligned}\boldsymbol{\rho}_q = \Big\{&\rho^{cw}_{\boldsymbol{i}_0}, \rho^{ccw}_{\boldsymbol{i}_0},\\ &\rho^{cw}_{\boldsymbol{i}_{0l}}, \rho^{ccw}_{\boldsymbol{i}_{0l}}, \rho^{cw}_{\boldsymbol{i}_{0r}}, \rho^{ccw}_{\boldsymbol{i}_{0r}},\\ &\rho^{cw}_{\boldsymbol{i}_{1l}}, \rho^{ccw}_{\boldsymbol{i}_{1l}}, \rho^{cw}_{\boldsymbol{i}_{1r}}, \rho^{ccw}_{\boldsymbol{i}_{1r}},\\ &\vdots\\ &\rho^{cw}_{\boldsymbol{i}_{(q-1)l}}, \rho^{ccw}_{\boldsymbol{i}_{(q-1)l}}, \rho^{cw}_{\boldsymbol{i}_{(q-1)r}}, \rho^{ccw}_{\boldsymbol{i}_{(q-1)r}}\Big\},\end{aligned} \tag{2.12}$$

where $\rho^{cw}_{\boldsymbol{i}}$ and $\rho^{ccw}_{\boldsymbol{i}}$ denote the extent of replication in the clockwise and counter-clockwise directions, respectively, of the chromosome with the label $\boldsymbol{i}$. For example, replication state 2 in Figure 2A is a replicating chromosome with replication proceeding from the *Ori* to the *Ter* in both clockwise and counter-clockwise directions. For notational convenience, it is assumed that $\boldsymbol{i}_q$ includes all variations of labels in the q-th generation. For example, $\boldsymbol{i}_2$ includes {*mll*, *mlr*, *mrl*, *mrr*} and $\boldsymbol{i}_{2l}$ includes all 4 possible left daughters originating from the chromosomes with these labels. The number of dimensions of $\boldsymbol{\rho}_q$ increases geometrically as a function of the number of considered generations as 2^q. We purposefully neglect to include the terms for replication extents deeper in the binary tree that are trivially zero.

The replication microstates are subject to two constraints. First, the extent of replication of the daughter chromosome with label $\boldsymbol{i}_s$ may not exceed that of its mother with label $\boldsymbol{i}_{s(-)}$, i.e.,

$$\rho^{cw}_{\boldsymbol{i}_s} < \rho^{cw}_{\boldsymbol{i}_{s(-)}} \quad \text{and} \quad \rho^{ccw}_{\boldsymbol{i}_s} < \rho^{ccw}_{\boldsymbol{i}_{s(-)}}. \tag{2.13}$$

this constraint is included because it is physically impossible for a daughter to replicate DNA sequences that do not yet exist. Second, the total replication extent, $\rho_{\boldsymbol{i}_s}$, must be less than or equal to the total genomic content of the chromosome, i.e.,

$$\rho_{\boldsymbol{i}} = \rho^{cw}_{\boldsymbol{i}} + \rho^{ccw}_{\boldsymbol{i}} \leq 1. \tag{2.14}$$

these two constraints guarantee that only physically realistic replication states are permitted by the model. A change in replication microstate is denoted as

$$\Delta\boldsymbol{\rho} = \left\{\Delta\rho_i^{cw} = a, \Delta\rho_i^{ccw} = b, \Delta\rho_j^{cw} = c, \Delta\rho_j^{ccw} = d, \ldots\right\}, \tag{2.15}$$

where only the forks with a nonzero change are included. Changes that lead to replication states not satisfying the two constraints are instead completed up to the maximum extent at which the constraints are still satisified.

We have previously presented a formal definition of replication microstates, we now turn to characterizing replication macrostates using state variables that correspond to experimental measurements. We begin this by defining a number of quantities that are measurable by experiments for the replication microstates. The total DNA content of a replication microstate relative to the DNA content of a single, unreplicated chromosome is given by

$$G\left(\boldsymbol{\rho}_q\right) = 1 + \sum_{p=0}^{q-1}\sum_{i\in\boldsymbol{I}_p}\left(\rho_i\right) \tag{2.16}$$

and corresponds to experimental measurements of the DNA content, such as fluorescent intensity of stained DNA. The number of *Ori*s in a replication microstate is given by

$$N_{Ori}\left(\boldsymbol{\rho}_q\right) = 1 + \sum_{p=0}^{q-1}\sum_{i\in\boldsymbol{I}_p}\Theta\left(\rho_i^{cw} + \rho_i^{ccw}\right), \tag{2.17}$$

where Θ is again the Heaviside step-function. The number of *Ter*s in a replication microstate is given by

$$N_{Ter}\left(\boldsymbol{\rho}_q\right) = 1 + \sum_{p=0}^{q-1}\sum_{i\in\boldsymbol{I}_p}\left[\Theta\left(\rho_i^{cw} - 1/2\right) + \Theta\left(\rho_i^{ccw} - 1/2\right)\right], \tag{2.18}$$

the ratio of the most-replicated region to the least-replicated region in a replication microstate is the number of *Ori*s divided by the number of *Ter*s and is given by

$$\Upsilon\left(\boldsymbol{\rho}_q\right) = N_{Ori}\left(\boldsymbol{\rho}_q\right)/N_{Ter}\left(\boldsymbol{\rho}_q\right) \tag{2.19}$$

and corresponds to experimental measurements comparing the relative quantities of target sequences, such as qPCR. Given experimental measurements of the DNA content, $G_{\text{exp.}}$ and $\Upsilon_{\text{exp.}}$, in a population of cells, and a maximum possible generation, p, we wish to determine the distribution of replication microstates, $\mathsf{P}(\boldsymbol{\rho})$, whose ensemble averages ($\langle G\rangle$ and $\langle\Upsilon\rangle$) match these experimental constraints. In other words, find $\mathsf{P}(\boldsymbol{\rho})$ such that

$$1 = \langle 1\rangle = \sum_{\{\boldsymbol{\rho}\}}\mathsf{P}\left(\boldsymbol{\rho}\right) \tag{2.20}$$

$$G_{\text{exp}} = \langle G\rangle = \sum_{\{\boldsymbol{\rho}\}}G\left(\boldsymbol{\rho}\right)\mathsf{P}\left(\boldsymbol{\rho}\right) \tag{2.21}$$

$$\Upsilon_{\text{exp}} = \langle\Upsilon\rangle = \sum_{\{\boldsymbol{\rho}\}}\Upsilon\left(\boldsymbol{\rho}\right)\mathsf{P}\left(\boldsymbol{\rho}\right) \tag{2.22}$$

are satisfied.

2.6 Train-track model of replication

In the "train-track" model of bacterial DNA replication (Gogou et al., 2021), the replisomes are thought to independently traverse the opposite arms of the mother chromosome while replicating the DNA (Dingman, 1974). Recent work has provided additional evidence for the train-track model by imaging independently moving replisomes using fluorescently labeled β-clamps (DnaN) in *E. coli* cells with synchronized replication initiation (Japaridze et al., 2020). We assume that DNA replication in Syn3A obeys the train-track model due to the aforementioned experimental evidence and the absence of multi-protein regulatory systems coded for in the minimized genome (Breuer et al., 2019; Thornburg et al., 2022).

In our implementation of the train-track model, monomers are added to the left and right daughter chromosomes following replication events by creating pairs of additional monomers centered about the location of the mother's corresponding monomer (Figure 2B). For convenience, we will denote the spatial coordinates of the i-th monomer of mother, left daughter, and right daughter as $\boldsymbol{x}_i^m$, $\boldsymbol{x}_i^l$, and $\boldsymbol{x}_i^r$, respectively, and similarly denote the orientation quaternions as $\boldsymbol{q}_i^m$, $\boldsymbol{q}_i^l$, and $\boldsymbol{q}_i^r$. The coordinates of the newly-replicated left and right daughter monomers are

$$\boldsymbol{x}_i^l = \boldsymbol{x}_i^m + r_{\text{DNA}}\left[\boldsymbol{q}_i^m\hat{\boldsymbol{e}}_y\left(\boldsymbol{q}_i^m\right)^{-1}\right]$$
$$\text{and}$$
$$\boldsymbol{x}_i^r = \boldsymbol{x}_i^m - r_{\text{DNA}}\left[\boldsymbol{q}_i^m\hat{\boldsymbol{e}}_y\left(\boldsymbol{q}_i^m\right)^{-1}\right], \tag{2.23}$$

where $\hat{\boldsymbol{e}}_y$ is a quaternion whose scalar component is zero and vector component is the unit basis vector in the y-direction. The orientations of the newly-replicated left and right daughter monomers are

$$\boldsymbol{q}_i^l = \boldsymbol{q}_i^m \quad \text{and} \quad \boldsymbol{q}_i^r = \boldsymbol{q}_i^m. \tag{2.24}$$

This method is applicable to nontrivial replication states (Figure 2A), efficiently replicates the chromosome in crowded environments (Figure 2B), and can occur mid-simulation (Supplementary Video SV2). Additionally, because this method is based on the binary tree model, it can be applied for replication events involving multiple forks (e.g., $\Delta\boldsymbol{\rho} = \{\Delta\rho_m^{cw} = 10, \Delta\rho_m^{ccw} = 10, \Delta\rho_{ml}^{cw} = 5, \Delta\rho_{ml}^{cw} = 5\}$) by hierarchically replicating the new monomers. The number of monomers that will be replicated can range from 0 up to the number of monomers of unreplicated DNA along the mother chromosome.

For the purposes of this model, we neglect to include the difference in the leading-versus lagging-strands, and model the fork itself as a standard DNA monomer. We add a harmonic angle potential of the form

$$U_{\text{fork}}^b = k_{\text{fork}}\left(\theta - \theta_0\right)^2 \tag{2.25}$$

between the following triplets of particles formed by the fork (f) and three bonded monomers, mother (m), left (l), and right (r): (m-f-l), (m-f-r), and (l-f-r). The parameters are $\theta_0 = 2\pi/3$ radians and $k_{\text{fork}} \times (1\text{ radian})^2 = \kappa_b$. Additionally, there are no torsional interactions between (f-m), (f-r), or (f-l).

2.7 Chromosome segregation calculations

Given a pair of replication forks producing left and right daughters, each of which may themselves be potentially

undergoing replication, the sets of N_l and N_r replicated monomers belonging to the left and right daughters are $\{\boldsymbol{x}_i^l\}$ and $\{\boldsymbol{x}_i^r\}$, respectively. For example, in state 4 shown in Figure 2A, the daughter sizes are $N_l = 60$ and $N_r = 90$ for fork m and $N_l = 30$ and $N_r = 30$ for fork mr. Segregation of the daughter chromosomes can be investigated using these sets of coordinates for all pairs of forks in a system with a nontrivial replication state by analyzing the disentanglement and the partitioning.

2.7.1 Disentanglement

The number of monomers belonging to the same (s) daughter within a radius, R, of the i-th replicated monomer of the left/right (l/r) daughter are

$$n_i^{s,(l/r)}(R) = \sum_{j=1,j\neq i}^{N_{(l/r)}} \Theta\left(R - \left|\boldsymbol{x}_i^{(l/r)} - \boldsymbol{x}_j^{(l/r)}\right|\right) \tag{2.26}$$

and the number of monomers belonging to the opposite (o) daughter within that radius are

$$n_i^{o,(l/r)}(R) = \sum_{j=1}^{N_{(r/l)}} \Theta\left(R - \left|\boldsymbol{x}_i^{(l/r)} - \boldsymbol{x}_j^{(r/l)}\right|\right). \tag{2.27}$$

the fraction of monomers on the same daughter within the radius about the i-th monomer is

$$\varphi_i^{(l/r)}(R) = \frac{n_i^{s,(l/r)}(R)}{n_i^{s,(l/r)}(R) + \left[N_{(l/r)}/N_{(r/l)}\right] \times n_i^{o,(l/r)}(R)}, \tag{2.28}$$

the average fraction for each daughter is

$$\bar{\varphi}^{(l/r)}(R) = \frac{1}{N_{(l/r)}} \sum_{i=1}^{N_{(l/r)}} \varphi_i^{(l/r)}(R), \tag{2.29}$$

and the degree of disentanglement (DoD) is a function of these

$$\mathrm{DoD}(R) = f\left(\bar{\varphi}^l(R), \bar{\varphi}^r(R)\right). \tag{2.30}$$

we use the harmonic mean for this function as it provides a conservative estimate, then shift and scale the result such that the range of the degree of disentanglement is [0,1]

$$f\left(\bar{\varphi}^l(R), \bar{\varphi}^r(R)\right) = 2 \times \left(\frac{2 \times \left[\bar{\varphi}^l(R) \times \bar{\varphi}^r(R)\right]}{\bar{\varphi}^l(R) + \bar{\varphi}^r(R)} - \frac{1}{2}\right). \tag{2.31}$$

using this definition, 0 corresponds to a fully entangled system that overlaps everywhere and 1 corresponds to a disentangled system whose constituent parts are separated by at least a distance R. When calculating the DoD for our system we use $R = 4r_{\mathrm{DNA}}$.

2.7.2 Partitioning

We evaluate the extent to which the daughter chromosomes are partitioned by calculating the distance between their centers of mass (CoM)

$$d_{\mathrm{CoM}} = |\boldsymbol{X}_{\mathrm{CoM}}^l - \boldsymbol{X}_{\mathrm{CoM}}^r|, \tag{2.32}$$

where

$$\boldsymbol{X}_{\mathrm{CoM}}^{(l/r)} = \frac{1}{N_{(l/r)}} \sum_{i=1}^{N_{(l/r)}} \boldsymbol{x}_i^{(l/r)}. \tag{2.33}$$

the d_{CoM} was then compared to a length-scale characteristic of what we will refer to as an "ideal partitioning". In an ideal partitioning, we assume the daughters will occupy volumes that are proportional to their relative sizes, N_l and N_r, in units of monomers and share a planar interface with minimal surface area. Given a radius of the spherical confinement, r, we then determine the distance between their centers of mass in this ideal scenario, which we will refer to as $L_{\mathrm{partition}}(N_l, N_r, r)$ (Supplementary Material).

2.8 Intra- and inter-daughter contact calculations

While the interactions between equivalent loci on daughter chromosomes are distinguishable in the *in silico* model, they are indistinguishable to most sequence-based experimental techniques, such as 3C methods. However, efforts have been made to resolve these interactions in eukaryotic systems with sister-chromatid-sensitive Hi-C (Mitter et al., 2020) and bacterial systems with recombinase assays (Lesterlin et al., 2012; Espinosa et al., 2020; Oomen et al., 2020). We extend our methodology for *in silico* contact maps (Supplementary Material) to the case of replicating chromosomes by using the relative position within the bond topology (Figure 2A) of the monomer identified as the *Ori* to determine equivalent loci containing identical DNA sequences on daughter chromosomes (Figure 3A). If F is the true contact map encoding the entirety of all intra- and inter-daughter interactions, then we will denote the sequence-equivalent map reflecting 3C observations as $\tilde{\mathsf{F}}$. The sequence-equivalent map is determined by summing the contributions for each of the possible interactions (Figure 3B) before rebalancing the resulting matrix. Additionally, this methodology can be further extended to address the determination of *in silico* contact maps that represent a mixture of chromosomes in different replication states (Supplementary Figures S6, S7) by calculating weighted averages of sequence-equivalent maps, which outside of isolated cases (Nagano et al., 2013; Ramani et al., 2017; Kos et al., 2021), are what 3C libraries are ultimately measuring within a population of unsynchronized cells.

2.9 Martini model preparation

Simulating a Martini model of the Syn3A chromosome requires CG starting coordinates and a CG topology that specifies all the bonded and non-bonded interactions of the DNA model (Uusitalo et al., 2015). In traditional protocols, both are generated by forward mapping an all-atom structure to Martini resolution (Uusitalo et al., 2015; 2017; Kroon et al., 2022). However, given the size of the chromosome, this approach becomes infeasible. Thus we follow a strategy that splits the generation of topology and coordinates into two separate steps. First, we generate starting coordinates at Martini resolution directly from the polymer model's coordinates using a new backmapping protocol. In the second step, the chromosome topology is generated from the genome sequence. Both steps are implemented in a Python package, Polyply, which focuses on

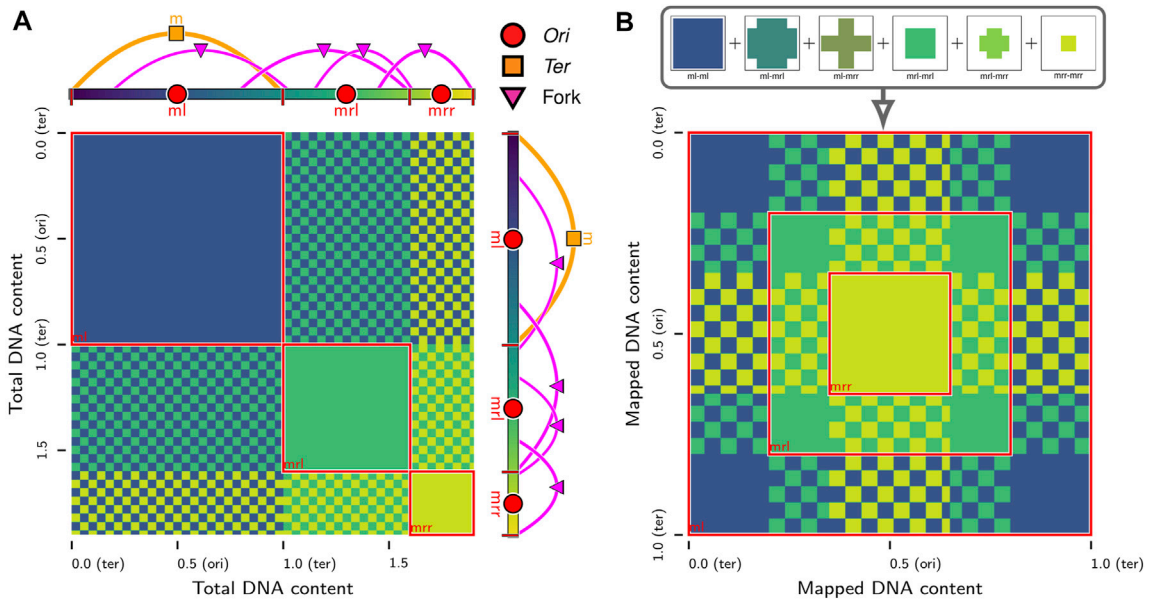


FIGURE 3
In silico contact calculations for replicating chromosomes: **(A)** The true contact map, F, of intra- and inter-daughter interactions in a replicating chromosome system with a nested theta structure. The system presented is replication state 4 in Figure 2A. Solid-colors indicate self-interactions of daughter chromosomes and checkerboard patterns indicate interactions between pairs of daughter chromosomes. **(B)** Mapping of loci in contact map of replicating chromosome system in **(A)** to equivalent loci with identical sequences in unreplicated system. The overlapping patterns are summed to produce the sequence-equivalent contact map, $\bar{F}$, that is equivalent to contact maps observed by sequence-based experimental methods.

facilitating the setup of MD simulation of complex polymer systems (Grünewald et al., 2022).

2.9.1 Generation of the starting coordinates

The protocol for constructing coordinates for the chromosome at Martini resolution starts with interpolating the 10 bp per monomer polymer model generated as previously described (Figure 4, step 1). To this end, a periodic B-spline, $\boldsymbol{m}(s)$, is fitted to the monomer positions, $\{\boldsymbol{x}_i\}$, which represents the chromosome's helical axis (Dierckx, 1996; Virtanen et al., 2020). Along the helical axis, the bp positions, $\{\boldsymbol{m}_j\}$, are sampled such that each segment of the curve between monomer centers contains 10 bp spaced equidistantly. Next, we align bp template coordinates at the Martini level using the resulting bp positions. To properly align the templates, we have to define the internal coordinates $(\hat{\boldsymbol{u}}_j, \hat{\boldsymbol{f}}_j, \hat{\boldsymbol{v}}_j)$ for all the sampled positions (Figure 4, step 2).

In order to construct these internal coordinates, we use a rotation minimizing frame (RMF). An RMF is a reference frame that does not rotate around the instantaneous tangent of the curve $\boldsymbol{m}(s)$, which is defined continuously along any B-spline. The stability of an RMF is ideal for our application since discontinuities in the orientation of consecutive bases will lead to an unrealistic chromosome geometry. The RMF is constructed along the sequence of bp positions, $\{\boldsymbol{m}_j\}$, using the double reflection method outlined by (Wang W. et al., 2008). The paper describes a simple and fast algorithm for approximating our chromosome's RMF with a global error in the order of $\mathcal{O}(h^4)$, where h is the distance between consecutive bps.

To apply the double reflection method and construct the RMF, we first calculate the instantaneous tangent $\hat{\boldsymbol{u}}_j$ on the bp positions using numerical differentiation. To ensure the double reflection method's accuracy, the approximation error of the tangents $\hat{\boldsymbol{u}}_j$ to the true tangent vector, $\boldsymbol{m}'(s)$, must be of the order $\mathcal{O}(h^5)$. Given an arbitrary starting reference vector, the RMF can now be constructed along the entire helical axis.

In order to transform the RMF to the internal coordinates of the chromosome, we must apply two additional transformations to the RMF. Since Syn3A′s chromosome is circular, the additional boundary condition that has to be satisfied is the continuity between the first and last bp's internal coordinates. This condition is realized by applying an additional twist per bp, i.e., a rotation over $\hat{\boldsymbol{u}}_i$, to compensate for a possible discontinuity in the RMF. Additionally, to incorporate the intrinsic helical pitch of B-DNA, an additional twist of 34.3° per bp is applied to each frame (Sinden, 1994). Approximating the DNA's intrinsic structure by a uniformly twisting double helix is justified by the absence of NAPs in Syn3A, resulting in the chromosome not sustaining any significant supercoiling (Gilbert et al., 2021). Finally, using a rigid transformation, templates of the Martini bps are placed on the sampled positions and aligned to the corresponding internal coordinates, building the starting coordinates of the Martini chromosome model (Figure 4, step 3).

2.9.2 Generation of the chromosome topology

The topology at the Martini level comprises the bead-type assignments (i.e., non-bonded interactions), the bonded interactions, and possibly structural biases such as an elastic network. The typical frameworks for generating topology files at the Martini level take an all-atom structure as input (Brooks et al.,

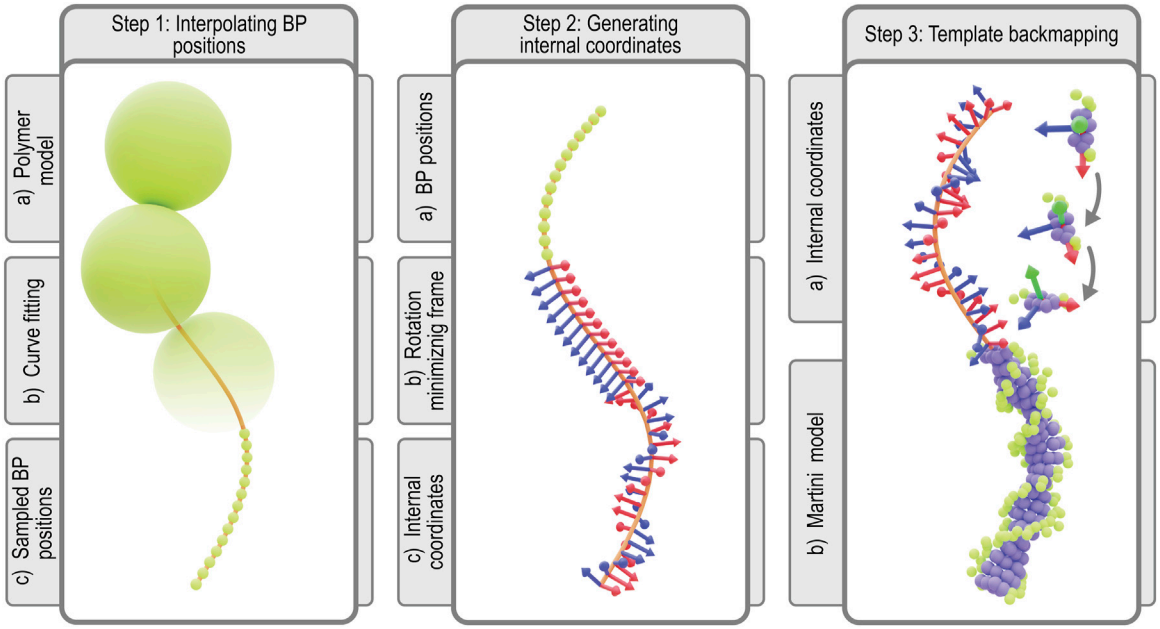


FIGURE 4
Martini backmapping protocol: Schematic depicting the steps in the protocol used to generate coordinates in the Martini representation. By backmapping a dsDNA polymer model, the protocol efficiently creates a near-atomistic model of the entire chromosome. In the final output Martini model at the far-right, each bp is represented by 7 purple beads for the nucleobases (3 per pyrimidine, 4 per purine) and 3 green beads for each backbone (2 per sugar, 1 per phosphate), for a total of 13 Martini beads per bp (Uusitalo et al., 2015).

1983; Liwo et al., 1997; Case et al., 2005; Phillips et al., 2005; de Jong et al., 2013; Machado and Pantano, 2016; Kroon et al., 2022; Abraham et al., 2023). Subsequently, a connectivity graph is generated from the distance matrix and valency-based rules. From this graph, using the all-atom to Martini correspondence defined in the mapping of the nucleobases, the Martini topology is created. This process is called resolution transformation. Using the complete all-atom connectivity graph makes procedures invariant to molecular topology and allows the identification of chemical modifications (e.g., methylation) on the fly. However, the underlying subgraph isomorphism is an NP-complete problem. Thus, while this procedure is very rigorous, it is not very efficient.

Instead, we extended the multiscale graph matching protocol implemented in Polyply to dsDNA. In essence, the protocol performs a resolution transformation from the residue graph to target resolution, in this case, Martini. Utilizing the residue graph gives the needed speed-up to handle polymers of the size of the chromosome. Even though the algorithm still uses a subgraph isomorphism, it is faster since it only works on the residue graph instead of the full molecule graph. Using this algorithm, the molecule topology is generated in two steps: 1) From a set of provided building blocks, all bonded interactions and bead-type assignments are determined for the individual nucleobases (i.e., intra-residue). 2) Bonded interactions, which span multiple residues, are assigned by finding all valid subgraph isomorphisms between graph fragments that describe these inter-residue interactions and the target graph at the residue level. For each match, the bonded interactions are added to the topology. Furthermore, the bead-types are also modified to account for the links between residues where needed. The second strand is generated in the same way by running the algorithm on the complementary single-strand sequence.

The intra- and inter-residue graph fragments, referred to as blocks and links, need to be provided to Polyply as input files. Thus we have extended the Polyply library with data files that describe DNA parameters for Martini2 (Uusitalo et al., 2015). Furthermore, for convenience, Polyply was extended with a parser for .fasta and .ig data files that describe DNA sequences. Most importantly, an automatic recognition of circular DNA is possible when provided with an .ig data file.

Finally, we note that all Martini DNA needs a secondary structure stabilization (i.e., elastic network). Informed by the generated starting coordinates of the Martini chromosome, an elastic network connects nearby beads with harmonic bonds. A simple auxiliary script was used to add the elastic network to the already existing topology generated with Polyply.

2.9.3 Additional structural components

In addition to modeling the intrinsic dynamics of the chromosomal DNA, the polymer model also captures the DNA interacting with the cell membrane and ribosomes. For our Martini chromosome model, these contributions can also explicitly be taken into account with the same near-atomistic resolution. To model the ribosomes, we use a bacterial homolog previously published by (Uusitalo et al., 2017). Initially, we attempt to align the ribosomes with their counterparts in the polymer model. In this step, steric clashes with the chromosome can occur, which we resolve by applying small random rigid body transformations to the ribosomes. The translation length in this transformation acts as a fudge factor, which slowly increases per failed iteration. Lastly, a

realistic cell membrane is constructed using the TS2CG tool, including both a realistic lipid composition and a representative membrane protein density (Pezeshkian et al., 2020).

3 Results

3.1 Diffusion of ribosomes and DNA monomers

The spatial heterogeneity of macromolecules and complexes within the cell and the need for them to encounter one another *via* diffusion strongly contribute to the stochastic nature of gene expression. For example, a RNA polymerase (RNAP) must diffuse to a gene to perform transcription and a mRNA and ribosome must diffuse to one another to perform translation. Some of these reactions can become coupled with one another, such as multiple ribosomes reading the same mRNA (polysomes) or a ribosome reading a nascent mRNA that is still being transcribed from a RNAP (expressomes - coupled transcription and translation) (O'Reilly et al., 2020). These couplings have been observed to varying extents in multiple bacteria. The proportion of ribosomes found in polysomes in *E. coli* has been reported as high as 80% (Bremer and Dennis, 2008), and the proportion in an organism related to Syn3A, *Mycoplasma pneumoniae*, has been reported as 26% (Xue et al., 2022). Expressomes have been observed to a lesser extent, the proportion of ribosomes participating in one only being 3% of ribosomes in *M. pneumoniae* (O'Reilly et al., 2020). Based on cryo-ET we estimated the proportion of ribosomes in polysomes in Syn3A is 25%–40% and from prior simulations we estimate the proportion of ribosomes in close enough proximity to the DNA to form an expressome to be roughly 20% (Gilbert et al., 2021). In the WS-WCM of Syn3A, polysomes were shown to be a critical factor in accurately doubling the proteome over the course of a cell cycle (Thornburg et al., 2022). Before we try to quantify how the effects of these coupled mechanisms affect the spatial organization and diffusion of the chromosome and ribosomes (Mondal et al., 2011), here we quantify how the chromosome and complete, intact ribosomes affect the diffusion of one another at the scale of a whole Syn3A cell.

Simulations were performed on 50 replicate systems of representative Syn3A cells with a radius of 200 nm, each of which contained 500 uniformly distributed ribosomes and a randomly-generated configuration of a single unreplicated chromosome. Following an initial energy minimization of the standard polymer model of the chromosome, bond (U_i^s), bending (U_i^b), and twisting (U_i^a and U_i^t) interactions between all DNA monomers were added/removed from the system for two test cases, which we will refer to as "with bonds" and "without bonds", respectively. We analyzed the diffusion of DNA monomers and ribosomes in two regions of the cell: 1) a central spherical volume extending to 150 nm within which surface effects are assumed to be negligible (Śmigiel et al., 2022) and 2) an outer concentric spherical shell extending from 150 nm to 200 nm. Particles are assigned to these shells using their initial coordinates at $t = 0$. Mean-squared displacements of the DNA monomers and ribosomes were calculated as ensemble averages within each of the regions for each replicate system, these are the transparent time-traces (Figures 5A, B), respectively. Least-squares fits were then used to determine the Brownian diffusion constants, D, and the power-law exponent, α, for the case of anomalous diffusion (Barkai et al., 2012; Oliveira et al., 2019; Muñoz-Gil et al., 2021) for each replicate (Supplementary Material). The ensemble-averaged values across replicates are reported in the legends (Figures 5A, B).

In the absence of bonds, the DNA monomers move following nearly Brownian diffusion. Bonding the monomers causes their motion to become sub-diffusive with $\alpha \approx 0.79$ for both inner and outer regions (Figure 5A). Sub-diffusive motion is an expected result for monomers within polymers, but Rouse dynamics predict $\alpha = 0.5$ for times shorter than the relaxation time (Doi and Edwards, 1988). Our result agrees with theoretical predictions ($\alpha = 0.75$) for short-time segmental motion in stiff worm-like chains with contour lengths much longer than their persistence length (Berg, 1979) and experimental measurements ($\alpha = 0.75$) of large (relative to void) particle diffusion in networks of stiff filaments (Amblard et al., 1996). We repeated similar simulations using systems whose initial conditions were generated without ribosomes to probe the origin of DNA monomers' sub-diffusive behavior in our model. In the scenario without ribosomes the DNA monomers are less sub-diffusive with $\alpha \approx 0.85$ (Supplementary Figure S4), which suggests sub-diffusive motion is a result of the confined chromosome forming a stiff polymer network. Our model's deviation from observed sub-diffusive behavior ($\alpha = 0.4$) of chromosomal loci in *E. coli* (Weber et al., 2010a) is likely a result of neglecting the viscoelastic nature of the bacterial cytoplasm (Weber et al., 2010b). These results for the DNA are observed for both the inner and outer regions.

Ribosomes move sub-diffusively within the inner region of the system without bonds and approach Brownian diffusion in the outer region of the system without bonds, where the DNA density is lower. When bonds are added to the system the ribosomes in the inner region undergo motion closer to Brownian diffusion. Comparing the radial distribution functions (Patrone and Rosch, 2017) of DNA monomers about the ribosomes (Figure 5C; Supplementary Material), we determined that this was a result of the system with bonds creating a polymer mesh with persistent voids (Sorichetti et al., 2020; Xiang et al., 2021) for the ribosomes to diffuse within, in contrast to the case without bonds where the DNA monomers rapidly diffuse and are closely crowded around the ribosomes. It should be noted that the asymptotic approach of the radial distribution functions in the outer shell approaching a value less than one is expected due to the cutoff radius including empty volumes outside the boundaries of the cell. The Brownian diffusion constants of ribosomes in systems with bonds is within the range of experimental measurements in other bacteria (Bakshi et al., 2012; Sanamrad et al., 2014). No significant correlations between the Brownian/anomalous diffusion of the DNA monomers and ribosomes were observed, as can be seen by the covariance ellipsoids and Pearson correlation coefficients reported in the legends (Figure 5C). These were not repeated for the case of chromosomes with loops and topoisomerase due to the non-equilibrium nature of those simulations.

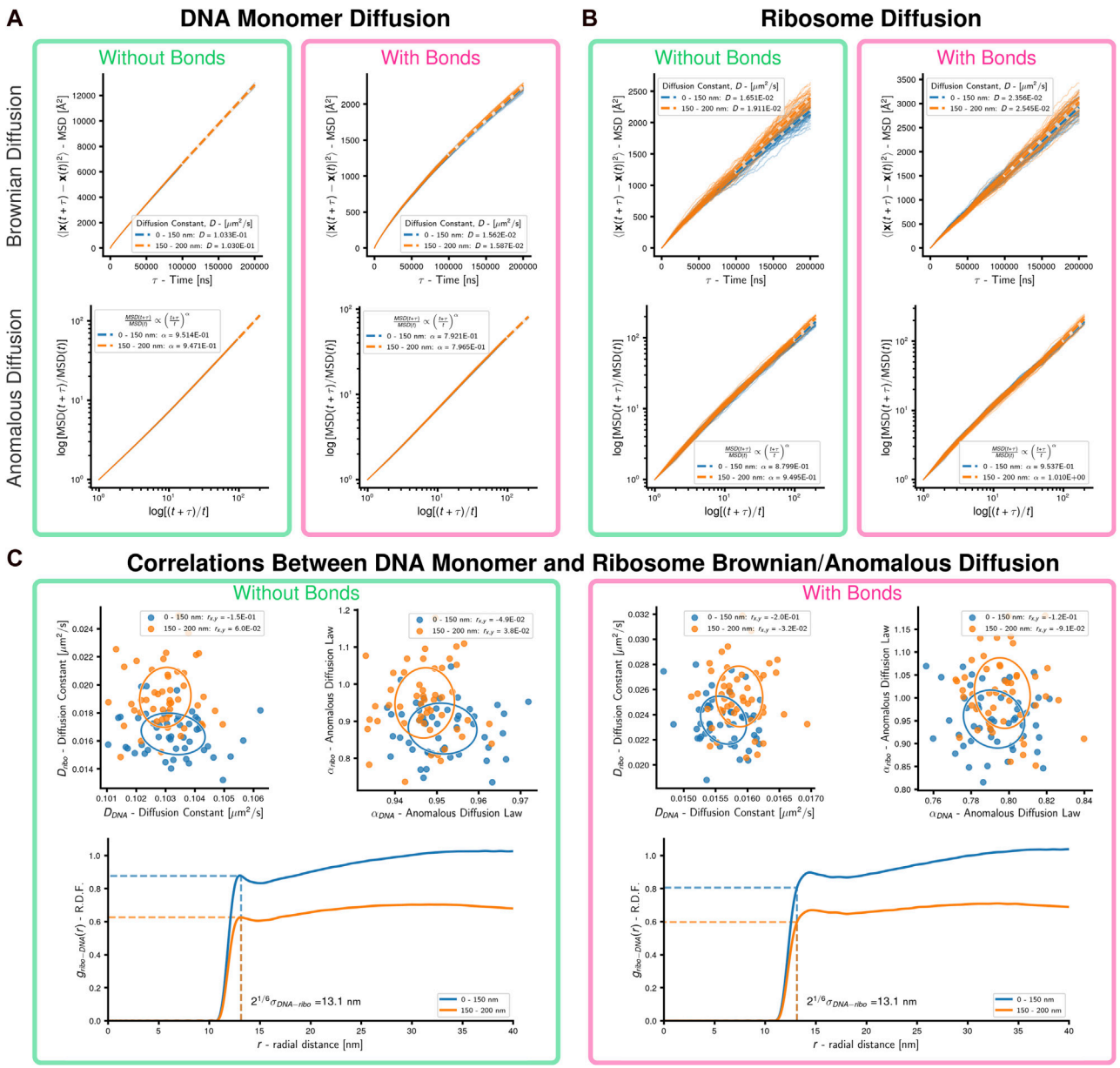


FIGURE 5
Cell-scale diffusion of ribosomes and DNA monomers: 50 replicates of a system with an unreplicated 54,338 monomer chromosome in 200 nm radius cell containing 500 uniformly distributed ribosomes were simulated. **(A)** Brownian and anomalous diffusion of DNA monomers with and without bonds forming DNA polymer. For the two concentric spherical shells, dashed lines are the least-squares fits for the Brownian diffusion constant (linear-linear) and anomalous diffusion power-law (log-log), respectively. **(B)** Brownian and anomalous diffusion of ribosomes with and without bonds forming DNA polymer. For the two concentric spherical shells, dashed lines are the least-squares fits for the Brownian diffusion constant (linear-linear) and anomalous diffusion power-law (log-log), respectively. **(C)** Correlations between DNA and ribosome diffusion with and without bonds forming DNA polymer. Above are scatter plots with covariance ellipses for Brownian and anomalous diffusion, Pearson correlation coefficients are reported in the legends. Below are estimates of the radial distribution function of DNA monomers about ribosomes, the dashed lines indicate the cutoff for WCA interactions. Results are shown for the two concentric spherical shells.

3.2 Chromosome segregation

There is experimental evidence of chromosome segregation during replication (Nielsen et al., 2006), and furthermore, segregation of replicating chromosomes in nontrivial replication states in *E. coli* (Youngren et al., 2014). For the purposes of this study, we separate chromosome segregation into two effects: A) the disentanglement of daughter chromosomes and B) the partitioning of the daughter chromosomes' centers of mass into different regions of the mother cell. Both chromosome disentanglement through the influence of compaction (Goloborodko et al., 2016b) caused by DNA-looping (Marko, 2009; 2011; Goloborodko et al., 2016a; Brahmachari and Marko, 2019) and the partitioning of chromosomes through entropic repulsion of polymer topologies

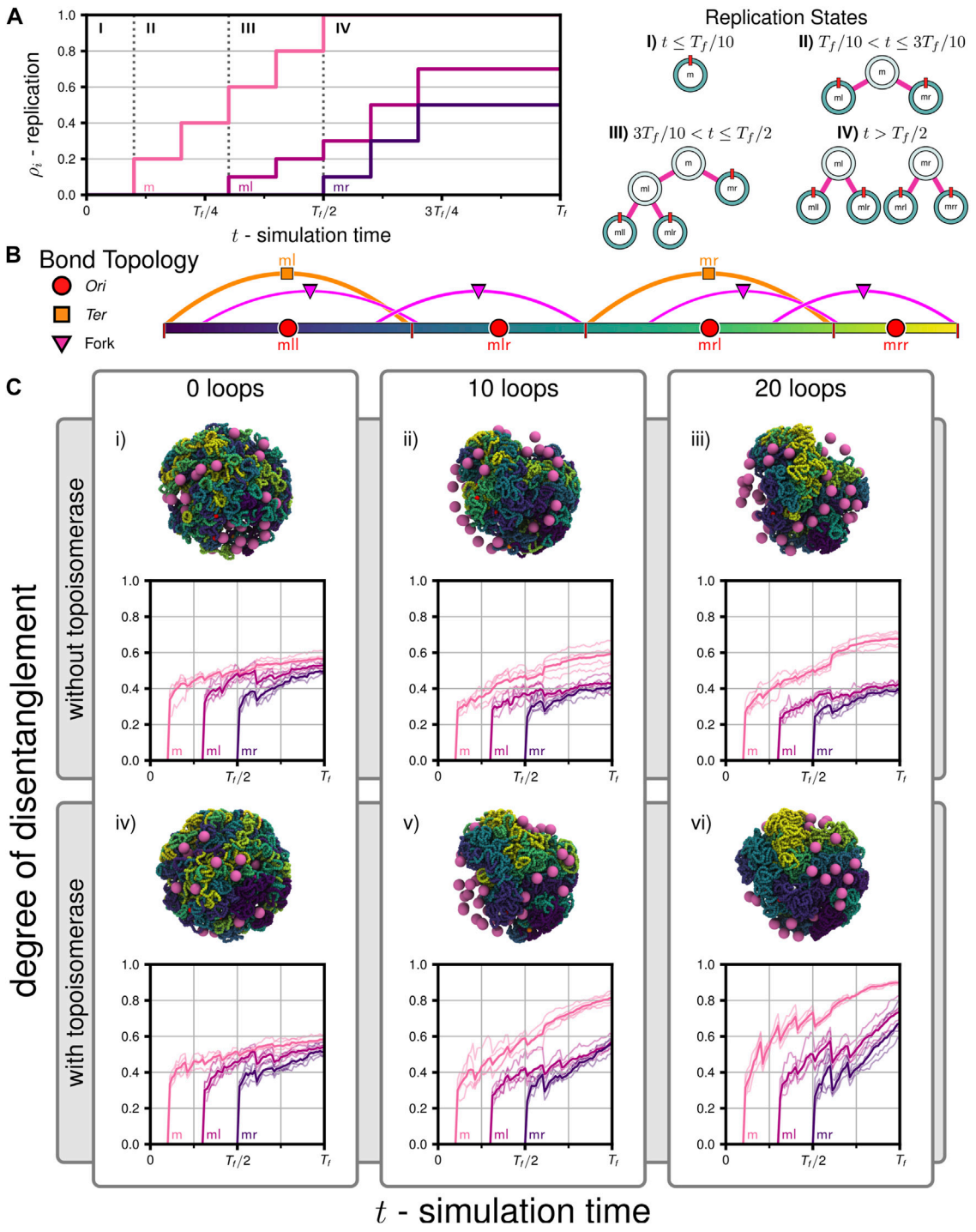


FIGURE 6
Disentanglement of daughter chromosomes during replication: **(A)** Replication progress ($\rho_i = \rho_i^{CW} + \rho_i^{CCW}$) as a function of time for the set of simulations testing the influence of loop extrusion and topoisomerases on disentanglement. The corresponding binary tree representations of the replication states are shown on the right. **(B)** Bond topology of the replicated system at $t = T_f$. **(C)** Mean degree of disentanglement as a function of simulation time for the six cases (i-vi) considered (solid line), five replicate systems were simulated for each case (faint lines). The trace labeled *m* corresponds to the entanglement of *ml* and its descendents (*mll*-purple,*mlr*-blue) with *mr* and its descendents (*mrl*-green,*mrr*-yellow), *ml* corresponds to the entanglement of the replicated region of *ml*, i.e., the regions of *mll* (purple) and *mlr* (blue) connected by forks, and *ml* corresponds to the entanglement of the replicated region of *mr*, i.e., the regions of *mrl* (green) and *mrr* (yellow) connected by forks. Snapshots of the final configurations at $t = T_f$ are shown above each plot, respectively.

within confinements (Jun and Mulder, 2006; Jung and Ha, 2010; Jung et al., 2012; Junier et al., 2013; Wasim et al., 2021; Mitra et al., 2022a) have been previously been studied in computational settings.

We probed chromosome segregation using a toy system approximately one-tenth the volume of a Syn3A cell with similar number densities (90 nm radius, a single unreplicated 50,000 bp chromosome, and 50 ribosomes). We carried out a series of simulations to probe the essential nature of proteins hypothesized to be necessary for simultaneous chromosome segregation during replication. Over the course of the simulations, the 5,000 monomer chromosome was replicated and the *Ori* to *Ter* ratio changed in the following sequence I) 1:1, II) 2:1, III) 3:1, and IV) 4:2 (Figure 6A). The final replication state is that of two fully replicated daughter chromosomes, each of which are themselves in the process of replication (Figure 6B), where the DNA content has more than tripled to 16,000 monomers (160,000 bp). The number of loops present in the systems were varied between 0, 10, and 20, and these systems were then simulated with and without the action of topoisomerases, for a total of six cases (i-vi in Figure 6C). Five independently generated initial conditions were used to prepare five replicate simulations per case, for a total of thirty simulations. Each simulation was run until the final time of T_f = 2.0E+7 timesteps using the looping and topoisomerase algorithm and parameters described in Section 2.4, which corresponds to 2,000 extrusion events for each loop present in the system. At every timestep, we used the binary tree model to group monomers into left/right daughters and their descendants, each with N_l and N_r monomers, respectively, and used those groupings to analyze the disentanglement and partitioning of the daughter chromosomes about each set of replication forks (*m*,*ml*,*mr*). We have completed an equivalent proof-of-concept simulation on the full system with 54,338 monomers in a 200 nm cell containing 500 ribosomes (Supplementary Video SV2).

3.2.1 Disentanglement of daughter chromosomes

We calculated a metric describing the relative number of contacts between different daugther chromosomes, which we will refer to as the degree of disentanglement, as a function of simulation time (Figure 6C) for all six cases. First, we note that for all cases the degree of disentanglement exhibits abrupt decreases when portions of the chromosome are replicated, i.e., each abrupt decrease is a result of the step-wise increases in the replication state (Figure 6A). This result was anticipated because daughter chromosomes are in close spatial proximity as they are generated using the train-track model (Figure 2B) and is consistent with experimental observations of daughter(/sister) chromosome cohesion due to precatenanes in the wake of the replication fork (Wang X. et al., 2008; Cebrián et al., 2015). This effect would be less-pronounced if a smaller fraction of the genome was replicated in each step. We find that both topoisomerase and loop-extruding SMC protein complexes are necessary for daughter chromosomes to be disentangled as replication occurs. In cases i-iii without topoisomerases, topological constraints cannot be resolved and the system remains entangled (Figure 6C). Interestingly, while adding loops in cases ii and iii assists in disentangling *ml* and *mr* about fork *m*, the presence of loops increases the entanglements of *mll* with *mlr* about fork *ml* and *mrl* with *mrr* about fork *mr*, respectively (Figure 6C). Within our model, looping in the absence of topoisomerases is deletrious for subsequent rounds of replication because enhanced compaction increases the likelihood that topological constraints are introduced during replication. However, including solely topoisomerase in case iv is not effective at disentangling the chromosome (Figure 6C). We hypothesize that this is because diffusive motion is insufficient to cross strands when the soft potential emulating topoisomerases in our model is active and that loop-extrusion assists to isolate possible strand-crossings before completing the crossings in subsequent extrusion steps to resolve topological constraints. In cases v and vi, we find the greatest degrees of disentanglement (Figure 6C). When comparing the disentanglement of *ml* and *mr* about fork *m* between cases ii-iii and v-vi, we find that a plateau is reached in cases ii-iii when the topological constraints cannot be resolved (Figure 6C). In summary, we find that systems require both topoisomerase and loops to simultaneously disentangle all daughter chromosomes as they are being replicated. Furthermore, increasing the number of loops increases the rate of disentaglement, as seen in case vi *versus* v. The trends quantified by the degree of disentanglement can also be qualitatively observed in the snapshots of the final configurations at $t = T_f$ (Figure 6C). The degree of disentanglement was calculated for the proof-of-concept simulation of the full chromosome (Supplementary Video SV2) and shows the same behavior as the cases (v and vi) with both SMC and topoisomerase (Supplementary Figure S8).

3.2.2 Partitioning of daughter chromosomes

We calculated the Euclidean distance separating the daughters' centers of mass relative to an ideal partitioning, $L_{\text{partition}}(N_l, N_r, R_{\text{sphere}})$, to assess the extent to which the daughter chromosomes had been partitioned to different volumes within the cytoplasmic space (Figure 7). If the daughters and their descendants have an equal number of monomers ($N_l = N_r$), ideal partitioning would correspond to them occupying identical hemispherical volumes (Supplementary Figure S2). The daughters' centers of mass would then be found at the centroids of the hemispheres and separated by $3R_{\text{sphere}}/4$. The functional dependence of the ideal partitioning on N_l and N_r accounts for possible asymmetries in nontrivial replication states, such as states III and IV (Figure 6A). Similar to the results of the degree of disentanglement (Figure 6C), we find that partitioning was the most complete in case vi with topoisomerase and the greatest number of loops (Figure 7). However, over the timescales simulated, the distance separating the daughters' centers of mass is still relatively insignificant as compared to the size of the confining volume. This can be observed qualitatively in the manner in which the compacted globules of the disentangled daughters are folded around one another (Figure 6C). Based on this, we conclude that disentanglement is necessary for partitioning to occur, and due to the necessity of topoisomerase and loops for disentanglement, successful partitioning is also dependent on topoisomerase and loops. However, absent a regulatory system introducing a spatial heterogeneity or active force, the partitioning in our model proceeds over a much longer time-scale than the disentanglement. This can be seen in case vi, where the degree of disentanglement about fork *m* is reaching a plateau near one, indicating that all that remains is an interface between the now disentangled daughters (Figure 6C), while the extent of partitioning has yet to reach half of the ideal distance, $L_{\text{partition}}$ (Figure 7).

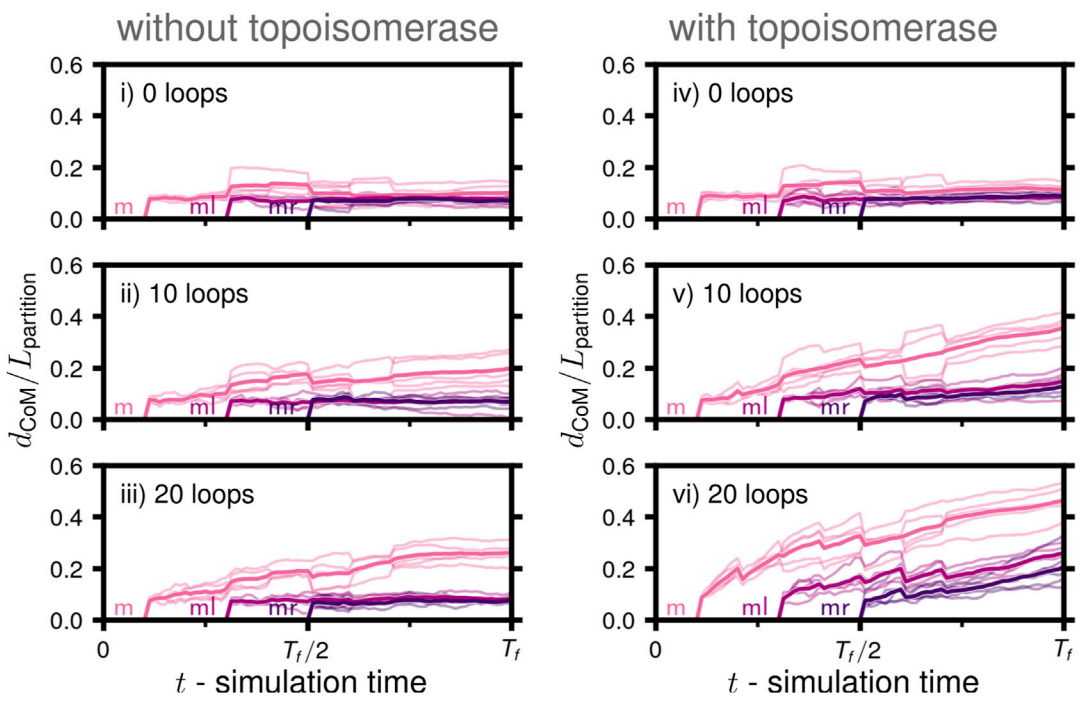


FIGURE 7
Partitioning of daughter chromosomes during replication: Mean separation of daughters' centers of mass (d_{CoM}) relative to the length-scale of ideal partitioning, $L_{partition}(N_l, N_r, R_{sphere})$, in a spherical volume of daughters with N_l and N_r monomers as a function of simulation time for the six cases (i-vi) considered (solid line), five replicate systems were simulated for each case (faint lines). The trace labeled *m* corresponds to the separation of *ml* and its descendents (*mll*,*mlr*) with *mr* and its descendents (*mrl*,*mrr*), *ml* corresponds to the separation of the replicated region of *ml*, i.e., the regions of *mll* and *mlr* connected by forks, and *ml* corresponds to the separation of the replicated region of *mr*, i.e., the regions of *mrl* and *mrr* connected by forks (Figure 6B).

3.2.3 Contact maps between daughter chromosomes

Chromosome segregation was also investigated using chromosome contact maps of the same replicating chromosome systems. Contact maps were calculated at 250 bp resolution using the configurations from $3T_f/4 \le t \le T_f$ (*i.e.*, when the replication state is constant) averaged over the five replicates for each case. We will denote the true contact maps for cases iii (Figure 8A) and vi (Figure 8B) as A and B and the sequence-equivalent maps as $\tilde{\mathsf{A}}$ and $\tilde{\mathsf{B}}$, respectively. For both cases we can observe inter-daughter interactions indicated by the increased contact frequency within the off-diagonal regions of the true maps. The inter-daughter contacts are enriched in the case iii, where the system lacks topoisomerase, particlularly between the *Ter*s of *mll* (unreplicated region of *ml*) and *mrl* (unreplicated region of *mr*), which is consistent with our findings when using the degree of disentanglement (Figure 6C) and partitioning (Figure 7), and agree with experimental observations of topo-IV modulating daughter(/sister) cohesion (Lesterlin et al., 2012; Conin et al., 2022). Additionally, we can calculate sequence-equivalent maps to determine how these inter-daughter interactions would be represented in an experimental contact map generated from a 3C library of cells in this replication state, and under these topoisomerase conditions. The sequence equivalent maps, $\tilde{\mathsf{A}}$ and $\tilde{\mathsf{B}}$, retain the characteristic primary diagonal and peaks at opposite corners indicative of circular chromosomes (inset Figure 8C). The effect of inter-daughter interactions are analyzed by comparing the average rates of loci self-interactions between the true and sequence-equivalent maps. The average loci self-interactions in true maps A and B are

$$\frac{\sum_i \mathsf{A}_{ii}}{N} = 0.083 \quad \text{and} \quad \frac{\sum_i \mathsf{B}_{ii}}{N} = 0.091, \tag{3.1}$$

respectively. The average loci self-interactions in sequence-equivalent maps $\tilde{\mathsf{A}}$ and $\tilde{\mathsf{B}}$ are

$$\frac{\sum_i \tilde{\mathsf{A}}_{ii}}{\tilde{N}} = 0.146 \quad \text{and} \quad \frac{\sum_i \tilde{\mathsf{B}}_{ii}}{\tilde{N}} = 0.121, \tag{3.2}$$

respectively.

Confoundingly, while one might anticipate a higher rate of loci self-interactions in $\tilde{\mathsf{B}}$ relative to $\tilde{\mathsf{A}}$ given the loci self-interactions in the true maps, the opposite case is true due to contributions from the inter-daughter interactions (Figure 8C), which are the result of precatenanes in the replicated daughters (Wang X. et al., 2008; Cebrián et al., 2015). This simple example is illustrative of how experimental contact maps not only encode an ensemble of chromosomes with different configurational states (Junier et al., 2015; Sekelja et al., 2016), but also encode an ensemble extending across an extra set of dimensions corresponding to the space of replication states. Sequence-equivalent maps have the further benefit of allowing one to observe changes in chromosome organization as a system follows a trajectory in configurational and replication state space. Using the simulations of case vi (Figure 6C), contact maps

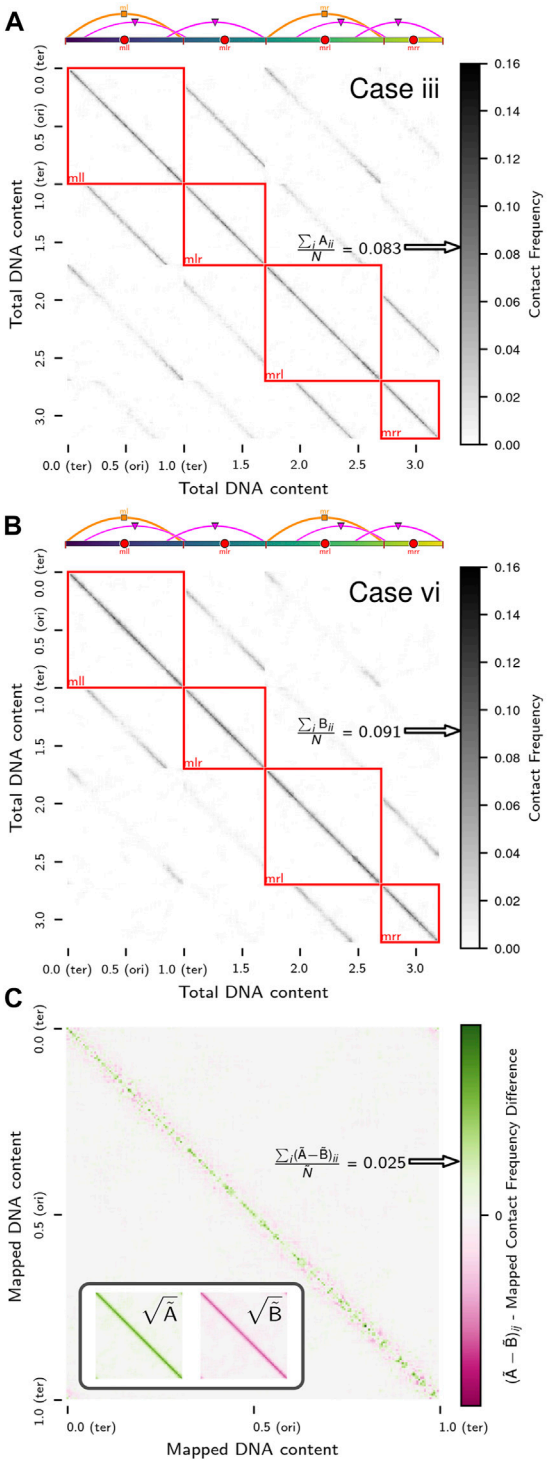


FIGURE 8
Contact maps of replicating chromosomes: **(A)** True contact map, A, for case iii (20 loops, without topoisomerase) of replicating chromosome system in Figure 6C. **(B)** True contact map, B, for case vi (20 loops, with topoisomerase) of replicating chromosome system in Figure 6C. **(C)** Difference in sequence-equivalent contact maps $\tilde{A}$ and $\tilde{B}$, which are created by mapping A and B, respectively, using the procedure illustrated in Figure 3B. Inset are the sequence-equivalent maps, displayed after taking an element-wise square-root to enhance visual clarity. Within A-C, the arrows along the colorbar indicate the average value of the diagonal elements representing loci self-interactions.

were calculated for ten time intervals of equal length (Supplementary Figures S6, S7). We see that the approach to a plateau in the degree of disentanglement (Figure 6C), which indicates the system is approaching a decatenaned state, is reflected in reduced differences in the sequence-equivalent contact maps (Supplementary Figures S6, S7).

3.3 Martini model

Using our new backmapping protocol, a Martini model of the Syn3A′s chromosome is constructed (Figure 9). With the aim of performing a molecular dynamics (MD) simulation, both starting configuration and topology are generated based on the previously described polymer model and the genome's sequence. The resulting Martini model contains around 7 million Martini beads, representing the 34 million atoms constituting the chromosome.

The chromosome model is energy minimized in vacuum using Gromacs-2023 (Abraham et al., 2023). However, running an MD simulation, additionally requires the solvation and charge neutralization of the model. This step dramatically increases the number of particles in the simulation to over 500 million Martini beads. At the current stage, Gromacs can not handle systems of this size, which restrains us from further exploring the dynamics of the system.

However, to illustrate our DNA backmapping protocol, we model and simulate the previously described toy chromosome system of approximately one-tenth the size of the Syn3A. Before applying our chromosome modeling protocol to this toy model, we first sample an artificial 50 kbp sequence with the same relative nucleobase frequency as the Syn3A genome. The resulting Martini model is solvated in a 185 nm cubic box, neutralized, and subsequently, a physiological salt concentration of 0.15 M NaCl is added to the system. To incorporate the confinement effect of the membrane on the chromosome, an additional spherical boundary potential with a radius of 90 nm is added to the model. Note that in the Martini version of the toy system, we omitted to model the ribosomes.

The final simulation consists of approximately 50 million Martini beads, representing over 500 million atoms (Figure 10A). First, we energy minimize and equilibrate the system before starting the production simulation, which is stable at a 20 fs timestep. In total, the system is simulated for 50 ns We note that on this short timescale, the chromosome will not fully equilibrate. Nevertheless, we have the ability to confirm that our backmapped model is consistent with the intended structure and observed sub-diffusive motion ($\alpha \approx 0.87$) of 10 bp segments of the Martini dsDNA (Supplementary Figure S5) that is consistent with the Brownian dynamics simulations of the full chromosome in the absence of ribosomes (Supplementary Figure S4).

A direct comparison between the polymer and Martini simulations is possible by analyzing the models' persistence lengths, l_p. For the Martini simulation, we determine the persistence length of the chromosomal DNA by calculating the orientational correlation of the bond vectors, $\hat{\boldsymbol{u}}_i$, connecting the centers of consecutive bps. In an idealized worm-like chain approximation, we expect the bond vectors to decorrelate exponentially along the chain, $\langle \hat{\boldsymbol{u}}_i \cdot \hat{\boldsymbol{u}}_{i+j} \rangle = e^{-j\langle l_0 \rangle / l_p}$, where $\langle l_0 \rangle$ is the mean distance between bps.

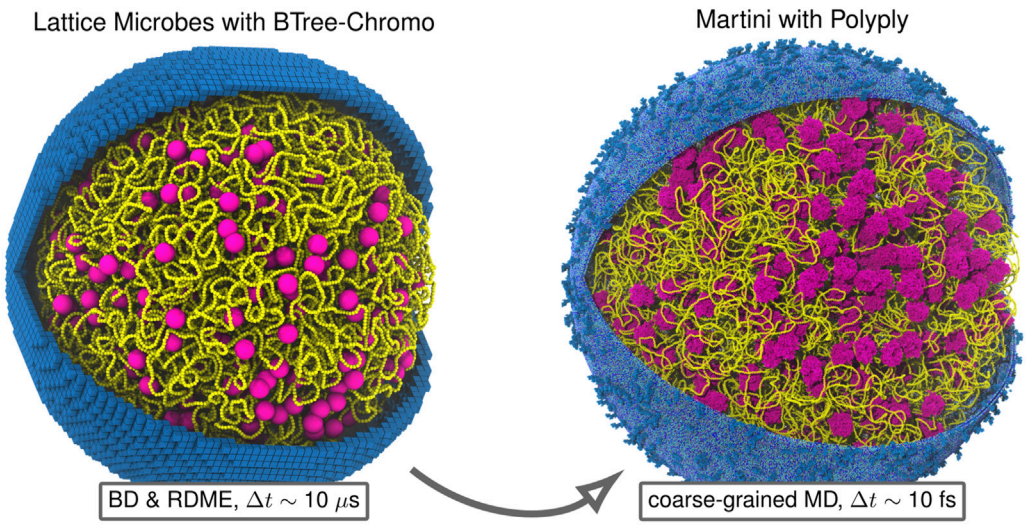


FIGURE 9
Backmapping Martini model of entire Syn3A cell: Example backmapping of polymer model of 200 nm radius Syn3A cell with a single unreplicated chromosome to near-atomistic resolution Martini representation using Polyply. For both representations we show the chromosome (yellow), ribosomes (magenta), and membrane (blue). The membrane in the Lattice Microbes representation is shown using the 8 nm cubic subvolumes used for reaction-diffusion master equation (RDME) simulations and the membrane in the Martini representation, which includes the lipid composition and membrane proteins of Syn3A (Thornburg et al., 2022), was generated using TS2CG (Pezeshkian et al., 2020). The two representations are complementary in that the combined polymer-RDME model resolves cell-wide chemical transformations over timescales comparable to the cell-cycle by neglecting detailed physical interactions among particles, while the Martini model alternatively resolves these detailed physical interactions among macromolecules over shorter timescales.

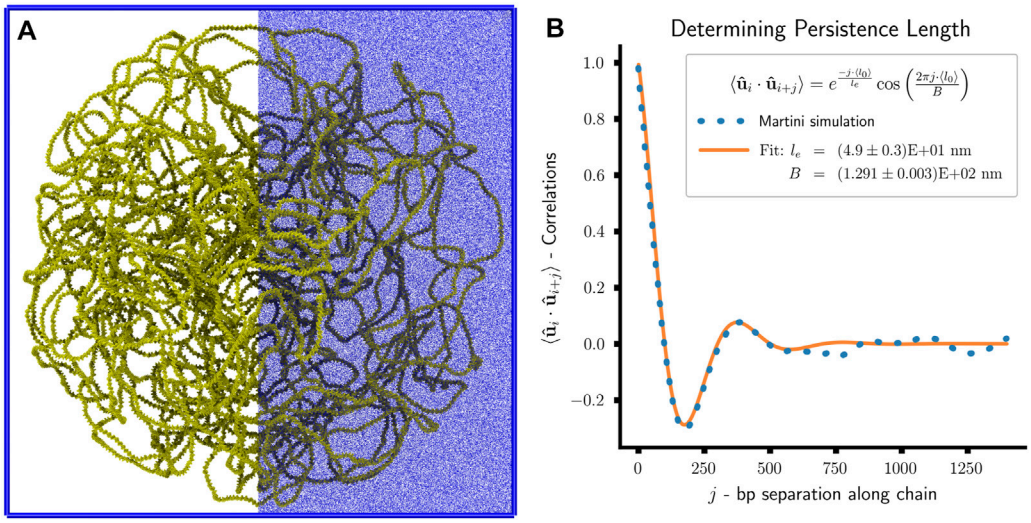


FIGURE 10
Martini simulation of toy system: **(A)** Snapshot of Martini simulation of toy system. The system consists of approximately 50 million Martini beads—chromosome 650,000 (yellow), water 50,528,240 (not shown), chloride ions 571,949 (blue), and sodium ions 671,949 (blue). The ions are only displayed on the right-half to enhance visual clarity. **(B)** Plot of bond vector correlations as a function of bp separation along the polymer chain and the least-squares fit of the effective persistence length, l_e, and confinement length scale, B.

However, calculating the bond vector correlations for the last 25 ns of the Martini simulation (Figure 10B) reveal a clear deviation from this idealized model. An additional oscillatory contribution is observed in the decay of the bond vector correlations, which can be attributed to the geometric confinement of the chromosome by the cell wall (Liu and Chakraborty, 2008; Cifra and Bleha, 2010; Castro-Villarreal and Ramírez, 2021). The resulting decay trend is well-captured by

$$\langle \hat{\boldsymbol{u}}_i \cdot \hat{\boldsymbol{u}}_{i+j} \rangle = e^{\frac{-j\langle l_0 \rangle}{l_e}} \cdot \cos\left(\frac{2\pi j \langle l_0 \rangle}{B}\right), \tag{3.3}$$

where l_e is the effective persistence length of the DNA, and B is a length scale related to the confinement size (Liu and Chakraborty, 2008). By performing a least-squares fit of the model to our simulation, we find l_e = (4.9 ± 0.3)E+01 nm and B = (1.291 ± 0.003)E+02 nm. Considering the 45 nm persistence length of the polymer model, which is a chosen model parameter, we observe a qualitative agreement between the two models. The quantitative deviation can be attributed to the confinement reducing the chromosome's conformational space and increasing its effective rigidity. In general, the measured l_e will be greater than or equal to l_p under confinement. However, the small amplitude of the fluctuations in the measured bond vector correlations indicates a moderate confinement regime, suggesting that l_e and l_p are comparable (Liu and Chakraborty, 2008).

4 Discussion

4.1 Study overview and methods

We developed a computational framework to investigate the minimal required components for chromosome replication and segregation in a genetically minimal bacterium, Syn3A. This framework is built around six major components: 1) a method to fold chromosomes around ribosome distributions originating from cryo-ET or other experimental measurements (Supplementary Figure S1), 2) an implementation of a 10 bp per monomer polymer model of dsDNA that includes its intrinsic mechanical properties (bending and twisting stiffness) and can be simulated using Brownian dynamics (Figures 1A–C), 3) algorithms that emulate the effect of known essential proteins that manipulate the chromosome — DNA-looping SMC complexes and strand-crossing type-II topoisomerases (Figure 1D; Supplementary Algorithms S1, S2), 4) a binary tree model to systematically describe nontrivial replication states and create accompanying 3D physical structures obeying the polymer model (Figure 2), 5) *in silico* chromosome contact maps of replicating chromosomes that capture intra- and inter-daughter interactions (Figure 3), and 6) a procedure mapping the chromosome to equivalent higher-resolution Martini whole-cell models using Polyply (Figure 9).

4.2 Key findings

Using the binary tree model of replication states, we have created a means to systematically describe nontrivial replication states that are known to be present in bacteria (Cooper and Helmstetter, 1968; Bremer and Dennis, 2008; Youngren et al., 2014). Previous simulations of replicating chromosomes have used either a set of fixed replication states (Wasim et al., 2021; 2023; Mitra et al., 2022b) or a pre-defined replication protocol (Mitra et al., 2022a). Our software implementation of this model enables users to create physical models of these states with the bond topology of nested theta structures (Figure 2A) and modify the states using computational equivalents of biological processes (Figure 2B). Furthermore, the aspects of the program used to create, manipulate (replicate asymmetrically at specific forks, replicate under well-stirred assumption), query (export bond topology, loci for true and sequence-equivalent maps, counts of genomic regions, etc.), and save replication states may be used independently from simulations of a physical model, which allows other researchers to use the program as a tool.

By combining the binary tree model with the Brownian dynamics model of the chromosomal dsDNA, we have developed a method to generate physics-based models of replicating chromosomes at 10 bp resolution, and simulate their time-evolution while undergoing diffusive motion and non-equilibrium replication events. Cryo-ET of Syn3A demonstrated that the ribosome distribution is near-uniform and the cytoplasm appears denser than other bacteria (Gilbert et al., 2021) and the chromosome itself, through excluded volume interactions with other macromolecular complexes (Dersch et al., 2022) and spatially localized transcription (Llopis et al., 2010), potentially represents the greatest influence on spatially heterogeneous reaction-diffusion processes within simulations of Syn3A (Thornburg et al., 2022).

After folding chromosomes organized as a fractal globule around ribosomes positions from cryo-ET (Gilbert et al., 2021), we measured the diffusion of complete 70S bacterial ribosomes. We find that configurations of the chromosomes create polymer meshworks that have voids containing ribosomes. Within these voids the ribosomes undergo nearly Brownian motion with diffusion constants lower than those observed in *E. coli* (Bakshi et al., 2012). We find that non-specific DNA-looping in the absence of a parABS system compacts the chromosome, with the assumed number of loops based on proteomics of SMC-scpAB components (Table 1) reducing the radius of gyration of 100-monomer segments by approximately 35% (Figure 1E). Although this compaction is substantial, the chromosome can be still be replicated using our implementation of the train-track model without issues.

In the context of our model, we find that both DNA-looping and strand-crossings are necessary for the segregation of daughter chromosomes during and after replication, which is in agreement with Syn3A′s gene essentiality data for SMC-complexes and type-II topoisomerases from transposon mutagenesis experiments (Breuer et al., 2019). We analyzed the time-course of chromosome segregation in a toy system by dividing it into two processes, disentanglement of the daughter chromosomes (Figure 6) and partitioning of the daughter chromosomes into distinct volumes (Figure 7). The system cannot be disentangled when no loops are present. Increasing the number of loops leads to disentanglement of the first generation of daughters, but that process will stall if topoisomerase is absent and the topological restraints cannot be resolved (Figure 6C), which is in agreement with experiments (Wang X. et al., 2008). Additionally, if there are loops and no topoisomerase, subsequent generations will be even more entangled due to replication occurring in the daughters already compacted by loops (Figure 6C). This coordinated role between SMC complexes and topo-IV has been observed in *E. coli* (Zawadzki et al., 2015; Nolivos et al., 2016; Mäkelä and Sherratt, 2020). Identical behavior is observed in the partitioning of the daughters (Figure 7), but the partitioning occurs over a slower

timescale than the disentanglement, with the partitioning less than 50% complete on average in case vi, where the daughters are almost completely disentangled. Based on this, successful disentanglement is necessary in our model for partitioning to proceed. It is qualitatively clear that partitioning lags behind disentanglement in the proof-of-concept simulation of the full chromosome undergoing simultaneous replication and segregation (Supplementary Video SV2), but we are encouraged by the preliminary result for the degree of disentanglement demonstrating that SMC complexes and topo-IV are sufficient at the chromosome-scale (Supplementary Figure S8).

Overall, these findings regarding the influence of SMC complexes and topoisomerases on chromosome segregation are consistent with computational studies of eukaryotic sister chromatids (Goloborodko et al., 2016a) and show that the same mechanisms are capable of segregating nested theta structures in bacteria. While we model the chromosome as a homopolymer rather than a heteropolymer, the energy landscape picture of proteins within a funnel (Bryngelson et al., 1995; Onuchic et al., 1997) is relevant when interpreting the process of chromosome segregation. The ATP-consuming process of loop-extrusion isolates knots and causes the system to approach energetic barriers representing these topological restraints within the system. Our model's periodic action of topoisomerases then lowers the barriers and loop-extrusion drives the system over the lowered barriers. We found that neither of these effects is sufficient in isolation, and the combination of ATP-consuming driving forces and lowered barriers enable the departure from a local energy minimum with a more-knotted topology into a new energy minimum with a less-knotted topology, which is consistent with previous computational studies on knotted chromosome topologies (Racko et al., 2018; Orlandini et al., 2019). These processes are akin to the role of protein-folding chaperones in resolving kinetically trapped misfolded proteins in a rugged energy landscape (Todd et al., 1996; Thirumalai et al., 2019).

Previous studies have calculated *in silico* chromosome contact maps of replicating bacterial chromosomes (Wasim et al., 2021; 2023), but to the best of our knowledge, did not include inter-daughter contacts. Using our model, we have created a procedure to calculate true maps that include inter-daughter contacts and convert those maps extending over the full DNA content of the replicating chromosome system back to the sequence-equivalent maps that would be measured by experimental 3C methods (Figure 3). This not only elucidates variations in the sequence-equivalent maps due to differing spatial organization of chromosomes in identical replication states (Figure 8), but also enables the comparison of maps originating from chromosomes in different replication states and the creation of maps representing a mixture of replication states. Features in contact maps that are attributed to processes during replication and chromosome segregation have been previously reported in synchronized *Caulobacter crescentus* cells (Le et al., 2013) and *E. coli* topo-IV knockout studies (Conin et al., 2022).

Using Polyply, we showed that we can obtain a starting structure of the entire Syn3A chromosome at near-atomic resolution, ready for subsequent sampling of its configuration space using molecular dynamics. Previous dynamics simulations of entire chromosomes are either based on simplified (1-2 bead per bp) models or are restricted to simulating smaller, viral genomes and nanostructures (Maffeo and Aksimentiev, 2020; Sengar et al., 2021).

4.3 Limitations

There is no sequence-specificity in the homopolymer model of replicating chromosomes beyond specific landmark monomers such as *Ori*s and *Ter*s, and there is no means to represent ssDNA. This limitation precludes us from modeling the unique molecular structures of the bubble during replication initation (Shimizu et al., 2016) and replisome during replication (Maffeo et al., 2022). The essentiality of HU in Syn3A despite its reduced proteomics count and high-affinity for structurally deformed DNA (Kamashev, 2000) suggests a role in DNA replication, which is further supported by the *Ori*:*Ter* ratio of *B. subtilis* being reduced upon HU deletion (Karaboja and Wang, 2022). However, in contrast to *E. coli* where HU/IHF has a well-defined role of stabilizing bent dsDNA in DnaA-based replication at an *oriC* (Yoshida et al., 2023), there is a lack of clarity regarding HU's role in Syn3A′s more minimalistic *oriC* (Richardson et al., 2019; Thornburg et al., 2019). In a similar vein, although the binary tree model fully describes topologies of nontrivial replication states that may be undergoing asymmetric replication, the absence of ssDNA prevents us from making the distinction between leading and lagging strands, which would be at the extreme end in opposite directions (clockwise vs. counter-clockwise) on the left and right daughters.

In the chromosome-scale polymer model, we neglected hydrodynamic interactions and did not directly include electrostatics beyond the parameterization of the persistence length, we feel the ability to backmap the system to a Martini representation with near-atomistic detail helps resolve this deficiency by providing information about the effect of neglecting those interactions. In particular, to address the viscoelastic nature of the medium, which was neglected in the Brownian dynamics model, one could simulate the polymer model using dissipative particle dynamics (DPD) (Español and Warren, 2017), where the memory function encoding non-Markovian dynamics due to the medium is constructed (Klippenstein et al., 2021) from whole-cell Martini simulations (Stevens et al., 2023). The Brownian dynamics timesteps (Δt = 0.1 ns) are much smaller than the timescales of loop-extrusion events (~1 s) (Ryu et al., 2021), and vastly smaller than Syn3A′s cell-cycle (~6600 s) (Breuer et al., 2019; Thornburg et al., 2022). To circumvent this, we used energy minimizations to relax the chromosome after non-equilibrium loop-extrusion steps, which helped to accelerate the simulations. However, this came at the cost of disconnecting the Brownian dynamics simulation time from the biological time of the loop-extrusion events. The current implementation of the code calls LAMMPS (Thompson et al., 2022) to run the Brownian dynamic simulations using multiple CPU-threads with OpenMP. Although this approach was sufficiently fast, in the course of the study it has become clear that moving the simulations to the GPU would offer a significant improvement.

4.4 Future directions

Now that we have created a computational model of Syn3A′s chromosome that includes replication and segregation of nontrivial replication states, we intend to integrate it with the 4D-WCM of Syn3A (Thornburg et al., 2022) to extend its predictive capabilities to the full cell-cycle. Information concerning the spatial coordinates of the replicating genome will be sent to the 4D-WCM and information regarding reaction events will be returned, similar approaches have been used by other researchers (Popov et al., 2016). Two immediate applications are the modeling of DnaA filamentation leading to formation of the replication bubble and the dynamic formation of polysomes based on translational activity. Given the absence of regulatory elements in Syn3A, an open-question is if the arrangement of genes can serve as a means of regulation (Chatterjee et al., 2021; Geng et al., 2022) as a result of the mechanochemical coupling between transcription and supercoiling (Chong et al., 2014; Kim et al., 2019). Following transcription events in the 4D-WCM, dynamically applying torsional strain to the chromosome model would enable local configurational changes in genes, thereby modulating their transcriptional propensity.

While the methods described in this study enable us to calculate *in silico* chromosome contact maps whose *Ori*:*Ter* ratio matches experimental qPCR measurements of 3.4 (Thornburg et al., 2022) by using a mixture of replication states with different ratios, there is a lack of clarity about relative weights of these states. Furthermore, there are a vast multitude of compatible replication microstates for each *Ori*:*Ter* ratio. Given that we now have a means to generate sequence-equivalent *in silico* contact maps of chromosomes in different replication states, this motivates the development of a protocol to deconvolve experimental maps generated from populations of unsynchronized cells (Junier et al., 2015; Sefer et al., 2016; Carstens et al., 2020; Zhou et al., 2021; Rowland et al., 2022) to determine the subpopulations of cells in different replication states. The respective replication states would then be found by an inversion of subpopulation contact maps (Supplementary Figures S6, S7) from their sequence-equivalent form to the true contact maps (Figure 3). We note that this proposed methodology faces two challenges: 1) the solution requires knowledge of sequence-equivalent contact maps for replication microstates and 2) even with that information, the problem likely remains underdetermined if only subject to the example set of constraints (Section 2.5) and not more informative constraints such as DNA abundance distributions (Bhat et al., 2022). Assuming further performance improvements of the simulation software, this study helps to address the first issue, but the second issue will need to be resolved.

All simulations of chromosome segregation in this study used a spherical confinement reflecting the observed morphology of SynX-series (Gilbert et al., 2021; Pelletier et al., 2021) and *M. mycoides* (Rideau et al., 2022) cells. Varying confinement over the cell-cycle will allow us to test entropic segregation in shapes with long aspect-ratios (Jun and Mulder, 2006; Jung and Ha, 2010; Jung et al., 2012; Youngren et al., 2014).

Simulations with the Martini model are limited in the description of DNA strand hybridization. To keep the strands together, an elastic network is used. Ongoing efforts are directed to include additional (virtual) bead types that provide a more accurate description of the directed hydrogen bonds that give rise to specific base pairing. Another challenge is to capture the replicating chromosome when creating whole-cell Martini models of different stages of the cell cycle. To this end, a Martini model of a complete replisome has to be constructed and integrated into our chromosome modeling protocol. As part of our DNA backmapping algorithm, we plan to support the incorporation of protein-DNA complexes, thereby facilitating the construction of complete replication forks.

Data availability statement

All software used for simulations and analysis in this study is open-source and listed in Supplementary Table S1. Software used for visualization is publicly available and listed in Supplementary Table S1.

Author contributions

BG: conceptualization, methodology, software, formal analysis, writing—original draft. ZT: methodology, validation, visualization. TB: methodology, validation. JAS: methodology, software, formal analysis, writing—original draft. FG: methodology, software. JES: visualization. SM: conceptualization, resources, writing—review and editing, funding acquisition. ZL-S: conceptualization, resources, writing—review and editing, funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

BG, ZT, TB, and ZL-S: we acknowledge partial support from NSF MCB 1818344 and 2221237, and “The Physics of Living Systems Student Research Network” NSF PHY 2014027. JAS, FG, and SM: we acknowledge funding from the ERC with the Advanced grant 101053661 (“COMP-O-CELL”) and from NWO through the NWA grant “The limits to growth: The challenge to dissipate energy” and BaSyc (“Building a Synthetic Cell”) consortium. JES: Visual Molecular Dynamics (VMD) was developed by the NIH Center for Macromolecular Modeling and Bioinformatics at the Beckman Institute at UIUC, with support from NIH P41-GM104601 and R24-GM145965.

Acknowledgments

We would like to thank Christopher Maffeo and Tyler Earnest for insightful discussions.

Conflict of interest

JES was employed by the company NVIDIA Corporation.

The remaining 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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1214962/full#supplementary-material

References

Abebe, A. H., Aranovich, A., and Fishov, I. (2017). HU content and dynamics in escherichia coli during the cell cycle and at different growth rates. *FEMS Microbiol. Lett.* 364, fnx195. doi:10.1093/femsle/fnx195

Abraham, M., Alekseenko, A., Bergh, C., Blau, C., Briand, E., Doijade, M., et al. (2023). Gromacs 2023 source code. doi:10.5281/ZENODO.7588619

Alipour, E., and Marko, J. F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* 40, 11202–11212. doi:10.1093/nar/gks925

Amblard, F., Maggs, A. C., Yurke, B., Pargellis, A. N., and Leibler, S. (1996). Subdiffusion and anomalous local viscoelasticity in actin networks. *Phys. Rev. Lett.* 77, 4470–4473. doi:10.1103/physrevlett.77.4470

Badrinarayanan, A., Le, T. B., and Laub, M. T. (2015). Bacterial chromosome organization and segregation. *Annu. Rev. Cell. Dev. Biol.* 31, 171–199. doi:10.1146/annurev-cellbio-100814-125211

Bakshi, S., Siryaporn, A., Goulian, M., and Weisshaar, J. C. (2012). Superresolution imaging of ribosomes and RNA polymerase in live *escherichia coli* cells. *Mol. Microbiol.* 85, 21–38. doi:10.1111/j.1365-2958.2012.08081.x

Banigan, E. J., van den Berg, A. A., Brandão, H. B., Marko, J. F., and Mirny, L. A. (2020). Chromosome organization by one-sided and two-sided loop extrusion. *eLife* 9, e53558. doi:10.7554/elife.53558

Barkai, E., Garini, Y., and Metzler, R. (2012). Strange kinetics of single molecules in living cells. *Phys. Today* 65, 29–35. doi:10.1063/pt.3.1677

Berg, O. G. (1979). Brownian motion of the wormlike chain and segmental diffusion of DNA. *Biopolymers* 18, 2861–2874. doi:10.1002/bip.1979.360181114

Bhat, D., Hauf, S., Plessy, C., Yokobayashi, Y., and Pigolotti, S. (2022). Speed variations of bacterial replisomes. *eLife* 11, e75884. doi:10.7554/elife.75884

Bianchi, D. M., Peterson, J. R., Earnest, T. M., Hallock, M. J., and Luthey-Schulten, Z. (2018). Hybrid CME–ODE method for efficient simulation of the galactose switch in yeast. *IET Syst. Biol.* 12, 170–176. doi:10.1049/iet-syb.2017.0070

Birnie, A., and Dekker, C. (2020). Genome-in-a-box: Building a chromosome from the bottom up. *ACS Nano* 15, 111–124. doi:10.1021/acsnano.0c07397

Bonato, A., and Michieletto, D. (2021). Three-dimensional loop extrusion. *Biophysical J.* 120, 5544–5552. doi:10.1016/j.bpj.2021.11.015

Brackley, C. A., Morozov, A. N., and Marenduzzo, D. (2014). Models for twistable elastic polymers in brownian dynamics, and their implementation for LAMMPS. *J. Chem. Phys.* 140, 135103. doi:10.1063/1.4870088

Brahmachari, S., and Marko, J. F. (2019). Chromosome disentanglement driven via optimal compaction of loop-extruded brush structures. *Proc. Natl. Acad. Sci. U.S.A.* 116, 24956–24965. doi:10.1073/pnas.1906355116

Bremer, H., and Dennis, P. P. (2008). Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus* 3, 1–48. doi:10.1128/ecosal.5.2.3

Breuer, M., Earnest, T. M., Merryman, C., Wise, K. S., Sun, L., Lynott, M. R., et al. (2019). Essential metabolism for a minimal cell. *eLife* 8, e36842. doi:10.7554/elife.36842

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4, 187–217. doi:10.1002/jcc.540040211

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Genet.* 21, 167–195. doi:10.1002/prot.340210302

Buenemann, M., and Lenz, P. (2010). A geometrical model for DNA organization in bacteria. *PLoS ONE* 5, e13806. doi:10.1371/journal.pone.0013806

Carstens, S., Nilges, M., and Habeck, M. (2020). Bayesian inference of chromatin structure ensembles from population-averaged contact data. *Proc. Natl. Acad. Sci.* 117, 7824–7830. doi:10.1073/pnas.1910364117

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., et al. (2005). The amber biomolecular simulation programs. *J. Comput. Chem.* 26, 1668–1688. doi:10.1002/jcc.20290

Castro-Villarreal, P., and Ramírez, J. E. (2021). Semiflexible polymer enclosed in a 3d compact domain. *Front. Phys.* 9. doi:10.3389/fphy.2021.642364

Cebrián, J., Castán, A., Martínez, V., Kadomatsu-Hermosa, M. J., Parra, C., Fernández-Nestosa, M. J., et al. (2015). Direct evidence for the formation of precatenanes during DNA replication. *J. Biol. Chem.* 290, 13725–13735. doi:10.1074/jbc.m115.642272

Chatterjee, P., Goldenfeld, N., and Kim, S. (2021). DNA supercoiling drives a transition between collective modes of gene synthesis. *Phys. Rev. Lett.* 127, 218101. doi:10.1103/physrevlett.127.218101

Chodavarapu, S., Felczak, M. M., Yaniv, J. R., and Kaguni, J. M. (2007). *Escherichia coli* DnaA interacts with HU in initiation at the e. coli replication origin. *Mol. Microbiol.* 67, 781–792. doi:10.1111/j.1365-2958.2007.06094.x

Chong, S., Chen, C., Ge, H., and Xie, X. S. (2014). Mechanism of transcriptional bursting in bacteria. *Cell.* 158, 314–326. doi:10.1016/j.cell.2014.05.038

Cifra, P., and Bleha, T. (2010). Shape transition of semi-flexible macromolecules confined in channel and cavity. *Eur. Phys. J. E* 32, 273–279. doi:10.1140/epje/i2010-10626-y

Cocco, S., Marko, J. F., and Monasson, R. (2002). Theoretical models for single-molecule DNA and RNA experiments: From elasticity to unzipping. *Comptes Rendus Phys.* 3, 569–584. doi:10.1016/s1631-0705(02)01345-2

Conin, B., Billault-Chaumartin, I., Sayyed, H. E., Quenech'Du, N., Cockram, C., Koszul, R., et al. (2022). Extended sister-chromosome catenation leads to massive reorganization of the *E. coli* genome. *Nucleic Acids Res.* 50, 2635–2650. doi:10.1093/nar/gkac105

Cooper, S., and Helmstetter, C. E. (1968). Chromosome replication and the division cycle of *escherichia coli*. *J. Mol. Biol.* 31, 519–540. doi:10.1016/0022-2836(68)90425-7

Dame, R. T., Rashid, F.-Z. M., and Grainger, D. C. (2019). Chromosome organization in bacteria: Mechanistic insights into genome structure and function. *Nat. Rev. Genet.* 21, 227–242. doi:10.1038/s41576-019-0185-4

Dame, R. T., and Tark-Dame, M. (2016). Bacterial chromatin: Converging views at different scales. *Curr. Opin. Cell. Biol.* 40, 60–65. doi:10.1016/j.ceb.2016.02.015

Dame, R. T. (2005). The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Mol. Microbiol.* 56, 858–870. doi:10.1111/j.1365-2958.2005.04598.x

Davidson, I. F., and Peters, J.-M. (2021). Genome folding through loop extrusion by SMC complexes. *Nat. Rev. Mol. Cell. Biol.* 22, 445–464. doi:10.1038/s41580-021-00349-7

de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., et al. (2013). Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* 9, 687–697. doi:10.1021/ct300646g

Dekker, J., and Mirny, L. (2016). The 3d genome as moderator of chromosomal communication. *Cell.* 164, 1110–1121. doi:10.1016/j.cell.2016.02.007

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science* 295, 1306–1311. doi:10.1126/science.1067799

Delong, S., Usabiaga, F. B., and Donev, A. (2015). Brownian dynamics of confined rigid bodies. *J. Chem. Phys.* 143, 144107. doi:10.1063/1.4932062

Denker, A., and de Laat, W. (2016). The second decade of 3c technologies: Detailed insights into nuclear organization. *Genes. & Dev.* 30, 1357–1382. doi:10.1101/gad.281964.116

Dersch, S., Rotter, D. A., and Graumann, P. L. (2022). Heterogeneity of subcellular diffusion in bacteria based on spatial segregation of ribosomes and nucleoids. *Microb. Physiol.* 32, 177–186. doi:10.1159/000526846

Di Pierro, M., Cheng, R. R., Aiden, E. L., Wolynes, P. G., and Onuchic, J. N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode

genome architecture. *Proc. Natl. Acad. Sci.* 114, 12126–12131. doi:10.1073/pnas.1714980114

Diebold-Durand, M.-L., Lee, H., Avila, L. B. R., Noh, H., Shin, H.-C., Im, H., et al. (2017). Structure of full-length SMC and rearrangements required for chromosome organization. *Mol. Cell.* 67, 334–347.e5. doi:10.1016/j.molcel.2017.06.010

Dierckx, P. (1996). *Monographs on numerical analysis*. repr edn. Oxford: Clarendon.Curve and surface fitting with splines

Dingman, C. W. (1974). Bidirectional chromosome replication: Some topological considerations. *J. Theor. Biol.* 43, 187–195. doi:10.1016/s0022-5193(74)80052-4

Doi, M., and Edwards, S. F. (1988). "The theory of polymer dynamics," in *International series of monographs on physics* (Oxford, England: Clarendon Press).

Dorman, C. J. (2019). DNA supercoiling and transcription in bacteria: A two-way street. *BMC Mol. Cell. Biol.* 20, 26. doi:10.1186/s12860-019-0211-6

Español, P., and Warren, P. B. (2017). Perspective: Dissipative particle dynamics. *J. Chem. Phys.* 146, 150901. doi:10.1063/1.4979514

Espinosa, E., Paly, E., and Barre, F.-X. (2020). High-resolution whole-genome analysis of sister-chromatid contacts. *Mol. Cell.* 79, 857–869.e3. doi:10.1016/j.molcel.2020.06.033

Fiorillo, L., Musella, F., Conte, M., Kempfer, R., Chiariello, A. M., Bianco, S., et al. (2021). Comparison of the hi-c, GAM and SPRITE methods using polymer models of chromatin. *Nat. Methods* 18, 482–490. doi:10.1038/s41592-021-01135-1

Fournier, A., Fussell, D., and Carpenter, L. (1982). Computer rendering of stochastic models. *Commun. ACM* 25, 371–384. doi:10.1145/358523.358553

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016). Formation of chromosomal domains by loop extrusion. *Cell. Rep.* 15, 2038–2049. doi:10.1016/j.celrep.2016.04.085

Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., et al. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science* 360, 102–105. doi:10.1126/science.aar7831

Geggier, S., Kotlyar, A., and Vologodskii, A. (2010). Temperature dependence of DNA persistence length. *Nucleic Acids Res.* 39, 1419–1426. doi:10.1093/nar/gkq932

Geng, Y., Bohrer, C. H., Yehya, N., Hendrix, H., Shachaf, L., Liu, J., et al. (2022). A spatially resolved stochastic model reveals the role of supercoiling in transcription regulation. *PLOS Comput. Biol.* 18, e1009788. doi:10.1371/journal.pcbi.1009788

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52–56. doi:10.1126/science.1190719

Gilbert, B. R., Thornburg, Z. R., Lam, V., Rashid, F.-Z. M., Glass, J. I., Villa, E., et al. (2021). Generating chromosome geometries in a minimal cell from cryo-electron tomograms and chromosome conformation capture maps. *Front. Mol. Biosci.* 8, 644133. doi:10.3389/fmolb.2021.644133

Giorgetti, L., and Heard, E. (2016). Closing the loop: 3c versus DNA FISH. *Genome Biol.* 17, 215. doi:10.1186/s13059-016-1081-2

Goel, V. Y., and Hansen, A. S. (2020). The macro and micro of chromosome conformation capture. *WIREs Dev. Biol.* 10, e395. doi:10.1002/wdev.395

Gogou, C., Japaridze, A., and Dekker, C. (2021). Mechanisms for chromosome segregation in bacteria. *Front. Microbiol.* 12, 685687. doi:10.3389/fmicb.2021.685687

Goldberg, A. P., Szigeti, B., Chew, Y. H., Sekar, J. A., Roth, Y. D., and Karr, J. R. (2018). Emerging whole-cell modeling principles and methods. *Curr. Opin. Biotechnol.* 51, 97–102. doi:10.1016/j.copbio.2017.12.013

Goloborodko, A., Imakaev, M. V., Marko, J. F., and Mirny, L. (2016a). Compaction and segregation of sister chromatids via active loop extrusion. *eLife* 5, e14864. doi:10.7554/elife.14864

Goloborodko, A., Marko, J. F., and Mirny, L. A. (2016b). Chromosome compaction by active loop extrusion. *Biophysical J.* 110, 2162–2168. doi:10.1016/j.bpj.2016.02.041

Goodsell, D. S., Autin, L., and Olson, A. J. (2018). Lattice models of bacterial nucleoids. *J. Phys. Chem. B* 122, 5441–5447. doi:10.1021/acs.jpcb.7b11770

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi:10.1038/nrg.2016.49

Grünewald, F., Alessandri, R., Kroon, P. C., Monticelli, L., Souza, P. C. T., and Marrink, S. J. (2022). Polyply; a python suite for facilitating simulations of macromolecules and nanomaterials. *Nat. Commun.* 13, 68. doi:10.1038/s41467-021-27627-4

Guo, M. S., Kawamura, R., Littlehale, M. L., Marko, J. F., and Laub, M. T. (2021). High-resolution, genome-wide mapping of positive supercoiling in chromosomes. *eLife* 10, e67236. doi:10.7554/elife.67236

Haas, D., Thamm, A. M., Sun, J., Huang, L., Sun, L., Beaudoin, G. A. W., et al. (2022). Metabolite damage and damage control in a minimal genome. *mBio* 13, e0163022. doi:10.1128/mbio.01630-22

Hacker, W. C., Li, S., and Elcock, A. H. (2017). Features of genomic organization in a nucleotide-resolution molecular model of the *escherichia coli* chromosome. *Nucleic Acids Res.* 45, 7541–7554. doi:10.1093/nar/gkx541

Hallock, M. J., Stone, J. E., Roberts, E., Fry, C., and Luthey-Schulten, Z. (2014). Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations. *Parallel Comput.* 40, 86–99. doi:10.1016/j.parco.2014.03.009

Heo, L., Sugita, Y., and Feig, M. (2022). Protein assembly and crowding simulations. *Curr. Opin. Struct. Biol.* 73, 102340. doi:10.1016/j.sbi.2022.102340

Higashi, T. L., Pobegalov, G., Tang, M., Molodtsov, M. I., and Uhlmann, F. (2021). A brownian ratchet model for DNA loop extrusion by the cohesin complex. *eLife* 10, e67530. doi:10.7554/elife.67530

Hirano, T. (2006). At the heart of the chromosome: SMC proteins in action. *Nat. Rev. Mol. Cell. Biol.* 7, 311–322. doi:10.1038/nrm1909

Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., et al. (2016). Design and synthesis of a minimal bacterial genome. *Science* 351, aad6253–aad6253–11. doi:10.1126/science.aad6253

Ilie, I. M., Briels, W. J., and den Otter, W. K. (2015). An elementary singularity-free rotational brownian dynamics algorithm for anisotropic particles. *J. Chem. Phys.* 142, 114103. doi:10.1063/1.4914322

Japaridze, A., Gogou, C., Kerssemakers, J. W. J., Nguyen, H. M., and Dekker, C. (2020). Direct observation of independently moving replisomes in *escherichia coli*. *Nat. Commun.* 11, 3109. doi:10.1038/s41467-020-16946-7

Jun, S., and Mulder, B. (2006). Entropy-driven spatial organization of highly confined polymers: Lessons for the bacterial chromosome. *Proc. Natl. Acad. Sci.* 103, 12388–12393. doi:10.1073/pnas.0605305103

Jung, Y., and Ha, B.-Y. (2010). Overlapping two self-avoiding polymers in a closed cylindrical pore: Implications for chromosome segregation in a bacterial cell. *Phys. Rev. E* 82, 051926. doi:10.1103/physreve.82.051926

Jung, Y., Jeon, C., Kim, J., Jeong, H., Jun, S., and Ha, B.-Y. (2012). Ring polymers as model bacterial chromosomes: Confinement, chain topology, single chain statistics, and how they interact. *Soft Matter* 8, 2095–2102. doi:10.1007/s00726-011-0946-7

Junier, I., Boccard, F., and Espéli, O. (2013). Polymer modeling of the e. coli genome reveals the involvement of locus positioning and macrodomain structuring for the control of chromosome conformation and segregation. *Nucleic Acids Res.* 42, 1461–1473. doi:10.1093/nar/gkt1005

Junier, I., Spill, Y. G., Marti-Renom, M. A., Beato, M., and le Dily, F. (2015). On the demultiplexing of chromosome capture conformation data. *FEBS Lett.* 589, 3005–3013. doi:10.1016/j.febslet.2015.05.049

Kamashev, D., and Rouviere-Yaniv, J. (2000). The histone-like protein HU binds specifically to DNA recombination and repair intermediates. *EMBO J.* 19, 6527–6535. doi:10.1093/emboj/19.23.6527

Karaboja, X., and Wang, X. (2022). HBsu is required for the initiation of DNA replication in bacillus subtilis. *J. Bacteriol.* 204, e0011922. doi:10.1128/jb.00119-22

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, 389–401. doi:10.1016/j.cell.2012.05.044

Kempfer, R., and Pombo, A. (2019). Methods for mapping 3d chromosome architecture. *Nat. Rev. Genet.* 21, 207–226. doi:10.1038/s41576-019-0195-2

Khan, S. R., Mahaseth, T., Kouzminova, E. A., Cronan, G. E., and Kuzminov, A. (2016). Static and dynamic factors limit chromosomal replication complexity in *escherichia coli*, avoiding dangers of runaway overreplication. *Genetics* 202, 945–960. doi:10.1534/genetics.115.184697

Kim, E., Kerssemakers, J., Shaltiel, I. A., Haering, C. H., and Dekker, C. (2020). DNA-loop extruding condensin complexes can traverse one another. *Nature* 579, 438–442. doi:10.1038/s41586-020-2067-5

Kim, S., Beltran, B., Irnov, I., and Jacobs-Wagner, C. (2019). Long-distance cooperative and antagonistic RNA polymerase dynamics via DNA supercoiling. *Cell.* 179, 106–119.e16. doi:10.1016/j.cell.2019.08.033

Klenin, K., Merlitz, H., and Langowski, J. (1998). A brownian dynamics program for the simulation of linear and circular DNA and other wormlike chain polyelectrolytes. *Biophysical J.* 74, 780–788. doi:10.1016/s0006-3495(98)74003-2

Klippenstein, V., Tripathy, M., Jung, G., Schmid, F., and van der Vegt, N. F. A. (2021). Introducing memory in coarse-grained molecular simulations. *J. Phys. Chem. B* 125, 4931–4954. doi:10.1021/acs.jpcb.1c01120

Kos, P. I., Galitsyna, A. A., Ulianov, S. V., Gelfand, M. S., Razin, S. V., and Chertovich, A. V. (2021). Perspectives for the reconstruction of 3d chromatin conformation using single cell hi-c data. *PLOS Comput. Biol.* 17, e1009546. doi:10.1371/journal.pcbi.1009546

Kroon, P. C., Grünewald, F., Barnoud, J., van Tilburg, M., Souza, P. C. T., Wassenaar, T. A., et al. (2022). *Martinize2 and vermouth: Unified framework for topology generation*. arXiv. doi:10.48550/ARXIV.2212.01191

Le, T. B. K., Imakaev, M. V., Mirny, L. A., and Laub, M. T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* 342, 731–734. doi:10.1126/science.1242059

Lee, H., Noh, H., and Ryu, J.-K. (2021). Structure-function relationships of SMC protein complexes for DNA loop extrusion. *BIODESIGN* 9, 1–13. doi:10.34184/kssb.2021.9.1.1

Lesterlin, C., Gigant, E., Boccard, F., and Espéli, O. (2012). Sister chromatid interactions in bacteria revealed by a site-specific recombination assay. *EMBO J.* 31, 3468–3479. doi:10.1038/emboj.2012.194

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi:10.1126/science.1181369

Lioy, V. S., Cournac, A., Marbouty, M., Duigou, S., Mozziconacci, J., Espéli, O., et al. (2018). Multiscale structuring of the *e. coli* chromosome by nucleoid-associated and condensin proteins. *Cell.* 172, 771–783.e18. doi:10.1016/j.cell.2017.12.027

Lioy, V. S., Junier, I., and Boccard, F. (2021). Multiscale dynamic structuring of bacterial chromosomes. *Annu. Rev. Microbiol.* 75, 541–561. doi:10.1146/annurev-micro-033021-113232

Lioy, V. S., Junier, I., Lagage, V., Vallet, I., and Boccard, F. (2020). Distinct activities of bacterial condensins for chromosome management in pseudomonas aeruginosa. *Cell. Rep.* 33, 108344. doi:10.1016/j.celrep.2020.108344

Liu, L. F., and Wang, J. C. (1987). Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci.* 84, 7024–7027. doi:10.1073/pnas.84.20.7024

Liu, L., Liu, C., and Alberts, B. (1980). Type II DNA topoisomerases: Enzymes that can unknot a topologically knotted DNA molecule via a reversible double-strand break. *Cell.* 19, 697–707. doi:10.1016/s0092-8674(80)80046-8

Liu, Y., and Chakraborty, B. (2008). Shapes of semiflexible polymers in confined spaces. *Phys. Biol.* 5, 026004. doi:10.1088/1478-3975/5/2/026004

Livny, J., Yamaichi, Y., and Waldor, M. K. (2007). Distribution of centromere-like parS sites in bacteria: Insights from comparative genomics. *J. Bacteriol.* 189, 8693–8703. doi:10.1128/jb.01239-07

Liwo, A., Ołdziej, S., Pincus, M. R., Wawak, R. J., Rackovsky, S., and Scheraga, H. A. (1997). A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* 18, 849–873. doi:10.1002/(sici)1096-987x(199705)18:7<849::aid-jcc1>3.0.co;2-r

Llopis, P. M., Jackson, A. F., Sliusarenko, O., Surovtsev, I., Heinritz, J., Emonet, T., et al. (2010). Spatial organization of the flow of genetic information in bacteria. *Nature* 466, 77–81. doi:10.1038/nature09152

Lua, R., Borovinskiy, A. L., and Grosberg, A. Y. (2004). Fractal and statistical properties of large compact polymers: A computational study. *Polymer* 45, 717–731. doi:10.1016/j.polymer.2003.10.073

Luthey-Schulten, Z., Thornburg, Z. R., and Gilbert, B. R. (2022). Integrating cellular and molecular structures and dynamics into whole-cell models. *Curr. Opin. Struct. Biol.* 75, 102392. doi:10.1016/j.sbi.2022.102392

Machado, M. R., and Pantano, S. (2016). SIRAH tools: Mapping, backmapping and visualization of coarse-grained models. *Bioinformatics* 32, 1568–1570. doi:10.1093/bioinformatics/btw020

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, eaav3751. doi:10.1126/science.aav3751

Maffeo, C., and Aksimentiev, A. (2020). MrDNA: A multi-resolution model for predicting the structure and dynamics of DNA systems. *Nucleic Acids Res.* 48, 5135–5146. doi:10.1093/nar/gkaa200

Maffeo, C., Chou, H.-Y., and Aksimentiev, A. (2022). Single-molecule biophysics experiments *in silico*: Toward a physical model of a replisome. *iScience* 25, 104264. doi:10.1016/j.isci.2022.104264

Mäkelä, J., and Sherratt, D. J. (2020). Organization of the escherichia coli chromosome by a MukBEF axial core. *Mol. Cell.* 78, 250–260.e5. doi:10.1016/j.molcel.2020.02.003

Manning, G. S. (2006). The persistence length of DNA is reached from the persistence length of its null isomer through an internal electrostatic stretching force. *Biophysical J.* 91, 3607–3616. doi:10.1529/biophysj.106.089029

Mantelli, S., Muller, P., Harlepp, S., and Maaloum, M. (2011). Conformational analysis and estimation of the persistence length of DNA using atomic force microscopy in solution. *Soft Matter* 7, 3412. doi:10.1039/c0sm01160f

Marbouty, M., Gall, A. L., Cattoni, D. I., Cournac, A., Koh, A., Fiche, J.-B., et al. (2015). Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by hi-c and super-resolution imaging. *Mol. Cell.* 59, 588–602. doi:10.1016/j.molcel.2015.07.020

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, 167351. doi:10.1016/j.jmb.2021.167351

Marko, J. F. (2009). Linking topology of tethered polymer rings with applications to chromosome segregation and estimation of the knotting length. *Phys. Rev. E* 79, 051905. doi:10.1103/physreve.79.051905

Marko, J. F. (2011). Scaling of linking and writhing numbers for spherically confined and topologically equilibrated flexible polymers. *J. Stat. Phys.* 142, 1353–1370. doi:10.1007/s10955-011-0172-4

Marrink, S. J., Monticelli, L., Melo, M. N., Alessandri, R., Tieleman, D. P., and Souza, P. C. T. (2022). Two decades of martini: Better beads, broader scope. *WIREs Comput. Mol. Sci.* 13. doi:10.1002/wcms.1620

Marucci, L., Barberis, M., Karr, J., Ray, O., Race, P. R., de Souza Andrade, M., et al. (2020). Computer-aided whole-cell design: Taking a holistic approach by integrating synthetic with systems biology. *Front. Bioeng. Biotechnol.* 8, 942. doi:10.3389/fbioe.2020.00942

McKie, S. J., Neuman, K. C., and Maxwell, A. (2021). DNA topoisomerases: Advances in understanding of cellular roles and multi-protein complexes via structure-function analysis. *BioEssays* 43, 2000286. doi:10.1002/bies.202000286

Messelink, J. J. B., van Teeseling, M. C. F., Janssen, J., Thanbichler, M., and Broedersz, C. P. (2021). Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales. *Nat. Commun.* 12, 1963. doi:10.1038/s41467-021-22189-x

Mitra, D., Pande, S., and Chatterji, A. (2022a). Polymer architecture orchestrates the segregation and spatial organization of replicating *E. coli* chromosomes in slow growth. *Soft Matter* 18, 5615–5631. doi:10.1039/d2sm00734g

Mitra, D., Pande, S., and Chatterji, A. (2022b). Topology-driven spatial organization of ring polymers under confinement. *Phys. Rev. E* 106, 054502. doi:10.1103/physreve.106.054502

Mitter, M., Gasser, C., Takacs, Z., Langer, C. C. H., Tang, W., Jessberger, G., et al. (2020). Conformation of sister chromatids in the replicated human genome. *Nature* 586, 139–144. doi:10.1038/s41586-020-2744-4

Mondal, J., Bratton, B. P., Li, Y., Yethiraj, A., and Weisshaar, J. C. (2011). Entropy-based mechanism of ribosome-nucleoid segregation in e. coli cells. *Biophysical J.* 100, 2605–2613. doi:10.1016/j.bpj.2011.04.030

Mosconi, F., Allemand, J. F., Bensimon, D., and Croquette, V. (2009). Measurement of the torque on a single stretched and twisted DNA using magnetic tweezers. *Phys. Rev. Lett.* 102, 078301. doi:10.1103/physrevlett.102.078301

Muñoz-Gil, G., Volpe, G., Garcia-March, M. A., Aghion, E., Argun, A., Hong, C. B., et al. (2021). Objective comparison of methods to decode anomalous diffusion. *Nat. Commun.* 12, 6253. doi:10.1038/s41467-021-26320-w

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., et al. (2013). Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature* 502, 59–64. doi:10.1038/nature12593

Nielsen, H. J., Li, Y., Youngren, B., Hansen, F. G., and Austin, S. (2006). Progressive segregation of the escherichia coli chromosome. *Mol. Microbiol.* 61, 383–393. doi:10.1111/j.1365-2958.2006.05245.x

Nolivos, S., Upton, A. L., Badrinarayanan, A., Müller, J., Zawadzka, K., Wiktor, J., et al. (2016). MatP regulates the coordinated action of topoisomerase IV and MukBEF in chromosome segregation. *Nat. Commun.* 7, 10466. doi:10.1038/ncomms10466

Nomidis, S. K., Carlon, E., Gruber, S., and Marko, J. F. (2022). DNA tension-modulated translocation and loop extrusion by SMC complexes revealed by molecular dynamics simulations. *Nucleic Acids Res.* 50, 4974–4987. doi:10.1093/nar/gkac268

Nunez, R. V., Avila, L. B. R., and Gruber, S. (2019). Transient DNA occupancy of the SMC interarm space in prokaryotic condensin. *Mol. Cell.* 75, 209–223.e6. doi:10.1016/j.molcel.2019.05.001

Oliveira, F. A., Ferreira, R. M. S., Lapas, L. C., and Vainstein, M. H. (2019). Anomalous diffusion: A basic mechanism for the evolution of inhomogeneous systems. *Front. Phys.* 7. doi:10.3389/fphy.2019.00018

Olivi, L., Berger, M., Creyghton, R. N. P., Franceschi, N. D., Dekker, C., Mulder, B. M., et al. (2021). Towards a synthetic cell cycle. *Nat. Commun.* 12, 4531. doi:10.1038/s41467-021-24772-8

Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997). Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* 48, 545–600. doi:10.1146/annurev.physchem.48.1.545

Oomen, M. E., Hedger, A. K., Watts, J. K., and Dekker, J. (2020). Detecting chromatin interactions between and along sister chromatids with SisterC. *Nat. Methods* 17, 1002–1009. doi:10.1038/s41592-020-0930-9

O'Reilly, F. J., Xue, L., Graziadei, A., Sinn, L., Lenz, S., Tegunov, D., et al. (2020). In-cell architecture of an actively transcribing-translating expressome. *Science* 369, 554–557. doi:10.1126/science.abb3758

Orlandini, E., Marenduzzo, D., and Michieletto, D. (2019). Synergy of topoisomerase and structural-maintenance-of-chromosomes proteins creates a universal pathway to simplify genome topology. *Proc. Natl. Acad. Sci.* 116, 8149–8154. doi:10.1073/pnas.1815394116

Páll, S., Zhmurov, A., Bauer, P., Abraham, M., Lundborg, M., Gray, A., et al. (2020). Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *J. Chem. Phys.* 153, 134110. doi:10.1063/5.0018516

Park, P. J. (2009). ChIP–seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680. doi:10.1038/nrg2641

Parry, B. R., Surovtsev, I. V., Cabeen, M. T., O'Hern, C. S., Dufresne, E. R., and Jacobs-Wagner, C. (2014). The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity. *Cell.* 156, 183–194. doi:10.1016/j.cell.2013.11.028

Patrone, P. N., and Rosch, T. W. (2017). Beyond histograms: Efficiently estimating radial distribution functions via spectral Monte Carlo. *J. Chem. Phys.* 146, 094107. doi:10.1063/1.4977516

Pelletier, J. F., Sun, L., Wise, K. S., Assad-Garcia, N., Karas, B. J., Deerinck, T. J., et al. (2021). Genetic requirements for cell division in a genomically minimal cell. *Cell.* 184, 2430–2440.e16. doi:10.1016/j.cell.2021.03.008

Pelletier, J., Halvorsen, K., Ha, B.-Y., Paparcone, R., Sandler, S. J., Woldringh, C. L., et al. (2012). Physical manipulation of the *Escherichia coli* chromosome reveals its soft nature. *Proc. Natl. Acad. Sci.* 109, E2649–E2656. doi:10.1073/pnas.1208689109

Pezeshkian, W., König, M., Wassenaar, T. A., and Marrink, S. J. (2020). Backmapping triangulated surfaces to coarse-grained membrane models. *Nat. Commun.* 11, 2296. doi:10.1038/s41467-020-16094-y

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26, 1781–1802. doi:10.1002/jcc.20289

Popov, K., Komianos, J., and Papoian, G. A. (2016). Medyan: Mechanochemical simulations of contraction and polarity alignment in actomyosin networks. *PLOS Comput. Biol.* 12, e1004877. doi:10.1371/journal.pcbi.1004877

Pountain, A. W., Jiang, P., Yao, T., Homaee, E., Guan, Y., Podkowik, M., et al. (2022). Transcription-replication interactions reveal principles of bacterial genome regulation. *bioRxiv*. doi:10.1101/2022.10.22.513359

Racko, D., Benedetti, F., Goundaroulis, D., and Stasiak, A. (2018). Chromatin loop extrusion and chromatin unknotting. *Polymers* 10, 1126. doi:10.3390/polym10101126

Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., et al. (2017). Massively multiplex single-cell hi-c. *Nat. Methods* 14, 263–266. doi:10.1038/nmeth.4155

Richardson, T. T., Stevens, D., Pelliciari, S., Harran, O., Sperlea, T., and Murray, H. (2019). Identification of a basal system for unwinding a bacterial chromosome origin. *EMBO J.* 38, e101649. doi:10.15252/embj.2019101649

Rickard, M. M., Zhang, Y., Gruebele, M., and Pogorelov, T. V. (2019). In-cell protein–protein contacts: Transient interactions in the crowd. *J. Phys. Chem. Lett.* 10, 5667–5673. doi:10.1021/acs.jpclett.9b01556

Rideau, F., Villa, A., Belzanne, P., Verdier, E., Hosy, E., and Arfi, Y. (2022). Imaging minimal bacteria at the nanoscale: A reliable and versatile process to perform single-molecule localization microscopy in mycoplasmas. *Microbiol. Spectr.* 10, e0064522. doi:10.1128/spectrum.00645-22

Roberts, E., Stone, J. E., and Luthey-Schulten, Z. (2012). Lattice microbes: High-performance stochastic simulation method for the reaction-diffusion master equation. *J. Comput. Chem.* 34, 245–255. doi:10.1002/jcc.23130

Rosa, A., and Zimmer, C. (2014). "Computational models of large-scale genome architecture," in *International review of cell and molecular Biology* (Elsevier), 275–349. doi:10.1016/b978-0-12-800046-5.00009-6

Rowland, B., Huh, R., Hou, Z., Crowley, C., Wen, J., Shen, Y., et al. (2022). Thunder: A reference-free deconvolution method to infer cell type proportions from bulk hi-c data. *PLOS Genet.* 18, e1010102. doi:10.1371/journal.pgen.1010102

Ryu, J.-K., Rah, S.-H., Janissen, R., Kerssemakers, J. W. J., Bonato, A., Michieletto, D., et al. (2021). Condensin extrudes DNA loops in steps up to hundreds of base pairs that are generated by ATP binding events. *Nucleic Acids Res.* 50, 820–832. doi:10.1093/nar/gkab1268

Sanamrad, A., Persson, F., Lundius, E. G., Fange, D., Gynnå, A. H., and Elf, J. (2014). Single-particle tracking reveals that free ribosomal subunits are not excluded from the *escherichia coli* nucleoid. *Proc. Natl. Acad. Sci.* 111, 11413–11418. doi:10.1073/pnas.1411558111

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* 112, E6456–E6465. doi:10.1073/pnas.1518552112

Schramm, F. D., and Murray, H. (2022). HU knew? bacillus subtilis HBsu is required for DNA replication initiation. *J. Bacteriol.* 204, e0015122. doi:10.1128/jb.00151-22

Sefer, E., Duggal, G., and Kingsford, C. (2016). Deconvolution of ensemble chromatin interaction data reveals the latent mixing structures in cell subpopulations. *J. Comput. Biol.* 23, 425–438. doi:10.1089/cmb.2015.0210

Sekelja, M., Paulsen, J., and Collas, P. (2016). 4d nucleomes in single cells: What can computational modeling reveal about spatial chromatin conformation? *Genome Biol.* 17, 54. doi:10.1186/s13059-016-0923-2

Sengar, A., Ouldridge, T. E., Henrich, O., Rovigatti, L., and Šulc, P. (2021). A primer on the oxDNA model of DNA: When to use it, how to simulate it and how to interpret the results. *Front. Mol. Biosci.* 8, 693710. doi:10.3389/fmolb.2021.693710

Shi, G., and Thirumalai, D. (2023). A maximum-entropy model to predict 3d structural ensembles of chromatin from pairwise distances with applications to interphase chromosomes and structural variants. *Nat. Commun.* 14, 1150. doi:10.1038/s41467-023-36412-4

Shimizu, M., Noguchi, Y., Sakiyama, Y., Kawakami, H., Katayama, T., and Takada, S. (2016). Near-atomic structural model for bacterial DNA replication initiation complex and its functional insights. *Proc. Natl. Acad. Sci.* 113, E8021–E8030. doi:10.1073/pnas.1609649113

Sinden, R. R. (1994). *DNA structure and function*. Elsevier. doi:10.1016/C2009-0-02451-9

Śmigiel, W. M., Mantovanelli, L., Linnik, D. S., Punter, M., Silberberg, J., Xiang, L., et al. (2022). Protein diffusion in *escherichia coli* cytoplasm scales with the mass of the complexes and is location dependent. *Sci. Adv.* 8, eabo5387. doi:10.1126/sciadv.abo5387

Snook, I. (2007). "Langevin and generalised Langevin dynamics," in *The Langevin and generalised Langevin approach to the dynamics of atomic, polymeric and colloidal systems* (Elsevier), 107–132. doi:10.1016/b978-044452129-3/50007-9

Sorichetti, V., Hugouvieux, V., and Kob, W. (2020). Determining the mesh size of polymer solutions via the pore size distribution. *Macromolecules* 53, 2568–2581. doi:10.1021/acs.macromol.9b02166

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

Strzałka, A., Kois-Ostrowska, A., Kędra, M., Łebkowski, T., Bieniarz, G., Szafran, M. J., et al. (2022). Enhanced binding of an HU homologue under increased DNA supercoiling preserves chromosome organisation and sustains *Streptomyces* hyphal growth. *Nucleic Acids Res.* 50, 12202–12216. doi:10.1093/nar/gkac1093

Sutormin, D. A., Galivondzhyan, A. K., Polkhovskiy, A. V., Kamalyan, S. O., Severinov, K. V., and Dubiley, S. A. (2021). Diversity and functions of type II topoisomerases. *Acta Naturae* 13, 59–75. doi:10.32607/actanaturae.11058

Szatmári, D., Sárkány, P., Kocsis, B., Nagy, T., Miseta, A., Barkó, S., et al. (2020). Intracellular ion concentrations and cation-dependent remodelling of bacterial MreB assemblies. *Sci. Rep.* 10, 12002. doi:10.1038/s41598-020-68960-w

Takaki, R., Dey, A., Shi, G., and Thirumalai, D. (2021). Theory and simulations of condensin mediated loop extrusion in DNA. *Nat. Commun.* 12, 5865. doi:10.1038/s41467-021-26167-1

Taylor, J., and Garnier, R. (2009). *Discrete mathematics*. 3 edn. Boca Raton, FL: CRC Press.

Thirumalai, D., Lorimer, G. H., and Hyeon, C. (2019). Iterative annealing mechanism explains the functions of the GroEL and RNA chaperones. *Protein Sci.* 29, 360–377. doi:10.1002/pro.3795

Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., et al. (2022). Lammps - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* 271, 108171. doi:10.1016/j.cpc.2021.108171

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

Thornburg, Z. R., Melo, M. C. R., Bianchi, D., Brier, T. A., Crotty, C., Breuer, M., et al. (2019). Kinetic modeling of the genetic information processes in a minimal cell. *Front. Mol. Biosci.* 6, 130. doi:10.3389/fmolb.2019.00130

Tiana, G., and Giorgetti, L. (Editors) (2019). *Modeling the 3D conformation of genomes* (Boca Raton, FL: CRC Press). doi:10.1201/9781315144009

Todd, M. J., Lorimer, G. H., and Thirumalai, D. (1996). Chaperonin-facilitated protein folding: Optimization of rate and yield by an iterative annealing mechanism. *Proc. Natl. Acad. Sci.* 93, 4030–4035. doi:10.1073/pnas.93.9.4030

Tran, N. T., Laub, M. T., and Le, T. B. (2017). SMC progressively aligns chromosomal arms in *caulobacter crescentus* but is antagonized by convergent transcription. *Cell. Rep.* 20, 2057–2071. doi:10.1016/j.celrep.2017.08.026

Trussart, M., Yus, E., Martinez, S., Baù, D., Tahara, Y. O., Pengo, T., et al. (2017). Defined chromosome structure in the genome-reduced bacterium *mycoplasma pneumoniae*. *Nat. Commun.* 8, 14665. doi:10.1038/ncomms14665

Uusitalo, J. J., Ingólfsson, H. I., Akhshi, P., Tieleman, D. P., and Marrink, S. J. (2015). Martini coarse-grained force field: Extension to DNA. *J. Chem. Theory Comput.* 11, 3932–3945. doi:10.1021/acs.jctc.5b00286

Uusitalo, J. J., Ingólfsson, H. I., Marrink, S. J., and Faustino, I. (2017). Martini coarse-grained force field: Extension to RNA. *Biophysical J.* 113, 246–256. doi:10.1016/j.bpj.2017.05.043

Verma, S. C., Harned, A., Narayan, K., and Adhya, S. (2023). Non-specific and specific DNA binding modes of bacterial histone, HU, separately regulate distinct physiological processes through different mechanisms. *Mol. Microbiol.* 119, 439–455. doi:10.1111/mmi.15033

Verma, S. C., Qian, Z., and Adhya, S. L. (2019). Architecture of the escherichia coli nucleoid. *PLOS Genet.* 15, e1008456. doi:10.1371/journal.pgen.1008456

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat. Methods* 17, 261–272. doi:10.1038/s41592-019-0686-2

von Koch, H. (1904). On a continuous curve without a tangent, obtained by an elementary geometrical construction. *Ark. Mat. Astron. Fys.* 1, 681–702.

Wang, J. C. (1991). DNA topoisomerases: Why so many? *J. Biol. Chem.* 266, 6659–6662. doi:10.1016/s0021-9258(20)89545-3

Wang, W., Jüttler, B., Zheng, D., and Liu, Y. (2008a). Computation of rotation minimizing frames. *ACM Trans. Graph.* 27, 1–18. doi:10.1145/1330511.1330513

Wang, X., Reyes-Lamothe, R., and Sherratt, D. J. (2008b). Modulation of *Escherichia coli* sister chromosome cohesion by topoisomerase IV. *Genes. & Dev.* 22, 2426–2433. doi:10.1101/gad.487508

Wasim, A., Gupta, A., Bera, P., and Mondal, J. (2023). Interpretation of organizational role of proteins on e. coli nucleoid via hi-c integrated model. *Biophysical J.* 122, 63–81. doi:10.1016/j.bpj.2022.11.2938

Wasim, A., Gupta, A., and Mondal, J. (2021). A hi–c data-integrated model elucidates *e. coli* chromosome's multiscale organization at various replication stages. *Nucleic Acids Res.* 49, 3077–3091. doi:10.1093/nar/gkab094

Weber, S. C., Spakowitz, A. J., and Theriot, J. A. (2010a). Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm. *Phys. Rev. Lett.* 104, 238102. doi:10.1103/physrevlett.104.238102

Weber, S. C., Spakowitz, A. J., and Theriot, J. A. (2012). Nonthermal ATP-dependent fluctuations contribute to the *in vivo* motion of chromosomal loci. *Proc. Natl. Acad. Sci.* 109, 7338–7343. doi:10.1073/pnas.1119505109

Weber, S. C., Theriot, J. A., and Spakowitz, A. J. (2010b). Subdiffusive motion of a polymer composed of subdiffusive monomers. *Phys. Rev. E* 82, 011913. doi:10.1103/physreve.82.011913

Xiang, Y., Surovtsev, I. V., Chang, Y., Govers, S. K., Parry, B. R., Liu, J., et al. (2021). Interconnecting solvent quality, transcription, and chromosome folding in escherichia coli. *Cell.* 184, 3626–3642.e14. doi:10.1016/j.cell.2021.05.037

Xue, L., Lenz, S., Zimmermann-Kogadeeva, M., Tegunov, D., Cramer, P., Bork, P., et al. (2022). Visualizing translation dynamics at atomic detail inside a bacterial cell. *Nature* 610, 205–211. doi:10.1038/s41586-022-05255-2

Yamamoto, T., Izumi, S., and Gekko, K. (2006). Mass spectrometry of hydrogen/deuterium exchange in 70s ribosomal proteins from *e. coli*. *FEBS Lett.* 580, 3638–3642. doi:10.1016/j.febslet.2006.05.049

Yoshida, R., Ozaki, S., Kawakami, H., and Katayama, T. (2023). Single-stranded DNA recruitment mechanism in replication origin unwinding by DnaA initiator protein and HU, an evolutionary ubiquitous nucleoid protein. *Nucleic Acids Res.* 51, 6286–6306. doi:10.1093/nar/gkad389

Youngren, B., Nielsen, H. J., Jun, S., and Austin, S. (2014). The multifork *Escherichia coli* chromosome is a self-duplicating and self-segregating thermodynamic ring polymer. *Genes. & Dev.* 28, 71–84. doi:10.1101/gad.231050.113

Yu, I., Mori, T., Ando, T., Harada, R., Jung, J., Sugita, Y., et al. (2016). Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *eLife* 5, e19274. doi:10.7554/elife.19274

Zawadzki, P., Stracy, M., Ginda, K., Zawadzka, K., Lesterlin, C., Kapanidis, A. N., et al. (2015). The localization and action of topoisomerase IV in escherichia coli chromosome segregation is coordinated by the SMC complex, MukBEF. *Cell. Rep.* 13, 2587–2596. doi:10.1016/j.celrep.2015.11.034

Zechiedrich, E. L., Khodursky, A. B., and Cozzarelli, N. R. (1997). Topoisomerase IV, not gyrase, decatenates products of site-specific recombination in *Escherichia coli*. *Genes. & Dev.* 11, 2580–2592. doi:10.1101/gad.11.19.2580

Zhou, T., Zhang, R., and Ma, J. (2021). The 3d genome structure of single cells. *Annu. Rev. Biomed. Data Sci.* 4, 21–41. doi:10.1146/annurev-biodatasci-020121-084709

frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 25 August 2023
DOI 10.3389/fcell.2023.1198359

OPEN ACCESS

EDITED BY
Michael Blinov,
UCONN Health, United States

REVIEWED BY
Johannes Fürnkranz,
Johannes Kepler University of Linz, Austria
David Covell,
National Institutes of Health (NIH), United States

*CORRESPONDENCE
Carlos F. Lopez,
clopez@altoslabs.com

RECEIVED 01 April 2023
ACCEPTED 07 August 2023
PUBLISHED 25 August 2023

CITATION
Glazer BJ, Lifferth JT and Lopez CF (2023), Automatic mechanistic inference from large families of Boolean models generated by Monte Carlo tree search. *Front. Cell Dev. Biol.* 11:1198359. doi: 10.3389/fcell.2023.1198359

Automatic mechanistic inference from large families of Boolean models generated by Monte Carlo tree search

Bryan J. Glazer[1], Jonathan T. Lifferth[2] and Carlos F. Lopez[3,4]*

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, United States, [2]Department of Human Genetics, Vanderbilt University, Nashville, TN, United States, [3]Department of Biochemistry, Vanderbilt University, Nashville, TN, United States, [4]Altos Labs, Redwood City, CA, United States

Many important processes in biology, such as signaling and gene regulation, can be described using logic models. These logic models are typically built to behaviorally emulate experimentally observed phenotypes, which are assumed to be steady states of a biological system. Most models are built by hand and therefore researchers are only able to consider one or perhaps a few potential mechanisms. We present a method to automatically synthesize Boolean logic models with a specified set of steady states. Our method, called MC-Boomer, is based on Monte Carlo Tree Search an efficient, parallel search method using reinforcement learning. Our approach enables users to constrain the model search space using prior knowledge or biochemical interaction databases, thus leading to generation of biologically plausible mechanistic hypotheses. Our approach can generate very large numbers of data-consistent models. To help develop mechanistic insight from these models, we developed analytical tools for multi-model inference and model selection. These tools reveal the key sets of interactions that govern the behavior of the models. We demonstrate that MC-Boomer works well at reconstructing randomly generated models. Then, using single time point measurements and reasonable biological constraints, our method generates hundreds of thousands of candidate models that match experimentally validated *in-vivo* behaviors of the *Drosophila* segment polarity network. Finally we outline how our multi-model analysis procedures elucidate potentially novel biological mechanisms and provide opportunities for model-driven experimental validation.

KEYWORDS

MCTS algorithm, Boolean model, model inference, *Drosophila* development, segment polarity network, multi-model inference

1 Introduction

Technological advances in high throughput sequencing have significantly increased the amount of data available to biologists. However, the systems of molecular interactions that generate many cellular phenotypes remain poorly understood. This lack of understanding is a particularly pressing problem for diseases such as cancer, in which small genetic perturbations can have drastic clinical consequences. In order to understand and potentially intervene in the mechanisms by which cellular systems become dysregulated, one must first create a hypothesis of the system's interactions.

Given the complexity and non-linearity of many biological systems, computational models are a key tool for hypothesis generation and testing, allowing *in silico* perturbation and experimentation. Much previous work has shown the value of computational models of cellular systems for both understanding mechanisms and predicting cellular response to perturbation (Sáez-Rodríguez et al., 2007; Schlatter et al., 2009; Béal et al., 2019).

However, manually creating these computational models can be time consuming and difficult for several reasons. First, selecting a set of interactions that lead to the desired behavior is challenging due to the vast number of possible interactions. Further, introducing a new interaction can create feedback loops that change the model's behavior in unintuitive ways. Finally, data is often limited, only covering a limited set of conditions. Thus, many possible model configurations may have behavior that matches the (limited) data equally well. In order to have a reasonable chance of finding a model that captures an accurate representation of the biological system, including in conditions outside the given data, one must create many models.

Thus, automated model synthesis is desirable as it alleviates the difficulty of manually constructing a wide variety of models that are consistent with data. However, an efficient search algorithm is required to synthesize data-consistent models from the vast space of possible Boolean models. In this work, we focus on automatic synthesis of Boolean models (Kauffman, 1969).

Approaches to inferring Boolean models with data-consistent behavior can be divided into two categories: constraint solving and optimization. Constraint solving based methods pose the problem as a series of mathematical constraints, e.g., that the update functions must be consistent with steady states described in the data. These constraints are typically encoded as Boolean logic equations or in a more abstract formalism such as answer set programming (ASP) (Chevalier et al., 2020; 2019) or satisfiability modulo theories (SMT) problems (Fisher et al., 2015; Yordanov et al., 2016). Specialized solvers then find a set of models which satisfy all the constraints specified by the data and the modeling assumptions.

Optimization methods use general purpose discrete optimization algorithms to generate Boolean models, which are then scored according to a user-defined objective function (incorporating, e.g., similarity to data or model complexity). The optimization algorithms then generate new models which are variations of the best scoring models (Terfve et al., 2012; Lim et al., 2016).

Inspired by recent work in reinforcement learning for games, which also have combinatorially large search spaces, we investigate Monte Carlo Tree Search (MCTS) for Boolean model synthesis. Our method uses MCTS to iteratively build Boolean models by adding interactions to the model's update rules, similar to the way this algorithm is used to select moves in the games of chess or Go (Gelly et al., 2006). We show that MCTS works well for a wide variety of input data and model structures by testing the algorithm's ability to recover randomly generated Boolean models. Further, we show that it works for a more biologically realistic scenario: generating multi-cellular models of the *Drosophila* segment polarity network. Our method generated hundreds of thousands of models of the segment polarity network that are all consistent with experimental observations.

Having created a large collection of data-consistent models, one must derive some insight into the key interactions or mechanisms which drive their behavior. This is itself a challenging pattern recognition problem, which we address by developing data driven methods to extract mechanisms from models. Specifically, we present methods for clustering models based on the structure of their interactions. Using the structural clustering, our methods reveal the key interactions that control model behavior. We use this analysis to develop a novel hypothesis for the mechanism of regulation of the *wg* gene by isoforms of *CI* in *Drosophila*.

We call this pipeline of automated model generation and mechanism exploration MC-Boomer, or Monte Carlo Boolean Modeler.

Our method differs from previous approaches in several key ways. First, we use a heuristic optimization method, in contrast to linear programming or satisfiability solver based approaches. This allows us to trivially encode more complex model dynamics (e.g., multi-cellularity) and constraints on the form of update rules. Further, our optimization approach requires simulation of all models, giving us a view into the state spaces of our models. This allows us to characterize models according their behavior between initial conditions and steady states, yielding greater insight into populations of models that all have similar steady states. This comes at the cost of greater required computational resources compared to methods based on specialized constraint solvers. However, our method is trivially parallelizable, which we exploit to find large numbers of data-consistent models in a reasonable time frame. Finally, our optimization based approach immediately generates models that are partial matches to the experimental data. In contrast, constraint solvers may neglect useful models that do not perfectly satisfy constraints, even when those constraints are mis-specified or based on noisy data. In the worst case, constraint solvers may yield zero models after a lengthy search, while our approach yields a spectrum of models of varying complexity and goodness of fit to the data.

More generally, the computational problem that MC-Boomer solves can be framed as follows: Boolean models are comprised of mathematical, logical equations that are instantiated and simulated as computer programs. Our approach constructs the update equations of a Boolean model, simulates its behavior, and compares this behavior to a reference dataset. Following this definition, Boolean model synthesis can also be considered a particular form of the more general problems of program synthesis or symbolic regression. These fields are concerned with generating programmatic or mathematical expressions whose behavior is consistent with a given data set. More broadly, this fits into the category of non-linear discrete optimization problems. Consequently, we note MCTS has been shown to perform well for program synthesis, comparable to established search algorithms such as genetic programming (Lim and Yoo, 2016). Further, previous empirical comparisons of MCTS and genetic algorithms in two discrete optimization problems show that while MCTS is not strictly better performing, it does produce good results more quickly (Höfer, 2020) and produces more diverse solutions (Bosc et al., 2018) than genetic algorithms. These two features of MCTS are critical in allowing MC-Boomer to generate a large number of diverse Boolean models of biological systems. This is a key advancement of MC-Boomer compared to conceptually similar

TABLE 1 Example Simulation of Boolean Model. This shows the states of a four step simulation of the Boolean model shown in Equation 1.

	t = 0	t = 1	t = 2	t = 3
A	1	0	1	1
B	0	1	1	1
C	1	1	1	1

optimization based approaches to Boolean model synthesis such as BTR (Lim et al., 2016) and PRUNET (Rodriguez et al., 2015). These previous approaches to Boolean model synthesis focus on finding a single model with a good fit to the data. In contrast, the efficiency of MCTS allows MC-Boomer to find large numbers of models that fit the data well. Thus, we are able to make inferences about possible mechanisms of biological systems that are based on families of thousands of potential models. Another previous approach (Saez-Rodriguez et al., 2009) considers the relative probabilities of individual interactions, based on the whole population of data-consistent models. However, we investigate model structures with more sophisticated and fine-grained analyses, such as structure-based clustering and clustering interpretability methods.

2 Boolean models

Here we provide a brief introduction to Boolean models. Boolean models are two-state, discrete dynamical systems, with the state update equations defined by Boolean logic. We provide a simple example below, which has three species and their corresponding update rules.

$$\begin{aligned} A^{t+1} &= B^t \ and \ C^t \\ B^{t+1} &= C^t \\ C^{t+1} &= A^t \ or \ not \ B^t \end{aligned} \tag{1}$$

Each node has a state, which can be false or true (equivalently zero or one). The next state of the system, at time $t + 1$, is determined by the value of the update equations applied to the current state (time t) of the species. Updating every state at every time step is called synchronous updating. Repeatedly applying synchronous updates gives a simulation trajectory, which is guaranteed to converge to an attractor state or a cyclic attractor (Albert et al., 2008). An attractor is a fixed point: a state which does not change after the update equations are applied. A cyclic attractor is a cycle of states, which periodically repeats as the update equations are applied. Applying synchronous updating to the example three species model (with an arbitrary initial state) gives the four step simulation trajectory as shown in Table 1, with the last two steps representing an attractor with all node states equal to one.

We restrict the form of our Boolean update functions to only "dominant inhibition", having the form:

$$x^{t+1} = (act_0^t \text{ or } act_1^t \text{ or } \ldots act_n^t) \text{ and not } (inh_0^t \text{ or } inh_1^t \text{ or } \ldots inh_m^t)$$

Here, x is the species in the model that will be updated, $act^t_{1\ldots n}$ and $inh^t_{1\ldots m}$ are the states (at time t) of other species in the network that regulate the target node. Both act_i and inh_i can be a single species or composites of two or more species connected by an *and* clause, e.g., (a and b). A node is activated at $t + 1$ only if one or more of its activators is active and no inhibitors are active at t. In the rest of the paper, we use green arrows to show activating interactions and red arrows to show inhibiting interactions in model figures.

The goal of our framework, MC-Boomer, is to automatically generate these models so that their attractor states are similar to observed or reference gene expression data.

3 Methods

Here we describe the components of our framework for automated generation and exploration of mechanistic hypotheses, which we call MC-Boomer (Monte Carlo Boolean Modeler). As shown in Figure 1, our framework consists of three steps: gathering data and prior knowledge (Figure 1, left), using Monte Carlo Tree Search to generate and test model hypotheses (Figure 1, middle), and finally analyzing the model collection using data-driven and multi-model inference approaches (Figure 1, right). The first step involves collecting data describing the state of a biological system (e.g., RNA or protein expression), as well as delineating constraints on the possible interactions between components of the biological system. In this section, we primarily describe the second step, the algorithmic components involved in generating models. We describe the third step, analysis of the models generated by MCTS, in more detail in the Results (Section 4), as part of our analysis of the segment polarity network.

We separate our discussion of model generation (Figure 1, middle) into three sections: simulation, scoring, and search. We simulate our models with Boolean update rules, introduce a novel edit distance for scoring, and use Monte Carlo Tree Search (MCTS) for search. Below we will describe each component in more detail.

3.1 Simulation

A Boolean model is composed of logic rules that determine the state of each species in the system at the next step. We use synchronous updating which updates the state of every species of the model at each step. Synchronous updating is deterministic and is guaranteed to reach either a single stable attractor state or a sequence of periodically repeating states, called a cyclic attractor (Albert et al., 2008). We detect both stable and cyclic attractors by recording the simulation state history and halting the simulation when the current state matches a previously simulated state.

Each Boolean model generated by MCTS is simulated once from each initial state specified by the user. Each simulation proceeds from its initial state until it converges to an attractor state (s_i). This attractor state s_i is represented by a bit vector containing the Boolean state (0/1, False/True) of each species in the model. Each initial state may converge to a unique attractor or several may converge to the same attractor. Thus, each attractor state observed in the simulations has an occurrence count (c_i^M), indicating the number of initial states which converge to this attractor. Similarly, the states observed in the reference data set must have associated occurrence counts (c_j^D) indicating the number of times they were observed in the data.

A more comprehensive review of simulating biological systems with Boolean networks can be found in Albert et al. (2008).

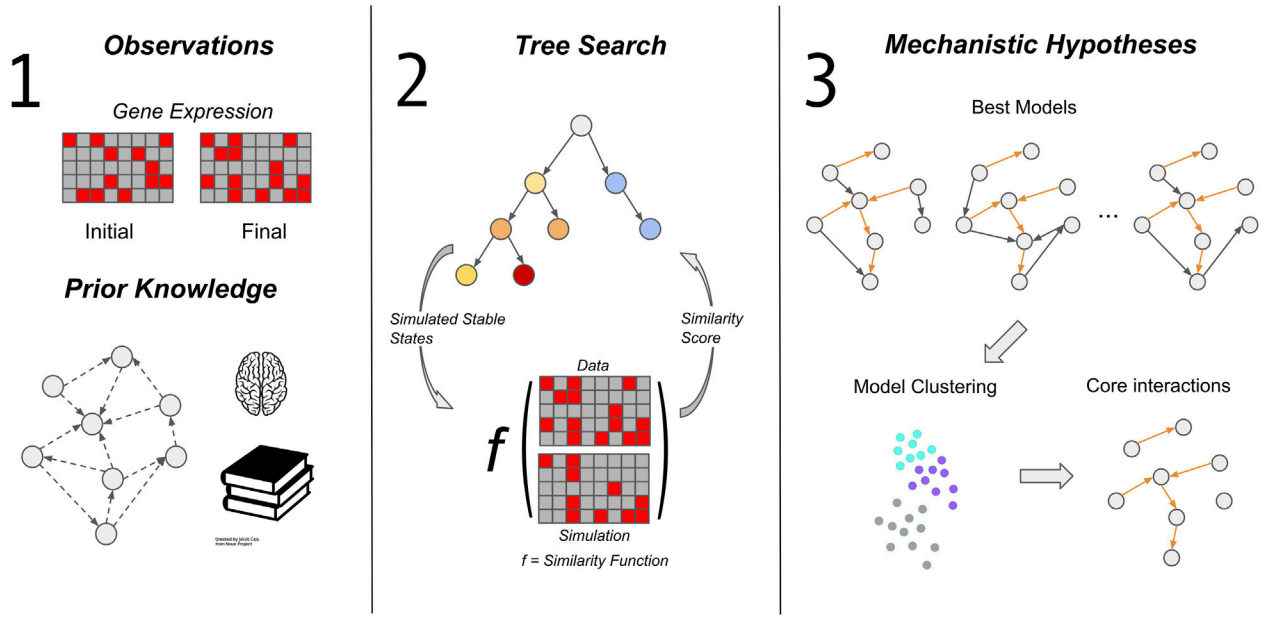


FIGURE 1
MC-Boomer workflow. The MC-Boomer workflow consists of three steps. The first is to gather data of the steady state expression levels of the genes of interest. Additional prior knowledge about known relationships between genes can guide and constrain the second step, model generation. We use Monte Carlo Tree Search to generate models (Section 3.3; Figure 3). The objective of search is to find models that have simulated attractor states that are similar to data, as measured by an edit distance, described in Section 3.2. We test this algorithm's ability to recover random models in Section 3.4. We further apply the method to a more biologically realistic model: the *Drosophila* segment polarity network, described in Section 4.1. We analyze the generated segment polarity models further in Sections 4.2 to extract structural features.

3.2 Scoring

We implemented an edit distance that compares reference data to model steady states (described in Figure 2). This distance is used to guide the MCTS search algorithm towards models that generate steady states that are similar to the data.

Given a model, we simulate it as described above in Section 3.1. This yields a set of attractor states and their occurrence counts. In addition, we assume that the user has provided a set of reference states and observation counts, derived from data or other observations. Our similarity score calculates the total number of state changes that are needed to transform the simulated attractor states to be equivalent to the reference states and occurrence counts. We describe our algorithm for calculating this similarity score below.

At each step of the distance calculation, we calculate the cost of transforming (i.e., editing) each simulated attractor state into each reference state. An edit consists of changing the value of the species in an simulated attractor state so that the simulated state becomes equivalent to a state in the reference. The size of an edit is the Manhattan distance between the bit vectors representing the state of the individual species in each attractor, i.e., how many species have different values in the simulated and reference states. The total "cost" C of an edit is the size of the edit multiplied by the difference in the occurrence counts between the simulated and reference states. This gives the total number of state bit vector changes required to transform a simulated attractor state into a reference state. At each step in the scoring algorithm we apply the edit with minimum cost. We apply the minimal cost edit by changing the count of the edited simulation state and increasing the total cost by C. We then repeat the process until all occurrence counts are equal between simulation and data. By accumulating edit costs at each step we obtain a total edit distance between simulated and measured attractor sets. This is normalized between (0,1) by dividing by the maximum possible edit distance $|s_i| \cdot N_c$, where $|s_i|$ is the number species in the model and N_c is the sum of occurrence counts.

$$\mathcal{D}_{edit}\left(A^{sim}, A^{obs}\right) = \frac{\sum_{k=1}^{N_e} C_k}{|s_i| N_c}$$

where N_e is the number of edits required, and C_k is the cost of the edit at step k. A graphical example of the edit distance calculation is shown in Figure 2.

Note that algorithm described above assumes that the simulated attractors are each single state attractors, rather than cycles. When the model reaches a cyclic attractor state, we simply average all the states in the cycle to obtain a single non-binary fractional state, which is then used normally in the scoring algorithm. We justify the choice to average cycles by assuming that the measurements used as inputs for MC-Boomer are noisy snapshots of cellular states. Thus, an average is a reasonable representation for multiple measurements of variable, noisy processes.

3.3 Monte Carlo tree search

The core task of MC-Boomer is generating the update rules of Boolean models such that the simulated attractor states of the

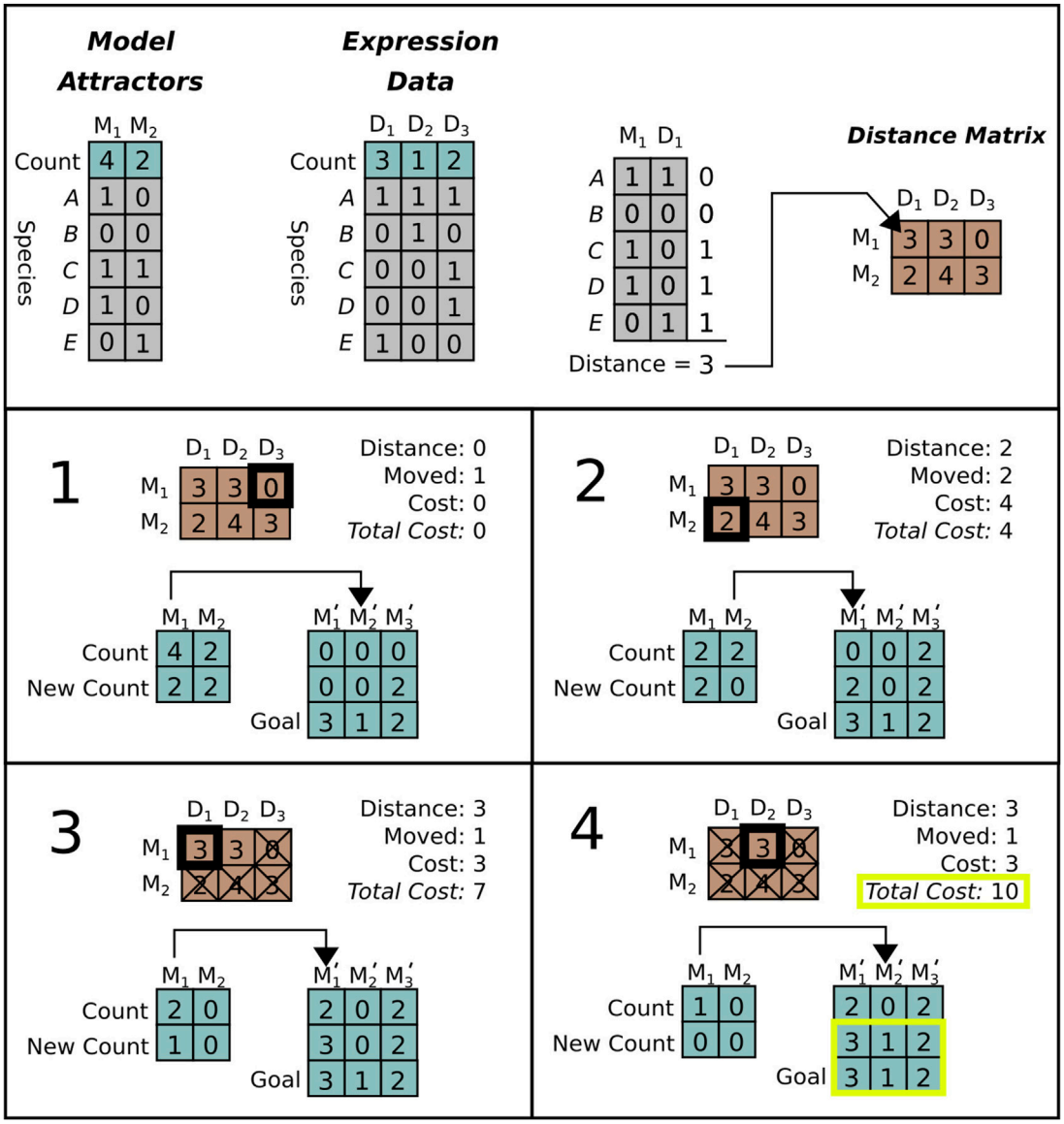


FIGURE 2
Distance calculation for a system with five genes. *Top row:* Model attractors (left) are generated from model simulations. The model attractors will be compared to attractors derived from the data (e.g., RNA expression) (middle left). Note that the data has three unique attractor states denoted D_i while the simulation only has two, denoted M_i. To calculate the first entry in the distance matrix (right) attractor states M_1 and reference states D_1 are compared. Differences are assigned a "1" while matches are assigned a "0." As shown, the distance between states M_1 and D_1 is "3" because they differ at three genes (C,D,E). *Bottom boxes:* Sequence of edits required to calculate the distance between simulation attractors (M) and data attractors (D). In the first step (1), we choose an edit by selecting the smallest valid distance from the distance matrix. This edit changes one of the M_1 attractors to D_2, but these are already identical, so the cost is zero. In step two (2) we select the next smallest distance (M_2 to D_1, with distance two) and change two attractors for a total cost of four. In step three (3) and four (4) we continue the same process. Note that in step three we remove multiple edits involving M_2 from consideration, as all of the available M_2 states have been edited already. In step four, the new state exactly equals D, so we halt the process with a final edit distance of ten.

generated models are similar to the observed, reference data. MC-Boomer employs Monte Carlo Tree Search (MCTS) to search the space of Boolean logic update rules. At each iteration, MCTS probabilistically selects a new term to add to one update equation of the model. In biological terms, this corresponds to adding an activating or inhibiting interaction between two genes in a regulatory network. MC-Boomer maintains a list of valid interactions between genes (update equation terms) that MCTS selects from, and this list is regenerated after each iteration so that MCTS can not add terms to the update equations that would result in biologically implausible or invalid models. See Section 3.1 for details on the mathematical form of the Boolean model update rules. After adding the new interaction to the model, MC-Boomer then simulates the model until it reaches a steady state (see Section 3.1) and compares its similarity to data (Section 3.2). Thus, the Boolean model update rules are constructed by adding individual interactions to the model, with the tree search guided by the simulation and similarity scoring of each model.

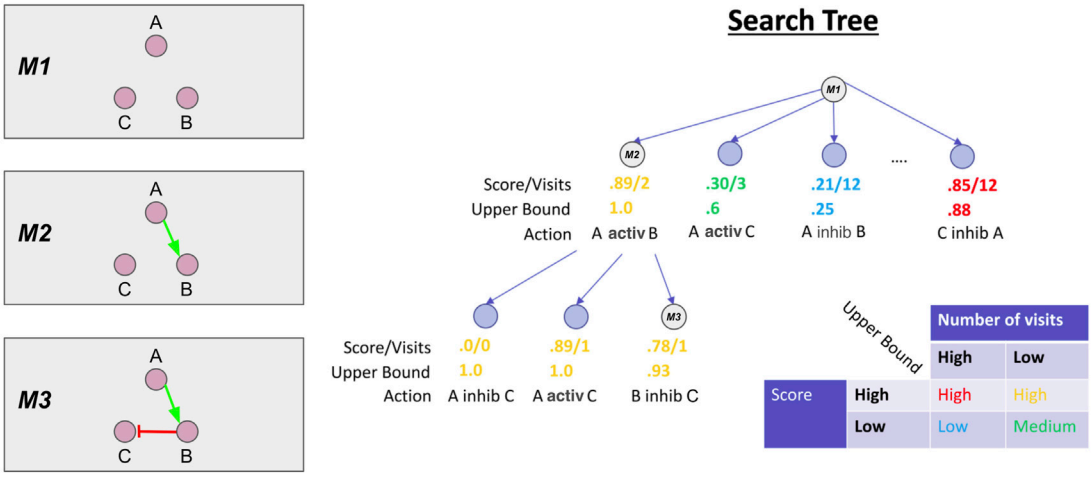


FIGURE 3
Monte Carlo Tree Search overview. On the left are the Boolean models corresponding to the branch of the search tree shown on the right, denoted M1, M2, M3. At each node in the tree, we also show the average score of models on the branch and the number of times the MCTS algorithm has visited the branch. These statistics are used to calculate the upper bound. In the bottom right, we show a conceptual overview of the functional form of the upper bound. In short, MCTS will aggressively explore branches with high scores but low number of visits. More exploration (i.e., a higher visit count) will progressively lower the upper bound until MCTS chooses another branch to explore.

Each unique combination of interactions is represented by a branch of the search tree. We show this graphically in Figure 3, where each branch of the search tree is annotated with the unique set of interactions that comprise the corresponding model. Multiple rule proposals are enumerated during the search (Figure 3 left, labels M1-M3). MCTS probabilistically chooses which branches to continue expanding, based on a statistical upper bound on the similarity score of models from each branch. The upper bound is called the Upper Confidence bound for Trees (UCT). The upper bound is approximated by tracking the number of times a branch has been explored (visit count) and the average similarity scores of models on each branch of the search tree. These statistics and an example upper bound are shown for each node in the search tree in Figure 3.

MCTS uses the upper bound to balance exploration of different rules *versus* exploitation of rules that have already produced high scores (Figure 3). The leftmost branch is relatively unexplored but models on that branch have high average similarity to the data. Thus, this branch has a high upper bound and the MCTS algorithm will preferentially explore and expand it. In contrast, the middle left branch has low average similarity scores but a low visit count so the upper bound is moderate, suggesting that MCTS may return to further explore this branch. The middle right branch has low similarity, but has been explored several times, yielding a very low upper bound. This effectively prunes the branch from the search, as the low upper bound corresponds to a low selection probability for further exploration. This pruning is not absolute, as MCTS will probabilistically explore all branches with a non-zero upper bound, given enough iterations. Finally, the rightmost branch has high scores, but has been visited many times, and so the upper bound is close to the average score.

We implemented several modifications to standard MCTS that have been shown to improve the algorithm's performance. Notably we used RAVE, a simple modification to the MCTS algorithm that shares value estimates of actions across all branches of the search tree (Gelly and Silver, 2011). Nested search uses the actions from the best random rollout to choose the next step, rather than selecting based on upper confidence bound (Rosin, 2011). Branch retention keeps the upper confidence bound from previous search iterations and reuses them for every subsequent search step. These methods are described in more detail in the Supplementary Material.

3.4 Validation experiments

We performed two experiments to demonstrate MC-Boomer for inferring Boolean models. In Section 3.4.2 and Section 3.4.3, we randomly generated Boolean models of various sizes, then tested MC-Boomer's ability to recover the structure and behavior of the random models. Then, in Section 4.1.1, we tested MC-Boomer's ability to recover the structure and behavior of the *Drosophila* segment polarity network, a complex multicellular model that accurately recapitulates key aspects of *drosophila* embryo morphogenesis (Albert and Othmer, 2003).

3.4.1 Random model generation

We first tested whether MC-Boomer could find models with a wide variety of behaviors and structures. We tested this by randomly generating models, simulating them, and then applying MC-Boomer to generate models matching their steady states. We randomly generated Boolean models with dominant inhibition update rules by sampling uniformly from a list of all possible interactions between sets of 8 or 16 species. Following this procedure, we generated 80 random networks.

Before testing MC-Boomer on the randomly generated models, we ensured that the attractor states of the random models had realistic, diverse characteristics. The attractors reached by the

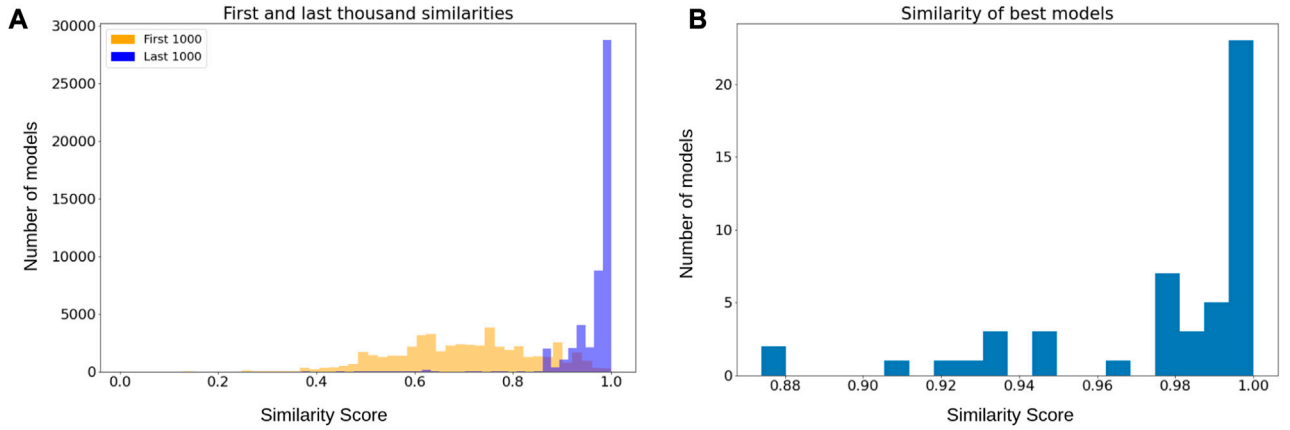


FIGURE 4
(A) Orange histogram depicts distribution of similarities from the first one thousand models sampled during the MCTS search. Blue histogram is the distribution of similarities from the last thousand models. The blue distribution is significantly shifted towards higher rewards, indicating that MCTS was systematically sampling good models. **(B)** Distribution of highest reward obtained by each independent search process. Most searches found models with >90% similarity.

random models do not collapse to an all active or inactive state, and instead have roughly one-third active species (as shown in Supplementary Table S2). We consider the characteristics of these attractors to be biologically relevant, similar to data that might be obtained from an experiment. Thus, good performance on these randomly generated models indicates that MC-Boomer can generalize to a realistic variety of input distributions.

3.4.2 Steady state behavioral similarity

We applied MC-Boomer to attempt to recover these random models using only their initial states and attractors as input data. Figure 4A shows that the models generated by MC-Boomer at the beginning of the search process poorly matched the behavior of the ground truth models. This is expected, as the MCTS algorithm is effectively a random search process during the initial steps. However, by the end of the search, MC-Boomer reliably found models that had steady states with high similarity to the ground truth models. Across all model sizes, MC-Boomer was able to find several exact behavioral matches, with a majority having >95% similarity, as shown in Figure 4B.

3.4.3 Rule set similarity

In addition to the steady state behavior of the models, we are also concerned with the content of the update rules generated by MC-Boomer. Many possible rule sets can have the same steady state behavior. However, many of these rule sets may be significantly different from each other and, most importantly, different from the underlying biological system. Under novel perturbations or conditions, these models may behave in radically different ways. Thus, we would like MC-Boomer to find models that match both the steady state behavior and the "interaction topology" of the underlying system. To validate MC-Boomer in this regard, we tested its ability to generate models with interactions that are similar to the reference models. In our tests, we quantified similarity by converting the update rules to sets of interactions for both the reference (randomly generated) model and the model generated by MC-Boomer. We then find the Jaccard index between the two interaction sets. This process is illustrated in Figure 5. Higher Jaccard indexes indicate that the MC-Boomer model matches the reference topology well.

With no restriction on the interactions selected by the model search process, MC-Boomer was able to find models with behavior that exactly matched the steady states of the reference models, but using rule sets that differed by as much as 80%. This corresponds to the left-most column of Figure 6, with zero reduction in search space, indicating that MC-Boomer was generating models using all possible interactions and no bias towards the true reference interactions.

We next investigated the effect of utilizing "prior knowledge" on MC-Boomer's ability to recover correct rules. As noted above, model inference is an underconstrained problem with many possible models having data-consistent behavior, and so ruling out infeasible interactions can reduce the number of spurious models. We simulated varying levels of prior knowledge by randomly removing incorrect interactions from MC-Boomer's action list, while retaining all of the correct interactions. We repeated the search five times, removing 10%, 25%, 50%, 75%, and then 90% of incorrect interactions from a set of 80 models. The aggregated Jaccard similarities for each percentage are shown in Figure 6. For models with both 8 and 16 species, increasing prior knowledge increased the Jaccard similarity to the reference data, as expected. Note that most protein-protein interaction databases are much sparser than our highest tested level of prior knowledge. For example, BioGRID (version 4.4.2021) has 26 k genes and 806 k interactions, which corresponds to a 99.9% reduction from all possible interactions (Oughtred et al., 2021). Thus, our tests simulate a very difficult scenario, relying on much less prior knowledge than is available in biochemical interaction databases.

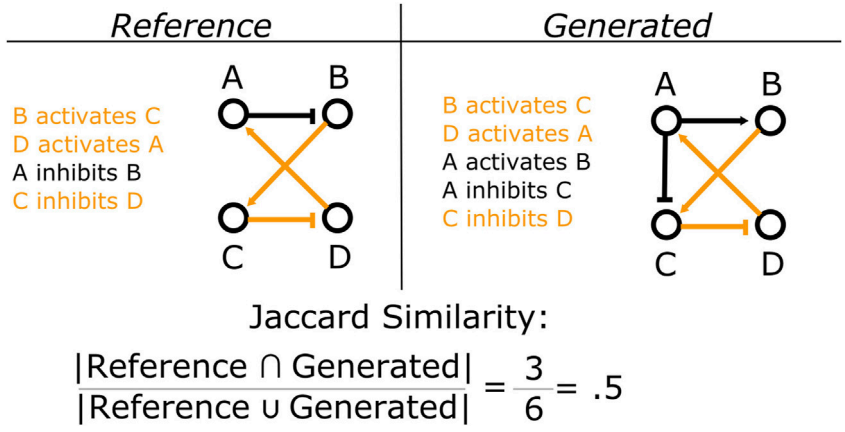


FIGURE 5
Example calculation of Jaccard similarity. We compare the structural similarity between two Boolean models by computing the Jaccard similarity between their sets of interactions. Here, shared interactions between the two models are colored orange, while interactions that are unique to each model are in black. In this example, the two models share three interactions in common, but have three more that are unique to each model. Thus they have a Jaccard similarity of 3/(3+3)=3/6=0.5.

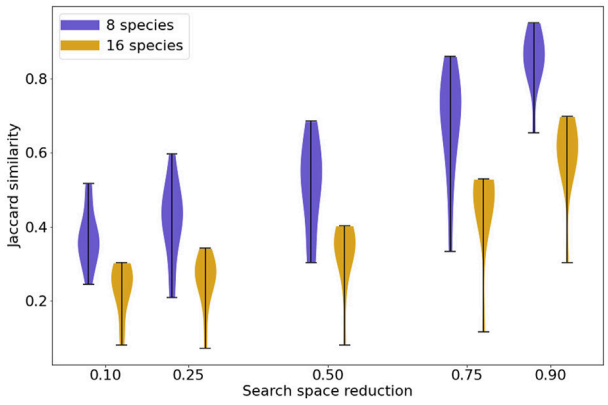


FIGURE 6
Jaccard similarity between synthetic reference and generated models with varying levels of prior knowledge. The violin plots show the distribution of Jaccard similarities achieved by MC-Boomer for synthetic models. The horizontal axis shows varying proportions of incorrect interactions randomly removed from the list of actions that MC-Boomer can choose when generating models. Removal of incorrect edges simulates the effect of prior knowledge, for example, using only interactions from a database of validated biochemical interactions. As expected, higher levels of prior knowledge lead to higher Jaccard similarities, as MC-Boomer has a higher probability of choosing correct interactions from a smaller list.

4 Results

Here we show the result of applying MC-Boomer to the segment polarity network (SPN). In Section 4.1.1 and Section 4.1.2, we describe the SPN and show MC-Boomer can generate models that are structurally similar to it, automatically discovering interactions that were previously manually selected by experts. Additionally we describe the large collection of alternate mechanisms generated by MC-Boomer, analyzing several in detail.

4.1 Segment polarity network (SPN)

As shown in the previous sections, MC-Boomer is able to generate models that are behaviorally and structurally similar to a variety of synthetically generated reference systems. While this was useful for validation, we also applied MC-Boomer to a more realistic setting to demonstrate the usefulness of the proposed framework. To that end, we employed MC-Boomer to build models of the *Drosophila* Segment Polarity Network (SPN), which is a gene circuit that controls the formation of borders and directionality of body segments during development of the *Drosophila* embryo. As a reference, we have chosen a well-studied model by Albert and Othmer (Albert and Othmer, 2003). Briefly, this model comprises 4 cells, with several distinct components, including genes, proteins, membranes, protein isoforms, and complexes. A diagram of the SPN interactions is shown in Figure 7 and a complete listing of the reference rules are shown in Supplementary Table S3. Albert and Othmer provided binarized expression levels for wild type conditions as well as three gene knockouts, shown in Supplementary Table S4. We applied MC-Boomer with these expression profiles to automatically generate models of the SPN.

4.1.1 Model generation

First, we will describe how we initialized the model and performed the search.

We applied several constraints to the search process so that MC-Boomer would only generate biologically plausible models. Membrane proteins (*WG*, *PTC*, *SMO*, *PH*, *HH*) could interact with membrane proteins only on adjacent cells. Internal proteins (*EN*, *SLP*, *CI*, *CIR*, *CIA*) could interact with other internal proteins, membrane proteins in the same cell, and genes in the same cell. Genes (*en*, *ci*, *ptc*, *hh*) could only activate their corresponding protein, and these gene-protein activating interactions were pre-specified in our search process. We generated all possible interactions that conform to these constraints, resulting in 334 possible interactions. We did not use any prior knowledge

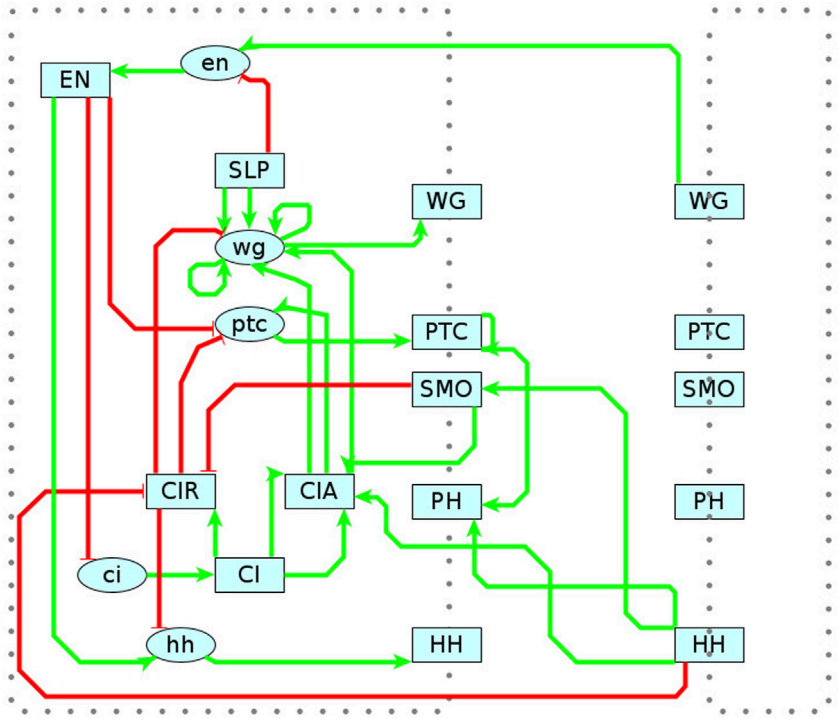


FIGURE 7
Reference Model for Segment Polarity Network. Diagram of the interactions in Albert and Othmer's model of the segment polarity network (Albert and Othmer, 2003). Green edges indicate activating interactions. Red are inhibiting. Lower case ovals indicate genes and upper case indicate proteins. The dotted border indicates the cell membrane, with membrane proteins straddling the border. On the right is the adjacent cell, with several interactions spanning between cells. Albert and Othmer's model has four cells with the same interactions inside each cell. Interactions between cells are symmetric, though only one direction is shown in the diagram to maintain clarity.

about the possible interactions beyond the basic biological knowledge described above, simulating a scenario in which a user does not bias the search to previously described interactions between genes. This tests MC-Boomer's ability to recover the reference model without assistance from biological prior knowledge, as well as its ability to generate novel, interesting hypotheses about the possible structure of the regulatory network.

All interactions added during the search process were repeated across all 4 cells. Multi-cellular membrane interactions were symmetric, added in both directions between neighboring cells.

The reference SPN model specified initial and stable states for the wild type network as well as initial and attractor states for knockouts of *wg*, *hh*, and *en* (see Supplementary Figure S4 for details).

We applied MC-Boomer to search for models that matched the behavior of the reference SPN model across the wild-type and three knockout conditions. At each search iteration, we simulated the model across all conditions, calculated the edit distance between simulated and reference steady states, then averaged all conditions' similarity scores to get a final score for the iteration. We implemented knockouts by removing all interactions to and from the *hh*, *en* and *wg* genes across all cells.

We ran 1,500 searches in batches of 30 in parallel on our institution's computing cluster. In each search step, MC-Boomer simulated 10 k model variations before adding the best interaction to the model and starting the next step. We restricted the search to terminate after 30 steps, but not before completing 8 steps. Every search was run with RAVE, nested search, and branch retention enabled with the same uniformly random sampled parameter distributions as in the synthetic data experiments. The complete search process took 41 h and simulated 430 million unique models. Eleven of the 1,500 search processes found models with exactly the same steady states as the reference model for all four conditions. Collectively, these eleven search processes generated $>202k$ models with perfect consistency to the attractor data.

4.1.2 Visualizing the set of data-consistent models

Given the large size of our collection of models with consistent steady state behavior, we were motivated to develop methods for visualization and exploration of large numbers of models.

First, we applied dimensionality reduction and clustering methods to visualize similarities between the models. We randomly sampled fifty thousand of the $202k$ data-consistent models and clustered them with the UMAP algorithm (McInnes et al., 2018) using the interaction set Jaccard distance between models, as illustrated in Figure 5. Model sampling was necessary because UMAP requires computation of a pairwise distance matrix that would have been infeasible for the full data set. Multiple different samples all gave similar results, thus giving us confidence that the sample analyzed here was representative of the overall model population.

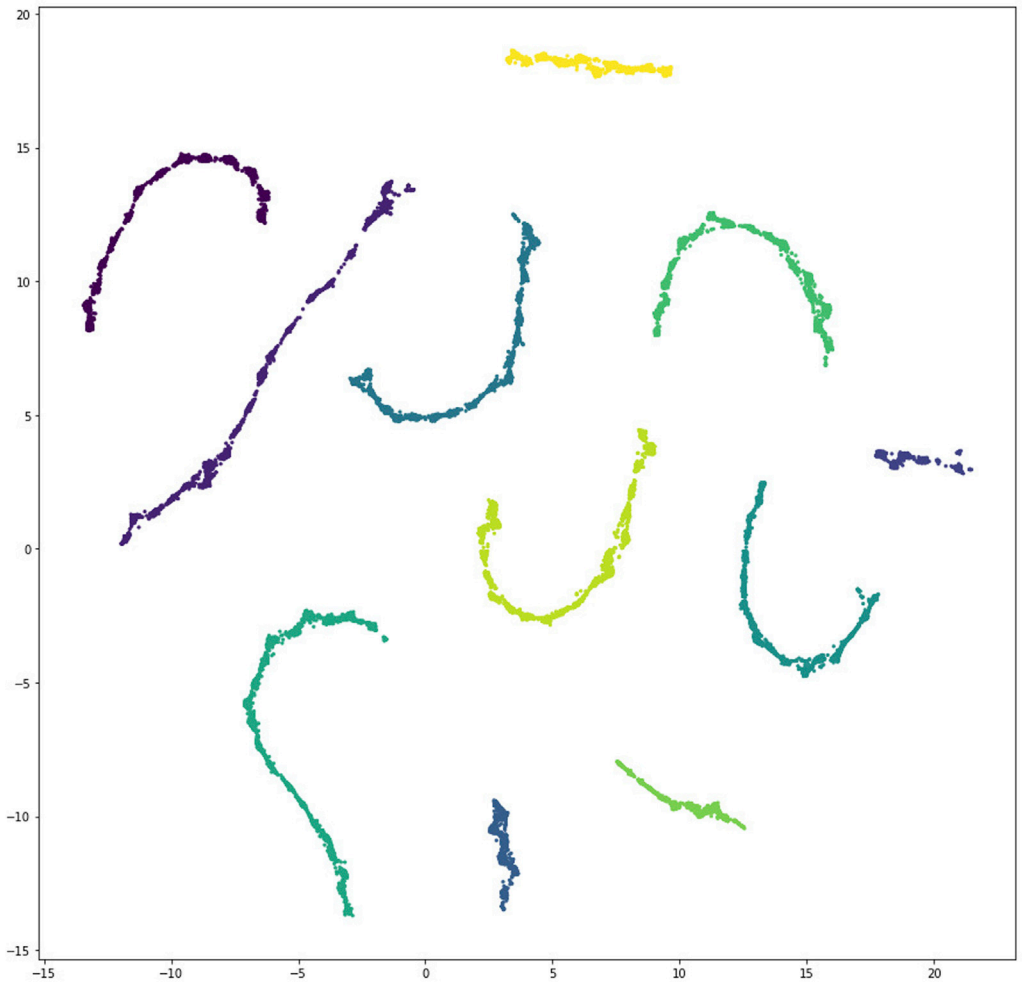


FIGURE 8
Scatter plot depicting clustering of unique data-consistent segment polarity models after UMAP projection to two dimensions. There are eleven well separated clusters, corresponding to the eleven independent search processes that found data-consistent models.

Applying UMAP with the Jaccard distance yielded the result shown in Figure 8 with eleven well separated clusters, corresponding to the eleven independent searches that produced data-consistent models.

4.1.3 Structural similarity between clusters and reference

We then compared the interactions in each MC-Boomer generated model with the interactions in the reference model's update rules to find the "structural" similarity.

The update rules of the reference model had 26 total interactions. We manually pre-specified eleven of the interactions in the reference segment polarity network. That is, all models generated by MC-Boomer included these interactions as "prior knowledge." This included all the interactions in which a gene activated its corresponding protein, as well as four interactions that did not fit the dominant inhibition dynamics of the rest of the network (Supplementary Figure S1). Our tests evaluated MC-Boomer's ability to discover models that included the remaining 15 interactions in the reference model.

Within each cluster of MC-Boomer models, we computed the mean, median, and maximum size of the intersection between the cluster's models' interactions and the reference model's interactions, as shown in Table 2.

Comparison across clusters revealed a wide disparity in accuracy, with cluster 3 having, on average, three rules in common with the reference SPN model. We note that while the models in cluster 3 had low structural similarity to the reference SPN model, all of the models in every cluster have the same steady state attractors as the reference. Cluster 7 had the highest average intersection, with several models in the cluster having 11 out of 15 rules in common with reference model. For cluster 7, we found the most common rules, i.e., those shared by $> 90\%$ of the models in the cluster. Figure 9A shows these common rules and Figures 9B,C shows "false positive" and "false negative" rules, respectively. False positive rules were present in MC-Boomer models but not in the reference and false negative rules were in the reference but not the MC-Boomer models. In the following sections, we investigate two of these interactions, one of which was not present in the reference model.

TABLE 2 Structural Intersection with Reference Model. For each cluster of models shown in Figure 8 we computed the intersection between these common rules and the reference model. We show the mean, median, and maximum intersection between each cluster's models and the reference. Cluster 7 has the highest intersection across all three statistics, while cluster 3 shares the fewest interactions with the reference. We further investigate the most common interactions in Cluster 7 in Figure 9.

Cluster	Intersection			Cluster size
	Mean	Median	Max	
0	4.80	5	8	5321
1	6.03	6	9	6566
2	6.32	6	9	1654
3	3.23	3	6	2057
4	6.87	7	9	5673
5	6.49	6	9	4805
6	3.30	3	6	6710
7	8.68	9	11	5493
8	5.22	5	8	2277
9	6.90	7	10	6413
10	8.27	8	11	3031

4.2 Mechanism identification

In the following sections, we investigate the specific interaction patterns or mechanisms that MC-Boomer generates. We first focus specifically on individual interactions that are present across all data-consistent models, proposing a novel hypothesis for the biological mechanism encoded by the Boolean logic of the interaction. Then, we propose data-driven methods to extract a diverse set of mechanisms from the large collections of models generated by MC-Boomer.

4.2.1 Investigating common mechanisms

The high-level clustering analysis shows that MC-Boomer generates models with a wide variety of structures but identical steady state behavior. However, this analysis is too broad to elucidate the precise nature of the mechanisms that these models use to generate this behavior. Accordingly, we more closely investigated two key interactions that are present in every model generated by MC-Boomer. Specifically, we consider "*EN* inhibits *ci*" and "*CIR* inhibits *CIA*", which are present in 100% of the data-consistent models.

First, we look at *EN* inhibiting *CI*, which is present in all of our models and also present in the reference model. This indicates that this interaction is a crucial link across the very diverse mechanisms

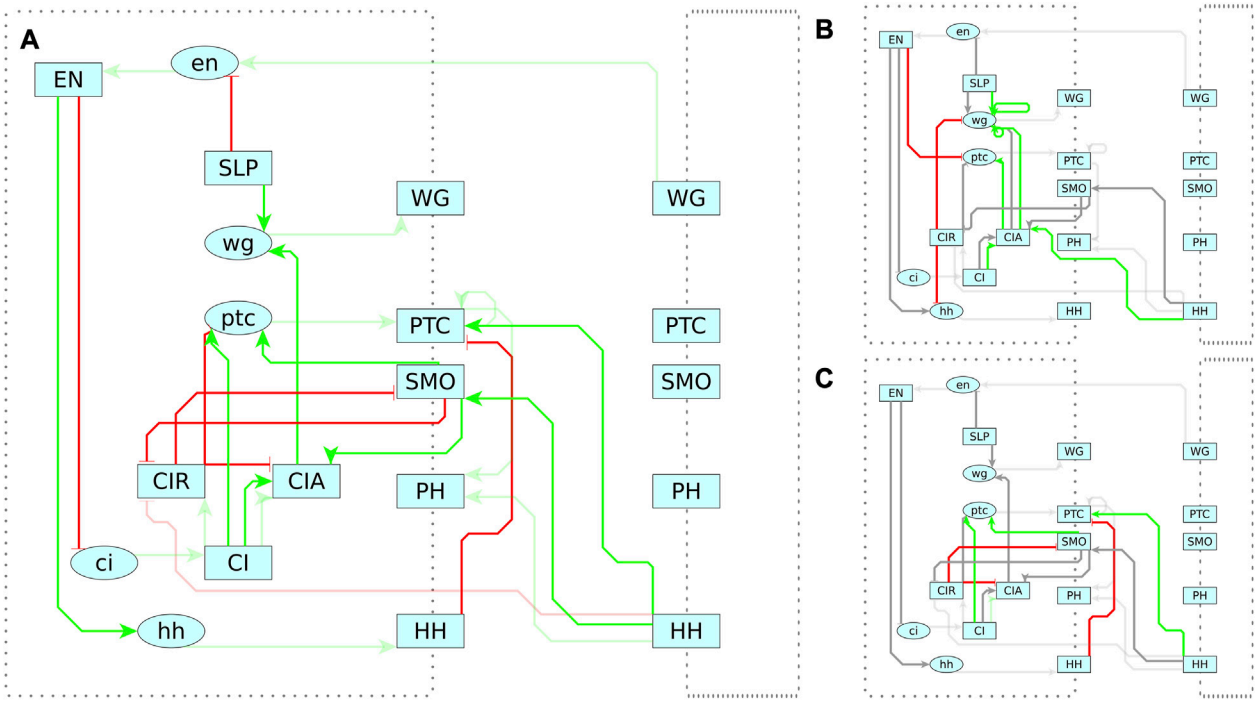


FIGURE 9
(A) Most common interactions in cluster 7. These interactions that are found in >90% of models in cluster 7. Transparent lines represent interactions that were pre-specified as prior knowledge, while more opaque interactions were generated by MC-Boomer. **(B)** Interactions highlighted in red and green are common in cluster 7 but are not present in the reference model. Grey interactions are shared between the reference model and cluster 7. Note the red inhibiting interaction between CIR and CIA, which is investigated in more detail in Section 4.1.2.**(B)** The reference model is shown here, highlighting interactions that are in not present in cluster 7. Again, note the red inhibiting interaction between CIR and wg. Section 4.1.2 proposes an alternative mechanism for inhibition of wg by CIR.

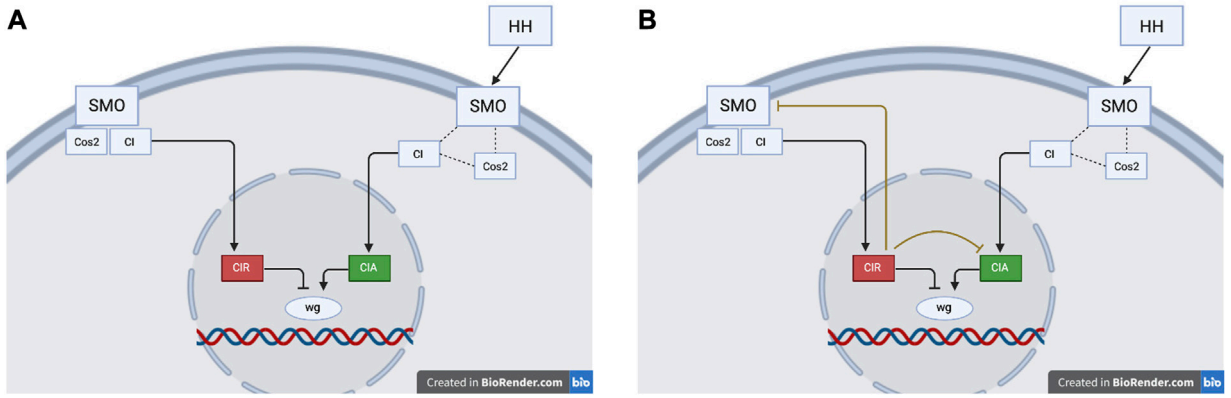


FIGURE 10
(A) The reference model depicts the modification of CI as a forked pathway, where the resulting product is determined by the activation state of SMO. In the reference model, active SMO promotes CIA and inhibits CIR. **(B)** The MC-Boomer model, in contrast, includes two novel interactions where CIR inhibits CIA and SMO. Figures partially based on Figure 3 from Hooper and Scott (2005)

employed by the eleven clusters of models and the reference model. Simulating a random sample of one thousand models with this interaction knocked out resulted in a 28% average absolute reduction in similarity to the reference steady state data. We observed that knocking out the *EN* to *ci* interaction in the reference model also reduced similarity to the reference data by 28%. Again, this indicates that the while the models are structurally diverse, they share a similar reliance on this particular interaction of *EN* and *ci*.

On the contrary, CIR inhibition of CIA is not present in the reference model. This interaction is shared by more than two hundred thousand unique models generated by MC-Boomer. The high frequency of the CIR inhibiting CIA interaction motivated further investigation into CIR and CIA's role in regulation of the wg gene.

To provide necessary background for our discussion of wg regulation, we briefly describe the key genes in this pathway. CIA is an activated, nuclear transported form of the CI protein, while CIR is a proteolytically cleaved form of CI which represses wg transcription. In the absence of HH, SMO forms a complex with CIA and Cos2, a kinesin-like protein that binds and sequesters CIA, preventing its nuclear translocation and permitting its cleavage into CIR. In the presence of HH, SMO is activated and Cos2 releases CIA, which is then transported to the nucleus, where it activates wg (Lum et al., 2003; Kalderon, 2004; Ranieri et al., 2012). The exact mechanisms and network dynamics behind CI activation, cleavage, and nuclear translocation have long remained a point of debate and uncertainty (Ruel et al., 2003).

In addition to CIR inhibiting CIA, MC-Boomer also suggests (in 34% of models) an inhibitory interaction between CIR and SMO. The novel inhibition of CIA and SMO by CIR can be interpreted in at least two ways, as shown in Figure 10.

1) These interactions do not represent real signaling mechanisms. In accordance with the reference model, the bi-directional inhibitory loop between CIR and SMO may simply reflect the normal activation states of these proteins. When SMO is active, CIR cannot be produced because SMO destabilizes Cos2 and therefore all CI is available as CIA. Conversely, when SMO is inactive, Cos2 binds CI and conversion to CIR occurs. Therefore, the inhibition of CIA and SMO by CIR may not represent genuine biochemical interactions, but may simply be artifacts of MC-Boomer's automated model generation process.
2) These interactions do represent real, redundant signaling mechanisms. The novel inhibition of CIA and SMO by CIR may represent redundant signals which prevent the possibility of competition at the target gene binding site. This type of redundancy is a feature observed in other biological signaling networks (Albert et al., 2011) CIR inhibition of CIA and SMO in the cytosol ensures that CIR can bind and inhibit wg in the nucleus without interference from CIA. In this interpretation, CIR is not just a passive cleavage product, but also an active participant in a feedback loop that inhibits the activity of CIA.

This second interpretation describes an instance of signaling redundancy. If CIR inhibits SMO and CIA, this helps to ensure a full transition between on and off network states and prevents any potential binding competition at the target gene.

Overall, these observations show that the proposed method can both reproduce the known biological features as well as provide novel insight into the segment polarity network by generating new mechanistic hypotheses, which require further investigation through experiments.

4.2.2 Identifying unique mechanisms in model clusters

We are able to analyze these two interactions in detail because they are shared across all models and their limited scope eases their interpretation. However, our clustering analysis showed that there

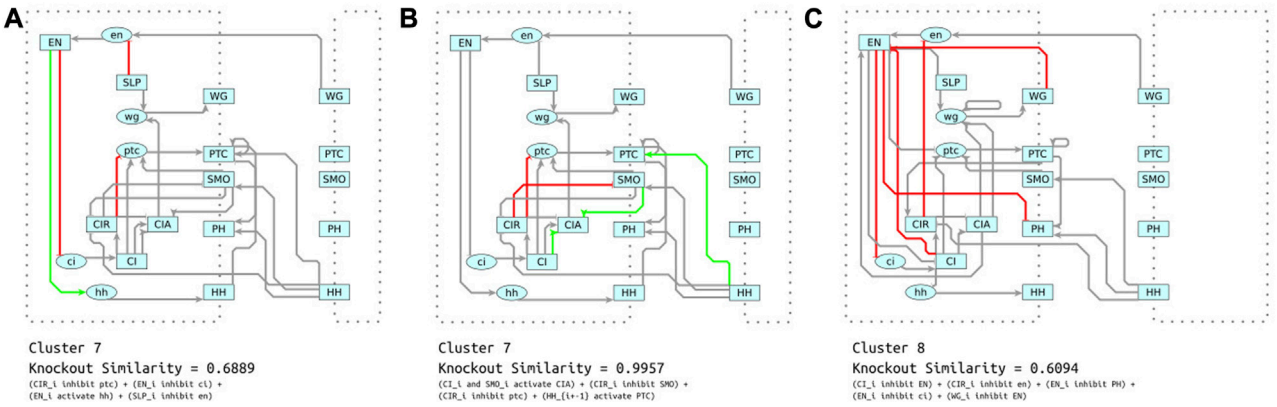


FIGURE 11
(A) Shown in red and green is the maximally disruptive set for cluster 7. These are interactions that are common in cluster 7, but are very uncommon in other clusters. Additionally, knocking out these interactions reduces similarity to the reference steady states more than other sets of common interactions (shown in grey). **(B)** Another distinguishing set of interactions, but these are minimally disruptive. Knocking them out only reduces similarity to the reference data by a negligible amount. **(C)** Maximally disruptive set for cluster 8. These reduce similarity to reference data to a similar degree as the most disruptive set in cluster 7, but these interactions utilize a different mechanism.

are at least eleven groups of models with widely differing structures. Accordingly, we also investigated the role of interactions that are specific to individual clusters of models. We searched for sets of up to 5 interactions that are present in a high proportion of models in each cluster, while not being present in models in other clusters. We call these distinguishing sets. We found between 25 and 21,570 distinguishing sets per cluster.

Given the large number of distinguishing sets for some clusters, we needed a measure of which sets are most important to the function of the models in the cluster. We quantified this by simulating knock outs of each distinguishing set in a sample of 100 models from their respective clusters and calculating the reduction in similarity to the reference data caused by the knockouts. We refer to distinguishing sets with the largest reduction in similarity as the maximally disruptive sets. These maximally disruptive sets identify the unique mechanisms that the models in each cluster most highly rely on to generate their behavior.

Comparing the interactions in the maximally disruptive sets revealed heterogeneity across the clusters. Most of the maximally disruptive sets shared two or fewer interactions in common. For example, the maximally disrupting sets for cluster 7 (Figure 11A) and cluster 8 (Figure 11C) only share a single interaction in common. Simulated knockouts of cluster 7 and 8's maximally disruptive sets reduced similarity to reference data by 31% and 39%, respectively. This indicates that models in these two clusters depend, to a similar degree, on these distinct sets of interactions for generating correct behavior. Inspection reveals that while the two mechanisms are not similar by a direct comparison, they share functional similarity in primarily modulating the connectivity and activity of *EN*. This corresponds with our previous analysis showing that *EN* interactions are crucial for correct model behavior across our whole collection of models. However, the actual mechanism by which *EN* activity is directed is quite distinct. The interactions in cluster 7 (Figure 11A) give *EN* a mixed activating/inhibiting role, while cluster 8 (Figure 11B) relies on several inhibitory feedback loops centered on *EN*.

Similar to the case of CIR described in Section 4.2.1, many of the distinguishing sets do not have any effect on the behavior of the model; one such example is illustrated by Figure 11B. One perspective is that these interactions are redundant and only increase the complexity of the model. Accordingly, several previous approaches (Terfve et al., 2012; Lim et al., 2016) penalize models with more interactions. Another perspective is that these redundant connections may confer robustness, i.e., an ability to recover from aberrant initial conditions or losses of function, or as with CIR they could help ensure full response to inhibition or activation.

5 Discussion

Biology is inherently complex, yet our measurements capture only a limited slice of the true activity within a cell. Current assay technology can only describe a subset of biomolecules at low time resolution and with significant noise. From this blurry view researchers must synthesize a model that can both describe the phenomena under investigation and predict the system's behavior in novel circumstances. Synthesizing a model can be made easier by choosing the simplicity of the Boolean logic modeling formalism to represent the system. Nonetheless, even for a small number of interacting species, the number of possible Boolean models is vast. Consequently, a typical researcher, creating models through trial and error, may only find one or perhaps a few models whose behavior is consistent with the observed data. However, as we have shown in Section 4.1.1, even in a small system with multiple measurements and

reasonable prior assumptions on model structure, there are hundreds of thousands of models that are all consistent with the data.

This observation was made possible by using an efficient search technique, Monte Carlo Tree Search, to build models. We demonstrate the power of MCTS to synthesize models with the correct steady-state behavior and the correct interactions in Section 3.4. While previous studies have shown that similar optimization methods (e.g., tabu search in Aghamiri and Delaplace (2020)) are effective for finding data-consistent models, they have focused on finding a single model that is "best" in terms of both complexity and fit to the data. In contrast, we retain every model that fits the data well and in Section 4.2.1 and Section 4.2.2 we develop a set of techniques for making sense of this large collection of models. We approach this from a data-driven perspective, in the sense that our MCTS algorithm generates data about the space of valid hypotheses. By clustering models based on their structural features, we can find recurrent motifs across the whole collection of models, as well as distinct motifs that discriminate the structure of groups of models. Simulated knockouts of these motifs then reveal that some are critical to the models' correct behavior.

5.1 Using MC-Boomer to design experiments

As we describe in Section 4.2.1, analysis of the models generated by MC-Boomer pointed us towards an alternate hypothesis for the mechanism by which *CIR* and *CIA* regulate expression of the wingless gene (*wg*) in the segment polarity network. An investigator using MC-Boomer to study this pathway may propose that CIA activation of *wg* depends on both SMO (Smoothened) stabilization and, as MC-Boomer suggests, the absence of CIR. The existence of these novel inhibitory relationships could be experimentally validated by introducing CIR into cells in which HH signaling has already activated SMO and CIA. Reduced concentrations of active CIA or SMO would indicate that CIR does, in fact, inhibit the activity of CIA and SMO.

5.2 Limitations and future work

Previous work (Fauré et al., 2006) has suggested that the general asynchronous updating scheme yields more biologically realistic results for Boolean network simulations. While our current approach uses synchronous updating, extending MC-Boomer to work with asynchronous updating would be straightforward.

The current approach is limited in its scalability to models with large numbers of interacting species by several key bottlenecks. First, this approach requires simulation of every synthesized model, and simulation becomes prohibitively expensive for large models. This could be alleviated through partial or approximate simulations of the models. While this would yield an approximation of the model's similarity to data, the UCT upper bound allows MCTS to tolerate some noise in the search process. Second, the search space scales exponentially with the number of species in the model. We show that restricting the search space through prior knowledge constraints on model structure is an effective strategy for improving structural and behavioral accuracy of synthesized models. The efficiency of the search algorithm could further be improved by using deep learning to guide MCTS. This is similar to the approach used by the AlphaZero algorithm (Silver et al., 2018), that proved to be exceptionally effective at searching the combinatorially large space of moves in games like chess and Go. We are currently exploring each of research directions as potential optimizations of the MC-Boomer algorithm.

6 Conclusion

Our work demonstrates that automated Boolean model inference can generate many alternative, hypothetical regulatory networks that each explain a systems's steady state behavior equally well. We observe that Monte Carlo Tree Search is effective at this task for both synthetic and real-world data, as it balances exploration of novel models with exploitation to generate multiple variations of high performing models. By using data analysis techniques on the huge collections of models that result from tree search, we find families of models and the core regulatory structures underlying their common behavior. Applying this analysis to a well known model of *Drosophila* development revealed previously known regulatory mechanisms as well as suggesting a novel role for the *CI* gene in *wg* regulation. This demonstrates that Boolean model inference should not be treated as a search for a single best performing model, but instead as a process of hypothesis generation and comparison.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

BJG developed the concept, wrote the software, performed analysis, and contributed to writing the manuscript. JTL performed model analysis, contributed to literature review and to writing. CFL provided guidance on concepts and analysis and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by U.S. National Library of Medicine grant number T15LM007450 (BG) and NSF CAREER award MCB1942255, National Institutes of Health (NIH) U01-CA215845 (CL).

Acknowledgments

We are grateful to Vito Quaranta for his commitment to this work. We thank Becca Creed for her help in submitting the paper.

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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1198359/full#supplementary-material

References

Aghamiri, S. S., and Delaplace, F. (2020). *TaBooN – boolean network synthesis based on tabu search*. arXiv:2009.03587 [cs, q-bio]. arXiv: 2009.03587.

Albert, I., Thakar, J., Li, S., Zhang, R., and Albert, R. (2008). Boolean network simulations for life scientists. *Source Code Biol. Med.* 3 (1), 16. doi:10.1186/1751-0473-3-16

Albert, R., DasGupta, B., Hegde, R., Sivanathan, G. S., Gitter, A., Gürsoy, G., et al. (2011). Computationally efficient measure of topological redundancy of biological and social networks. *Physical Review E.* 84 (3), 036117. doi:10.1103/PhysRevE.84.036117

Albert, R., and Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J. Theor. Biol.* 223 (1), 1–18. doi:10.1016/s0022-5193(03)00035-3

Béal, J., Montagud, A., Traynard, P., Barillot, E., and Calzone, L. (2019). Personalization of logical models with multi-omics data allows clinical stratification of patients. *Front. Physiology* 9, 1965. doi:10.3389/fphys.2018.01965

Bosc, G., Boulicaut, J.-F., Raïssi, C., and Kaytoue, M. (2018). Anytime discovery of a diverse set of patterns with Monte Carlo tree search. *Data Min. Knowl. Discov.* 32 (3), 604–650. doi:10.1007/s10618-017-0547-5

Chevalier, S., Froidevaux, C., Paulevé, L., and Zinovyev, A. (2019). "Synthesis of boolean networks from biological dynamical constraints using answer-set programming," in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), 34–41. ISSN: 2375-0197. doi:10.1109/ICTAI.2019.00014

Chevalier, S., Noël, V., Calzone, L., Zinovyev, A., and Paulevé, L. (2020). "Synthesis and simulation of ensembles of boolean networks for cell fate decision," in 18th International Conference on Computational Methods in Systems Biology (CMSB), volume 12314 of Lecture Notes in Computer Science (Online, Germany: Springer), 193–209.

Fauré, A., Naldi, A., Chaouiya, C., and Thieffry, D. (2006). Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinforma. Oxf. Engl.* 22 (14), e124–e131. doi:10.1093/bioinformatics/btl210

Fisher, J., Köksal, A. S., Piterman, N., and Woodhouse, S. (2015). "Synthesising executable gene regulatory networks from single-cell gene expression data," in Computer aided verification, *lecture notes in computer science*. Editors D. Kroening and C. S. Păsăreanu (Springer International Publishing), 544–560.

Gelly, S., and Silver, D. (2011). Monte-Carlo tree search and rapid action value estimation in computer Go. *Artif. Intell.* 175 (11), 1856–1875. doi:10.1016/j.artint.2011.03.007

Gelly, S., Wang, Y., Munos, R., and Teytaud, O. (2006). *Modification of UCT with patterns in monte-carlo Go*.

Höfer, D. (2020). Comparing MCTS with genetic Algorithms for optimizing multigrid methods. *Master's thesis*. FRIEDRICH-ALEXANDER-UNIVERSITÄT ERLANGEN-NÜRNBERG.

Hooper, J. E., and Scott, M. P. (2005). Communicating with hedgehogs. *Nat. Rev. Mol. Cell Biol.* 6 (4), 306–317. Number: 4 Publisher: Nature Publishing Group. doi:10.1038/nrm1622

Kalderon, D. (2004). Hedgehog signaling: costal-2 bridges the transduction gap. *Curr. Biol.* 14 (2), R67–R69. doi:10.1016/j.cub.2003.12.047

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* 22 (3), 437–467. doi:10.1016/0022-5193(69)90015-0

Lim, C. Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., et al. (2016). Btr: training asynchronous boolean models using single-cell expression data. *BMC Bioinforma.* 17 (1), 355. doi:10.1186/s12859-016-1235-y

Lim, J., and Yoo, S. (2016). "Field report: applying Monte Carlo tree search for program synthesis," in Search based software engineering, *lecture notes in computer science*. Editors F. Sarro and K. Deb (Cham: Springer International Publishing), 304–310.

Lum, L., Zhang, C., Oh, S., Mann, R. K., von Kessler, D. P., Taipale, J., et al. (2003). Hedgehog signal transduction via smoothened association with a cytoplasmic complex scaffolded by the atypical kinesin, costal-2. *Mol. Cell* 12 (5), 1261–1274. doi:10.1016/s1097-2765(03)00426-x

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: uniform manifold approximation and projection. *J. Open Source Softw.* 3 (29), 861. Publisher: The Open Journal. doi:10.48550/arXiv.1802.03426

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. A Publ. Protein Soc.* 30 (1), 187–200. doi:10.1002/pro.3978

Ranieri, N., Ruel, L., Gallet, A., Raisin, S., and Thérond, P. P. (2012). Distinct phosphorylations on kinesin costal-2 mediate differential hedgehog signaling strength. *Dev. Cell* 22 (2), 279–294. doi:10.1016/j.devcel.2011.12.002

Rodriguez, A., Crespo, I., Androsova, G., and Sol, A. d. (2015). Discrete logic modelling optimization to contextualize prior knowledge networks using PRUNET. *PLOS ONE* 10 (6), e0127216. Publisher: Public Library of Science. doi:10.1371/journal.pone.0127216

Rosin, C. D. (2011). *Nested rollout policy adaptation for Monte Carlo tree search*. IJCAI.

Ruel, L., Rodriguez, R., Gallet, A., Lavenant-Staccini, L., and Thérond, P. P. (2003). Stability and association of smoothened, Costal2 and fused with cubitus interruptus are regulated by hedgehog. *Nat. Cell Biol.* 5 (10), 907–913. Number: 10 Publisher: Nature Publishing Group. doi:10.1038/ncb1052

Saez-Rodriguez, J., Alexopoulos, L. G., Epperlein, J., Samaga, R., Lauffenburger, D. A., Klamt, S., et al. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol. Syst. Biol.* 5, 331. doi:10.1038/msb.2009.87

Sáez-Rodríguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., et al. (2007). A logical model provides insights into t cell receptor signaling. *PLoS Comput. Biol.* 3, e163. doi:10.1371/journal.pcbi.0030163

Schlatter, R., Schmich, K., Vizcarra, I. A., Scheurich, P., Sauter, T., Borner, C., et al. (2009). ON/OFF and beyond - a boolean model of apoptosis. *PLOS Comput. Biol.* 5 (12), e1000595. Publisher: Public Library of Science. doi:10.1371/journal.pcbi.1000595

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 362 (6419), 1140–1144. doi:10.1126/science.aar6404

Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Goncalves, E., Morris, M. K., et al. (2012). CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst. Biol.* 6 (1), 133. doi:10.1186/1752-0509-6-133

Yordanov, B., Dunn, S.-J., Kugler, H., Smith, A., Martello, G., and Emmott, S. (2016). A method to identify and analyze biological programs through automated reasoning. *npj Syst. Biol. Appl.* 2 (1), 16010–16016. Number: 1 Publisher: Nature Publishing Group. doi:10.1038/npjsba.2016.10

frontiers | Frontiers in Cell and Developmental Biology

TYPE Brief Research Report
PUBLISHED 28 September 2023
DOI 10.3389/fcell.2023.1209589

OPEN ACCESS

EDITED BY
Susan Mertins,
Leidos Biomedical Research, Inc.,
United States

REVIEWED BY
Pavel Kraikivski,
Virginia Tech, United States
Lendert Gelens,
KU Leuven, Belgium

*CORRESPONDENCE
Jacques-Alexandre Sepulchre,
Jacques-Alexandre.Sepulchre@univ-cotedazur.fr
Alejandra C. Ventura,
alejvent@fbmc.fcen.uba.ar

RECEIVED 20 April 2023
ACCEPTED 11 September 2023
PUBLISHED 28 September 2023

CITATION
Marrone JI, Sepulchre J-A and Ventura AC (2023), Pseudo-nullclines enable the analysis and prediction of signaling model dynamics.
Front. Cell Dev. Biol. 11:1209589.
doi: 10.3389/fcell.2023.1209589

Pseudo-nullclines enable the analysis and prediction of signaling model dynamics

Juan Ignacio Marrone[1,2], Jacques-Alexandre Sepulchre[3]* and Alejandra C. Ventura[1,2]*

[1]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física. Ciudad Universitaria, Buenos Aires, Argentina, [2]CONICET—Universidad de Buenos Aires, Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE). Ciudad Universitaria, Buenos Aires, Argentina, [3]Institut de Physique de Nice, CNRS UMR7010, Université Côte d'Azur, Nice, France

A powerful method to qualitatively analyze a 2D system is the use of nullclines, curves which separate regions of the plane where the sign of the time derivatives is constant, with their intersections corresponding to steady states. As a quick way to sketch the phase portrait of the system, they can be sufficient to understand the qualitative dynamics at play without integrating the differential equations. While it cannot be extended straightforwardly for dimensions higher than 2, sometimes the phase portrait can still be projected onto a 2-dimensional subspace, with some curves becoming pseudo-nullclines. In this work, we study cell signaling models of dimension higher than 2 with behaviors such as oscillations and bistability. Pseudo-nullclines are defined and used to qualitatively analyze the dynamics involved. Our method applies when a system can be decomposed into 2 modules, mutually coupled through 2 scalar variables. At the same time, it helps track bifurcations in a quick and efficient manner, key for understanding the different behaviors. Our results are both consistent with the expected dynamics, and also lead to new responses like excitability. Further work could test the method for other regions of parameter space and determine how to extend it to three-module systems.

KEYWORDS

pseudo-nullclines, oscillations, bistability, MAPK, signaling, bifurcations, cell cycle

1 Introduction

In cell signaling, mathematical modeling plays an important role in analyzing and predicting different systems behavior. The range of complexity is vast, with examples as different as the two-dimensional Fitzhugh–Nagumo model (FitzHugh, 1961) and a description of the MAPK cascade with 23 equations (Kochańczyk et al., 2017).

In general, it is well known that most nonlinear differential equations modeling biological systems are not analytically solvable. Therefore, the goal of qualitative analysis of dynamical systems is to provide information about its possible behaviors without having access to its analytical solutions. In this context, a powerful method to analyze qualitatively a planar (i.e., 2D) system is the use of nullclines. These are curves where the derivative of one of the variables is equal to zero. These curves separate regions of the plane where the sign of the derivatives is constant. Moreover, their intersections correspond to steady states of the dynamics. This information can provide a quick way to sketch the phase portrait of the system, like for instance the aforementioned Fitzhugh–Nagumo model. Thus, the technique

of nullclines is sometimes sufficient to understand the qualitative dynamics of the system without integrating their differential equations.

However, this technique cannot be extended straightforwardly to phase spaces of dimension higher than 2 because the geometrical objects corresponding to the nullclines are no longer curves but more generally (hyper-)surfaces of codimension-1. Nevertheless, there are cases where the phase portrait of the system can still be projected onto a 2-dimensional subspace, with some curves playing the role of pseudo-nullclines. When applicable, phase plane analysis, and in particular the concept of nullclines, has been one of the most useful tools for the qualitative analysis of dynamical systems. Since the main limitation of the nullcline method is its restricted application to a 2-dimensional phase space, any extension of said method to a higher number of dimensions should be valuable.

In this work, we study signaling models of dimension higher than 2, where pseudo-nullclines are defined and used to qualitatively analyze the system dynamics. The first one is an early cell cycle model in *Xenopus laevis* embryo (Tsai et al., 2014). The authors study the change in the oscillatory behavior during this developmental phase, which is present across different phyla. The second example we analyze corresponds to a subsystem of the Mitogen Activated Protein Kinase (MAPK) cascade, found in all eucaryotic cells. Signals from growth factors in cell surface receptors activate three sequential levels of proteins, with the output of the cascade responsible for the phosphorylation of multiple transcription factors. This leads to its involvement in responses like proliferation and differentiation (Lewis, et al., 1998; Schaeffer and Weber, 1999; Kochańczyk et al., 2017). The well-studied model by Huang and Ferrell consists of 22 equations describing the three-level cascade (Huang and Ferrell, 1996). The last two levels, corresponding to double phosphorylation (DP) cycles, constitute the motif that we study in this work.

Our method applies when a system can be decomposed into 2 modules which are mutually coupled through 2 scalar variables. We show that, by projecting the whole dynamics onto the subspace subtended by the two scalar variables, we can define curves that play the role of pseudo-nullclines. Intersections of these pseudo-nullclines correspond to steady states of the full system. Although the use of these pseudo-nullclines is more limited than with true nullclines, we show that this approach can be useful to figure out the onset of oscillations, and other dynamical behaviors like excitability, for a system whose actual phase space dimension is larger than 2. Other works use pseudo-nullclines to analyze different cell cycle motifs (Tyson and Novák, 2022), by using specific features only applicable to those models. We propose a more systematic approach based on the modularity of the analyzed systems.

We illustrate that situations where the pseudo-nullclines intersect transversely or tangentially enable the distinction of phase portraits of oscillations described respectively by supercritical Hopf or by SNIC bifurcations, while also pointing toward Saddle-Homoclinic bifurcations. On the other hand, we show that these pseudo-nullclines admit a natural interpretation in terms of response functions of each module submitted to a constant input of the other module.

2 Methods

The idea of the method is to decompose the system in 2 modules, assuming that the coupling between the modules is one-dimensional. This means that if the variables of the modules are denoted respectively by two sets of real variables, i.e., $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$ and $\boldsymbol{y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_m})$, the model equations can be written as:

$$\begin{aligned} \frac{\boldsymbol{dx}}{\boldsymbol{dt}} &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha}(\boldsymbol{y})) \\ \frac{\boldsymbol{dy}}{\boldsymbol{dt}} &= \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\beta}(\boldsymbol{x})) \end{aligned} \tag{1}$$

where $\boldsymbol{\alpha}(\boldsymbol{y})$ and $\boldsymbol{\beta}(\boldsymbol{x})$ are two real-valued functions. Such a system can be seen as a first module, described by equations $\frac{\boldsymbol{dx}}{\boldsymbol{dt}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{a})$, where a is some input parameter, interconnected with a second module whose equations are $\frac{\boldsymbol{dy}}{\boldsymbol{dt}} = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{b})$, with b being the corresponding input parameter. The interconnection comes from replacing the input a of the first module by the function $\boldsymbol{\alpha}(\boldsymbol{y})$, and the input b of the second module by $\boldsymbol{\beta}(\boldsymbol{x})$. Decomposing a system into two interconnected modules has been considered in the literature by (Angeli et al., 2004).

To simplify the presentation and the notations in what follows we will continue with a basic example, where the coupling functions are simply $\boldsymbol{\alpha}(\boldsymbol{y}) = \boldsymbol{y_1}$ and $\boldsymbol{\beta}(\boldsymbol{x}) = \boldsymbol{x_1}$. The extension to a more general function is easy and is included at the end of the Supplementary Material, along with a sketch of the general scheme.

Thus, now a stationary state $(\boldsymbol{x^*}, \boldsymbol{y^*})$ of system (1) is a solution of the system of equations:

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y_1}) &= \boldsymbol{0} \\ \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x_1}) &= \boldsymbol{0} \end{aligned}$$

Suppose that the solutions of this system of equations can be written as follows:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{X}(\boldsymbol{y_1}) \\ \boldsymbol{y} &= \boldsymbol{Y}(\boldsymbol{x_1}) \end{aligned} \tag{2}$$

Then, by projecting these functions on the plane of coordinates $(\boldsymbol{x_1}, \boldsymbol{y_1})$, we define pseudo-nullclines of the system as two curves C_1 and C_2 whose graphs are respectively given by the parametrizations $(\boldsymbol{X_1}(\boldsymbol{y_1}), \boldsymbol{y_1})$ and $(\boldsymbol{x_1}, \boldsymbol{Y_1}(\boldsymbol{x_1}))$. The first curve can be seen as the response function of (component 1 of) the first module with respect to its input parameter y_1. Similarly, one can interpret the second pseudo-nullcline as the response function of the second module submitted to its input parameter x_1. One advantage of this definition is that by construction said stationary states of the couple modules must be found among the intersections of the two pseudo-nullclines. Indeed, by definition $(\boldsymbol{x_1^*}, \boldsymbol{y_1^*})$ can be written in two ways, either $(\boldsymbol{X_1}(\boldsymbol{y_1^*}), \boldsymbol{y_1^*})$, or $(\boldsymbol{x_1^*}, \boldsymbol{Y_1}(\boldsymbol{x_1^*}))$, thus belonging to the two graphs of C_1 and C_2.

Conversely, if $(\boldsymbol{x_1^*}, \boldsymbol{y_1^*})$ belongs to the intersection set of the pseudo-nullclines C_1 and C_2, then:

$$\begin{aligned} \boldsymbol{x_1^*} &= \boldsymbol{X_1}(\boldsymbol{y_1^*}) \\ \boldsymbol{y_1^*} &= \boldsymbol{Y_1}(\boldsymbol{x_1^*}) \end{aligned}$$

And by construction, the functions X and Y satisfy the steady state equations:

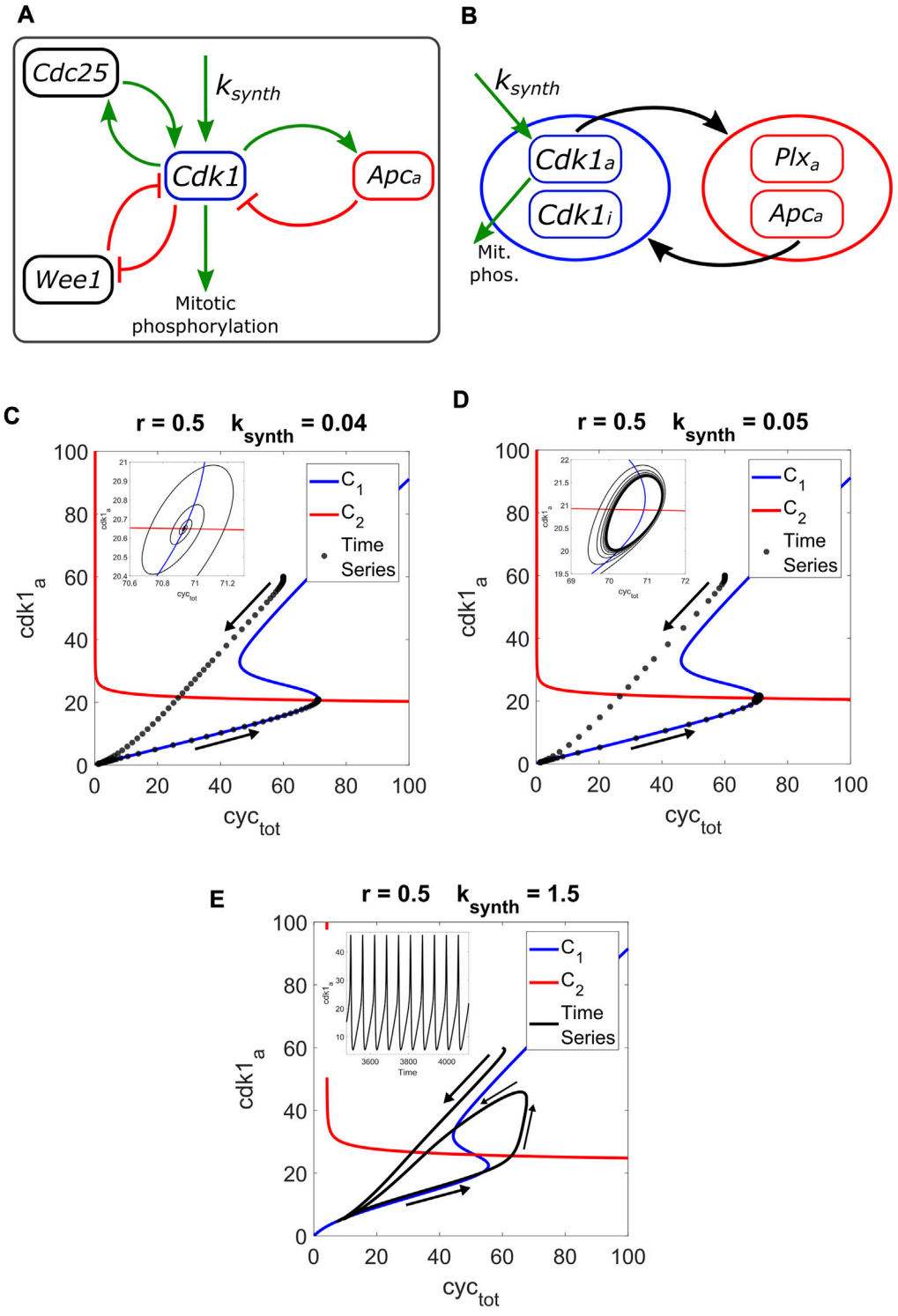


FIGURE 1
(A) Scheme of the cell cycle model, based on the one found in (Tsai et al., 2014). The parameter k_{synth} (synthesis rate of the cyclin) acts as the input of the system. Two positive feedbacks (with Cdc25 and Wee1) and one negative feedback (with Apc_a) govern the motif. The output Cdk1 (active) is involved in the mitotic phosphorylation. **(B)** Motif scheme based on the pseudo-nullcline method, separating the two modules, and representing how they are interconnected. More details on the equations can be found in the Supplementary Material. **(C)** Pseudo-nullclines C_1 (in blue) and C_2 (in red) for r = 0.5 and input = 0.04, with the corresponding time series (in black). Arrows denote the trajectory taken by the system, from $cdk1_a$ = 60 to a steady state represented by the curves intersection. **(D)** The input is now 0.05, leading to a small limit cycle around the intersection. **(E)** With input = 1.5, the limit cycle grows in amplitude, following the lower branch of C_1 but not the upper one, as shown in (Tsai et al., 2008).

$$f(X(y_1^*), y_1^*) = 0$$
$$g(Y(x_1^*), x_1^*) = 0$$

In other words, $x^* = X(y_1^*)$ and $y^* = Y(x_1^*)$ constitute a steady state of the coupled system since they satisfy the system of equations:

$$f(x^*, y_1^*) = 0$$
$$g(y^*, x_1^*) = 0$$

Another advantage of this geometrical method is that it is able to reveal a limit point bifurcation, like a saddle-node bifurcation. As it shown in the Supplementary Material, this occurs when a steady state corresponds to a tangential intersection of the pseudo-nullclines. In particular, this feature enables to distinguish between a SNIC bifurcation or a Hopf bifurcation because in the first case oscillations appear through a tangent bifurcation, whereas in the second case the pseudo-nullclines intersect transversely. Both cases are illustrated by applying our method to different signaling motifs studied in the Results section.

3 Results

3.1 Pseudo-nullclines method applied to a cell cycle model combining positive and negative feedback loops

In (Tsai et al., 2014), the authors study an oscillatory cell cycle model in *X. laevis* embryos, where the period and shape of the oscillation change between the first mitotic cycle and the subsequent cycles. They analyze the system obtaining experimental data and running computational simulations. A scheme of the model is presented in Figure 1A, showing the two positive feedback loops and the negative one involved.

The system of equations is as follows, written in a more generic manner (see Supplementary Material for the equations in detail):

$$\frac{d[cdk1_a]}{dt} = f_1([cdk1_a], [cdk1_i], [apc_a])$$
$$\frac{d[cdk1_i]}{dt} = f_2([cdk1_a], [cdk1_i], [apc_a])$$
$$\frac{d[plx_a]}{dt} = g_1([plx_a], [cdk1_a])$$
$$\frac{d[apc_a]}{dt} = g_2([apc_a], [plx_a])$$

The first module, with f_1 and f_2, consists of two equations depending on three variables. The last of these, Apc_a, is the only one belonging to the second module and thus treated as an input parameter. This results in two equations with two variables: for each value of Apc_a, a solution can be found. With both equations equal to zero, one can reach an expression that determines the first pseudo-nullcline:

$$F([cdk1_a], [apc_a]) = 0$$

In the second module, with g_1 and g_2, we also have two equations and three variables. The input parameter from the other module is $Cdk1_a$. As before, taking both equations equal to zero, one can reach an expression for the second pseudo-nullcline:

$$G([apc_a], [cdk1_a]) = 0$$

Finally, we do for both curves a change of variables from Apc_a to Cyc_{tot} (total cyclin, the sum of active and inactive complexes), and work in the (Cyc_{tot}, $Cdk1_a$) phase space (see Supplementary Material for details).

In Figure 1B, we present a scheme of the model following this modular description, as a comparison to the previous scheme based on (Tsai et al., 2014). All parameter values are presented in the Supplementary Material. The parameters changed are reported in the following text and in the Figures.

In Figure 1C, we present the pseudo-nullclines for the system and the corresponding time series trajectory (starting from Cdk1a = 60 nM) for k_{synth} = 0.04, which is just outside the oscillatory range (see Supplementary Material for a bifurcation diagram with k_{synth} as the input, showing two supercritical Hopf bifurcations). The parameter that controls the positive feedback strength, r, is equal to 0.5 (used in the Tsai et al. work). The system trajectory drops and then ascends following the lower branch of C_1, forming a spiral before ending at a fixed point. The intersection of pseudo-nullclines and the fixed point are within a very small distance of each other, meaning that the intersection represents the stable steady state of the system.

In Figure 1D, k_{synth} = 0.05, which corresponds to a limit cycle of relatively small amplitude. The intersection of curves occurs within the cycle, representing the unstable steady state, and is located just below the fold of C_1. In a 2D system analyzed with true nullclines, it would be expected for oscillations to occur only when the intersection is located between the two folds of the S-shaped curve. The crossing between our pseudo-nullclines taking place close but below the fold, plus the minimal distance between the intersection and the end point of the time series in the previous case, reflect the "pseudo" character of our method while still showing its usefulness.

In Figure 1E, k_{synth} = 1.5, the value used in the work of Tsai et al. Once again, the trajectory follows the lower branch of C_1 but not the upper one. This is consistent with results showed by the authors in (Tsai et al., 2008).

Given the bistable shape of the pseudo-nullcline for the Cdk1 module, there is the question of whether both curves could be brought together in a tangential manner. The results shown so far only deal with transversal intersections, with one stable fixed point or limit cycles around an unstable point, born through Hopf bifurcations. A tangency would represent a saddle-node bifurcation, which could act as a SNIC or indicate the existence of a Saddle-Homoclinic (SHom) bifurcation, since there would be a saddle (by virtue of the SN) and a limit cycle (as already established). These global bifurcations would allow more control over the period than what is possible with Hopf bifurcations.

Considering the shape of C_1, the distance with C_2, and the composition of Hill functions that goes into C_2, we performed a few modifications in the model with the goal of bringing about a tangency. First, we added an extra parameter into the differential equation for Apc_a (see Supplementary Material). Since low values of Apc_a correspond to high values of Cyc_{tot} (outside of the plot scale), adding the extra parameter can bring C_2 to a drop in $Cdk1_a$ close to the right-hand fold of C_1. At the same time, it could represent basal activity of Apc_a in absence of Plx_a.

With k_{synth} = 1.5, r = 10 (value used in (Tsai et al., 2008)) and increasing ec50 for Plx_a to adjust the threshold of C_2, we arrived at

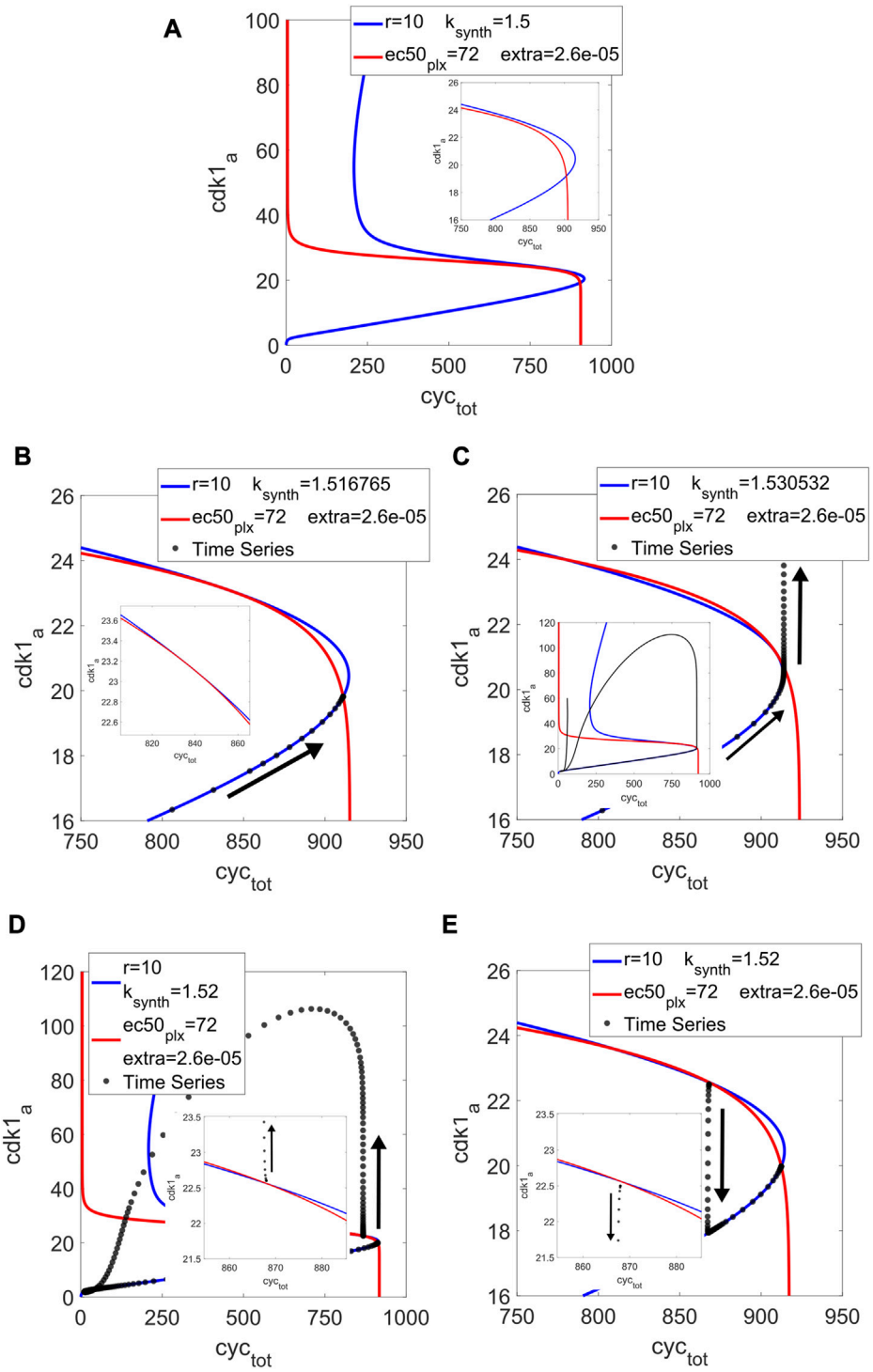


FIGURE 2
Cell cycle model: pseudo-nullclines C_1 (in blue) and C_2 (in red) for r = 10, $ec50_{plx}$ = 72, extra = 2.6e-5 and different input values, with their corresponding time series (in black) for all panels except the first one. **(A)** Input = 1.5, the pseudo-nullclines are close to a tangency, taking advantage of the right-hand fold in C_1. **(B)** Input = 1.516765, tangency at a distance of the fold, with the time series ending at the lower intersection (the stable steady state). **(C)** Input = 1.530532, tangency close to the fold, intersection between the folds (an unstable steady state), and a limit cycle develops. **(D)** Input = 1.52, taking advantage of the saddle one can choose different initial conditions to obtain excitability (or not). Above the saddle, the system goes around the phase space before ending at the stable steady state. **(E)** Input = 1.52, with an initial condition below the saddle it goes directly to the stable steady state.

Figure 2A. The distance between the pseudo-nullclines close to the right-hand fold of C_2 is small enough that a tangency seems possible. We ran the model in MatCont and found two SN at k_{synth} = 1.516765 and 1.530532. Figures 2B,C show the pseudo-nullclines at these values. The tangential behavior of the curves can be appreciated. For the lower input value, the tangency occurs between the C_1 folds and the

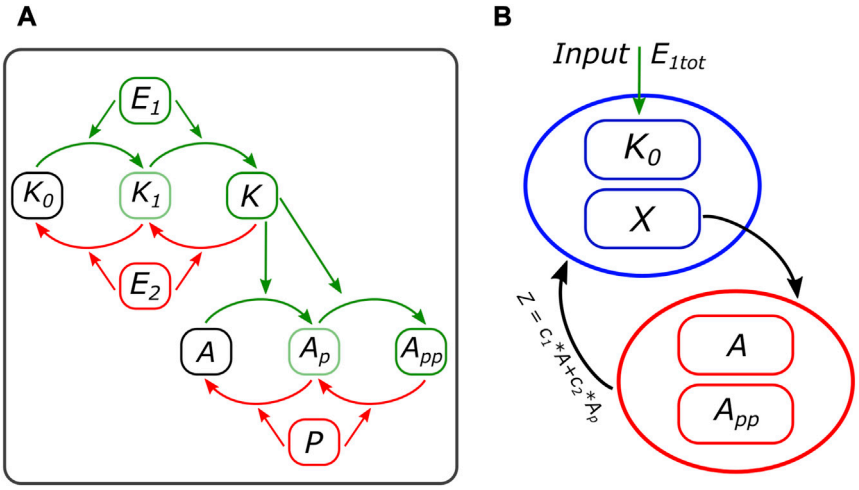


FIGURE 3
(A) Scheme of the 2 + 2 model. The input kinase E_1 activates the first-level kinase, which goes through two steps before phosphorylating, also in two steps, the second-level substrate. **(B)** Motif scheme based on the pseudo-nullcline method, representing the first- and second-level modules and how they are interconnected through X and Z. More details on the equations can be found in the Supplementary Material.

transversal intersection outside of them, corresponding to a stable steady state. For the higher input, the tangency is much closer to the C_1 fold while the intersection is between the folds, representing an unstable fixed point around which the limit cycle takes place. Between these two SN for the full system, an SHom bifurcation was found at $k_{synth} = 1.527$.

In Figure 2D, we show a case for $k_{synth} = 1.52$, which is outside of the oscillatory range but between the two SN. Depending on the initial condition, the system can 1) go around the phase space describing one output peak in time before ending at the steady state or 2) take a shorter path to said fixed point. With the initial condition of Figure 2D, just above the saddle point represented by the middle intersection, it goes around. In Figure 2E, it starts from below the saddle, and so it goes directly to lower intersection, corresponding to the stable steady state. The model displays excitability in this region of parameter space, well described by the pseudo-nullclines.

In all, not only our method was consistent with bifurcations born from the original parameter set, but it also allowed us to find a new bifurcation through the manipulation of the two pseudo-nullclines. Moving one parameter at a time facilitates an exploration where the intersections between the curves can change and lead to new findings. In this particular system, the use of Hill functions shows a useful path for the exploration, by modifying the amplitude and threshold of C_2. We argue that, since Hill functions are prevalent in system biology, this example could serve as inspiration for the analysis of many other cases. At the same time, for any model, the pseudo-nullclines will provide a visual guide for finding new behaviors.

3.2 MAP kinase subsystem where both modules are capable of bistability

The second model studied in this work corresponds to the last two levels of the MAPK cascade. It consists of a DP cycle where its output, the double phosphorylated substrate, acts as the kinase for another DP cycle. We will call it the 2 + 2 system, following the double modification process in each level. A scheme is presented in Figure 3A. This motif is of interest for our work, taking the application of the method to a subsystem in an important and well-studied model in biology. But also, there are two important differences with the cell cycle motif from the previous subsection: it is of higher dimension (17 variables versus 4) and capable of bistability in both modules (Markevich et al., 2004).

The parameter set we chose comes from our previous work (Marrone et al., 2023), where the DP cycle displayed bistability when scanning the input kinase. This was a necessary condition to obtain oscillations in the motifs studied and valuable for this work since the presence of oscillations in the model and bistability in each of the two modules (emergent through SN or fold bifurcations) will test the pseudo-nullclines method.

We work with a reduced version of the 2 + 2 system, which can be written as follows (see Supplementary Material for the detailed reduction from the original 17 equations):

$$\frac{d[K_0]}{dt} = f_1([K_0], [K_1], X, Z)$$
$$\frac{dX}{dt} = f_2([K_0], [K_1], X, Z)$$
$$\frac{d[A]}{dt} = g_1([A], [A_p], [A_{pp}], X, Z)$$
$$\frac{d[A_{pp}]}{dt} = g_2([A], [A_p], [A_{pp}], X, Z)$$

X and Z are functions of some of the original variables:

$$X = [K] + [AK] + [A_pK]$$
$$Z = c_1[A] + c_2[A_p]$$

These two are the coupling functions of the model, one for each module, connecting the first and second DP cycle. All parameter values are presented in the Supplementary Material. The input

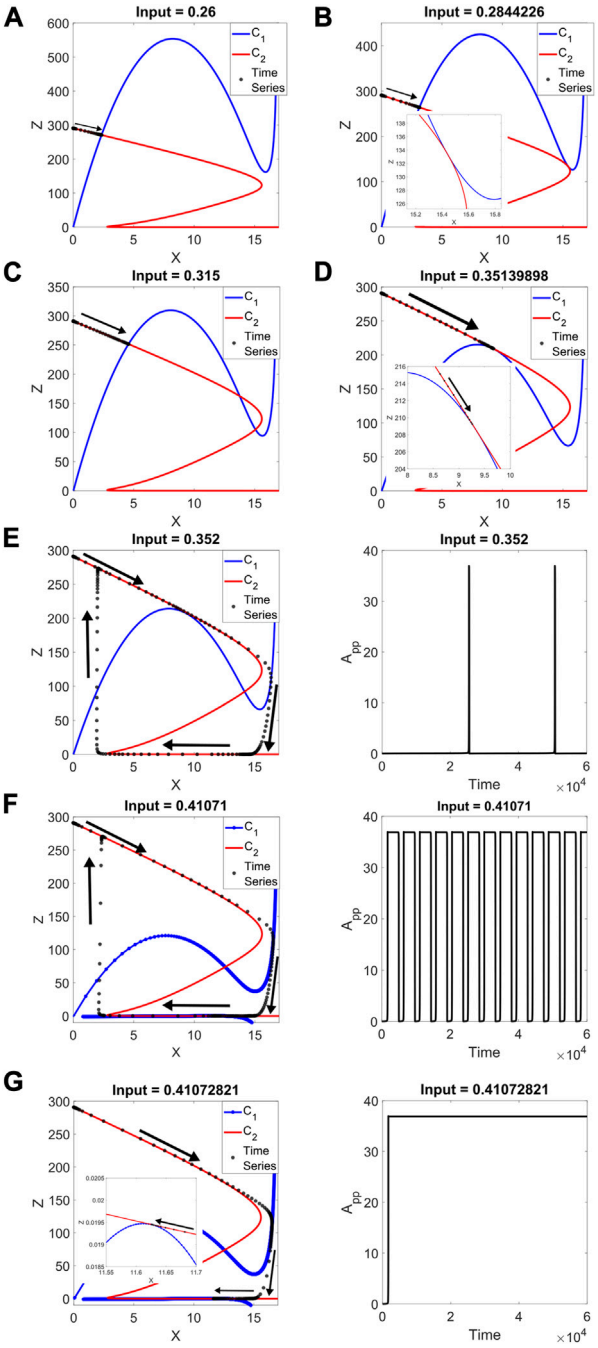


FIGURE 4
2 + 2 model: pseudo-nullclines C_1 (in blue) and C_2 (in red) for different input values, with their corresponding reduced-model time series (in black). Arrows denote the trajectory taken by the system. **(A)** Input = 0.26, only one intersection, where the time series ends. **(B)** Input = 0.2844226, a tangency takes place between the curves, where an SN is located in the full system. **(C)** Input = 0.315, three intersections are found, with the time series going to the representative of the stable fixed point **(D)** Input = 0.35139898, a second tangency for the second full-system SN. **(E)** Input = 0.352. Left: the limit cycle takes full advantage of C_2's amplitude. Only one intersection remains, the unstable steady state. Right: the peaks are narrow compared to the time the output is off. **(F)** Input = 0.41071. Left: close to the next SN, the system continues to oscillate with relatively unchanged amplitude, and the first-level pseudo-nullcline shows up for low values of Z (vertical axis plotted from a negative value for clarity). The curves are close to each other. Right: the output is on

(*Continued*)

FIGURE 4 (Continued)
for a longer time, with brief drops. **(G)** Input = 0.41072821, the tangency at low Z occurs, where the next full-system SN is located. Right: the time series no longer is an oscillation.

parameter is reported in the following text and in the Figures. Also in the Supplementary Material, a bifurcation diagram for the full system (of 17 equations) with the input E_{1tot} as the parameter, showing four SN bifurcations and two Hopf bifurcations.

In Figure 4A, we show the results at input = 0.26, including the time series for the reduced system. Only one intersection exists, corresponding to a stable fixed point, where the series culminates. It is important to remark once again that both modules are capable of bistability, so it is within the bounds of expectation for both pseudo-nullclines to have folds. In Figure 4B, the input reaches an SN (in the full-system bifurcation diagram), and the two curves are tangent to one another. The new two intersections in Figure 4C represent the new steady states that come after the SN.

In Figure 4D, the input takes the system to a second SN, and the tangency between pseudo-nullclines occurs at a higher value of Z than in Figure 4B. This is coherent with the output A_{pp} being lower on this SN. When Z (combination of A and A_p) is high, A_{pp} is low and *vice versa*. It is also worth noting that this second tangency takes place near a different fold of the first-level curve.

Starting from this input value, oscillations are found, as shown in Figure 4E. There is only one curve intersection, representing an unstable steady state. This point is located between the two folds of both C_1 and C_2. The output spends most of each period at a low level, with brief peaks of activity.

In Figure 4F, the system is close to the next SN. The curves are close to a tangency at a value just above Z = 0. The previous intersection between the folds remains, and two new intersections are close to occur. The output now spends more time at a high level, with relatively brief drops.

The nature of these oscillations comes from the system's proximity to global bifurcations. When the pseudo-nullclines are almost tangent and the behavior is oscillatory, the trajectory of the system slows down in the vicinity of the almost-tangency. For the input of Figure 4E, the almost-tangency occurs for high Z, low A_{pp}. The system can spend a relatively long time in this area. In Figure 4F, at low Z, high A_{pp}, the high-level time can be extended with precise manipulations of the input, leaving narrow drops in output.

An interesting aspect of this case is that we have not been able to confirm the presence of SNIC bifurcations via MatCont for the full system (even though SN bifurcations are found when the tangencies occur), while the reduced system cannot be analyzed due to the implicit equations for the conservations (see Supplementary Material). We argue that our method provides further evidence of global bifurcations when a well-known software for analyzing bifurcations falls short of confirmation.

Once the input reaches the next SN, in Figure 4G, the curves are tangent, and the time series stops at that point. At this tangency, the oscillations disappear. The range for stable limit cycles appears limited by two SN bifurcations, with the limit cycle taking advantage of C_2's amplitude all along the oscillatory range.

Further scanning of the input shows what is expected, with two new intersections and the time series stopping at the lowest one in Z (the

highest in A_{pp}). Eventually, the last SN point of the full system is represented by a new tangency close to the left-hand fold of the second-level curve (see Supplementary Material for these last results).

Even though, throughout Figure 4, we are plotting the trajectory of the reduced system, one can find similar results when integrating the full system. And while we cannot obtain with MatCont a bifurcation diagram for the reduced system (as mentioned, due to the implicit nature of the conservation equations), we selected input values following the bifurcations in the full system, with consistent results.

4 Discussion

In this work, we applied our pseudo-nullclines method on two models, one corresponding to the embryonic cell cycle and another to a subsystem of the MAPK cascade. They represent two well-known and important examples in systems biology. The parameter sets involved different bifurcations and behaviors, with the purpose of testing the method.

For the Tsai et al. motif, not only we found consistency in our results using the authors' parameter values, but we were also able to manipulate the pseudo-nullclines toward different bifurcations and therefore, new behaviors. The use of Hill functions for the differential equations was convenient in this regard, and their recurrent use in mathematical modelling of biological systems means that the pseudo-nullclines could be useful for dynamical analysis.

The 2 + 2 motif, unlike the first case, displayed folds for both pseudo-nullclines, representing the underlying bistability in each DP cycle and therefore expanding the pseudo-nullclines application to a bistability-in-both-modules example. The method proved consistent with the motif behavior even though a reduction of the system equations was first necessary, and also helped tracked bifurcations that were not confirmed on MatCont. It remains to be seen whether the method continues to provide useful and consistent results for other regions of parameter space, and how it can be extended to the full MAPK cascade, which involves three modules.

A 2021 work by De Boeck et al. studies the embryonic cell cycle through two bistable switches (a three-equation system), finding high amplitude oscillations with increased robustness: a larger oscillatory region of parameter space than in the case with one bistable switch (De Boeck et al., 2021). Our results, coming from a cell cycle motif (with one bistable module) and a system composed of two bistable modules, could be further developed in this area of cell biology considering the advantages from the work by De Boeck et al. (correct cell cycle progression) and our own (consistent and different behaviors with a four-equation system). In particular, recent work by Parra-Rivas et al. presents a very detailed bifurcation study of various cell cycle models, including the combination of two bistable switches (Parra-Rivas et al., 2023). Our pseudo-nullclines method could be useful for further interpretation in the origin of said bifurcations, which include those of the global type (like the two motifs studied in our present work).

One can find cases in the literature for which our method cannot be applied, like in (Kraikivski et al., 2015) where the system in question, a large cell cycle model in yeast, is divided into a high number of modules, some of them having more than one connection to the rest. It is possible that some type of model reduction or approximation is first necessary to analyze it through pseudo-nullclines. On the other hand, other candidates in the literature are found for applying the pseudo-nullclines method. In (Perez-Carrasco et al., 2018), the authors combine two simple motifs to arrive at a system capable of different behaviors, not obtained with each motif in isolation. The three-equation description is such that two modules are readily determined, each one depending on the other through their coupling variables. The same can be said of the motifs in (Ananthasubramaniam and Herzel, 2014), where the authors lower the degree of cooperativity necessary for oscillations to occur by adding positive feedbacks on three-component negative feedback loops. We believe that the method can be of great value in systems biology, with useful analysis and potential findings in experimental biology.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

J-AS and AV designed the project. JM performed all mathematical analysis and simulations. JM, J-AS and AV analyzed the results. JM, J-AS and AV prepared the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Grants from the Argentine Agency of Research and Technology (PICT 2019-01681 and PICT 2019-1455) to AV, and by the French-Argentinian IRP LICOQ for the work stays of J-AS and JM.

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1209589/full#supplementary-material

References

Ananthasubramaniam, B., and Herzel, H. (2014). Positive feedback promotes oscillations in negative feedback loops. *PLoS ONE* 9 (8), e104761. doi:10.1371/journal.pone.0104761

Angeli, D., Ferrell, J. E., and Sontag, E. D. (2004). Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proc. Natl. Acad. Sci. U. S. A.* 101 (7), 1822–1827. doi:10.1073/pnas.0308265100

De Boeck, J., Rombouts, J., and Gelens, L. (2021). A modular approach for modeling the cell cycle based on functional response curves. *PLoS Comput. Biol.* 17 (8), e1009008. doi:10.1371/journal.pcbi.1009008

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical J.* 1 (6), 445–466. doi:10.1016/S0006-3495(61)86902-6

Huang, C. Y. F., and Ferrell, J. E. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc. Natl. Acad. Sci. U. S. A.* 93 (19), 10078–10083. doi:10.1073/pnas.93.19.10078

Kochańczyk, M., Kocieniewski, P., Kozłowska, E., Jaruszewicz-Błońska, J., Sparta, B., Pargett, M., et al. (2017). Relaxation oscillations and hierarchy of feedbacks in MAPK signaling. *Sci. Rep.* 7, 38244. doi:10.1038/srep38244

Kraikivski, P., Chen, K. C., Laomettachit, T., Murali, T. M., and Tyson, J. J. (2015). From START to FINISH: computational analysis of cell cycle control in budding yeast. *Npj Syst. Biol. Appl.* 1, 15016. doi:10.1038/npjsba.2015.16

Lewis, T. S., Shapiro, P. S., and Ahn., N. G. (1998). Signal transduction through MAP kinase cascades. *Adv. Cancer Res.* 74, 49–139. doi:10.1016/s0065-230x(08)60765-4

Markevich, N. I., Hoek, J. B., and Kholodenko, B. N. (2004). Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J. Cell Biol.* 164 (3), 353–359. doi:10.1083/jcb.200308060

Marrone, J. I., Sepulchre, J.-A., and Ventura, A. C. (2023). A nested bistable module within a negative feedback loop ensures different types of oscillations in signaling systems. *Sci. Rep.* 13 (1), 529. doi:10.1038/s41598-022-27047-4

Parra-Rivas, P., Ruiz-Reynés, D., and Gelens, L. (2023). Cell cycle oscillations driven by two interlinked bistable switches. *Mol. Biol. Cell* 34 (6), ar56. doi:10.1091/mbc.E22-11-0527

Perez-Carrasco, R., Barnes, C. P., Schaerli, Y., Isalan, M., Briscoe, J., and Page, K. M. (2018). Combining a toggle switch and a repressilator within the AC-DC circuit generates distinct dynamical behaviors. *Cell Syst.* 6 (4), 521–530. doi:10.1016/j.cels.2018.02.008

Schaeffer, H. J., and Weber, M. J. (1999). Mitogen-activated protein kinases: specific messages from ubiquitous messengers. *Mol. Cell. Biol.* 19 (4), 2435–2444. doi:10.1128/mcb.19.4.2435

Tsai, T. Y.-C., Yoon, S. C., Ma, W., Pomerening, J. R., Tang, C., and Ferrell, J. E., Jr (2008). Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science* 321 (5885), 126–129. doi:10.1126/science.1156951

Tsai, T. Y.-C., Theriot, J. A., and Ferrell, J. E. (2014). Changes in oscillatory dynamics in the cell cycle of early Xenopus laevis embryos. *PLoS Biol.* 12 (2), e1001788. doi:10.1371/journal.pbio.1001788

Tyson, J. J., and Novák, B. (2022). Time-keeping and decision-making in the cell cycle. *Interface Focus* 12, 20210075. doi:10.1098/rsfs.2021.0075

frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 03 November 2023
DOI 10.3389/fcell.2023.1233808

OPEN ACCESS

EDITED BY
Michael Blinov,
UConn Health, United States

REVIEWED BY
Silas Boye Nissen,
Stanford University, United States
Lee Bardwell,
University of California, Irvine,
United States

*CORRESPONDENCE
Jeremy Gunawardena,
jeremy@hms.harvard.edu

†PRESENT ADDRESS
Kee-Myoung Nam,
Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, United States

RECEIVED 02 June 2023
ACCEPTED 02 October 2023
PUBLISHED 03 November 2023

CITATION
Nam K-M and Gunawardena J (2023), The linear framework II: using graph theory to analyse the transient regime of Markov processes.
Front. Cell Dev. Biol. 11:1233808.
doi: 10.3389/fcell.2023.1233808

The linear framework II: using graph theory to analyse the transient regime of Markov processes

Kee-Myoung Nam† and Jeremy Gunawardena*

Department of Systems Biology, Harvard Medical School, Boston, MA, United States

The linear framework uses finite, directed graphs with labelled edges to model biomolecular systems. Graph vertices represent chemical species or molecular states, edges represent reactions or transitions and edge labels represent rates that also describe how the system is interacting with its environment. The present paper is a sequel to a recent review of the framework that focussed on how graph-theoretic methods give insight into steady states as rational algebraic functions of the edge labels. Here, we focus on the transient regime for systems that correspond to continuous-time Markov processes. In this case, the graph specifies the infinitesimal generator of the process. We show how the moments of the first-passage time distribution, and related quantities, such as splitting probabilities and conditional first-passage times, can also be expressed as rational algebraic functions of the labels. This capability is timely, as new experimental methods are finally giving access to the transient dynamic regime and revealing the computations and information processing that occur before a steady state is reached. We illustrate the concepts, methods and formulas through examples and show how the results may be used to illuminate previous findings in the literature.

KEYWORDS

linear framework, graph theory, Matrix-Tree theorems, rational functions, Markov processes, first-passage times

1 Introduction

The linear framework is a graph-theoretic approach to analysing biomolecular systems (Gunawardena, 2012; Mirzaev and Gunawardena, 2013; Gunawardena, 2014). A recent review (Nam et al., 2022) described how the framework has been used to study systems at steady state, in contexts such as post-translational modification and gene regulation. The present paper is a sequel to this review, which describes how the graph-theoretic approach can be extended to the transient regime, prior to the steady state being reached, for systems that are Markov processes. These new results were introduced in the first author's Ph.D. thesis (Nam, 2021) and full details with complete proofs are being published separately (Nam and Gunawardena, 2023). The purpose of the present paper is to provide an elementary introduction to this circle of ideas for a wider readership in cell and developmental biology. We hope this will be of interest to anyone who wants to explore the transient regime for biological systems that can be modelled by Markov processes.

Linear framework graphs (hereafter, "graphs") are finite, simple, directed graphs with labelled edges. (A simple graph is one in which there is at most one edge between any two

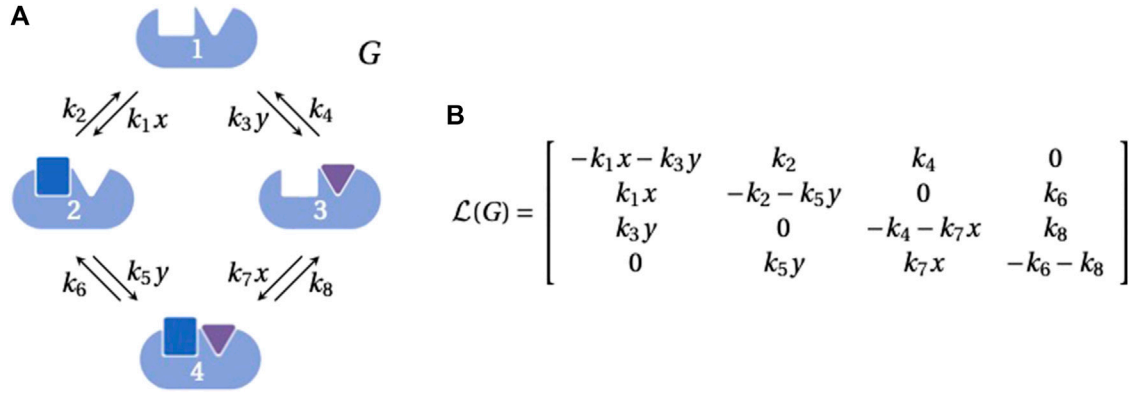


FIGURE 1
Linear framework graph and Laplacian matrix. **(A)** An example graph, G, representing the binding of two ligands, each to one site, on a biomolecule, with vertices indexed 1, . . ., 4 as shown. The labels on the edges $1 \to 2$ and $3 \to 4$ include the concentration, x, of the blue ligand that binds to the first site and the labels on the edges $1 \to 3$ and $2 \to 4$ include the concentration, y, of the purple ligand that binds to the second site. The parameters k_1, k_3, k_5 and k_7 are on-rates for binding, with dimensions of (concentration × time)$^{-1}$; the other parameters are simple rates with dimensions of (time)$^{-1}$. Graphics were generated using BioRender.com. **(B)** The Laplacian matrix, $\mathcal{L}(G)$, for the graph in panel **A**.

distinct vertices and there are no self-loops.) Graph vertices, usually denoted 1, 2, 3, . . ., represent chemical species or molecular states; edges, denoted $i \to j$, represent reactions or transitions; and edge labels, denoted $\ell(i \to j)$, represent rates which are positive and have dimensions of (time)$^{-1}$. Importantly, the labels may include expressions that describe how the underlying system is interacting with its environment. For example, the graph in Figure 1A shows how ligand binding gives rise to concentration terms in the edge labels.

A graph yields a linear dynamics, from which the linear framework gets its name. The dynamics is most simply described by imagining that the edges are chemical reactions with the edge labels as the rate constants for mass-action kinetics. Since each reaction has only a single substrate, the resulting dynamics is necessarily linear and can be expressed in matrix form as

$$\frac{du(t)}{dt} = \mathcal{L}(G) \cdot u(t). \tag{1}$$

Here, $u(t) = (u_1(t), \ldots, u_N(t))^T$ is the column vector of concentrations at each of the N vertices, and $\mathcal{L}(G)$ is the Laplacian matrix of the graph (Figure 1B). Graph Laplacians are defined with varying conventions and scalings and they may be interpreted as discrete versions of the classical Laplacian differential operator (Chung, 1997). From this viewpoint, Eq. 1 is a discretised diffusion equation. Since matter is neither created nor destroyed during the dynamics, there is a conservation law,

$$u_1(t) + \cdots + u_N(t) = u_{\text{tot}}. \tag{2}$$

Eq. 2 manifests itself in the column sums of the Laplacian being zero, $1 \cdot \mathcal{L}(G) = 0$ (Figure 1B), where 1 denotes the all-ones row vector of the appropriate dimension.

The framework is typically used in two contexts: for bulk biochemistry of reacting chemical species, where $u(t)$ in Eq. 1 describes the deterministic time evolution of species concentrations; and for individual molecular systems that exhibit stochastic transitions, where $u(t)$ describes the deterministic time evolution of the probabilities of the molecular states. In the latter case, since probabilities sum to 1, $u_{\text{tot}} = 1$. It is interesting that the same mathematics describes both contexts. Here, we will be working in the context of individual molecules and stochastic transitions. From now on, $u(t)$ will be the vector of probabilities and we will assume that $u_{\text{tot}} = 1$.

The graph formulation allows nonlinear biochemistry, which often arises from ligand binding, to be disentangled into a linear part carried by the linear dynamics in Eq. 1 and a nonlinear part that comes through the edge labels (Nam et al., 2022). The terms appearing in the labels, such as ligand concentrations (Figure 1A), have to be dealt with separately. They may be specified by separate conservation laws or by other graphs (Nam et al., 2022). For the present paper, we will assume that any ligands that are interacting with a graph are present in "reservoirs" (Nam et al., 2022, §4), similar to thermodynamic reservoirs, so that their free concentrations do not change upon binding. Accordingly, edge labels are treated as constants over the timescale of the dynamics in Eq. 1. In this case, for the stochastic context described above, the graph specifies the infinitesimal generator for a finite-state, continuous-time, time-homogeneous Markov process, $X(t)$, (hereafter, a "Markov process"), so that the edge labels are given by,

$$\ell(i \to j) = \lim_{h \to 0} \frac{\Pr(X(t+h) = j \mid X(t) = i)}{h},$$

whenever the right-hand side is nonzero and therefore positive. (A zero infinitesimal rate does not yield an edge.) Conversely, any such Markov process with an infinitesimal generator is specified by a graph (Mirzaev and Gunawardena, 2013, Theorem 4). The Laplacian dynamics in Eq. 1, with $u_{\text{tot}} = 1$, becomes the master equation for the forward evolution of the vertex probabilities, $u(t)$. The linearity of the linear framework is perhaps less surprising now, as master equations are, indeed, linear (van Kampen, 1992). We see that, within reservoir assumptions, the linear framework provides a graph-theoretic way to define and study the Markov processes that have been widely used to model biological systems.

Surprisingly, the graph rarely makes an appearance in the Markov process literature. This may be because the graph theory has so far primarily been used to study steady states of the Laplacian dynamics (Nam et al., 2022), which may not have been of much mathematical interest outside of applications in biology. Since Eq. 1 is linear, it can readily be solved in terms of the eigenvalues and eigenvectors of $\mathcal{L}(G)$. Recall that if $\mathcal{L}(G)\cdot v = \lambda v$, for some vector v and some scalar λ, then v is an eigenvector for the eigenvalue λ (Strang, 2022). By definition, the steady state of Eq. 1, which we will denote by $u^{\infty}(G)$, satisfies $du^{\infty}(G)/dt = 0$, so it follows from Eq. 1 that $\mathcal{L}(G)\cdot u^{\infty}(G) = 0$. In other words, $u^{\infty}(G)$ is an eigenvector for the zero eigenvalue.

When G is *strongly connected* (see below), the steady state, $u^{\infty}(G)$ is unique. This particular eigenvector can be calculated from $\mathcal{L}(G)$ using the determinants of principal sub-matrices, or the *first minors* of $\mathcal{L}(G)$, which thereby have terms of alternating sign (Strang, 2022). It is a remarkable property of Laplacian matrices that extensive cancellations take place so that their minors can be written as *manifestly positive polynomials* in the edge labels (Eq. 5). A polynomial is a sum of *monomials*, where a monomial is an algebraic expression consisting solely of a product of variables and a numerical coefficient, like $5a^3bc^2$ (Barbeau, 1989). A polynomial is manifestly positive if the numerical coefficient of each monomial is positive. (A polynomial like $a^2 - 2ab + b^2 = (a-b)^2$ is positive for any distinct positive values of a and b, but it is not manifestly positive.) A *rational function* or *rational expression* is the ratio of two polynomials and is itself manifestly positive if both its numerator and denominator polynomials are manifestly positive.

The algebra that gives rise to manifestly positive polynomials is controlled by appropriate subgraphs of G, described in the classical Matrix-Tree theorem (MTT), which goes back to 19th century work on electrical circuits (Kirchhoff, 1847; Mirzaev and Gunawardena, 2013); the manifest positivity is exactly what is required for parametric dependence in biology. Steady-state probabilities thereby emerge as manifestly positive rational functions of the edge labels (Eq. 4). This representation has proved very useful in giving mathematical access to steady states (Nam et al., 2022).

An important feature of this rational expression for steady-state probabilities is that it holds for systems that do not necessarily reach a steady state of thermodynamic equilibrium. Briefly, graphs that can reach thermodynamic equilibrium must be *reversible*, so that, given any edge $i \to j$, there is an edge $j \to i$ that represents the reverse process, and must satisfy the *cycle condition*: the product of the label ratios along any cycle of reversible edges is always 1 (Nam et al., 2022, §4). The cycle condition is equivalent to *detailed balance* or *microscopic reversibility*. In this case, a considerable simplification can be made in describing steady-state probabilities and the resulting expressions turn out to be equivalent to those of equilibrium statistical mechanics (Nam et al., 2022, §4). One great advantage of the linear framework is that it provides a restricted context in which non-equilibrium statistical mechanics can be exactly solved in rational algebraic terms. The functional significance of energy expenditure is a very interesting problem in cellular information processing (Estrada et al., 2016) but lies outside the scope of the present paper. We will mention some of the questions that arise in the Discussion.

A distinguishing feature of the linear framework is that the graph is treated, not just as a description or as a vehicle for doing Matrix-Tree calculations, but as a mathematical entity in its own right, in terms of which general theorems can be formulated. The graph provides a rigorous language in which salient biological features can be precisely expressed while others can be left largely unspecified, thereby allowing some general principles to emerge from behind the overwhelming molecular complexity that is ever present. Among the areas for which this approach has yielded insights are input-output responses (Wong et al., 2018; Yordanov and Stelling, 2018), post-translational modifications (Dasgupta et al., 2014; Nam et al., 2020), allostery (Biddle et al., 2021) and gene regulation (Estrada et al., 2016; Biddle et al., 2019).

Since the initial development of the linear framework, we had long thought that only steady states could be expressed as rational functions of the edge labels. However, as we will show here, important properties of the transient regime, such as first-passage times, can also be calculated as rational functions of the edge labels. The capability to analyse transient behaviour using graph-theoretic methods is particularly welcome because real-time and single-molecule experimental methods are finally giving access to the transient regime within living cells (Kleine Borgmann et al., 2013; Liao et al., 2015; Jones et al., 2017; Loffreda et al., 2017; Chen et al., 2018; Dufourt et al., 2018; Mir et al., 2018; Volkov et al., 2018; Nandan et al., 2022). Much of our understanding of biochemical behaviour has relied on steady-state assumptions, which are not always explicitly stated. The rich complexity of transient behaviours which are beginning to emerge suggests that the time is ripe to develop a more fundamental understanding of the kinds of biochemical computations and information processing that can be achieved transiently. For this, the mathematical methods described here may be of some value.

2 Results

2.1 Steady states and spanning trees

As preparation for discussing first-passage times, we briefly explain how steady-state probabilities are calculated in terms of the graph; see (Nam et al., 2022, §2) for more details. If we have a graph G, we noted in the Introduction that the steady state, $u^{\infty}(G)$, satisfies $\mathcal{L}(G)\cdot u^{\infty}(G) = 0$, so that, in linear algebra terms, $u^{\infty}(G)$ lies by definition in the *kernel* of the Laplacian matrix: $u^{\infty}(G) \in \ker\mathcal{L}(G)$. If G is *strongly connected*—i.e., if, for any pair of distinct vertices i and j, there is a directed path of edges from i to j—then this kernel is one-dimensional (Gunawardena, 2012),

$$\dim\ker\mathcal{L}(G) = 1. \tag{3}$$

(The structure of $\ker\mathcal{L}(G)$ is well understood for non-strongly connected graphs (Mirzaev and Gunawardena, 2013). We will not need this for steady states but we will encounter non-strong connectivity when discussing first-passage times in the next section.) Eq. 3 means that if $z \in \ker\mathcal{L}(G)$ is any nonzero vector, then any other vector in the kernel, such as $u^{\infty}(G)$, is a scalar multiple of z: $u^{\infty}(G) = \lambda z$, for some number λ.

The classical Matrix-Tree theorem (MTT) yields a formula for a canonical basis vector, $\rho(G) \in \ker\mathcal{L}(G)$. We will describe this formula shortly but note first that, as just mentioned, $u^{\infty}(G)$ must be a scalar multiple of $\rho(G)$, so that $u_i^{\infty}(G) = \lambda\rho_i(G)$ for

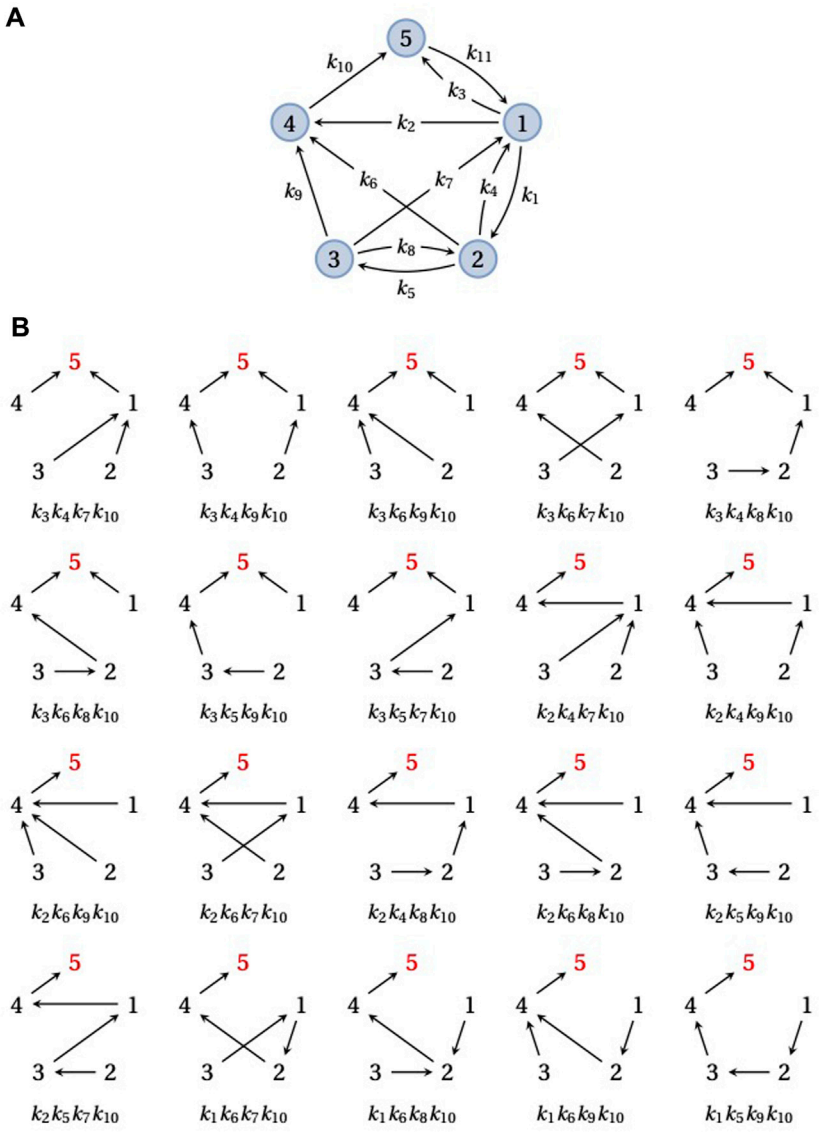


FIGURE 2
Spanning trees and steady-state probabilities. **(A)** An example graph, G, on five vertices, $\{1,\ldots,5\}$, with 11 edges, labeled $k_1,\ldots,k_{11}$. G is strongly connected. **(B)** The 20 spanning trees of G rooted at vertex 5 (red), each with its corresponding monomial product of edge labels. The sum of these 20 edge label products gives $\rho_5(G)$ in Eq. 5.

$i = 1, \ldots, N$. Using the conservation law in Eq. 2 and recalling that $u_{\text{tot}} = 1$ for probabilities, λ may be removed by normalising, so that,

$$u_i^{\infty}(G) = \frac{\rho_i(G)}{\rho_1(G) + \cdots + \rho_N(G)}. \tag{4}$$

We need some terminology to explain how $\rho(G)$ is determined from G. A *spanning forest*, F, of G is a subgraph that contains all vertices in G ("spanning"), lacks cycles when edge directions are ignored ("forest"), and has at most one outgoing edge from each vertex. The vertices with no outgoing edges are called the *roots* of F. If F has only one root, it is called a *spanning tree*. A forest consists of separate trees, although the forest is upside down, with each tree ascending to its root. Given any non-empty subset of vertices, $\varnothing \neq U \subseteq \{1, \ldots, N\}$, let $\Phi_U(G)$ denote the set of spanning forests of G that are rooted at U. Finally, given any subgraph H of G, let $w(H)$ denote the product of all the edge labels in H: $w(H) = \prod_{i \to j \in H} \ell(i \to j)$. As a matter of convention, if H has no edges, then $w(H) = 1$. Then, $\rho_i(G)$ is obtained by summing $w(F)$ over all spanning trees F of G that are rooted at i,

$$\rho_i(G) = \sum_{F \in \Phi_{\{i\}}(G)} w(F). \tag{5}$$

$\rho_i(G)$ is a manifestly positive polynomial in the edge labels, with each $w(F)$ being a monomial with coefficient +1. The steady-state probabilities, $u^{\infty}(G)$, can be recovered from $\rho_i(G)$ by using Eq. 4. Figure 2 illustrates this calculation for an example graph with five vertices and $i = 5$. Spanning trees are sufficient to calculate steady-state probabilities in Eq. 5 but spanning forests are also needed for the transient quantities considered below (Eqs. 6, 7).

Eq. 5 is a consequence of the classical MTT. The MTT is one of a family of theorems that describe the relationship between the minors of $\mathcal{L}(G)$ and spanning forests of G. The details of how Eq. 5 arises from the MTT, along with a statement and proof of the MTT itself, are given in Mirzaev and Gunawardena (2013).

Since a strongly connected graph contains at least one directed path from each vertex to every other vertex, there is always at least one spanning tree rooted at each vertex. Therefore, the right-hand side of Eq. 5 is never empty and has at least one term for any choice of i. However, the number of rooted spanning trees may depend on the vertex: in Figure 2, there are 20 spanning trees rooted at vertex 5 but the reader can check that there is only one spanning tree rooted at vertex 3. The size of $\rho_i(G)$ can vary markedly with i, depending on the structure of G.

It follows from Eq. 4 that $u^{\infty}(G)$ is a manifestly positive rational function of the labels and is also always nonzero, irrespective of the values of the labels. It is well known in probability theory that the steady-state probabilities of a Markov process are always positive when the corresponding graph is strongly connected, and here we not only see why this is so but also how to calculate these probabilities in terms of the transition rates.

Manifest positivity is what we would want for a formula that yields a steady-state probability. It is a striking fact that many well-known mathematical formulas of molecular biology, such as those of Michaelis–Menten and King–Altman in enzyme kinetics, Monod–Wyman–Changeux and Koshland–Némethy–Filmer in protein allostery and Ackers–Johnson–Shea in gene regulation, all have the structure of manifestly positive rational functions. However, they are typically derived in entirely different ways. In fact, all these rational functions can be shown to arise from Eqs. 4, 5 applied to appropriate linear framework graphs (Gunawardena, 2012; Wong et al., 2018; Nam et al., 2022), thereby revealing a surprising mathematical unity underlying the complexity of molecular biology.

2.2 First-passage times and spanning forests

We turn now from the steady state to the transient regime and specifically to *first-passage times* (FPTs) (Iyer-Biswas and Zilman, 2016). Given a graph G, the FPT from one vertex, i, to a distinct target vertex, $j \neq i$, is the random variable for the time it takes the underlying Markov process, $X(t)$, to reach j for the first time when starting from i. Formally,

$$\Theta_{i,j}(G) = \inf\{t > 0 : \; X(t) = j \mid X(0) = i\}.$$

Of interest are the mean and higher moments of the FPT distribution. *Recurrence times* for the process returning to i after leaving i can be treated similarly, as can FPTs for reaching a subset of target states from a distinct subset of initial states, but we will leave these refinements aside so as not to complicate the discussion.

For the kinds of stochastic molecular systems considered here, FPTs have been used to quantify several properties: the completion time of an enzymatic turnover (Fisher and Kolomeisky, 1999; Kou et al., 2005; Shaevitz et al., 2005; Kolomeisky and Fisher, 2007; Chemla et al., 2008; Garai et al., 2009; Bel et al., 2010; Moffitt et al., 2010; Cao, 2011; Moffitt and Bustamante, 2014); the speed with which an enzyme can discriminate between correct and incorrect substrates (Banerjee et al., 2017; Cui and Mehta, 2018; Mallory et al., 2019); the statistical structure of transcriptional bursting (Lammers et al., 2020); and the time by which a regulated molecule crosses an abundance threshold (Co et al., 2017; Ghusinga et al., 2017; Gupta et al., 2018). We briefly discuss two examples by way of motivation before proceeding to the technical details.

The development of single-molecule techniques for visualising transcription in live cells (Fukaya et al., 2016; Dufourt et al., 2018) has revealed that transcription is often characterised by transient "bursts" of mRNA expression interspersed by periods of inactivity. Efforts to explain how such bursting arises have focussed on stochastic transitions between transcriptionally active and inactive states in a Markovian setting (Peccoud and Ycart, 1995; Lammers et al., 2020). In active states, successive mRNAs are produced in a burst, which is terminated when the system makes a transition to an inactive state. The FPT to reach an active state from an inactive one provides an estimate of the time between bursts, which can be measured experimentally. As noted by Lammers et al. (2020), comparing the distributions of such FPTs offers a sensitive means to discriminate between different gene regulatory models.

FPTs have also been used to quantify the time at which a regulated molecule reaches a specific abundance threshold (Co et al., 2017; Ghusinga et al., 2017; Gupta et al., 2018). An example of this type of system is bacterial lysis by phage λ. Upon infecting *Escherichia coli*, phage λ expresses a protein, holin S105, that accumulates in the inner cell membrane until a threshold concentration is reached, at which point the holin molecules abruptly initiate lysis by puncturing the membrane with large irregular holes (White et al., 2010). Various other cellular processes, such as bacterial sporulation (Piggot and Hilbert, 2004), cell cycle progression (Liu et al., 2015) and cell migration during development (Gupta et al., 2018), rely on similar thresholding mechanisms. The FPT analysis undertaken by Ghusinga et al. (2017) shows the impact of different regulatory strategies on the variance in the FPT to reach the threshold and gives insight into the regulatory mechanism of bacterial lysis.

Despite their broad usefulness in biology, FPTs have often been calculated by numerical simulations (Lammers et al., 2020) or by analytical methods that rely on the special structure of the model (Ghusinga et al., 2017). We describe here a systematic graph-theoretic scheme, similar to that in Eq. 5, by which the moments of the FPT distribution can be expressed as rational functions of the edge labels.

Since $\Theta_{i,j}(G)$ measures the time taken by $X(t)$ to reach j from i for the first time, the distribution of $\Theta_{i,j}(G)$ does not depend on the outgoing edges from j or their labels. Therefore, one can remove from G the edges leaving j without affecting the distribution of $\Theta_{i,j}(G)$. For example, the distribution of $\Theta_{i,5}(G)$ is the same for the strongly connected graph in Figure 2A and for the graph in Figure 3A, which is formed by removing the edges leaving 5 from the graph in Figure 2A. In consequence, it is convenient when working with FPTs to deal with graphs that may not be strongly connected, for which some additional terminology is helpful.

A graph G always has a unique decomposition into *strongly connected components* (SCCs), which can be thought of as the maximal strongly connected subgraphs; see Mirzaev and Gunawardena (2013) for the full details. The directed edges

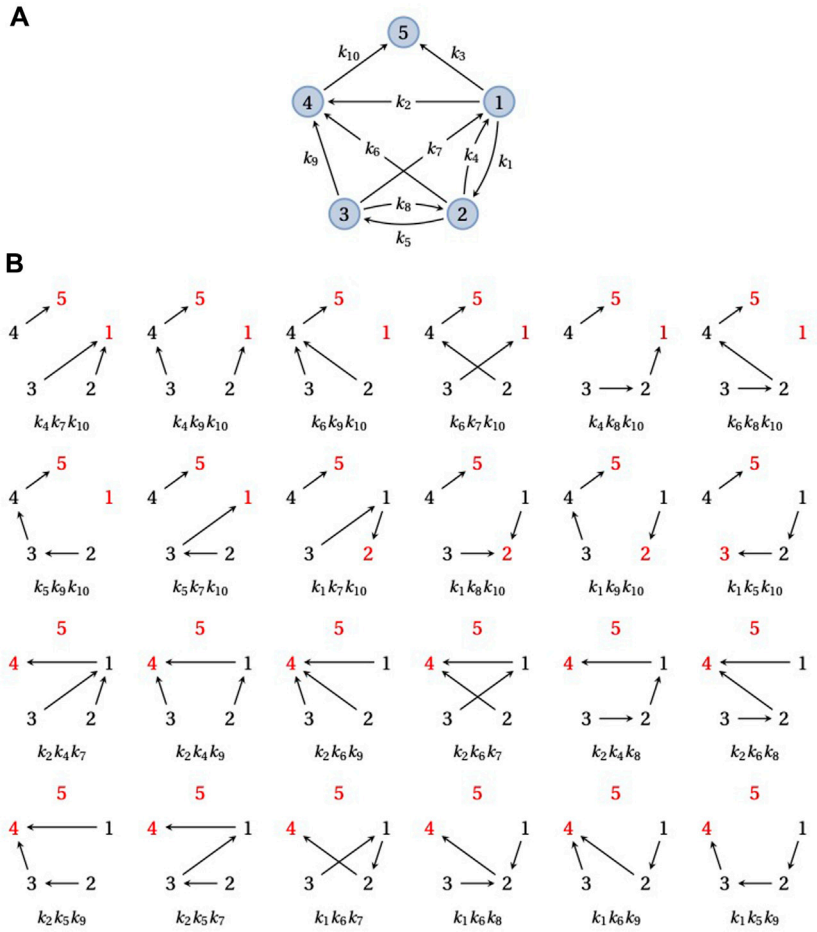


FIGURE 3
Spanning forests and FPTs. **(A)** An example graph, G, obtained by taking the graph in Figure 2A and removing the outgoing edge from vertex 5. G has a single terminal SCC containing the single vertex 5. **(B)** The 24 doubly-rooted spanning forests of G in which 5 is a root (red font) and there is a path from 1 to the other root (also in red font), each with its corresponding product of edge labels. The sum of these 24 edge label products is equal to the numerator of $\tau^{(1)}_{1,5}(G)$ in Eq. 6.

which leave these SCCs give rise to a *partial order* on the set of SCCs. Those SCCs which are maximal in the partial order are called *terminal*. For example, the graph in Figure 2A is strongly connected and therefore has only a single SCC, but if the edge 5 → 1 is removed, to yield the graph in Figure 3A, this graph has 3 SCCs in the partial order $\{1, 2, 3\} \preceq \{4\} \preceq \{5\}$. Let us consider the special case where G has a unique terminal SCC that contains just one vertex, say, $q \in \{1, \ldots, N\}$, like the graph in Figure 3A. This is what happens upon removal of the edges leaving a vertex, q, in a strongly connected graph, as in Figure 2A: q forms a unique terminal SCC, $\{q\}$, with only one vertex. If the underlying Markov process $X(t)$ starts from any other vertex, say i, then the probability that $X(t)$ eventually reaches q is 1. There may, of course, be trajectories of the process along which q is never reached but these form a set of probability zero.

We need just a bit more notation. The quantities we want to calculate are the kth moments of the probability distribution of the FPT from i to q,

$$\tau^{(k)}_{i,q}(G) = \langle\, \Theta_{i,q}(G)^k \,\rangle,$$

where $\langle - \rangle$ denotes the average over the underlying sample space of trajectories. Let $\mathcal{I}$ denote the subset of non-terminal vertices, $\mathcal{I} = \{1, \ldots, N\} \backslash \{q\}$. Given any non-empty subset of vertices, $\varnothing \neq U \subset \{1, \ldots, N\}$, and vertices $j \in \{1, \ldots, N\}$ and $r \in U$, let $\Phi_{U:j \rightsquigarrow r}(G)$ denote the set of spanning forests of G that are rooted at U and contain a directed path of edges from j to the root r, specified by $j \rightsquigarrow r$. By convention, there is always a (trivial) directed path from any vertex to itself, so that $r \rightsquigarrow r$. Then, for the mean FPT, we have (Nam and Gunawardena, 2023),

$$\tau^{(1)}_{i,q}(G) = \frac{\sum_{j \in \mathcal{I}} \sum_{F \in \Phi_{\{j,q\}:\, i \rightsquigarrow j}(G)} w(F)}{\sum_{F \in \Phi_{\{q\}}(G)} w(F)}. \tag{6}$$

The numerator in Eq. 6 runs over all doubly-rooted spanning forests of G in which q is one root and there is a directed path of edges from i to the other root. Figure 3B demonstrates this calculation for the graph in Figure 3A. The denominator in Eq. 6 runs over all spanning trees of G rooted at q and is similar in that respect to the right-hand side of Eq. 5.

The combinatorics become more complicated for the higher moments of $\Theta_{i,q}(G)$. Choose k-tuples of non-terminal vertices,

$$(j_1, \ldots, j_k) \in \underbrace{\mathcal{I} \times \cdots \times \mathcal{I}}_{k \text{ times}},$$

and set $j_0 = i$. Then, for the kth moment, we have (Nam and Gunawardena, 2023),

$$\tau_{i,q}^{(k)}(G) = \frac{k! \sum_{(j_1,\ldots,j_k)} \left(\prod_{u=1}^{k} \left(\sum_{F \in \Phi_{\{j_u,q\}:\, j_{u-1} \rightsquigarrow j_u}(G)} w(F) \right) \right)}{\left(\sum_{F \in \Phi_{\{q\}}(G)} w(F) \right)^k}. \quad (7)$$

The product in the numerator of Eq. 7 again involves doubly-rooted spanning forests, in which q is one of the roots and the other root shifts along the k-tuple from j_1 to j_k, with j_{u-1} having a directed path to j_u as u runs from 1 to k. Eq. 7 reduces to Eq. 6 when $k = 1$.

Note that a spanning forest, or the special case of a spanning tree, that has q as a root cannot include any outgoing edge from q. Hence, the spanning forests or trees with $q = 5$ as a root are the same for the strongly connected graph in Figure 2A as for the graph in Figure 3A, in which $\{q\}$ has become the unique terminal SCC by removing the edges that leave q. Accordingly, both the numerator and denominator in Eqs. 6, 7 give the same result for $q = 5$ in either graph. This is the graph-theoretic consequence of the fact, mentioned above, that the probability distribution of $\Theta_{i,5}(G)$ is the same for the graphs in Figure 2A and Figure 3A.

Eq. 7 and, by specialisation, Eq. 6 can be derived, after some manipulations, from the All-Minors Matrix-Tree theorem, a more recent generalisation of the classical MTT (Nam and Gunawardena, 2023).

As a sanity check on Eq. 7, we note that if G has N vertices, then any spanning forest with r roots has $N - r$ edges, as can be checked for the examples in Figure 2B and Figure 3B. It follows from Eq. 7 that $\tau_{i,q}^{(k)}(G)$ has dimensions of (time)^k, as expected for the kth moment of an FPT.

Let us see what Eq. 7 tells us for the graph G consisting of just two vertices, 1 and 2, with $\ell(1 \to 2) = a$ and $\ell(2 \to 1) = b$. If we consider $\tau_{1,2}^{(k)}(G)$, then, for the denominator of Eq. 7, we need the spanning trees rooted at 2, given by $\Phi_{\{2\}}(G)$. There is only one such tree F, for which $w(F) = a$. As for the numerator, we need the spanning forests rooted at j_u and 2, given by $\Phi_{\{j_u,2\}:\, j_{u-1} \rightsquigarrow j_u}(G)$. Since the roots have to be distinct, the only possibility is that $j_u = 1$. But then the only forest, F, with these roots has just these vertices and no edges. Recalling the convention for what happens when there are no edges, we find that $w(F) = 1$. It follows that Eq. 7 collapses to the simple conclusion that

$$\tau_{1,2}^{(k)(G)} = \frac{k!}{a^k}.$$

In particular, the mean FPT is $1/a$ and the variance, which is $\tau_{1,2}^{(2)}(G) - (\tau_{1,2}^{(1)}(G))^2$, is $1/a^2$. Only the rate a is relevant, as we would expect, since the rate b is the label on an edge that leaves the target vertex. Because this example is so simple, the moments of the FPT distribution can be readily calculated without the paraphernalia of Eq. 7. The case of a longer pipeline of vertices is more demanding, as we will see below (Figure 5).

Eq. 7 gives a general and systematic method to calculate FPTs from the linear framework graph associated with a Markov process. It can be used to calculate exact formulas in simple graphs and to avoid estimating FPT moments by cumbersome numerical simulations of the Markov process. The combinatorics rapidly become formidable as the graph becomes larger or less symmetric, as is perhaps already evident in Figure 2B and Figure 3B. The broader value of Eq. 7 is that it reveals the mathematical structure of the FPT moments as manifestly positive rational functions of the edge labels. This can often be informative in its own right, as we will see in discussing enzyme kinetics below. We will say more about ways of dealing with the combinatorial complexity in the Discussion.

2.3 Splitting probabilities and conditional FPTs

In the previous section, we considered the FPT distribution from a given vertex i to a single target vertex. It is, however, often the case that there are several target vertices and one wants to know the probability of reaching a particular target vertex or the FPT to that vertex conditioned on the Markov process actually reaching it. (If target vertices lie in different SCCs that are not related in the partial order, then a trajectory that reaches one target can never reach any other target, so that the mean FPT to each target becomes infinite. Conditioning on reaching the target is therefore essential.) Let us suppose, therefore, that G is a graph with one or more terminal SCCs, each of which consists of a single vertex. Let $\mathcal{T} \subset \{1, \ldots, N\}$ be the subset consisting of these terminal vertices. Given $i \in \{1, \ldots, N\}$ and $q \in \mathcal{T}$, define the *splitting probability from i to q*, denoted $\pi_{i,q}(G)$, to be the probability that the underlying Markov process, when started from i, eventually reaches q, as opposed to any other terminal vertex. Then we have (Nam and Gunawardena, 2023),

$$\pi_{i,q}(G) = \frac{\sum_{F \in \Phi_{\mathcal{T}:\, i \rightsquigarrow q}(G)} w(F)}{\sum_{F \in \Phi_{\mathcal{T}}(G)} w(F)}. \quad (8)$$

The denominator in Eq. 8 runs over all spanning forests of G rooted at $\mathcal{T}$, and the numerator runs over the subset of those spanning forests in which there is a directed path of edges from i to the root q. Accordingly, the right-hand side of Eq. 8 must lie between 0 and 1, as expected for a probability. If $i \in \mathcal{T}$ and $i \neq q$, then there is no directed path from i to q and so Eq. 8 gives 0, while if $i = q$, then every spanning forest has a (trivial) path of directed edges from i to q and so Eq. 8 gives 1. If G contains only one terminal vertex, then every spanning forest of G rooted at $\mathcal{T} = \{q\}$ has a path of directed edges from i to q, and so Eq. 8 again gives 1. Figure 4 illustrates the calculation of the splitting probability from $i = 1$ to $q = 5$ on a six-vertex graph with two terminal vertices, 5 and 6.

Let us turn now to the conditional FPT for reaching a particular target vertex, $q \in \mathcal{T}$, from the vertex $i \in \mathcal{I}$, where, as before, $\mathcal{I}$ is the subset of non-terminal vertices, $\mathcal{I} = \{1, \ldots, N\} \backslash \mathcal{T}$. For the mean conditional FPT from $i \in \mathcal{I}$ to $q \in \mathcal{T}$, denoted by $\chi_{i,q}^{(1)}(G)$, we find that (Nam and Gunawardena, 2023),

$$\chi_{i,q}^{(1)}(G) = \frac{\sum_{j \in \mathcal{I}} \left(\sum_{F \in \Phi_{\mathcal{T} \cup \{j\}:\, i \rightsquigarrow j}(G)} w(F) \right) \left(\sum_{F \in \Phi_{\mathcal{T}:\, j \rightsquigarrow q}(G)} w(F) \right)}{\left(\sum_{F \in \Phi_{\mathcal{T}:\, i \rightsquigarrow q}(G)} w(F) \right) \left(\sum_{F \in \Phi_{\mathcal{T}}(G)} w(F) \right)}. \quad (9)$$

If there is only one terminal vertex, so that $\mathcal{T} = \{q\}$, then the mean conditional FPT, $\chi_{i,q}^{(1)}(G)$, as given by Eq. 9, is equal to the mean FPT,

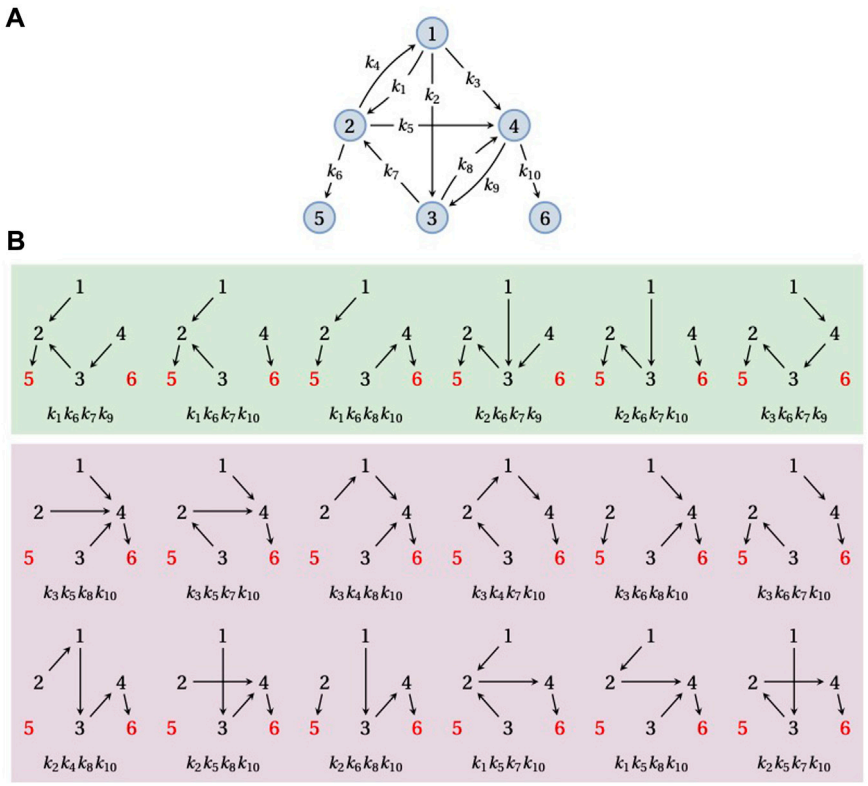


FIGURE 4
Splitting probabilities. **(A)** An example graph, G, on six vertices, $\{1, \ldots, 6\}$, with three SCCs. The partial order is given by $\{1, 2, 3, 4\} \preceq \{5\}$ and $\{1, 2, 3, 4\} \preceq \{6\}$, with $\{5\}$ and $\{6\}$ being the two terminal SCCs. **(B)** The 18 spanning forests of G rooted at vertices 5 and 6 (red font), with those containing a path from 1 to 5 in the green box and those containing a path from 1 to 6 in the purple box. Each spanning forest is shown with its corresponding product of edge labels. The sum of all 18 edge label products is equal to the denominator of $\pi_{1,5}(G)$ in Eq. 8; the sum of the six edge label products in the green box is equal to the numerator of $\pi_{1,5}(G)$ in Eq. 8.

$\tau_{i,q}^{(1)}(G)$, as given by Eq. 6. Formulas for the higher moments of the conditional FPT can be obtained in a similar way.

Evidently, the unconditional mean FPT to reach any terminal vertex in $\mathcal{T}$ from i, denoted $\psi_i^{(1)}(G)$, is now given by,

$$\psi_i^{(1)}(G) = \sum_{p \in \mathcal{T}} \pi_{i,p}(G)\, \chi_{i,p}^{(1)}(G).$$

Combining Eqs. 8, 9, we can show that this mean FPT can also be expressed in terms of the spanning forests of G, as

$$\psi_i^{(1)}(G) = \frac{\sum_{j \in \mathcal{I}} \left(\sum_{F \in \Phi_{\mathcal{T} \cup \{j\}:\, i \to j}(G)} w(F) \right)}{\left(\sum_{F \in \Phi_{\mathcal{T}}(G)} w(F) \right)}, \tag{10}$$

which specialises to Eq. 6 when there is only a single terminal vertex.

Splitting probabilities and conditional FPTs have not been as widely used as have the unconditional FPTs described in the previous section. This reflects the relatively simple models that have been formulated so far in the literature. However, as we have shown here, there is no greater difficulty in dealing with these more complex quantities, at least within the graph-theoretic approach that we have outlined here. All the quantities we have considered are manifestly positive rational functions of the edge labels. This mathematical accessibility should allow deeper analysis of transient stochastic properties.

2.4 Single-molecule enzyme kinetics

Single-molecule experimental methods have given unprecedented access to the stochastic kinetics of individual enzymes and have stimulated the development of theoretical models to account for the resulting data. This literature offers a convenient setting to illustrate the ideas introduced above.

A frequently used model in enzyme kinetics corresponds to a *pipeline* graph (Figure 5) (Fisher and Kolomeisky, 1999; Kou et al., 2005; Kolomeisky and Fisher, 2007; Chemla et al., 2008; Garai et al., 2009; Moffitt et al., 2010; Moffitt and Bustamante, 2014). Such a graph consists of vertices 1, …, N, representing different conformations of the enzyme, with nearest-neighbour transitions, $i \to i+1$ or $i \to i-1$. Substrate may bind at any forward transition, $i \to i+1$, so that $\ell(i \to i+1)$ incurs a concentration term that we will denote by x, and binding is assumed to be reversible, so that $i+1 \to i$. The final transition, $N-1 \to N$, is usually treated as an irreversible catalytic step, with the enzyme returning to its initial conformation, so that vertex N corresponds to vertex 1 in the next enzymatic cycle. A pipeline may be thought of as partitioned into reversible "blocks" that are separated by sequences of irreversible transitions. Figures 5A, C show pipeline graphs with 1 and 3 reversible blocks, respectively.

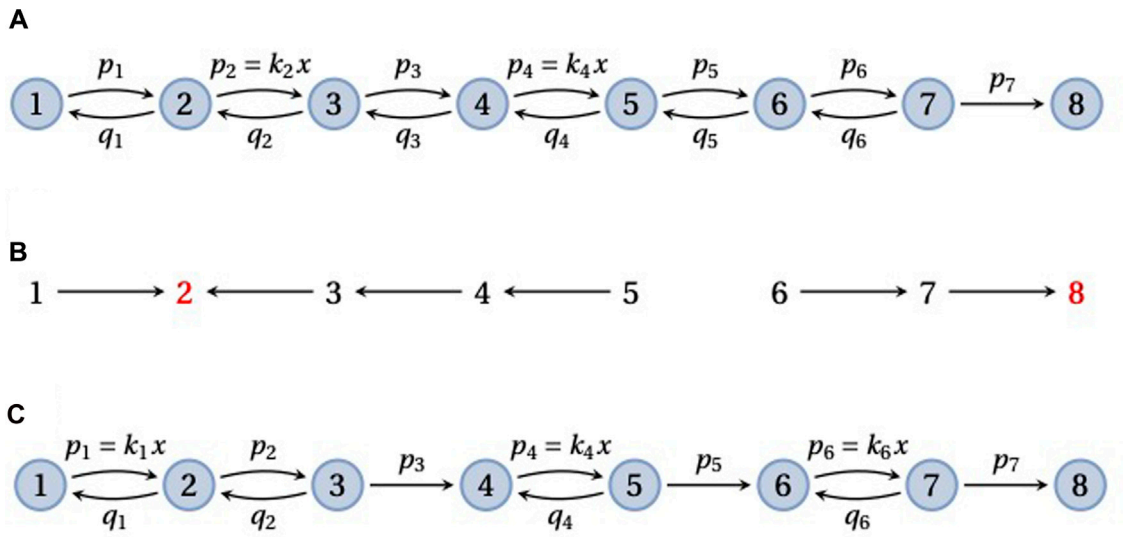


FIGURE 5
Pipeline graphs. **(A)** A pipeline graph on 8 vertices that consists of a single reversible block, with substrate binding with concentration x at the edges 2 → 3 and 4 → 5, followed by a single irreversible transition, 7 → 8. **(B)** The spanning forest $F(2, 6, 8)$, in the notation described in the text, for the graph in panel **A**. The two roots, 2 and 8, are in red font. **(C)** A pipeline graph with three reversible blocks, in each of which the substrate binds once. As explained in the text, the mean FPT, $\tau_{1,8}^{(1)}(G)$, has a reciprocal Michaelis–Menten dependence on the substrate concentration, x, as in Eq. 15.

The mean FPT for reaching vertex N from vertex 1 is a measure of the enzyme's completion time. Bustamante and colleagues have emphasised how the substrate dependence of $\tau_{1,N}^{(1)}(G)$ and $\tau_{1,N}^{(2)}(G)$ contains information about the enzyme mechanism, and they have built on previous studies (Derrida, 1983) to analyse this theoretically (Moffitt et al., 2010). This amounts to studying $\tau_{1,N}^{(1)}(G)$ and $\tau_{1,N}^{(2)}(G)$ as functions of x, which falls directly into the scope of the results described above. We will show how the graph-theoretic methods introduced here provide a straightforward way to recover some of these previous findings. We do not intend to be exhaustive and there is much more of interest in the cited references. We hope, rather, to show the advantages of the graph-theoretic approach over the variety of approaches used previously, such as recursive solution of the master equation (Derrida, 1983) or Fourier transformation and determinants (Chemla et al., 2008).

Consider first a pipeline graph, G, with a single reversible block consisting of the vertices $1, \ldots, N-1$ and recall Eq. 6 for the mean FPT, where the terminal vertex is $q = N$. An example is shown in Figure 5A with the notation that we will use for the edge labels, $\ell(i \to i+1) = p_i$ and $\ell(i+1 \to i) = q_i$. It is evident that there is only a single spanning tree, $T \in \Phi_{\{N\}}(G)$, consisting of all the forward edges, so that $w(T) = p_1 \cdots p_{N-1}$. This gives the denominator of $\tau_{1,N}^{(1)}(G)$. As for the doubly-rooted spanning forests of $\Phi_{\{j,N\}}(G)$ in the numerator, they can be indexed as $F(j, k, N)$, where $j < k \leq N$ and k is the vertex with the smallest index that has a directed path to the root N (Figure 5B). Furthermore, each such forest has a directed path from 1 to the root j, so that $\Phi_{\{j,N\}:1 \rightsquigarrow j}(G) = \Phi_{\{j,N\}}(G)$. We see from the labels in Figure 5B that

$$w(F(j,k,N)) = p_1 \cdots p_{j-1} q_j \cdots q_{k-2} p_k \cdots p_{N-1}, \tag{11}$$

where the "missing" label, between vertices $k-1$ and k, corresponds to the gap between the tree rooted at j and the tree rooted at N in the forest. If we divide by the denominator, we see that each spanning forest $F(j, k, N)$ contributes a rational function of the labels that we may write in the form,

$$\frac{w(F(j,k,N))}{w(T)} = \frac{1}{p_j} \prod_{u=j}^{k-2} \frac{q_u}{p_{u+1}}.$$

The spanning forests in $\Phi_{\{j,N\}}(G)$ therefore contribute the sum,

$$\frac{\sum_{F \in \Phi_{\{j,N\}:1 \rightsquigarrow j}(G)} w(F)}{w(T)} = \frac{\Delta(j,N)}{p_j},$$

where,

$$\Delta(j,N) = \sum_{k=j+1}^{N} \left(\prod_{u=j}^{k-2} \frac{q_u}{p_{u+1}} \right). \tag{12}$$

Note that, in Eq. 12, the empty product for $k = j+1$ is by convention taken to be 1. It follows from Eq. 6 that the enzyme completion time is given by,

$$\tau_{1,N}^{(1)}(G) = \sum_{j=1}^{N-1} \frac{\Delta(j,N)}{p_j}. \tag{13}$$

With some notational translation, Eq. 13 can be seen to be the same as (Moffitt et al., 2010, Eq. S2). The quantity $\Delta(j, N)$ in Eq. 12 first appears in Derrida's derivation of the velocity and diffusion constant of a Markov particle on a periodic pipeline (Derrida, 1983, Eq. 24); $\Delta(j, N) = \Gamma(j+1, N-1)$, where Γ is the quantity defined in Eq. S3 of Moffitt et al. (2010). The calculation above, using the general formula for the mean FPT in Eq. 6, is hopefully more transparent.

Suppose now that substrate binds at s forward transitions in the pipeline graph, with concentration x. We will refer to terms other than x in the edge labels as "kinetic parameters," which thereby include both simple rates and on-rates. Since we can exclude the final catalytic transition from substrate binding, it follows that $1 \leq$

$s \leq N-2$. Eq. 11 then shows that the enzyme completion time has the following structure as a rational algebraic function of x,

$$\tau_{1,N}^{(1)}(G) = \frac{a_0 + a_1 x + \cdots + a_s x^s}{b x^s}. \tag{14}$$

Here, the coefficients $a_0, \ldots, a_s$ and b are all manifestly positive polynomials in the kinetic parameters. In particular, the forest $F(N-1, N, N)$ includes all the substrate-binding transitions, which confirms that $a_s > 0$. If the substrate-binding transitions are specified, these polynomials may be explicitly calculated using Eq. 11. Eq. 14 already provides some insight. In the limit of low substrate, the completion time diverges at an order, $1/x^s$, that depends on the number of substrate-binding transitions. In contrast, in the limit of high substrate, the completion time asymptotes to the positive value a_s/b. If substrate binds at only one transition in the pipeline, so that $s = 1$, then the completion time exhibits a reciprocal Michaelis–Menten form (Kou et al., 2005; Garai et al., 2009; Moffitt et al., 2010; Moffitt and Bustamante, 2014) (Discussion),

$$\tau_{1,N}^{(1)}(G) = \frac{a_0 + a_1 x}{b x}. \tag{15}$$

The higher moments of the FPT, as specified by Eq. 7, are more complicated to calculate but the doubly-rooted spanning forests that are needed for the numerator, which are contained in $\Phi_{\{j_u,N\}:\, j_{u-1} \rightsquigarrow j_u}(G)$, have already been enumerated by the forests $F(j, k, N)$ introduced above (Figure 5B). It seems reasonable to conclude from Eq. 7 that $\tau_{1,N}^{(k)}(G)$ has a similar rational algebraic structure as shown in Eq. 14 but with a degree of ks for both the numerator and the denominator. In particular, if substrate binds at only one transition, so that $s = 1$, the second moment of the FPT is a quadratic rational function (Moffitt et al., 2010).

In their study of the packaging motor for the φ29 bacteriophage, Bustamante and colleagues consider a more general pipeline graph, G, that consists of multiple reversible blocks separated by single irreversible transitions (Figure 5C) (Moffitt et al., 2010). The packaging motor is a pentameric ring of identical ATPase units that compacts the φ29 double-stranded DNA into the assembling viral capsid. It has been found to do this in a burst of four ATP-consuming steps per cycle. ATP hydrolysis during the catalytic step is typically irreversible under physiological conditions and a pipeline with 4 reversible blocks serves as a model for the motor (Moffitt et al., 2010, Figure 4A).

If the Markov process takes an irreversible transition in G, it cannot subsequently visit the preceding reversible blocks. Also, every irreversible transition must be taken to reach N. Hence, any trajectory that begins at 1 and reaches N must take each irreversible transition exactly once. It follows from this that the FPT from 1 to N is just the sum of the FPTs for each reversible block considered separately and these FPTs are all independent of each other. Suppose there are m reversible blocks which start at the vertices $e_0, e_1, \ldots, e_{m-1}$, where $1 = e_0 < e_1 < e_2 < \cdots < e_{m-1} < N$. Let G_i be the subgraph consisting of the vertices from e_{i-1} to e_i, which includes the ith reversible block and the immediately following irreversible transition. It follows that,

$$\tau_{1,N}^{(k)}(G) = \tau_{1,e_1}^{(k)}(G_1) + \tau_{e_1,e_2}^{(k)}(G_2) + \cdots + \tau_{e_{m-1},N}^{(k)}(G_m). \tag{16}$$

If substrate binds at the same number of transitions in each reversible block, then Eq. 7 shows that the $\tau_{e_{i-1},e_i}^{(k)}(G_i)$ all have the same rational algebraic structure with the same degrees in both the numerator and the denominator. It follows from Eq. 16 that $\tau_{1,N}^{(k)}(G)$ must also have this same rational algebraic structure. For the case of the φ29 packaging motor, ATP binds at only one transition in each reversible block, so the completion time has the reciprocal Michaelis–Menten form of Eq. 15 and the resulting curve may be fitted to the experimental data (Moffitt et al., 2010, Figure 3A). Bustamante and colleagues make use of the reciprocal of the coefficient of variation,

$$n_{\min} = \frac{\left(\tau_{1,N}^{(1)}(G)\right)^2}{\tau_{1,N}^{(2)}(G) - \left(\tau_{1,N}^{(1)}(G)\right)^2},$$

which is readily seen from the discussion above to be a quadratic rational function of x, and they also fit this curve to the experimental data (Moffitt et al., 2010, Figure 3B). A theorem due to Aldous and Shepp (1987), which is of independent interest, tells us that, for an arbitrary graph with N vertices, $n_{\min} < N$.

An interesting question arises as to whether $n_{\min}$ itself is also manifestly positive, as might be expected of a coefficient of variation, given that this is true for both $\tau_{1,N}^{(1)}(G)$ and $\tau_{1,N}^{(2)}(G)$. A further point made by Moffitt et al. (2010) is that the quadratic structure of $n_{\min}$ may not be limited to pipeline graphs but may be true also for some graphs with branches and parallel pathways. If so, the graph-theoretic methods described here offer a way to generalise their findings.

3 Discussion

We have reviewed here how the graph-theoretic linear framework, as applied to continuous-time Markov processes, can be used to show that the moments of the FPT distribution (Eqs. 6, 7), splitting probabilities (Eq. 8) and conditional mean FPTs (Eq. 9) can be exactly expressed as manifestly positive rational algebraic functions of the edge labels or transition rates. This reveals that not only steady-state probabilities but also transient properties of Markov processes have this same algebraic structure, thereby substantially expanding the mathematical scope of the linear framework.

The formulas given here can be used to obtain closed-form solutions for simple graphs, as we showed for the pipeline graphs used in enzyme kinetics (Eq. 13). However, this is a little misleading because enumeration of spanning forests becomes rapidly intractable as the graph becomes larger or less symmetric. Moreover, as is evident by examining the algebraic terms in Figure 2B and Figure 3B, every label in the graph can appear in the formulas. There is both a combinatorial explosion and a global parametric dependence. These challenges have long been recognised when dealing with steady-state probabilities (Nam et al., 2022), before the transient regime became mathematically accessible, and several strategies have emerged for dealing with them.

First, when properties of interest are treated as functions of substrate concentration, a great deal can be said about the resulting rational algebraic structure, even when it is hard to calculate the coefficients explicitly in terms of the edge labels (Thomson and Gunawardena, 2009; Nam et al., 2022). As we saw with Eq. 14, the algebraic structure for the mean FPT, $\tau_{1,N}^{(1)}(G)$, is highly informative,

especially with respect to the limits of low or high concentration, which may also be experimentally accessible. The Michaelis–Menten structure, or its reciprocal in Eq. 15, arises in a remarkably wide range of biological contexts that are far removed from the 3-vertex pipeline graph considered, in effect, by Michaelis and Menten (Michaelis and Menten, 1913). The linear framework allows general theorems to be proved, which characterise many of the contexts in which the Michaelis–Menten structure does appear (Wong et al., 2018). In this respect, the context discussed above, of a pipeline graph with multiple reversible blocks, in which substrate binds once in each block, falls outside the scope of the theorems in Wong et al. (2018). As suggested by Moffitt et al. (2010), it seems plausible that the Michaelis–Menten structure may also arise for more complicated graphs and an interesting problem arises in characterising this new context.

Second, the question of when the Michaelis–Menten structure arises is closely related to whether or not the graph satisfies the cycle condition and can thereby reach a steady state of thermodynamic equilibrium. If it can, there is a necessary and sufficient condition for the emergence of the Michaelis–Menten structure; if it cannot, and the graph reaches a non-equilibrium steady state, then only partial sufficient conditions are known (Wong et al., 2018). Of course, the pipeline example just mentioned cannot reach thermodynamic equilibrium, as it contains irreversible transitions (Figure 5A). If the cycle condition is satisfied, the complexity problem is substantially reduced, insofar as calculating steady-state probabilities is concerned. It is possible to find an alternative basis element to $\rho(G)$ in $\ker\mathcal{L}(G)$ (Eq. 5), which is based on paths rather than spanning trees, for which the combinatorial explosion disappears and the parametric dependence becomes local, not global (Nam et al., 2022). It is a very interesting question as to whether transient quantities like FPTs show any similar reduction in complexity for graphs that satisfy the cycle condition.

Aside from the calculational complexity, the thermodynamic issues also have a deep impact on biological function. The role of energy expenditure in force generation or pattern formation has been widely studied (Kolomeisky and Fisher, 2007; Karsenti, 2008) but its significance for cellular information processing has been more elusive (Wong and Gunawardena, 2020). In the latter domain, unlike the two former ones, information processing can take place at thermodynamic equilibrium, for instance, through binding and unbinding. However, there is a limit to how well this can be done, as first pointed out by Hopfield (1974). We have introduced the concept of the *Hopfield barrier*, as the limit to how well a given information processing task can be undertaken by a mechanism that operates at thermodynamic equilibrium (Estrada et al., 2016). For example, the Hill function with Hill coefficient n is the universal Hopfield barrier for the sharpness of input-output responses with n binding sites for the input (Nam et al., 2022; Martinez-Corral et al., 2023). Another interesting question arises as to whether there are also Hopfield barriers in the transient regime. That is, if a graph satisfies the cycle condition and can reach a steady state of thermodynamic equilibrium, are there limits on the moments of the FPT distribution, $\tau_{i,q}^{(k)}(G)$, which can only be exceeded if energy is expended to break the cycle condition, allowing the system to reach a non-equilibrium steady state?

Third, the algebraic complexity of non-equilibrium steady states can be reorganised to make the complexity more tractable (Çetiner and Gunawardena, 2022). This breakthrough has enabled steady-state calculations to be undertaken that were previously out of reach. It is conceivable that similar kinds of reorganisation may also throw light on the calculation of transient quantities. Finally, a fourth potential approach to overcoming the complexity is to exploit the recursive technique for enumerating spanning forests that was developed by Chebotarev and Agaev (2002). While this technique looks promising, it has yet to be properly exploited.

The methods outlined here bring the FPTs of Markov processes into focus as manifestly positive rational algebraic functions of the transition rates. This gives mathematical access to them in a way that has been lacking in previous treatments, which have not exploited graph theory and the Matrix-Tree theorems. We hope this review will encourage more use of the linear framework in cell and developmental biology. We anticipate that, as we have found for steady states, this exploration will lead to further general principles and mathematical theorems that rise above the molecular complexity that confronts us in biology.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

K-MN undertook most of the work described here in his Ph.D. thesis, which was supervised by JG. K-MN and JG wrote the paper together. All authors contributed to the article and approved the submitted version.

Funding

K-MN and JG were supported in part by NIH grant R01GM122928.

Acknowledgments

We thank Michael Blinov for the invitation to submit a paper to this research topic and for his encouragement and patience; two reviewers for their constructive suggestions; and members of the Gunawardena lab for their comments.

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

Aldous, D., and Shepp, L. (1987). The least variable phase-type distribution is Erlang. *Commun. Stat. Stoch. Models* 3, 467–473. doi:10.1080/15326348708807067

Banerjee, K., Kolomeisky, A. B., and Igoshin, O. A. (2017). Elucidating interplay of speed and accuracy in biological error correction. *Proc. Natl. Acad. Sci. U. S. A.* 114, 5183–5188. doi:10.1073/pnas.1614838114

Barbeau, E. J. (1989). *Polynomials*. Springer-Verlag.

Bel, G., Munsky, B., and Nemenman, I. (2010). The simplicity of completion time distributions for common complex biochemical processes. *Phys. Biol.* 7, 016003. doi:10.1088/1478-3975/7/1/016003

Biddle, J. W., Nguyen, M., and Gunawardena, J. (2019). Negative reciprocity, not ordered assembly, underlies the interaction of Sox2 and Oct4 on DNA. *eLife* 8, e41017. doi:10.7554/eLife.41017

Biddle, J. W., Martinez-Corral, R., Wong, F., and Gunawardena, J. (2021). Allosteric conformational ensembles have unlimited capacity for integrating information. *eLife* 10, e65498. doi:10.7554/eLife.65498

Cao, J. (2011). Michaelis–Menten equation and detailed balance in enzymatic networks. *J. Phys. Chem. B* 115, 5493–5498. doi:10.1021/jp110924w

Çetiner, U., and Gunawardena, J. (2022). Reformulating non-equilibrium steady states and generalized hopfield discrimination. *Phys. Rev. E* 106, 064128. doi:10.1103/PhysRevE.106.064128

Chebotarev, P., and Agaev, R. (2002). Forest matrices around the Laplacian matrix. *Lin. Alg. Appl.* 356, 253–274. doi:10.1016/s0024-3795(02)00388-9

Chemla, Y. R., Moffitt, J. R., and Bustamante, C. (2008). Exact solutions for kinetic models of macromolecular dynamics. *J. Chem. Phys. B* 112, 6025–6044. doi:10.1021/jp076153r

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J. B., and Gregor, T. (2018). Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.* 50, 1296–1303. doi:10.1038/s41588-018-0175-z

Chung, F. R. K. (1997). Spectral graph theory. *No. 92 in regional conference series in mathematics*. Providence, RI, USA: American Mathematical Society.

Co, A. D., Lagomarsino, M. C., Caselle, M., and Osella, M. (2017). Stochastic timing in gene expression for simple regulatory strategies. *Nucleic Acids Res.* 45, 1069–1078. doi:10.1093/nar/gkw1235

Cui, W., and Mehta, P. (2018). Identifying feasible operating regimes for early T-cell recognition: the speed, energy, accuracy trade-off in kinetic proofreading and adaptive sorting. *PLOS ONE* 13, e0202331. doi:10.1371/journal.pone.0202331

Dasgupta, T., Croll, D. H., Owen, J. A., Vander Heiden, M. G., Locasale, J. W., Alon, U., et al. (2014). A fundamental trade-off in covalent switching and its circumvention by enzyme bifunctionality in glucose homeostasis. *J. Biol. Chem.* 289, 13010–13025. doi:10.1074/jbc.M113.546515

Derrida, B. (1983). Velocity and diffusion constant of a periodic one-dimensional hopping model. *J. Stat. Phys.* 31, 433–450. doi:10.1007/bf01019492

Dufourt, J., Trullo, A., Hunter, J., Fernandez, C., Lazaro, J., Dejean, M., et al. (2018). Temporal control of gene expression by the pioneer factor Zelda through transient interactions in hubs. *Nat. Commun.* 9, 5194. doi:10.1038/s41467-018-07613-z

Estrada, J., Wong, F., DePace, A., and Gunawardena, J. (2016). Information integration and energy expenditure in gene regulation. *Cell.* 166, 234–244. doi:10.1016/j.cell.2016.06.012

Fisher, M. E., and Kolomeisky, A. B. (1999). The force exerted by a molecular motor. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6597–6602. doi:10.1073/pnas.96.12.6597

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer control of transcriptional bursting. *Cell.* 166, 358–368. doi:10.1016/j.cell.2016.05.025

Garai, A., Chowdhury, D., Chowdhury, D., and Ramakrishnan, T. V. (2009). Stochastic kinetics of ribosomes: single motor properties and collective behavior. *Phys. Rev. E* 80, 011908. doi:10.1103/PhysRevE.80.011908

Ghusinga, K. R., Dennehy, J. J., and Singh, A. (2017). First-passage time approach to controlling noise in the timing of intracellular events. *Proc. Natl. Acad. Sci. U. S. A.* 114, 693–698. doi:10.1073/pnas.1609012114

Gunawardena, J. (2012). A linear framework for time-scale separation in nonlinear biochemical systems. *PLOS ONE* 7, e36321. doi:10.1371/journal.pone.0036321

Gunawardena, J. (2014). Time-scale separation: Michaelis and Menten's old idea, still bearing fruit. *FEBS J.* 281, 473–488. doi:10.1111/febs.12532

Gupta, S., Varennes, J., Korswagen, H. C., and Mugler, A. (2018). Temporal precision of regulated gene expression. *PLOS Comput. Biol.* 14, e1006201. doi:10.1371/journal.pcbi.1006201

G. Strang (Editor) (2022). *Introduction to linear algebra*. 6 edn (Wellesley, MA, USA: Wellesley-Cambridge Press).

Hopfield, J. J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. U. S. A.* 71, 4135–4139. doi:10.1073/pnas.71.10.4135

Iyer-Biswas, S., and Zilman, A. (2016). "First-passage processes in cellular biology," in *Advances in chemical physics*. Editors S. A. Rice and A. R. Dinner (John Wiley & Sons Inc.), 261–306.

Jones, D. L., Leroy, P., Unoson, C., Fange, D., Ćurić, V., Lawson, M. J., et al. (2017). Kinetics of dCas9 target search in *Escherichia coli*. *Science* 357, 1420–1424. doi:10.1126/science.aah7084

Karsenti, E. (2008). Self-organization in cell biology: a brief history. *Nat. Rev. Mol. Cell. Biol.* 9, 255–262. doi:10.1038/nrm2357

Kirchhoff, G. (1847). Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wind. *Ann. Phys. Chem.* 148, 497–508. doi:10.1002/andp.18471481202

Kleine Borgmann, L. A., Ries, J., Ewers, H., Ulbrich, M. H., and Graumann, P. L. (2013). The bacterial SMC complex displays two distinct modes of interaction with the chromosome. *Cell. Rep.* 3, 1483–1492. doi:10.1016/j.celrep.2013.04.005

Kolomeisky, A. B., and Fisher, M. E. (2007). Molecular motors: a theorist's perspective. *Annu. Rev. Phys. Chem.* 58, 675–695. doi:10.1146/annurev.physchem.58.032806.104532

Kou, S. C., Cherayil, B. J., Min, W., English, B. P., and Xie, X. S. (2005). Single-molecule Michaelis–Menten equations. *J. Phys. Chem. B* 109, 19068–19081. doi:10.1021/jp051490q

Lammers, N. C., Kim, Y. J., Zhao, J., and Garcia, H. G. (2020). A matter of time: using dynamics and theory to uncover mechanisms of transcriptional bursting. *Curr. Opin. Cell. Biol.* 67, 147–157. doi:10.1016/j.ceb.2020.08.001

Liao, Y., Schroeder, J. W., Gao, B., Simmons, L. A., and Biteen, J. S. (2015). Single-molecule motions and interactions in live cells reveal target search dynamics in mismatch repair. *Proc. Natl. Acad. Sci. U. S. A.* 112, E6898–E6906. doi:10.1073/pnas.1507386112

Liu, X., Wang, X., Yang, X., Liu, S., Jiang, L., Qu, Y., et al. (2015). Reliable cell cycle commitment in budding yeast is ensured by signal integration. *eLife* 4, e03977. doi:10.7554/eLife.03977

Loffreda, A., Jacchetti, E., Antunes, S., Rainone, P., Daniele, T., Morisaki, T., et al. (2017). Live-cell p53 single-molecule binding is modulated by C-terminal acetylation and correlates with transcriptional activity. *Nat. Commun.* 8, 313. doi:10.1038/s41467-017-00398-7

Mallory, J. D., Kolomeisky, A. B., and Igoshin, O. A. (2019). Trade-offs between error, speed, noise, and energy dissipation in biological processes with proofreading. *J. Phys. Chem. B* 123, 4718–4725. doi:10.1021/acs.jpcb.9b03757

Martinez-Corral, R., Nam, K.-M., DePace, A. H., and Gunawardena, J. (2023). *The Hill function is the universal Hopfield barrier for sharpness of input-output responses*. In preparation

Michaelis, L., and Menten, M. (1913). Die kinetik der Invertinwirkung. *Biochem. Z* 49, 333–369.

Mir, M., Stadler, M. R., Ortiz, S. A., Hannon, C. E., Harrison, M. M., Darzacq, X., et al. (2018). Dynamic multifactor hubs interact transiently with sites of active transcription in *Drosophila* embryos. *eLife* 7, e40497. doi:10.7554/eLife.40497

Mirzaev, I., and Gunawardena, J. (2013). Laplacian dynamics on general graphs. *Bull. Math. Biol.* 75, 2118–2149. doi:10.1007/s11538-013-9884-8

Moffitt, J. R., and Bustamante, C. (2014). Extracting signal from noise: kinetic mechanisms from a Michaelis–Menten-like expression for enzymatic fluctuations. *FEBS J.* 281, 498–517. doi:10.1111/febs.12545

Moffitt, J. R., Chemla, Y. R., and Bustamante, C. (2010). Mechanistic constraints from the substrate concentration dependence of enzymatic fluctuations. *Proc. Natl. Acad. Sci. U. S. A.* 107, 15739–15744. doi:10.1073/pnas.1006997107

Nam, K.-M., and Gunawardena, J. (2023). *Algebraic formulas for first-passage times of Markov processes in the linear framework*. In preparation.

Nam, K.-M., Gyori, B. M., Amethyst, S. V., Bates, D. J., and Gunawardena, J. (2020). Robustness and parameter geography in post-translational modification systems. *PLOS Comput. Biol.* 16, e1007573. doi:10.1371/journal.pcbi.1007573

Nam, K.-M., Martinez-Corral, R., and Gunawardena, J. (2022). The linear framework: using graph theory to reveal the algebra and thermodynamics of biomolecular systems. *Interface Focus* 12, 20220013. doi:10.1098/rsfs.2022.0013

Nam, K.-M. (2021). *Algebraic approaches to molecular information processing*. Ph.D. thesis. Harvard University.

Nandan, A., Das, A., Lott, R., and Koseska, A. (2022). Cells use molecular working memory to navigate in changing chemoattractant fields. *eLife* 11, e76825. doi:10.7554/eLife.76825

Peccoud, J., and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* 48, 222–234. doi:10.1006/tpbi.1995.1027

Piggot, P. J., and Hilbert, D. W. (2004). Sporulation of Bacillus subtilis. *Curr. Opin. Microbiol.* 7, 579–586. doi:10.1016/j.mib.2004.10.001

Shaevitz, J. W., Block, S. M., and Schnitzer, M. J. (2005). Statistical kinetics of macromolecular dynamics. *Biophys. J.* 89, P2277–P2285. doi:10.1529/biophysj.105.064295

Thomson, M., and Gunawardena, J. (2009). The rational parameterization theorem for multisite post-translational modification systems. *J. Theor. Biol.* 261, 626–636. doi:10.1016/j.jtbi.2009.09.003

van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*. Amsterdam, The Netherlands: Elsevier.

Volkov, I. L., Lindén, M., Rivera, J. A., Ieong, K.-W., Metelev, M., Elf, J., et al. (2018). tRNA tracking for direct measurements of protein synthesis kinetics in live cells. *Nat. Chem. Biol.* 14, 618–626. doi:10.1038/s41589-018-0063-y

White, R., Chiba, S., Pang, T., Dewey, J. S., Savva, C. G., Holzenburg, A., et al. (2010). Holin triggering in real time. *Proc. Natl. Acad. Sci. U. S. A.* 108, 798–803. doi:10.1073/pnas.1011921108

Wong, F., and Gunawardena, J. (2020). Gene regulation in and out of equilibrium. *Annu. Rev. Biophys.* 49, 199–226. doi:10.1146/annurev-biophys-121219-081542

Wong, F., Dutta, A., Chowdhury, D., and Gunawardena, J. (2018). Structural conditions on complex networks for the Michaelis-Menten input-output response. *Proc. Natl. Acad. Sci. U. S. A.* 115, 9738–9743. doi:10.1073/pnas.1808053115

Yordanov, P., and Stelling, J. (2018). Steady-state differential dose response in biological systems. *Biophys. J.* 114, 723–736. doi:10.1016/j.bpj.2017.11.3780

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).

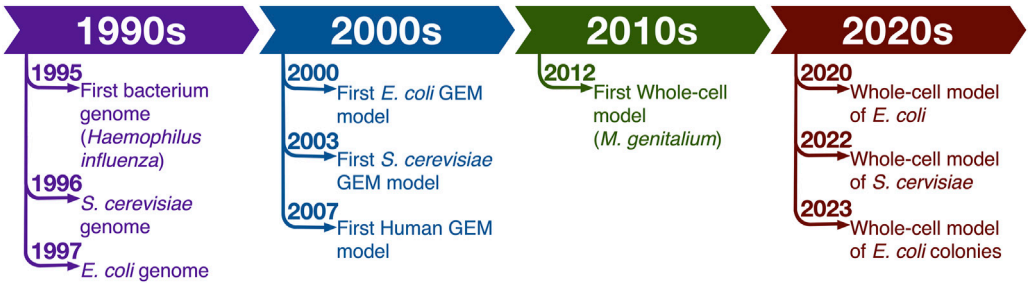


FIGURE 1
Timeline of some of the important milestones in development of whole-cell models.

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

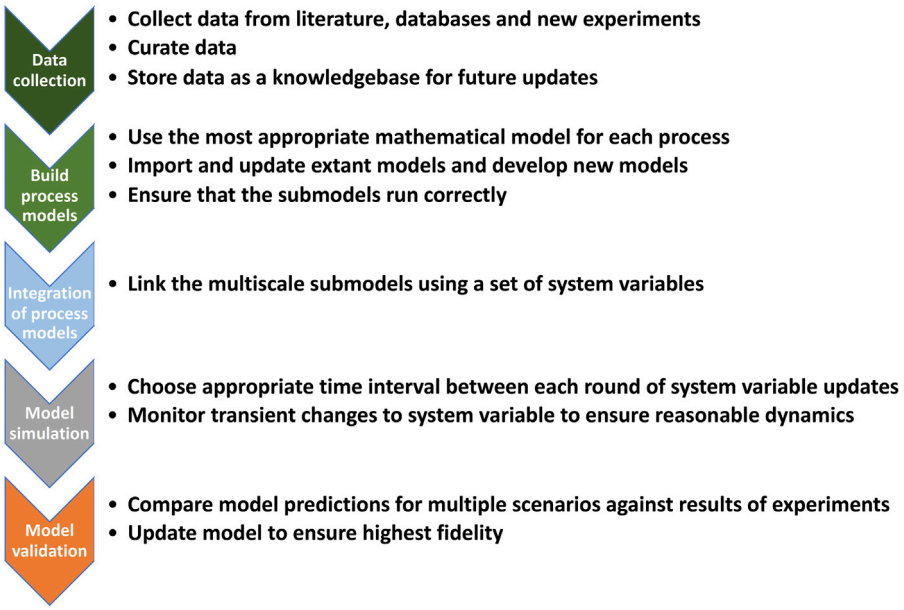


FIGURE 2
Assembly process for whole-cell models.

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

frontiers | Frontiers in Cell and Developmental Biology

TYPE Methods
PUBLISHED 24 November 2023
DOI 10.3389/fcell.2023.1235116

OPEN ACCESS

EDITED BY
Susan Mertins,
Leidos Biomedical Research, Inc.,
United States

REVIEWED BY
Samik Ghosh,
Systems Biology Institute, Japan
Sravan Kumar Miryala,
Northeastern University, United States

*CORRESPONDENCE
Paola Lecca,
Paola.Lecca@unibz.it

RECEIVED 05 June 2023
ACCEPTED 13 November 2023
PUBLISHED 24 November 2023

CITATION
Lecca P, Lombardi G, Latorre RV and Sorio C (2023), How the latent geometry of a biological network provides information on its dynamics: the case of the gene network of chronic myeloid leukaemia.
Front. Cell Dev. Biol. 11:1235116.
doi: 10.3389/fcell.2023.1235116

How the latent geometry of a biological network provides information on its dynamics: the case of the gene network of chronic myeloid leukaemia

Paola Lecca[1]*, Giulia Lombardi[2], Roberta Valeria Latorre[3] and Claudio Sorio[3]

[1]Faculty of Engineering, Free University of Bozen-Bolzano, Bolzano, Italy, [2]Department of Mathematics, University of Trento, Trento, Italy, [3]General Pathology Division, Department of Medicine, University of Verona, Verona, Italy

Background: The concept of the latent geometry of a network that can be represented as a graph has emerged from the classrooms of mathematicians and theoretical physicists to become an indispensable tool for determining the structural and dynamic properties of the network in many application areas, including contact networks, social networks, and especially biological networks. It is precisely latent geometry that we discuss in this article to show how the geometry of the metric space of the graph representing the network can influence its dynamics.

Methods: We considered the transcriptome network of the Chronic Myeloid Laeukemia K562 cells. We modelled the gene network as a system of springs using a generalization of the Hooke's law to n-dimension ($n \geq 1$). We embedded the network, described by the matrix of spring's stiffnesses, in Euclidean, hyperbolic, and spherical metric spaces to determine which one of these metric spaces best approximates the network's latent geometry. We found that the gene network has hyperbolic latent geometry, and, based on this result, we proceeded to cluster the nodes according to their radial coordinate, that in this geometry represents the node popularity.

Results: Clustering according to radial coordinate in a hyperbolic metric space when the input to network embedding procedure is the matrix of the stiffnesses of the spring representing the edges, allowed to identify the most popular genes that are also centres of effective spreading and passage of information through the entire network and can therefore be considered the drivers of its dynamics.

Conclusion: The correct identification of the latent geometry of the network leads to experimentally confirmed clusters of genes drivers of the dynamics, and, because of this, it is a trustable mean to unveil important information on the dynamics of the network. Not considering the latent metric space of the network, or the assumption of a Euclidean space when this metric structure is not proven to be relevant to the network, especially for complex networks with hierarchical or modularised structure can lead to unreliable network analysis results.

KEYWORDS

network geometry, graph embedding, dynamical systems, spring systems, chronic myeloid leukaemia, systems biology

1 Introduction

With the emergence of systems biology around the year 2000, the representation of a system of interacting biological entities, such as proteins, molecules, functional complexes, *etc.*, in the form of a network or graph has become preponderant and an unreliable prerequisite of any mathematical model regarding both the static and dynamic properties of the network. This representation of the components of a system as network nodes and their interactions as arcs between the nodes proved to be easy to understand as it is intuitive and also an excellent tool for organising data. However, the immediacy of understanding such a representation comes at the price of its low informational power, its susceptibility to misinterpretation and its use that often takes place under tacit or even unconscious assumptions. Particularly in the graph representation of a network, it is natural to think of the concept of distance between nodes as the number of arcs separating the nodes, or, if the weights of the arcs are known, as the weighted sum of the number of arcs separating the nodes. In doing so, it is implicitly assumed that the distance between two nodes is a Euclidean distance, or, in other words, that the metric space in which the network resides is flat Euclid space. This implicit assumption on a measure as important as the distance between nodes, used in multiple contexts as a measure of the intensity of an interaction between nodes, if not of the propensity of the interaction itself, may not only be reductive or approximate, but may even be incorrect. An erroneous assumption about the metric space that represents the geometry of the network carries serious risks, one of which is that of not being able to grasp the organizational principles of the typology and consequently the dynamics of the network. Indeed, the distribution of widely used centrality metrics like as degree and clustering coefficient reflect the features of the metric space, which defines the network's geometry. For example, heterogeneous degree distributions and significant clustering emerge naturally as reflections of the underlying hyperbolic geometry's negative curvature and hyperbolic metric characteristic (Krioukov et al., 2010). On the opposite, if a network has some metric structure and a heterogeneous degree distribution, the network has an effective hyperbolic geometry below (Krioukov et al., 2010).

It is often said to indicate the metric space of a network that the graph representing the network is "embedded" in a metric space, which is called the latent geometry of the network. The adjective "latent" is justified by the fact that the graph representation of a network does not make visible the characteristics of the metric space in which the coordinates of the nodes are actually defined. The verb "to embed", on the other hand, although commonly used, we condemn somewhat misleadingly, since the network, if endowed with a metric structure, is in fact not embedded in a metric space as if it were a distinct entity that fits into it, but is itself a portion of it, more precisely a discrete version of the continuous metric space that represents it. The use of the verb "to embed" stems from the procedures dedicated to understanding what the latent geometry of the network might be and based on tests in which the network is considered to have metrics of a different nature and then the distortion that the new metric has with respect to the original metric defined by the network's similarity matrix (i.e., weighted adjacency matrix) is assessed.

The latent geometry of a network is an important area of study in network science. We refer the reader to Boguñá et al. (2021); Jhun (2022) or an overview of the studies and fields of application of the study of the latent geometry of a network. In Jhun (2022), it is reported that latent geometry has been used to travel networks efficiently (Kleinberg, 2000; Boguñá et al., 2010) detect missing links (Liben-Nowell and Kleinberg, 2007; Clauset et al., 2008), map the brain (Allard and Serrano, 2020), and analyse proximity network (Papadopoulos and Flores, 2019). Interestingly, it has been shown that the map of contagions of various pandemics develops through paths defined on the latent geometry of the network of contacts and movements of individuals (Taylor et al., 2015). Systems biology has also benefited from the results of latent network geometry analysis, in particular the study of genetic networks and protein-protein interaction networks as reported in (Alanis-Lobato et al., 2018; Härtner et al., 2018; Pio et al., 2019; Klimovskaia et al., 2020; Sun et al., 2021; Lecca and Re, 2022; Lecca, 2023; Seyboldt et al., 2022).

While we can say that the relationship between latent geometry and static topological properties of the network, such as those measured by the centrality indices, is well established, the relationship between latent geometry and network dynamical properties is little investigated. A recent attempt in this direction was made by Rand et al. (2021). In this paper, Rand et al. study embryonic development. From egg to adult, embryonic development results in the reproducible and organised manifestation of complexity. In this process, the activity of gene networks culminates in the sequential differentiation of distinct cell types that construct this complexity, which has been likened by Conrad Waddington metaphor (Fard et al., 2016; Squier, 2017; Sánchez-Romero and Casadesús, 2021) to a flow through a landscape with valleys representing alternative destinies. Geometric approaches enable the formal description of such landscapes and codify the types of behaviours produced by differential equation systems.

With this study of ours, we wish to make a contribution in this still very unexplored field of the relations between latent geometry and the evolution of a network, particularly a biological network. We propose a method to infer the equations governing the dynamics of a network of genes previously identified by the authors (Lombardi et al., 2022) as involved in the development and progression of Chronic Myeloid Leukaemia (CML). The method consists of two steps: i) the determination of the latent geometry of the network through embedding of the network in three models of metric space (Euclidean, hyperbolic, and spherical), and ii) the determination of the dynamic equations describing this metric space. If the result of the step i) is the hyperbolic metric, the parameter of the dynamics of the interactions in the network conceived as a subspace of a hyperbolic space will depend on the hyperbolic distance between the interacting partners. Similarly, if the result of step i) is a spherical metric, the dynamics of the network will be parametrized by distance of the interacting nodes in the spherical space, and finally, if the result of step i) is an Euclidean metric, the network will be a dynamical systems whose parameters will depend on Euclidean distance between the interacting nodes.

In this study, we conceive of a network as a system of springs, in which the nodes constitute the masses and the arcs the springs that connect these masses/nodes. The spring constant represents the transmission efficiency of the interaction between the nodes. The interaction between a node A and a node B is seen as a change in the state of A causing a change in B. In accordance with the spring model, the interaction between nodes is seen as a propagation of the alteration of A's state through the spring to B, which absorbs the alteration of A in turn changing its state. The vibrational states of the networks nodes are governed by a generalization of the Hook law. According to this law, the spring constant is calculated by dividing the force required to stretch or compress a spring by the lengthening or

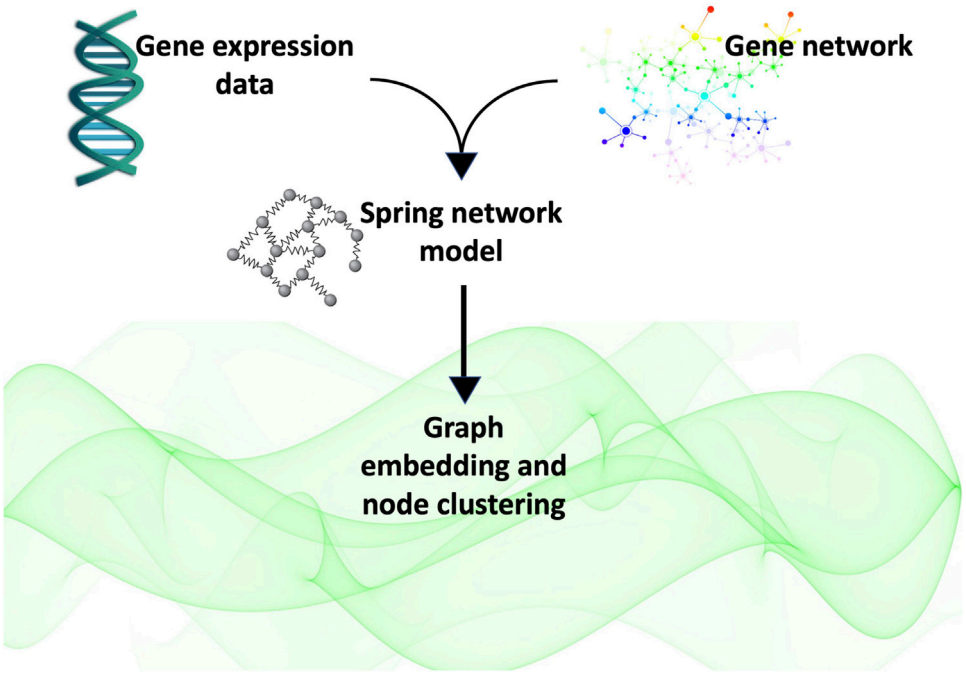


FIGURE 1
In this study, we obtained the gene network of interest by querying the Pathway Commons database with the list of genes of interest. We represented the network as a system of springs whose masses are the expression level of the genes as measured in our experiments in (Lombardi et al., 2022). We calculated the weighted adjacency matrix of the network as that matrix whose entries are given by the spring constant calculated at equilibrium. Finally, we used this matrix to embed the graph into three spaces (flat, positively curved and negatively curved)space) in order to determine which of them best represented the network's latent geometry. Finding the hyperbolic space fits best the latent geometry of the network, we proceeded to cluster the nodes according to their radial coordinate, that representative of the node popularity (Papadopoulos et al., 2012).

shortening of the spring. It is stated mathematically as $k = -F/\Delta x$, where Δx is the displacement of the mass, F is the force applied over x, and k is the spring constant (also known as *spring stiffness*). The propagation velocity of the elastic wave in a spring stressed by a force is directly proportional to the square root of the spring's elastic constant. A stiffer spring has a greater spring stiffness, and *vice versa*. As a consequence, a high spring stiffness is interpreted as high efficiency and thus greater ease in the transmission of interaction between nodes. The elastic constant metaphor, in network metric space, corresponds to a measure of similarity between nodes, such that nodes connected by harder springs are closer nodes in terms of similarity. In the model of network we present here, the elastic constants of the springs are obtained from a generalization of the Hook's law for a system with N masses and E springs (N corresponding to the number of nodes and E corresponding to the number of edges), where the mass of the node is given by its total degree and the change in the position of the node is given by the index of vibrational centrality proposed by Estrada and Hatano (2010). The matrix of elastic constants is used in network embedding procedures in three types of space, Euclidean, hyperbolic, and spherical. The metric space for which the embedding of the network shows a minimum distortion of the values of this matrix is considered as the best approximation for the metric space of the network. The distances of the nodes in this metric space constitute the parameters of the network dynamics, which we describe here in terms of mass action law.

The article is organised as follows: in Section 2 we introduce the three types of isotropic spaces considered in this study and the embedding techniques we used to identify which of the metric spaces considered best represents the network's latent geometry. In Section 3 we describe the data and the gene network of the case study. In Section 4, we describe the mathematical modelling of the gene network as a system of springs, and finally in Section 5 we report the results obtained. This is followed by some concluding remarks and a recapitulation of the study performed (Section 6). In Figure 1 we illustrate the main steps of the analysis presented in this study.

2 Network geometry and methods of embedding

A graph embedding consists in the determination of the coordinates of the graph node in a given metric space in such a way that the graph similarity matrix is reproduced with as little error as possible. The embedding of a graph thus consists of the problem of finding the coordinates of the nodes in a given metric space from the similarity matrix of the graph, which is a measure of the distances between nodes. In the final analysis, embedding consists of finding coordinates of points given their distance. Isotropic spaces can only be classified as Euclidean (flat), elliptic (having positive curved), or hyperbolic (having negative curvature). In the following sections, we will recall some basic definitions, such as that of inner product and distance for these three types of spaces, and briefly mention the mathematical techniques of graph embedding, of which there are many variants in literature. We

also recall how the latent geometry is related to the structure and organizational principles of the network (e.g., presence of communities, hierarchical organization, etc.).

2.1 Euclidean space

The Euclidean geometry is based on the following five postulates: (i) Any two points can be joined by a straight line segment. (ii) Any portion of a straight line can be stretched forever. (iii) Any straight line segment can be used as the radius of a circle, with one endpoint serving as the centre. (iv) All right angles are congruent. (v) When two lines are drawn so that they intersect a third in a fashion that results in a side where the total of the inner angles is less than two right angles, the two lines will always cross each other if they are extended far enough.

More formally, an Euclidean space, is a real vector space (i.e., a vector space whose field of scalars is $\mathbb{R}$) E equipped with a positive definite symmetric bilinear form φ: $E \times E \to \mathbb{R}$. The real number $\varphi(x, y)$ is called the *inner product* between the vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, that is defined as

$$\begin{aligned}\varphi(\mathbf{x}, \mathbf{y}) &= (x_1, x_2, \ldots, x_n) \cdot (y_1, y_2, \ldots, y_n) \\ &= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n. \end{aligned} \tag{1}$$

Usually the inner product of two vectors $\mathbf{x}$, $\mathbf{y}$ is dented with the angular bracket $\langle \mathbf{x}, \mathbf{y} \rangle$. In the Euclidean space $\mathbb{R}^n$ the distance between the points whose coordinates are given by the vectors $\mathbf{x}$ and $\mathbf{y}$ is

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sqrt{\langle \mathbf{x}, \mathbf{y} \rangle} = \sqrt{\varphi(\mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x})} \\ &= \sqrt{(y_1 - x_1)(y_2 - x_2) \cdots (y_n - x_n)} \equiv = \|\mathbf{y} - \mathbf{x}\|. \end{aligned} \tag{2}$$

To embed a graph into an Euclidean space, we used the classical (metric) multidimensional scaling algorithm that, given as an input the pairwise dissimilarities matrix $\{d_{ij}\}$, reconstructs a map that preserves distances. The algorithm implements the following steps.

1. Find a random arrangement of points, for example, by taking a sample from a normal distribution.
2. Determine the distances between the points.
3. Find the best monotonic transformation for the proximity to get the best scaled data.
4. Find a new arrangement of points to reduce the stress between the optimally scaled data and the distances. The stress of the embedding in Euclidean space is defined by the following residual sum of squares

$$\text{Stress}(x_1, x_2, \ldots, x_n) = \sqrt{\sum_{i \neq j = 1, \ldots, n} \left(d_{ij}^{(\text{input})} - d_{ij}^{(\text{embedding})} \right)^2} \tag{3}$$

where, in the case of Euclidean embedding, $d_{ij}^{(\text{embedding})} = \|x_i - x_j\|$.

5. Compare the stress to a certain standard. If the stress is too low, stop the algorithm; otherwise, go back to step 2.

We implemented the embedding in R (R Core Team, 2021), using the function `cmdscale` (Gower, 2015) of the library `stats`. Theoretical foundations and details about multidimensional scaling techniques can be found in many text books and review paper [see, for example, (Borg and Groenen, 2005; Cox and Cox, 2008; Zhang and Takane, 2010)].

2.2 Hyperbolic geometry and the Poincaré model

Hyperbolic geometry accepts the first four axioms of Euclidean geometry but rejects the fifth, namely, that there exists a line and a point not on the line with at least two parallels to the given line crossing through the provided point. This is equivalent to performing geometry on a surface with a constant negative curvature. This geometry differs greatly from the more conventional Euclidean geometry, and are hard to visualise. The main reason is that by the Hilbert's theorem (Hilbert, 1933) the hyperbolic plane cannot be isometrically embedded into Euclidean 3D-space (isometric means preserving the length of every curve). We must flatten the curvature to display the hyperbolic plane. In doing this, many of the straight lines in hyperbolic space end up being curved as a result. The French mathematician Henri Poincaré is responsible for one of the widely accepted theories for flattening the hyperbolic plane and the n-dimensional ball model (Poincaré disk in 2D) (Anderson, 2005).

The Poincaré n-dimensional ball $\mathbb{B}_{\mathbb{R}}^n$ ($\mathbb{B}_{\mathbb{R}}^n = \{\mathbf{x} \mid \|\mathbf{x}\|^2 < 1\}$) is a model for n-dimensional hyperbolic geometry in which lines are represented by circle diameters or by arcs of a circle with ends perpendicular to the boundary of the ball. (Figure 2). If $n = 2$ the Poincaré model is a unit open disc. We briefly summarize here the method in Conn (2010) to calculate the distances in the unit disc model Consider the fractional linear transformation S that sends $\infty \mapsto i$ and $\pm 1 \mapsto \pm 1$. S sends the real axis to the boundary of the unit disc and, since fractional linear transformations preserve the orientation of circles, it sends the upper half-plane to the disc's interior. The H_2-distance between two points a, b in the unit disc is the H_1-distance between their preimages $S^{-1}(a)$, $S^{-1}(b)$ in the upper half-plane (Conn, 2010), and in this way the unit disc inherits a metric from the metric of the upper half-plane.

Let D^1 denote the interior of the unit disc and suppose

$$\gamma\colon [0, 1] \to D^1$$

is a piecewise continuously differentiable curve. If $H_k(\gamma)$ ($k = 1, 2$) denotes the length of the curve γ, then

$$H_2(\gamma) = H_1(S^{-1} \circ \gamma). \tag{4}$$

Writing $S^{-1} \equiv T$, we have

$$\begin{aligned} H_1(T \circ \gamma) &= \int_{T \circ \gamma} \frac{1}{\text{Im}(z)} |dz| = \int_0^1 \frac{1}{\text{Im}((T \circ \gamma)(t))} \left| (T \circ \gamma)'(t) \right| dt \\ &= \int_0^1 \frac{1}{\text{Im}(T(\gamma(t)))} \left| T'(\gamma(t)) \right| \left| \gamma'(t) \right| dt \\ &= \int_\gamma \frac{1}{\text{Im}(T(z))} \left| T'(z) \right| |dz|. \end{aligned} \tag{5}$$

Since T has the form

$$T(z) = \frac{iz - 1}{-z + i}. \tag{6}$$

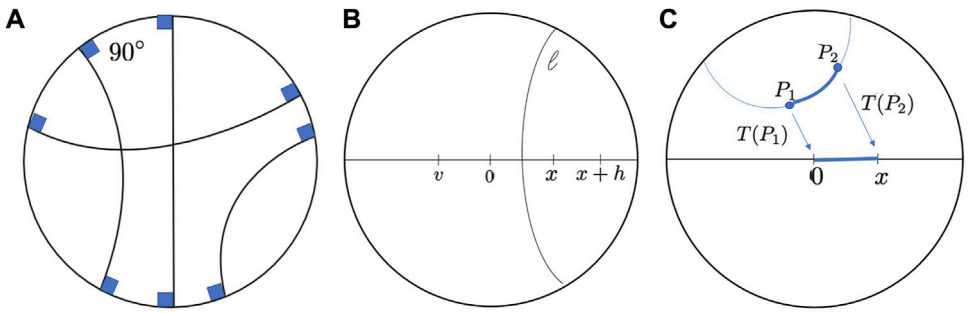


FIGURE 2
(A). Geodesics in Poincaré disk. **(B)**. Reflections in Poincaré disk geometry. 0, x, $x + h$ and v are points on the diameter. 0 is the reflection of x, and v is the reflection of $x + h$ with respect to the hyperbolic segment l. **(C)**. The distance between two generic points P_1 and P_2 can be found first transforming P_1 tp 0 and P_2 to x.

we have that

$$\mathrm{Im}\,(T(z)) = \mathrm{Im}\left(\frac{(iz-1)(-\bar{z}-i)}{|-z+i|^2}\right) = \frac{1-|z|^2}{|-z+i|^2}, \tag{7}$$

and

$$|T'(z)| = \frac{2}{|-z+i|^2}, \tag{8}$$

we obtain

$$H_2(\gamma) = \int_\gamma \frac{2}{1-|z|^2}|dz| \tag{9}$$

that is general formula calculating distances in the Poincaré disc. $2/(1-|z|^2)dz$ is the element of arc length. Consequently, the distance between two points $a, b \in \mathbb{C}$ on Poincaré disc is

$$d(a,b) = H_2(a,b) = \log\frac{|1-\bar{a}b| + |b-a|}{|1-\bar{a}b| - |b-a|}. \tag{10}$$

Any diameter of the unit disc is a geodesic, so if z is a point in the unit disc, then the Euclidean segment from 0 to z is also a hyperbolic segment from 0 to z. We have hence that

$$H_2(0,z) = \int_0^{|z|} \frac{2}{1-t^2}dt = 2\tanh^{-1}(|z|) = \log\left(\frac{1+|z|}{1-|z|}\right). \tag{11}$$

Complex networks connect different nodes. This diversity indicates that there is at least some element taxonomy, meaning that all nodes can be classified in some way. This classification means that nodes can be separated into large groups that are made up of smaller subgroups that are made up of even smaller sub-subgroups, and so on. The relationships between such groups and subgroups can be approximated by treelike structures, which illustrate hidden hierarchies in networks. Krioukov et al. demonstrated that the metric structures of trees and hyperbolic spaces are equivalent (Krioukov et al., 2010; Kurkofka et al., 2021; Lecca, 2023; Lecca and Re, 2022). It is not necessary for the node classification hierarchy to be exactly a tree, but rather approximately a tree. When a network can be approximated by a tree, its latent geometry is negatively curved (Gromov, 2007).

To perform the embedding into a hyperbolic space (Poincaré model), we used the function `hydraPlus` of the R library `hydra` (HYperbolic Distance Recovery and Approximation) (Keller-Ressel, 2019), that uses a strain-minimizing hyperbolic embedding based on reduced matrix eigendecomposition (Keller-Ressel and Nargang, 2020). The stress of embedding in hyperbolic space is then given by formula (3), where $d^{(\text{embedding})}$ is given by the output of `hydra`.

2.3 Spherical geometry and embedding

Spherical geometry is the geometry of a hypersphere's surface. The hypersphere can be easily immersed in euclidean space; for example, the embedding of a three-dimensional sphere of radius r is well known relation $x^2 + y^2 + z^2 = r^2$, with $\mathbf{x} = (r\sin u \sin v, r\cos u \sin v, r\cos v)^T$. A simple extension of this is the embedding of a $(n-1)$-dimensional sphere in n-dimensional space:

$$\sum_{i=1}^{n} x_i^2 = r^2. \tag{12}$$

There is a constant sectional curvature of $1/r^2$ throughout this curved surface. The length of the shortest curve that lies in the space and connects the two points is the geodesic distance between two points in a curved space. The geodesic on the hypersphere is a perfect circle for a spherical space. The distance is equal to the width of the arc that connects the two locations on the great circle.

If two points in the hypersphere's centre form an angle with θ_{ij}, then the distance between them is

$$d_{ij} = r\theta_{ij}. \tag{13}$$

A point can be represented by a position vector $\mathbf{x}_i$ of length r with the coordinate origin at the origin of the hypersphere. We can also write

$$d_{ij} = r\,\cosh\frac{\langle \mathbf{x}_i, \mathbf{x}_j\rangle}{r^2} \tag{14}$$

since the inner product is $\langle \mathbf{x}_i, \mathbf{x}_j\rangle = r^2\cos\theta_{ij}$.

To perform the embedding of the graph in a hypersphere, we used the method proposed by Wilson et al. (2014a), and the Matlab code that this authors made available in Wilson et al. (2014b). We summary briefly the core of embedding method in this way.

Given a dissimilarity matrix D, we want to determine the set of points on a hypersphere that give the same distance matrix. Because the curvature of the space is unknown, we must also determine the

radius of the hypersphere. We have n items of interest, thus we would ordinarily look for a $n-1$ dimensional Euclidean space. A coordinate system with the origin at the centre of the hypersphere is considered. A matrix $\mathbf{X}$ of point positions vectors is constructed in such a way that

$$\mathbf{X}\mathbf{X}^T = \mathbf{Z} = \left\{z_{ij}\right\} = r^2 \cos\frac{d_{ij}}{r} \tag{15}$$

$\mathbf{Z}$ is a $n \times n$ matrix that is positive semi-definite and has rank $n-1$ since the embedding space has dimension $n-1$, $\mathbf{X}$ is made up of n points that are located in a space of dimension $n-1$, and so does the embedding space. This means that the $\mathbf{Z}$'s eigenvalues are positive, with only one being zero. This observation can be used to calculate the radius of curvature. Then, in order to find r, Wilson et al. (2014a) proposed to create $\mathbf{Z}(r)$ and identify the smallest eigenvalue λ_1, to calculate then the optima radius of curvature as

$$r^* = \arg\min_r |\lambda_1[\mathbf{Z}(r)]|. \tag{16}$$

The stress of embedding in hyperbolic space is then given by formula (3), where $d^{(\text{embedding})}$ is given by the elements of the matrix $\mathbf{Z}$.

2.4 How latent geometry influences network dynamics

By the term "dynamic" of a system, we mean the *time and space* evolution of the system as described by differential and/or algebraic equations whose variables are quantitative features of the system's actors, and whose mathematical form model the topological system's organization. The equations of the dynamics are parameterized by the dynamical properties of the system itself (such as frequency of oscillation, if the system is oscillatory, elastic constant, if the system is assimilated to a spring system, specific rate of reaction, *etc.*) There are interesting studies showing how the geometry of complex networks affects the dynamics. To cite a relevant contribution to the field, we mention the work of Millán et al. (2018) which shows that the latent geometry of a network has a significant impact on the synchronization dynamics. Unlike Millán et al. work, which is more focused on the dynamic properties of the system (i.e., parameters and synchronization laws), here we focus on the influence that latent geometry can have on network organization. And since from the network organization, the dynamics of the network is derived, we can expect latent geometry to influence the dynamics. In particular, the geometry of the network determines the presence or absence of functional modules containing highly cooperative nodes. The identification of these possible functional clusters can be done correctly only if the metric space of the network is identified. In fact, this space defines the distance between nodes, the measure on which clustering algorithms are based. A clustering in Euclidean space may lead to a different result from clustering in hyperbolic space, the distance computed in this space being different from the distance computed in Euclidean space. The correct dynamics is one whose parameters and functional modules are established by the latent geometry for at the network under consideration.

In this study, we conceived a network as a spring system. Through the identification of the most appropriate latent geometry of the network under consideration, i.e., that geometry that most closely reproduces the values of the spring constants of the edges thought of as springs, we were able to identify cluster of gene drivers for the network dynamics. The role of drivers of these genes was validated through functional analysis of them. In the next sections, the data from which we built the network, as well as the model and analysis of the network itself are reported.

3 Data and gene network

We use here the data of gene expression relevant to the landscape of Chronic Myeloid Leukemia K562 cells. We refer the reader to a recent publication by the authors (Lombardi et al., 2022)], where we describe the experimental activity implemented for data measurement and algorithmic procedures for selecting differentially expressed genes. For the reader's convenience we summarise it briefly below.

On an Agilent whole human genome oligo microarray (#G4851A, Agilent Technologies, Palo Alto, CA), the RNAs from the samples were hybridised. This microarray consists of 60,000 distinct human transcripts represented by 60-mer DNA probes created using SurePrint technology. The manufacturer's recommended protocol was followed when one-color gene expression was carried out. In a nutshell, samples were used to extract the total RNA fraction using the Trizol Reagent (Invitrogen). Agilent Technologies' Agilent 2100 Bioanalyzer was used to evaluate the quality of the RNA samples. RNAs with low integrity (RNA integrity number less than 7) were not included in the microarray analysis. Using the Low Imput Quick-Amp Labelling Kit, one colour (Agilent Technologies) in the presence of cyanine 3-CTP, labelled cRNA was produced from 100 ng of total RNA. In a revolving oven, hybridizations were carried out for 17 h at 65°C. Agilent's scanner produced images with a 3 μm resolution, and Agilent Technologies' Feature Extraction 10.7.3.1 software was utilised to extract the microarray raw data. The GeneSpring GX 11 programme (Agilent Technologies) was then used to analyse the microarray results. Data transformation was used to normalise all of the data's negative raw values to 1.0 using the 75th percentile. Only the probes expressed in at least one sample (marked as Marginal or Present) were retained using a filter on low gene expression.

The data used in this work come from the aforementioned examination of the CML cell transcriptome (K562) using microarray hybridization under various settings. The cells were transfected with full-length PTPRG and compared to three controls: cells transfected with the empty vector, cells transfected with a PTPRG inactive mutant with a mutation in the catalytic domain (D1028A), and cells treated with Imatinib, which targets the oncogene BCR/ABL1. The complete dataset is publicly available at the GitLab repository. https://gitlab.inf.unibz.it/Paola.Lecca/chronic-myeloid-leukemia-genes.

Here, from the entire dataset available at this link, we only considered the gene expression levels of the untreated group (empty vector and inactive mutant domain D1028A) and those of the treatment group expressing full-length PTPRG. We then selected the genes, that, according to the analysis in Lombardi et al. (2022),

result to be differentially expressed between the two groups. To construct the gene network, we queried PathwaysCommons (PathwayCommons.All.hgnc repository) (Cerami et al., 2010a; Cerami et al., 2010b) by providing as input for the search the list of gene names we considered in this study. The obtained gene network is a representation of molecular associations specified through nodes (genes) and edges (molecular interactions or statistical relationships). Among the various format, PathwaysCommons gives as output result of the query the gene networks also in SIF (Simple Interaction format), which is a table providing details on gene-gene interactions. This format offers various levels of detail such as: interaction type, reference data source, Pubmed id, reference pathways, and mediators id. Our analyses focused on the most granular level of information, namely, the interactions between pairs of genes, listed in the SIF table as"Participant A" and "Participant B" (we refer the reader to the public repository of our data to view the data format). The types of interactions included in the network are as follows: interacts-with, in-complex-with, catalysis-precedes, controls-state-change-of, controls-transport-of, controls-transport-of-chemical, controls-expression-of, controls-phosphorylation-of, controls-production-of, chemical-affects, consumption-controlled-by.

As a final result of querying to common pathways and selecting differentially expressed genes on the two groups (treated and untreated), we obtained a network that is a non-planar multi-edge graph with 2,080 nodes and 3,745 edges, that we simplify to a non-multi-edge graph with 2,080 nodes and 3,464 edges.

3.1 Graph embedding in presence of noise of input data: Some remarks

The presence of noise on the data in the adjacency matrix used as input to the graph embedding procedures could be a vexing problem if embedding stresses in different metric spaces are to be compared to identify which metric space is best representative of the latent geometry of the network. Noise, for example, may not allow weak edges to be distinguished from the absence of nodes and may affect the reliability of the measurement of even the most robust arcs (i.e., those with the greatest weight). Data analysis frequently faces the challenge of distinguishing between real weak edges and noise-induced low-weight edges. To solve this issue, noise is typically either eliminated or studied in the absence of data.

In the specific case of our study, the experimental data from which we start to construct the weighted adjacency matrix of the graph are very accurate. Our dataset was validated comparing the outcome of the cDNA microarray with the analysis of a specific set of genes chosen for being informative and for being predicted up and downregulated. Validation was performed in triplicate with quantitative PCR on a new, independent, preparation of cDNA derived from the same cell lines, thus ensuring that the results present in our dataset represent a true variation in mRNA levels. Notably the analysis permitted to predict a shift to erythroid differentiation of the cells that was confirmed also at protein level. All supporting data are reported on the publication (Lombardi et al., 2022).

Interesting and noteworthy works elucidating the role and the influence of noise in graph embedding has been done recently by Maddalena et al. (2022) and Okuno and Shimodaira (2019). The treatment of the presence of noise is in fact so complex that it deserves the implementation of a focused study and consequently the writing of a separate article. It is out of the scope of this study, give the high quality of the data we used here, but it is our intention to explore this issue further in a forthcoming study.

To the best of our knowledge at present, we find of particular interest the study of Blevins et al. (2021). Instead, by analysing the structure of noisy, weak edges that have been artificially added to model networks, the authors explored how noise and data coexist in this work. They discovered that there are qualitative classifications of noise structure that arise, and that these noisy edges can be used to categorise the model networks. The authors state that the structure of low-weight, noisy edges varies depending on the topology of the model network to which they are added. Interestingly, Blevins et al. showed that noise is a complex, topology-dependent, and even valuable phenomenon in characterising higher-order network interactions rather than a monolithic annoyance.

4 Mathematical model of the gene network

To estimate the weights of the network arcs, we conceptualise the network as a system of masses (representing the nodes) and springs (representing the edges), as in Figure 3. Estrada and Hatano (2010) has brought a remarkable contribution to spring-like network models. In a complex network, Estrada and Hatano suggested a new metric for measuring node vulnerability. The metric is based on an analogy where the network's nodes are represented by masses and its edges by springs. They defined the measure as the node displacement, or the amplitude of vibration of each node, under variation caused by the thermal bath in which the network is intended to be immersed, and that represents the environment from which stimuli may possibly come. The Estrada index for the vibrational centrality of the node i is defined as the node displacement $(\Delta x)_{ii}$

$$(\Delta x)_{ii} = \sqrt{\frac{T}{k}\mathbf{L}^{+}_{ii}}, \tag{17}$$

where T is the temperature of the external bath, and k is the spring stiffness. Estrada and Hatano assumed that the network edges are identified with springs with a common spring stiffness k.

Instead in the network spring model of Lecca and Re (2020), the authors postulated that weights of the arcs are given by the stiffness of the springs representing the arcs, so each arc may have a different stiffness/weight. The harder the spring, the more efficiently the signal is transmitted from node to node; the softer the spring, the less quickly the signal is transmitted from node to node. According to this metaphor, edges characterised by high values of the stiffness of the hypothetical spring joining them are nodes that interact more effectively than nodes whose spring stiffness joining them is lower. The stiffness of the spring is thus interpreted as the efficiency of the interaction. Next, we briefly summarize the computational method developed by Lecca and Re (2020), and used in this study, to calculate the stiffnesses of the springs.

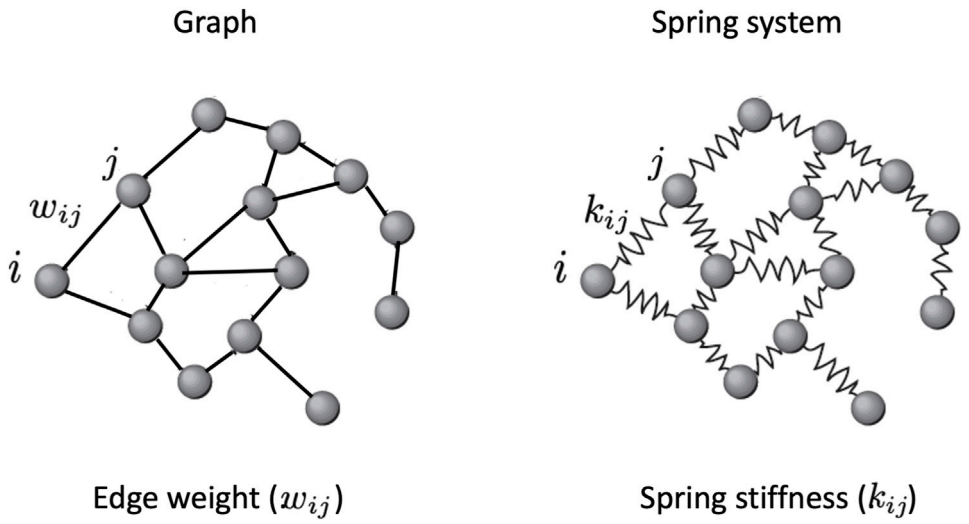


FIGURE 3
A spring system, also known as a spring network, is a model of physics used in engineering and physics that is represented as a graph with a mass at each vertex and a spring with a specific stiffness and length along each edge. Extending the Hooke's law to higher dimensions (see the Section ?), it is possible to calculate the spring stiffnesses that in this model represent the edge weights.

The elastic force applied to the nodes by the springs according to the generalised Hooke's law for a system of N springs is

$$F_{\text{elastic}} = -\boldsymbol{K}\Delta\boldsymbol{x}, \tag{18}$$

where $\boldsymbol{K}$ is the matrix providing the stiffnesses of all of the springs, and the elements of $\Delta\boldsymbol{x}$ are the vibrational centralities of the nodes. We obtain the force on the nodes by multiplying F_{elastic} by the transpose of the graph weighted incidence matrix $\boldsymbol{C}^{\top}$, where, in general, the weights are given by the node masses, i.e.,

$$\boldsymbol{C} = \boldsymbol{AM} \tag{19}$$

where $\boldsymbol{A}$ is the unweighted incidence matrix. We should remark that weighting the incidence matrix with node mass values means taking into consideration the nodes' inertia to the propagation of the elastic force through the springs incident to them (Lecca and Re, 2020). In this study, the nodes' masses are given by the nodes' total degree.

The force on node is then defined by

$$F_{\text{nodes}} = -\boldsymbol{C}^{\top}\boldsymbol{K}\Delta\boldsymbol{x} \tag{20}$$

At the equilibrium $F_{\text{nodes}} = \boldsymbol{0}$, i.e.,

$$\boldsymbol{C}^{\top}\boldsymbol{K}\Delta\boldsymbol{x} = \boldsymbol{0}, \tag{21}$$

where $\boldsymbol{K}$ is obtained as the nullspace (or kernel) of $\boldsymbol{C}^{\top}$, in formula:

$$\boldsymbol{K} = \text{Ker}\left(\boldsymbol{C}^{\top}\right). \tag{22}$$

Indeed, all vectors $\boldsymbol{K}$ that have the properties that $\boldsymbol{C}^{\top}\boldsymbol{K} = 0$ and $\boldsymbol{K}$ are not zero make up the null space of any matrix $\boldsymbol{C}^{\top}$.

Once $\mathbf{K}$ is obtained, we construct the dissimilarity matrix of the graph, which is then used as input for the embedding algorithms, as follows

$$d_{ij} = \frac{1}{1 + k_{ij}} \tag{23}$$

where k_ij are the elements of the matrix $\mathbf{K}$. Thus, nodes connected by a spring with a high value of the elastic constant have a lower dissimilarity value than nodes connected by a spring with a low value of the elastic constant. This reflects the situation where the propagation speed of the interaction along a spring with high stiffness is higher than along a spring with low stiffness.

Of particular interest is in case the system is not in equilibrium. In fact, $\boldsymbol{K}$ is independent on Δx only when the system is at equilibrium, i.e., when $F_{\text{nodes}} = \boldsymbol{0}$ and Eq. 21. In non-equilibrium conditions, we have instead that $F_{\text{nodes}} = \boldsymbol{C}^{\top}\boldsymbol{K}\Delta\boldsymbol{x} \neq \boldsymbol{0}$. Suppose that we know the forces F_{nodes} acting on nodes. For example, this could be the case in which perturbation experiments are implemented to measure and analyse the responsiveness of the network nodes to stimuli and/or stresses, or, assimilating forces on nodes to white noise distributed over all nodes, noise always present in biological systems at the micro-scale given their inherent stochastic dynamics). To calculate the matrix $\boldsymbol{K}$, in this case, the requirements are that the matrices $\boldsymbol{C}^{\top}$ and Δx are invertible, so that

$$\boldsymbol{K} = \left(\boldsymbol{C}^{\top}\right)^{-1} F_{\text{nodes}} (\Delta x)^{-1}. \tag{24}$$

Note, Δx is invertible if and only if all the entries on its main diagonal are non-zero, which means that little to much all nodes have a significantly non-zero response to stress.

5 Results

We embedded the gene network in the three metric spaces considered by considering different dimension values. We started with dimension 3, since the network is not planar. As shown in Figure 4, the embedding that produces the least amount of stress on the dissimilarity matrix - obtained as in Section 4 - is the hyperbolic embedding. The network is then

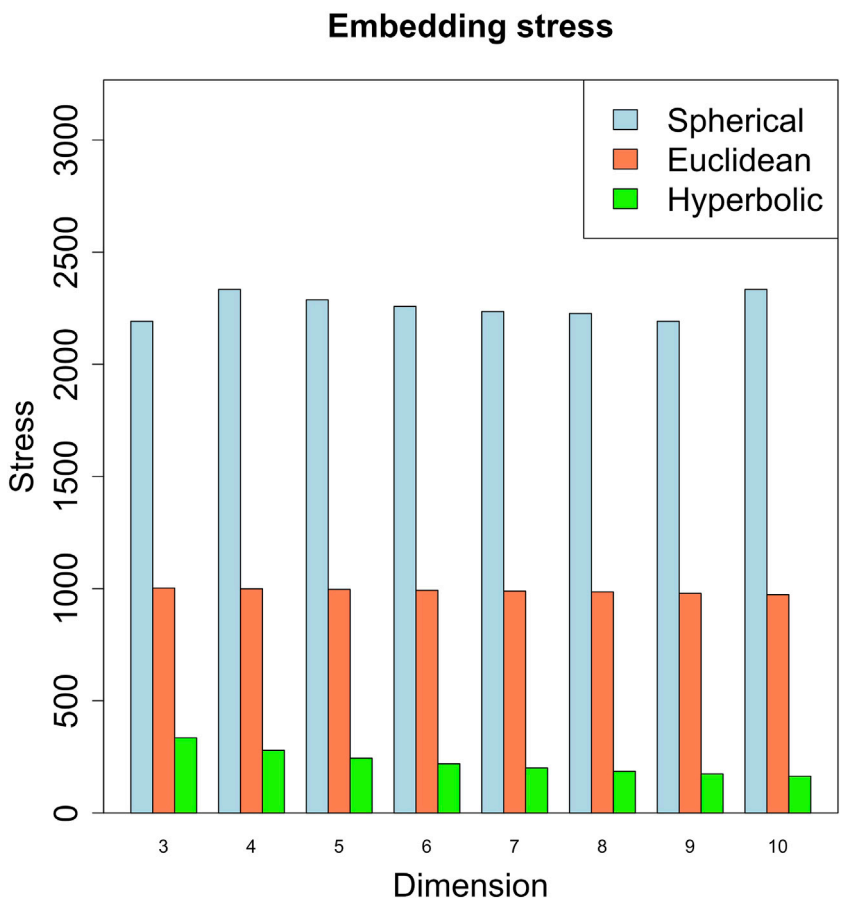


FIGURE 4
Embedding stress vs. metric space dimensions. The embedding with the least stress is the hyperbolic one, revealing a putative hyperbolic latent geometry of the gene network.

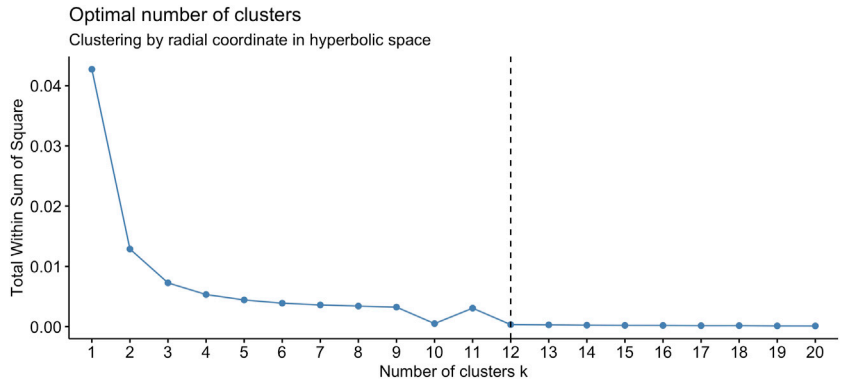


FIGURE 5
The optimal number of clusters of the set of radial coordinates of the points (node) on the Poincaré disk, according to the Elbow method, is 12.

characterized by a power-law degree, and by a hierarchical structure reflected also in the presence of clusters in the radial coordinates of the points, that is known to represent the node popularity (Papadopoulos et al., 2012; Yang and Rideout, 2020; Kovács and Palla, 2021). By nodes having high popularity, we mean nodes that are related to the majority of the other nodes in the graph [see also (Lecca et al., 2023) for a short review of the Papadopoulos et al. definition of node popularity]. These nodes can aid in the efficient spreading of information throughout the network.

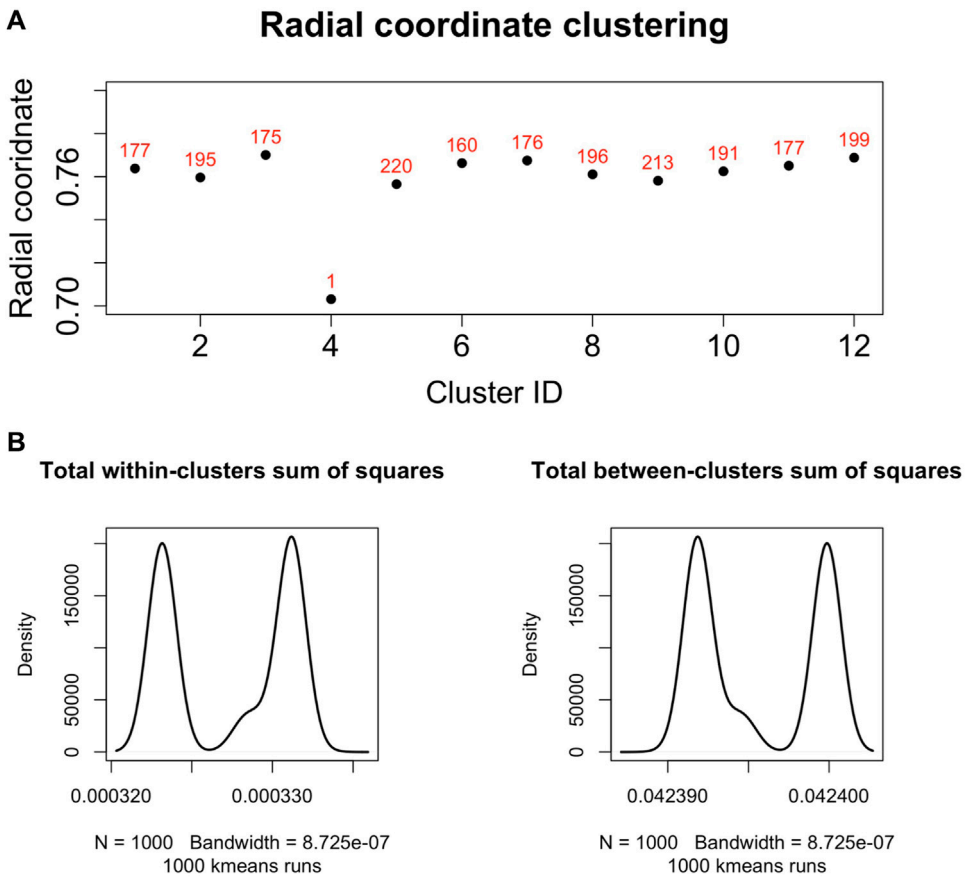


FIGURE 6
(A). Centroids of the radial coordinate clusters *versus* cluster identifier in a single run of k-means algorithm. The number of genes belonging to each cluster is shown in red. The gene belonging to the cluster number four is ZRANB1, a gene characterized by low immune cell specificity, and belong to NK-cells immune cell expression cluster (Uhlén et al., 2022b). **(B)**. We performed 1,000 runs of the k-means for the clustering of the radial co-ordinate of the nodes of the network in hyperbolic space and drew the distributions of the within- and between-clusters sum of squares. The within-cluster sum of squares quantifies the internal cohesion inside each cluster. The between-cluster sum of squares quantifies the external separation between clusters. The figure shows that for the k-means clustering of the radial coordinates the between-clusters sum of squares is two orders of magnitude greater than the within-cluster sum of squares, revealing the accuracy and then reliability of the clustering results.

We found that the set of radial coordinates, whose values range in [0.7036133, 0.7709305], is characterized by 12 clusters as determined by the Elbow method (Umargono et al., 2020) (see Figure 5). The range of radial coordinate values is. In Figure 6A we report the cluster ID and the size of the 12 clusters of the radial coordinates as obtained by a single run of the k-means algorithm. We found that the gene with the smallest popularity (i.e., with the smallest radial coordinate) is ZRANB1. This gene allows for K63-linked polyubiquitin modification-dependent protein binding and thiol-dependent deubiquitinase activity. Involved in a variety of functions, including the positive control of the Wnt signalling pathway, protein deubiquitination, and cell morphogenesis regulation (NIH, 2023). However, according to the date in The Human Atlas of Proteins is a low immune cell specificity gene (Pontén et al., 2008; Uhlen et al., 2017; Uhlén et al., 2022a; Uhlén et al., 2022b). To assess the stability and the quality of the clustering, we repeated the k-means 1,000 times and graphed the distributions of the within- and between-sum of squares (see Figure 6B) that show that the first is two order of magnitude smaller then the second. The skewness of the distributions and the disproportion between within- and between-cluster sum of squares indicate the stability and accuracy of the clustering, respectively.

Nodes with similar radial co-ordinate have similar popularity, so clustering according to the radial co-ordinate identifies communities of nodes with similar popularity. However, the radial co-ordinate, in addition to representing the popularity of a node, i.e., its degree of connectivity with other nodes in the network, identifies the distance from the origin in the Poincaré ball. In a network with hyperbolic latent geometry, in its representation in the Poincaré ball, the mean degree of a node is a negative exponential function of the node's radial coordinate (Krioukov et al., 2010). Thus, the average degree of a node decreases exponentially with increasing distance of the node from the origin of the Poicaré ball, or, in other terms, the higher the radial co-ordinate of a node, the lower its degree on average. The area inside the unit ball represents the infinite hyperbolic plane, and, consequently, nodes with radial co-ordinate equal 1 are points at infinity. Clustering according to the radial co-ordinate thus identifies bands of points (nodes) that are concentric on the Poincaré disc and that have a decreasing degree of connectivity as one moves away from the origin. This is why we say that

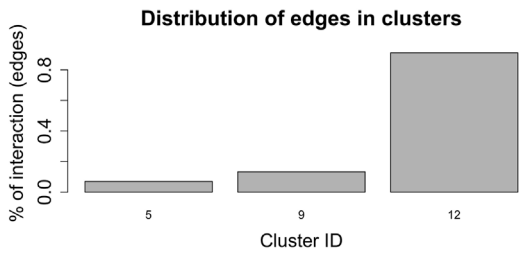


FIGURE 7
Barplot showing the percentage of edges joining nodes belonging to the same cluster according to the radial co-ordinate value. Only clusters 5, 9 and 12 contain nodes that are connected by an edge and belong to the cluster. The remaining edges connect nodes belonging to different clusters.

clustering according to radial co-ordinates allows clusters of strongly interconnected nodes to be identified (if any). The cluster of nodes closest to the origin identifies not only nodes with high connectivity, but also nodes that are close to each other (this second characteristic also applies to clusters far from the origin). The coexistence of two characteristics such as high degree and small distance between nodes is typical of a cluster of nodes with efficient interactivity and greater inertia to perturbations induced by external stimuli, such as variation of expression level, interactions with drugs, *etc.* The short inter-node distance reflects the high efficiency of communications, the high connectivity may be responsible of the node robustness.

Nodes that are thus highly interconnected and close in network metric space are potential drivers of network dynamics. This conjecture is demonstrated in the case where the distribution of the stiffness of the arcs in the cluster to which these nodes belong is similar to the distribution of the stiffness of the arcs in the overall network. Stiffness is in fact a dynamic property of the system. The cluster of nodes and arcs with dynamic properties that are reflected in the dynamic properties of the entire network can thus be considered a cluster of driver nodes, a characteristic that designates it as a prime candidate for further wet experiments. In the case of our study, experiments and data from the literature support the hypotheses formulated by the computational analysis, as we shall see below. Indeed, the results we report below are intended to demonstrate these statements.

Figure 7, shows the barplot of the percentage of edges connecting nodes belonging to the same cluster of radial coordinates. Of the 2,162,160 total edges of the graph 1,512 belong to cluster 5, 2,877 to cluster 9, and 19,701 to cluster 12. The remaining 2,138,069 arcs connect nodes belonging to different clusters. In order to understand whether and, if so, how clustering according to radial co-ordinate is reflected in the distribution of spring stiffness, we produced the graph in Figure 8, showing the density plots of the spring stiffness of the interactions between node within the three clusters (5, 9 and 12) compared with the density plot of all spring stiffness of the network. To make the results easier to read and understand, we rescaled the spring stiffness values obtained by formula (22) within a range between 0 and 1 and applied formula (23) to the values obtained in this range.

Of interest we find as shown in this Figure 8 the two peaks of the density plot in red colour corresponding to the stiffness of the interactions between the nodes belonging to cluster number 12. Of the three clusters of radial node distance, number 12 is the one that best reflects the density plot of total spring stiffness. The interactions between nodes belonging to cluster 12 are markedly clustered as is the distribution of stiffnesses across all the arcs of the graph. We interpret this result as the fact that cluster 12 contains nodes that share similar popularity values and are involved in driver interactions of the network dynamics, since the distribution of spring stiffnesses of the arcs of these nodes reproduce the distribution of spring stiffnesses of the entire network.

Cluster 12 contains 199 genes, which a functional analysis implemented with the `enrichGO` function of R library `clusterProfiler` for the Gene Ontology (GO) Enrichment Analysis (Yu, 2012; Yu et al., 2012) finds to have the molecular functions shown in barplot of Figure 9 and the ontologies of the cellular compartments as in Figure 10. The list of the gene names of

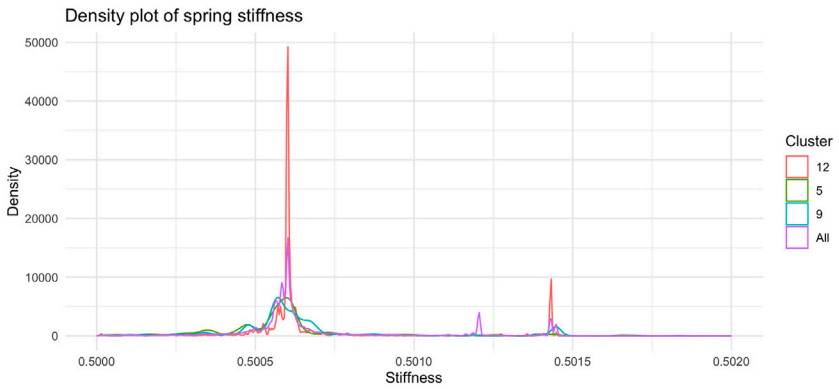


FIGURE 8
Density plot of the distribution of spring stiffness in the tree clusters (5, 9, and 12) and in whole dataset. The trend of the red curve for Cluster 12 is the one that most accurately reproduces the global density plot of the stiffness of the arches of the entire network. The nodes in Cluster 12 recapitulate the overall dynamics of the network through their interactions and can therefore be regarded as putative drivers of the dynamics.

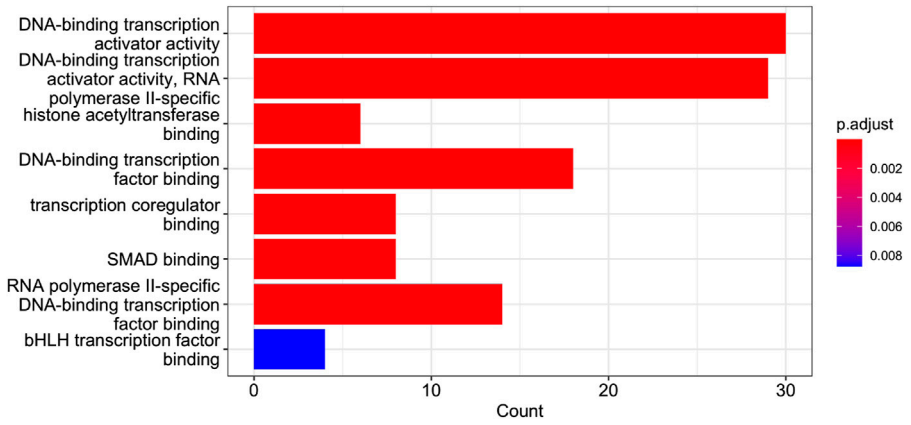


FIGURE 9
GO Enrichment Analysis of the gene set of Cluster 12. The barplot shows the enrichment GO categories of molecular functions after false discovery rate control. See also the verbose tabular outbut in Cluster_12_GSEA_results_EnrichGO_MF.xlsx provided in Supplementary Material.

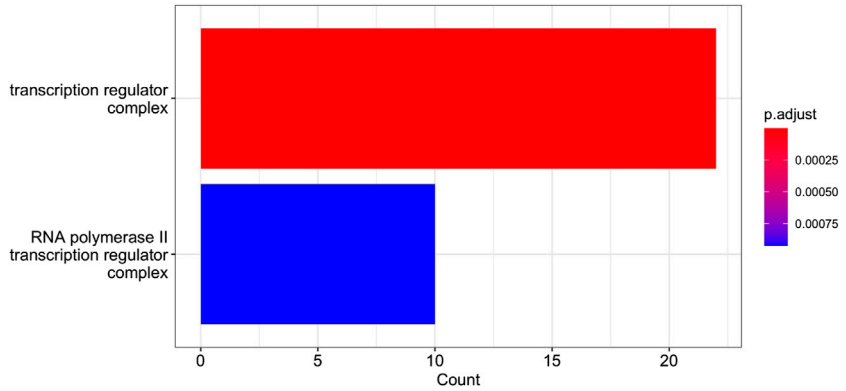


FIGURE 10
GO Enrichment Analysis of the gene set of Cluster 12 (obtained with the R function `enrichGO`). The barplot shows the enrichment GO categories of cellular compartment after false discovery rate control. See also the verbose tabular outbut in Cluster_12_GSEA_results_EnrichGO_CC.xlsx provided in Supplementary Material.

Cluster 12, as well as the summary of `enrichGO` and of `gost` function of the R library `gprofiler2` (Raudvere et al., 2019; Kolberg et al., 2020; Raudvere et al., 2023) are available in the Supplementary Material. To give a more complete view of the results of the gene set enrichment analysis of Cluster 12, we show in Figure 11 the Manhattan-plots of the gene set enrichment analysis, of which we also give an interactive version in the Supplementary Material.

Of interest is a result shown in Figure 9, namely, the presence in Cluster 12 of genes co-involved in the molecular processes of "SMAD binding". Smad proteins, are central mediators of the signal transduction of TGF-β family members were identified in the dataset analysed. A Cross-talk between TGF-β/Smad pathway and Wnt/β-catenin pathway in pathological scar formation has been described suggesting a complicated interaction between the two signal pathways in pathological scar formation (both synergy and antagonism) (Sun et al., 2015). More recently TGF-β/SMAD, Hippo/YAP/TAZ, and Wnt/β-catenin signalling pathways, major inducers of transcriptional reprogramming, were shown to converge at several levels and were all required for a proliferative-to-invasive phenotype switch in melanoma development (Lüönd et al., 2021). We already described in a previous study the involvement of Wnt/β-catenin signalling pathway in the tumour suppressor effect driven by PTPRG in CML (Tomasello et al., 2020) and the current data reporting the involvement of SMAD pathway is in line with a complex cellular reprogramming induced by PTPRG expression whose key role in the haematopoietic differentiation program was already described (Sorio et al., 1997). This complex reprogramming is supported by the large number pathways involved in DNA binding/transcription reported on Cluster 12 GSEA. In particular, in our previous study (Lombardi et al., 2022), we validated the SMAD1 gene. Specifically, qRT-PCR was used to assess gene mRNA levels, and the relative fold changes were calculated between K562 expressing PTPRG and the untreated control group (control and D1028A). The endogenous control was GAPDH. We found that the fold change of SMAD1 is markedly greater in the case of the control [see Figure 3 of Lombardi et al. (2022)].

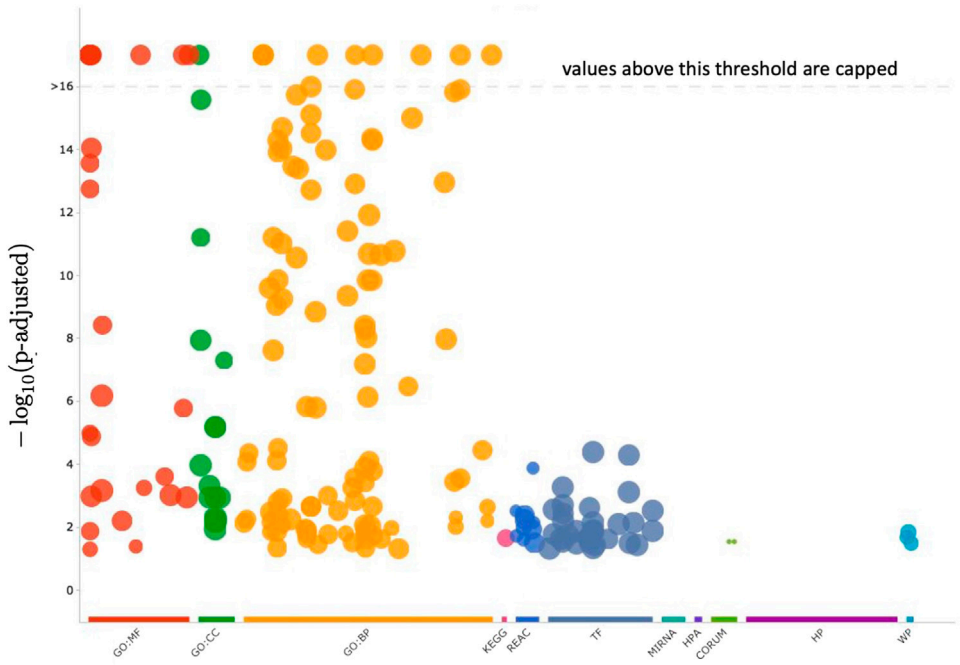


FIGURE 11
In this figure, the enrichment results of the gene set of Cluster 12 are visualized with a Manhattan-like-plot using the function `gostplot` (Raudvere et al., 2023). The *x*-axis depicts functional terms that are colour-coded and categorised according to data sources and positioned in the fixed "source_order." The order is set up so that terms that are close together in the source hierarchy are also close together in the Manhattan plot. The modified *p*-values are displayed on the *y*-axis in negative $\log_{10}$ scale. Every circle represents one phrase and is proportional to the term size, i.e., larger terms have larger circles. The Supplementary Material includes an interactive version of this plot (`Manhattan_plot_GSEA_Gost.html`). Hovering over the circle in the interactive plot will display the appropriate information. If the $-\log_{10}$ (*p*-values) exceed 16, they are capped at 16. This adjusts the *y*-axis scale to keep Manhattan plots from different queries similar, and it is also intuitive because statistically, *p*-values less than that can all be summarized as highly significant.

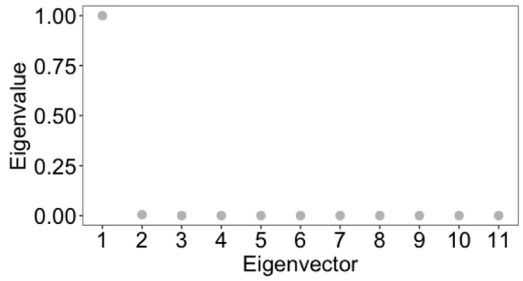


FIGURE 12
Eigengap heuristic: the optimal number of clusters, *k*, that maximises the eigengap (difference between consecutive eigenvalues of the Laplacian matrix of the graph). The optimal number of clusters is that k such that λ_{k+1} is reasonably large but all other eigenvalues, $\lambda_1, \ldots, \lambda_k$, are very small. The closer the eigenvectors of the ideal case are, and hence the better spectral clustering performs, the wider this eigengap is.

5.1 Comparison with spectral clustering

We compared the results of the clustering by radial coordinate in hyperbolic space with the spectral clustering method. This method is widely used to identify communities of nodes in a network by examining the edges that connect them, i.e., taking as an input the weighted adjacency matrix of the graph. It is a well-established method with theoretical foundations in graph theory (we refer the reader to JingMao and YanXia (2015); von Luxburg (2007) for a review and a tutorial on this popular spectral clustering methods). Processing directly the weighted adjacency matrix of the graph, that is the same input as our embedding procedure, we consider spectral clustering to be the most appropriate method to deal with, compared to clustering methods based on graph centrality measures, or on statistical correlation measures between nodes, who do not into account directly distance measures between nodes. Before applying spectral clustering, we estimated the optimal number of clusters with eigengap heuristics [appropriate procedure for estimating the number of clusters for spectral clustering methods (von Luxburg, 2007)], obtaining that the optimal number of clusters is 2 (see Figure 12). Cluster 1 contains 1,731 nodes and cluster 2 contains 349 nodes. Using the R script Spectral_clustering.R to implement spectral clustering - available in GitLab repository, we found that the within cluster sum of squares by cluster is 1.1079409 and 0.2348562, whereas the between sum of squares is 0.6502382. As a consequence, we conclude that the results of the spectral clustering are not reliable. This result highlights how taking into account the latent geometry of the network and with it the clustering according to the spatial co-ordinate of the nodes/points of the network resulted in a much better quality of clustering, compared to a clustering which, as in our

approach, processes the weighted adjacency matrix, but does not consider the latent geometry of the network expressed by the position of the points in the optimal embedding space and the distance defined by the metric in this space.

That various clustering methods are not appropriate for graphs with geometry has also been pointed out by Avrachenkov et al. (2021) that states that while it has been demonstrated that spectral clustering is consistent in some geometric graphs, a cut-based technique (such as spectral clustering) can also be significantly hindered by the geometric structure. It is possible to divide space into regions in such a way that there is relatively little interaction between nodes in two different regions. Therefore, the Fiedler vector of a geometric graph may only be linked to a geometric arrangement and contain no information regarding the labelling of the latent community. Furthermore, because the regions of space can include a balanced number of nodes, the widely used regularisation strategy (Zhang and Rohe, 2018), which seeks to penalise small size communities in order to bring back the vector associated with the community structure in the second position, would not function in geometric graphs.

6 Conclusion

In this study, we modelled the transcriptome network of the of Chronic Myeloid Laeukemia K562 cells overexpressing the tumour suppressor gene PTPRG, as a physical system of springs and then deduced the spring constant from topological properties of the nodes, such as total degree. To represent the network, we considered the dissimilarity matrix consisting of the values of the spring's elastic constant, which in our model quantifies the efficiency of information transmission between nodes. Through network embedding procedures that processed the dissimilarity matrix to derive the coordinates of the nodes in a metric os pact we determined the optimal latent geometry of the network is hyperbolic. This important information made it possible to proceed with the classification of nodes according to radial co-ordinates (which is the geometric equivalent of the 'physical' concept of node popularity) and to identify a set of candidate driver genes for network dynamics.

This methodology aimed at analysing a network without ignoring the existence of its metric space with a geometry other than the Euclidean one usually imposed or taken for granted, shows how latent geometry can determine a classification of nodes according to their relevance in the network's evolutionary processes, ultimately its dynamics. In the particular case study presented here we obtained that the network has hyperbolic latent geometry, and based on this we proceeded to utilise the concept that in this type of geometry the radial coordinate is a fundamental variable for clustering nodes. Geometries other than hyperbolic are characterised by other spatial variables that can be considered discriminating for the purpose of identifying driver nodes of the dynamics. What is presented in the paper, besides being a concrete result on a specific case study, is also a proposal for a method of analysing a network in order to reveal information about the dynamics of the network itself.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

PL conceived the paper, the physical and mathematical modelling of the network analysed in the study, and implemented the R code for the graph embedding and node clustering. GL performed the differential expression analysis and produced the network that includes the differentially expressed genes found. RL and CS produced the experimental data. All authors contributed to the article and approved the submitted version.

Funding

This study has been supported by the fund of the DAQETA-CML (Detecting and quantifying (side-)effects of recent experimental therapies against Chronic Myeloid Leukemia) Interdisciplinary Project 2020, earmarked by the Free University of Bozen/Bolzano.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1235116/full#supplementary-material

References

Alanis-Lobato, G., Mier, P., and Andrade-Navarro, M. (2018). The latent geometry of the human protein interaction network. *Bioinformatics* 34, 2826–2834. doi:10.1093/bioinformatics/bty206

Allard, A., and Serrano, M. Á. (2020). Navigable maps of structural brain networks across species. *PLOS Comput. Biol.* 16, e1007584. doi:10.1371/journal.pcbi.1007584

Anderson, J. W. (2005). *Hyperbolic geometry*. 2 edn. London, England: Springer.

Avrachenkov, K., Bobu, A., and Dreveton, M. (2021). Higher-order spectral clustering for geometric graphs. *J. Fourier Analysis Appl.* 27, 22. doi:10.1007/s00041-021-09825-2

Blevins, A. S., Kim, J. Z., and Bassett, D. S. (2021). Variability in higher order structure of noise added to weighted networks. *Commun. Phys.* 4, 233. doi:10.1038/s42005-021-00725-x

Boguñá, M., Bonamassa, I., Domenico, M. D., Havlin, S., Krioukov, D., and Serrano, M. Á. (2021). Network geometry. *Nat. Rev. Phys.* 3, 114–135. doi:10.1038/s42254-020-00264-4

Boguñá, M., Papadopoulos, F., and Krioukov, D. (2010). Sustaining the internet with hyperbolic mapping. *Nat. Commun.* 1, 62. doi:10.1038/ncomms1063

Borg, I., and Groenen, P. J. F. (2005). *Modern multidimensional scaling*. 2 edn. New York, NY: Springer.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2010a). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 39, D685–D690. doi:10.1093/nar/gkq1039

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2010b). Pathway commons: a resource for biological pathway analysis — pathwaycommons.org. Avaliable at: https://www.pathwaycommons.org/(Accessed May 27, 2023).

Clauset, A., Moore, C., and Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 98–101. doi:10.1038/nature06830

Conn, Z. (2010). Distances in the hyperbolic plane and the hyperbolic Pythagorean theorem. Avaliable at: http://www.zachconn.com/.

Cox, M. A. A., and Cox, T. F. (2008). *Multidimensional scaling*. Berlin, Heidelberg: Springer Berlin Heidelberg, 315–347. doi:10.1007/978-3-540-33037-0_14

Estrada, E., and Hatano, N. (2010). A vibrational approach to node centrality and vulnerability in complex networks. *Phys. A Stat. Mech. its Appl.* 389, 3648–3660. doi:10.1016/j.physa.2010.03.030

Fard, A. T., Srihari, S., Mar, J. C., and Ragan, M. A. (2016). Not just a colourful metaphor: modelling the landscape of cellular development using hopfield networks. *npj Syst. Biol. Appl.* 2, 16001. doi:10.1038/npjsba.2016.1

Gower, J. C. (2015). "Principal coordinates analysis," in *Wiley StatsRef: Statistics Reference Online* (Wiley), 1–7. doi:10.1002/9781118445112.stat05670.pub2

Gromov, M. (2007). *Metric structures for Riemannian and non-Riemannian spaces*. Boston: Birkhäuser. doi:10.1007/978-0-8176-4583-0

Härtner, F., Andrade-Navarro, M. A., and Alanis-Lobato, G. (2018). Geometric characterisation of disease modules. *Appl. Netw. Sci.* 3, 10. doi:10.1007/s41109-018-0066-3

Hilbert, D. (1933). "Über flächen von konstanter gaußscher krümmung," in *Algebra invariantentheorie geometrie* (Berlin Heidelberg: Springer), 437–448. doi:10.1007/978-3-642-52012-9_30

Jhun, B. (2022). Topological analysis of the latent geometry of a complex network. *Chaos Interdiscip. J. Nonlinear Sci.* 32, 013116. doi:10.1063/5.0073107

JingMao, Z., and YanXia, S. (2015). "Review on spectral methods for clustering," in 2015 34th Chinese Control Conference (CCC), Hangzhou, China, 28-30 July 2015 (IEEE), 3791–3796. doi:10.1109/ChiCC.2015.7260226

Keller-Ressel, M. (2019). hydra: hyperbolic Embedding — cran.r-project.org. Avaliable at: https://cran.r-project.org/web/packages/hydra/index.html (Accessed May 27, 2023).

Keller-Ressel, M., and Nargang, S. (2020). Hydra: a method for strain-minimizing hyperbolic embedding of network- and distance-based data. *J. Complex Netw.* 8. doi:10.1093/comnet/cnaa002

Kleinberg, J. M. (2000). Navigation in a small world. *Nature* 406, 845. doi:10.1038/35022643

Klimovskaia, A., Lopez-Paz, D., Bottou, L., and Nickel, M. (2020). Poincaré maps for analyzing complex hierarchies in single-cell data. *Nat. Commun.* 11, 2966. doi:10.1038/s41467-020-16822-4

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., and Peterson, H. (2020). gprofiler2 – an r package for gene list functional enrichment analysis and namespace conversion toolset g:profiler. *F1000Research* 9, ELIXIR-709. doi:10.12688/f1000research.24956.2

Kovács, B., and Palla, G. (2021). The inherent community structure of hyperbolic networks. *Sci. Rep.* 11, 16050. doi:10.1038/s41598-021-93921-2

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M. (2010). Hyperbolic geometry of complex networks. *Phys. Rev. E* 82, 036106. doi:10.1103/PhysRevE.82.036106

Kurkofka, J., Melcher, R., and Pitz, M. (2021). Approximating infinite graphs by normal trees. *J. Comb. Theory, Ser. B* 148, 173–183. doi:10.1016/j.jctb.2020.12.007

Lecca, P. (2023). Uncovering the geometry of protein interaction network: The case of SARS-CoV-2 protein interactome. *AIP Conf. Proc.* 2872 (1), 030008. doi:10.1063/5.0163052

Lecca, P., and Re, A. (2020). "Stiffness estimate of information propagation in biological systems modelled as spring networks," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 16-19 December 2020 (IEEE). doi:10.1109/bibm49941.2020.9313294

Lecca, P., and Re, A. (2022). "Checking for non-euclidean latent geometry of biological networks," in 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 06-08 December 2022 (IEEE). doi:10.1109/bibm55620.2022.9995274

Lecca, P., Re, A., Lombardi, G., Latorre, R. V., and Sorio, C. (2023). "Graph embedding of chronic myeloid leukaemia k562 cells gene network reveals a hyperbolic latent geometry," in *Proceedings of eighth international congress on information and communication technology* (Singapore: Springer Nature), 979–991. doi:10.1007/978-981-99-3091-3_80

Liben-Nowell, D., and Kleinberg, J. (2007). The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* 58, 1019–1031. doi:10.1002/asi.20591

Lombardi, G., Latorre, R. V., Mosca, A., Calvanese, D., Tomasello, L., Boni, C., et al. (2022). Gene expression landscape of chronic myeloid leukemia k562 cells overexpressing the tumor suppressor gene PTPRG. *Int. J. Mol. Sci.* 23, 9899. doi:10.3390/ijms23179899

Lüönd, F., Pirkl, M., Hisano, M., Prestigiacomo, V., Kalathur, R. K., Beerenwinkel, N., et al. (2021). Hierarchy of TGFβ/SMAD, hippo/YAP/TAZ, and wnt/β-catenin signaling in melanoma phenotype switching. *Life Sci. Alliance* 5, e202101010. doi:10.26508/lsa.202101010

Maddalena, L., Giordano, M., Manzo, M., and Guarracino, M. R. (2022). "Whole-graph embedding and adversarial attacks for life sciences," in *Trends in biomathematics: stability and oscillations in environmental, social, and biological models. BIOMAT 2021*. Editor R. P. Mondaini (Springer, Cham). doi:10.1007/978-3-031-12515-7_1

Millán, A. P., Torres, J. J., and Bianconi, G. (2018). Complex network geometry and frustrated synchronization. *Sci. Rep.* 8, 9910. doi:10.1038/s41598-018-28236-w

NIH (2023). ZRANB1 zinc finger RANBP2-type containing 1 [*Homo sapiens* (human)]. Avaliable at: https://www.ncbi.nlm.nih.gov/gene/54764 (Accessed May 30, 2023).

Okuno, A., and Shimodaira, H. (2019). "Robust graph embedding with noisy link weights," in *The 22nd international conference on artificial intelligence and statistics, AISTATS 2019*. Editors K. Chaudhuri and M. Sugiyama (Naha, Okinawa, Japan: PMLR), 664–673.

Papadopoulos, F., and Flores, M. A. R. (2019). Latent geometry and dynamics of proximity networks. *Phys. Rev. E* 100, 052313. doi:10.1103/PhysRevE.100.052313

Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguñá, M., and Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature* 489, 537–540. doi:10.1038/nature11459

Pio, G., Ceci, M., Prisciandaro, F., and Malerba, D. (2019). Exploiting causality in gene network reconstruction based on graph embedding. *Mach. Learn.* 109, 1231–1279. doi:10.1007/s10994-019-05861-8

Pontén, F., Jirström, K., and Uhlen, M. (2008). The human protein atlas—a tool for pathology. *J. Pathology* 216, 387–393. doi:10.1002/path.2440

Rand, D. A., Raju, A., Sáez, M., Corson, F., and Siggia, E. D. (2021). Geometry of gene regulatory dynamics. *Proc. Natl. Acad. Sci.* 118, e2109729118. doi:10.1073/pnas.2109729118

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198. doi:10.1093/nar/gkz369

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2023). Gene list functional enrichment analysis and namespace conversion with gprofiler2. Avaliable at: https://cran.r-project.org/web/packages/gprofiler2/vignettes/gprofiler2.html (Accessed June 02, 2023).

R Core Team (2021). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sánchez-Romero, M. A., and Casadesús, J. (2021). Waddington's landscapes in the bacterial world. *Front. Microbiol.* 12, 685080. doi:10.3389/fmicb.2021.685080

Seyboldt, R., Lavoie, J., Henry, A., Vanaret, J., Petkova, M. D., Gregor, T., et al. (2022). Latent space of a small genetic network: geometry of dynamics and information. *Proc. Natl. Acad. Sci.* 119, e2113651119. doi:10.1073/pnas.2113651119

Sorio, C., Melotti, P., D'Arcangelo, D., Mendrola, J., Calabretta, B., Croce, C. M., et al. (1997). Receptor protein tyrosine phosphatase gamma, ptp gamma, regulates hematopoietic differentiation. *Blood* 90, 49–57. doi:10.1182/blood.v90.1.49

Squier, S. M. (2017). *Epigenetic landscapes: drawings as metaphor*. Durham, NC, United States: Duke University Press.

Sun, N., Pei, S., He, L., Yin, C., He, R. L., and Yau, S. S. T. (2021). Geometric construction of viral genome space and its applications. *Comput. Struct. Biotechnol. J.* 19, 4226–4234. doi:10.1016/j.csbj.2021.07.028

Sun, Q., Guo, S., Wang, C. C., Sun, X., Wang, D., Xu, N., et al. (2015). Cross-talk between TGF-β/Smad pathway and Wnt/β-catenin pathway in pathological scar formation. *Int. J. Clin. Exp. Pathol.* 8, 7631–7639.

Taylor, D., Klimm, F., Harrington, H. A., Kramár, M., Mischaikow, K., Porter, M. A., et al. (2015). Topological data analysis of contagion maps for examining spreading processes on networks. *Nat. Commun.* 6, 7723. doi:10.1038/ncomms8723

Tomasello, L., Vezzalini, M., Boni, C., Bonifacio, M., Scaffidi, L., Yassin, M., et al. (2020). Regulative loop between β-catenin and protein tyrosine receptor type γ in chronic myeloid leukemia. *Int. J. Mol. Sci.* 21, 2298. doi:10.3390/ijms21072298

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2022a). The human protein atlas — proteinatlas.org. Avaliable at: https://www.proteinatlas.org/(Accessed May 30, 2023).

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2022b). ZRANB1 protein expression summary - the Human Protein Atlas — proteinatlas.org. Avaliable at: https://www.proteinatlas.org/ENSG00000019995-ZRANB1 (Accessed May 30, 2023).

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science* 357, eaan2507. doi:10.1126/science.aan2507

Umargono, E., Suseno, J. E., and Vincensius Gunawan, S. K. (2020). "K-means clustering optimization using the elbow method and early centroid determination based-on mean and median," in Proceedings of the International Conferences on Information System and Technology – CONRIST, Yogyakarta, Indonesia (Setúbal, Portugal: SciTePress) 1, 234–240. doi:10.5220/0009908402340240

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics Comput.* 17, 395–416. doi:10.1007/s11222-007-9033-z

Wilson, R. C., Hancock, E. R., Pekalska, E., and Duin, R. P. (2014a). Spherical and hyperbolic embeddings of data. *IEEE Trans. Pattern Analysis Mach. Intell.* 36, 2255–2269. doi:10.1109/TPAMI.2014.2316836

Wilson, R. C., Pekalska, E. R. H. E., and Duin, R. P. (2014b). Classical (metric) multidimensional scaling. Avaliable at: https://www.cs.york.ac.uk/cvpr/post/sphericalembedding/(Accessed May 23, 2023).

Yang, W., and Rideout, D. (2020). High dimensional hyperbolic geometry of complex networks. *Mathematics* 8, 1861. doi:10.3390/math8111861

Yu, G. (2012). enrichGO function - RDocumentation — rdocumentation.org. Avaliable at: https://www.rdocumentation.org/packages/clusterProfiler/versions/3.0.4/topics/enrichGO (Accessed June 02, 2023).

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zhang, Y., and Rohe, K. (2018). "Understanding regularized spectral clustering via graph conductance," in *Proceedings of the 32nd international conference on neural information processing systems* (Red Hook, NY, USA: Curran Associates Inc.), 10654–10663.

Zhang, Z., and Takane, Y. (2010). "Multidimensional scaling," in *International encyclopedia of education*. Editors P. Peterson, E. Baker, and B. McGaw (Oxford: Elsevier), 304–311. doi:10.1016/B978-0-08-044894-7.01348-8

frontiers | Frontiers in Cell and Developmental Biology

TYPE Hypothesis and Theory
PUBLISHED 26 June 2024
DOI 10.3389/fcell.2024.1240384

OPEN ACCESS

EDITED BY
Susan Mertins,
Leidos Biomedical Research, Inc., United States

REVIEWED BY
Chen Zhao,
Nanjing Medical University, China
Reka Albert,
The Pennsylvania State University (PSU), United States

*CORRESPONDENCE
Ravi Iyengar,
ravi.iyengar@mssm.edu

RECEIVED 14 June 2023
ACCEPTED 30 May 2024
PUBLISHED 26 June 2024

CITATION
Hansen J, Jain AR, Nenov P, Robinson PN and Iyengar R (2024), From transcriptomics to digital twins of organ function.
Front. Cell Dev. Biol. 12:1240384.
doi: 10.3389/fcell.2024.1240384

From transcriptomics to digital twins of organ function

Jens Hansen[1], Abhinav R. Jain[1], Philip Nenov[1,2], Peter N. Robinson[3] and Ravi Iyengar[1]*

[1]Department of Pharmacological Science and Institute for Systems Biomedicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [2]College of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, United States, [3]Berlin Institute of Health at Charité Rahel Hirsch Center for Translational Medicine, Berlin, Germany

Cell level functions underlie tissue and organ physiology. Gene expression patterns offer extensive views of the pathways and processes within and between cells. Single cell transcriptomics provides detailed information on gene expression within cells, cell types, subtypes and their relative proportions in organs. Functional pathways can be scalably connected to physiological functions at the cell and organ levels. Integrating experimentally obtained gene expression patterns with prior knowledge of pathway interactions enables identification of networks underlying whole cell functions such as growth, contractility, and secretion. These pathways can be computationally modeled using differential equations to simulate cell and organ physiological dynamics regulated by gene expression changes. Such computational systems can be thought of as parts of digital twins of organs. Digital twins, at the core, need computational models that represent in detail and simulate how dynamics of pathways and networks give rise to whole cell level physiological functions. Integration of transcriptomic responses and numerical simulations could simulate and predict whole cell functional outputs from transcriptomic data. We developed a computational pipeline that integrates gene expression timelines and systems of coupled differential equations to generate cell-type selective dynamical models. We tested our integrative algorithm on the eicosanoid biosynthesis network in macrophages. Converting transcriptomic changes to a dynamical model allowed us to predict dynamics of prostaglandin and thromboxane synthesis and secretion by macrophages that matched published lipidomics data obtained in the same experiments. Integration of cell-level system biology simulations with genomic and clinical data using a knowledge graph framework will allow us to create explicit predictive models that mechanistically link genomic determinants to organ function. Such integration requires a multi-domain ontological framework to connect genomic determinants to gene expression and cell pathways and functions to organ level phenotypes in healthy and diseased states. These integrated scalable models of tissues and organs as accurate digital twins predict health and disease states for precision medicine.

KEYWORDS

digital twin, dynamical modeling, systems biology, networks, transcriptomics

Introduction

Accurate multiscale computational models of physiological functions of different organs within the human body have the potential to revolutionize our understanding of human biology and greatly advance the practice of medicine. Vast amounts of data are being collected in different domains of genomics, biochemistry, cell biology and physiology and clinical sciences. It will be necessary to bring together these data to understand the physiology of organ systems. Physiology is dynamics (Rubin et al., 2019). Understanding how the function of organs changes over time is essential for understanding both homeostasis for health and disease origins and progression. The functions of organs arise from cell-level physiological activity. Examples include the heart, where ability of cardiomyocytes to contract in a rhythmic and coordinated fashion underlie the beating of heart, and the kidney where ability of different cell types of the nephron to filter large molecules and reabsorb ions, water and small molecules underlie our ability to regulate water balance, excrete end products of metabolism, maintain pH balance in blood, and control blood pressure. Thus, to generate accurate predictive models of organ function, the first step is to build accurate models of whole cell functions. Such models should consider the key components and pathways within the cell; the networks that arise from interactions between pathways and pathway components; the topological features of the networks including the feedback loops, feedforward loops and bifans (Milo et al., 2002) which enable processing of information within the cell (Ma'ayan et al., 2005); and state changes driven by bistable switches (Bhalla and Iyengar, 1999; Tanaka and Augustine, 2008).

To go from cell-based models to organ level models we need to consider how the different cell types in the organ function and interact as well as the role of the extracellular matrix in controlling the mechanical and signaling properties of the organ. Multiple anatomical structures make up each organ. Blood vessels are one example of tissue components contributing to an organ's physiology. Blood vessels have vascular smooth muscle cells, fibroblasts, endothelial cells (Sturtzel, 2017) that line the wall of the blood vessels and make up the capillaries, as well as pericytes (Lee and Chintalgattu, 2019) in some organs. The latter two cell types are often the source of important signaling molecules and sense mechanical forces such as the pressure from blood flow to control organ function.

Changes in cell state are driven by changes in gene expression patterns that control whole cell responses. Transcriptomic profiles represent cell identity as well as cell state. Hence, we hypothesize that changes in gene expression patterns can be used to predict dynamic physiological capabilities. We describe our initial approach to test this hypothesis and provide preliminary evidence that the approach we propose could work. Our approach consists of two sets of operations that integrate two different modeling approaches. First, we take a ranked list of genes, typically differentially expressed mRNAs indicative of two different conditions (states) the cells or organs are in and create networks using pathway information from prior knowledge databases. These interacting pathways are enriched for the differentially expressed genes and could account for change in activity. Going from genes to pathways using prior knowledge is a very widely used statistical modeling approach called gene-set enrichment analysis (Subramanian et al., 2005). Second, the reactions participating in identified pathways that together make up edges in directed subgraphs or graphs are readily converted to systems of coupled differential equations. These systems of coupled differential equations are dynamical models that can be used to run simulations to predict how cell biochemical or physiological functions change with time. Here, we describe how this two-step algorithm can work, and eventually become part of a larger algorithm for a digital twin. In biology, digital twins can be thought of multi-scale computational models that can predict physiological events from genomic and molecular data. Such predictions may be at the cell level, tissue/organ level or at the whole organism level. In this review we consider the cell and organ levels.

Computational approaches to modeling dynamics

To support widespread use of single cell transcriptomics multiple approaches to conduct trajectory analyses from time series and single timepoint experiments have been published, and these approaches are described and compared in a review article (Ding et al., 2022). This approach has been particularly useful in mapping trajectories during developmental processes and provide useful insight into precursor and differentiated cell types in many organ systems. However, all these approaches provide pseudo-time series outputs that can only be constrained by experimental time series analyses. Pseudo time series order entities with respect to one another to infer trajectories. For example, ligand activation of receptor and stimulation of membrane effectors occur prior to activation of protein kinases. This information can be used to develop trajectories from receptors to physiological effectors such as channels and metabolic enzymes. Pseudo time series analyses do have value in understanding the progression of biological states and we had used pseudo time series in a 2005 study (Ma'ayan et al., 2005) to understand the role of regulatory motifs such as feedforward and feedback loops in signal propagation from receptor to transcription factors to control the duration of transcription factor activation. Orthogonal experimental approaches such as single nucleus ATAC Seq and CRISPR/Cas9 mediated gene modification provide mechanistic insights into trajectory analyses and together they may help define realistic time-dependent predictions in the future. The limitation of pseudo-time series to capture physiological dynamics lies in its inability to be scalable and hence is likely to be of limited value in realistic digital twins.

A combination of proteomic and phenotypic feature measurements to identify new drug combinations that would work on drug resistant cells uses differential equation-based modeling to develop predicted responses of cancer cells (Frohlich et al., 2018). The approach is similar to the PK/PD modeling widely used in pharmacology that is a mainstay in the drug discovery process. Such approaches that integrate perturbation data with prior pathway information can predict drug responses, especially responses to combination therapy. The Cell Box Software suite (Yuan et al., 2021) provides a useful tool set for such analyses including network development in a purely data driven manner. Limitations of such a modeling approach is that the captured

perturbation dynamics depend on many undefined reactions and rate constants and hence it is uncertain whether such an approach will work under different physiological states and conditions without specific large-scale gathering of experimental data for each condition.

An integrative dynamical model using coupled differential equations that are solved in a standard solver using MATLAB has been developed to predict macrophage polarization (Zhao et al., 2021; Zhao and Popel, 2021). The scope of the model is extensive and impressive, although surprisingly the prostaglandin biosynthesis and signaling pathways are missing. Nevertheless, the model represents an important step in the development of the virtual macrophage that can predict macrophage polarization and functions in various physiological states. Such models could well be adapted to describe other types of blood cells although and their trajectories in health and disease. Beyond cell level models, these researchers have proposed approaches that integrate omics data and dynamical models for tissue level angiogenesis models that represent communication between different cell types (Zhang et al., 2022). Such approaches are likely to be useful in developing digital twins for angiogenesis.

The approach we propose here has some similarities and differences with these previously described models. Our approach is focused on getting the cell-level molecular and pathway details "right" and then determining if dynamical models based on granular biochemical and biophysical reactions can be used to predict and understand physiological behaviors at the cell level and at the organ level. The pros and cons of this approach and its use as the core of digital twins of organs are discussed below.

Advantages and challenges in the use of numerical analyses to predict physiological dynamics

Modeling biochemical and physiological processes using standard chemical kinetics is better than most other approaches because this is the most realistic representation of these processes including those involved in generation and sensing of forces. We have long favored the use of chemical kinetics representations and shown that we can make non-intuitive experimentally verifiable predictions. Our model using systems of ordinary differential equations (Bhalla and Iyengar, 1999) that predicted the existence bistable positive feedback loops that can enable switching cellular states has been experimentally validated by others in cerebellar long-term depression (Tanaka and Augustine, 2008). Our spatial partial differential equation model predicting selective cAMP accumulation in dendrites as compared to cell body of neurons (Neves et al., 2008) was validated using a cAMP biosensor in mouse brain slice tissues by Castro et al. (2010). We have continued to use this approach to develop predictive models of interactions between subcellular processes. We predicted that dynamic balance between membrane vesicle transport and microtubule growth is required for neurite outgrowth (Yadaw et al., 2019). We used gene knockdown of vesicle transport and docking protein to demonstrate the validity of our prediction (Yadaw et al., 2019; Hansen et al., 2022). Despite these successes, challenges have always been present. Initially some of the challenges were computational, such as computational costs and propagation of errors. With the exponential increase in computational capability these challenges have become less of a barrier. However, the biological challenges persist. The cellular concentrations of most proteins have yet to be explicitly measured in most cell types of the human body, although it is often possible to estimate or guesstimate them from the vast biochemistry and cell biology literature. Also, reaction rates are often not known. Databases such as BRENDA (Schomburg et al., 2017) are useful, although kinetic information regarding mammalian systems is limited. Another useful resource is Bionumbers which contains many "average" values used to set up the models for numerical simulations (Milo et al., 2010).

Gene expression changes to neurite outgrowth, a whole cell response: identifying and modeling cell regulatory pathways and networks

In a recent study, we have shown how transcriptional patterns can be used by cells to drive cell state changes and whole cell responses to external signals through well-known canonical pathways (Hansen et al., 2022). Although our study is based on bulk transcriptomics and discovery proteomics obtained from only one cell line cultured in isolation, our analysis strategy should be applicable to single cell transcriptomics and other omics technologies as well. Briefly, we treated the neuronal cell line Neuro2A (N2A) with an agonist for the cannabinoid receptor 1 (CB1R) to induce neurite outgrowth. Differentially expressed genes and proteins induced after different stimulation periods were subjected to pathway enrichment analysis (Figure 1), using the Molecular Biology of the Cell Ontology (MBCO), a cell biology focused ontology that was generated in our lab (Hansen et al., 2017).

We identified many subcellular processes (SCPs), which are commonly thought of as constitutive pathways that are operational in many, if not all cell types. While these SCPs such as alternative splicing, pyrimidine salvage and membrane protein synthesis are universal, the ability of the extracellular signal to regulate them in a coordinated manner gives the cell additional capacity to mount the whole cell response. Our data documents that the canonical SCPs are activated in a chronological order that matches their dependencies (Figure 2). It can be readily seen that many cellular pathways in different organelles such as the nucleus, endoplasmic reticulum (ER), cytosol and growing neurite compartments are involved. Although shown in an abstracted form for clarity, each of the SCPs shown in Figure 2 contains multiple interacting proteins that come together to form larger functional networks. The different pathways must work in a highly coordinated fashion and imbalances in their coordination can lead to stoppage of the cellular responses. This conclusion is supported by dynamical modeling of one set of SCPs involved in transporting newly synthesized membranes as cytosolic vesicles from the Trans-Golgi network (TGN) through the neurite shaft to the growing tip at the end of the neurite. The new membrane is needed to build the axonal shaft as neurite grows. The importance of dynamics is inferred from the multicompartment ordinary differential equation (ODE) model that simulates the movement of newly synthesized vesicles from the TGN in the cell body to the growing tip. After

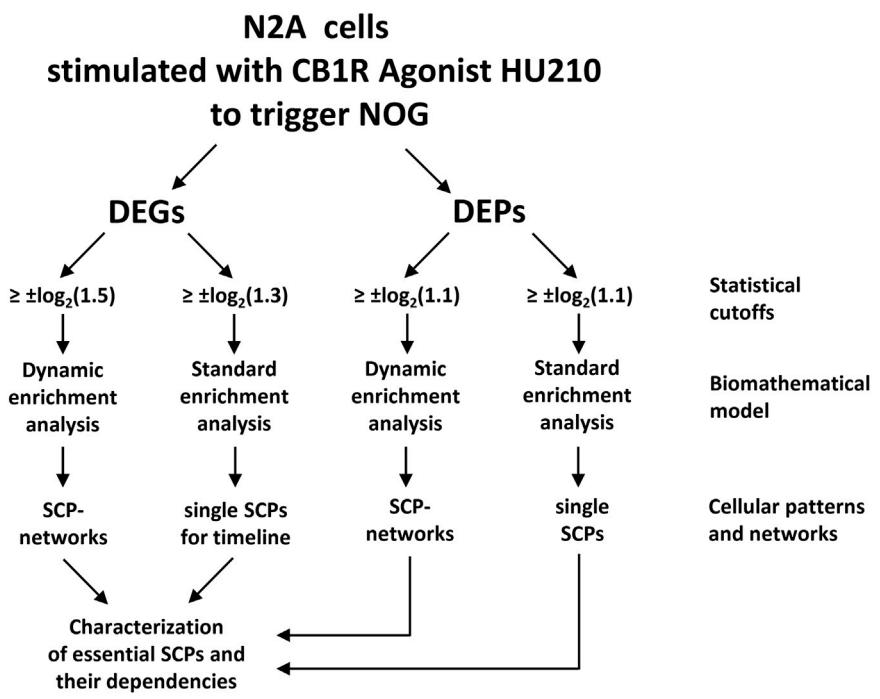


FIGURE 1
A flow chart showing the steps used for building networks of subcellular pathways (SCPs) underlying neurite outgrowth (NOG). Standard and dynamic enrichment analysis refer to methods used inferring pathway from differentially expressed genes (DEGs) or proteins (DEPs). Reproduced from and for details see Hansen et al. (2022).

developing an analytical solution for the prediction of parameter settings that allow neurite outgrowth at a given velocity and literature-curated model constraints with high accuracy, we could show how multiple pathways interact with each other to generate the whole cell response. Our analysis revealed that increased neurite outgrowth depends on increased backward vesicle traffic from the neurite tip to the TGN (Figure 3). This initially counter intuitive dependency ensures back transport of components needed for forward vesicle traffic. Such focused simulations within the larger overall computational model are likely to be critical parts for verification of the underlying pathways and validation of mechanisms at the subcellular levels could also be used to parameterize and identify the uncertainty in how interactions between SCP subnetworks as well as interactions with the cell and extracellular matrix lead to dynamics of organ level functions.

Predicting cardiomyocyte electrophysiology and contractility from transcriptomic changes

The ability to develop a dynamical model for cell functions is dependent on the pathways and networks inferred from the DEGs and DEPs. Once these networks are identified, pathway activities can be readily connected to systems of coupled differential equations that can be used for multi-compartment ODE models or PDE models. Although most models that capture biochemical SCPs use a pathways framework, biophysical models can also capture changes in gene expression to predict responses to perturbation. In a recent study using cardiomyocytes differentiated from healthy human subjects, gene expression changes induced by tyrosine kinase inhibitor drugs that are effective cancer therapeutics was used to develop computational models that predict arrhythmogenic responses to cancer drug therapy in individuals (Shim et al., 2023). Changes in levels of gene expression of different channel proteins by drugs were scaled and incorporated as changes in level of channel proteins into a multicompartment ODE model of cardiomyocyte action potential and contractility. Experimental measurements of cardiomyocyte action potentials, intracellular calcium, and contraction in the cardiomyocytes demonstrated that modeling predictions were mostly (80%) accurate. The simulations were also able to predict responses to drugs and a second perturbation such as hypokalemia (low potassium). Together the biochemical and biophysical models demonstrate the ability of numerical simulations to use transcriptomic data for predictions.

An integrated algorithm to go from differentially expressed genes to biochemical dynamics: eicosanoid biosynthesis network in macrophages

We developed a computational pipeline that integrates a canonical model of interest with transcriptomic or proteomic data – either bulk or single cell - to develop cell-type selective dynamical models for the prediction of cell-type selective whole cell responses (Figure 4). The canonical model would involve all known enzymes and reactions described for any cell type of the same

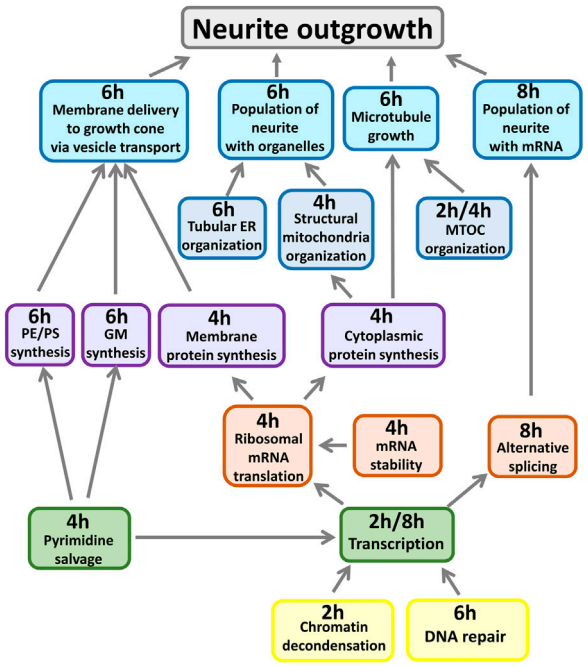


FIGURE 2
Whole cell response requires deep and distributed responses. All shown SCPs are related to growth of the neurite shaft or scaffold. Described functions summarize pathway activities predicted at indicated time points from gene expression profiles induced by CB1R stimulation. MTOC: Microtubule organization center, PE/PS: Phosphatidylethanolamine/-choline, GM: Ganglioside, ER: Endoplasmic reticulum. Adapted from and for details see Hansen et al. (2022).

organism. Like others (Frohlich et al., 2018), we assume that the reaction rate parameters are canonical as well, i.e., they are the same for all cell types within an organism. Experimental confirmation of our spatial cAMP models (Neves et al., 2008) by others (Castro et al., 2010) indicate that this is likely to be true. Starting sources for the construction of a canonical model could be the KEGG metabolic networks (Kanehisa et al., 2017) or Reactome pathways (Gillespie et al., 2022) for reaction schemas and BRENDA database (Schomburg et al., 2017) for reaction rate parameters. Using transcriptomic and/or proteomic data our computational pipeline adds cell-type selectivity to canonical dynamical models by adjusting steady-state or time-dependent concentrations of enzymes and other proteins to experimentally observed levels.

In more detail, our computational script converts the canonical model into cell-type selective models by first removing enzymes that are not expressed and all reactions that as a consequence lost connection to precursor metabolites because of interrupted substrate flow. In the case of transcriptomic data, our pipeline automatically adds translation and protein degradation reactions to each proteoform in each compartment that can be linked to experimentally determined mRNA levels. Canonical models can be updated based on new knowledge, and our pipeline will generate updated cell-type selective models as well. The individualization of dynamical models from cell-type selective omic datasets has been implemented by other authors as well, studying drug effects on the survival of cancer cell lines (Frohlich et al., 2018). Currently, our algorithm allows compartmentalization of the cell and is capable of predicting metabolite profiles in addition to protein states. After generation of cell-type selective models, our script writes functional MATAB code for each cell type, allowing simulation of cell-type selective responses using standard ODE solvers. Our algorithm can be readily modified to write code for modeling software such as Octave, or Python ODE solvers.

To test our algorithm, we selected arachidonic acid (AA) metabolism that is operative in many cell types and organs. The metabolites generated by this network are important signaling mediators with physiological effects on kidney, uterus and blood vessels as well as other organ systems. Due to the availability of proteomic, transcriptomic and metabolomic datasets from the same experiments, we selected a macrophage cell line, bone-marrow derived macrophages (BMDM), to develop the model and assess its predictive capability.

Our canonical model (Figure 4) focused on the synthesis of the major derivatives of AA, i.e., prostaglandins, prostacyclins, thromboxane, leukotrienes and the products of 12- and 15-lipoxygenases (Wang et al., 2021). AA is generated from intracellular membranes by cytosolic phospholipase A2 that is recruited to the site of action by an intracellular calcium peak induced by macrophage activation (Leslie, 2015). Canonical reaction parameters were curated from the literature, if available (PENTACON, 2023). To generate a cell-type-selective dynamic model, we used freely available transcriptomic, proteomic and lipidomic datasets generated from BMDM. The proteomic data described protein expression values in unstimulated BMDMs (Qie et al., 2022) and was used to determine protein expression values at baseline. The published transcriptomic and lipidomic data was generated after BMDM activation by sequential stimulation with Lipid A, an LPS analogue and ATP (Kihara et al., 2014). Both ligands work through cell surface receptors. We used the transcriptomic data to predict how the enzyme expression levels obtained from the proteomic data change in response to macrophage activation. After individualization of the canonical model our script wrote the related MATLAB code that allowed simulation of metabolite profiles after macrophage activation (Figure 5).

The researchers who generated the transcriptomic and lipidomic datasets also published a dynamical model of arachidonic acid metabolism that predicts experimental lipid profiles with high accuracy and showed functional coupling between cyclooxygenases and the terminal synthases (Kihara et al., 2014). We outline the major differences between their and our approaches. These researchers a) simulated reactions using flux dynamics, where fluxes depend on enzyme-specific rate parameters as well as time-dependent enzyme and substrate concentrations. Our equations are based explicitly on Michaelis Menton Kinetics. b) They assumed enzyme protein concentrations follow gene expression values with a delay of 4 h. We use mRNA translation and protein degradation rates to simulate changes of baseline enzyme expression that we predicted from proteomics data. In our model using translation and degradation rates protein expression profiles follow the gene expression profiles with only a short delay. c) The original study focused on the reactions downstream of AA and use the experimental AA time course as a given input for their reactions. Our model includes simulation of AA production and recycling. d) our model contains multiple subcellular compartments, i.e., cytoplasm, endoplasmic reticulum/nuclear membranes, and the Golgi apparatus whose sizes are determined from experimental data. Inclusion of multiple different

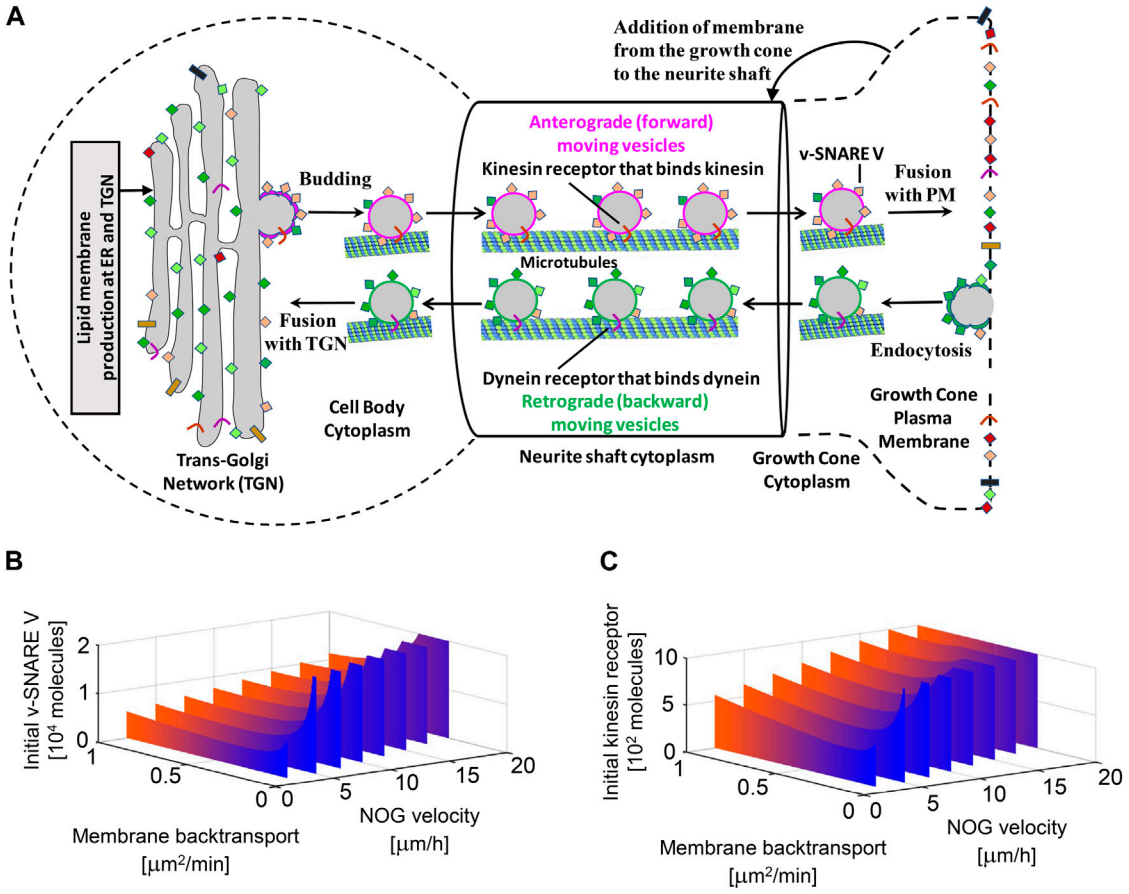


FIGURE 3
Multicompartment ODE model showing the necessity for vesicle recycling (i.e., membrane back transport) for neurite growth. Recycling is required to maintain the dynamic concentrations of the motor protein kinesin and the fusion protein v-SNARE for the whole cell response. TGN, Trans-Golgi Network; PM, Plasma Membrane. Reproduced from Hansen et al. (2022). For details see Hansen et al. (2022) and Yadaw et al. (2019).

compartments allows consideration of different intracellular localizations of downstream enzymes (Yuan and Smith, 2015; Calder, 2020), simulation of enzyme membrane recruitments triggered by the calcium peak (Leslie, 2015) as well as vesicular enzyme trafficking (Yuan and Smith, 2015). These realistic details allow for better specification of cell type identity. Overall, our automated algorithm works well (Figure 5). Generally, if initial simulations are substantially different from experimental observations, the model can be revised to add additional cell biological details such as post-translational regulation or additional subcellular compartments. Such variations on a canonical theme model provide a feasible approach to model cell type selective metabolic changes and can be readily adapted to single cell transcriptomic data.

Dynamical models from single cell transcriptomic data—use of ML-AI approaches

The rapid advances in transcriptomics at the single cell has greatly enhanced our understanding of tissue and organ function. Single cell transcriptomics not only allows us to document the abundances of the different cell types and subtypes in an organ, but also to estimate their capacities for physiological functions. Further, in disease states, single cell transcriptomic measurements enable us to identify infiltrating immune cells and the mechanisms by which they control inflammation and organ responses that can drive disease initiation and progression. Developing accurate computational models of physiological dynamics at the single cell level will be a necessary first step in creating digital twins to understand how organ function changes in disease states. Once an ML or AI algorithm is trained on a particular model, its use can significantly decrease the time needed for simulation with a previously untested sets of expression levels, without loss of quality of the predictions (Nilsson et al., 2022). Such models can also be used to understand the molecular and cellular basis of organ robustness, wherein the organ remains resilient to damage from different types of perturbation including external insults. ML and AI algorithms can also be trained to generate predictions in the opposite direction, i.e., to predict the underlying expression levels from the observed output of the dynamical system. ML and AI algorithms could also help to identify suited drug combinations that generate the desired effect in one cell type, while avoiding the unwanted side effect in another cell type.

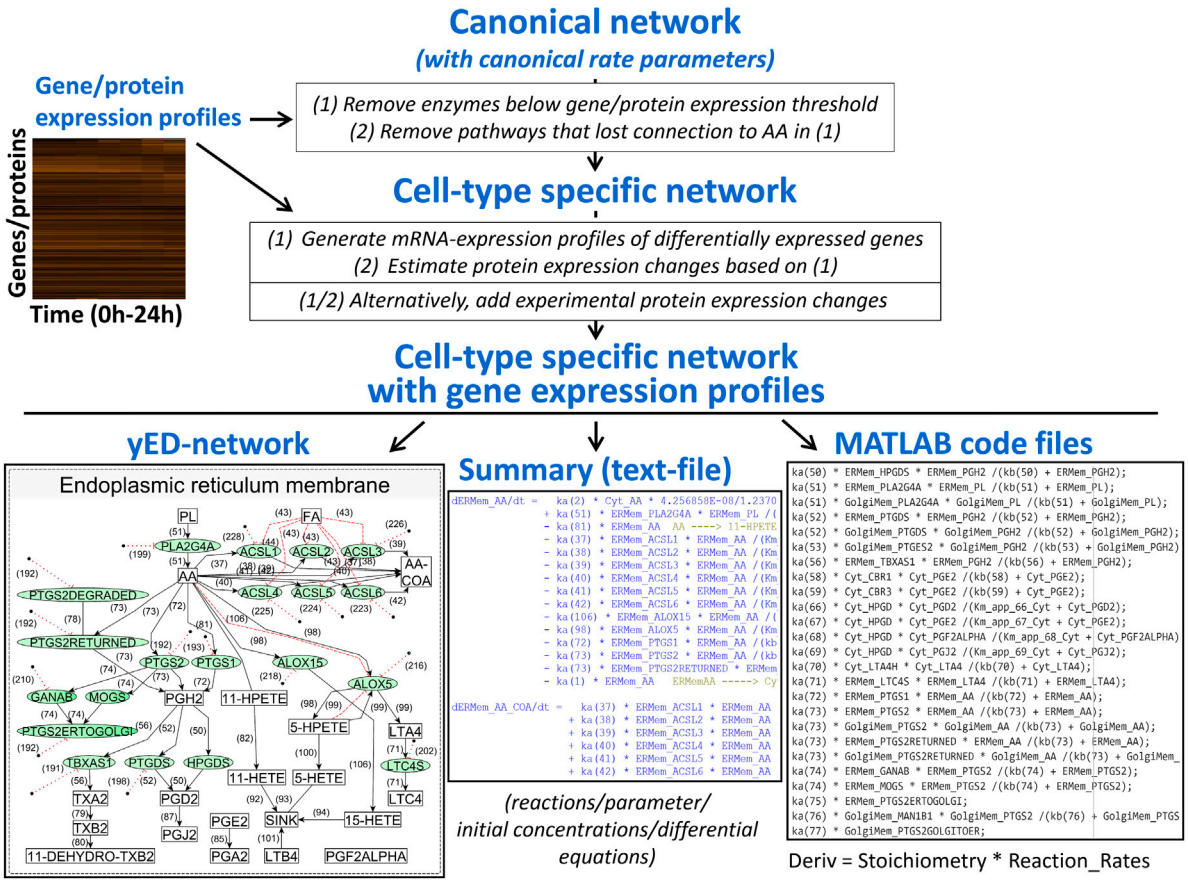


FIGURE 4
An algorithm that integrates canonical networks of subcellular processes with gene expression profiles to produce cell type specific networks and systems of reactions that can be used for dynamical modeling. AA, arachidonic acid.

Advances in hardware technologies including the development of increasingly fast GPU processers have made the running of thousands to millions of models both cheap and fast. Commercial software such as MATLAB or freeware such as Octave offers programs that that can be used for such simulations. The barriers to using these technologies are mostly at the biological level. The overall biological knowledge of the system being simulated should be utilized to constrain the development of the large-scale simulations with flexibility. Such an approach would prevent simulation of the proverbial spherical cow, but at the same time allow detection of black swans - rare variations in whole cell functions with high impact on physiology.

To fully utilize the knowledge from single cell transcriptomic data, a systematic approach to build organ level dynamical models from single cell transcriptomic data starts with building reasonable models for each cell type and each cell assigned to a cell type (Figure 6). Single cell transcriptomic data indicate that different components of a pathway are expressed at varying levels in individual cells. Model simulations can generate outputs for all observed expression profiles. Additional synthetic training data can be generated by introduction of random variations in enzyme concentrations that lie within biologically reasonable constraints. If the model contains equations describing drug actions, their concentration can be varied in the synthetic and experimental training data using the same rules. Overall, such an approach could allow the generation of thousands or even millions of different models, each of which will link its own enzyme and drug profile to its simulated molecular response profile. Training of ML and AI algorithms on all profiles can unveil relationship patterns between individual molecules or groups of molecules across the three different profiles.

Machine learning approaches are already used as an alternative to classical dynamical modeling for signaling pathways from receptors to transcription factors (Nilsson et al., 2022). Using a genome-scale artificial neural network and synthetic data based on canonical pathways and parameters the model predicted with reasonable accuracy the relationship between ligand receptor interactions and transcription factor activation in macrophages as assessed by transcriptomics.

In other fields that use numerical simulation extensively, neural networks have been successfully used to develop models and make reasonably accurate predictions. Adaptation of graph neural networks that use a “encode-process-decode” approach as described by the authors has been used to develop accurate medium range weather predictions (Lam et al., 2023). This machine learning approach uses network framework where system states (e.g., reactant identity, reactant concentrations) are

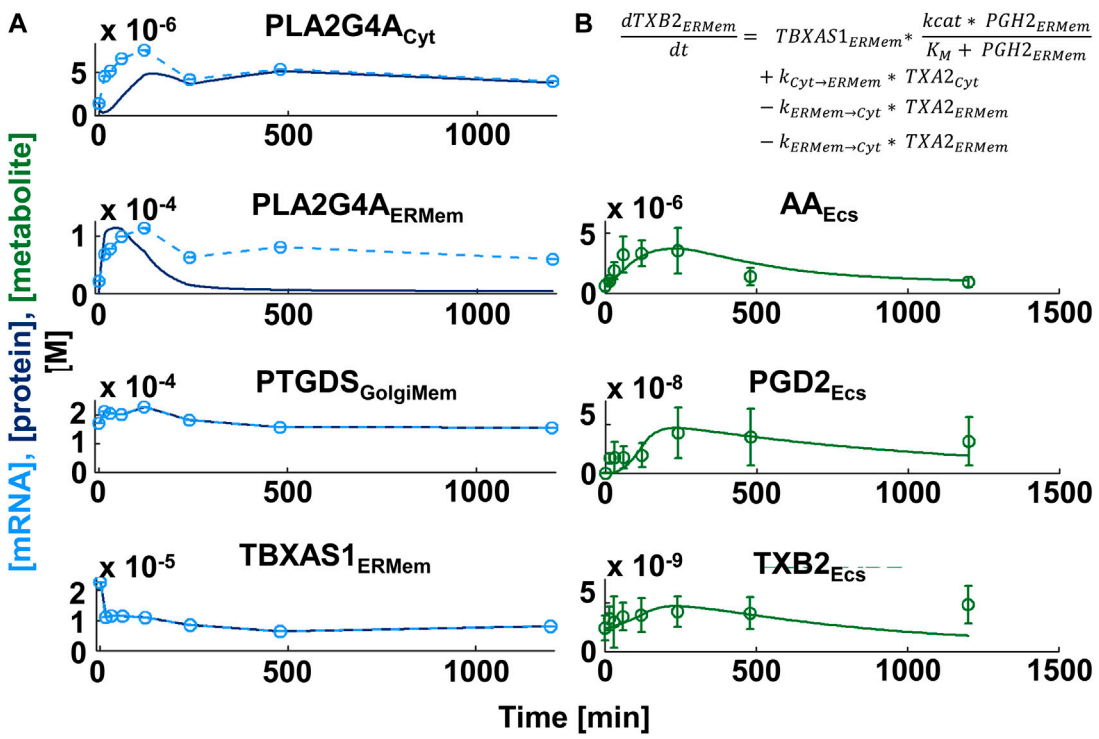


FIGURE 5
Comparison of simulation and experimental data for production of lipid messengers in macrophages. **(A)** Gene expression profiles (light blue circles in left figure column, n ≥ 3) induced by sequential treatment of Bone-marrow-derived Macrophages with the LPS analogue Lipid A and ATP (Kihara et al., 2014) were mapped to the canonical network of Arachidonic Acid Metabolism and subjected to spline interpolation (light blue dashed lines). Assuming high turnover rates, protein expression time series (dark blue lines) were predicted from mRNA profiles. To allow direct comparison we adjusted the mRNA profile values to lie within the same range as the protein concentrations. ATP stimulation generates a cytoplasmic calcium burst that triggers translocation of multiple eicosanoid enzymes, including cytoplasmic phospholipase A2 (PLA2G4A) from the cytoplasm (Cyt) to intracellular membranes, e.g., the endoplasmic reticulum or Golgi membranes (ERMem, GolgiMem, respectively). Simulated concentrations of the lipid messengers (PGD2 and TXB2 - green lines in right column) agree with lipid messengers measured in the extracellular space (Ecs) (culture medium) in the same experiment (green circles and standard deviations). AA: Arachidonic Acid, PGD2: Prostaglandin D2, TXB2: Thromboxane B2. PTGDS: Prostaglandin D2 synthase, TBXA1: Thromboxane A1 synthase 1. **(B)** Enzyme kinetics in our model are simulated by Michaelis Menton kinetics.

represented as nodes and dynamics are approximated by message-passing between these nodes. Such systems do not require explicit formulation of the system in terms of differential equations (Sanches-Gonalez et al., 2020), nevertheless are able to learn and produce complex simulations with mesh-based systems (Sanchez-Gonzalez and Battaglia, 2021). Although we have not yet seen the use of such graph neural systems-based models for dynamics from single cell transcriptomics data, it is likely that such simulations will be useful to extract deep knowledge as we accumulate spatial transcriptomic data at the single cell level.

Cell to tissue models and disease states–integrating with clinical and pathology phenotypes

Cell models as cores of digital twins presume a middle-out format. This format uses a cell centric approach in going from genes to organ level physiological functions. The components (mostly proteins) of pathways and functional units within cells can be connected to genes and their genomic and epigenetic determinants at one end and organ physiology and organismal phenotypes at the other end. Changes in cellular components in different physiological and pathophysiological states are experimentally identified from omics analyses. To make these connections in an explicit manner so distant functional relationships are not only computable but also findable at every scale of organization and traceable across scales we need knowledge graphs that connect components and features both within and between knowledge domains.

An example of framework that connects physiological and pathophysiological characteristics (phenotypes) to genomics at an individual level is the Global Alliance for Genomics and Health (GA4GH) Phenopacket schema (Jacobsen et al., 2022). This schema uses ontology terms across various domains such as genomic variants, pathology, clinical measurements, and therapeutic actions to connect features from one domain to another. Developing Phenopacket-like schemas as knowledge graphs will be the next challenge to be solved to connect cell level physiology to organ phenotypes. In addition to pathways and processes within each cell type at a single cell level, such connections will have to include molecular details of cell-cell interactions and cell-matrix interaction. Technologies advances in spatial transcriptomics, metabolomics, and proteomics, at the single cell level are making it possible to identify and map spatial relationships between individual cells in a single cell type, and between different cell

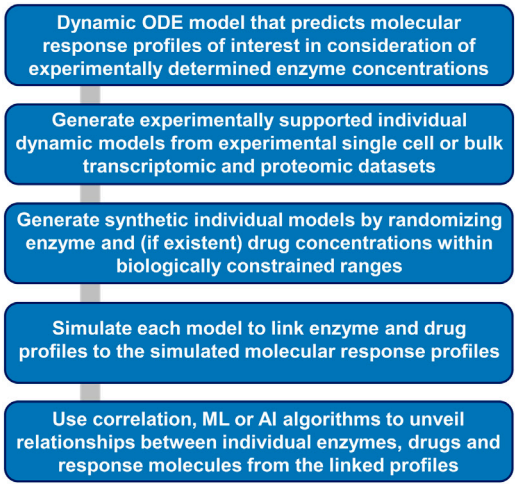


FIGURE 6
Workflow for ML and AI-based extraction of molecular relationships from simulations of dynamic models. Dynamic models that incorporate experimental or synthetic enzyme concentrations as independent variables allow generation of dependent large-scale simulated response profiles. Statistical, ML and AI algorithms can allow identification of hidden relationships between individual enzyme, drug, and response molecule concentrations.

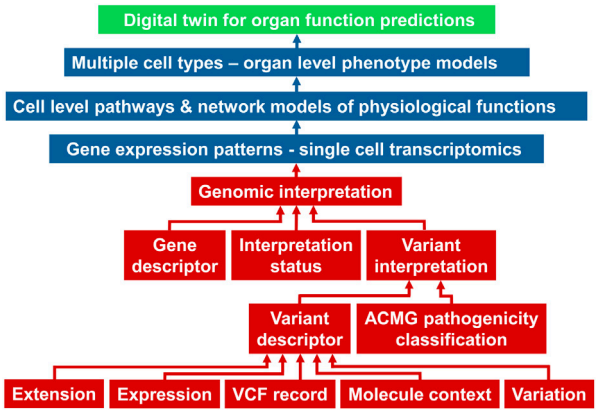


FIGURE 7
A simplified schema adopted from (Jacobsen et al., 2022) of how genomic descriptors (red boxes) within the Phenopacket schema can be connected to single cell transcriptomic data and models to develop digital twins of organ function. ACMG, American College of Medical Genetics; VCF, variant call format.

types. A simplified schematic of a digital twin for organ function prediction is shown in Figure 7. The genomic interpretation workflow is taken from the Phenopackets schema (Jacobsen et al., 2022).

Relationships between multicellular structures within an organ such as blood vessels and nephrons in kidney or blood vessels and chambers (e.g., left ventricle) in the heart will be specified in terms of molecular interactions. This spatial knowledge will have to be incorporated into functional models to accurately simulate how cell-level physiology functions enable the emergence of organ phenotypes that are clinically measured, such as estimated glomerular filtration rates for kidney and left ventricular ejection fraction for heart. At the other end of a multilayered knowledge graph, we will have to connect the transcriptomic data in different cell types and subtypes to genomic determinants such as single nucleotide polymorphisms, copy number variations and other features. We will also have to connect epigenetic determinants to transcriptomic profiles. The effects of non-coding RNAs in controlling transcription will have to be mapped to the knowledge graph to fully describe the various modes of regulation that control mRNA levels for translation.

Cell endowment is a concept that emerges from single cell transcriptomics. Cell endowment states that normal function of organ level physiological functions is dependent on the levels of key cell types. Single cell transcriptomic data sets provide information regarding the number of cells in each cell type and subtype in addition to the gene expression profiles and this information will be the basis for important parameters that connect cell physiological events to organ phenotypes. This information can be captured in the knowledge graphs as node attributes at the cell level and used in a quantitative fashion in the numerical models The ability to encode cell endowment within the graph structure is a good example of power of graphs in representing multidimensional biological systems. For such graphs to be properly constructed it is essential that the semantic frameworks within different domains are appropriately and correctly harmonized and that ontology integration is an early focus in development of digital twins.

Conclusion and perspective

Challenges in building realistic digital twins for organ function

Organ structure

The conversion of cell-level physiology into organ function is in part controlled by the spatial organization of the different cell types within the organ in the context of the extracellular matrix. Additionally, both local and global geometries in the organ will shape biophysical forces that in turn control cell-level physiology through mechanotransduction. Here, we have to account both for the contributions of the extracellular matrix to the overall biomechanical properties of the tissue and organ as well as the interactions of matrix proteins with cell membrane proteins to communicate both biomechanical and biochemical signals to the different cell types. It is likely that these properties will vary from organ to organ and even within regions of an organ. How these similarities and differences are encoded in the knowledge graphs is a challenge that needs to be addressed.

Cell biological rules

Physiological functions at the whole cell level are governed by a myriad of rules including those that specify constitutive properties. Such rules need to consider the regulation by the vast signaling networks that transduce external and intracellular signals to control effector functions, such as cytoskeletal dynamics or intracellular degradation pathways. Rules governing the relationship between mRNA and protein levels are of importance as well, when building

functional networks from single cell or bulk transcriptomic data. Rules for protein turnover and location are also important and need to be appropriately coded as node attributes. Although there is general concordance between mRNA and protein levels (Buccitelli and Selbach, 2020) this needs to be ascertained for individual proteins of interest and can be done by parameter variation exercises in dynamical models.

Not every cellular function is required for simulation of whole cell physiology that drives organ phenotype. However, for an organ function of interest, it is essential to generate rules on how to simulate the activities of relevant pathways and their functional interactions. For example, for simulating organ functions such as nutrient absorption in the intestines (Kellett et al., 2008), glucose reabsorption in the kidney proximal tubule cells (Chichger et al., 2016) or water reabsorption in the kidney principal cells (Zhao et al., 2023) it is essential that rules governing trafficking (i.e., transport and recycling) of the appropriate transporters, channels and pumps are specified for the cell types of interest. Many of these rules can be generated from the vast experimental literature in cell biology, biochemistry and physiology that have studied individual processes in depth. The rules can be encoded as edge specifications. However, in using prior knowledge, it is important to have strict guidelines in interpreting the experiments to avoid artefactual conclusions. A common example is the caution we need to exercise in extracting rules from studies that overexpress proteins of interest in exogenous systems to obtain insight into native physiological functions.

Parameters for interactions

For building dynamical models, obtainment of kinetic parameters for the reactions and concentrations of reactants has remained among the most intractable problems, although databases such as BRENDA (Schomburg et al., 2017) offer great help for this task. Since our early work on bistable switches for cell states in the late nineties (Bhalla and Iyengar, 1999) till today, 25 years later, no systematic effort to develop catalogs of quantitative parameters has been undertaken. This lack of data sets has led us to estimate and guesstimate parameters (Bhalla and Iyengar, 1999; Rangamani et al., 2011) or calculate parameter dependencies (Yadaw et al., 2019) over the years. Others have used the Hill equation approximation (Ryall et al., 2012) which provides biologically relevant simulations as assessed by experiments that test simulation predictions.

Specification of reaction rates is complicated by the fact that often post-translational modifications such as phosphorylation change reaction rates. Hence, these rates need to be specified for different states of the same proteins (proteoforms) (Melani et al., 2022). Additionally, initial concentrations of protein reactants arise from mixtures of these proteoforms and knowledge of the relative proportions of the proteoforms is very valuable in accurately specifying initial concentrations for a group of reactions. Such detailed knowledge exists for very few pathways within the mammalian cell but can be estimated from experimentally obtained overall profiles of pathways activities.

The issues regarding kinetic parameters can lead one to conclude that dynamical models are often not worth the effort. However, this is not so. Dynamical models are important because physiology is dynamics. Unless we can develop and integrate dynamical models with the growing array of informatics and statistical reasoning models we will not achieve the full predictive capability that current large datasets can enable. Artificial intelligence (AI) and machine learning (ML) algorithms that sort through vast arrays of parameter variations in a combinational manner can help. Steady state behavior of stimulated signaling networks has already been successfully modeled with high computational performance using recurrent neuronal networks that reflect network topologies and approximate protein interactions with a perturbation-specific activation function (Nilsson et al., 2022). AI and ML algorithms incorporated at the interface of transcriptomic data derived networks and their casting as dynamical models can help sort through both the rules required to specify and constrain and the parameters needed to run the simulations. Initially such integration will be by trial and error. However, as we develop large libraries of models that predict a range of organ physiological behaviors, we will be able to select well-constrained models for understanding and predicting an organ state or function of interest.

Error propagation, uncertainty and accuracy of predictions

The advances in data gathering and enormous growth in computing capability have brought us to the cusp of building accurate computational representations of many organ systems in our body. Integration of the different modeling approaches will ensure that we do not produce spherical cows, rather multiscale models with zoom-in zoom out capabilities where macroscopic functions of the whole organs can be understood and predicted from genomic characteristics underlying molecular and cellular properties. While at 30,000 feet view the ability to develop digital twins that predict organ behavior from genomic information based on mechanistic functions at the molecular and cell level appear achievable given the vast amounts of data in different domains cheap high-performance computing and current advance in machine learning and artificial intelligence algorithms, the picture at the ground level is considerably more complex. There are multiple levels of uncertainty that can lead to propagation of errors resulting in diminishing the accuracy of predictions. At a minimum there are many types of uncertainty 1) within a data domain there can be uncertainty regarding node size and attributes 2) within molecular interaction domains uncertainty regarding the existence of edge and edge strength 3) uncertainty in connections between edges and potential interdomain edges being affected by distal domains. 4) errors in computations arising from methods of simulations, such as errors due to large time steps in ODE models. There is a need to develop methods to quantify each of these uncertainties and error generating steps and develop an overall numerical score that reflects the reliability and accuracy of prediction. It is likely this will be a separate sub-field in the development of digital twins for organs.

It is commonly understood that each individual is different from others, but nevertheless belongs to groups or categories of physiological functions such that disease states in these groups can be treated with similar therapeutic approaches. It is also commonly observed in clinical practice that some individuals within a therapeutically defined group need to have a personalized therapeutic strategy that is optimal to control their pathophysiology. Currently this is done empirically by trial and error. As accurate digital twins are developed, we should be able to predict the clinical responses of these individuals for optimal therapeutic benefits.

Code availability statement

Modeling code for the reproduction of the simulations shown in Figure 5 can be found at 'github.com/SBCNY/Integration_transcriptomics_dynamicModels' in the folder 'Arachidonic_acid_metabolism'.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Kihara et al. (2014) and Qui et al. (2022).

Author contributions

JH and RI contributed equally to conceptual development of integrated dynamical modeling approaches. JH and AJ developed the directed graphs, dynamical models and ran the simulations shown here. PN is working on approaches to integrate ML approaches to the dynamical models shown here. PN and PR are focused on development of integrated ontologies for multiscale models. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Funding by NIH Grants GM 137056 for development for integrated modeling approaches and model curation was supported by U54HL117798 to Garret Fitzgerald.

Acknowledgments

We thank Dr. Shankar Subramaniam for providing detailed tables of the data from Kihara et al. (2014). We thank Dr. William Smith (University of Michigan), Dr. Robert Murphy (University of Colorado) and Kara Dolinsky (Princeton University) for guidance during the curation of the eicosanoid network.

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

Bhalla, U. S., and Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science* 283, 381–387. doi:10.1126/science.283.5400.381

Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* 21, 630–644. doi:10.1038/s41576-020-0258-4

Calder, P. C. (2020). Eicosanoids. *Essays Biochem.* 64, 423–441. doi:10.1042/EBC20190083

Castro, L. R., Gervasi, N., Guiot, E., Cavellini, L., Nikolaev, V. O., Paupardin-Tritsch, D., et al. (2010). Type 4 phosphodiesterase plays different integrating roles in different cellular domains in pyramidal cortical neurons. *J. Neurosci.* 30, 6143–6151. doi:10.1523/JNEUROSCI.5851-09.2010

Chichger, H., Cleasby, M. E., Srai, S. K., Unwin, R. J., Debnam, E. S., and Marks, J. (2016). Experimental type II diabetes and related models of impaired glucose metabolism differentially regulate glucose transporters at the proximal tubule brush border membrane. *Exp. Physiol.* 101, 731–742. doi:10.1113/EP085670

Ding, J., Sharon, N., and Bar-Joseph, Z. (2022). Temporal modelling using single-cell transcriptomics. *Nat. Rev. Genet.* 23, 355–368. doi:10.1038/s41576-021-00444-7

Frohlich, F., Kessler, T., Weindl, D., Shadrin, A., Schmiester, L., Hache, H., et al. (2018). Efficient Parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Syst.* 7, 567–579. doi:10.1016/j.cels.2018.10.013

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., et al. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* 50, D687–D692. doi:10.1093/nar/gkab1028

Hansen, J., Meretzky, D., Woldesenbet, S., Stolovitzky, G., and Iyengar, R. (2017). A flexible ontology for inference of emergent whole cell function from relationships between subcellular processes. *Sci. Rep.* 7, 17689. doi:10.1038/s41598-017-16627-4

Hansen, J., Siddiq, M. M., Yadaw, A. S., Tolentino, R. E., Rabinovich, V., Jayaraman, G., et al. (2022). Whole cell response to receptor stimulation involves many deep and distributed subcellular biochemical processes. *J. Biol. Chem.* 298, 102325. doi:10.1016/j.jbc.2022.102325

Jacobsen, J. O. B., Baudis, M., Baynam, G. S., Beckmann, J. S., Beltran, S., Buske, O. J., et al. (2022). The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol.* 40, 817–820. doi:10.1038/s41587-022-01357-4

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi:10.1093/nar/gkw1092

Kellett, G. L., Brot-Laroche, E., Mace, O. J., and Leturque, A. (2008). Sugar absorption in the intestine: the role of GLUT2. *Annu. Rev. Nutr.* 28, 35–54. doi:10.1146/annurev.nutr.28.061807.155518

Kihara, Y., Gupta, S., Maurya, M. R., Armando, A., Shah, I., Quehenberger, O., et al. (2014). Modeling of eicosanoid fluxes reveals functional coupling between cyclooxygenases and terminal synthases. *Biophys. J.* 106, 966–975. doi:10.1016/j.bpj.2014.01.015

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. *Science* 382, 1416–1421. doi:10.1126/science.adi2336

Lee, L. L., and Chintalgattu, V. (2019). Pericytes in the heart. *Adv. Exp. Med. Biol.* 1122, 187–210. doi:10.1007/978-3-030-11093-2_11

Leslie, C. C. (2015). Cytosolic phospholipase A2: physiological function and role in disease. *J. Lipid Res.* 56, 1386–1402. doi:10.1194/jlr.R057588

Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., et al. (2005). Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science* 309, 1078–1083. doi:10.1126/science.1108876

Melani, R. D., Gerbasi, V. R., Anderson, L. C., Sikora, J. W., Toby, T. K., Hutton, J. E., et al. (2022). The Blood Proteoform Atlas: a reference map of proteoforms in human hematopoietic cells. *Science* 375, 411–418. doi:10.1126/science.aaz5284

Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* 38, D750–D753. doi:10.1093/nar/gkp889

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* 298, 824–827. doi:10.1126/science.298.5594.824

Neves, S. R., Tsokas, P., Sarkar, A., Grace, E. A., Rangamani, P., Taubenfeld, S. M., et al. (2008). Cell shape and negative links in regulatory motifs together control spatial information flow in signaling networks. *Cell* 133, 666–680. doi:10.1016/j.cell.2008.04.025

Nilsson, A., Peters, J. M., Meimetis, N., Bryson, B., and Lauffenburger, D. A. (2022). Artificial neural networks enable genome-scale simulations of intracellular signaling. *Nat. Commun.* 13, 3069. doi:10.1038/s41467-022-30684-y

PENTACON (2023). The personalized NSAID therapeutics consortium. Available at: https://pentaconhq.org/.

Qie, J., Liu, Y., Wang, Y., Zhang, F., Qin, Z., Tian, S., et al. (2022). Integrated proteomic and transcriptomic landscape of macrophages in mouse tissues. *Nat. Commun.* 13, 7389. doi:10.1038/s41467-022-35095-7

Rangamani, P., Fardin, M. A., Xiong, Y., Lipshtat, A., Rossier, O., Sheetz, M. P., et al. (2011). Signaling network triggers and membrane physical properties control the actin cytoskeleton-driven isotropic phase of cell spreading. *Biophys. J.* 100, 845–857. doi:10.1016/j.bpj.2010.12.3732

Rubin, D. M., Letts, R. F. R., and Richards, X. L. (2019). Teaching physiology within a system dynamics framework. *Adv. Physiol. Educ.* 43, 435–440. doi:10.1152/advan.00198.2018

Ryall, K. A., Holland, D. O., Delaney, K. A., Kraeutler, M. J., Parker, A. J., and Saucerman, J. J. (2012). Network reconstruction and systems analysis of cardiac myocyte hypertrophy signaling. *J. Biol. Chem.* 287, 42259–42268. doi:10.1074/jbc.M112.382937

Sanches-Gonalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., and Battaglia, P. W. (2020) *Learning to simulate complex physics with graph networks*. arXiv.

Sanchez-Gonzalez, A., and Battaglia, P. W. (2021) *Learning mesh-based simulation with graph networks*. arXiv.

Schomburg, I., Jeske, L., Ulbrich, M., Placzek, S., Chang, A., and Schomburg, D. (2017). The BRENDA enzyme information system-From a database to an expert system. *J. Biotechnol.* 261, 194–206. doi:10.1016/j.jbiotec.2017.04.020

Shim, J. V., Xiong, Y., Dhanan, P., Dariolli, R., Azeloglu, E. U., Hu, B., et al. (2023). Predicting individual-specific cardiotoxicity responses induced by tyrosine kinase inhibitors. *Front. Pharmacol.* 14, 1158222. doi:10.3389/fphar.2023.1158222

Sturtzel, C. (2017). Endothelial cells. *Adv. Exp. Med. Biol.* 1003, 71–91. doi:10.1007/978-3-319-57613-8_4

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Tanaka, K., and Augustine, G. J. (2008). A positive feedback signal transduction loop determines timing of cerebellar long-term depression. *Neuron* 59, 608–620. doi:10.1016/j.neuron.2008.06.026

Wang, B., Wu, L., Chen, J., Dong, L., Chen, C., Wen, Z., et al. (2021). Metabolism pathways of arachidonic acids: mechanisms and potential therapeutic targets. *Signal Transduct. Target Ther.* 6, 94. doi:10.1038/s41392-020-00443-w

Yadaw, A. S., Siddiq, M. M., Rabinovich, V., Tolentino, R., Hansen, J., and Iyengar, R. (2019). Dynamic balance between vesicle transport and microtubule growth enables neurite outgrowth. *PLoS Comput. Biol.* 15, e1006877. doi:10.1371/journal.pcbi.1006877

Yuan, B., Shen, C., Luna, A., Korkut, A., Marks, D. S., Ingraham, J., et al. (2021). CellBox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell Syst.* 12, 128–140 e4. doi:10.1016/j.cels.2020.11.013

Yuan, C., and Smith, W. L. (2015). A cyclooxygenase-2-dependent prostaglandin E2 biosynthetic system in the Golgi apparatus. *J. Biol. Chem.* 290, 5606–5620. doi:10.1074/jbc.M114.632463

Zhang, Y., Wang, H., Oliveira, R. H. M., Zhao, C., and Popel, A. S. (2022). Systems biology of angiogenesis signaling: computational models and omics. *WIREs Mech. Dis.* 14, e1550. doi:10.1002/wsbm.1550

Zhao, C., Medeiros, T. X., Sove, R. J., Annex, B. H., and Popel, A. S. (2021). A data-driven computational model enables integrative and mechanistic characterization of dynamic macrophage polarization. *iScience* 24, 102112. doi:10.1016/j.isci.2021.102112

Zhao, C., and Popel, A. S. (2021). Protocol for simulating macrophage signal transduction and phenotype polarization using a large-scale mechanistic computational model. *Star. Protoc.* 2, 100739. doi:10.1016/j.xpro.2021.100739

Zhao, X., Liang, B., Li, C., and Wang, W. (2023). Expression regulation and trafficking of aquaporins. *Adv. Exp. Med. Biol.* 1398, 39–51. doi:10.1007/978-981-19-7415-1_3

frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 30 July 2024
DOI 10.3389/fcell.2024.1310265

OPEN ACCESS

EDITED BY
Susan Mertins,
Leidos Biomedical Research, Inc., United States

REVIEWED BY
Romas Baronas,
Vilnius University, Lithuania
Michael Blinov,
UConn Health, United States

*CORRESPONDENCE
Mehdi Sadeghi,
sadeghi@nigeb.ac.ir

†These authors have contributed equally to this work

RECEIVED 09 October 2023
ACCEPTED 02 July 2024
PUBLISHED 30 July 2024

CITATION
Eidi Z, Khorasani N and Sadeghi M (2024), Correspondence between multiple signaling and developmental cellular patterns: a computational perspective.
Front. Cell Dev. Biol. 12:1310265.
doi: 10.3389/fcell.2024.1310265

Correspondence between multiple signaling and developmental cellular patterns: a computational perspective

Zahra Eidi[1†], Najme Khorasani[1†] and Mehdi Sadeghi[2*]

[1]School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran, [2]National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran

The spatial arrangement of variant phenotypes during stem cell division plays a crucial role in the self-organization of cell tissues. The patterns observed in these cellular assemblies, where multiple phenotypes vie for space and resources, are largely influenced by a mixture of different diffusible chemical signals. This complex process is carried out within a chronological framework of interplaying intracellular and intercellular events. This includes receiving external stimulants, whether secreted by other individuals or provided by the environment, interpreting these environmental signals, and incorporating the information to designate cell fate. Here, given two distinct signaling patterns generated by Turing systems, we investigated the spatial distribution of differentiating cells that use these signals as external cues for modifying the production rates. By proposing a computational map, we show that there is a correspondence between the multiple signaling and developmental cellular patterns. In other words, the model provides an appropriate prediction for the final structure of the differentiated cells in a multi-signal, multi-cell environment. Conversely, when a final snapshot of cellular patterns is given, our algorithm can partially identify the signaling patterns that influenced the formation of the cellular structure, provided that the governing dynamic of the signaling patterns is already known.

KEYWORDS

developmental pattern, signaling, cell tissue, self-organization, regenerative therapy, Turing dynamics

1 Introduction

The duality of variety and organization is among the canonical concerns in biology. During the course of development in multicellular organisms, although successive cell divisions lead to the creation of diverse cells, it does not result in colony-like accumulation of piled-up cells. Although, in principle, the genetic material of every single cell of an organism is the same, influenced by variant stimulants, they are capable of generating highly complex spatial patterns (Liu and Warmflash, 2021; Dubrulle et al., 2015; Heemskerk et al., 2019; and van Boxtel et al., 2015). A diverse range of chemical stimuli, as underlying drivers of non-genetic variations, act at multiple scales (Shahbazi et al., 2019). These stimuli play a crucial role in directing cell fate determination in stem cells at the individual cell level (Britton et al., 2021). On the other hand, collective processes such as tissue homeostasis, wound healing, angiogenesis, and tumorigenesis are intimately linked with competing environmental chemical cues (Schweisguth and Corson, 2019). Understanding the

mechanisms underlying the generation and maintenance of these ordered spatial assemblies could potentially aid in the development of novel strategies for controlling tissue organization and function *in vitro* and *in vivo*.During the development of multicellular organisms, tissues are created through the spatial arrangement of differentiated cells. Although modeling the formation of a spatial arrangement from a single stem cell is complex, it becomes even more complicated in reality as tissues are formed from the spatial arrangement of cells from different stem cells. This process requires intercellular signal transmission, which affects gene expression regulation and intracellular decision-making.Internal mechanisms are responsible for generating the right proportion of different types of specialized cells, distributing them in their right position, and maintaining the organized structure in the presence of intercellular chemical signaling agents (Khorasani and Sadeghi, 2022). Cells also sense and respond to mechanical stimuli and the physical properties of their environment via induced downstream genetic regulatory networks (Valet et al., 2022; Lenne et al., 2021; Wagh et al., 2021). Several multi-stable regulatory networks play their role as the internal decision-makers of dividing cells (Khorasani and Sadeghi, 2022). This study investigates the impact of various chemical signals on the mechanism by which multiple stem cells generate intricate tissue structures and tries to provide a deeper understanding of the mechanisms behind morphological variations. In reality, the formation of intermediate structures during embryo development or the formation of a tissue consisting of cells with different phenotypes and with organization in their spatial arrangement without a previous template is a complex problem, and modeling them using the simplest possible assumptions can lead to a better comprehension of the development process in multicellular organisms.We would like to answer these questions, or, more realistically, get any enlightenment about the following: first, in the presence of variant positional cues, how can spatially organized populations give rise to and maintain large-scale inhomogeneities starting from an initially roughly homogeneous mass of intermixed stem cell populations? Second, how do individual stem cells perceive and interpret their surrounding spatial information to make decisions about their developmental pathway in response to the local concentration of these stimulants? Finally, is it possible to infer information about the specific form of the signals that created them from the final structure of cell populations?

The basis of cellular pattern formation is mounted on the interaction of the mediating nonlinear diffusive signaling components (Murray, 2001). For the spontaneous construction of patterns during development, as proposed by Turing's classic theory, the system requires two diffusive chemical compounds: an activator compound and an inhibitor compound (Turing, 1990). The latter locally undergoes an autocatalytic reaction to generate more of itself and also activates the formation of the inhibitor compound in some way. Meanwhile, the former inhibits the formation of more activator compounds. The key element for obtaining spatial patterns is that the activator and the inhibitor components diffuse through the reaction medium at different rates. Thus, the effective ranges of their respective influences are different. Accordingly, if the inhibitor agent diffuses faster than the activator one, a stable pattern can emerge from a homogeneous background merely by the amplification of small perturbations. The patterns generally take the form of spots (and reverse spots) or stripes based on the choice of model parameters (Murray, 2001). The dynamic elaborates different possible pattern formation processes in a variety of developmental situations. The related examples span from the regeneration of hydras (Meinhardt, 2003) to animal coating patterns (Koch and Meinhardt, 1994). Wave phenomena can also generate patterns of spatiotemporal type (Cotterell et al., 2015; Eidi et al., 2021). Since the typical characteristic time of cell division is higher than that of a traveling wave, here, we exclude the formation of cellular patterns induced by spatiotemporal signaling patterns. Recently, Marcon et al. (2016) proposed a new development in classical Turing models, indicating that the essential prerequisite of varied diffusion rates for mobile signaling molecules is not essential for pattern formation. Remarkably, specific networks are capable of creating patterns using signals without the constraint of relative diffusion rates.

Here, we assume that there are two multipotent stem cells as resources of variation generation, each of which is potentially capable of constructing its own organized structure in the absence of the other. Although the cells do not directly interact, they have an intracellular signal-dependent tri-stable switch that affects their reproduction rates in response to multiple signals in the environment. We present a computational model for their internal mechanism in the presence of each other to form an organized population consisting of whole descendants. We see that signaling messengers play a significant and irreplaceable role as regulatory agents in communication between different cell types. Our results indicate that the association of variant environmental signaling messengers and intracellular decision-making switches grants a diverse range of cellular patterns. Furthermore, having the ultimate arrangement of cellular organization, one can approximately indicate the signaling patterns based on which the cellular patterns have been established, provided that the prior assumption of the pattern is given.

2 Materials and methods

In this model, we consider a scenario where a plane is initially populated by two types of stem cells, SC_1 and SC_2. These stem cells can both renew themselves and divide into their corresponding differentiated cells. When they divide into specialized cells, SC_1 can give rise to either *A* or *B*, while SC_2 can give rise to either *C* or *D*; see Figure 1A. In this case, to simplify the computational process and maintain the essence of the scenario, we will disregard any intermediate stages and assume a direct division of stem cells into their offspring. The division outcomes of each cell are influenced by the amount of signaling agent that the mother cell receives (Khorasani et al., 2020; Khorasani and Sadeghi, 2022, Khorasani and Sadeghi, 2024). Here, the main idea is that in the absence of cellular displacement, competition between existing chemical signals in the environment plays the principal role in the pattern formation process at the population level. To model the underlying mechanism, we need to answer the following questions:

- What type of signal does the model refer to?
- How can a mixture of different signals impact the fate of an individual cell?

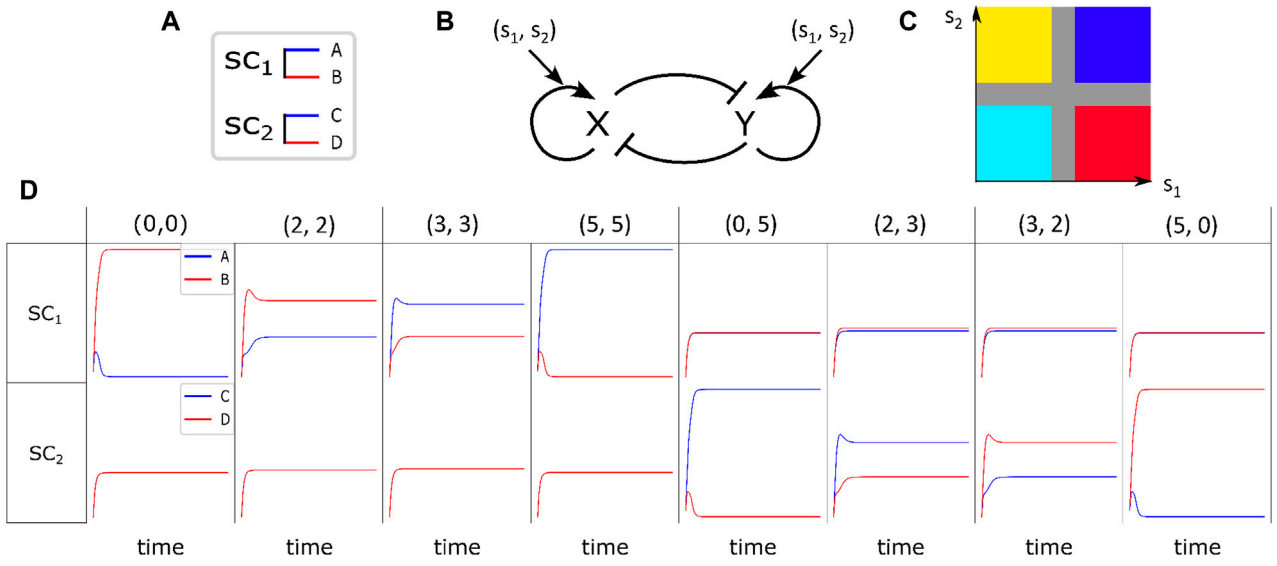


FIGURE 1
Fate determination of stem cells under the influence of environmental signaling agents. **(A)** Differentiation of SC_1 leads to either phenotype *A* or *B*, while SC_2 can differentiate into phenotype *C* or *D* depending on the amount of signaling agent exposure. **(B)** Regulation switch present in each stem cell contributes to its development toward a specific fate, which is influenced by environmental stimulant pairs (s_1, s_2). **(C)** The cruciform shape in the (s_1, s_2) plane represents the phase field of possible developed cells. Each phenotype is color-coded, with blue representing *A*, cyan representing *B*, yellow representing *C*, and red representing *D*. The values of s_1 and s_2 have been scaled up to fall within the range of (0,5). Each quadrant in the plane corresponds to a specific phenotype. The values of (s_1, s_2) directly affect the determinants within the cell and, thereby, influence the outcome of cell division, as described by Equations 3, 4. The gray cross denotes a comparable concentration of signaling agents, within which randomness plays a significant role in determining cell division outcomes. **(D)** Fate of each stem cell SC_i is influenced by a specific combination of (s_1, s_2) pairs. There are different pairs (s_1, s_2) that can affect SC_1 and leave SC_2 unaffected. For example, columns 1 to 4 can influence SC_1 and leave SC_2 neutral, while columns 5 to 8 have the opposite effect.

- What is the effect of the signals on the offspring at the population level?

The materials and methods is structured as follows: first, we introduce different possible dynamics for propagating extracellular signals in the environment, including positional information in Section 2.1.1 and reaction–diffusion dynamics in Section 2.1.2. Subsequently, in Section 2.2, we propose a regulatory switch that allows an individual cell to determine its fate influenced by the uptake of different environmental signals. Finally, in Section 2.3, we describe an algorithm for predicting the final cellular pattern of a system that is initially composed of multiple signaling agents and dividing cells.

2.1 Signals

Let us assume that the stem cells in a medium are exposed to spatial chemical information, we refer to them as signals, which are captured and interpreted by the cells to develop the spatial organization. There are various ways to provide spatial patterns in biology, among which, positional information and reaction–diffusion dynamics are the most prominent (Green and Sharpe, 2015).

2.1.1 Positional information dynamic

Generally, positional information dynamic refers to the development of the spatial cellular organization in the embryo differentiating at specific positions based on their response to the gradient of environmental signals (Schweisguth and Corson, 2019). For example, embryonic organizer centers secrete morphogens that specify the emergence of germ layers and the establishment of the body's axes during embryogenesis (De Santis et al., 2021). In the current study, by positional information, we mean any external chemical cues whose procedure of setting up is immaterial for us, and we merely focus on their impact on the regulation of internal switches. To illustrate the relationship between different signals, Figure 2 exemplifies the simultaneous presence of two signal profiles of Gaussian type (the first column), a Gaussian profile and a sinusoidal one (the second and third columns), and two sinusoidal with different frequencies (the fourth column). In each column, the final cellular pattern resulting from the process of cell division and self-renewal of competing stem cells is represented by the third row. Initially, the stem cells are randomly distributed in an environment that contains upper-row signals. In all cases, the final pattern can be distinguished by six different colors. The colors magenta and green represent Sc_1 and Sc_2, respectively. The colors blue, cyan, yellow, and red are used to represent the offspring *A*, *B*, *C*, and *D*, respectively. The pattern formation process is implemented using Algorithm 1.

2.1.2 Signaling through the reaction–diffusion dynamic

To generate two independent signaling agents in the medium, we consider a system that consists of two independent reaction–diffusion processes. Each process involves two

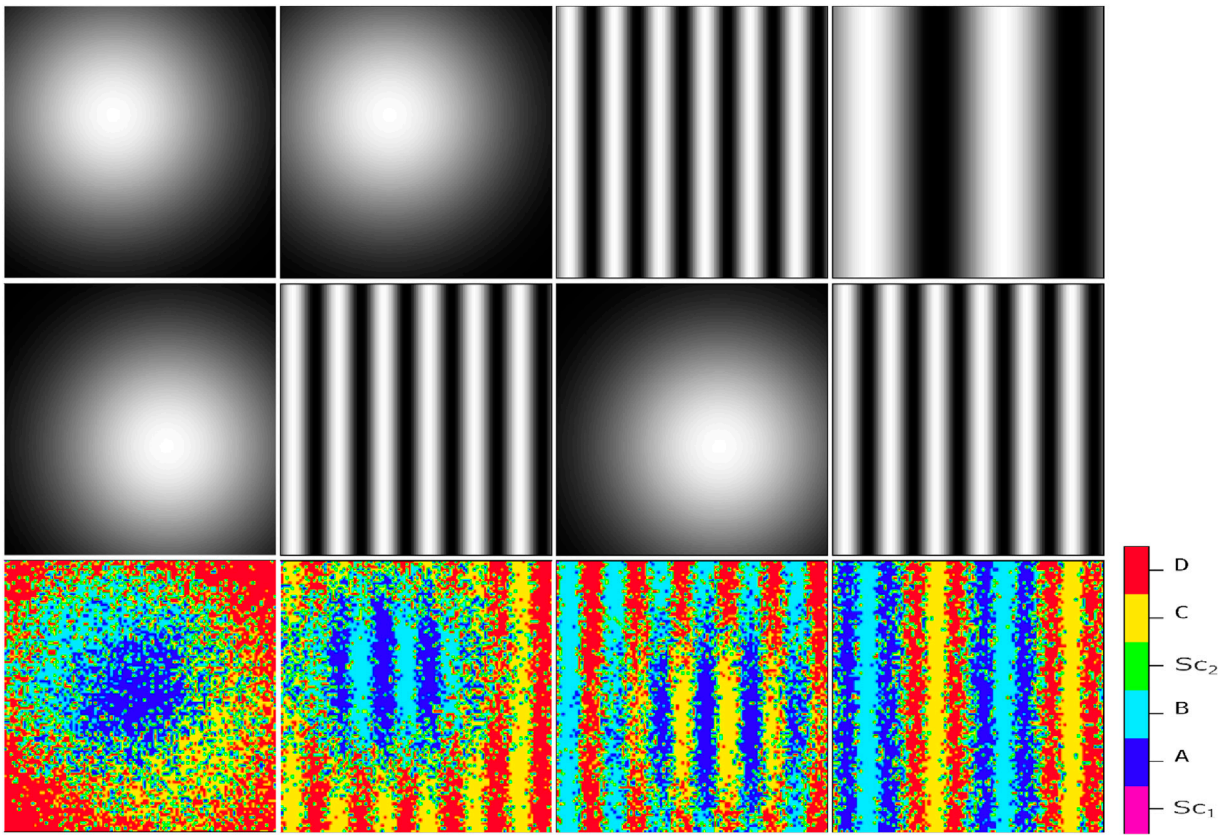


FIGURE 2
Resulting cellular pattern (third row) of a system influenced by two independent static signal profiles (first and second rows). The signal profiles consist of a Gaussian profile, described by the exponential function $\exp[-(2\sigma^2)^{-1}((x-x^*)^2+(y-y^*)^2)]$, and a sinusoidal profile $sin(kx)$. The third row in each column displays the final cellular pattern resulting from stem cell division and self-renewal. Initially, stem cells are randomly distributed in an environment containing the signal profiles from the upper row. The final pattern, distinguishable by six colors, reveals specific cell types: magenta for Sc_1, green for Sc_2, blue for offspring *A*, cyan for offspring *B*, yellow for offspring *C*, and red for offspring *D*. The outcome depends on the comparison of signal concentrations at each point. The randomness involved in the patterns belongs to the areas where the concentration of positional signals is comparable. The first column: (top) $x^* = 40$, $y^* = 30$, and $\sigma = 2$, (middle) $x^* = 60$, $y^* = 30$, and $\sigma = 2$ (bottom) the developed pattern in consequence of the combination of its upper-head signals. The second column: (top) $x^* = 40$, $y^* = 30$, and $\sigma = 2$ (middle) and $k = 4.5$ (bottom) the developed pattern in consequence of the combination of its upper-head signals. The third column (top) $k = 4.5$, (middle) $x^* = 40$, $y^* = 30$, and $\sigma = 2$ (bottom) the developed pattern in consequence of the combination of its upper-head signals. The fourth column: (top) $k = 1.5$ (middle) and $k = 4.5$ (bottom) the developed pattern in consequence of the combination of its upper-head signals.

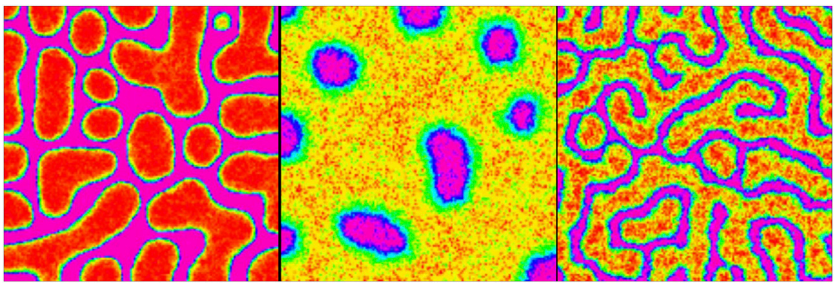


FIGURE 3
Possible signaling agent patterns s_i ($s_u^{(i)}$) following Equations 1, 2 with parameters: $\mathcal{A} = 0.9$, $\mathcal{B} = 1.2$, $\gamma = 10000$, and $\mathcal{C} = 0.2$ and lattice size $h = 0.01$. (Left) Spot pattern with $D_u = 1$ and $D_v = 20\ D_u$, (middle) reverse spot pattern with $D_u = 25$ and $D_v = 500\ D_u$, and (right) stripe pattern with $D_u = 1$ and $D_v = 20\ D_u$.

TABLE 1 Involved reactions and their corresponding propensity functions (reaction no. 1–6) generating signaling patterns and the reactions (reaction no. 7–10) involved in the production and degradation of intracellular determinants. In total, there are 20 reactions incorporated in the Gillespie algorithm, $i \in \{1, 2\}$.

Reaction no.	Reaction type	Propensity function
1	Production of $s_u^{(i)}$	$\gamma(\mathcal{A}s_u^{(i)} + \mathcal{C})$
2	Degradation of $s_v^{(i)}$	$\gamma(s_v^{(i)})$
3	Diffusion of $s_u^{(i)}$	D_u/h^2
4	Production of $s_u^{(i)}$	$\gamma(\mathcal{B}s_u^{(i)})$
5	Degradation of $s_v^{(i)}$	$\gamma(s_v^{(i)} + 1)$
6	Diffusion of $s_v^{(i)}$	D_v/h^2
7	Production of x_i	$\alpha_x^{(i)}\frac{x_i^n}{\beta^n + x^n} + k_1\frac{\beta^n}{\beta^n + y_i^n}$
8	Degradation of x_i	$\gamma_1 x_i$
9	Production of y_i	$\alpha_y^{(i)}\frac{y_i^n}{\beta^n + y_i^n} + k_2\frac{\beta^n}{\beta^n + x_i^n}$
10	Degradation of y_i	$\gamma_2 y_i$

interacting chemicals, namely, $s_u^{(i)}$ and $s_v^{(i)}$, where $i \in \{1, 2\}$. The spatial distribution of $s_u^{(i)}$ and $s_v^{(i)}$ is interdependent, as governed by their corresponding dynamics. Thus, there are two independent chemical variants produced by these two reaction–diffusion processes. We assume that the concentration field of $s_u^{(i)}$ in the medium, $i \in \{1, 2\}$, defines the dynamic profile of each independent signaling agent. Moreover, $s_u^{(i)}$ and $s_v^{(i)}$ are deemed to spread over the environment with $D_{s_u^{(i)}}$ and $D_{s_v^{(i)}}$, respectively. The governing equations for the propagation of $s_u^{(i)}$ and $s_v^{(i)}$ are as follows (Shoji et al., 2003):

$$\frac{\partial s_u^{(i)}}{\partial t} = \nabla^2 s_u^{(i)} + \gamma f\left(s_u^{(i)}, s_v^{(i)}\right), \tag{1}$$

$$\frac{\partial s_v^{(i)}}{\partial t} = d\nabla^2 s_v^{(i)} + \gamma g\left(s_u^{(i)}, s_v^{(i)}\right). \tag{2}$$

Here, by rescaling the space variable, the diffusion coefficient of $s_u^{(i)}$ and $s_v^{(i)}$ are set to 1 and d, respectively. Here, d is equal to $\frac{D_{s_v^{(i)}}}{D_{s_u^{(i)}}}$. Thus, assuming that $d \geq 1$, the diffusivity of $s_v^{(i)}$ is larger than that of $s_u^{(i)}$. In addition, $f(s_u^{(i)}, s_v^{(i)})$ and $g(s_u^{(i)}, s_v^{(i)})$ are reaction kinetics of the system represented with the following terms:

$$f\left(s_u^{(i)}, s_v^{(i)}\right) = \mathcal{A}s_u^{(i)} - s_v^{(i)} + \mathcal{C} \quad and \quad g\left(s_u^{(i)}, s_v^{(i)}\right) = \mathcal{B}s_u^{(i)} - s_v^{(i)} - 1.$$

Here, $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are the controlling parameters. The kinetics also constrains the variable $s_u^{(i)}$ within a finite range: $0 \leq s_u^{(i)} \leq s_{u_{max}}^{(i)}$. The parameter γ exhibits the relative strength of reaction kinematics. This dynamic with a reflective boundary condition can produce steady-state heterogeneous spatial patterns of chemical concentrations (Shoji et al, 2003). The diffusion process, with $d \geq 1$, in this context, is considered the main deriving process for the heterogeneity in the system. Moreover, $s_{u_{max}}^{(i)}$ is considered the controlling parameter, upon which the behavior of spatial patterns differs; see Figure 3. To simulate the dynamic, we implement the Gillespie method (Gillespie et al., 2007), which exhibits some degree of randomness in the simulation of chemical kinetics. The Gillespie algorithm is widely regarded as the "gold standard" for explaining the behavior of systems characterized by a limited number of determinants and driven by inherent fluctuations, all while avoiding the complexities of mathematical equations. This method generates a statistically possible solution of Equations 1, 2, for which the reaction rates are known. Defining the *propensity function* for every single reaction, including diffusion ones that are considered to be reducible to an analogous reaction, we have a measure to find out the time when the next chemical reaction takes place and determine which reaction is likely preferred by the system. The entire reactions of the system and their corresponding propensity functions are listed in Table 1. By updating the propensity functions at each step, one can track the changes in the corresponding cell-type population vector, which is induced by a single occurrence of a particular reaction. Repeating the algorithm simulates the whole behavior of the reaction–diffusion system stochastically. The complete algorithm implementation is detailed in Section 2.3. Before delving into that, it is crucial to explain how various chemical environmental signals impact the ultimate fate of an individual cell.

As previously mentioned, we assume that the concentration field of $s_u^{(i)}$ in the medium, where $i \in 1, 2$, represents the dynamic profile of each separate signaling agent. From this point forward, whenever we refer to s_i, we are referring to $s_u^{(i)}$.

2.2 Biased internal switch of determinants

Once we have identified the environmental signals that can influence the fate of stem cells, we can explore the subsequent question: how do simultaneous signals impact the destiny of a single cell?

Let us assume that within the cytoplasm of each stem cell SC_i $(i = 1, 2)$, there are two interacting chemical determinants, x_i and y_i, where $i \in \{1, 2\}$, whose values play a crucial role in determining the outcome of cell division. In this model, the interaction dynamics of these cytoplasmic determinants of the stem cell SC_i $(i = 1, 2)$ are controlled by a tri-stable regulatory switch. This switch controls the fate of cell division and determines whether the stem cell differentiates or self-renews. (Balázsi et al., 2011; Staff, 2017; Khorasani et al., 2020; Khorasani and Sadeghi, 2022, Khorasani and Sadeghi, (2024)). See Figure 1B.

$$\frac{\partial x_i}{\partial t} = \alpha_x^{(i)}\frac{x_i^n}{\beta^n + x^n} + k_1\frac{\beta^n}{\beta^n + y_i^n} - \gamma_1 x_i, \tag{3}$$

$$\frac{\partial y_i}{\partial t} = \alpha_y^{(i)}\frac{y_i^n}{\beta^n + y_i^n} + k_2\frac{\beta^n}{\beta^n + x_i^n} - \gamma_2 y_i. \tag{4}$$

In Figure 1B, the regulatory switch is shown. It involves mutual repression of x_i and y_i and their degradation effects, as well as their self-activation in the form of the Hill function. In the above equations, n is the Hill coefficient, β is the synthesis rate of determinants, $\alpha_x^{(i)}$ and $\alpha_y^{(i)}$ are the self-activation rates, $k_1 = k_2$ are the inhibition rates, and $\gamma_1 = \gamma_2$ are the degradation rates of x_i and y_i, respectively.

It has been demonstrated by Khorasani et al. (2020) that in the absence of stimulant signaling chemicals, when there is only one type of stem cell and the coefficients in Equations 3, 4 are constant,

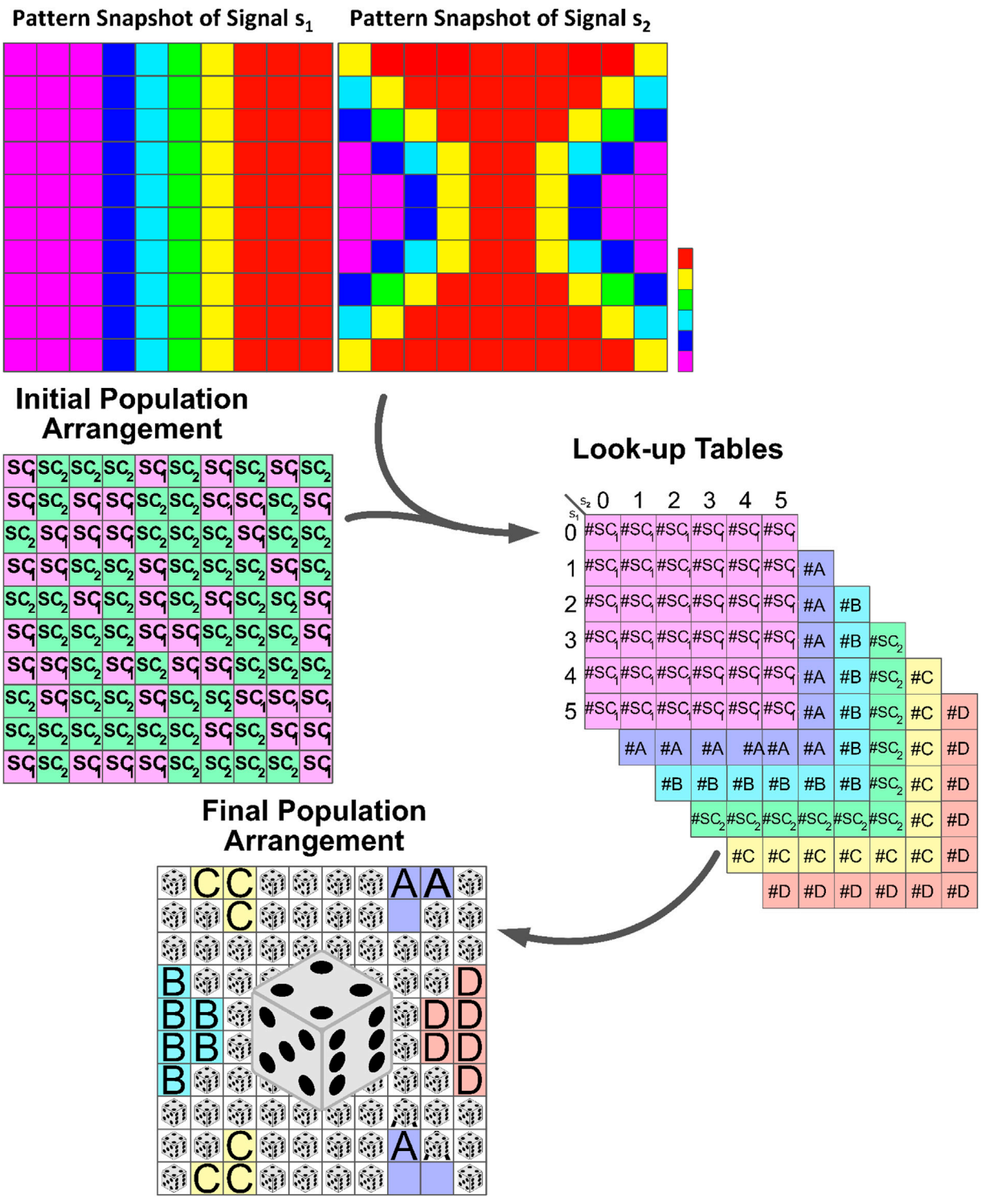


FIGURE 4
Schematic illustration of the formation of developmental patterns influenced by external signals. Stem cells respond to dynamic signals, leading to differentiation and self-renewal based on the number of determinants produced within their cytoplasm. Look-up tables summarize cell division outcomes based on the signal combination intervals. The highlighted section with dice in the lower panel represents regions with comparable signal concentrations. Here, the response of stem cells to these signals becomes unpredictable, and randomness plays an important role in determining the outcome. See Algorithm 1 for in-depth details.

there are three stable steady-states for each stem cell. These steady states correspond to three distinct cell fates: the stem cell itself and its corresponding differentiated cells. The cell's absorption to a specific attractor is determined by the values of x_i and y_i, which, in turn, defines the domains of the three attractors. Building upon previous research, we aim to investigate how variant environmental chemical

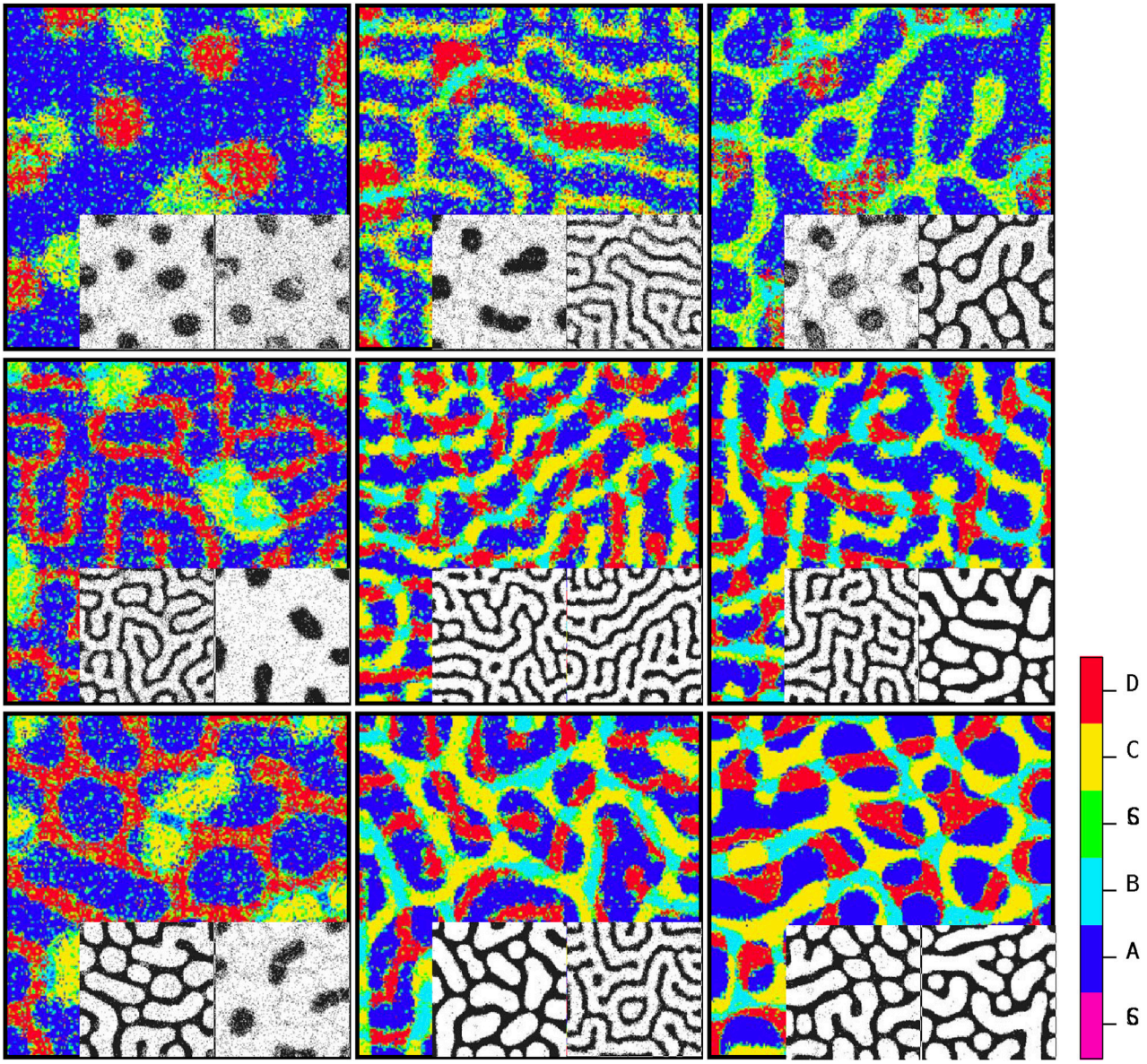


FIGURE 5
Steady-state patterns of developmental cellular arrangements in a multiple signaling field. The figures depict the patterns obtained using Algorithm 1, with each pattern governed by the dynamics of Equations 1, 2. The inline patterns, shown alongside, correspond to the signaling patterns predicted by Algorithm 2. By understanding the distribution of stimulating signals for stem cells, Algorithm 1 can determine the final cellular developmental pattern. Conversely, the inline patterns are generated according to the guidelines outlined in Algorithm 2 by analyzing the steady-state cellular pattern in each element as input. The color code reflects the cell types.

signals influence the concentrations of x_i and y_i. In this study, we consider two types of stem cells, SC_1 and SC_2, and assume the presence of two independent signals, s_1 and s_2, in the environment. The values of these signals evolve, and each stem cell (SC_i) can detect the presence of both s_1 and s_2. The cell then regulates its internal determinants based on the amount of these signals it receives, denoted as (s_1, s_2). We assume that the coefficients $\alpha_x^{(i)}$ and $\alpha_y^{(i)}$ are not fixed parameters, but rather, they are influenced by environmental signals s_1 and s_2. The behavior of $\alpha_x^{(i)}$ and $\alpha_y^{(i)}$ is governed by the following relations:

$$\alpha_j^{(1)} = \alpha_{0j}^{(1)} + \begin{cases} \eta(s_1)\eta(s_2) & if \quad j = x \\ \zeta(s_1)\zeta(s_2) & if \quad j = y \end{cases}, \tag{5}$$

$$\alpha_j^{(2)} = \alpha_{0j}^{(2)} + \begin{cases} \eta(s_2)\zeta(s_1) & if \quad j = x \\ \eta(s_1)\zeta(s_2) & if \quad j = y \end{cases} \tag{6}$$

Here, $\eta = \frac{s^2}{K_1^2+s^2}$ and $\zeta = \frac{K_2^2}{K_2^2+s^2}$. K_1 and K_2 are the fixed parameters. We see that different concentrations of s_1 and s_2 will lead to different levels of x_i and y_i, which will, in turn, influence the fate of the stem cells.In this study, the parameters of Equations 3, 4 were set as follows: $\gamma_1 = \gamma_2 = 0.38$, $\beta = 42$, $k_1 = k_2 = 30$, $\alpha_{0j}^{(1)} = \alpha_{0j}^{(2)} = 30$, and $n = 4$. Additionally, in the definition of η and ζ, both K_1 and K_2 were adjusted to equal 2.5. Finally, the parameters η and ζ were scaled up by a factor of 20. It is important to note that these parameters were determined through a trial and error process since the study was computational in nature.

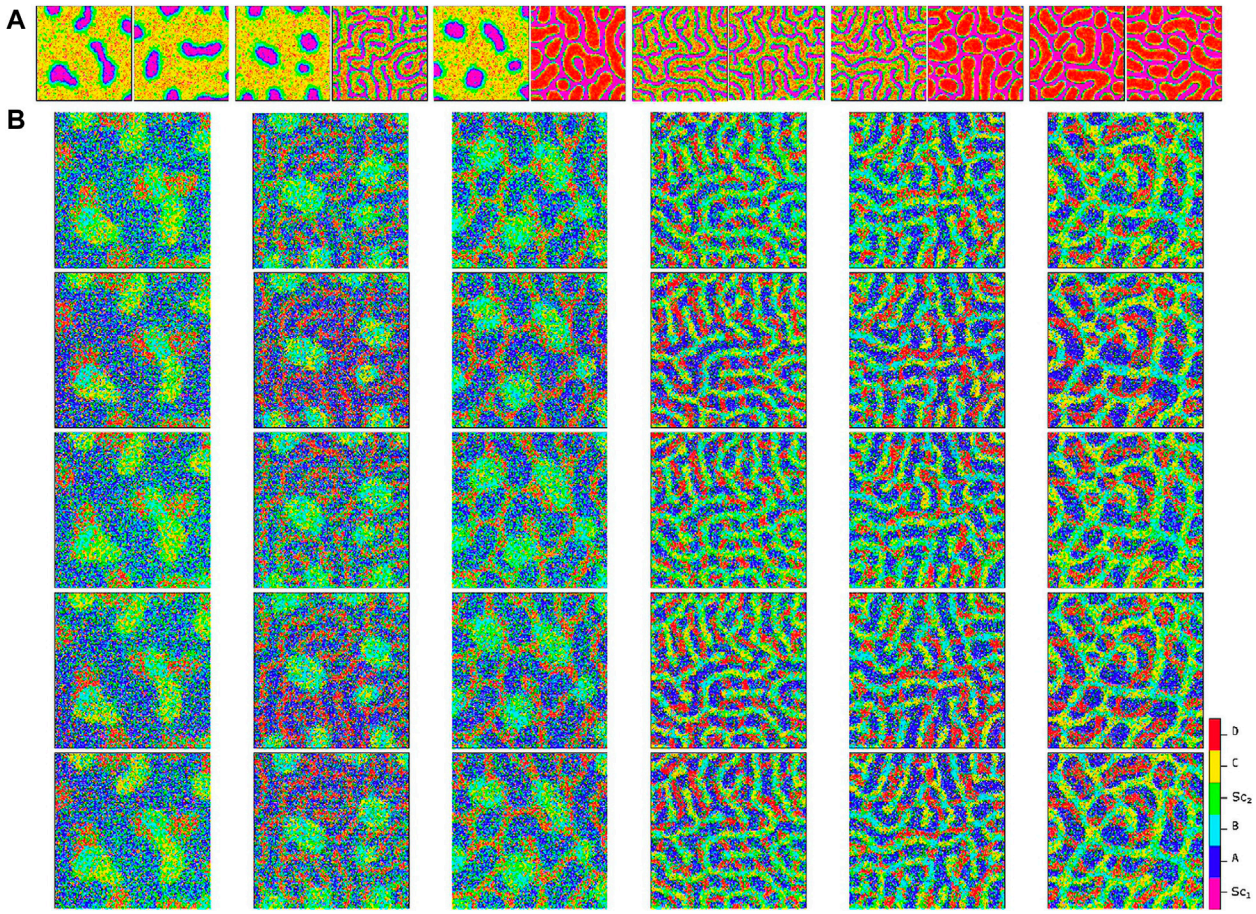


FIGURE 6
Comparison of cellular patterns for multiple pairs of signaling patterns simulated using the proposed algorithm. Panel **(A)** illustrates the signaling patterns pairs at the top of each column, while Panel **(B)** shows five independent simulated cellular patterns for each pair. The apparent similarity of the cellular patterns demonstrates the reproducibility of the method. Refer to Figure 7 for an analytical measure of the pattern similarity.

2.3 Patterns at the population level

In this section, we introduce a position-dependent procedure based on the Gillespie algorithm to simulate the development of differentiated cells in a population, as illustrated in Figure 4. The Gillespie algorithm is widely used for modeling systems with a small number of determinants or chemicals, taking into account inherent fluctuations. Previous studies have mainly focused on understanding the decision-making mechanism of a single type of stem cell. However, in this study, we extend our analysis to include multiple types of stem cells and various signaling stimulants in the system.

Consider a system comprising two types of stem cells, namely, SC_1 and SC_2. These stem cells possess the capability to undergo self-renewal and differentiate into their specialized cells. The differentiation process is regulated by the presence of signaling agents s_1 and s_2. SC_1 is capable of differentiating into A and B phenotypes, whereas SC_2 is competent to develop into C and D offspring; see Figure 1B. s_1 and s_2 independently propagate on the substrate via the reaction–diffusion dynamics of Equations 1, 2. The objective is to track the potential fate of stem cells at each location on a two-dimensional grid based on their exposure to two types of signals, s_1 and s_2. To achieve higher accuracy, the signal levels are scaled up to a range of 0–5. For each present cell type, a 6×6 lookup table is created at the start of the simulation, where each element in the table represents the potential number of cells of the corresponding cell type after division, assuming that a specific combination of s_1 and s_2 signals (s_1, s_2) exists at the mother cell's location.

The cell cycle span represents the average time interval in which each stem cell reaches the domain of one of its possible attractors: the stem cell itself or its differentiated offspring. Through trial and error, it has been determined that approximately 100 steps are necessary for the cells to reach a state of homeostasis. During this period, the values of intercellular signaling agents and intracellular determinants are updated using the Gillespie algorithm. Table 1 contains the list of reactions for these variants along with their corresponding propensity functions. Once this period is completed, the cells are ready to undergo division. At this point, we record the probable number of each possible fate based on the values of the signals and determinants. These numbers serve as the "virtual" destination of the stem cells and are recorded in their respective 6×6 look-up tables, as shown in Figure 4. The value 6 represents the resolution of the signal considered by the simulation for each cell. Consequently, the range of signal variations has been divided into six equal intervals. The selection of the element to enter the number of each probable fate in the table is directly dependent on the specific subinterval within which the values of s_1 and s_2 reside.

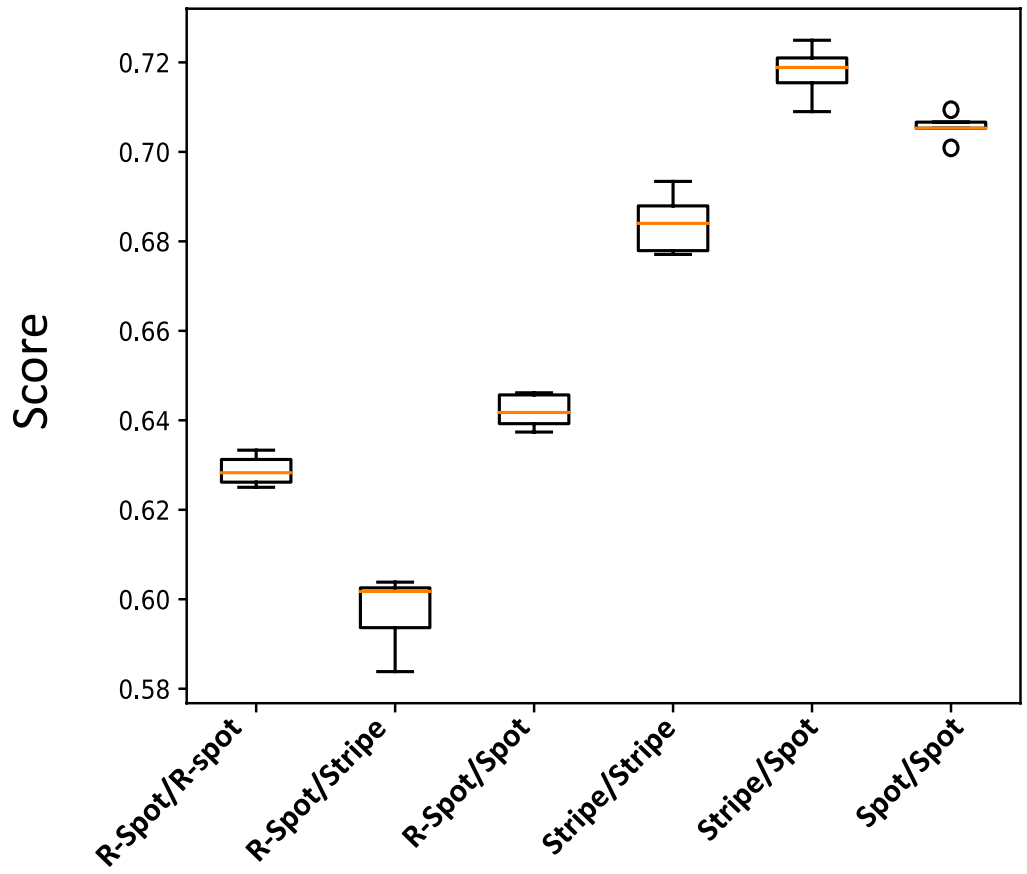


FIGURE 7
Box plot illustrating the normalized score values for the cellular patterns depicted in Figure 6. The score value measures the similarity of each pattern to a reference deterministic pattern created from their extreme signaling value pairs.

The process then repeats for another cell cycle duration, which is typically 100 steps. After collecting enough data in the look-up tables, for example after 1,000 steps, we can estimate the probability (Pb) of each of the six cell types being born. This is done by referring to the look-up tables and calculating the probability as the number of that particular cell type in the table divided by the total population size.

The entire process continues until the difference between two successive Pb values becomes smaller than a predefined tiny value, denoted as ϵ. This signifies that there is no significant change in the probability value and indicates the steady state of the pattern. The final pattern is constructed using the probabilities of creating each phenotype at the very last step.

To simulate the dynamic of the pattern formation through the division process provoked by the positional chemical information, we perform the following steps recurrently on a substrate of size $sz = 100$, on which SC_1 and SC_2 have been distributed randomly (Figure 4, panel of initial population arrangement).

1. For an adequate duration, such as 100 successive steps, let the dynamic of Equations 3, 4, upon which the number of determinant agents evolves, proceed. Here, we reckon that the signaling patterns of s_1 and s_2 simultaneously evolve based on Equations 1, 2 and provoke the stem cells toward a possible destiny.
2. Follow up the "potential" destiny of the stem cell located at each grid on the plane. Allocate a 6×6 look-up table for each of the present cell types (just once at the very first iteration). We scale up the amounts of signals to the range of (0, 5). This is the variation interval of the reverse spot signals. The rows of each table represent the number of subintervals that correspond to changes in s_1, while the columns represent the number of subintervals that correspond to changes in s_2. Next, we need to count the number of "virtual" offspring and renewed stem cells and categorize them based on the current amount of s_1 and s_2 in each location. Then, we insert the numbers into the row and column that correspond to the subinterval where they reside in the respective cell type; see Figure 4. It is crucial to highlight that, at this point, the fate of the cells is not determined. Instead, an assessment of their potential fate can be derived by taking into account the spatiotemporal value of (s_1, s_2).
3. Repeat the two previous steps 1,000 times, and record the corresponding classified data according to the above-mentioned method. In this way, one collects more data and, in consequence, the final predicted fate of the cells is closer to that of a real system.
4. Once every 1000 steps, assess the amount of s_1 and s_2 on every single grid of the main substrate and find out the corresponding number of the potential fate of the cell types on each of the six

look-up tables. Then, compute the probability of the virtual emergence of each cell type simply as the number which is associated with it in the look-up table, divided by the number of the whole population. After calculating the probability of all possible outcomes of the cell-fate random variable, we compare them with the same quantities for the 1,000 steps ahead and replace their maximum difference in the d variable, which is taken as an arbitrarily large value that guarantees that we will have enough repetitions in our simulations. Repeat the above sequence of instructions until the amount of d is less than that of $e = 0.0025$, which is adopted as an arbitrary and constant limit for the acceptable error in our simulations.

5. Finally, substitute the initial distribution of mother cells SC_1 and SC_2 on the substrate with the final pattern of daughter cells of each type based on their come-up probability. At this stage, every single grid of the main substrate is implanted with the cell type that is more likely conformed with the influence of the signal agent pair (s_1, s_2) on the internal switch of determinants; see Figure 4. For a summary of the ordered process, see Algorithm 1.

In a population of stem cells, the number of dividing cells remains constant. An alternative explanation for the above algorithm can be described as follows: once the mother cells reach a state of homeostasis after 100 steps, they divide. However, the algorithm disregards the differentiated cells as the algorithm focuses on studying the internal switch of the stem cells at this stage. Thus, we assume that only the stem cell daughter cells remain at each grid point. In the next iteration, the offspring stem cells explore the phase space of (x_i, y_i) by updating the values of x_i and y_i using the Gillespie algorithm. Then, these cells are absorbed into one of the three stable states of the internal switch: the stem cell itself or its differentiated offspring. As a result, the cells divide, and the results are recorded in the look-up tables. Again, the differentiated cells are disregarded, and the process is repeated for the stem cell offspring across the entire grid. The algorithm continues until completion.

The only additional assumption in this description is that every division always yields a stem cell as its daughter cell. The difference between the two descriptions lies in the fact that the first description defines potential cell fates, while the latter assumes that the divisions are real. Both descriptions aim to gather more data, resulting in a final predicted fate of the cells that are closer to that of a real system.

3 Results

3.1 Our signal-dependent tri-stable switch works

Figure 1D illustrates the solutions of Equations 3, 4 in the presence of variant pairs of (s_1, s_2). From different columns of the figure, it is evident that when the stem cells are exposed to different pairs of (s_1, s_2), the following fate of cell differentiation differs. The stem cells' response to the presence of signals, which is implemented via Equations 5, 6, depends on the amount of both s_1 and s_2. In other words, there are pair combinations of (s_1, s_2) that influence SC_1, while SC_2 remains neutral; *e.g.,* column 1 to 4 and vice versa (*.e.g.,* column 5–8). On the other hand, every single stem cell differentiates into one of its potential offspring based on the amount of (s_1, s_2) to which it has been exposed. The first row of the tabular Figure 1D displays the final course of action of SC_1 in the presence of variant combinations of (s_1, s_2). As it is seen in this row, in the presence of $(s_1, s_2) = (0, 0)$, B cell type is superior. The same trend is seen when $(s_1, s_2) = (2, 2)$ but with less difference between A and B production. In the presence of $(s_1, s_2) = (3, 3)$, the process is reversed, and A production becomes prior to that of B cells. When SC_1 experiences $(s_1, s_2) = (5, 5)$ pair signals, the A cell type becomes superior. The corresponding signal pairs of the last four columns have no impact on the preceding one of the cell types. Similarly, the second row illustrates the behavior of SC_2 in the presence of different pairs of (s_1, s_2). It is seen that the first four rows have no specific influence on altering production probabilities of C and D. Although in the presence of $(s_1, s_2) = (0, 5)$, the production rate of C is higher, the process becomes reversed when the (s_1, s_2) pair reaches (3,2). When s_2 vanishes and s_1 is on its highest value, *i.e.,* 5, the production probability of $D(C)$ is the maximum (minimum).

```
ϵ ← 0.0025. % a predefined small value.
d ← 10000. % the difference in emerging probabilities of
the six cell types between the successive steps.
  To ensure a sufficient number of iterations, the
  initial value of d is set as a large number.
co ← 0. % a dummy counter.
sz ← 100. % the number of grids on the plane.
pb_old[6][sz][sz] ← 0.
Construct the medium and plant stem cells, SC1 and SC2.
Form signal patterns s1 and s2.
while d ≥ ϵ do
  co ← co + 1.
  for i = 1 TO 100 do
    Update the system.
  end
  Let the stem cells be divided potentially, observe the
  offspring, and collect the data.
  if co%1000 == 0 then
    pb_new ←Compute the probability of the birth of each
    6 cell types based on the s1 and s2 values in their
    mother cells' grid.
    d ← Maximum value of |pb_new − pb_old|.
    pb_old ← pb_new.
  end
end
Design the corresponding medium based on the
collected data
```

Algorithm 1. The sequential instruction to form a complex cellular pattern based on a given signaling blueprint.

3.2 Individual cellular decisions lead to collective cellular patterns under the influence of combined signals

Figure 5 depicts the arrangement of the final developed cellular patterns induced as a result of variant possible combinations of signaling patterns governed by Eqs 1, 2. The color bar represents different cell types. Purple and green stand for SC_1 and SC_2,

respectively. Blue and cyan colors render phenotypes *A* and *B*, respectively, while yellow and red colors refer to *C* and *D* phenotypes, respectively. Each array of this arrangement corresponds to the combination of two members of the solution family of Eqs 1, 2, which are depicted inline in each case; see next paragraph. According to Figure 3, the solution family of these equations has three members: spot (left panel), reverse spot (middle panel), and stripe (right panel). From Figure 5, it is evident that the combination of these signaling patterns leads to a diverse collection of distinctive cellular pattern.

3.3 One can recognize the signal patterns from the final cellular arrangement, provided that the prior assumption of the pattern is given

```
sz ← 100.
pMat[sz][sz] ← the matrix corresponding to the population pattern, and the color code for different cell types.
s1[sz][sz] ← 0. %si[sz][sz], i ∈ {1,2} is the corresponding matrix of s_i on the plane.
s2[sz][sz] ← 0.
for i = 1 TO sz do
  for j = 1 TO sz do
    if pMat[i][j] == 1 then
      s1[sz][sz] ← 0.5.
      s2[sz][sz] ← 0.5.
    end
    if pMat[i][j] == 2 then
      s1[sz][sz] ← 1.
      s2[sz][sz] ← 1.
    end
    if pMat[i][j] == 3 then
      s1[sz][sz] ← 0.
      s2[sz][sz] ← 0.
    end
    if pMat[i][j] == 4 then
      s1[sz][sz] ← 0.5.
      s2[sz][sz] ← 0.5.
    end
    if pMat[i][j] == 5 then
      s1[sz][sz] ← 0.
      s2[sz][sz] ← 1.
    end
    if pMat[i][j] == 6 then
      s1[sz][sz] ← 1.
      s2[sz][sz] ← 0.
    end
  end
end
```

Algorithm 2. The sequential instructions for determining the form of triggering signaling patterns (spot, reverse spot, or stripe) associated with the final cellular arrangement in a system of two reproducing stem cells (SC_1 and SC_2) under the influence of two independent signals (s_1 and s_2) in the environment. The signals are generated through a Turing process with Equations 1, 2.

Figure 5 depicts the ultimate configurations of cellular arrangements resulting from different combinations of signaling patterns generated by Eqs 1, 2. The two corresponding acquired signaling patterns are displayed at the bottom right of each array. The key point here is that there is a dual relationship between the signal distribution and cell growth pattern. By understanding the distribution of signals that stimulate stem cells, algorithm 1 can be utilized to ascertain the final cell growth pattern. Conversely, by knowing the specific types of signals present, algorithm 2 can be employed to determine the parameters associated with the signal pattern based on the final cellular arrangement. In other words, if we are provided with a snapshot of the steady state of a developed cellular pattern and we assume that this pattern is influenced by two independent signals (s1 and s2) generated through a Turing process with Eqs 1, 2, algorithm 2 can predict the shape of each signal (spot, reverse spot, or stripe) based on the observed final cellular pattern. Recognition of signal patterns is a directional process. Algorithm 2: first, it is necessary to consider two blank planes, each of which is in accord with one of the signals to project its corresponding pattern onto it. Next, we go through every single pixel of the cellular pattern. Then, based on the color of the pixel, we map the projection of this color onto the signal planes. Let us assume that the color of a pixel is blue, meaning that this pixel is occupied with a cell of phenotype *A*. According to the relations (Eqs 5, 6) as well as Figure 4, this implies that at this spot, the concentration of both signals is approximately at its own summit. As a result, the projection of every blue pixel of the cellular pattern on both signal planes is a white point. Similarly, the cyan color in the cellular pattern corresponds to the *B* phenotype, whose occurrence is highly probable when the concentration of both signals is low. Accordingly, the map of each cyan pixel matches a corresponding black color on both signal planes. Likewise, the yellow (red) color represents the *C* phenotype (*D* phenotype), whose production rate is high when the concentration of s_1 is low (high), while that of s_2 is high (low). As a consequence, the projection of each yellow (red) pixel onto the corresponding point on the first signal plane is white (black), while its projection onto the similar point on the second signal plane is black (white). For a summary of the ordered process, see Algorithm 2.

4 Discussion

The positional stimuli have been emerging as key regulators of transcription and gene expression in diverse physiological contexts (Rulands et al., 2018). These environmental drivers engage in the phenotypic diversity and proliferation/differentiation balance of stem cells (Balázsi et al., 2011; Rulands and Simons, 2016; Blake et al., 2006). The regulation process of non-genetic diversity involves the interplay of intracellular and intercellular components to interpret positional cues (Çağatay et al., 2009; Acar et al., 2008). In a competing arena in which various chemical stimulants vie for affecting a cell's fate more, the process demands more robust and complex mechanisms. In order to specify and extend their offspring territory, the stem cells utilize a signaling process to communicate and collaborate with each other. This process ends in collective self-organized forms on length scales that are much larger than those of the individual units Chhabra et al. (2019).

In this study, first, we investigated the impact of multiple passive external signals on intracellular switches of a single stem cell. This provides us with a direct inspection of the connection between intracellular and extracellular dynamics. By mapping the environmental signaling patterns to the probability of the emergence of differentiated cell types, this model is capable of capturing any desired complex pattern, whether passive or active. The sort of models that recapitulates signaling dynamics and predicts cell fate patterning upon chemical perturbations precedingly has been investigated in non-competitive environments (Khorasani and Sadeghi, 2022; Khorasani et al., 2020; Sharifi-Zarchi et al., 2015; Chambers et al., 2007; Kalmar et al., 2009; Chen et al., 2010; Bergsmedh et al., 2011). Here, we focused on the behavior of each cell in the interaction with multiple signals. Figure 2 illustrates the resulting phenotypic cellular patterns of different combinations of two typical signal profiles of Gaussian and sinusoidal blueprints.

The environmental signals influence the fate of each stem cell, SCi (i = 1,2), by means of biasing the regulation of our tri-stable switch; see Equations 3, 4. Based on the definition of $\alpha_j^{(1)}$ and $\alpha_j^{(2)}$ in Equations 5, 6, it is evident that the pairs of (s_1, s_2) are relevant in controlling the decisions of this switch. This definition is advantageous in various aspects: first, it directs each stem cell's fate to the symmetric phase space of Figure 1C, where each of the patches correspond to one of the resulting phenotypes and there is no dominant domain between them. In addition, the representative patches are far enough apart to lead to distinctive outcomes in the occurring cellular pattern field. The narrow cruciform band, *i.e.,* the gray area in Figure 1C between these four patches, is where the fate of each cell is determined stochastically. From Figure 1D, it is evident that the regulatory switch plays either an active role or a neutral one based on the amount of existing signals (s_1, s_2) in each point, *i.e.,* combinations of (s_1, s_2), which effectively lead to offspring *A* or *B* from SC_1, have nothing to do with SC_2 and *vice versa*. In consequence, there is a smooth transition from left to right in each row of Figure 1D.

After investigating the impact of static environmental stimulants on the internal switch, we dealt with the active signaling between the sources that produce variant phenotypes. We took advantage of confined Turing models for two different signals secreted from each of stem cells (Shoji et al., 2003). The dynamic includes linear reaction terms and additional constraints that confine the two variables within a finite range. The resulting patterns of this dynamic are either stationary striped patterns or spotted patterns. The second pattern, in turn, consists of two forms: spotted and reverse spotted patterns. Here, the tuning parameter upon which the pattern type is specified is the maximum concentration of the activator $s_u^{(i)}$, where $i \in 1, 2$ (Shoji et al., 2003); see Figure 3. Based on this prior dynamic, nine distinct mutual patterns are generated by the two signals $s_u^{(1)}$ and $s_u^{(2)}$.

Stochasticity has been proven to be a non-genetic diversifying resource of variation in nature (Delbrück, 1940; McEntire et al., 2021; Acar et al., 2008; Kepler and Elston, 2001; Wu and Tzanakakis, 2012; Perez-Carrasco et al., 2016). It has been shown that controlled amount of randomness ends in phenotypic variation and, as a result, population heterogeneity (Losick and Desplan, 2008; Greulich and Simons, 2016; Khorasani et al., 2020). In this study, to reflect the non-deterministic portion of the signaling system, we implemented the Gillespie algorithm (Gillespie et al., 2007) by stepping in time to successive molecular reaction events according to the premises of the model of Shoji et al. (2003); see Equations 1, 2. Another aspect of incorporating randomness in our reductionist insight is simulating the emergence of every cell type in the look-up table based on the calculation of its corresponding probability; see Figure 4. Stochastic algorithms generally provide the chance to explore multiple solutions and potentially uncover a better one compared to a deterministic method, which may get stuck in a local minimum (Gillespie et al., 2007). Additionally, these algorithms can be easily tailored to different problems and constraints, making them adaptable for solving more complex issues. By utilizing stochastic methods, we can account for the inherent randomness and fluctuations present in natural systems. This strategy allows for controlled noise to be introduced into the system. As long as the level of randomness is controllable, the system's behavior remains predictable, and the resultant patterns are statistically reproducible.

In our algorithm, we evaluate the similarity between cellular patterns exposed to different pairs of signaling patterns by comparing them numerically to a cellular pattern constructed through a deterministic process while being exposed to the same pair of signaling patterns. This measure of similarity serves as an indicator of the reproducibility of cellular patterns using the algorithm proposed. To accomplish this, we first create two new 100×100 matrices, each corresponding to one of the signaling patterns. The size of the matrices corresponds to the plane on which the signals are distributed, with each element indicating the quantity of a specific signal at each grid location. The elements of these newly constructed matrices are either zero or the maximum value of that signal based on the corresponding elements in the original signaling matrices. If the original signal matrix element is less than half of its maximum number, the element in the new matrix is set to zero. If the element is greater than or equal to half of its maximum number, it is replaced by the maximum value of that signal. For example, when two reverse-spot type signaling patterns are distributed in the medium, each with a maximum value of 5, there are four possible pairs of extreme signals: (0,0),(0,5),(5,0), and (5,5). From Figure 1, it is evident that these pairs of signals lead to the emergence of *A*, *C*, *D*, and *B* types of cells, respectively.

By using these extreme signaling patterns, we can determine the fate of each stem cell in the medium and create a deterministic cellular pattern accordingly. We now have a reference pattern to assess the reproducibility of our algorithm and measure the resemblance of different patterns exposed to similar pairs of signaling patterns. Figure 6 illustrates five different realizations for various pairs of signaling patterns. The signaling patterns are shown above each column, and the resulting cellular population realizations are displayed below them in each column. We can compare each realization with its corresponding deterministic pattern, pixel by pixel. If the cell types in a pixel are identical, we assign a score of +1 for the resemblance of the pattern to the deterministic reference pattern. The normalized score function, which quantifies the resemblance, is the sum of all these +1's divided by the population size. Figure 7 shows the box plot of the

score variable for the cellular patterns in Figure 6 with respect to the different signaling pattern pairs. It is observed that for all the pairs, the median of the scores is above 0.6. We see that the synthesis of the signaling arrangement with the switch in the presence of controlled noise creates rich and highly reproducible organizations of differentiated cells. Figure 5 depicts the resulting patterns of the differentiated cells that have been exposed to various combinations of active signaling lay-outs of Figure 3. The procedure we dealt with in this study is one of the various known roots to construct an organized arrangement of cells. Mobility of cells (Gallagher et al., 2022), modulation of the physical and geometrical environment (Valet et al., 2022), and priming with chemical signals (Shahbazi et al., 2019) are among other intrinsic capacities of stem cells to make patterns. In practice, a combination of all these methods is incorporated to form an organization (Omid-Shafiei et al., 2023). Nevertheless, it is seen that solely following chemical environmental cues leads to the production of a rich and wide range of patterns.

In conclusion, this study demonstrates that the signal-dependent tri-stable switch can serve as a useful tool to bridge intracellular dynamics with intercellular structures. In this scenario, although the stem cells do not directly interact with each other, their reproduction rates are influenced by external signals in their environment through the switch mechanism within each cell. By studying individual cellular decisions and the influence of multiple signals, we observe how complex cellular patterns emerge. Although the algorithm utilized in this study simplifies certain aspects, such as considering the dynamic environment during cell division, apoptosis, and cell movement, the presented systematic approach allows for the simulation of complex cellular organizations based on fundamental biophysical processes, resulting in reproducible outcomes. Moreover, for any given complex cellular pattern, for which merely the prior class of signal patterns is known, the provided method closely concludes the signaling profile that sets off the cellular pattern. Overall, these findings highlight the potential of using signal-dependent switches for better comprehension and regulation of cellular behaviors in diverse scenarios.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

ZE: conceptualization, formal analysis, investigation, methodology, validation, visualization, and writing–original draft. NK: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, and writing–review and editing. MS: conceptualization, methodology, project administration, validation, and writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Acar, M., Mettetal, J. T., and Van Oudenaarden, A. (2008). Stochastic switching as a survival strategy in fluctuating environments. *Nat. Genet.* 40, 471–475. doi:10.1038/ng.110

Balázsi, G., Van Oudenaarden, A., and Collins, J. J. (2011). Cellular decision making and biological noise: from microbes to mammals. *Cell* 144, 910–925. doi:10.1016/j.cell.2011.01.030

Bergsmedh, A., Donohoe, M. E., Hughes, R.-A., and Hadjantonakis, A.-K. (2011). Understanding the molecular circuitry of cell lineage specification in the early mouse embryo. *Genes* 2, 420–448. doi:10.3390/genes2030420

Blake, W. J., Balázsi, G., Kohanski, M. A., Isaacs, F. J., Murphy, K. F., Kuang, Y., et al. (2006). Phenotypic consequences of promoter-mediated transcriptional noise. *Mol. Cell* 24, 853–865. doi:10.1016/j.molcel.2006.11.003

Britton, G., Chhabra, S., Massey, J., and Warmflash, A. (2021). "Fate-patterning of 2d gastruloids and ectodermal colonies using micropatterned human pluripotent stem cells," in *Programmed morphogenesis* (Springer), 119–130.

Çağatay, T., Turcotte, M., Elowitz, M. B., Garcia-Ojalvo, J., and Süel, G. M. (2009). Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell* 139, 512–522. doi:10.1016/j.cell.2009.07.046

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., et al. (2007). Nanog safeguards pluripotency and mediates germline development. *Nature* 450, 1230–1234. doi:10.1038/nature06403

Chen, L., Wang, D., Wu, Z., Ma, L., and Daley, G. Q. (2010). Molecular basis of the first cell fate determination in mouse embryogenesis. *Cell Res.* 20, 982–993. doi:10.1038/cr.2010.106

Chhabra, S., Liu, L., Goh, R., Kong, X., and Warmflash, A. (2019). Dissecting the dynamics of signaling events in the bmp, wnt, and nodal cascade during self-organized fate patterning in human gastruloids. *PLoS Biol.* 17, e3000498. doi:10.1371/journal.pbio.3000498

Cotterell, J., Robert-Moreno, A., and Sharpe, J. (2015). A local, self-organizing reaction-diffusion model can explain somite patterning in embryos. *Cell Syst.* 1, 257–269. doi:10.1016/j.cels.2015.10.002

Delbrück, M. (1940). Statistical fluctuations in autocatalytic reactions. *J. Chem. Phys.* 8, 120–124. doi:10.1063/1.1750549

De Santis, R., Etoc, F., Rosado-Olivieri, E. A., and Brivanlou, A. H. (2021). Self-organization of human dorsal-ventral forebrain structures by light induced shh. *Nat. Commun.* 12, 6768–6811. doi:10.1038/s41467-021-26881-w

Dubrulle, J., Jordan, B. M., Akhmetova, L., Farrell, J. A., Kim, S.-H., Solnica-Krezel, L., et al. (2015). Response to nodal morphogen gradient is determined by the kinetics of target gene induction. *Elife* 4, e05042. doi:10.7554/eLife.05042

Eidi, Z., Khorasani, N., and Sadeghi, M. (2021). Reactive/less-cooperative individuals advance population's synchronization: modeling of dictyostelium discoideum

concerted signaling during aggregation phase. *PloS one* 16, e0259742. doi:10.1371/journal.pone.0259742

Gallagher, K. D., Mani, M., and Carthew, R. W. (2022). Emergence of a geometric pattern of cell fates from tissue-scale mechanics in the drosophila eye. *Elife* 11, e72806. doi:10.7554/eLife.72806

Gillespie, D. T., et al. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* 58, 35–55. doi:10.1146/annurev.physchem.58.032806.104637

Green, J. B., and Sharpe, J. (2015). Positional information and reaction-diffusion: two big ideas in developmental biology combine. *Development* 142, 1203–1211. doi:10.1242/dev.114991

Greulich, P., and Simons, B. D. (2016). Dynamic heterogeneity as a strategy of stem cell self-renewal. *Proc. Natl. Acad. Sci.* 113, 7509–7514. doi:10.1073/pnas.1602779113

Heemskerk, I., Burt, K., Miller, M., Chhabra, S., Guerra, M. C., Liu, L., et al. (2019). Rapid changes in morphogen concentration control self-organized patterning in human embryonic stem cells. *Elife* 8, e40526. doi:10.7554/eLife.40526

Kalmar, T., Lim, C., Hayward, P., Munoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., et al. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol.* 7, e1000149. doi:10.1371/journal.pbio.1000149

Kepler, T. B., and Elston, T. C. (2001). Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical J.* 81, 3116–3136. doi:10.1016/S0006-3495(01)75949-8

Khorasani, N., and Sadeghi, M. (2022). A computational model of stem cells' decision-making mechanism to maintain tissue homeostasis and organization in the presence of stochasticity. *Sci. Rep.* 12, 9167–9217. doi:10.1038/s41598-022-12717-0

Khorasani, N., and Sadeghi, M. (2024). A computational model of stem cells' internal mechanism to recapitulate spatial patterning and maintain the self-organized pattern in the homeostasis state. *Sci. Rep.* 14, 1528–1615. doi:10.1038/s41598-024-51386-z

Khorasani, N., Sadeghi, M., and Nowzari-Dalini, A. (2020). A computational model of stem cell molecular mechanism to maintain tissue homeostasis. *Plos one* 15, e0236519. doi:10.1371/journal.pone.0236519

Koch, A., and Meinhardt, H. (1994). Biological pattern formation: from basic mechanisms to complex structures. *Rev. Mod. Phys.* 66, 1481–1507. doi:10.1103/revmodphys.66.1481

Lenne, P.-F., Munro, E., Heemskerk, I., Warmflash, A., Bocanegra-Moreno, L., Kishi, K., et al. (2021). Roadmap for the multiscale coupling of biochemical and mechanical signals during development. *Phys. Biol.* 18, 041501. doi:10.1088/1478-3975/abd0db

Liu, L., and Warmflash, A. (2021). Self-organized signaling in stem cell models of embryos. *Stem Cell Rep.* 16, 1065–1077. doi:10.1016/j.stemcr.2021.03.020

Losick, R., and Desplan, C. (2008). Stochasticity and cell fate. *science* 320, 65–68. doi:10.1126/science.1147888

Marcon, L., Diego, X., Sharpe, J., and Müller, P. (2016). High-throughput mathematical analysis identifies turing networks for patterning with equally diffusing signals. *Elife* 5, e14022. doi:10.7554/eLife.14022

McEntire, K. D., Gage, M., Gawne, R., Hadfield, M. G., Hulshof, C., Johnson, M. A., et al. (2021). Understanding drivers of variation and predicting variability across levels of biological organization. *Integr. Comp. Biol.* 61, 2119–2131. doi:10.1093/icb/icab160

Meinhardt, H. (2003). Models of biological pattern formation: common mechanism in plant and animal development. *Int. J. Dev. Biol.* 40, 123–134.

Murray, J. D. (2001). *Mathematical biology II: spatial models and biomedical applications*, 3. New York: Springer.

Omid-Shafiei, S., Hassan, M., Nu¨ssler, A. K., Najimi, M., and Vosough, M. (2023). A shadow of knowledge in stem cell science. *Cell J. (Yakhteh)* 25, 738–740. doi:10.22074/cellj.2023.2005680.1346

Perez-Carrasco, R., Guerrero, P., Briscoe, J., and Page, K. M. (2016). Intrinsic noise profoundly alters the dynamics and steady state of morphogen-controlled bistable genetic switches. *PLoS Comput. Biol.* 12, e1005154. doi:10.1371/journal.pcbi.1005154

Rulands, S., Lescroart, F., Chabab, S., Hindley, C. J., Prior, N., Sznurkowska, M. K., et al. (2018). Universality of clone dynamics during tissue development. *Nat. Phys.* 14, 469–474. doi:10.1038/s41567-018-0055-6

Rulands, S., and Simons, B. D. (2016). Tracing cellular dynamics in tissue development, maintenance and disease. *Curr. Opin. Cell Biol.* 43, 38–45. doi:10.1016/j.ceb.2016.07.001

Schweisguth, F., and Corson, F. (2019). Self-organization in pattern formation. *Dev. Cell* 49, 659–677. doi:10.1016/j.devcel.2019.05.019

Shahbazi, M. N., Siggia, E. D., and Zernicka-Goetz, M. (2019). Self-organization of stem cells into embryos: a window on early mammalian development. *Science* 364, 948–951. doi:10.1126/science.aax0164

Sharifi-Zarchi, A., Totonchi, M., Khaloughi, K., Karamzadeh, R., Araúzo-Bravo, M. J., Baharvand, H., et al. (2015). Increased robustness of early embryogenesis through collective decision-making by key transcription factors. *BMC Syst. Biol.* 9, 23–16. doi:10.1186/s12918-015-0169-8

Shoji, H., Iwasa, Y., and Kondo, S. (2003). Stripes, spots, or reversed spots in two-dimensional turing systems. *J. Theor. Biol.* 224, 339–350. doi:10.1016/s0022-5193(03)00170-x

Staff, P. C. B. (2017). Correction: intrinsic noise profoundly alters the dynamics and steady state of morphogen-controlled bistable genetic switches. *PLOS Comput. Biol.* 13, e1005563. doi:10.1371/journal.pcbi.1005563

Turing, A. (1990). The chemical basis of morphogenesis. *B. Jack Copel.* 52, 153–197. doi:10.1007/BF02459572

Valet, M., Siggia, E. D., and Brivanlou, A. H. (2022). Mechanical regulation of early vertebrate embryogenesis. *Nat. Rev. Mol. Cell Biol.* 23, 169–184. doi:10.1038/s41580-021-00424-z

van Boxtel, A. L., Chesebro, J. E., Heliot, C., Ramel, M.-C., Stone, R. K., and Hill, C. S. (2015). A temporal window for signal activation dictates the dimensions of a nodal signaling domain. *Dev. Cell* 35, 175–185. doi:10.1016/j.devcel.2015.09.014

Wagh, K., Ishikawa, M., Garcia, D. A., Stavreva, D. A., Upadhyaya, A., and Hager, G. L. (2021). Mechanical regulation of transcription: recent advances. *Trends Cell Biol.* 31, 457–472. doi:10.1016/j.tcb.2021.02.008

Wu, J., and Tzanakakis, E. S. (2012). Contribution of stochastic partitioning at human embryonic stem cell division to nanog heterogeneity. *PloS one* 7, e50715. doi:10.1371/journal.pone.0050715

Frontiers in
Cell and Developmental Biology

Explores the fundamental biological processes of life, covering intracellular and extracellular dynamics.

The world's most cited developmental biology journal, advancing our understanding of the fundamental processes of life. It explores a wide spectrum of cell and developmental biology, covering intracellular and extracellular dynamics.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact