5-2024

Single Cell Pharmacodynamic Modeling of Cancer Cell Lines

Arnab Mutsuddy
amutsud@g.clemson.edu

SINGLE CELL PHARMACODYNAMIC MODELING OF CANCER CELL LINES

A Dissertation

Presented to

The Graduate School of

Clemson University

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

Chemical Engineering

---

by

Arnab Mutsuddy

May 2024

Accepted by:

Dr. Marc Birtwistle, Committee Chair

Dr. Jessica Larsen

Dr. Adam Melvin

Dr. Jon Calhoun

Abstract

Cancer is one of the leading causes of disease related death worldwide. Since the discovery of the genomic origins of cancer, targeted therapy has been developed towards specific mutations implicated for oncogenic transformation. However, current standard-of-care for mapping cancer patients to efficacious drug combination is often inadequate. The pathophysiology of tumor progression relies on the dysregulation of biomolecular pathways of which the topology and the dynamics challenge prognosis. Moreover, the overall genomic instability involved in disease states and the resulting inter-patient as well as intra-tumoral heterogeneity challenge rationalization of therapy and clinical decision-making. It highlights the need for the use of quantitative methodologies that may forecast clinical outcomes considering the complex nature of disease progression.

In this work, we evaluated the use of single cell mechanistic modeling in predicting anticancer drug response. We begin our work with the foundation of one of the largest single cell models of stochastic proliferation and death signaling. It incorporates several signal transduction pathways which are implicated in oncogenic transformation and describes how the coordinated dynamics of these pathways drives stochastic outcomes of cellular processes such as proliferation or death in response to growth stimulus and drug dose. We addressed several aspects which may contribute to its future development towards a framework for generating unbiased drug response prediction with a more inclusive biological context encompassing multiple tumor types. At first, we focus on enhancing the accessibility and computational efficiency of the model by

introducing a scalable and modular format for its construction and potential expansion. Then, we developed a mechanistic cell population simulation framework based on the single cell simulation functionality of the model. This allowed us to generate representations of dynamic cell populations, bridging the gap between simulation outputs and experimental datasets, such as dose response for various drugs. A direct comparison of simulation outputs with experimental datasets enabled validation of the current modeled biology as well as identification of crucial knowledge gaps within the ERK signaling and cell cycle pathways. Furthermore, we developed a method to perform omics-informed context definition taking inputs of genomic, transcriptomic, and proteomic datasets for a number of cancer cell lines from one of the largest datasets of cancer cell characteristics, the Cancer Cell Line Encyclopedia. This allowed us to generate cell-line specific model variants as well as devise a strategy for the mechanistic exploration of drug sensitivity datasets generated for these cell lines. We believe the methods presented here will help provide guidance in attempting to build a deeper quantitative understanding of the dynamic and multivariate molecular complexities that currently challenge treatment efficacies in cancer.

Acknowledgements

First and foremost, I want to thank my parents for their love, understanding and guidance throughout my life. Their sacrifices and unwavering belief in me have been the cornerstone of my journey thus far.

I want to express my sincerest gratitude to my advisor, Dr. Marc Birtwistle, for the profound impact he had on my academic and personal journey. I shall remain ever grateful to him for his insightful advice and endless patience over the past five years. Throughout my PhD journey, he has always been a constant source of inspiration and encouragement. His guidance has been instrumental in shaping my growth as a researcher, and I am immensely grateful for having the opportunity to learn from him.

I want to thank Dr. Mehdi Bouhaddou for his exemplary work that laid the foundation for my research, and for the immensely helpful guidance he provided at the start of my journey as a graduate researcher. I am also indebted to Dr. Cemal Edem for his exceptional mentorship and guidance in computational modeling.

I want to extend my gratitude to Dr. Daniel Cook and Dr. Misha Salim for their helpful advice and technical support that made a large part of my dissertation work possible. I am also immensely grateful to Dr. Jon Calhoun for generously sharing his extraordinary expertise on high performance computation and parallel computing, without which I would not be able to complete this dissertation.

I would also want to extend my heartfelt gratitude to all former and current members of the Birtwistle Lab, with whom I've had the great pleasure of working, Alan Stern, Deepraj Sarmah, Madeline McCarthy, Dr. Xiaoming Lu, Aurore Amrit, Jonah Huggins, Daniel Pritko, Megan Abravanel, and Oluwaferanmi Ogunleye. Thank you all for making the lab such a great place to work. I would like to express my appreciation to all the undergraduate students who dedicated their time and effort to support my research, Micah Jordan, Will Dodd, Benjamin Usry, Isabel Leal and Benjamin Childs.

Finally, I want to express my sincere gratitude to my dissertation advisory committee, Dr. Jessica Larsen, Dr. Adam Melvin, and Dr. Jon Calhoun for their service and insightful assessment.

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Introduction

Cancer is a disease that continues to challenge human innovation. While advancements in modern medicine have helped eliminate the fatalities of numerous diseases in the past, expanding longevity and enhancing the quality of human lives, cancer persists as a formidable obstacle. Its relentless impact on human health stems from its adept manipulation of molecular physiology. By definition, cancer refers to a group of diseases characterized by uncontrolled and abnormal cell growth, primarily caused by accumulated genetic mutations[1,2]. Such aberrant cellular proliferation, often difficult to treat, may disrupt the organization and function of tissues and organs, ultimately leading to the patient's demise. Despite the recent advancements in modern medicine, cancer remains a leading cause of death by disease. According to recent statistics, more than 19 million people around the world are diagnosed with cancer every year, and about 10 million die of the disease[3].

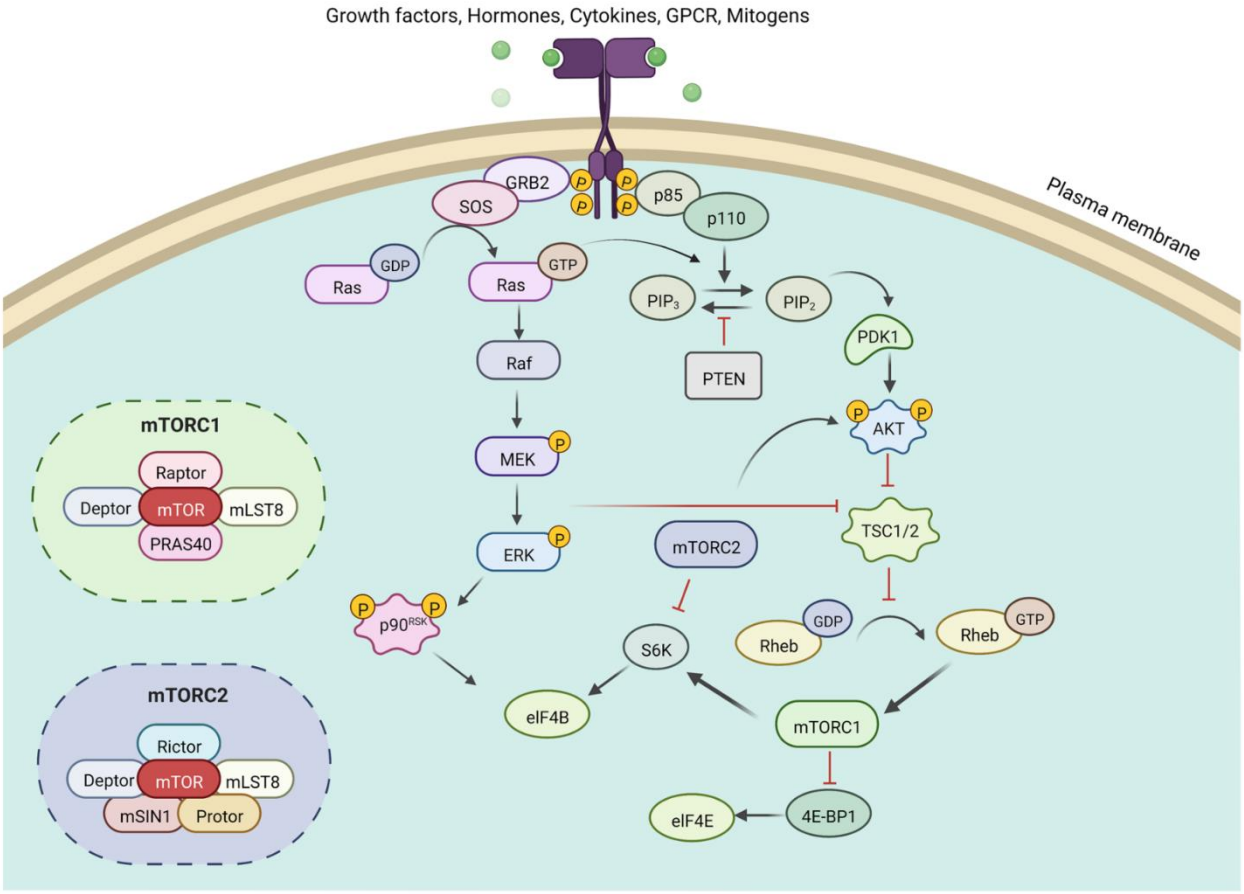
1.2 Difficulties in Cancer Treatment

Conventional treatment strategy for cancer included surgical resection, which is most effective at an early state of the disease. Following the discovery of x-ray, radiotherapy was introduced as an adjuvant. In 20th century, it was discovered that nitrogen mustard, a chemical warfare agent used in World War I had therapeutic effects on patients with non-Hodgkin's lymphoma[4]. This led to the widespread use of DNA alkylating agents such as chlorambucil and cyclophosphamide for cancer treatment

which became known as chemotherapy. However, success with conventional cancer treatment modalities was far from comprehensive. In many cases, tumor remissions were brief and incomplete. Radiation and chemotherapy were also known to damage healthy cells, organs and tissues[5]. Drug resistance in previously suppressed cancer cells due to reduced drug uptake and increased drug efflux was also a common problem[6].

In later years, discovery of genetic origin of cancer spurred the development of targeted therapy, whereby a drug engages with specific subcellular biomolecular targets, often originating from driver mutations in cancer cells. A prominent example is imatinib[7], which inhibits the BCR-ABL fusion protein in patients with chronic myelogenous leukemia. Other examples include sunitinib for renal cell carcinoma[8] and trastuzumab for Her2-positive breast cancer[9]. Despite the success stories, challenges persist even with targeted therapy. In many cases, patients with matching mutations fail to respond as anticipated by precision medicine[10]. Tumor cells are also known to exploit biomolecular pathways to develop resistance to targeted therapy[11]. All these problems stem from the complexity of the biomolecular processes underlying the pathophysiology. The inception and progression of cancer involves multi-scale and multivariate complexities. The cell uses circuits assembled from arrays of intercommunicating components, which are predominantly gene products such as proteins and RNAs, and metabolites. The interactions that these components engage in can be quite complex and their spatiotemporal dynamics are essential in maintaining various phenotypical functions.

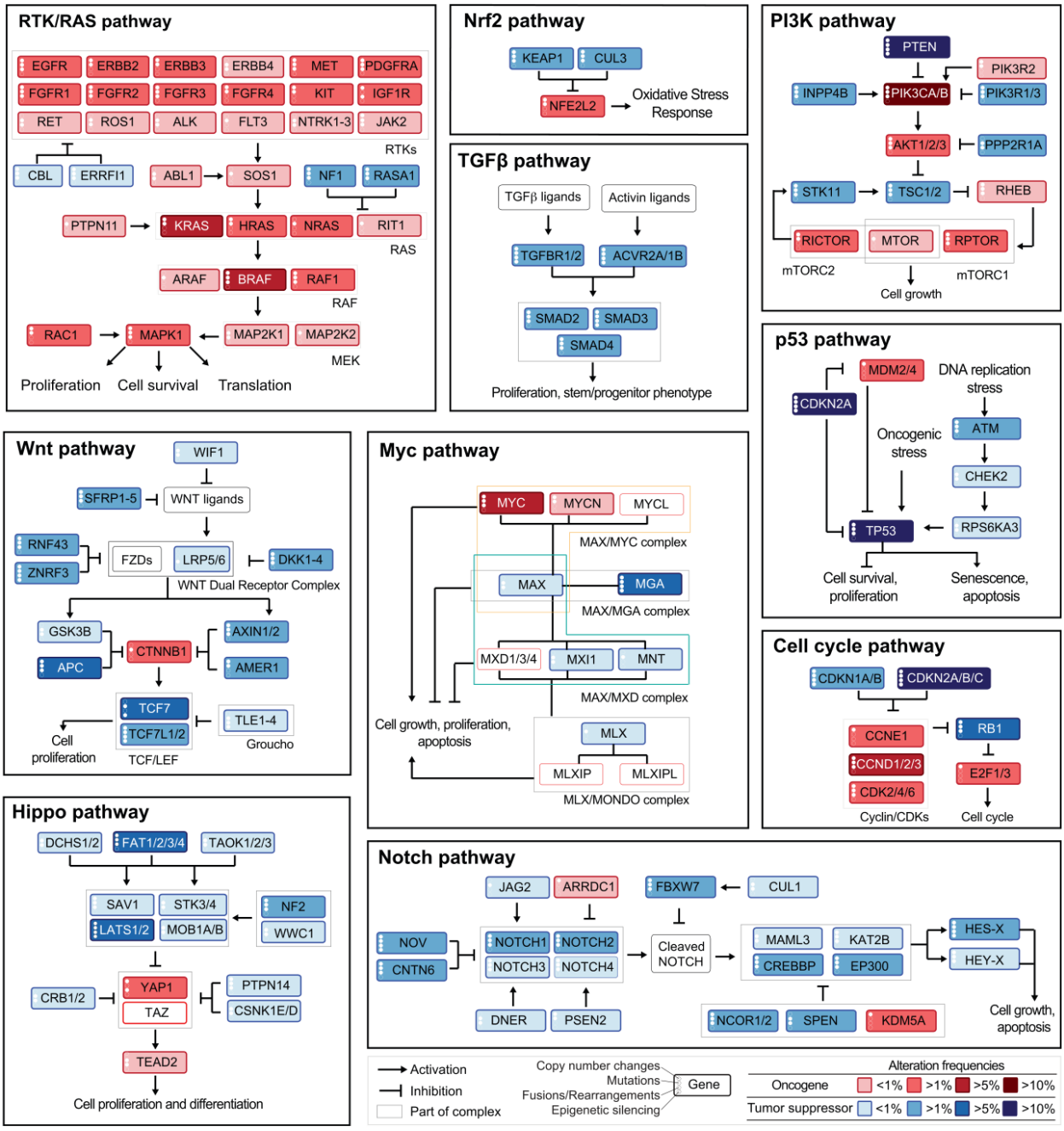
Growth factor receptors located at cell surface gather a wide variety of signals and funnel them into the cytoplasm. Throughout cytoplasm, a complex circuitry of signaling proteins interacting with precision and specificity transmit signals from upstream components and pass on to their intended downstream components. Outputs of these signal processing mechanisms are then transmitted to the nucleus, providing critical inputs to the regulatory mechanism governing cell proliferation. All these signaling pathways along with their numerous feedback and feed-forward loops and crosstalks make up the cellular signal transduction network, examples include the MAPK (Mitogen Activated Protein Kinase) pathway and PI3K-AKT-mTOR pathway[12] (Fig. 1.1). As per the molecular origins of tumor formation, dysregulation of signaling pathways occurs as a result of genetic mutations, epigenetic alterations, and other chromosomal abnormalities. Tumor progression hinges on the hijacking of these pathways, granting cancer cells a proliferative advantage over normal cells. The idea of molecular targeted therapy is to combat proliferation of tumor cells by intercepting aberrant signaling proteins. However, incomplete understanding of how signaling proteins are interconnected within the biomolecular network may result in unexpected outcomes in therapeutic procedures. For example, mTOR inhibition with rapamycin is known to result in increased Akt activity[13] indicating negative feedback mechanisms that were unaccounted for.



Adapted from Li et al. 2022 Molecular Biomedicine

Figure 1.1: Schematic of the PI3K/AKT/mTOR and Raf/MEK/ERK (MAP Kinase) signaling pathways.

Adapted from Li et al. (2022). Growth factors, hormones, cytokines, GPCRs, and mitogens activate receptor tyrosine kinases (RTKs) recruiting PI3K to attach to the plasma membrane. PI3K catalyzes PIP2 to PIP3, which then promotes AKT activation via the activity of PDK1 and mTORC2. RTK activation may also accelerate guanine exchange factors to load Ras with GTP. Ras-GTP dimers recruit RAFs to promote MEP activation. This leads to phosphorylation of ERK.



Adapted from Sanchez-Vega et al. 2018 Cell

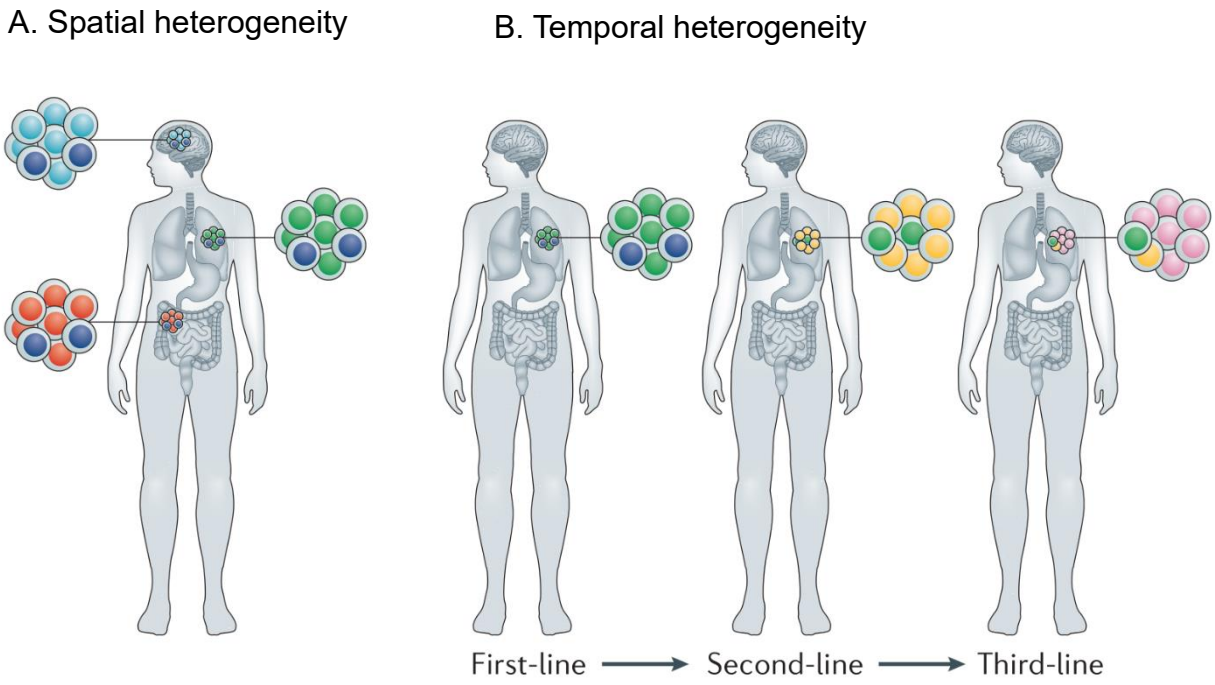
Figure 1.2: Curated chart of major altered pathways in the Cancer Genome Atlas. Adapted from Sanchez-Vega et al. (2018). Pathway members and their interactions in ten selected pathways. Average alteration frequency of genes are indicated with color intensity of red (oncogenic activations) or blue (tumor suppressor inactivations)

Beyond intricate dynamics observed within molecular biology, the heterogeneous landscape of cancer across patients and tumor types adds another layer of complexity to therapeutic decision-making. Extensive genomic and transcriptomic characterization of tumor samples across the world has revealed a wide range of driver mutations (Fig. 1.2), occurring across several major signaling pathways[14–16]. Driver mutations may also vary across patients, affecting their response to treatment[17].

In addition to the tumor heterogeneity observed across patients, patients are also known to exhibit intra-tumoral heterogeneity (Fig. 1.3) whereby any tumor may consist of a collection of cells with distinct biomolecular signatures with differential response to treatment[18]. Intra-tumoral heterogeneity may occur as spatial heterogeneity, with uneven distribution of genetically distinct subpopulations across disease sites. It may also occur as temporal heterogeneity, which refers to the dynamic variation of molecular signature of tumor cells over time.

The intricate nature of biomolecular processes and heterogeneous landscape of cancer are the main reasons for its poor prognosis in therapeutic procedure. It also necessitates the use of combination therapy[19] where efficacious treatment regimen is challenging to develop due to incomplete understanding of patient response and complexities associated with the large number of drug combinations that require consideration. There is an inadequate number of cancer patients available for clinical trial due to various demographic and socio-economic reasons[20]. If we consider the current number (72) of FDA approved therapeutic kinase inhibitor agents[21], more than 2500 two-way and almost 60,000 three-way combinations could be conceived which further limits the scope of discovery that clinical trials may deliver. As a result, it is

becoming increasingly difficult to make important decisions in clinical practice with regards to the expected outcome of a given drug or drug combination and stratifying patient groups based on disease state and prognosis.



Adapted from Dagogo-Jack et al. (2018) Nat. Rev. Clin. Oncol.

Figure 1.3: Conceptual schematic of spatial (A) and temporal heterogeneity (B) in a cancer patient.
Adapted from Dagogo-Jack et al. (2018). (A) Spatial heterogeneity refers uneven distribution of cancer subclones with unique genomic characteristics across different regions of primary tumor or metastatic sites. (B) Temporal heterogeneity is variation of molecular characteristics of a tumor over time, which can be due to natural progression of the tumor or a result of selective pressure due to therapy. Colors are a symbolic representation of unique molecular characteristics of tumor cells.

1.3 The Need for Computational Models

The aforementioned problems underscore the necessity for an innovative approach in clinical decision making, one that may lean more heavily on quantitative analysis to forecast efficacy. Considering the limits of basic human intuition in performing quantitative analysis of complex systems, computational models are an ideal candidate for this purpose. Mathematical models may serve as virtual laboratories with meticulously controlled conditions, enabling scientists and clinicians to investigate emergent clinical behaviors that result from insights into cell signaling biology and pharmacodynamics. If such systems can make reliable predictions, they may help evaluate new therapeutic strategies. Statistical modeling approaches have been demonstrated in classifying genomic signatures that correlate with treatment efficacy[22–25]. However, they lack consideration of biophysiological mechanism of tumorigenesis and causal mechanisms driving drug response, precluding their application as comprehensive support for clinical decision-making. Moreover, statistical models predominantly rely on machine learning to generate predictions while avoiding the need to understand complex mechanisms. To provide accurate predictions, these models require training with large scale datasets. The complex nature of causality involved with disease progression and drug response in cancer constitutes a high dimensional problem, massively escalating the requirements for training data which can be challenging to generate.

On the other hand, mechanistic models are derived from physical assumptions about a system, which can allow extrapolation outside of the context of its parameterization and provide insight into observed phenomena. These models may

incorporate formalisms from chemical kinetics theory and principles of mass action and mass balance to describe biochemical interactions between proteins. They represent biochemical mechanisms as rate laws that describe progression of reactions based on the concentrations of reactants and products. For any cellular entity, the rate of change in its amount is the sum of all rates that generate it, minus all rates that consume it. This gives rise to a system of ordinary differential equations (ODEs) describing the temporal evolution of the system as a whole. Over the years, mechanistic models based on systems of ODEs have been successfully applied to reproduce biologically correct behavior and hypothesize on underlying biological mechanism in a wide range of cellular systems and signaling pathways for processes such as receptor binding[26], cell cycle[27], DNA damage[28] and apoptosis[29], all of which are known to be implicated with driver mutations in oncogenic transformation. The pathway-centric models that have been formulated earlier have laid the foundation for building disease-centric models by combining them with pharmacodynamic profiles of relevant drugs with an aim to predict drug dosage required to reach certain desired outcomes[30,31].

1.4 Single Cell Mechanistic Pharmacodynamic Modeling

In a previous work accomplished at our laboratory, one of the largest single cell mechanistic models of stochastic proliferation and death signaling was built[32]. It incorporates several signal transduction pathways which are implicated in oncogenic transformation, such as receptor tyrosine kinase (RTK), RAF-MEK-ERK signaling, PI3K-AKT-mTOR signaling, cell cycle, DNA damage and apoptosis. The model describes expression of 141 genes that are included in the mentioned pathways and dynamic interactions of their resulting products, such as proteins, and protein complexes.

Furthermore, the model allows inclusion of the pharmacodynamics of any drug of interest as binding reactions with their intended pathway targets. Simulation of the model describes temporal evolution of a single cell whereby the coordinated dynamics of multiple pathways and possible drug actions may give rise to stochastic cell fate in terms of cell division or death. Initial study showcased the model's predictive capability with accurate predictions of synergistic effects of MEK-inhibitor and AKT-inhibitor drugs for MCF10A cells as well as differential sensitivity of U87 glioma cells to these inhibitors. The mathematical formalism employed in this model allows representation of some significant molecular level complexities involved in tumor formation and disease progression including dynamics of pathway activity and biomolecular heterogeneity at the single cell level in terms of cell-to-cell variability. Moreover, the model structure allows delineation of biological context through the integration of genomic, transcriptomic and proteomic data sourced from any cell line and may potentially be used for patient specificity[30]. For future work, we intend to evaluate the applicability of the model as a framework for predicting anticancer drug responses with a broader and potentially more inclusive biological context encompassing multiple tumor types. Enhancement of such nature may necessitate the incorporation of a broader array of signaling pathways, given that the initially integrated pathways constitute only a subset of those relevant to oncogenic transformation. When incorporating additional biomolecular processes, it is crucial to ensure that the model accurately represents the intended biology. One effective approach could involve validating the model's predicted outcomes against numerous experimental perturbation datasets, primarily accessible as drug sensitivity profiles across a diverse spectrum of cancer cell lines[33–35].

1.5 Thesis Overview

In this work, we sought to improve several aspects that challenge our goal of enhancing the predictive capability of single cell pharmacodynamic modeling for anticancer drug response, namely, (1) enhancing its accessibility by creating a scalable and modular software pipeline for model construction and potential expansion; (2) developing methods for validating its biology and identifying significant knowledge gaps by comparison with experimental drug dose response data; and (3) enhancing its adaptability with new biological context informed by genomic, transcriptomic, and proteomic data.

Starting with Chapter 2, we introduce SPARCED, a modular and scalable implementation of our single cell mechanistic pan-cancer driver pathway model. It streamlines the procedure for the modification of model structure, enabling efficient expansion of the model with pharmacodynamics for a broader range of drugs and new signaling pathways.

Continuation of our work with the SPARCED model relies on our ability to perform more complex and resource intensive computation. Therefore, in Chapter 3, we describe extensive performance benchmarking and improvement of our simulation algorithm, achieving at least 4-fold increase for stochastic and more than 200-fold increase for deterministic computation speed.

Dose response assays in general measure drug sensitivity or resistance by capturing cell population characteristics, such as viable cell counts at specific durations of treatment. In order to accomplish expansion and enhancement of single cell models

based on experimental dose response data, a linkage needs to be established between dynamic interactions at the cellular pathway level and their emergent outcomes at the cell population level. In Chapter 4, we describe the development of a mechanistic cell population simulation framework which attempts to reconcile results from dose response experiments with outcomes from mechanistic single cell models. Furthermore, we discuss results from simulations of dose response experiment using our cell population simulation framework in the context of MCF10A cells treated with four different anti-cancer drugs, namely, alpelisib, neratinib, trametinib and palbociclib. The results and subsequent analysis helped us validate effects of drug action on the MAPK pathway as well as identify some knowledge gaps in the representation of RTK and cell cycle pathways.

In Chapter 5, we describe the development of a robust pipeline for omics-informed context definition for our single cell model. For this purpose, we focus on the Cancer Cell Line Encyclopedia (CCLE), one of the largest and most comprehensive databases where more than 1000 cancer cell lines have been characterized with genomic, transcriptomic, and proteomic analyses as well as drug sensitivity profiles for 24 anticancer drugs. We initialized the single cell model with omics data for several cell lines originating from various tissue types, laying the groundwork for constructing virtual drug sensitivity profiles. These profiles may serve as a roadmap for expanding the biological context of our model.

In Chapter 6, we highlight certain limitations within our cell cycle submodel, notably the absence of proteomics informed cell cycle species levels and their

stochastic gene expression. We explore potential remedies, including revisiting and updating the cell cycle submodel to incorporate the latest pathway knowledge.

And finally, in Chapter 7, we draw the conclusions from the current work and discuss its broader impact in improving single cell pharmacodynamic for anticancer therapy. Additionally, we explore potential future avenues for research in this domain.

Chapter 2

DEVELOPMENT OF A MODULAR AND SCALABLE PIPELINE FOR A LARGE-SCALE MECHANISTIC MODEL OF SINGE CELL PROLIFERATION AND DEATH SIGNALING

2.1 Author Contribution

The work presented in this chapter is adapted from the following publication:

> Erdem, C., **Mutsuddy, A.**, Bensman, E.M., Dodd, W.B., Saint-Antoine, M.M., Bouhaddou, M., Blake, R.C., Gross, S.M., Heiser, L.M., Feltus, F.A. and Birtwistle, M.R., 2022. A scalable, open-source implementation of a large-scale mechanistic model for single cell proliferation and death signaling. *Nature communications*, 13(1), p.3555.

The contribution of the candidate involved curation of data, composition of model input files, development of software pipeline for model construction and context definition, and validation of model construction and simulation protocols.

2.2 Introduction

The ever-increasing availability and accumulation of FAIR[36] (findable, accessible, interoperable, and reproducible) and big (omics) datasets requires new computational methods and models to integrate, analyze, and interpret the underlying information[37–39]. How can we leverage the totality of available information not only to learn more about biology but also to make predictions, especially those that are clinically relevant? Advances in statistical and machine learning approaches enable (mostly) data-driven exploration and hypothesis generation from big datasets[40–43]. Trained on features of the input dataset(s), such models can be used for, as just a few examples, to predict drug responses[44–46] or decide tumor type/stage[47–50]. Although transformative, such machine learning and statistical models have shortcomings. Most notably, they often fail to explain predicted outcomes with detailed mechanistic reasoning[51–55] – a major scientific gap and a roadblock to reconciling and integrating such models.

Besides such “black-box” modeling approaches, an alternative and complementary vehicle for data integration are so-called “mechanistic models”[55]. Mechanistic models provide an interpretable integration of different data types, because they have explicitly modeled biophysical correlates, while enabling further exploration for underlying logic behind heterogeneous, nonlinear, and often unintuitive relationships across big datasets[56]. If mechanistic models are available towards the whole-genome or whole-single-cell scale, one can start to predict complex, multi-network, and emergent cellular behaviors[57,58], elucidate phenotypic responses to multiple perturbations[59,60] tailor and train on patient-specific data for personalized, pharmacologic decision making[61,62], or use them as “data integrators” for data consistency checking[63]. However,

most published mechanistic models are “small” scale; built for single pathways with a handful of genes, meant to interpret a single dataset[64–73]. Such small-scale mechanistic models provided important insights into processes such as yeast response to pheromones[70], lac operon regulation in E. coli[69], or phenotypic responses to different ligand stimulations[64]. However, the limited scope of small-scale models means they inherently will struggle to integrate multiple datasets. Large-scale mechanistic models[32,58,74,75] on the other hand, can provide a more extensive representation of cellular interactions and are thus well-poised for data integration that complement shortcomings of machine learning approaches.

One of the many ways of mechanistic model construction is the use and modification of existing models by inserting new species or interactions to explain new experimental observations[73,76,77]. Model merging, the act of stitching pre-existing models together, is an extension of this method for creating larger models. However, such an approach requires extensive detail checking and harmonizing species/parameter definitions. Often, unfortunately, sufficient annotation is not provided which makes this task harder. Moreover, while most mechanistic models are comprised of ordinary differential equations (ODEs), many large-scale models require multiple sub-modules of different mathematical formalisms. For example, metabolic processes are usually described by steady-state flux-balance models [78,79], gene expression events are stochastic[80–82], and protein signaling events are represented by a system of ODEs[64,65,73]. Thus, sorting out a single platform for different modeling formalisms to create a large-scale model is a daunting task. It is so far only achieved by creating highly custom-structured and custom-coded model-agglomerates that are not well-

suited to further alterations or re-use[32,58]. The latter, Bouhaddou2018 pan-cancer model[32], is previously published by our group to study single-cell responses to mitogens and drugs.

A second way of constructing models is to build them bottom-up by writing out every reaction one by one. In this regard, rule-based modeling (RBM) provides an innovative approach[83]. RBM software, such as BioNetGen[84,85], Kappa[86], and PySB[87], enables researchers to write “rules” for repeated reaction events following specific patterns. RBM software then creates the reaction network by propagating the rules from the initial set of species. Although RBM revolutionized large-scale model construction by minimizing manual equation scripting (i.e., writing out every differential equation), some limitations exist. First, it can generate a vast (even infinite) number of reactions from a small set of rules (usually called the curse of combinatorial complexity). This makes interpreting, analyzing, and debugging such models cumbersome, if possible. Tools like NFsim[88] can overcome such problems by simulating events based on the rules rather than a priori generating the entire reaction network. Thus, such software becomes advantageous when a small number of rules create a very large number of reactions, e.g., polymerization, aggregation, or multi-site phosphorylation[89]. However, such network-free simulators typically require an explicit representation of every molecule in the system, which dramatically increases the computational cost and renders such methods inefficient for large-scale mechanistic models. Secondly, current RBM implementations dictate that reactions taking place via the same rule have the same rate constant parameter values. Often, allostery or site cooperativity precludes this simplifying assumption, leading to manually writing out every such reaction in the model

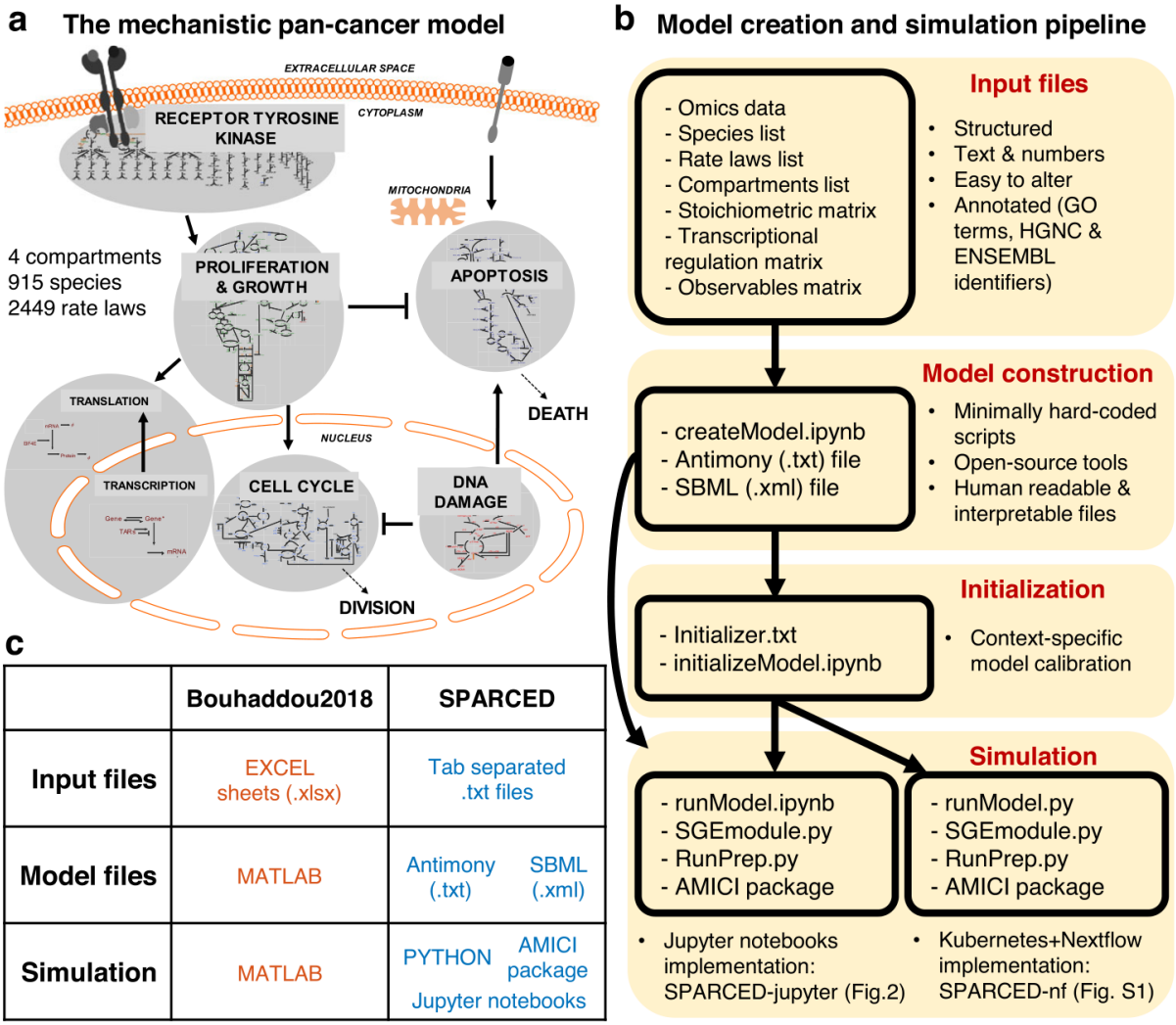
(or writing one rule for each reaction), which then obviates the advantages of RBM. Finally, with its capability of capturing biological complexity via simple rules, the RBM concept is quite powerful but additional efforts are needed to enable merging of existing non-rule-based models, creating a mixture of different modeling formats (i.e., mixed-grain modeling), and defining different simulation settings (i.e., hybrid modeling = deterministic + stochastic parts).

Regardless of how a large-scale model is constructed, it should have certain properties for FAIRness (findable, accessible, interoperable, and reproducible) and re-useability[90–92]. Porubsky et al.[91] recently summarized the best modeling practices and reinforced: providing metadata/annotations and model creation steps/files (Practices 1-5), using standard and cross-platform model files (Practice 3), and open-source, license-free, version-controlled, and reproducible model dissemination (Practices 8-9). As the size of the model increases, conforming to modeling standards (e.g., simulation type, simulation speed, software to use, scripting package to use, algorithm to use) gets harder. That is why most of the large-scale (many genes or whole-cell) models are necessarily custom-structured, are composed of multiple submodules, or are lacking sufficient annotations and metadata (e.g., ENSEMBL or HGNC identifiers)[32,58,74,75]. These custom-made models also do not yet follow a single standard format, a key property for easy distribution, re-use, and model merging and expansion with other models. The SBML (Systems Biology Markup Language) format[93,94] offers a long-established and well-defined way of specifying annotated model structures, with an explicit and structured definition of each element of a mechanistic model (species, reactions, volumes, initial concentrations, parameters, rules, events, equations). SBML

is an extensible, machine-readable markup language and not a simple text file. SBML has interfaces and packages in most programming languages (like Python, C++, Perl) and can be imported by most software (Python, MATLAB, COPASI[95], Virtual Cell[96], and another ~300 packages). However, it is non-trivial to write thousands of reactions in SBML standards, directly or with available GUI-based software. To circumvent this problem, there are efforts to convert other model formats to SBML, like Antimony[97]. The Antimony format is defined in simple text format and is human readable and interpretable. Regardless, any constructed mechanistic model, in SBML format or not, must be simulated with reasonable CPU time. Although simulating models on local machines is often done, High Performance (HPC) or Cloud Computing (CC) platforms are suitable for larger tasks such as parameter sensitivity/estimation or multiple single-cell simulations[98–100]. Therefore, another milestone for large-scale mechanistic models is inherent HPC/CC compatibility, especially for single-cell simulations and heterogeneous data integration.

Here, we provide a framework and model construction recipe for large-scale mechanistic modeling that converts our lab's previous large-scale pan-cancer model into a format that conveys several crucial properties noted above. First, we define a simple set of structured and annotated input text files that set model specifics: genes, species, reactions, reaction stoichiometry, cellular compartments, transcriptional regulations, input omics data, and parameter values (Fig. 2.1). These text files enable easy creation or alteration of the model network, with minimal coding or software usage requirements (but they are easily amenable to such things if desired). We then use Jupyter notebooks[101] to process the input files and to create a human-interpretable

Antimony file, which is then converted into an SBML (community gold-standard) model file. We simulate the model using SBML compatible Python packages including AMICI, specifically designed for efficient simulation of large-scale models[102,103], and our own Python submodule for stochastic gene expression that enables single-cell simulations. We also developed an HPC/CC (Kubernetes) compatible version of the pipeline that enables simulating large number of single cells and/or stimulation conditions. To apply our work, we re-create and extend our previous single mammalian cell mechanistic model of proliferation and death signaling and regulation[32], which we call SPARCED (SBML, Proliferation, Apoptosis, Receptor Tyrosine Kinases, Cell cycle, Expression, DNA damage). The pipeline and model are available on GitHub (github.com/birtwistlelab/SPARCED).



c

	Bouhaddou2018	SPARCED
Input files	EXCEL sheets (.xlsx)	Tab separated .txt files
Model files	MATLAB	Antimony (.txt) SBML (.xml)
Simulation	MATLAB	PYTHON AMICI package Jupyter notebooks

Figure 2.1: SPARCED is a structured, human interpretable, and easy to modify big mechanistic model. (a) The schematic of the underlying model for SPARCED. Image adapted from (30). (b) The pan-cancer mechanistic model Bouhaddou2018 is re-written in open-source and structured file format. The steps of model construction include input file creation and conversion into an SBML file. The optional initialization step calibrates model parameters for new cellular contexts and phenotypic behaviors. The annotated SBML model file and stochastic module are simulated together at single-cell level locally or by using cloud-computing. The benefits of the new SPARCED model include easy alteration and expansion capabilities through text file editing, human-readable annotated input files, and use of Jupyter notebooks for model creation and simulation. The modeling pipeline introduced here are inline with good practices of re-usable big mechanistic models (57). (c) The Bouhaddou2018 model file types are simplified and converted into open-source platforms.

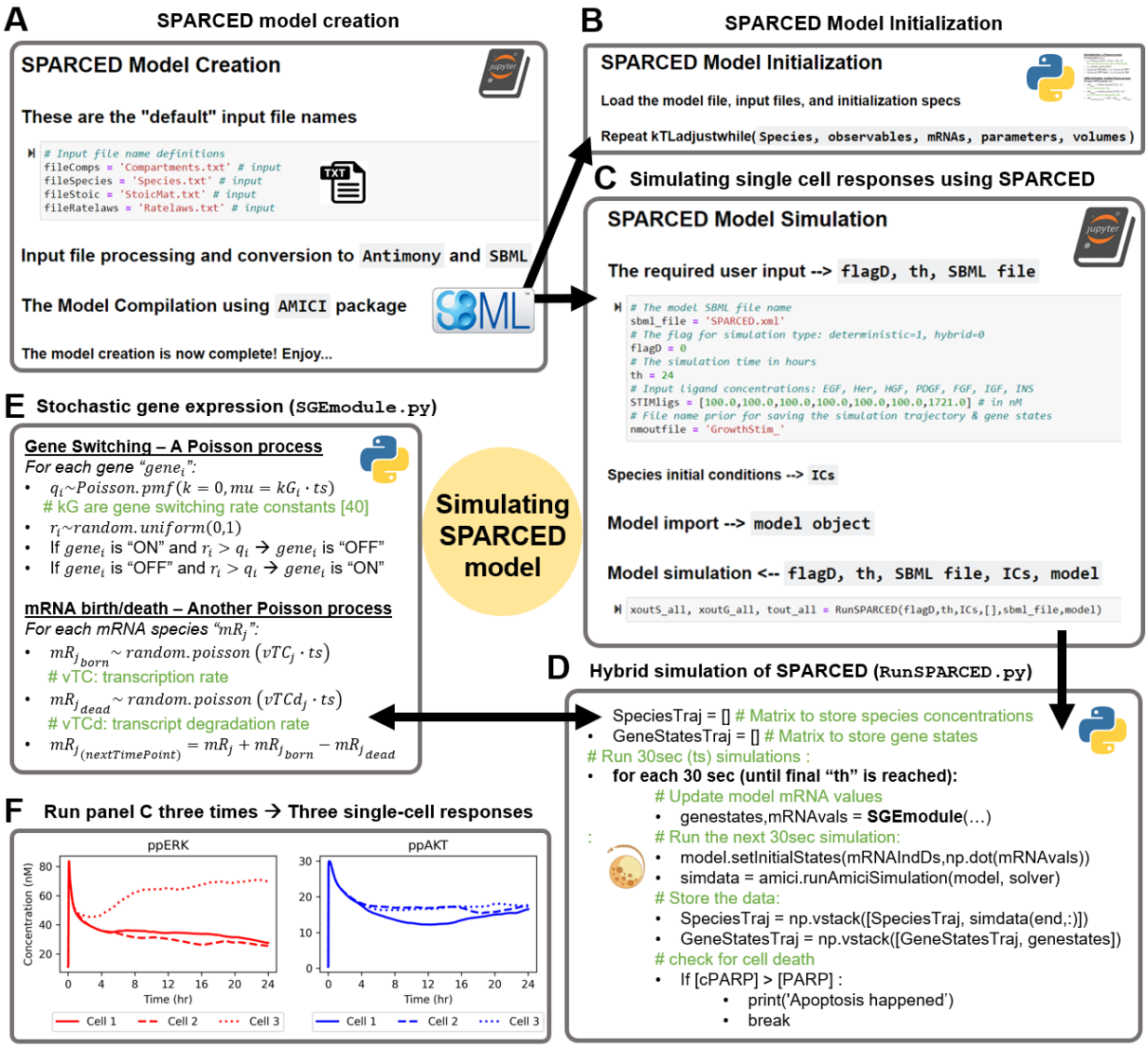


Figure 2.2: SPARCED-jupyter enables single-cell response simulations using Jupyter Notebooks.

Figure 2.2: SPARCED-jupyter enables single-cell response simulations using Jupyter Notebooks. (A) The model creation notebook processes the user defined input files and converts them into the model SMBL file. The model (SBML file) is compiled for simulations using the AMICI python package. (B) When a model is generated for a new cellular context (using new omics input data), the model creation step is followed by an initialization step to adjust protein translation rate constants and cell death related parameters. (C) The model simulation starts with specifying and importing the SPARCED model SBML (see panel A). The user defines the model file name and the sets four additional parameters: (i) The flag (1 or 0) to specify if the model should run in deterministic or in hybrid mode (see D and E), respectively. (ii) The time duration in hours for which the model should run. (iii) The vector of ligand concentrations (in nM) to stimulate the cells. (iv) The output file name. Next, the species initial conditions are, by default, read-in from the “Species” input file. Then, the model file is imported, and the model is simulated according to the specified input. The model outputs three matrices of species concentrations over time at every 30seconds, the activation states of genes over time (every 30 seconds), and the time points of simulation in seconds. The two former matrices are saved using the user define file name (iv). (D) The model is simulated iteratively for each 30 seconds, where the current species concentrations are inputs for the gene expression module, which then outputs new mRNA levels to update the SBML model states. The model is then run for another 30 seconds, until the total simulation time reaches the user input (th) or until the cell dies. The cell death is decided based on cleaved-PARP levels surpassing the PARP levels. (E) In the gene expression module, in hybrid mode, the model randomly decides which genes become active or inactive, and which mRNAs are transcribed or degraded. This SGEmodule.py script is called every 30 seconds with updated species concentrations, simulated using the models SBML with AMICI package. (F) When the model is hybrid-simulated three times, the different cell responses are observed. Shown are serum-starved average cells stimulated with full growth media for 24 hours. Plotted are free ppERK and ppAKT species concentrations (nM).

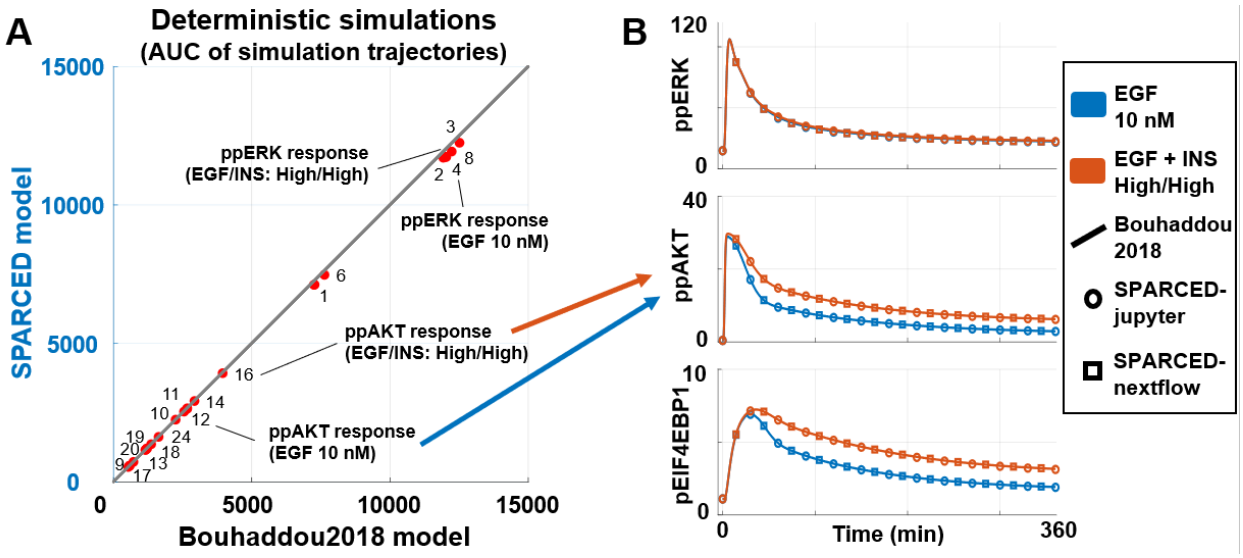


Figure 2.3: SPARCED model recapitulates deterministic simulation results of the Bouhaddou2018 model. (A) Summary of comparisons of SPARCED model deterministic simulations to Bouhaddou2018 model simulations. The area under the curve (AUC) values of each simulation (see Fig. 2.14) are calculated and plotted for the two model results. (B) Simulation results from Bouhaddou2018 model (line) and SPARCED-nf model (square) run on Kubernetes cluster workflow are the same as SPARCED model (circle) results. Comparisons of selected panels from (A) are shown only.

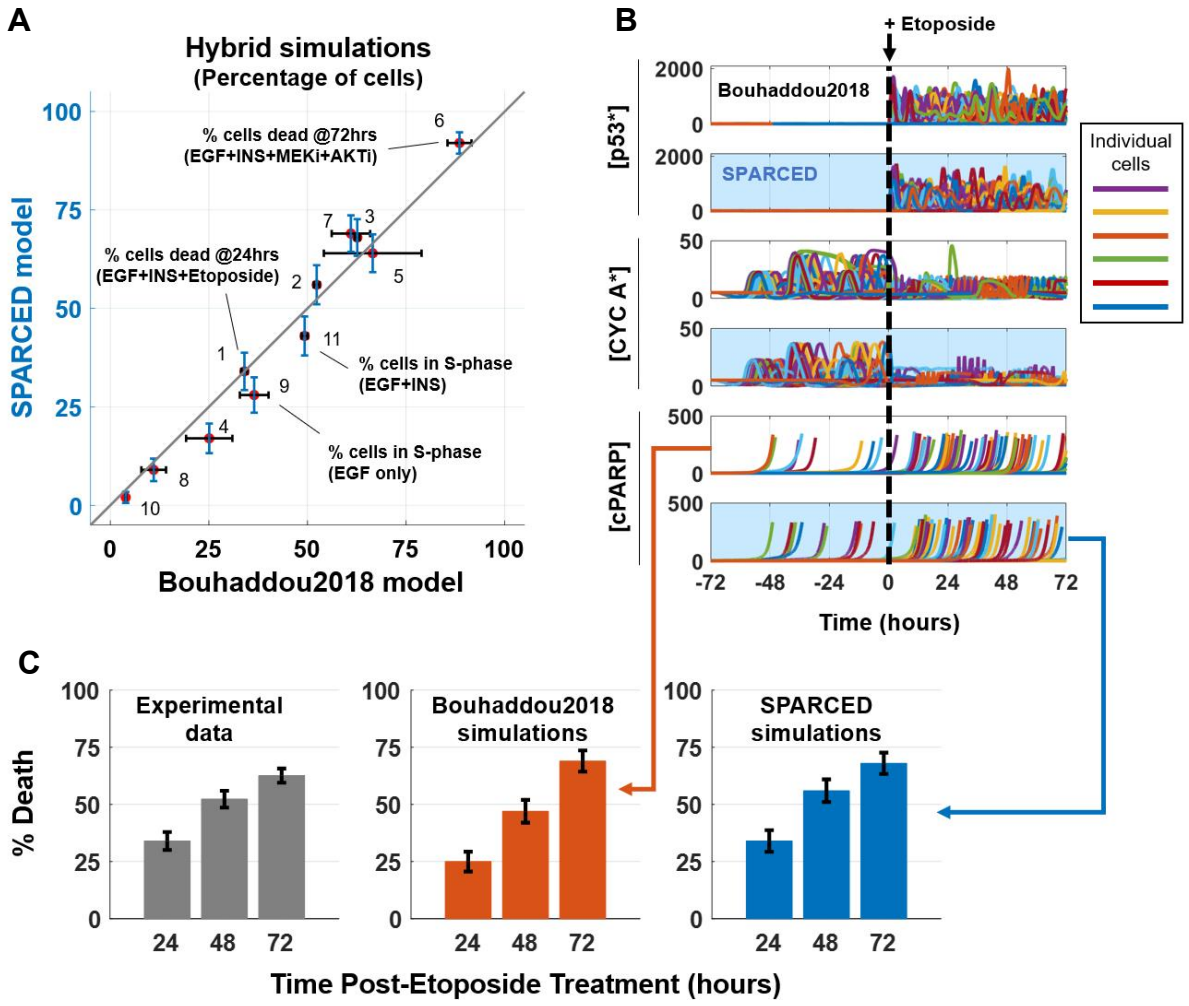


Figure 2.4: SPARCED model recapitulates experimental observations and hybrid (stochastic) simulation results of the Bouhaddou2018 model. (A) Experimental and stochastic simulation results from Bouhaddou 2018 model are reproduced by SPARCED model simulations. Each dot is a different condition, explained in Figure 2.16A. Error bars show experimental or simulation standard error of the mean. Simulations are of at least 100 cells, and three independent experimental observations where applicable. (B) Stochastic simulation of 100 cells recapture protein level trajectories (active p53, Cyclin A, and cPARP) from older model qualitatively. Panels with blue background are SPARCED simulations and white background panels are from Bouhaddou2018 model. 100 stochastic cells are stimulated with EGF+Insulin for 72 hours before Etoposide treatment for another 72 hours. Etoposide is stimulated also with EGF+Insulin. Results for Etoposide treatment without prior growth factor stimulations are shown in Figure 2.16B. (C) Quantification of results in (B) shows that SPARCED model simulations coincide with earlier observations in percentage of death induced by etoposide treatment. See Figure 2.16C for the effect of no growth factor stimulation before Etoposide treatment. Bars represent mean ± s.e.m.

2.3 Results:

2.3.1 SPARCED Model Construction and Unit Testing

Current large-scale mechanistic models are agglomerates of smaller models and tools, used mainly within the same research lab. Most such models also lack clear and satisfactory annotation and metadata, making them harder to understand and alter[32,58]. The goals of this work were (i) to build tools that help large-scale mechanistic model construction and alteration, that is simple, efficient, open-source, and cloud computing compatible; and (ii) to provide a scalable and re-useable big mechanistic model for a single mammalian cell.

We first created a set of simple input files and scalable processing scripts for one of the broadest cancer signaling models in the literature[32], called the Bouhaddou2018 model here (Fig. 2.1a). The input files are simple tab-separated text files (Fig. 2.1b-c), unlike licensed file formats with a mixture of hard coded information in multiple interconnected scripts commonly used in modeling literature. A Jupyter notebook processes the input files into an Antimony text file. The model creation code generates the SPARCED model file in SBML format using the Antimony text file and annotations from the model input files. When the model construction step is complete and the SBML file is created, it is imported and simulated using a Python package called AMICI[102,103]. For every new cell line model, a pre-calibration step called “Initialization” is employed to tune parameter values. Here, we ensure total protein levels match experimental observations and particular phenotypic criteria are met; for example, we specify that serum and growth factor starved cells on average do not traverse the cell cycle and do not die by apoptosis within 48 hours. The resulting initialized parameter values and

species concentrations are saved in a new SBML file, and the model is compiled for model testing and other simulations.

The result is what should be a replica of the Bouhaddou2018 model, which we call SPARCED. Like the Bouhaddou2018 model, the initial SPARCED model is based on non-transformed breast epithelial MCF10A cell line data. We annotated all the species in the model with HGNC gene identifiers, providing easier programmatic filtering and curation of species list, while keeping the user defined simpler names for complicated species structures. However, the extent to which the models are congruent was not yet clear, and thus we next set out to examine agreement between the two. We verified that the previous Bouhaddou2018 model simulations are reproducible and match expected experimental observations through the same unit test concept (Table 2.1) introduced for the original model[32] (40). Each unit test has a dedicated Jupyter notebook on GitHub repository (github.com/birtwistlelab/SPARCED/SPARCED_Brep). We illustrate select unit testing examples below, and all results are presented in Figures 2.3-2.16).

Table 2.1: List of SPARCED model unit testing and comparisons to Bouhaddou2018 model. The SPARCED model passed each test depicted below and recapitulated experimental and simulation observations reported by the Bouhaddou2018 model.

Descriptions of unit tests	Simulation type	Figure #	Original paper Figure #
Functional test to ensure the deterministic module is updated every 30 seconds with mRNA numbers generated by the stochastic module.	Hybrid	2.2	2B
Simulated ligand-receptor cooperativity coefficients for the receptor tyrosine kinases match experimental observations (negative cooperativity: EGF, FGF, IGF, INS; no cooperativity: HGF, NRG1, and positive cooperativity: PDGF).	Deterministic	2.3a	3A + S3A
Activated EGF receptors internalize and peak ~30 minutes after ligand treatment.	Deterministic	2.3b	S3B
EGF and insulin stimulation activates both ERK and AKT pathways. Dual stimulation with the two ligands induces prolonged AKT activation.	Deterministic	2.4-2.5	3B,C,D + S3C
Double and/or single stranded DNA damage activates p53 and DNA damage repair mechanisms represses its response.	Deterministic	2.6a	3E
Increasing DNA damage amount in single cells leads to higher number of activated p53 peaks.	Hybrid	2.6b-c	3F + S3E
Increasing simulated TRAIL dose decreases the time it takes to die for an average cell.	Deterministic	2.7a-b	3G
The fraction of surviving cells decreases as stimulated TRAIL dose increases.	Hybrid	2.7c	3H
Increasing ERK and AKT activity levels prolongs TRAIL induced time to death, whereas increasing PUMA and NOXA expression levels decreases the time it takes for cells to die.	Deterministic	2.7d	3I
Increasing Cyclin D mRNA levels induces proper cyclin-CDK complex progression and oscillations for cell cycle entry and progression.	Deterministic	2.8a	3J
Etoposide treatment induces cell cycle arrest and cell death. Cycling cells (with prior growth factor stimulation) show increased percentage of death to etoposide treatment, compared to non-cycling cells.	Hybrid	Fig. 3d-e	4A,B,C

Descriptions of unit tests	Simulation type	Figure #	Original paper Figure #
Inhibition of AKT and ERK pathways together synergistically increase cell death, in EGF and insulin stimulated cells.	Hybrid	Supp. Fig. 9	5A
ERK and AKT inhibition induced cell death mechanisms are predominantly BIM dependent, not BAD dependent.	Hybrid	Supp. Fig. 10a	5C
EGF and insulin cooperatively induce cell cycle entry, with insulin inducing very little cell cycle entry alone.	Hybrid	Supp. Fig. 10b	6B
Activation of both ERK and AKT pathways are required for robust cell cycle entry. Time averaged ppERK and ppAKT levels correlate with Cyclin D levels.	Deterministic	Supp. Fig. 11	6E
The number of ribosomes within the cell doubles within 24 hours.	Deterministic	Supp. Fig. 8b	S2D

2.3.2 SPARCED Model Simulation

Before presenting particular unit test applications, we wanted to provide an overview of model simulation. We built a Jupyter notebook called runModel.ipynb to simulate the SPARCED model (Fig. 2.2). This notebook requires the model SBML (from createModel.ipynb, Fig. 2.2a), along with the simulation duration (th), the ligand concentrations (if desired), the name for the output files, and whether the simulation should be deterministic only or hybrid mode (flagD). The “Initialization” calibration step is employed only when the model is being trained for a new set of omics data or for different phenotypic criteria (Fig. 2.2b). The rest of the runModel.ipynb notebook imports necessary packages and model files and runs the simulation (Fig. 2.2c).

As mentioned, the SPARCED model consists of two modules: deterministic and stochastic. The SBML file forms the basis of the deterministic module whereas the stochastic module describes gene states (active/inactive) and mRNA birth/death events for the genes (Fig. 2.2c). When run in the hybrid simulation mode, the deterministic and stochastic modules exchange information every 30 simulated seconds (Fig. 2.22d-e). The current levels of select protein states can induce changes in gene activation/deactivation or mRNA transcription/decay rates. The newly updated mRNA copy numbers change nascent protein translation rates in the deterministic module (Fig. 2.2d). When run deterministically, the model does not stochastically sample gene activation or mRNA transcription events, and such simulations correspond to an average cell state.

Individual cells (in vitro on a dish or in vivo) exhibit mRNA and protein expression variability, in part due to stochastic gene expression processes[81,82]. To capture this

phenomenon in silico, we ran simulations in hybrid mode. In this mode, each simulation has different initial mRNA and protein levels that are dictated by burst like expression processes, and the expression throughout the simulated time course follows suit. This leads to a natural and typically observed amount of variation in total protein levels. We hereafter refer to such settings and resulting trajectories as single-cell simulations. Virtual cell population responses are sets of multiple independent single-cell simulations, usually 100 cells. So, when the runModel.ipynb notebook is run multiple times in hybrid mode, different single-cell responses are simulated (Fig. 2.2f). For instance, the activation and phosphorylation of ERK (Fig. 2.2f left, red lines) and AKT (Fig. 2.2f right, blue lines) proteins in response to growth factor treatment will show variability across three example cells. Although the amplitude of initial response is similar for all three cells, the longer-term responses are quite different. Our previous analyses showed that such single-cell heterogeneity in the initial concentrations of these proteins could help predict cellular fate, namely cell division[32]. These jupyter notebooks provide a simple interface to interact with the SPARCED model.

2.3.3 SPARCED Model Unit Testing - Deterministic

We first tested agreement between deterministic Bouhaddou2018 and SPARCED model simulations. The SPARCED model simulations recapitulated the response of an average (deterministic) cell under different stimulation conditions, to within simulation error (Fig. 2.3a). As an example, we highlight SPARCED model simulations of the cell response (MCF10A cells) to treatment with EGF alone or EGF+insulin (Fig. 2.3b and Fig. 2.15). Treating growth factor and serum-starved MCF10A cells with EGF and insulin induces activation of ERK, AKT, and their downstream signaling partners, which

together influence cell proliferation[32,104,105]. The Bouhaddou2018 model showed that compared to single ligand treatments, EGF+insulin stimulation increases and prolongs AKT and its downstream EIF4EBP1 phosphorylation (Fig. 2.3b). The simulation results from the Bouhaddou2018 model (the solid lines) and SPARCED model (circles) are indistinguishable. The SPARCED-nf implementation, which runs on a high-performance cloud computing infrastructure, similarly reproduces the original simulation data (Fig. 2.3b, triangles). These results, together with all other deterministic tests in Table 2.1 (Figs. 2.5-11 and 14), confirm that the SPARCED model recapitulates the Bouhaddou2018 model simulations and unit tests in deterministic settings. Thus, the simple input file structure combined with automatic model generation is equivalent to the prior MATLAB instantiation in this regard.

2.3.4 SPARCED Model Unit Testing - Stochastic (Hybrid)

Next, we evaluated the SPARCED model for stochastic unit tests in single cell simulations. Each single simulated cell has different initial protein levels and dynamics due to stochastic gene expression, and thus may respond differently to the same treatment. A simulated cell population is a collection of multiple single cell simulations, usually 100 unless otherwise noted. The SPARCED model stochastic simulations closely matched Bouhaddou2018 model results, to within simulation error (Fig. 2.4a and Fig. 2.16a). As an example, we highlight here how single cells respond stochastically to DNA damage. Etoposide, a chemotherapy drug, induces double- and single-stranded DNA damage, causes cell cycle arrest, and leads to cell death (72). Previous experimental data[32] showed that in the absence of EGF and insulin (to promote cell cycle exit), there is minimal etoposide-induced cell death (Fig. 2.16b-c). However, in the

presence of EGF and insulin (to drive cell cycle progression), etoposide-induced cell death increases over time and reaches around 60% of the cells (Fig. 2.4b-c). Simulating etoposide treatment of cycling cells induces robust p53 pulses, disruption of Cyclin A dynamics/cell cycle arrest (Fig. 2.4b), and more cell death relative to non-cycling cells (Fig. 2.4c). The SPARCED simulation results closely match experimental data and Bouhaddau2018 simulations. We conclude that SPARCED model captures DNA damage induced single-cell death percentage and cell cycle state-dependent effect of etoposide. The SPARCED model also passed all other stochastic/hybrid unit tests (Table 2.1, Supplementary Figs. 2.5, 2.9, 2.10, 2.12, and 2.13).

2.3.5 SPARCED Model Unit Testing - Context Change

Different cell types have different mRNA and protein expression levels, and many mechanistic models assume that it is different expression levels that drive different phenotypes, as opposed to changes in biochemical rate constants. These constants are based on biophysical events like binding, which are based on molecular structures. Here, we tested the ability of the SPARCED model to be re-“initialized” to study different cell types by changing initial levels of total proteins and mRNAs without changing the model topology. Thus, we introduced a protocol to enable SPARCED model context change (Fig. 2.17). In short, OmicsData, Species, and Ratelaws input files are updated with new cell line information, including mRNA levels, protein/species levels, and constitutive translation rate constants. Then, the new model is created by running the “createModel” Jupyter notebook or by submitting a new SPARCED-nf job.

The re-calibration step for context change followed in Bouhaddou2018 model was called Initialization, where protein-specific translation rates and key parameters

important for cell decision making are estimated to ensure agreement with new omics datasets and expected phenotypic behavior with respect to proliferation and apoptosis. Here we also provide a new, python-based version of the Initialization procedure for SPARCED models (see Computational Methods), where the outputs are species concentrations and rate parameter values updated in a new SBML file. Here, to test the drug combination response differences in different cell lines, we changed SPARCED model context (i.e., parameter values and species concentrations) by initializing the model to the U87 glioma cell line. Following the protocol outlined in Supplementary Fig. 2.17, we replaced MCF10A cell line values in the input files with values from U87 cell line data.

U87 cells are PTEN-deficient and more sensitive to AKT inhibition compared to MCF10A cells[32]. Both cell lines show minimal sensitivity to MEK inhibition alone and AKT & MEK inhibitors are both needed to kill MCF10A cells. In contrast, AKT inhibition alone is sufficient to kill U87 cells. To simulate the U87 cell response to AKT and MEK inhibitors, we first updated the OmicsData input file using U87 mRNAseq data from Bouhaddou2018 model. Here, we did not use U87 cell line proteomic data and estimated the initial total protein levels using the new mRNA levels and gene-level mRNA/protein ratios from MCF10A data. We set the PTEN translation rate to zero and set values of rate parameters dictated by Initialization in the Ratelaws input file (Fig. 2.18A). Additionally, we provide an improved Python based initializer initializeModel.ipynb notebook, which re-creates the un-stimulated steady-state initial conditions for species and adjusts translation rate constants using cell-line specific initialization input file. We also updated the species initial conditions in the Species input

file using steady-state values for U87 cells from the Bouhaddou2018 model. We created a new model SBML file (SPARCED_U87) using the updated input files. SPARCED_U87 model simulations of response to MEK and AKT inhibitors reproduced the Bouhaddou2018 model results and experimental observations (Fig. 2.18B). We conclude that changing model context by changing input files is possible and contributes towards the goal of easy model alteration to study of different cell types.

When the cellular context (omics input data) for the SPARCED model is changed, all appropriate Unit Tests should be passed. We expect that addition and alteration of the list provided (Table 2.1) will accommodate increasingly different prior knowledge about the new context. Examples of such information include cell line mutations, growth condition differences, or tumor cell behavior.

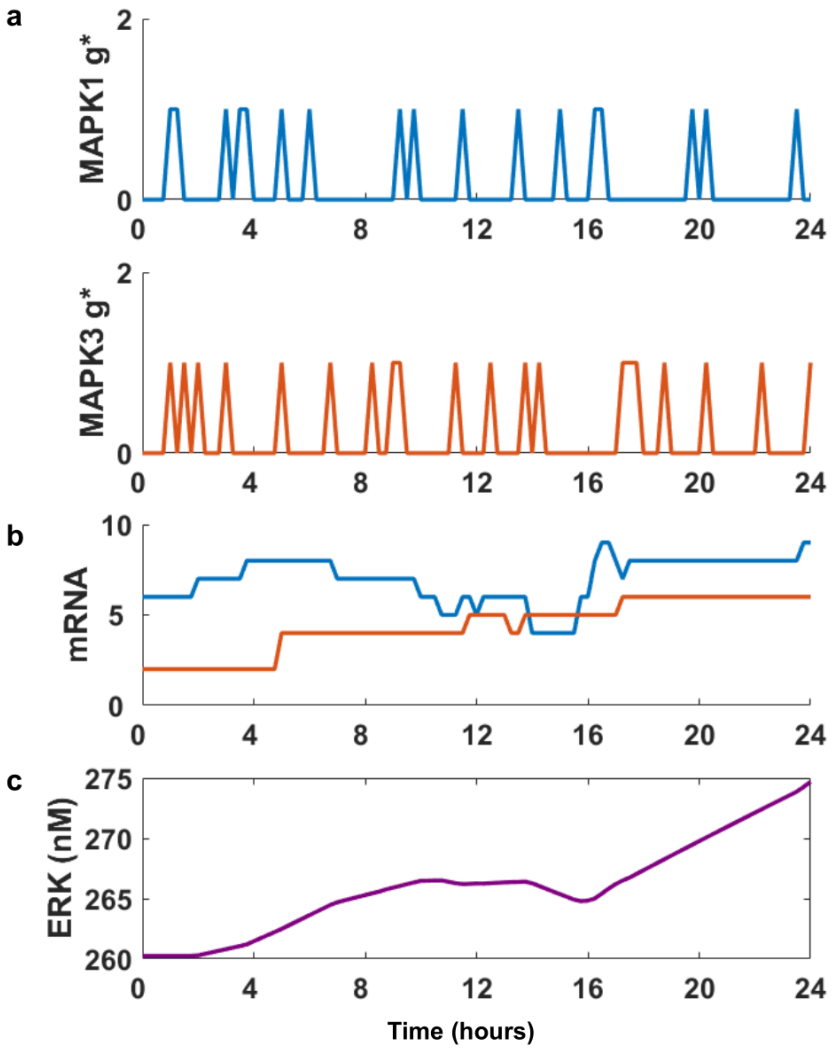


Figure 2.5: SPARCED model includes a stochastic gene expression module. (a) Two isoforms of ERK gene (MAPK1 and MAPK3) are activated randomly. (b) It leads to two distinct mRNA species. (c) The ERK1 and ERK2 mRNAs are translated into a single ERK protein. The trajectories are obtained from a stochastic single cell simulation with now growth factor for 24 hours.

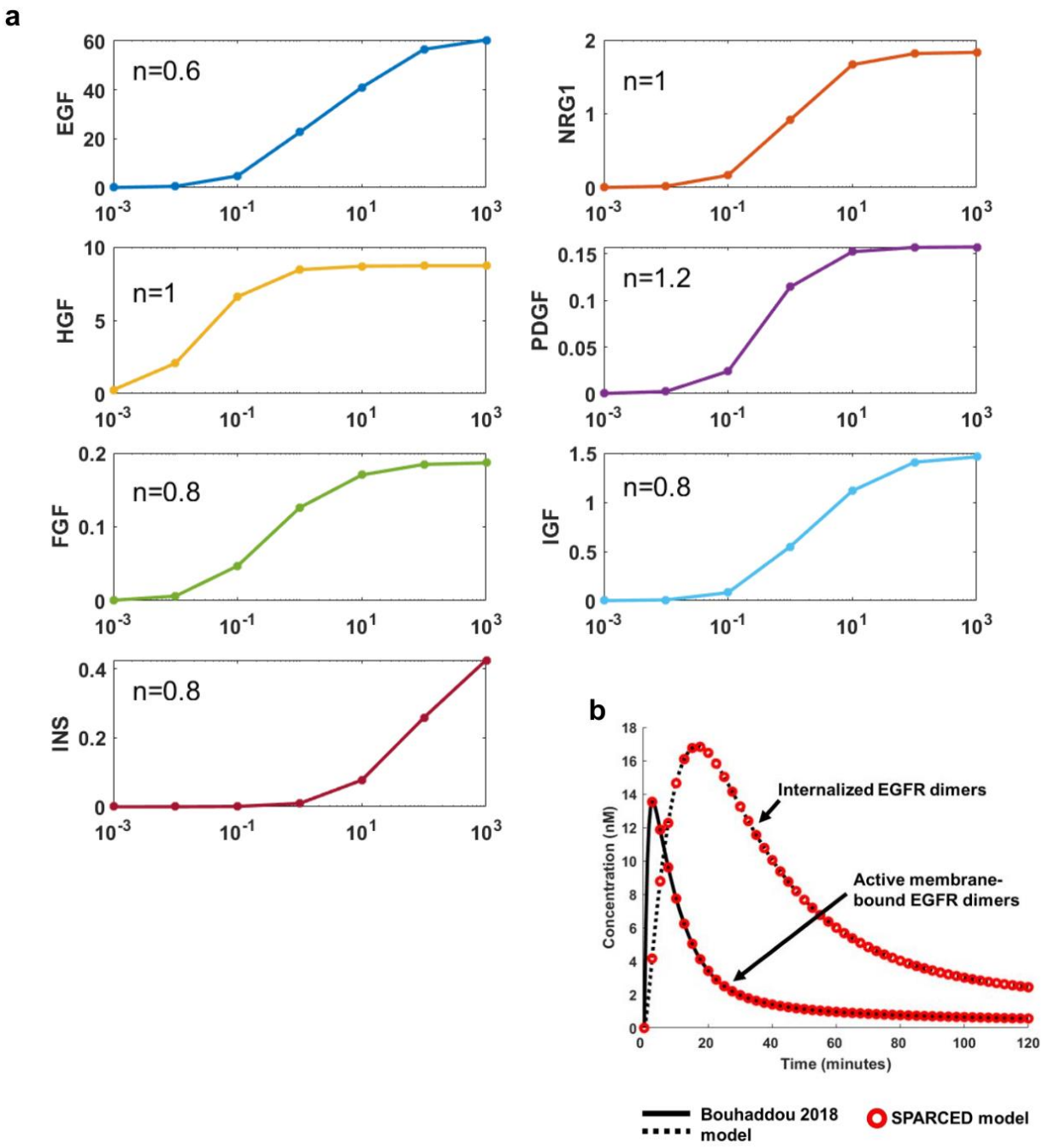


Figure 2.6: SPARCED model recapitulates ligand-receptor cooperativity observations (a) Hill coefficients for each ligand-receptor pair in MCF10A context. The simulations capture literature knowledge. (b) The dynamics of activated EGFR (membrane-bound and internalized) dimers are recaptured by the SPARCED model, compared to Bouhaddou2018 model.

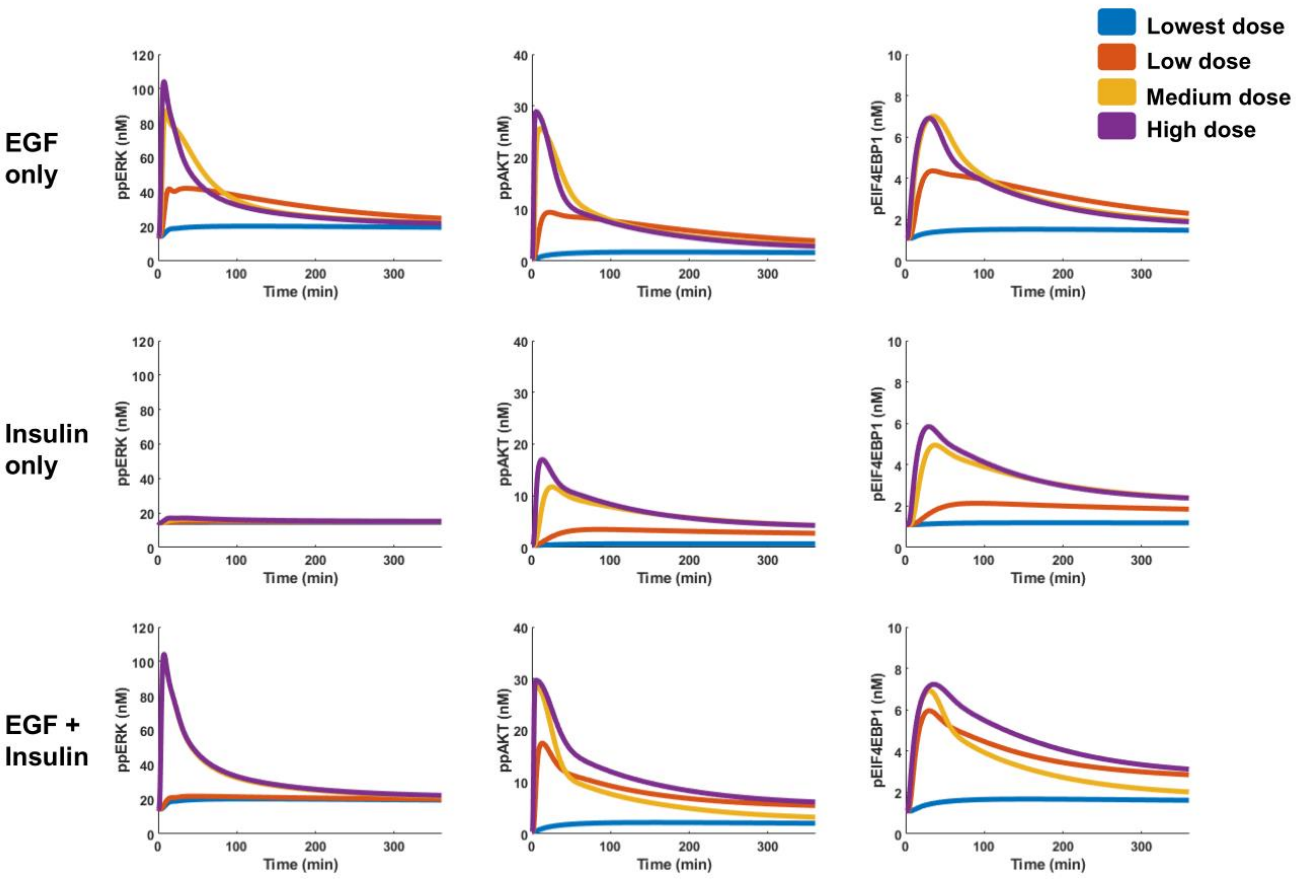


Figure 2.7: Model response to EGF and insulin. Signaling dynamics of ppERK, ppAKT, and pEIF4EBP1 induced by EGF, Insulin, or EGF+Insulin treatment for 6 hours. Serum-starved MCF10A cells stimulated with EGF (0.01, 0.1, 1.0 and 10.0 nM), Insulin (0.17, 1.7, 17.0, and 1721 nM), or EGF+Insulin (0.01+0.17, 10+0.17, 0.01+1721, and 10+1721 nM)

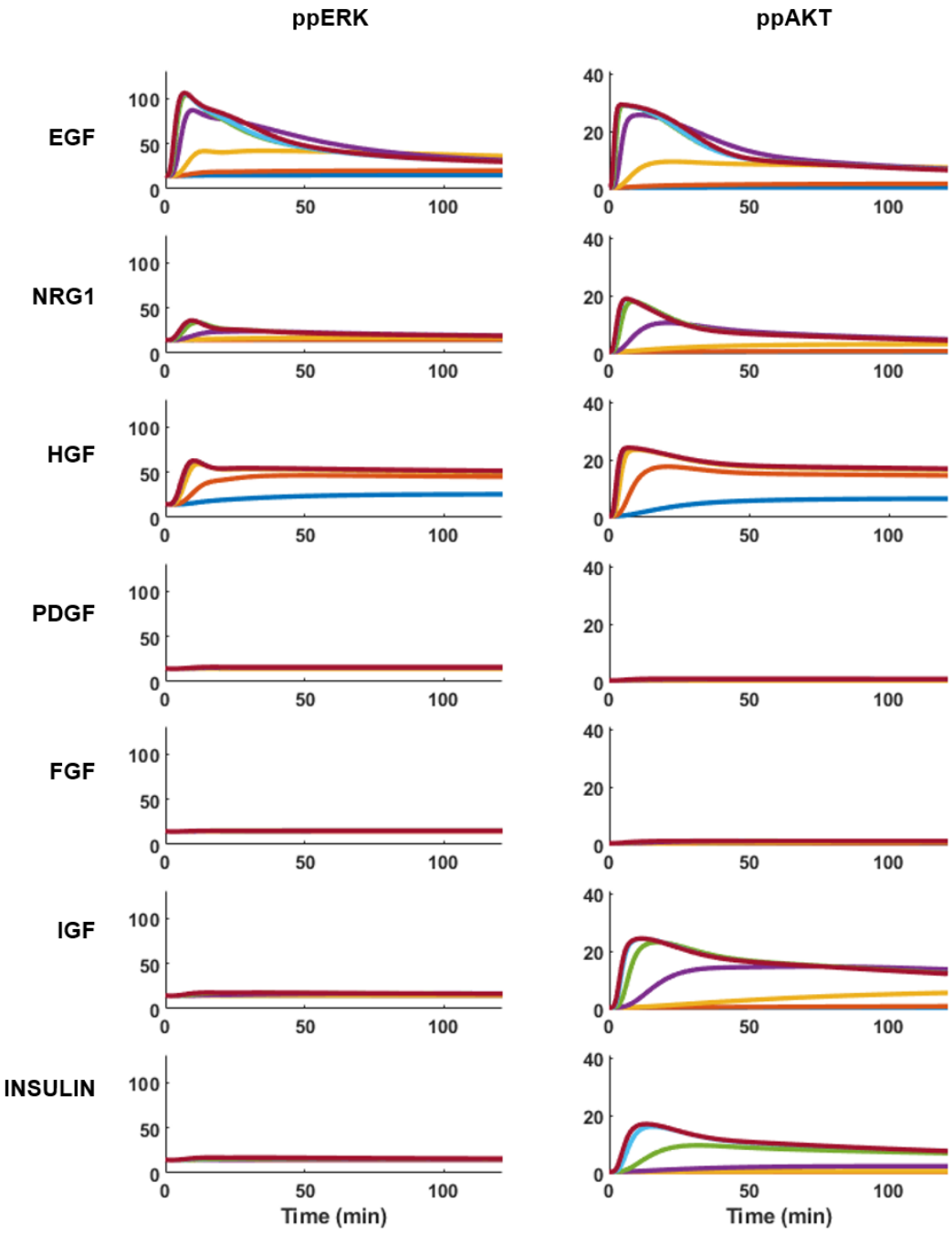


Figure 2.8: Signaling dynamics of ppERK and ppAKT induced by EGF, Heregulin (NRG1), HGF, PDGF, FGF, IGF, and Insulin treatment for 2 hours. Serum-starved MCF10A cells are stimulated with corresponding ligands at a dose range of 0.001 to 1000 nM.

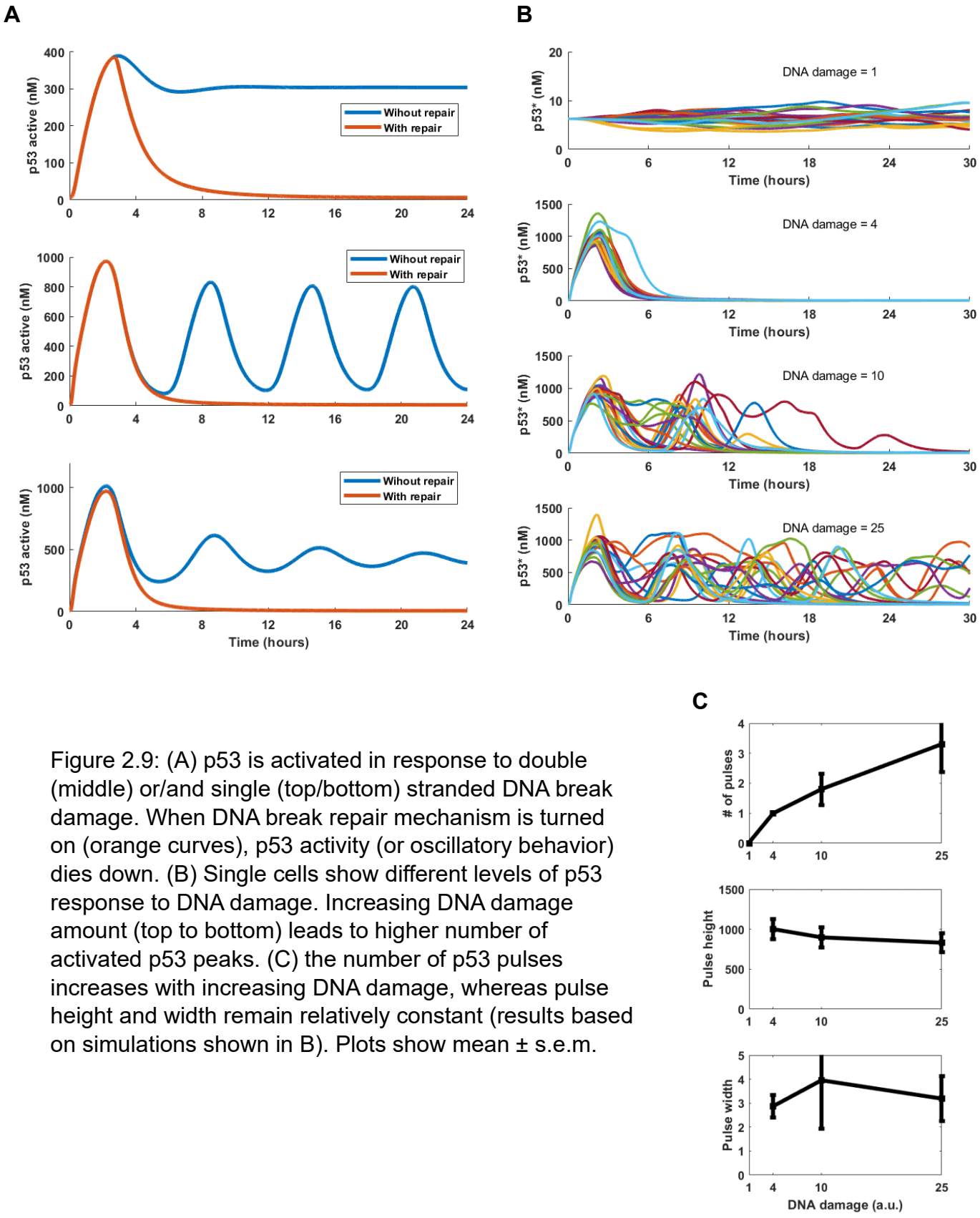


Figure 2.9: (A) p53 is activated in response to double (middle) or/and single (top/bottom) stranded DNA break damage. When DNA break repair mechanism is turned on (orange curves), p53 activity (or oscillatory behavior) dies down. (B) Single cells show different levels of p53 response to DNA damage. Increasing DNA damage amount (top to bottom) leads to higher number of activated p53 peaks. (C) the number of p53 pulses increases with increasing DNA damage, whereas pulse height and width remain relatively constant (results based on simulations shown in B). Plots show mean ± s.e.m.

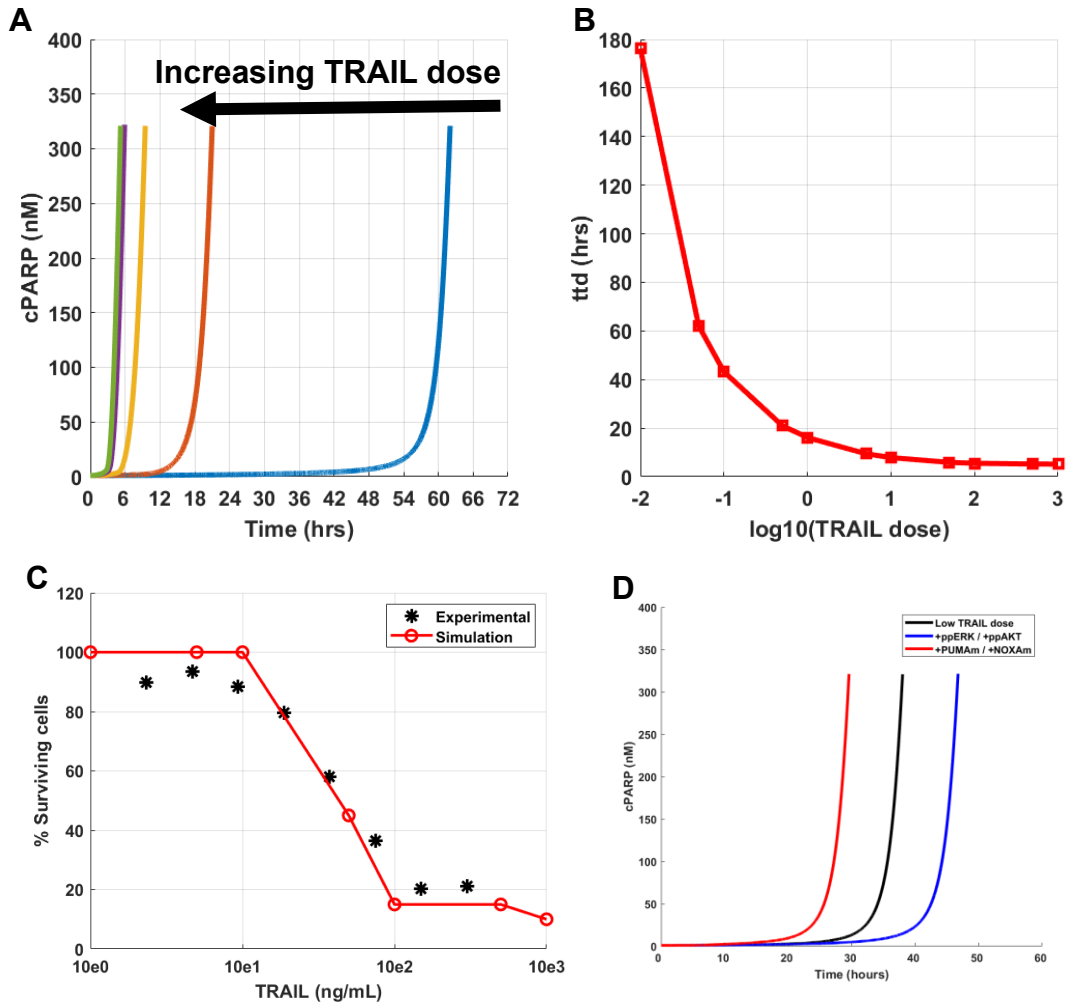


Figure 2.10: (A) Increasing TRAIL dose decreases the time it takes to die (ttd) for the average cell. Representative cells trajectories are shown, where the cells are simulated deterministically with different doses of TRAIL until they die (or up to 100 hours). The time of death is defined by the amount of cleaved PARP (cPARP, y-axis) surpassing the amount un-cleaved PARP. (B) Summary of ttd values for different TRAIL doses. (C) The fraction of surviving cells decreases as stimulated TRAIL dose increases. The red circles represent percentage of living cells when 20 stochastic single cells are simulated with specified TRAIL dosage for 5 hours. The black stars are experimental data from Bouhaddou 2018 model (D) Increasing ERK and AKT activity levels prolongs TRAIL induced time to death (blue curve), whereas increasing PUMA and NOXA expression levels decreases the time it takes for cells to die (red curve). Cells with specified alterations are compared to the cell stimulated with a low dose of TRAIL (black curve). cPARP levels are the proxy for cell death, where the cells go apoptosis when [cPARP]>[PARP].

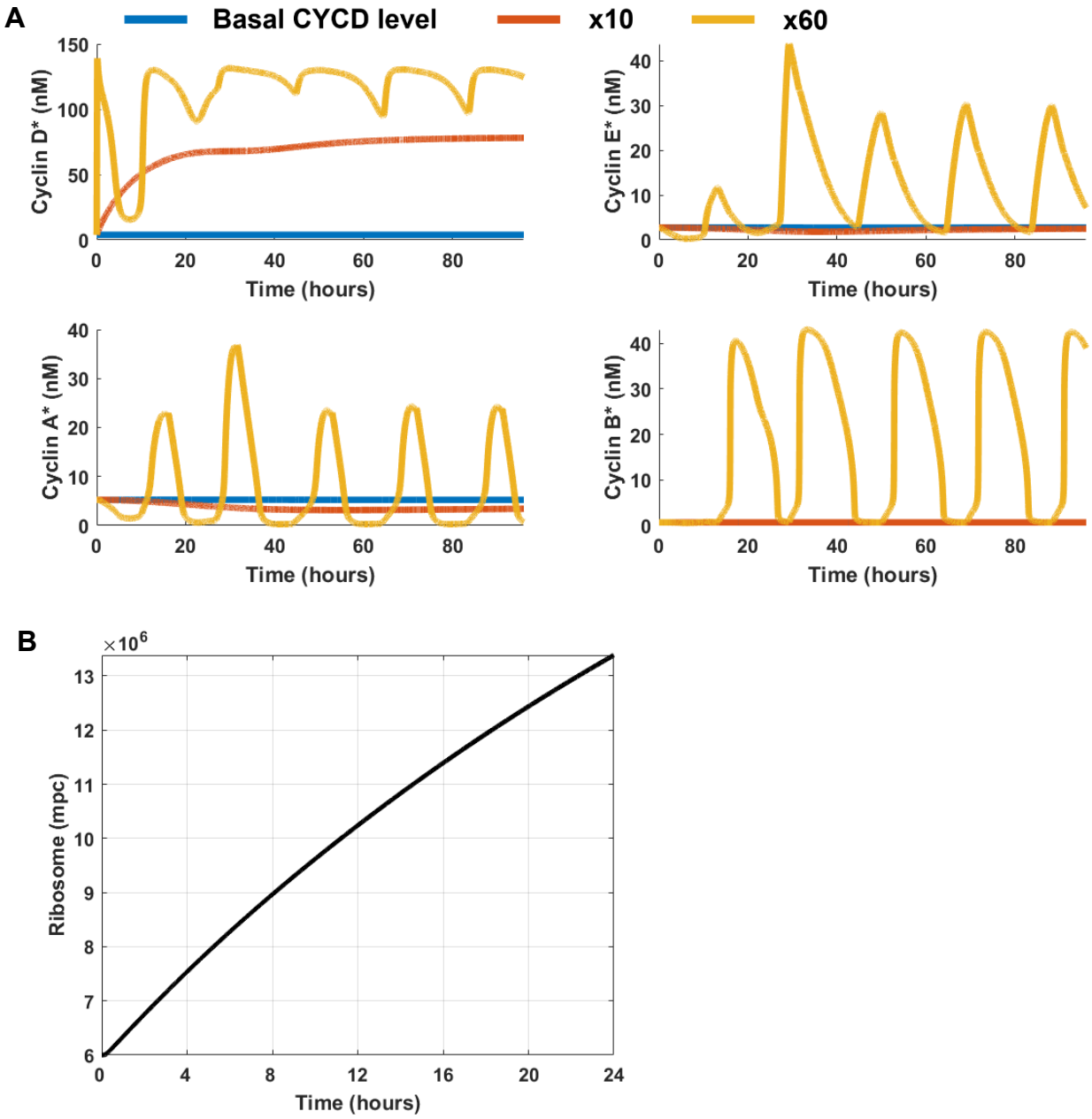


Figure 2.11: (A) Increasing Cyclin D mRNA levels induces proper cyclin-CDK complex progression and oscillations for cell cycle entry. Plots show Cyclin D, E, A, and B concentrations when basal (blue), 10X basal (dark orange), and 60X basal (light orange) levels of Cyclin D mRNA (CYCD) are simulated. (B) The number of ribosomes in the cell doubles around 20 hours. The cell is simulated with full growth condition (EGF=100 nM, NRG1=100 nM, HGF=100 nM, PDGF=100 nM, FGF=100 nM, IGF=100 nM, INS=100 nM).

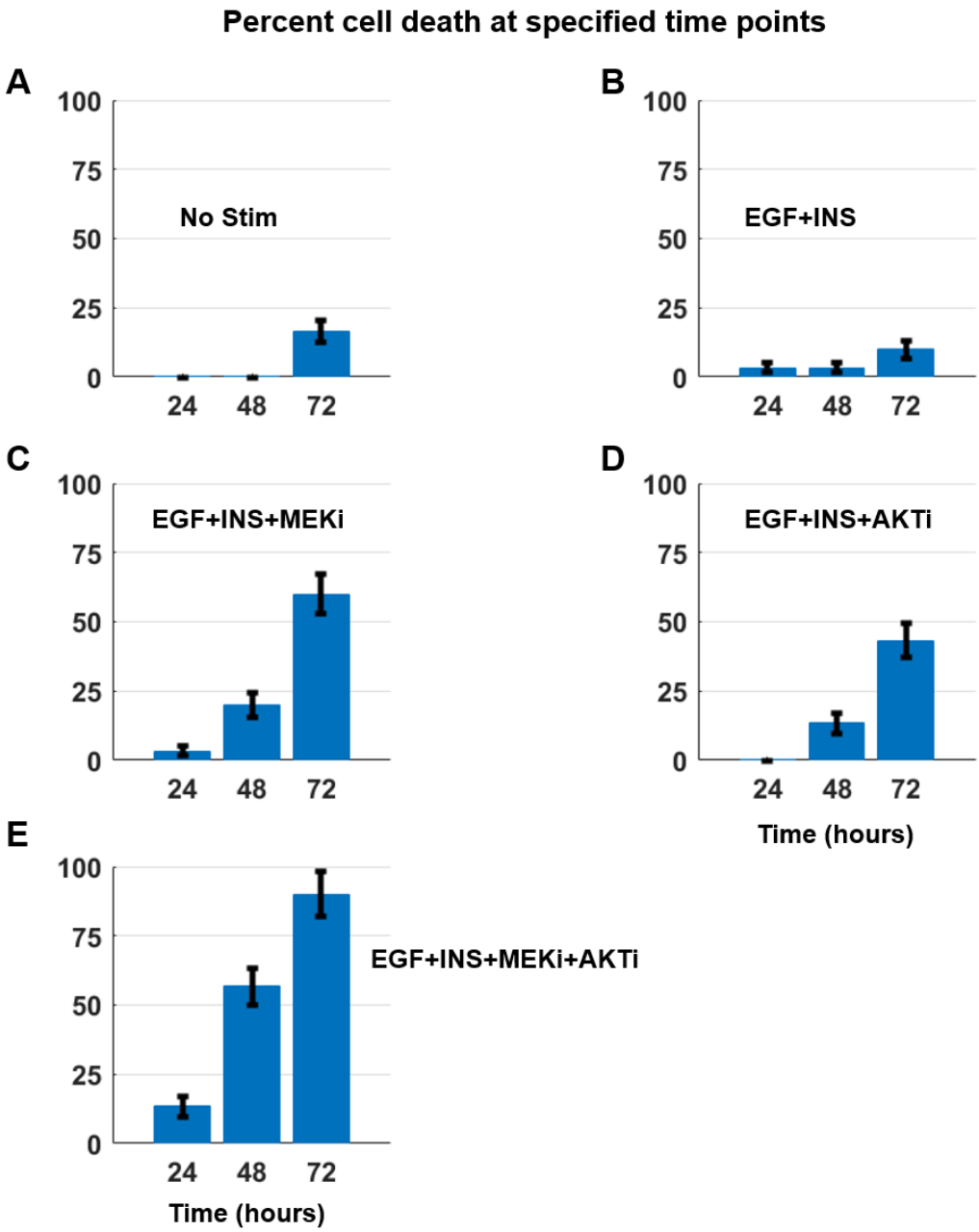


Figure 2.12: Inhibition of AKT and ERK pathways together synergistically increase cell death, in EGF and insulin stimulated cells. Serum-starved MCF10A cells are simulated with following conditions: (A) No stimulation, (B) EGF=20ng/mL + Insulin=10µg/mL, (C) EGF=20ng/mL + Insulin=10µg/mL + MEKi=10µM, (D) EGF=20ng/mL + Insulin=10µg/mL + AKTi=10µM, and (E) EGF=20ng/mL + Insulin=10µg/mL + MEKi=10µM + AKTi=10µM for up to 80 hours. The bar plots show mean ± s.e.m. of time to death for 30 cells. The ttd are captured by cPARP spikes.

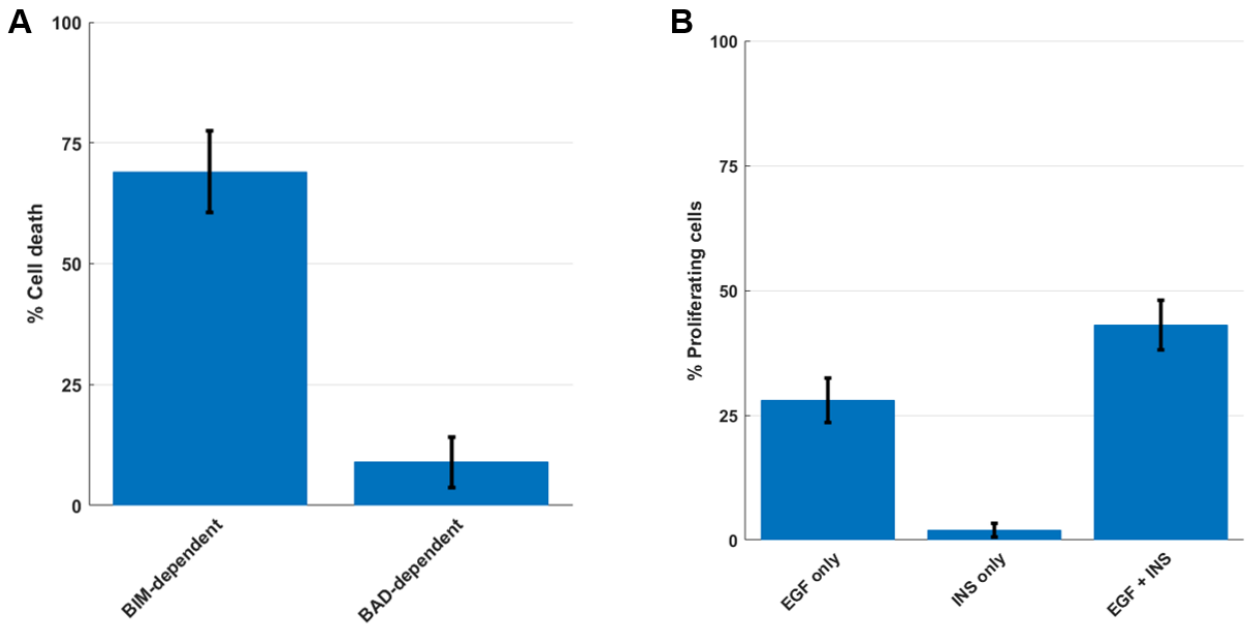


Figure 2.13: (A) Simulations where BIM-dependent or BAD-dependent mechanisms are switched off and percent death calculated in response to EGF + insulin at 48 hours. The results show that ERK and AKT inhibition induced cell death mechanisms are mostly BIM dependent, not BAD. Bars represent mean ± s.e.m. of 100 stochastic cell simulations. (B) EGF and insulin cooperatively induce cell cycle entry, with insulin inducing very little cell cycle entry alone. Cells are simulated with EGF (10nM), Insulin (1721nM), or EGF+Insulin (10nM+1721nM) for 30 hours and the percentage of cells entering S-phase are calculated. Cells are considered in S-phase when the sum of concentrations of Cyclin E, A, and B is greater than 20nM. Bars represent mean ± s.e.m. of 100 stochastic cell simulations.

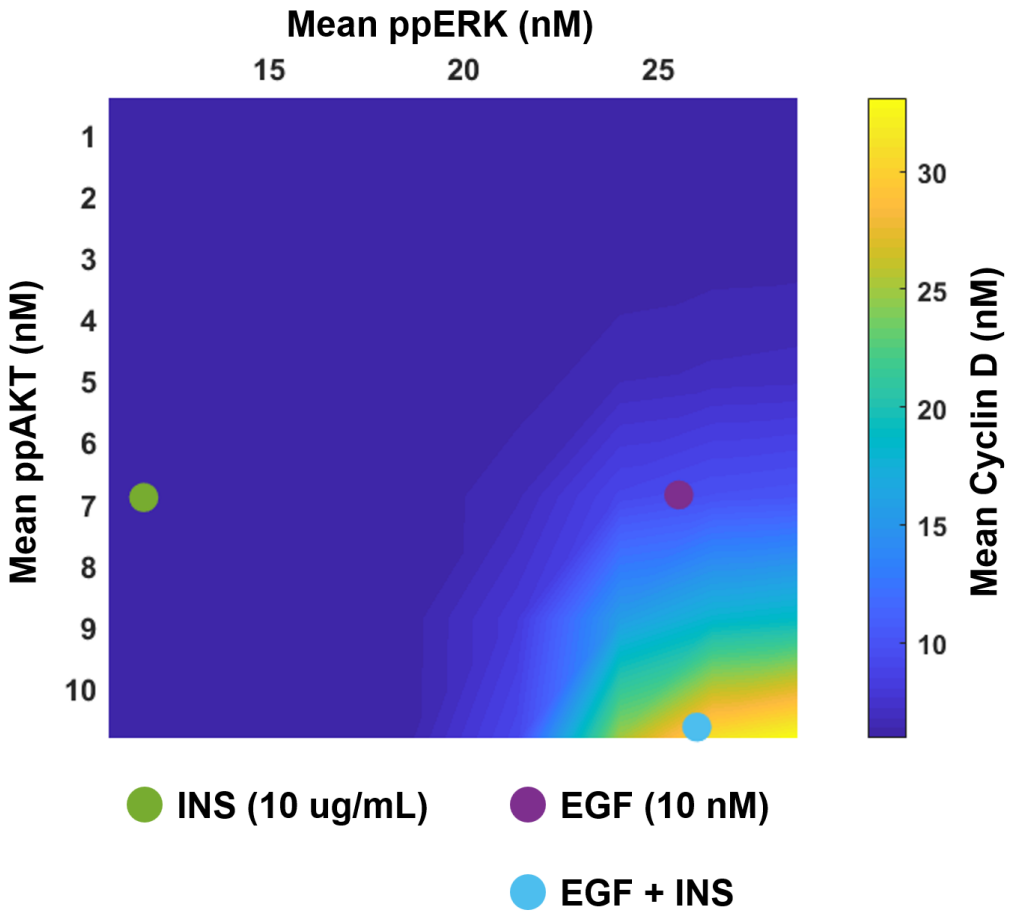


Figure 2.14: Activation of both ERK and AKT pathways are required for robust cell cycle entry. Time averaged ppERK and ppAKT levels correlate with Cyclin D levels. Basal levels of ppERK and ppAKT are increased (between 1X-20X) and each condition is simulated up to 6 hours. The time-averaged levels of ppERK and ppAKT are plotted against the time-averaged Cyclin D levels. Conditions representing EGF (10nM), Insulin (1721nM), and EGF+Insulin (10nM+1721nM) are shown with colored circles.

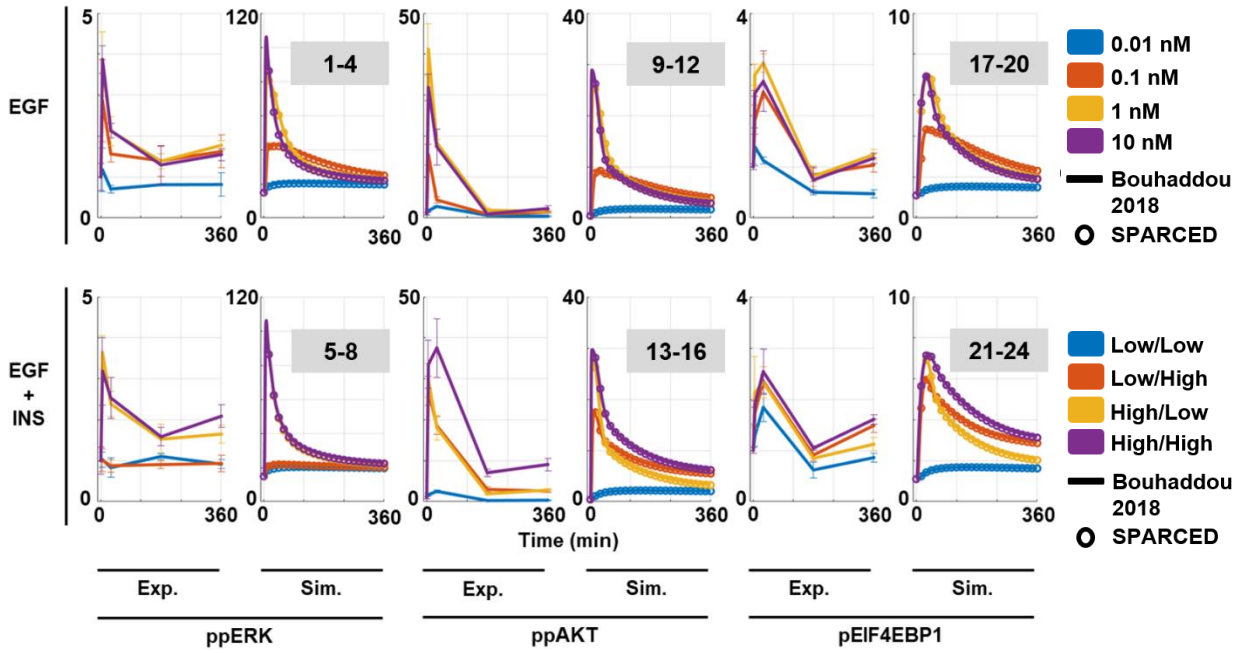


Figure 2.15: SPARCED model recapitulates downstream pathway activation by ligands and ligand combination treatments. Experimental data and simulation results from MATLAB (lines) and SPARCED (circles) models with EGF (top) and EGF+Insulin (bottom) stimulation for 6 hours. Plots show double-phosphorylated ERK (ppERK), serine-phosphorylated AKT (pAKT), and phospho-EIF4EBP1 (pEIF4EBP1) levels. . The numbers in gray shaded boxes represents numbering of conditions in Fig. 2.3A. Exp: Experimental data, Sim: Simulation.

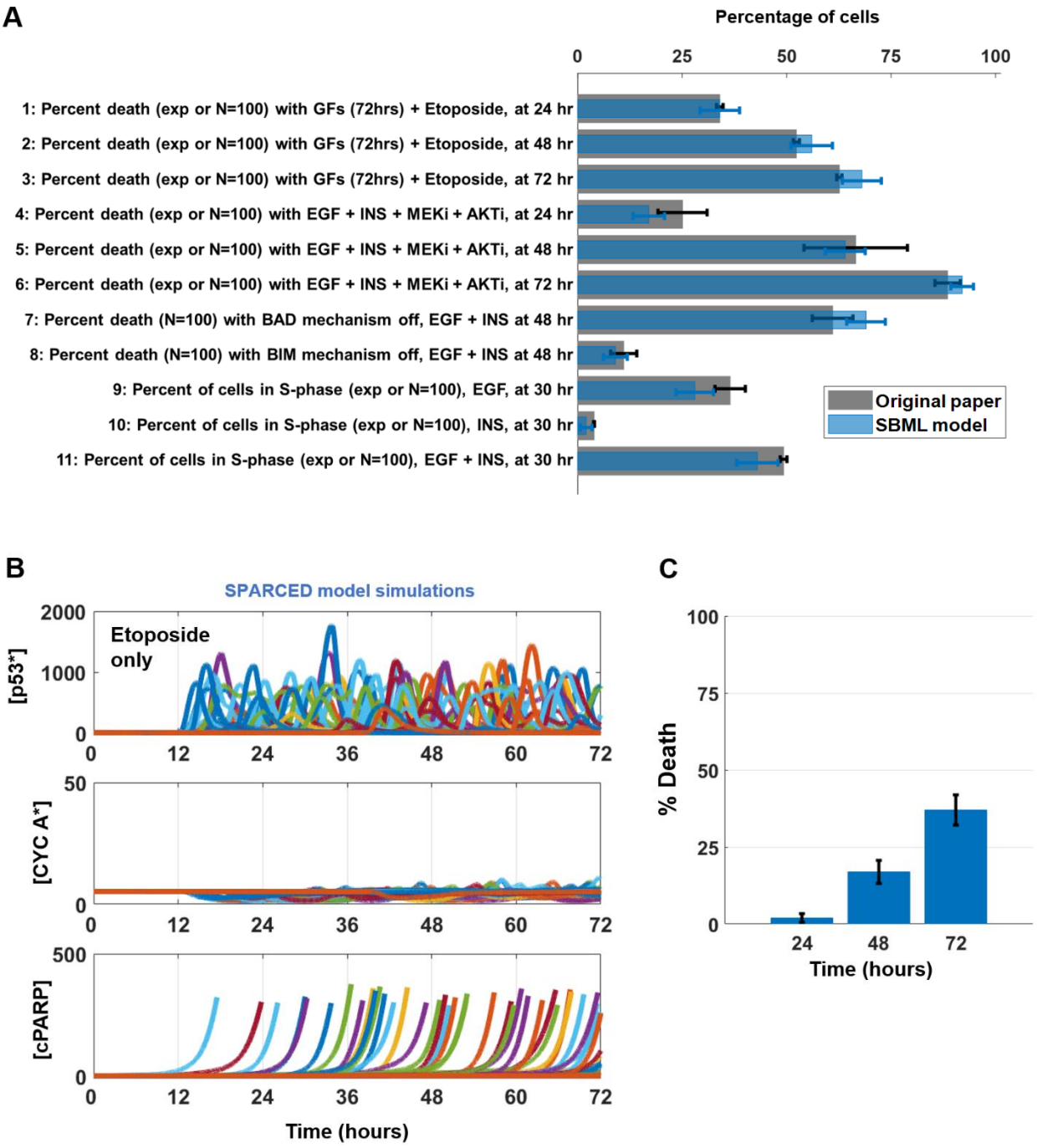


Figure 2.16: (A) Bar plots corresponding to the conditions shown in Fig. 2.3C. Gray bars are experimental or simulation data from Bouhaddou 2018 model and blue bars are simulation results of SPARCED model. Bars represent mean ± s.e.m. (B) Etoposide treatment alone induces lesser cell death compared to Etoposide + Growth Factor stimulation, shown in Fig. 2.3D-E. (C) Percentage of cell death of 100 cells shown in (B). Bars represent mean ± s.e.m.

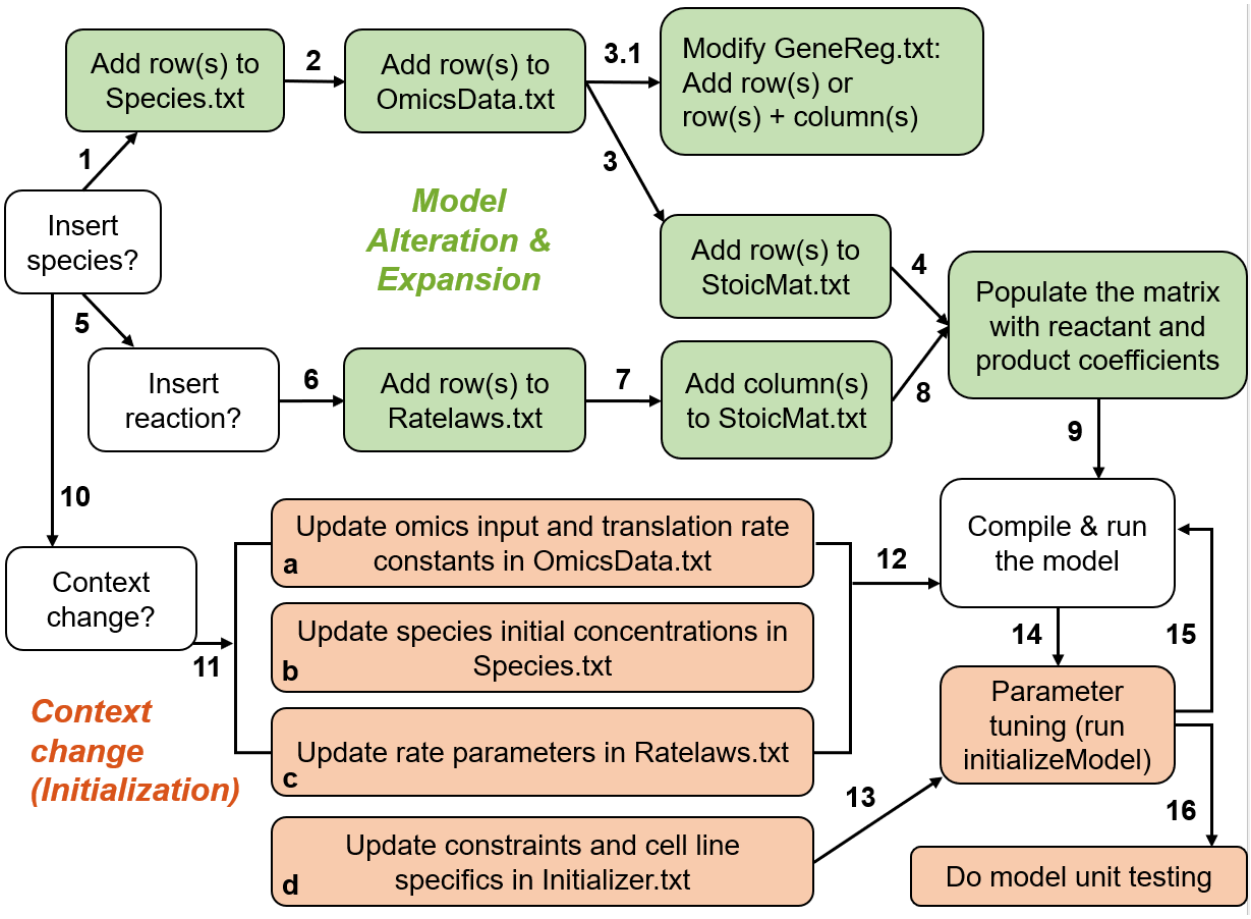


Figure 2.17: SPARCED model alteration guidelines - Steps of model expansion and context change procedures are listed. Refer to Table 2.2 for more details. Steps can be skipped if no changes are necessary.

A

Parameter category changed	Number of parameters changed (out of 3275 possible)	Input file
mRNA counts	141	OmicsData.txt
Gene copy numbers	141	OmicsData.txt
Constitutive mRNA translation rates	141	OmicsData.txt
Induced mRNA translation rates (from Bouhaddou2018)	141	Ratelaws.txt
Degradation rate constants (from Bouhaddou2018)	46	Ratelaws.txt
Species initial concentrations (from Bouhaddou2018)	914 (out of 914)	Species.txt

B

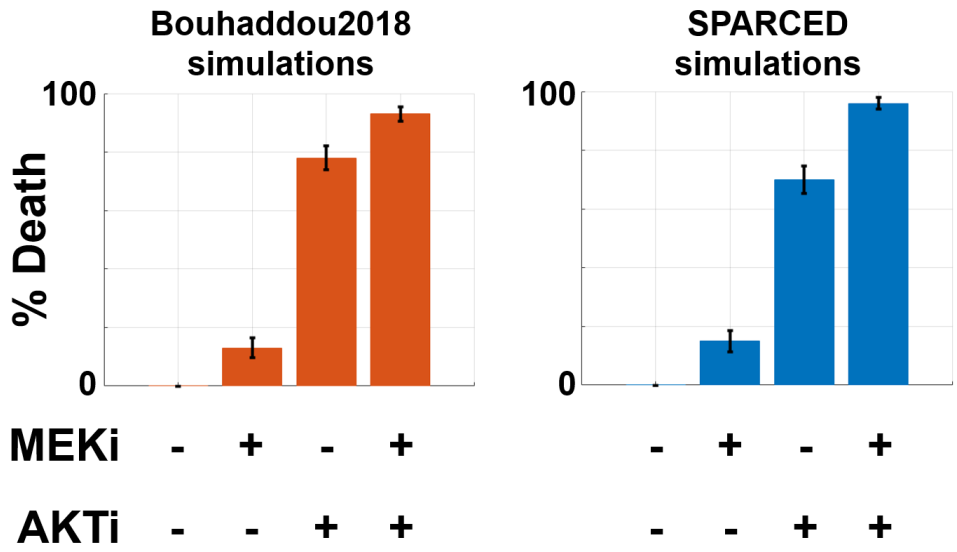


Figure 2.18: SPARCED model alteration for U87 context. (A) The list of parameters and species values modified for SPARCED model context change from MCF10A cells to U87 cells. (B) SPARCED_U87 model simulations reproduce previous observations, where U87 cells show increased response and sensitivity to AKT inhibition. MEKi: MEK inhibitor, AKTi: AKT inhibitor. Bars represent mean ± s.e.m. of 100 single cell simulations for each condition.

Table 2.2: SPARCED Model Alteration Steps

SPARCED model expansion protocol		**SPARCED model context change protocol**	
1	Decide if you need to add a new species, and a row to Species.txt for each new one.	**10**	If no changes in model topology are required, and a model context change (i.e. cell type) is desired, no new rows and columns are added.
2	Define if the new species has a new gene and add a row to OmicsData.txt file. If the new species is a product of existing gene, no alteration in OmicsData.txt required.	**11a**	Update the OmicsData.txt file with new data of mRNA levels, gene copy numbers, and protein levels. Also update the constitutive mRNA translation rate constants.
3	Add each new species as a new row to StoicMat.txt file.	**11b**	Update species initial concentrations in Species.txt file. The default concentrations here are steady-state values.
3.1	If the new species has a new gene, define if it has a transcriptional activator and/or repressor. If No, add only row(s) in GeneReg.txt, with all zeros as the column values. If Yes, add row(s) to define the new gene(s) and column(s) to define the effector species. Define the transcriptional regulation parameters by entering non-zero values in the new row(s)/column(s).	**11c**	Update rate-law parameter values in Ratelaws.txt. This step could include a small number of manual alterations, like turning-off a reaction.
4	Set the new species' stoichiometric coefficient for each reaction in StoicMat.txt file.	**11d**	Update experimental constraints and cell line specifics in Initializer.txt.
5	Decide if you need to add a new reaction to the model (if not, skip to 9).	**12**	Create the new SBML file and compile the new model using createModel.ipynb notebook.
6	Add a new row for each new reaction to the Ratelaws.txt and set the corresponding rate parameters.	**13**	Use the new cell line constraints (Initialization.txt) to calibrate the new model.
7	Add a new column for each new reaction to the StoicMat.txt.	**14**	Tune translation and degradation rate parameters using the initializeModel.ipynb notebook.
8	Define the reactants & products of each new reaction and set their stoichiometric coefficients StoicMat.txt file.	**15**	Save the new species concentrations and parameter values, then re-compile the model. The model is ready to run.
9	Create the SBML file and compile the model using createModel.ipynb notebook. The model is ready to be imported and run.	**16**	Do model unit re-testing within the new context.

2.4 Discussion

Here, we have re-created one of the largest mechanistic models in the literature, using our new python-based creation and simulation pipeline. Our modeling pipeline and model creation recipe are based on structured and easy to modify input text files and uses Jupyter notebooks or scripts (for scaled cloud-computing) to create and simulate model files. It also enables easier model alteration (species/rate law or parameter value changes), omics data integration, and model variant vs. hypotheses testing. Our exemplar model, called SPARCED, is available online on GitHub (github.com/birtwistlelab/SPARCED). While the pipeline we introduce is a recipe for large-scale model construction, the SPARCED model itself can serve as a basis for creating context-specific (personalized) model variants, studying virtual cell population responses, and as a building block towards whole-cell-scale models.

We showcased the use of SPARCED model by changing the cellular context of the model from MCF10A breast epithelial cells to U87 glioblastoma cells, by only replacing parameter values in a few input files. Although we used previously calculated values from the Bouhaddou2018 model to show reproducibility of the subsequent analyses, it is notable that SPARCED pipeline correctly creates and simulates a large-scale mechanistic model file only from an altered set of text-based input files. Additionally, we created a new version of the Initialization script that utilizes another cell-line specific input file to calibrate the model initial conditions. The initialization allows distribution of total protein and mRNA level omics data across all model species and estimates data-driven, cell-line specific translation rate constants. As a customizable set of steps, the initialization sustains user defined phenotypic responses, like cells not

going into apoptosis or cell cycle without growth factor stimulation. Importantly, the procedure accommodates mRNA input alone (without proteomics data) and calculates total protein levels using gene-level mRNA-to-protein ratios from the default MCF10A values. The outputs of the initialization procedure are species concentrations and parameter values deposited into a new SBML file, which is exported with a new name and re-compiled using AMICI.

Many existing big models are constructed in complicated and hard-coded ways and are not available in standard modeling formats, like SBML. For instance, the Bouhaddou2018 model we used as our starting point was custom coded in MATLAB with tens of different script files with thousands of lines. Although the model performance was optimized for its topology, alteration and expansion of the model was extremely difficult. However, models, especially the large-scale and clinically relevant mechanistic models, must become easy to formulate, understand, and disseminate for reproducibility, re-useability, and applicability in clinical decision making. Here, the model construction pipeline and the SPARCED model contributes to this need by being built upon structured and annotated input files, by using open-source packages, and by being available publicly on GitHub.

One key advantage of the SPARCED model format is its potential compatibility with RBM. The reactions and species created by RBM software can be incorporated (manually or programmatically) into the SPARCED model input files. Although existing RBM software can export models in SBML format and enable multiple features, the SPARCED models enable single rate parameter changes and inclusion/exclusion of individual rate laws at the input file level. Then, the SPARCED-nf pipeline can be used

to study large-scale variant analysis or to do parameter scanning. One main goal of the AMICI package[106] (95) is enabling large-scale parameter estimation, and our choice to use this package was to enable such future endeavors when needed. Combining this idea to test consistency across multiple datasets, users can search for best-fit models or pinpoint discrepant datasets given the model topology[63].

The SPARCED model encodes intrinsic stochasticity of total protein levels and mRNA numbers in its hybrid simulation mode, making it unique (together with the Bouhaddou2018 model) to offer stochastic as well as deterministic simulation settings within a single model of this biological and time scale. There are other tools such as COPASI that offer hybrid (deterministic + stochastic) simulation settings[95,107]. Our hybrid simulation approach treats the gene expression module as stochastic (events modelled as Poisson processes) and the protein signaling module as deterministic. COPASI uses next-reaction-method[108] for the part it determines as stochastic based on molecule numbers of the interacting species. However, as the developers stated, such implementations tend to be inefficient and take prolonged simulation wall-times. Indeed, COPASI (v4.25, build 207, on Windows 10 Pro) fails when we try to simulate our model. A next step for the SPARCED model is to combine the gene expression (scripted) and protein signaling (SBML file) modules into a single SBML model file. Such a change would enable broader cross-platform testing and usage of the SPARCED model. As stated above, even the current SPARCED model SBML is too big for most tools available to accurately simulate for the relevant time scales (24-72 hours). New numerical / algorithmic methods are required to simulate large-scale hybrid models[109]. The single-cell capability of SPARCED allows one to capture some important aspects of

cell line and tumor heterogeneity compared to an average cell condition (the way many mechanistic models are built). Users can leverage this feature to simulate virtual populations and study a cell population response to drug treatment, which is often a single-cell readout as are most cellular phenotypes. However, such simulation settings require larger computational resources and thus model compatibility for high performance computing environments. The SPARCED model is built to be compatible with cloud computing, where it can be used to simulate thousands of single cells with single job execution (see Methods).

There are, of course, some shortcomings and remaining challenges. Although we extensively showed that the SPARCED model creation/alteration is much easier compared to previous version, it still is a (careful) stitch-together of other models. There are certainly other models that can be substituted and tested. The tab-separated input files separate model details from the simulation itself and offers multiple advantages mentioned throughout this work. However, it can be seen cumbersome for some modelers. These input files include species and compartment annotations, but there are other recent efforts to standardize such metadata/annotation sharing, which we would conform to when fully developed [110]. The hybrid mode of SPARCED includes Python scripts for stochastic simulation of gene expression module. By defining this module in another SBML file, and using packages like SBML comp[111], one can start exploring other methods and tools for performance testing. However, a full comparison to other hybrid and stochastic methods requires computational tools that are yet to be developed that work with models of this size. Additionally, we do not provide scripts to merge new models into SPARCED, and the field of model merging is an active research

area[90,111,112]. For instance, researchers can write code to insert new reactions and species from other model files into existing input files, which would then also require to update OmicsData.txt input file with new gene information. Yet, in our experience, model merging often necessitates human interaction to define the mechanisms by which species interact with one another and what rate laws should define those interactions. Finally, if one desires to alter the model at the Antimony file stage, there are currently no automated ways we provide to map the changes onto input files. This would be possible if desired but was not studied here.

SPARCED is a large-scale pan-cancer pathway model that incorporates six major sub-modules, making it one of the state-of-the-art computational models for mammalian signaling. However, it does not yet include one of the hallmarks of cancer, the cellular metabolism mechanisms[113]. Additionally, the current version of the SPARCED simulation code does not explicitly track cell division events, although the cell cycle itself is modeled. However, this is ongoing work and will enable us to better capture and compare to experimental observations and data, such as traditional drug dose viability response experiments that are fundamentally related to tracking single cells and their division / death events.

One of the challenges is to explore simulations of spatially aware single cells. Currently, the SPARCED model captures intrinsic heterogeneity of cells (by having stochastic gene switching and mRNA birth/death events) but these cells cannot “talk” to each other. In the future, by having scenarios where spatial orientation of cells are recorded and the secreted or stimulated molecules are shared between them, we can better capture tissue microenvironment and heterogenous pharmacokinetics[114,115].

Related to this first task, the second challenge is to create and simulate scenarios with multiple cell types (i.e., models trained on data from different subtypes of cells) or defining events to capture differentiation of cells. For example, one may be able to use single cell RNAseq data to train SPARCED-like models to enable tissue-level simulations with the critical cell types in the proper geometric locations. This overall vision would enable spatially aware, single-cell level, large-scale mechanistic models trained on individual patient data for in silico drug screening. The pipeline presented here is an important step towards this goal.

Another challenge to achieve using large-scale models is a whole-cell level mechanistic model for mammalian cells[90,116]. With our approach, the SPARCED model can be enlarged using other small-scale models for pathways and mechanisms not currently included in the model. By utilizing the unit testing approach, one can then verify the model performance and get larger, more comprehensive models. The open-source framework presented here increasingly facilitates community contribution for model context-change and parameter tuning based on new experimental conditions.

Our introduced method of large-scale mechanistic model construction, and the SPARCED model as a basis, will enable researchers to more easily create and manipulate new model versions, test different mechanisms of action to interpret experimental observations, and change the model's cellular context. The models created by the SPARCED pipeline can incorporate multiple (omics) datasets, providing non-"black-box" data integration and modeling; however the extent to which a fixed "initialization" pipeline can be successfully applied to a variety of cell lines remains to be tested. These SPARCED models additionally provide single-cell level simulations,

compatibility with cloud computing, and human-interpretable & annotated model files in SBML format (as do other modeling tools, albeit not at this scale). The SPARCED model now can more easily be re-used as one of the largest mammalian-cell mechanistic model in the literature and serves a primer role in creation of context-specific, hypotheses testing, and expandable models. In conclusion, the SPARCED model format contributes towards important foundations of reusable big models, paving the way towards personalized mechanistic models for data integration.

2.5 Materials and Methods

2.5.1 Computational Methods

The Bouhaddou2018 model

The Bouhaddou2018 model (Fig. 2.1a) is one of the largest single-cell mechanistic models for mammalian cell signaling regulating proliferation and death. The first version of the model used as a test case in this work was written in MATLAB (The MathWorks, Inc.)[32]. The model is a hybrid of deterministic and stochastic modules. The deterministic module describes the concentration dynamics of 774 proteins, protein complexes, and post-translationally modified species through 2449 reactions using the Sundials CVODEs package for simulation[117]. The stochastic module describes gene state (active/inactive) and mRNA birth/death dynamics for 141 genes. The deterministic and stochastic modules exchange information every 30 simulated seconds. In short, the current levels of select protein states can induce changes in gene activation/deactivation and/or mRNA transcription/decay rates. The newly updated mRNA copy numbers change nascent protein translation rates in the deterministic module.

The SPARCED Model

We converted the Bouhaddou2018 model into a Python + SBML[93] format (Fig. 2.1b). The deterministic module is ultimately encoded in an SBML file (.xml) whereas the stochastic module is written in Python. A foundational and important feature of this recoding effort is that the SBML file is generated from a small set of simple structured input text files (Fig. 2.1c) via Python scripts. Introduction of such structured input files and associated Jupyter notebooks enables simple alteration of model structure and/or parameter values, for example turning on/off certain interactions. The input files also enable rigorous annotation of model features using, for example, ENSEMBL[118] and HGNC[119] identifiers, which is seldom done in such mechanistic modeling.

Input files:

There are six SPARCED model input text files (tab separated values), each with a defined structure as detailed below. The user can change these files to create and compile a new model.

(1) `OmicsData`: This file includes the gene copy number, mRNA copy number, and proteomic data. This input file also contains rate constants for the stochastic module and initialization procedure. Each row of the file corresponds to one gene and the columns are different data types. The first column is gene name (HGNC identifiers), the second column is gene copy number, the third column is mRNA molecule copy number per cell (mpc), the fourth and fifth columns are rate constants of gene inactivation and activation respectively (s^{-1}), the sixth column is constitutive transcription rate constants (molecules per second), the seventh column is maximal transcription rate constants (molecules per second), the eighth column is mRNA degradation rate

constants (s^{-1}), the ninth column is protein copy number (mpc), the tenth column is protein half-life parameters (seconds), and finally the eleventh column is the translation rate constants (s^{-1}). These latest set of rate constants are from literature and provided for genes for which our omics input lacked protein level data. All the rate constants are taken from the Bouhaddou2018 model. Users can add new rows to this file, using RNAseq data to estimate mRNA levels for the genes to be added[32]. When adding genes (rows) to the model, a reasonable starting point for rate constants (or other values), in the absence of any other data, is to use median values from the genes/parameters currently in the model.

(2) `Species`: This file contains information about the species in the deterministic module. Each row corresponds to one species (protein, protein complex, post-transcriptionally modified species). Transcripts (in nM) are also included in this file because they are regarded as species with updated concentrations in the stochastic module every 30 seconds and are used in translation rate laws. The first column is the species name. Names can be arbitrary so long as they are unique in the model. Importantly, the name list needs to match the first column in the `StoichiometricMatrix` file described below. The second column is the species home compartment. The home compartment of a species defines its cellular localization. A species can reside in a compartment defined in the Compartments input file: currently Cytoplasm, Mitochondria, Nucleus, or Extracellular. The third column is initial condition in nM units, with respect to the home compartment volume. These values are taken from the Bouhaddou2018 model, post-initialization. The fourth column

is a comma separated list of ENSEMBL gene identifiers corresponding to gene products present in the species.

(3) `Ratelaws`: This file has a row for each reaction in the deterministic module. The first column is the unique (arbitrary) name of each reaction. Currently, we named each reaction based on the related sub-module (e.g., vA1-87 for Apoptosis and vC1-104 for Cell Cycle). The number and order of rows in this file should match the columns in the `StoichiometricMatrix` input file defined below. The second column in this file contains the home compartments for the reactions. The designated compartments should be one defined in the Compartments input file: currently Cytoplasm, Mitochondria, Nucleus, or Extracellular. The home reaction compartments define the effective search volume for each reaction and are used to rescale concentrations when appropriate. Note that both species and reactions have home compartments defined, where a species can participate in a reaction defined in a different compartment. For instance, the EGF binding to EGFR reaction occurs in extracellular space (volume Ve), where EGF's home compartment is the extracellular space and EGFR's home compartment is cytoplasm (Vc). A volumetric correction for EGFR concentration in this rate law is done by multiplying by the ratio of Vc/Ve. The third column can have either a number or a reaction formula. If it is a number, it means the corresponding reaction is mass-action type, and the number is the rate constant for that reaction in units of nM and seconds. Note that the reactants and products are defined in the `StoichiometricMatrix` input file. If the third column is a formula, it means the reaction will follow that rate law, and the next set columns in that row are the values of each parameter defined in the formula in the third column, again in units of nM and

seconds. The rate law can include any species name described in the Species input file. The parameter names in the rate law should start with “k” and be unique in that formula. We distinguish multiple parameter names with an underscore and ascending list of numbers (e.g., kA_1, kA_2). During model generation, all parameter names in this file are re-named in an ascending order based on the number of rate laws. The full list of parameter name/value pairs are output into a new file (`ParamsAll`) for user reference.

(4) `StoichiometricMatrix`: This file defines the reaction stoichiometry, and therefore the reactants and products of model reactions. The rows correspond to the Species input file and the columns correspond to the rows in the `Ratelaws` input file. Here, the species and rate law names should match the names defined in Files `Species`, `Ratelaws` and `Observables`. Each element (starting at the second row and second column index) has a stoichiometric coefficient (typically -2, -1, 0, 1, or 2), where negative sign indicates reactants, and positive sign implicates products of a reaction.

We also provide an option to not use the stoichiometric matrix as an input file (see github.com/birtwistlelab/SPARCED/tree/noStoicMat). Instead, the reactions are defined within a new column in the updated `RatelawsNew` input file.

(5) `GeneReg`: This file describes transcriptional activation and inhibition interactions, where rows correspond to genes (the same order as the first column of `OmicsData`) and columns to species that are defined as activators or repressors of transcriptional activity. The first column is gene name (HGNC format). There are currently seven more columns in this file, each corresponding to one species defined as

an activator or a repressor (e.g., p53 induces p21 transcription or AP1 inhibits cFOS transcription). A single value of zero indicates no effect. A non-zero entry in row i and column j denotes that species j regulates gene i transcription. The non-zero entries have the form “A; B”, where “A” is the hill coefficient and “B” is the half-maximal concentration of the species “j” effect. To simplify the input file structure, we use positive values of “A” to denote activation, and negative “A” values to denote inhibition. This file is used by the stochastic module script to update mRNA levels. To add additional transcriptional regulators (activators or repressors) into the SPARCED model, users should add as many columns as new regulator species and populate the columns with corresponding rate constants.

(6) `Compartments`: This file contains the names of compartments in the model (first column), the volume of the compartment in liters (second column), and the corresponding GO-term of the compartment (third column). The compartment names should match the compartment names listed in `Species` and `Ratelaws` input files.

(7) `Observables`: This file contains information about model observables. Each observable corresponds to the compartmental-volume-corrected summation of all formats of a protein. There are 102 observables defined (columns) for the model species (rows) in this file. The entries are either 1 (the species in the row is part of the observable in the column) or 0 (otherwise). The “createModel” Jupyter notebook uses this file to define an observables variable as an input for the AMICI model compiler.

(8) `Initializer` (Optional): This file (Supplementary File 16) contains information used for model initialization. Species concentrations (columns 1-2), mRNA

level adjustments (columns 3-4), parameter values (columns 5-7), observables to exclude from translation rate adjustments (column 8), and single parameter scan range (columns 9-11) are populated for each step of initialization. The steps used here are shown to work to get a good starting point for serum starved MCF10A cells, which do not undergo apoptosis or enter cell cycle without growth factor stimulation, in deterministic simulation mode.

Dependencies

(1) Docker: As is the practice with Kubernetes-compatible workflows, all model dependencies and runtime environments are Dockerized into a downloadable image for self-contained model execution. This means when a job for SPARCED-nf is launched on the Kubernetes cluster, it will download the Docker image for SPARCED-nf and execute the model within that container. The Docker image for SPARCED-nf is built on the Ubuntu-18.04 operating system with python3 installed, as well as a few minor system utilities required for AMICI. The image can be found at https://hub.docker.com/repository/docker/birtwistlelab/sparced.

(2) `Nextflow` (nf): In this cloud-scalable version of the model, the Jupyter notebooks have been converted into python source code and re-modularized for greater parallel-simulation efficiency. The process of creating and executing the model is handled entirely by `Nextflow`, a workflow-management application and language for building resilient pipelines. When SPARCED-nf is launched, `Nextflow` begins by creating a head pod on the cluster to coordinate each of the jobs needed to run the model. The head pod

creates smaller jobs that each download the containerized dependencies from Dockerhub, pull the model source files from the SPARCED-nf GitHub repository, and run the assigned process. Once the model has completed execution, the output files are saved to a section of the Kubernetes cluster called the persistent volume claim (PVC), where they remain stored in the cloud for user download.

SPARCED-nf model simulation set-up

SPARCED-nf uses the same tab-separated-value input files as SPARCED. For SPARCED-nf to build and execute, the files are copied into the aforementioned PVC for workflow access. This is done with kube-runner (https://github.com/SystemsGenetics/kube-runner), a submodule for automating common PVC tasks with Kubernetes' kubectl tool. The kube-load.sh file is used to write new input to the PVC, and kube-login.sh is used to access and delete old input files from the cluster.

Along with its scalability, SPARCED-nf is also highly customizable. The nextflow.config configuration file is used to define the specifics of simulation scenarios.

(1) nextflow.config: This configuration file has two main sections. In the first section (called K8), users define the Kubernetes namespace specifics and folder configurations. In the second section (called params), users customize runtime arguments for simulation settings. The available parameters are input_dir_name (the directory name of the input files), flag_deterministic (flag=1 for deterministic or flag=0 for hybrid simulations), sim_time (simulation time in hours), Vol_nuclear (volume of nuclear compartment in liters),

Vol_cyto (volume of cytoplasmic compartment in liters), speciesVals (species names + initial concentration values to start from), ratelawVals (parameter names + values), and numCells (number of single cells if the simulations are hybrid). Importantly, the “speciesVals” and “ratelawVals” parameters allow users to pass in a formatted string to specify parameter sweeps. Using these in conjunction with the “numCells” parameter, the user can simulate thousands of cells in hundreds of different microenvironments in a single execution.

(2) SPARCED-nf:model_build: Analogous to “createModel.ipynb” in SPARCED-jupyter model, this phase of the Nextflow pipeline constructs all necessary files for the model simulation.

(3) SPARCED-nf:split_from_params: This is the major parallelizing step of SPARCED-nf. Having received the relevant model files from the last step, the workflow ingests the speciesVals, ratelawVals, and numCells arguments set by the user in the nextflow.config. Using the input files, it creates new input files to satisfy the user-specified parameter sweeps. Each new input file permutation is moved into its own new folder, and each such folder is duplicated numCells times.

(4) SPARCED-nf:model_run: This final step of the Nextflow workflow is responsible for model execution and output generation. Each folder created in the previous step above serves as the unique runtime environment in this step. The model pulls assigned simulation input files associated with the folder. Each instance of this step is run in parallel across different simulation

environments. Functionally, the code executed is very similar to the "runModel.ipynb" notebook and the model outputs are saved to the PVC.

When the models complete execution, each SPARCED-nf:model_run instance saves its output to a unique folder on the PVC. To download these folders to the local filesystem, users can employ kube-save.sh (from the kube-runner module).

Code Availability

The final model scripts, files, and information are available in Birtwistle Lab GitHub repository, github.com/birtwistlelab/SPARCED and github.com/birtwistlelab/SPARCED/tree/noStoicMat.

Computational standard-error-of-the-mean

We report the s.e.m. for simulations (Figs. 3c, 3e, Supplementary Figs. 9a-e, 10a-b, 13a, 13c, and 14c) using the ratio of binomial proportions. See equation (1) below, where the "Percentage of cells" corresponds to the percentage of cells showing the phenotypic readout (i.e., percentage of cells in S-phase, percent cell death) and the "Number of total cells" is the number of starting single-cell simulations, usually 100.

$$s.e.m. = \sqrt{\frac{Percentage\_of\_cells \cdot (100 - Percentage\_of\_cells)}{Number\_of\_total\_cells}}$$

Chapter 3

COMPUTATIONAL SPEED-UP OF LARGE-SCALE, SINGLE-CELL MODEL SIMULATIONS VIA A FULLY-INTEGRATED SBML-BASED FORMAT

3.1 Author Contribution

The work presented in this chapter has been adapted from the following publication:

> **Mutsuddy, A**., Erdem, C., Huggins, J.R., Salim, M., Cook, D., Hobbs, N., Feltus, F.A. and Birtwistle, M.R., 2023. Computational speed-up of large-scale, single-cell model simulations via a fully integrated SBML-based format. Bioinformatics Advances, 3(1), p.vbad039.

The contribution of the candidate involved curation of data, composition of model input files, development of software pipeline for model construction, validation of model construction and simulation protocols, software optimization, investigation, performance benchmarking, writing and visualizations.

3.2 Abstract

Large-scale and whole-cell modeling has multiple challenges, including scalable model building and module communication bottlenecks (e.g. between metabolism, gene expression, signaling, etc). We previously developed an open-source, scalable format for a large-scale mechanistic model of proliferation and death signaling dynamics, but communication bottlenecks between gene expression and protein biochemistry modules remained. Here, we developed two solutions to communication bottlenecks that speed up simulation by ~4-fold for hybrid stochastic-deterministic simulations and by over 100-fold for fully deterministic simulations. Fully deterministic speed-up facilitates model initialization, parameter estimation and sensitivity analysis tasks.

3.3 Introduction

Recapitulating the behavior of single cells in silico is a grand challenge not only for systems biology, but also for biology in general. Such an accomplishment would imply that we have a thorough understanding of all the cellular and sub-cellular processes that give rise to relevant phenotypes. Such models could enable rational engineering for biotechnology applications, or forward predictions in precision medicine[120–123]. Large-scale and whole-cell modeling is a suitable foundation for meeting such challenges[90,124,125]). The first such efforts focused on genome-scale metabolic modeling in multiple organisms[126,127]. Subsequent efforts focused on integrating multiple "modules" in addition to metabolism (e.g. gene expression, signaling, etc.) in single-celled organisms such as *M. genitalium*, *E. coli*, and *S. cerevisiae*[56,58,126,127], and a minimal lab-generated cell[128], but the lack of dedicated tools specifically for large-scale / whole-cell models presented roadblocks for reuse. Algorithmic developments included rule-based modeling to specify reactions more compactly[129], and model composition tools[87,95,130,131], but large-scale models often still presented challenges. More recent work has provided such tools like AMICI[106] that enables SBML-specified models to be simulated quickly, PEtab[132,133] and Datanator[134] (Roth et al., 2021) that specifies data formats for parameter estimation, formalisms that can help with unambiguous species naming[135], and composition approaches such as SBML merging[136] and ours that simplify model aggregation and expansion in ways that are compatible with efficient large-scale simulation algorithms and easy to reuse[137]. Not unexpectedly, however, there remains much work to be done to even technically enable large-scale and whole-cell modeling.

Here, we focused on improving communication between different modules as a major impediment for computation speed in large-scale modeling (Fig. 3.1). We used our recently published SPARCED model as a test case, a large-scale mechanistic model of proliferation and death signaling in single mammalian cells. This model consists of 141 genes, and 1196 unique biochemical species. It is built by translating a simple set of structured input text files into an SBML-compliant module that captures “protein biochemistry” (signaling) and is simulated using AMICI, and a module that captures “gene expression” using python. It can be simulated in a hybrid stochastic/deterministic mode, where gene expression dynamics follow Poisson-like processes, or a fully deterministic mode. Regardless of the mode of operation, computation speed was a major concern. Continuation of our work with the SPARCED model relies on our ability to perform more complex and resource-intensive computation, such as model initialization, parameter estimation and sensitivity analysis. Even though many such tasks require only deterministic operation, insufficient computation speed with multi-module deterministic formalism precluded further analysis, which motivated us to seek further improvement in computation speed for the general operation of the SPARCED model.

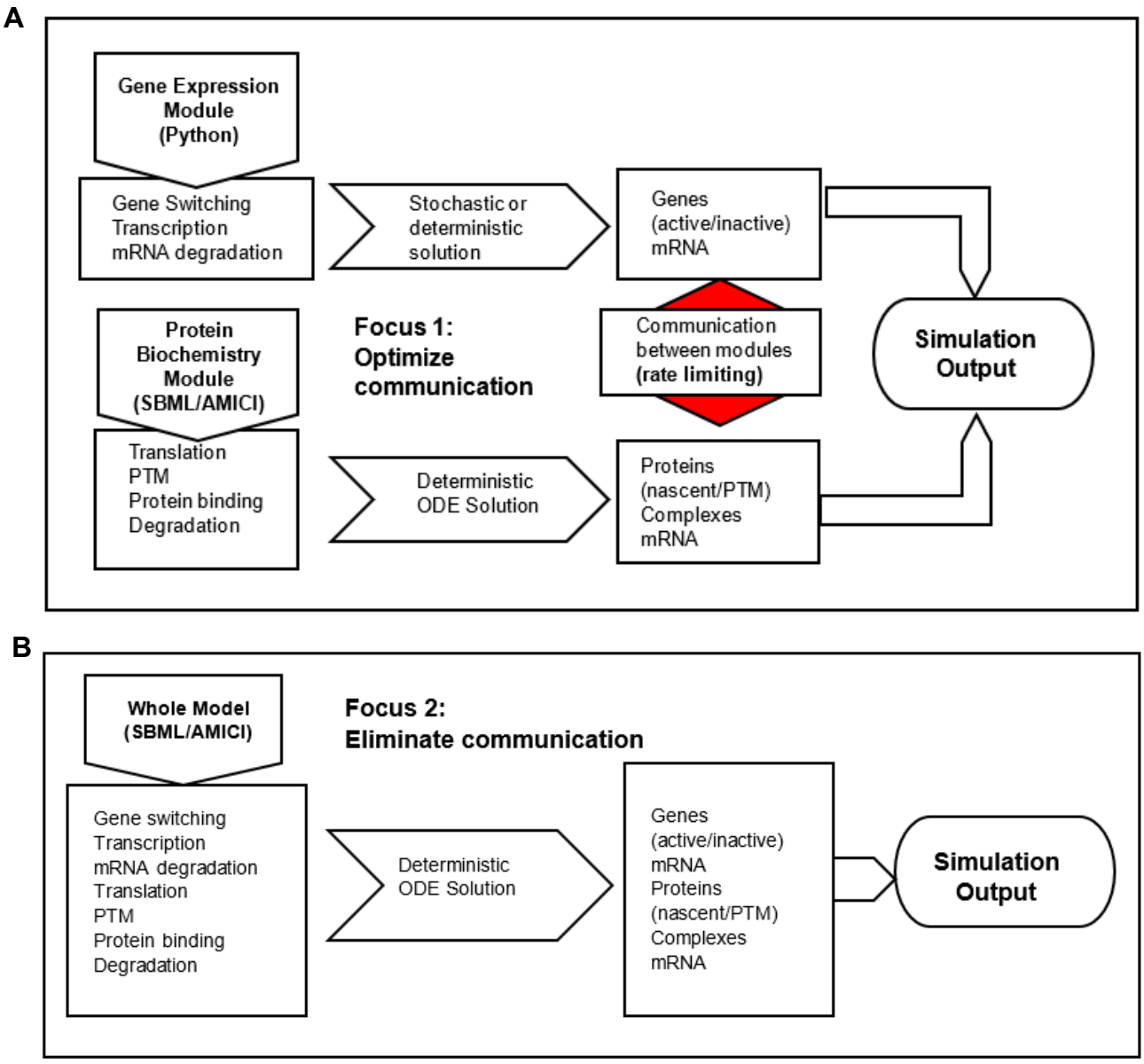


Figure 3.1: Workflow of the SPARCED model. (A) Simulation workflow of the original SPARCED model highlighting the bottleneck of communication between the gene expression module and the protein biochemistry module. One speedup reported here targets that bottleneck for faster stochastic simulations. (B) A new simulation workflow reported here that integrates the gene expression module with the protein biochemistry module using SBML, enabling large computational speed up for deterministic simulations. No solvers yet exist for stochastic simulations at this scale.

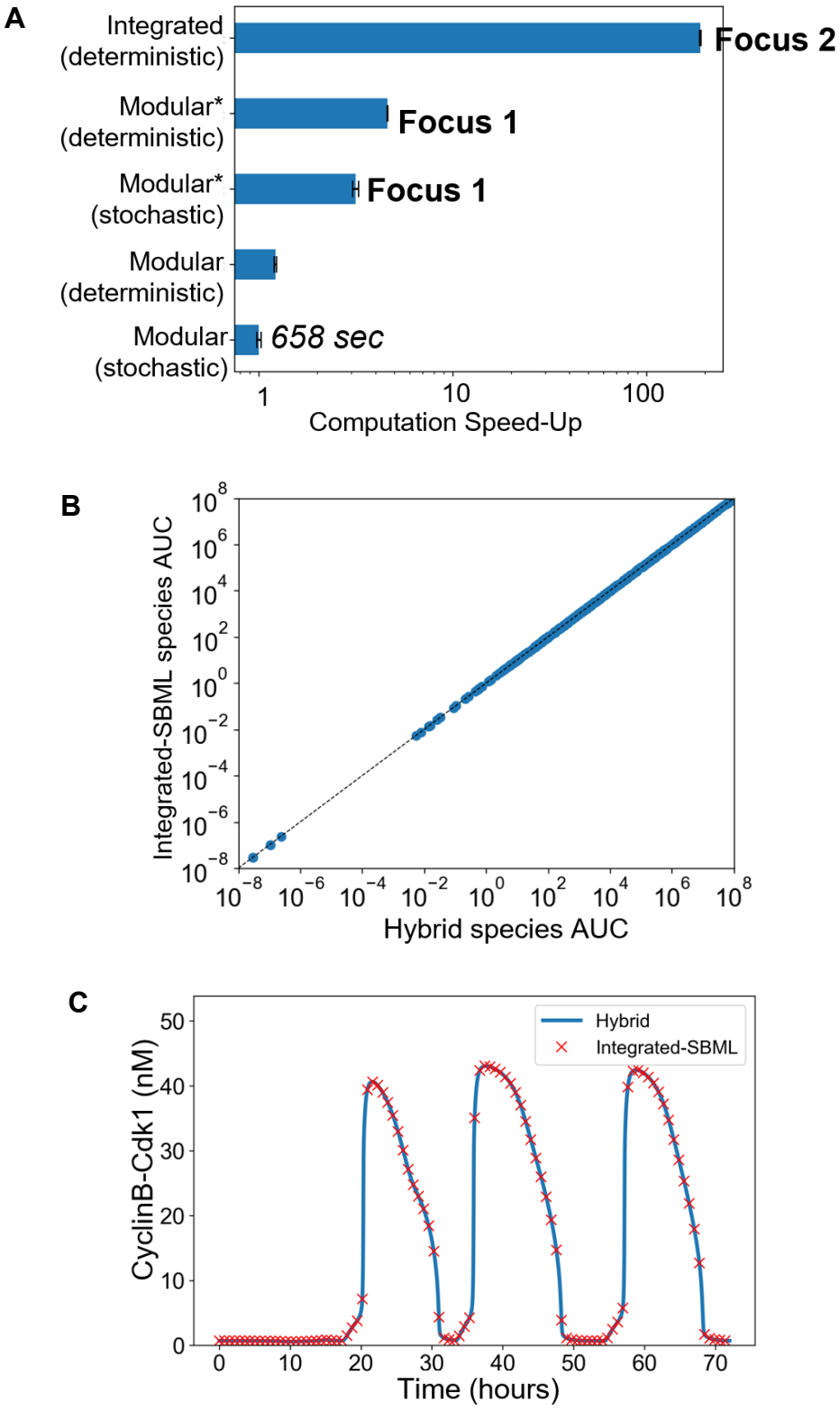


Figure 3.2: Computational speed-up of the SPARCED model

Figure 3.2: Computational speedup of the SPARCED model. (A) Computation speed-up enabled by improving communication between modules (denoted by *; Focus 1) and by integrated SBML of the gene expression module with the protein biochemistry module (Integrated; Focus 2). Improving communication (Focus 1) yields 3-4 fold speed-up, and eliminating communication (Focus 2) yields ~100-fold speed-up. A relative speed of 1 corresponds to 658 seconds. Error bars are from 10 replicate simulations. Simulations were performed on Palmetto (Clemson's HPC resource—Intel Xeon CPU 2.5 GHz). (B) Area-under-curve for the dynamics of all model species in the original modular deterministic formulation and the integrated formulation. Simulated serum-starved MCF10A cells were treated with 1 nM EGF and 0.005 nM HGF and observed for 72 hours. (C) An example trajectory for a biochemical correlate of cell division events from the simulations in D for both models, showing good agreement.

3.4 Results

As is typical for large-scale models, communication between modules was done at specified simulation time steps, in our case every 30 simulated seconds. Using Python's built-in code profiler tool, "cProfile" on our simulation code, we discovered that 94.5% of the total execution time was being spent running NumPy processes that store information in arrays. This outcome indicated a potential lack of efficiency in simulation output handling. We further sought a more detailed profiling with "line profiler", which tracks execution time for every line of code. These results showed 1.4% comprised stochastic gene expression, 15.8% comprised solving ODEs (AMICI), and 81.9% comprised storing solver results into a NumPy object for inter-module communication (with the remaining 0.9% on miscellaneous overhead). We thus focused on inter-module communication as a rate limiting step for simulation speed (Fig. 3.1A—Focus 1).

During each 30 second time step, results from the "protein biochemistry" module are saved in a "results" object defined within the AMICI library. However, accessing the state matrix via the Python object interface incurred expensive reconstructions of the full NumPy array from AMICI-managed memory. These overheads could be largely avoided, since only the last column of the state matrix (corresponding to the most recent timestep) was needed at each iteration. By using direct access to the SWIG pointer referencing these state variables, we were able to avoid re-reading state data, yielding a 3-4-fold simulation speedup (Fig. 3.2A—Focus 1).

However, we reasoned that a potentially better solution to improving inter-module communication was to eliminate it altogether. This required a structural reformulation of

the entire SPARCED model, whereby both modules are contained within a single SBML file. The drawback to this so-called integrated SBML model is that no efficient numerical solvers yet exist to perform stochastic simulations on such large models. Nevertheless, fully deterministic simulations are still of use in certain situations, like model initialization by which we convert the cellular context of the model using multi-omics data[30,32], parameter estimation, and sensitivity analysis. In contrast, a fully deterministic simulation with the previous formulation still required communication between the "protein biochemistry" ODEs, and the mean approximation of the stochastic "gene expression" module, subject to inter-module communication bottlenecks. In compliance with the original SPARCED model construction workflow, we re-designed the model building pipeline to use the same set of text-based input files to output two executable models, one of which is fully SBML-specified and the other retains the native hybrid multi-module formalism. (Fig. 3.1B—Focus 2). After implementing this change, an over 100-fold computational speed-up was observed (Fig. 3.2A—Focus 2). We verified that simulation results obtained with this "integrated SBML" model were identical to the original model to ensure that the reformulation of the model and its build process had not introduced any errors (Fig. 3.2B-C).

Table 3.1: Comparison of COPASI and SPARCED. Deterministic and stochastic simulations have been run on COPASI using the LSODA method with duration set to 259200(s), intervals to 8640 and interval size 30(s). The COPASI Time Course deterministic simulation was run using the default settings. Integrated Reduced Model was left unchecked. Relative Tolerance was set to 1e-6. Absolute Tolerance was set to 1e-12. Max Internal Steps was set to 100000. The Max Internal Step Size was set to 0. The COPASI Time Course hybrid simulation was run using the default settings. Max Internal Steps were set to 1000000. The Upper and Lower Limits were set to 1000 and 800, respectively. The Partitioning Interval was set at 1. Use Random Seed was left unchecked and random Seed was set to 1.

Computer specifications: CPU: AMD Ryzen 5 5600g (6 core) RAM: 64GB GPU: Nvidia GTX970 OS: Ubuntu 22.04	Execution time (COPASI) (minutes: seconds)	Execution time (SPARCED) (minutes: seconds)
	Deterministic	
Trial 1	06:03.13	00:01.29
Trial 2	06:03.29	00:01.31
Trial 3	06:38.61	00:01.31
Trial 4	06:03.37	00:01.37
	Stochastic	
Trial 1	> 20 minutes	01:45.58
Trial 2	> 20 minutes	02:01.19
Trial 3	> 20 minutes	02:04.34
Trial 4	> 20 minutes	01:57.22

We next sought to examine whether this faster simulation framework provides a superior alternative to more commonly available general purpose simulation tools, such as COPASI[95]. We imported the integrated-SBML model into the COPASI GUI environment. As a test case, we ran the same deterministic or hybrid stochastic simulations using both COPASI and SPARCED (serum-starved MCF10A treated with growth factors for 72 hours). Both deterministic and hybrid stochastic performance were slower in COPASI (Table 3.1).

3.5 Discussion

In conclusion, here we provide code that speeds up simulation of a large-scale model of cell behavior by ~4-fold for stochastic simulations and ~100-fold for deterministic simulations, by focusing on improving or eliminating communication between modules. The substantial technical improvement is predominantly impactful towards our previous work, since it will facilitate model initialization, parameter estimation, and sensitivity analysis. We do not present this work as a general-purpose tool for large scale simulation, however, we believe that certain generalities in our methods and solutions may provide helpful suggestions for improvement in computational models of similar scale and structure. Namely, decreasing inter-module communication bottlenecks in stochastic and deterministic operation of large-scale models with multi-module formalism via more efficient variable handling and acceleration of fully deterministic simulation of such models by amalgamation of multiple modules into a single body mathematical description. We expect this to be impactful as a general strategy to further enable large-scale and whole-cell modeling, and also spur

the development of simulation algorithms that can perform stochastic simulations using an integrated formulation.

Chapter 4

LINEAGE-RESOLVED MECHANISTIC MODELING OF STOCHASTIC SINGLE-CELL PROLIFERATION AND DEATH ENABLES DIRECT COMPARISON OF SIMULATIONS TO ANTI-CANCER DRUG DOSE RESPONSE DATA TO ILLUMINATE GAPS IN DRUG ACTION KNOWLEDGE

4.1 Abstract

There is a large publicly-available body of anti-cancer drug dose viability response data that could improve data-starved mechanistic computational models of how cancer cells respond to drugs. However, simulation algorithms that enable a direct comparison of model output to experiment readout, often cell number (or a proxy) after several days of drug treatment, through mechanistic description of drug action, have not yet been described. We present such a simulation framework that tracks mechanistically-detailed single cell lineages to connect simulated drug dose effects on stochastic division and death events to cell number assay readouts. As an application, we simulated drug dose response experiments for four targeted anti-cancer drugs (Alpelisib, neratinib, Trametinib and Palbociclib) and compared them to experimental data. Simulations are consistent with data for strong growth inhibition by Trametinib and overall lack of efficacy for Alpelisib, but are inconsistent with data for palbociclib and neratinib. Discrepancies with data suggest that (i) the importance and/or essentiality of CDK4/6 (Palbociclib target) for driving the cell cycle is likely overestimated, and (ii) the

cellular balance between basal (tonic) and ligand-induced signaling is a critical determinant of response to irreversible EGFR inhibitors (Neratinib). This work lays a foundation for application of mechanistic modeling to large-scale drug dose viability response data sets.

4.2 Introduction

One of the grand challenges of systems biology is to build a comprehensive and quantitative understanding of the structure and functionality of living cells. Towards this purpose, whole cell model, that can describe the function of every gene and its products in an organism is an attractive manifestation of such a goal. Even though first drafts have been described for some micro-organisms[58,120] such a feat is yet to be achieved for human cells. Nevertheless, a wide range of individual pathway[28,29,104,138–150] and integrative multiple pathway[32,75,151,152] human cell models have been published which may be leveraged to build a foundation for the future human whole cell model[116,153]. Such comprehensive models have the potential to contribute to solving biological challenges such as predicting multiscale phenotypes regulated by complex signal transduction and metabolic networks[154–158], diagnosis of disease states and their progression[159,160], and development of efficacious therapeutic procedures[161–163].

A major challenge in improving these models is the sheer complexity of biology and the limitations of our current understanding[164,165]. Conducting dose-response experiments on cell lines provides an effective means to perturb biological systems and uncover novel insights into biological pathways[166–168]. Numerous large-scale databases[33–35] exploring drug sensitivity of a wide array of cell lines are available in the literature, and assessing how well computational models based on current knowledge can explain this data can reveal existing knowledge gaps and inform next stages of research.

An obvious pre-requisite to leveraging such data sets for these purposes is the existence of robust methods for comparing simulations appropriately to experimental

readouts. The overwhelming majority of anti-cancer drug dose response viability assays measure cell number, or a proxy for cell number. It is challenging to generate a conceptual analogue of such metric using mechanistic models that describe events at a lower scale, such as single cell or subcellular processes. In recent years, more accurate representations of mechanistically informed dynamic cell populations have been simulated using agent based modeling in process control and optimization of production of therapeutic proteins using mammalian cell culture[169], analyzing the impact of cell population heterogeneity in colony and tissue context[170] and elucidating the role of heterogeneity in IFNβ signaling[171]. However, the computationally intensive nature of these modeling tools[172,173] and technical limitations to incorporate sufficient molecular-level details may challenge their expansion and applicability. Therefore, unlike pathway specific or multi-pathway mechanistic models, it is difficult for multi-scale agent-based models to capture cellular and subcellular processes at the same level of details. Moreover, applying such an approach to reconcile biological pathway models with dose-response experiment data remains unexplored.

In this work, we present an algorithm that combines detailed mechanistic descriptions of anti-cancer drug action with lineage tracking-based recording of individual division and death events to construct simulation outputs that are directly comparable to drug dose viability response assays. As a test case we use a previously developed large-scale model of single mammalian cell proliferation and death, SPARCED[174], and add to it mechanistic pharmacodynamic models based on known binding interactions between drugs and modeled targets. First, we describe the algorithm and the types of novel analytics that can be derived, such as cell population

dynamics, cross-generational biomarker tracking for cell lineages and cell population dendrograms. Then, we simulate dose responses to multiple drugs, namely, Alpelisib (PI-3K inhibitor), Trametinib (MEK inhibitor), Palbociclib (CDK4/6 inhibitor) and Neratinib (EGFR inhibitor). The results show agreement with experimental results for strong growth inhibition by Trametinib and overall lack of efficacy for Alpelisib, but substantial discrepancy for Palbociclib and Neratinib. Deeper analyses investigating the reasons for these differences suggests that (i) contemporary belief in the importance of CDK4/6 for driving cell cycle completion is likely to be overestimated, and (ii) the cellular balance between basal (tonic) and ligand-induced ERK signaling is a critical determinant of response to irreversible EGFR inhibitors. This work may serve as a foundation for mechanistic analysis of experimental drug dose viability response data sets.

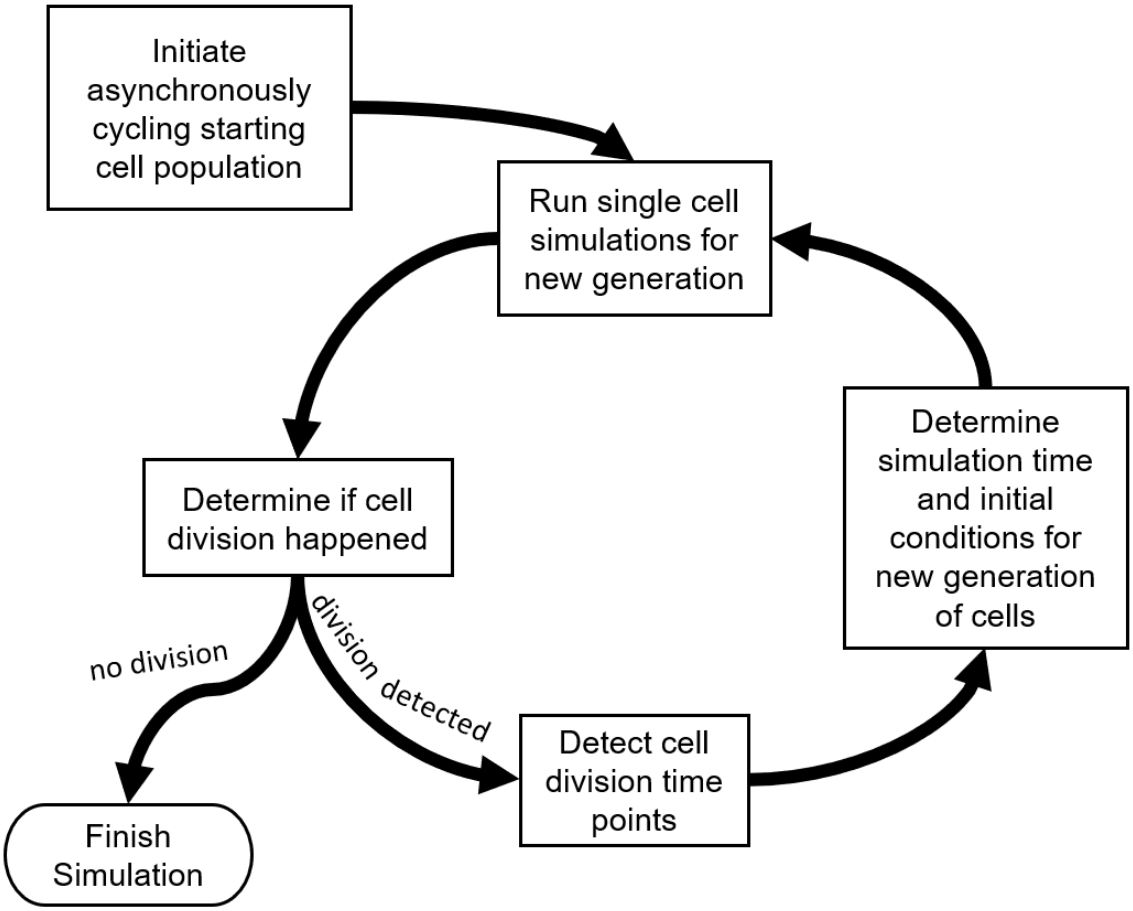


Figure 4.1: Computational workflow of the variable cell population simulation. Generation-specific multiple-step iterations of the stochastic SPARCED single cell model is performed to approximate a dynamic cell population. An asynchronously cycling starting cell population is initiated as the first generation. Upon the execution of each generation, observed division events or lack thereof determine the fate of next generation and variable number of generations are created until no new cell divisions are detected within simulation time.

4.3 Results

4.3.1 Lineage-resolved single-cell simulation framework

We first set out to construct a simulation algorithm that mirrors drug dose response viability assays (Fig. 4.1). These assays typically start with a population of asynchronously cycling cells that are treated with drug for ~3 days and then assayed for final cell number (or a metric proportional to it). The final cell number is related to the number of cell division and death events each initial single cell ultimately experienced. Thus, the simulation algorithm should start with a population of asynchronously cycling cells and be able to count the individual division and death events from each initial cell. Throughout this manuscript, we use our previously published mechanistic model of single cell proliferation and death signaling[174] representing MCF10A breast epithelial cells (SPARCED), although any single-cell resolved model with division and death event readouts is in principle compatible with the below-described algorithm.

We create a simulated population of asynchronously cycling, drug treated single cells via the following (Fig. 4.2). The initial model state is an average, serum-starved cell (non-cycling). The first step is to generate a population of cells with heterogeneous gene expression profiles, enabled by descriptions of intrinsic noise in gene expression as previously described[32]. We refer to this process as heterogenization. After 48 simulated hours (a period of time when the distribution of most protein levels across the cell population stabilizes), the addition of full growth media is simulated (in the case of MCF10A cells and this model—EGF and Insulin). Subsequently, synchronized cell cycle progression is observed in simulations for an additional 48 hours, creating so-called “Generation 0”. To convert these Generation 0 synchronized cells into Generation 1

asynchronously cycling cells, we sample random times during the 48 hour full growth media treatment window for each single cell. These selections become initial conditions for Generation 1, which is then subjected to simulated drug treatment for 72 hours.

Once these simulations are completed, the outputs are analyzed to determine cell division events (based on CyclinB-CDK1 peaks—see Methods) and the time point when each event occurred (Fig. 4.2). Based on the cell division time points, the remaining simulation time and (difference between division time and 72 hours), initial conditions for each daughter cell are determined for the next generation, and lineage information is recorded. Subsequently, simulations for the next generation are run and this cycle continues until no division events occur in a given generation. Detected cell death events (based on cleaved PARP dynamics), halt a lineage.

These simulations not only enable close replication of typical drug dose response experiments but also lineage-resolved analyses. For example, individual division and death events from a parental cell can be tracked (Fig. 4.3A). It also allows dynamic tracking of observables (such as ERK or AKT activity) across multiple generations of any single cell lineage (Fig. 4.3B-C). Such capability may generate hypotheses linking drug sensitivity or resistance with cell fates and lineage, or variations in biochemistry that predispose cells to response or resistance.

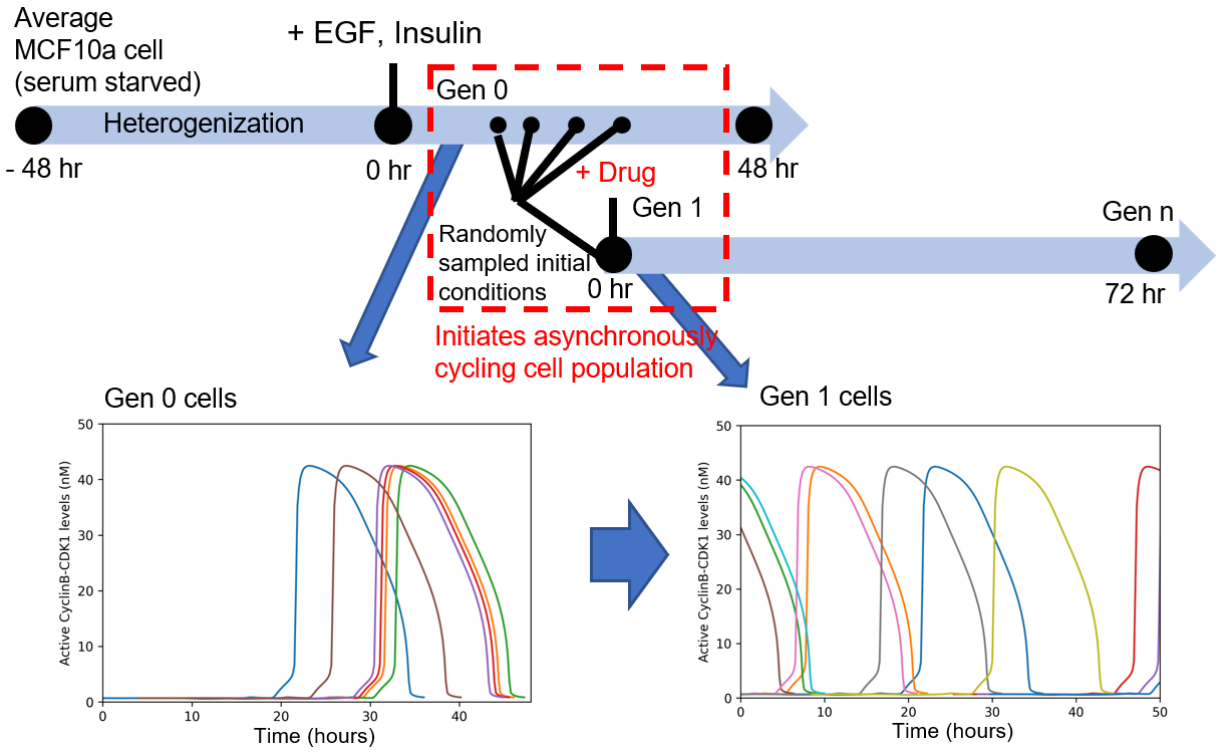


Figure 4.2: Initiation of asynchronously cycling variable cell population. Starting cell population is initiated using a pool of single cell simulations run with growth factor stimulation (generation 0), from which initial conditions are sampled from random time points to generate an asynchronously cycling cell population (generation 1). Generation 0 cells are initiated from a serum starved condition and can only enter cell cycle after a certain period of incubation, which is evident from their active CyclinB-CDK1 trajectories. Generation 1 cells may undergo various stages of cell cycle at the start of simulation time, which resembles the condition of cells in a dose response assay.

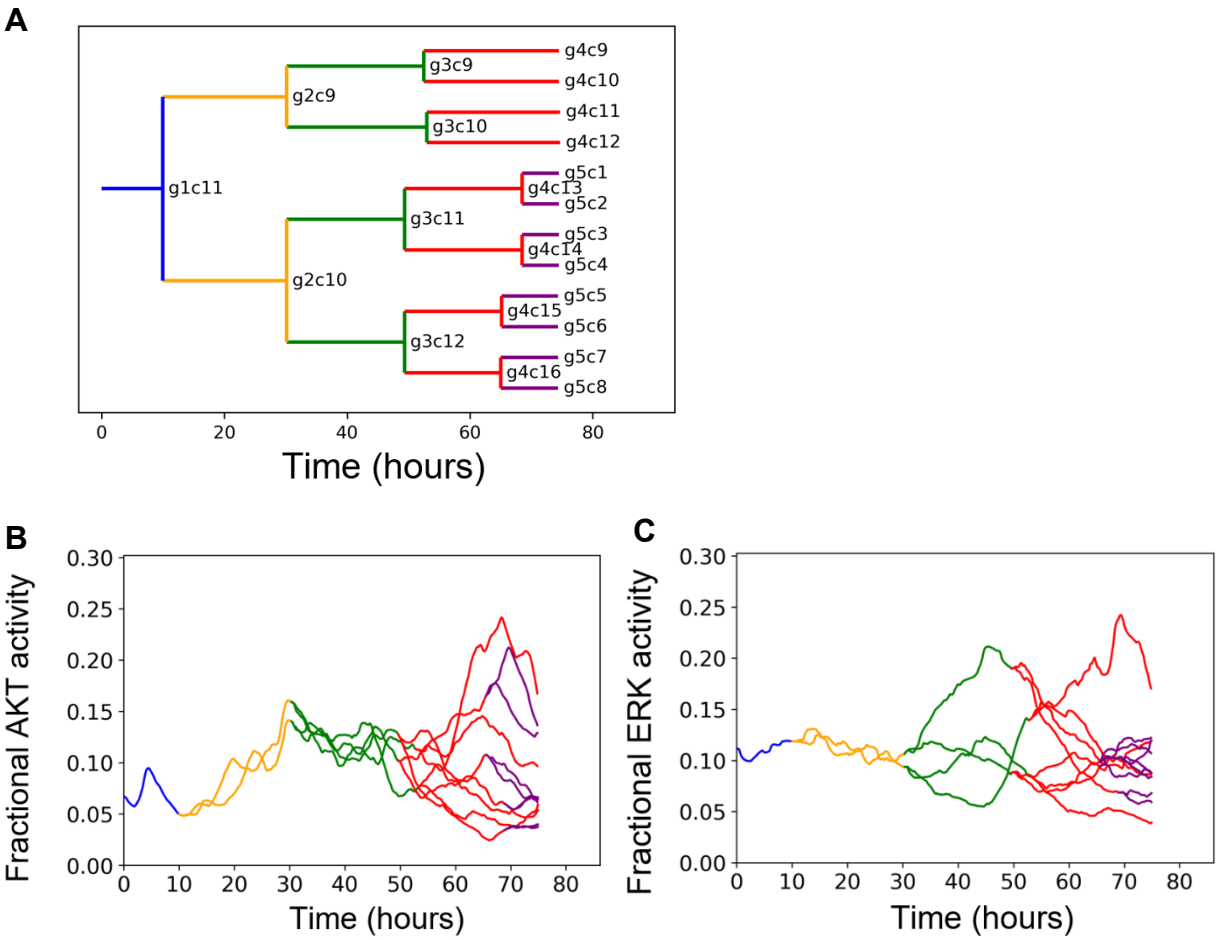


Figure 4.3: Visualizations generated from cell population simulations (A) In-silico lineage tracing capability demonstrated with a single cell lineage tree. Lines representing individual cells are labeled with generation and index. Changes in color indicate inception of a new generation at that time point. Here, blue = generation 1, yellow = generation 2, green = generation 3, red = generation 4, purple = generation 5 (B,C) Cross-generational trajectory of observed ERK and AKT activity from a generation 1 single-cell lineage, same color code as the cell lineage tree is applied to these visualizations as well.

4.3.2 Comparing Simulated Drug Dose Responses to Experimental Measurements

Now with an algorithm that enables single-cell mechanistic models to be simulated in ways analogous to drug dose viability response experiments, and a model (SPARCED) that describes mechanisms of cell proliferation and death signaling in MCF10A cells, we could compare model predictions to such experimental data[175–178]. Specifically, we focused on four previously-studied, targeted anti-cancer drugs for which our model includes primary and significant off-targets: Trametinib (MEK inhibitor), Alpelisib (PI-3K inhibitor), Neratinib (EGFR inhibitor), and Palbociclib (CDK4/6 inhibitor)[179].

First, we extended SPARCED by including known drug interactions with protein-level target species and leveraging previously described capabilities to robustly and easily increase model scope[174]. literature resources[180–183] were utilized to retrieve information about the binding affinity of the selected drugs for known proteins (which are included as model species) representing their on-target and off-target effects. Reactions depicting the drugs interacting with these proteins to form complexes were included in the model with parameter values calculated from the binding affinity data. These drug-protein complex formations represent the pharmacodynamics of individual drugs in terms of the extent to which they perturb the modeled biological pathways. Specific details of individual drugs and their actions as well as validations by simulation tests have been included in the methods section.

We then performed lineage-resolved simulations for various doses of the modeled drugs for which experimental data are available[179], with no adjustment to the SPARCED model. This framework allows direct simulation of the dynamic cell population in response to drug doses (Fig. 4.4A). Also, the effect of drug action over cell lineages can be visualized by dendrogram plots, with horizontally connected lines corresponding to one of the initial 100 single initial cells (Fig. 4.4B-D). The simulation outputs were used to calculate dose response using the growth-rate inhibition[184] metrics which is the same method applied to the available experimental dataset, allowing direct comparison of experimental and simulation results (Fig. 4.5). The simulation results for Trametinib (Fig. 4.5) demonstrate surprising agreement with experimental data. The simulations also captured the overall lack of efficacy for Alpelisib (Fig. 4.5), although there are some slight deviations between experiment and simulations yet to be explained. On the other hand, predicted Palbociclib and Neratinib responses were substantially different from experiments, indicating significant knowledge gaps in SPARCED and perhaps in the general signaling literature, which we investigate in the subsequent sections.

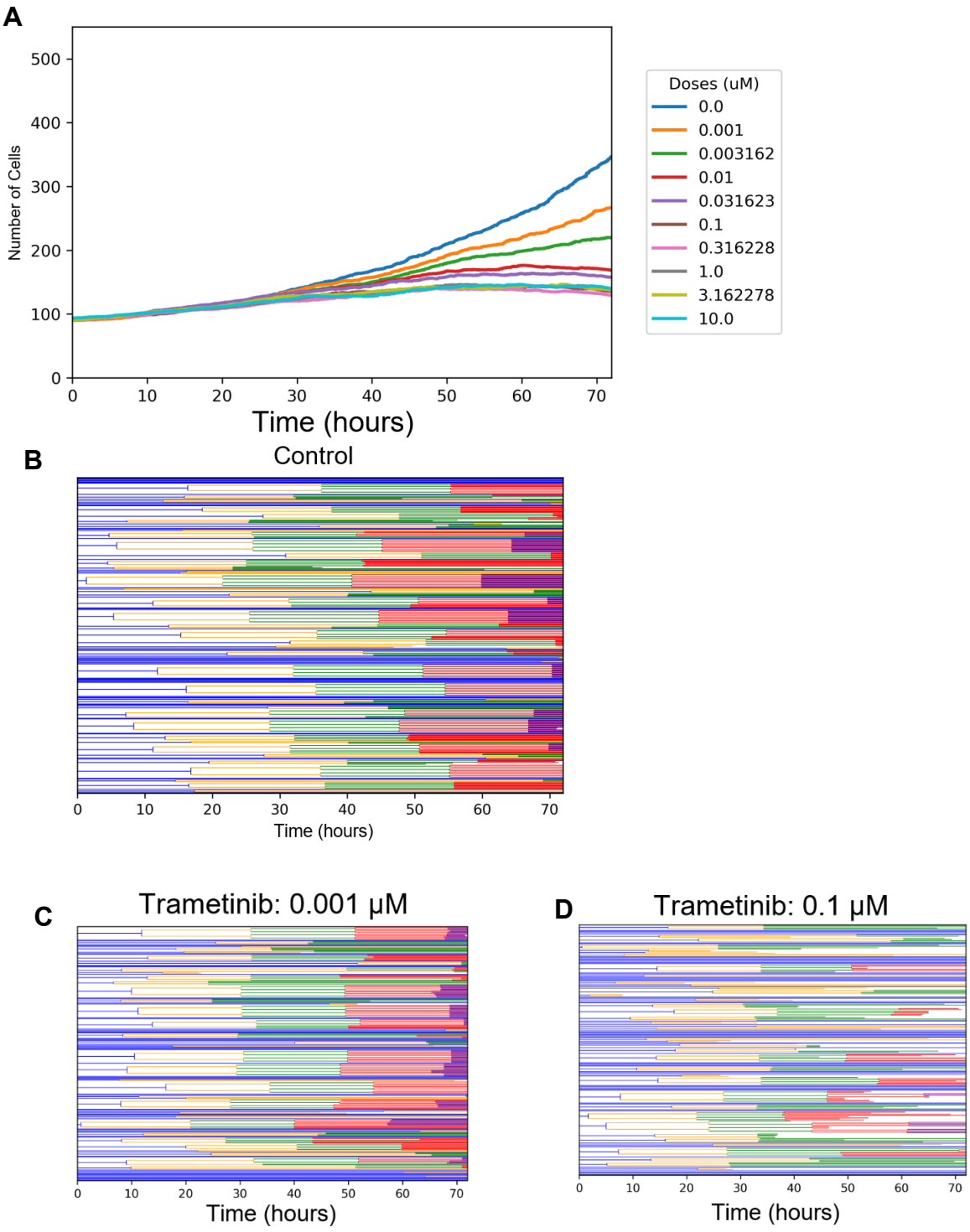


Figure 4.4: (A) Median cell population dynamics resulting from dose response simulations of an example drug (Trametinib)
(B,C,D) Cell population dendrogram from varying doses of Trametinib including control (B), low (C) and moderate (D) doses.

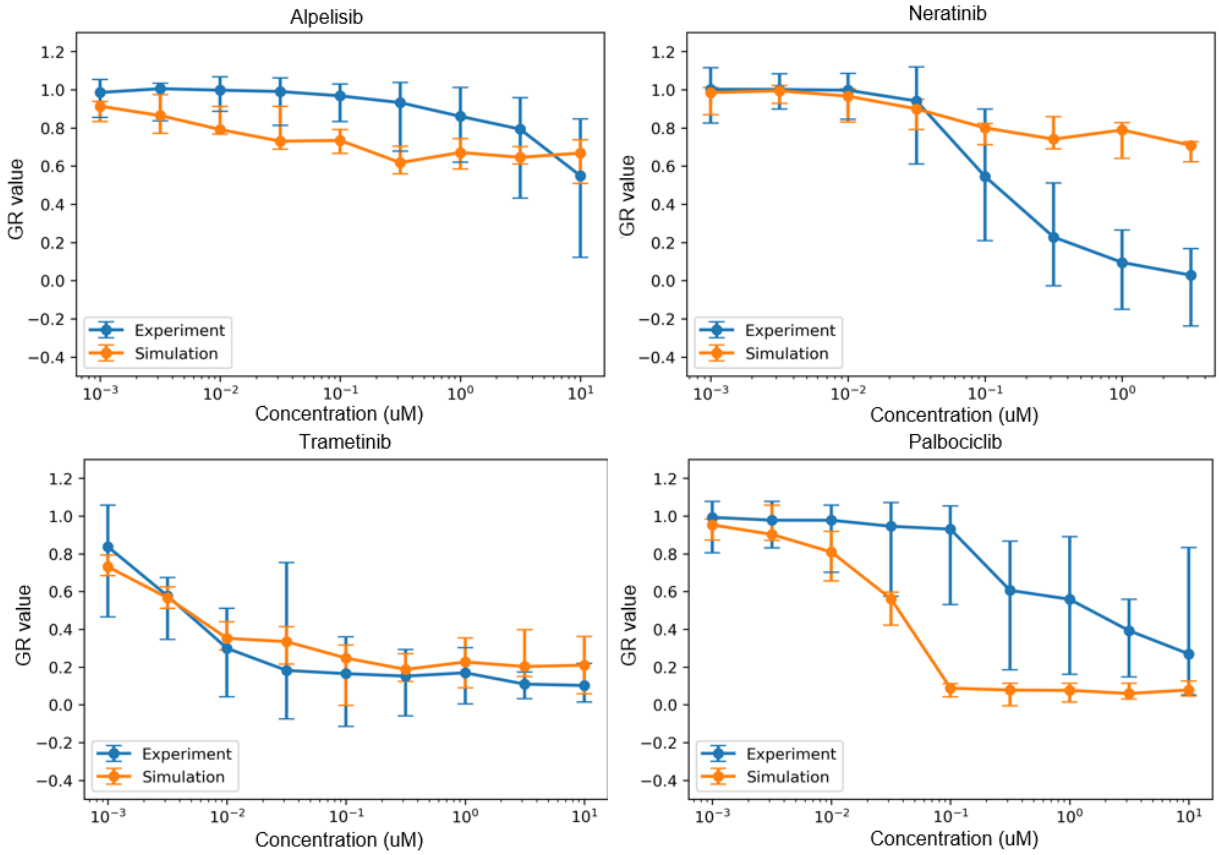


Figure 4.5: Simulated dose response measured in GR-value for four drugs compared to their experimental counterparts.

4.3.3 Palbociclib Dose Response Discrepancies Suggests CDK4/6 is Partially Redundant for Cell Cycle Progression

What could explain the experiment/simulation discrepancy for Palbociclib, a potent inhibitor of CDK4/6, a central mediator of cell cycle progression from G_0 and G_1 to S-phase (Fig. 4.5)?

Simulated palbociclib dose response starts to deviate from the experimental results at doses as low as 0.01 μM. Above 0.1 μM, the simulated dose response shows complete cytostasis. On the other hand, experimental results show minimal growth inhibition at 0.1 μM. Between doses 0.3 μM and 3.3 μM, median GR-scores from the experimental results are between 0.6 and 0.2, which indicates only a partial growth inhibition even at high doses of Palbociclib. The simulated potency of Palbociclib (Fig 4.6) suggests that at doses 0.3 μM and above, more than 80% of the target should be bound to the drug at any time point after 5 hours. Since the modeled drug-target interaction is based on the binding affinity observed in experimental drug-target binding assays, we can expect a similar potency in dose response experiments. The expected result of such potency as per the modeled function of CDK4/6 in cell cycle progression is a population-wide cytostasis at moderate and higher Palbociclib doses whereby only a negligible number of cells could be expected to complete cell cycle. The population dendrogram visualization for simulations confirms this (Figure 4.7B). This is clearly contrary to the experimental observations, which implies that the continuation and completion of cell cycle is less reliant on the activity of CDK4/6 than what the model estimates. The observation suggests the presence of intrinsic resistance mechanisms, bolstering the robustness of cell cycle to CDK4/6 inhibition. Such a mechanism could be

an undiscovered negative feedback transcriptional regulation of the drug target, or a feedforward regulation of a downstream activator of cell cycle which might explain how cell cycle can continue even when one of its key activators is inhibited.

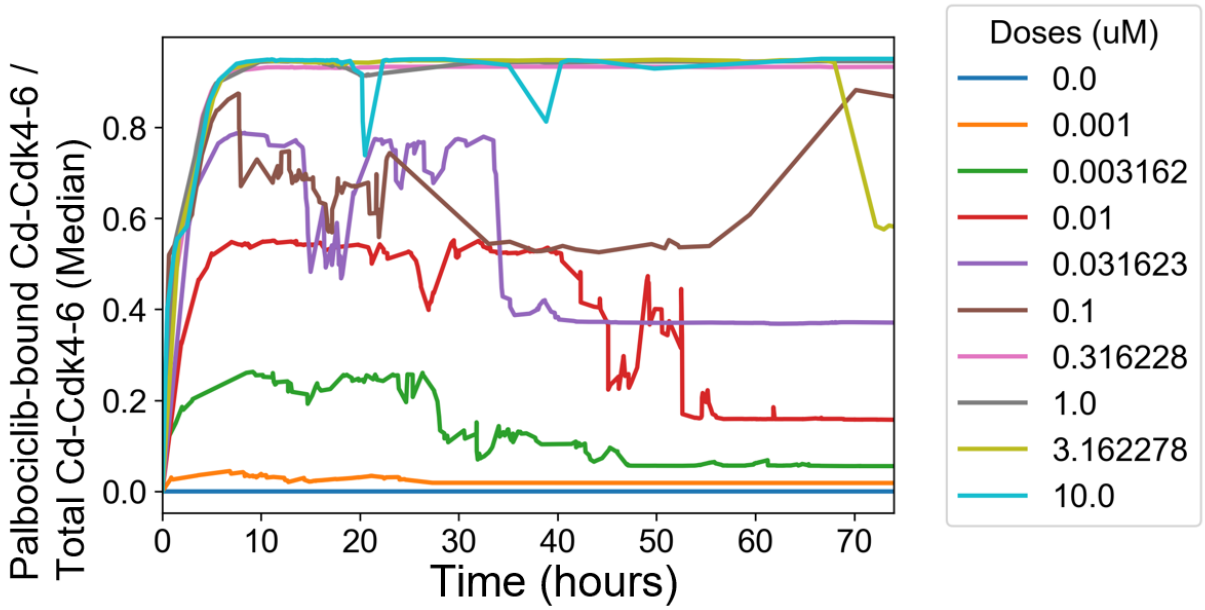


Figure 4.6: Investigation into Palbociclib dose response - Observed target engagement activity for various doses of Palbociclib

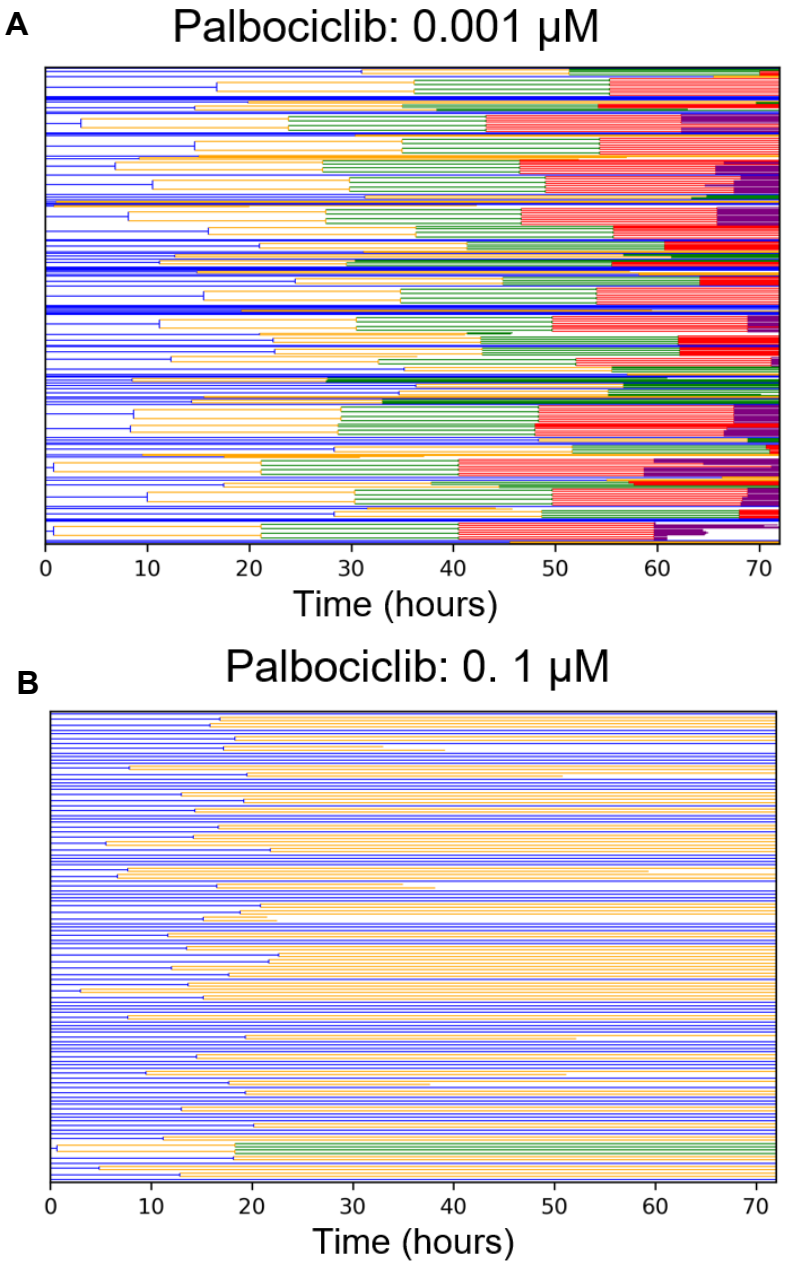


Figure 4.7: Cell population dendrograms for low (A) and moderate (A) Palbociclib doses

4.3.4 The Balance of Tonic Versus Ligand-Induced Growth Factor Signaling is Critical for Capturing Drug Effects

Neratinib is an irreversible inhibitor of the EGFR (with some off-target activity for the closely related ErbB2/HER2, Fig. 4.8A), a receptor tyrosine kinase that, upon ligand binding, activates the pro-proliferative and survival ERK and AKT pathways[185–187]. Hence, the drug action is expected to block ERK and Akt signaling when a ligand, such as EGF, binds to EGFR. The experimentally reported Neratinib dose response on MCF10A cells (Fig. 3F) show strong growth inhibition at doses above 0.1 μM and complete cytostasis at 3.16 μM. However, simulation-predicted growth inhibition within this range is significantly weaker, despite complete target engagement (Fig. 4.8B).

To explain this discrepancy, we considered that the current modeled balance of ligand-induced versus basal (also called tonic) signaling ERK signaling could be incorrect. Specifically, that basal ERK signaling was too strong and causes non-negligible proliferation in the absence of EGF. If cell cycling is initiated by basal signaling too strongly, coupled with the fact that Neratinib cannot inhibit basal signaling, this could explain some of the model-experiment discrepancy.

MCF10A cells are dependent upon EGF for cell cycle progression[188,189]. Thus, in simulations, cells dividing without EGF would support the above explanation. In lineage-resolved simulations where the growth media contained only insulin, numerous cell division events were observed (Fig. 4.9). Since the proliferative signaling activity that caused these divisions did not originate as a result of EGF-EGFR activity, during Neratinib dose response simulations these will be unaffected. This is inconsistent with

the experimentally observed cell behavior and hence may be a major cause of mismatch between simulation and experiment.

How could the model be changed to account for these mismatches? First, we attempted to ensure that basal ERK signaling in the presence of insulin minimally induces cell cycle progression. The basal Ras-GDP to Ras-GTP exchange rate (Fig. 4.10A) is the main reaction controlling basal ERK activity in the model. We reduced the value of the associated rate constant until the probability of cell division in the absence of EGF and presence of insulin was near zero (Fig 4.10B), and then simulated neratinib dose response again (Fig. 4.11). The new simulated neratinib dose responses show closer alignment with experiments. However, for all other drugs, experiment-model agreement became significantly worse, most likely now because the absolute levels of EGF-induced ERK signaling are altered. This result shows the close interacting nature of signaling mechanisms in the model for influencing broad features of drug response and cautions against developing models without considering comparison to a compendium of data.

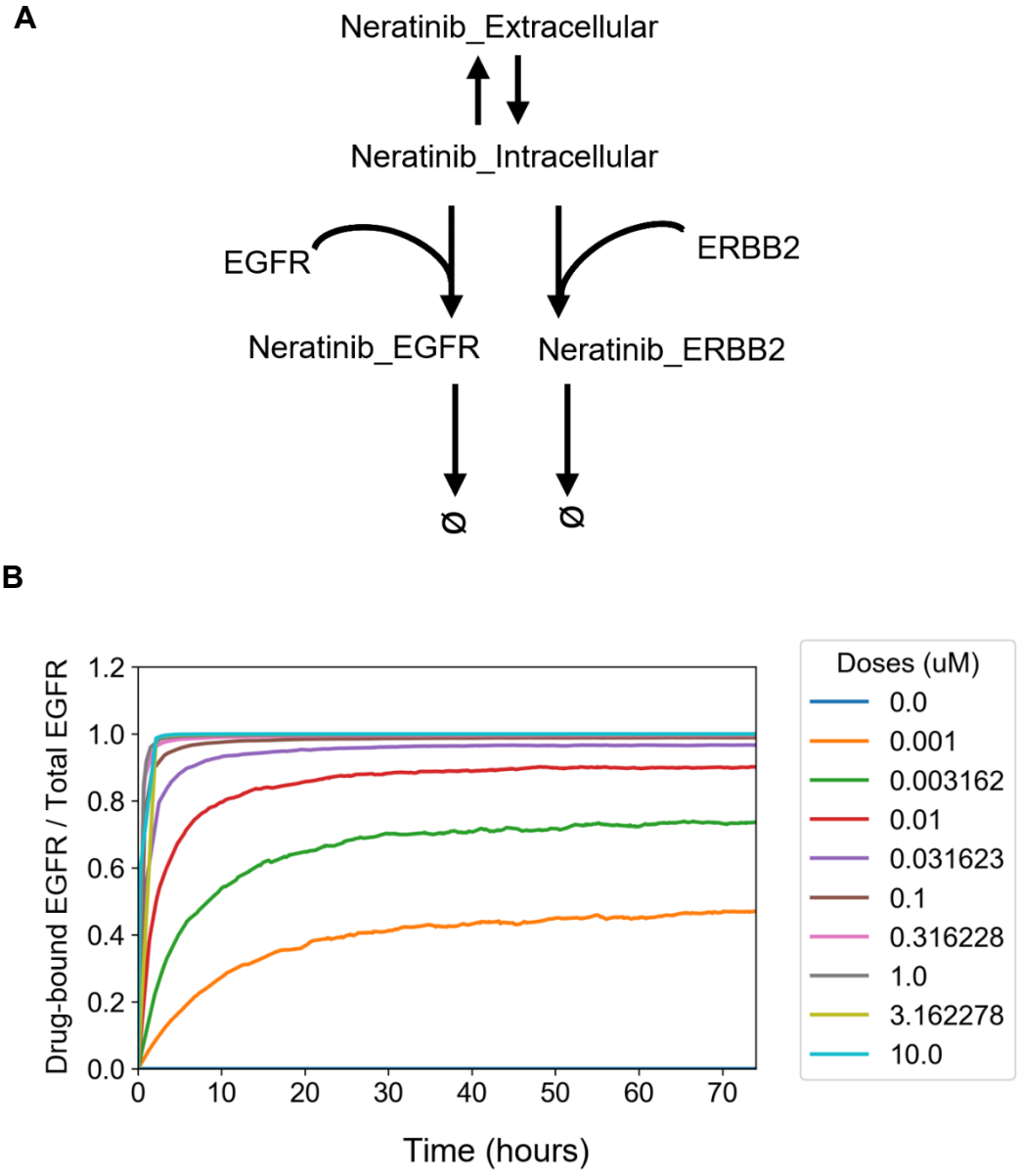


Figure 4.8: Investigation into Neratinib dose response (A) Neratinib drug action (B) Dynamic population median ratio of drug-bound EGFR to total EGFR amount showing expected target engagement due to modeled drug action

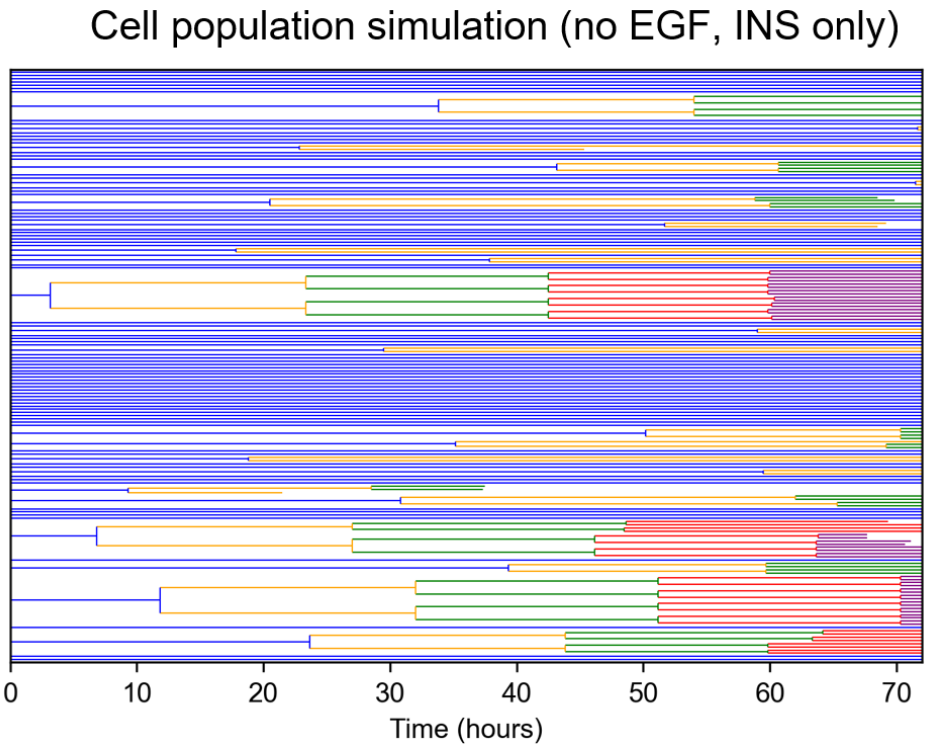


Figure 4.9: Cell population dendrogram from a simulation whereby the population was simulated only with INS in absence of EGF. Results indicate that cells are able to enter cell cycle without ligand induced ERK proliferative signaling.

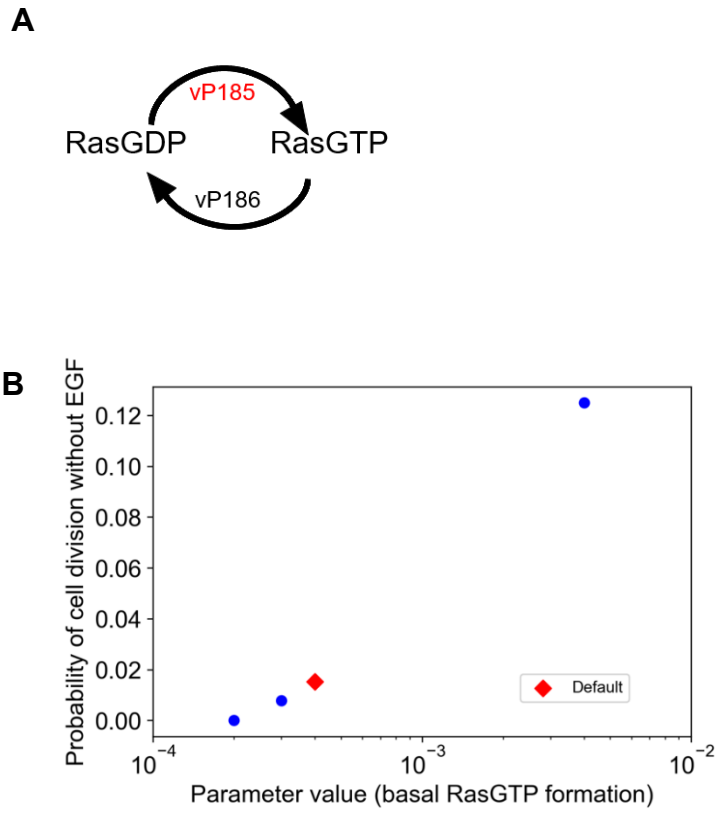


Figure 4.10: (A) Partial kinetic scheme showing the reaction rate parameters that can tune basal ERK activity.

(B) Cell division probability observed in single cell stochastic simulations as a function of basal RasGTP transformation rate

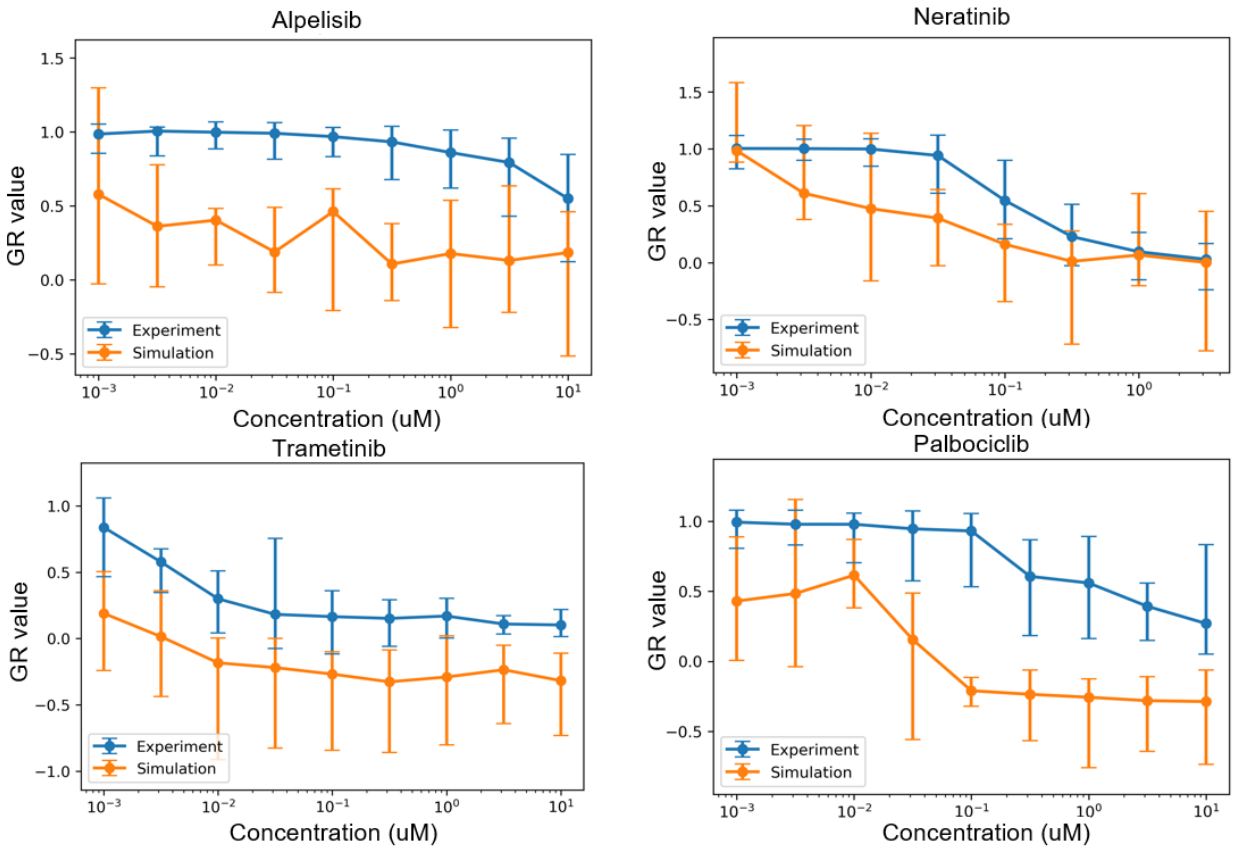


Figure 4.11: Alteration in the SPARCED model to address discrepancy in Neratinib dose response simulation. Resulting dose response in simulation compared to experimental results

4.4 Methods

4.4.1 SPARCED Pharmacodynamic Model

Reactions representing drugs binding to their reported targets with mass action rate laws were added to the SPARCED model. The assumptions and included drug actions for each individual drug are described below. To validate whether the individual drug action models generated the intended effect, we ran deterministic simulations for each drug. In these simulations, growth factors were added to a single cell incubated with a fixed dose of the drug or control, and observed the simulated trajectory of the drug, drug bound protein targets and certain downstream effectors.

Alpelisib: Alpelisib enters and leaves the cell with first order kinetics and binds reversibly to its intracellular targets, free PI3K and catalytic subunit bound PI3K dimer[180]. Furthermore, we assume the binding of Alpelisib to its target prevents its dimerization (to the regulatory subunit), but it can also bind a dimerized target. Reactions representing Alpelisib drug action are :

Free PI3K + Alpelisib => Alpelisib bound PI3K

Alpelisib bound PI3K1 => Free PI3K + Alpelisib

Alpelisib bound PI3K1 => Ø

PI3K dimer + Alpelisib => Alpelisib bound PI3K dimer

Alpelisib bound PI3K dimer => PI3K dimer + Alpelisib

Alpelisib bound PI3K dimer => Ø

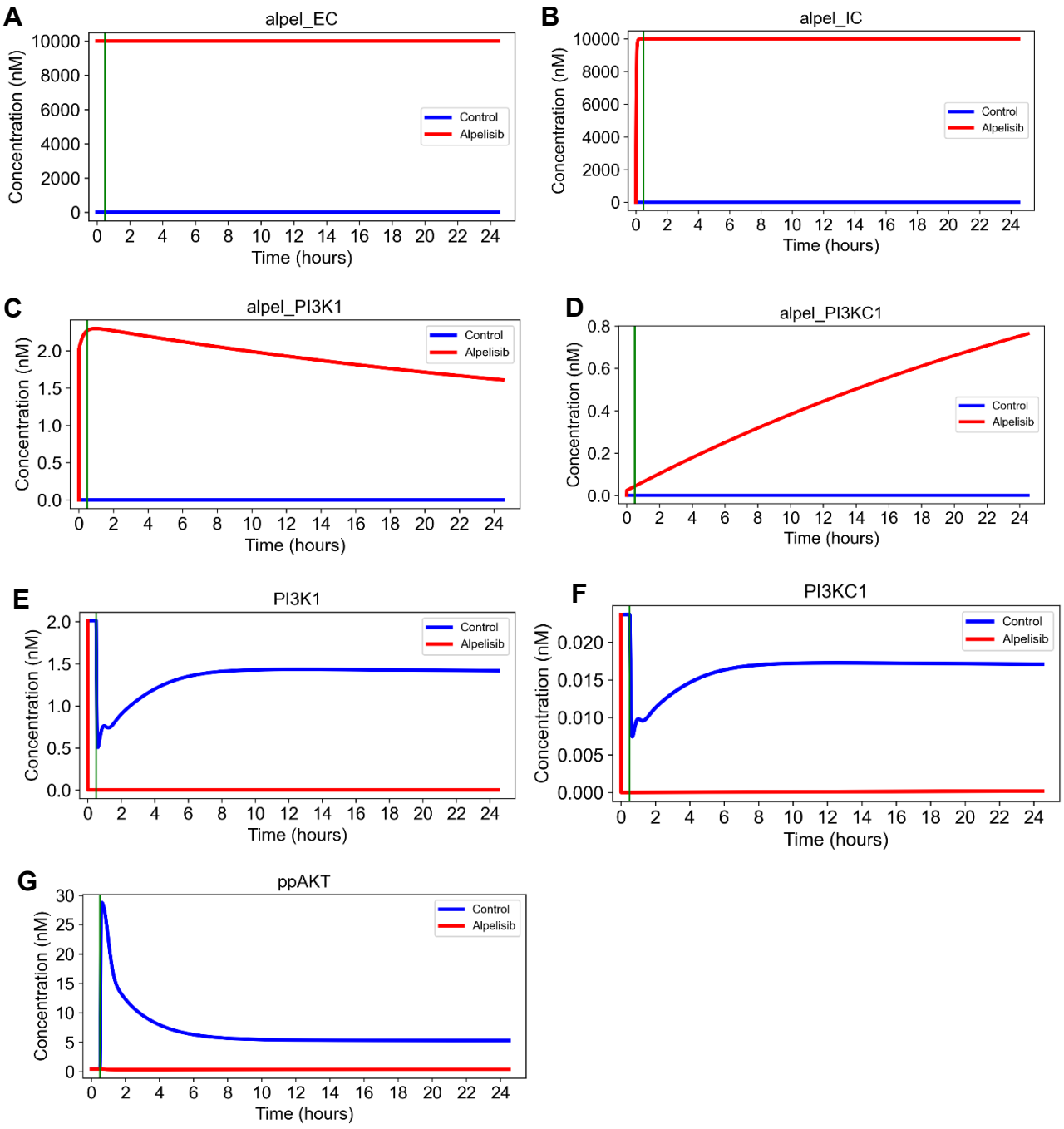


Figure 4.12: Validation simulation results for Alpelisib drug action. Results have been generated with single cell deterministic simulations in the MCF10A context. Serum-starved cell has been incubated for 30 minutes with 10 μM Alpelisib dose or control condition before addition of 3.3 nM EGF, 0.005 nM HGF and 1721.0 nM insulin. Simulation results confirm extracellular (A) and intracellular (B) distribution of Alpelisib dose, binding of Alpelisib with modeled targets (C,D) and reduction of free target concentrations (E,F), and reduction of downstream AKT activity due to Alpelisib drug action (G)

Palbociclib: Palbociclib enters and leaves the cell and nucleus with first order kinetics. Once inside the nucleus, it reversibly binds to its target, intranuclear CDK4/6[181]. A target bound to Palbociclib loses its phosphorylation activity. Reactions representing Palbociclib drug action are:

Palbociclib + Cd-CDK4/6 => Palbociclib bound Cd-CDK4/6

Palbociclib bound Cd-CDK4/6 => Palbociclib + Cd-CDK4/6

Palbociclib bound Cd-CDK4/6 => Ø

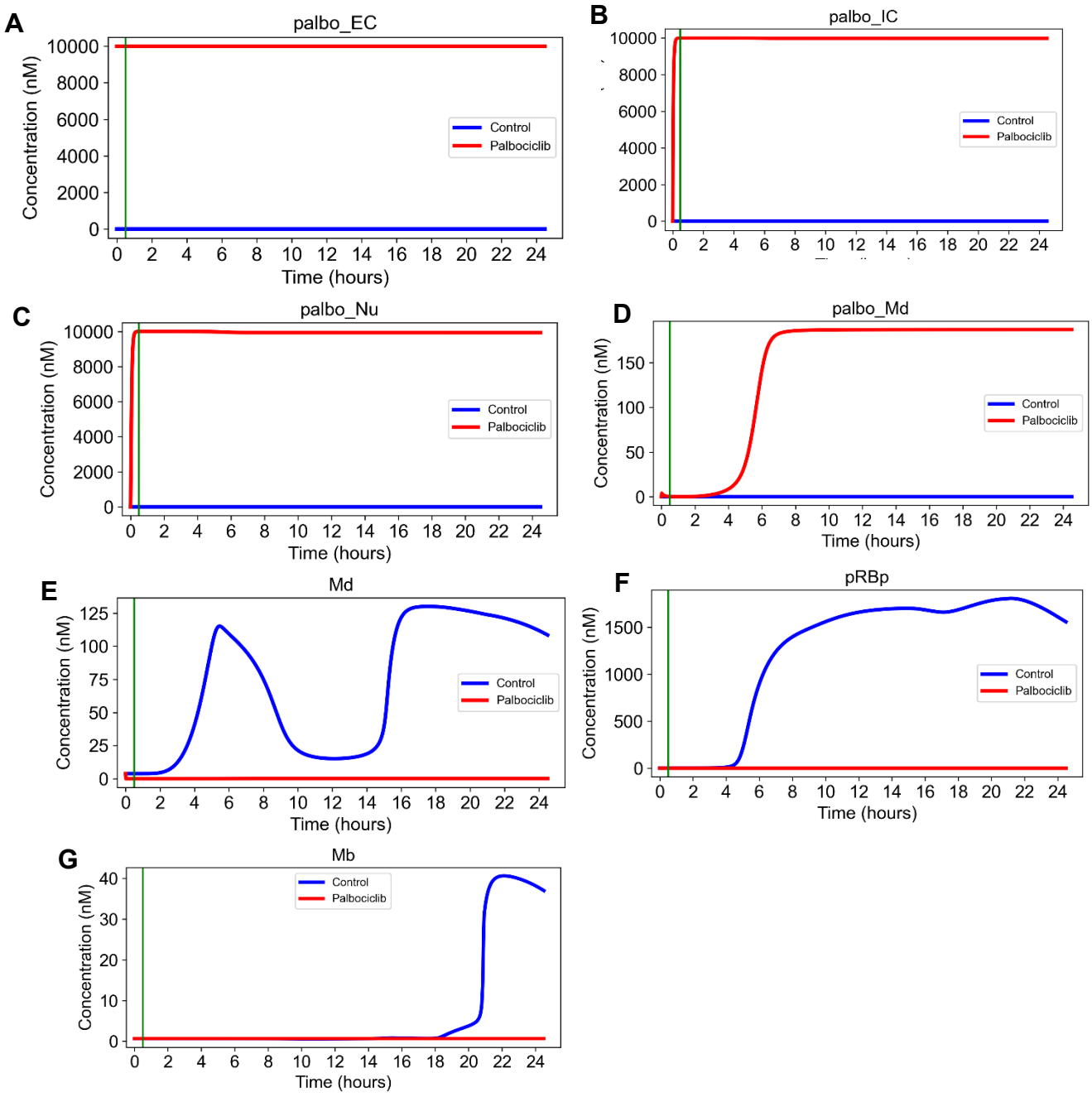


Figure 4.13: Validation simulation results for Palbociclib drug action. Results have been generated with single cell deterministic simulations in the MCF10A context. Serum-starved cell has been incubated for 30 minutes with 10 μM Palbociclib dose or control condition before addition of 3.3 nM EGF, 0.005 nM HGF and 1721.0 nM insulin. Simulation results confirm extracellular (A) and intracellular (B) and nuclear (C) distribution of Palbociclib dose, binding of Palbociclib with modeled target active cyclin D-CDK4/6 (D) and reduction of free target activity (E), and reduction of downstream pRB phosphorylation (F) and eventual halting of cell cycle (lack of active cyclin B/CDK1 peak) (G)

Trametinib: Trametinib enters and leaves the cell with first order kinetics. Once inside cytoplasm, it reversibly binds to its target, unphosphorylated MEK[182]. Furthermore, we assume binding of Trametinib to its target prevents its dimerization and phosphorylation and a dimerized or phosphorylated target cannot bind with Trametinib. The reactions representing Trametinib drug action are:

Trametinib + MEK => Trametinib bound MEK

Trametinib bound MEK => Trametinib + MEK

Trametinib bound MEK => Ø

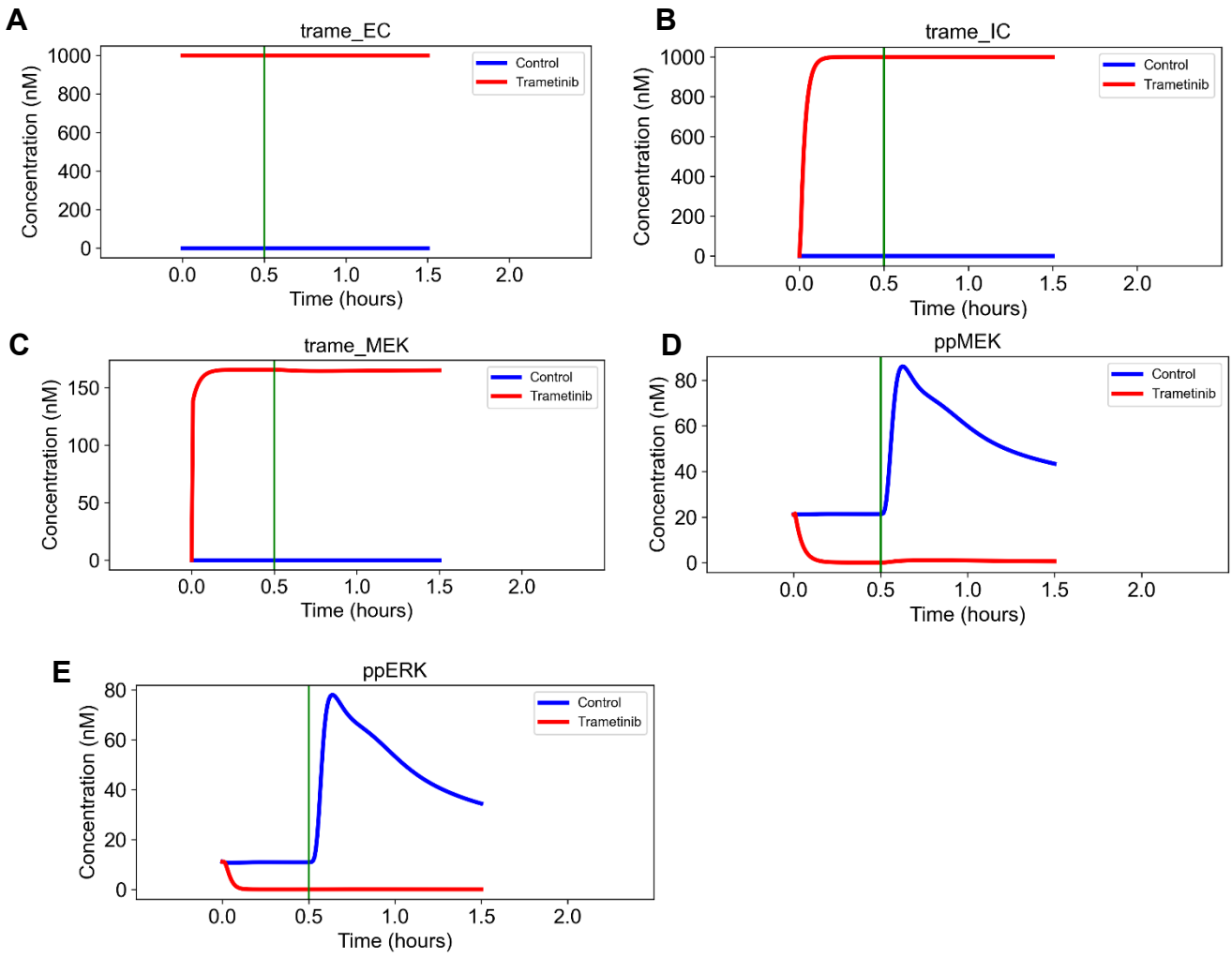


Figure 4.14: Validation simulation results for Trametinib drug action. Results have been generated with single cell deterministic simulations in the MCF10A context. Serum-starved cell has been incubated for 30 minutes with 10 μM Trametinib dose or control condition before addition of 3.3 nM EGF, 0.005 nM HGF and 1721.0 nM insulin. Simulation results confirm extracellular (A) and intracellular (B) distribution of Trametinib dose, binding of Trametinib with modeled target MEK (C) and reduction of MEK activity (D) and ERK activity (E)

Neratinib: We assume Neratinib enters and leaves the cell with first order kinetics. Neratinib binds to its targets of ErbB family of receptors (EGFR, HER2, HER4) on their intracellular domain[183]. Binding of Neratinib to its target is irreversible. It prevents the target receptors from binding EGF and a receptor simultaneously cannot bind Neratinib and EGF. The reactions representing Neratinib drug action are:

Neratinib + EGFR => Neratinib bound EGFR

Neratinib bound EGFR => Ø

Neratinib + HER2 => Neratinib bound HER2

Neratinib bound HER2 => Ø

Neratinib + HER4 => Neratinib bound HER4

Neratinib bound HER4 => Ø

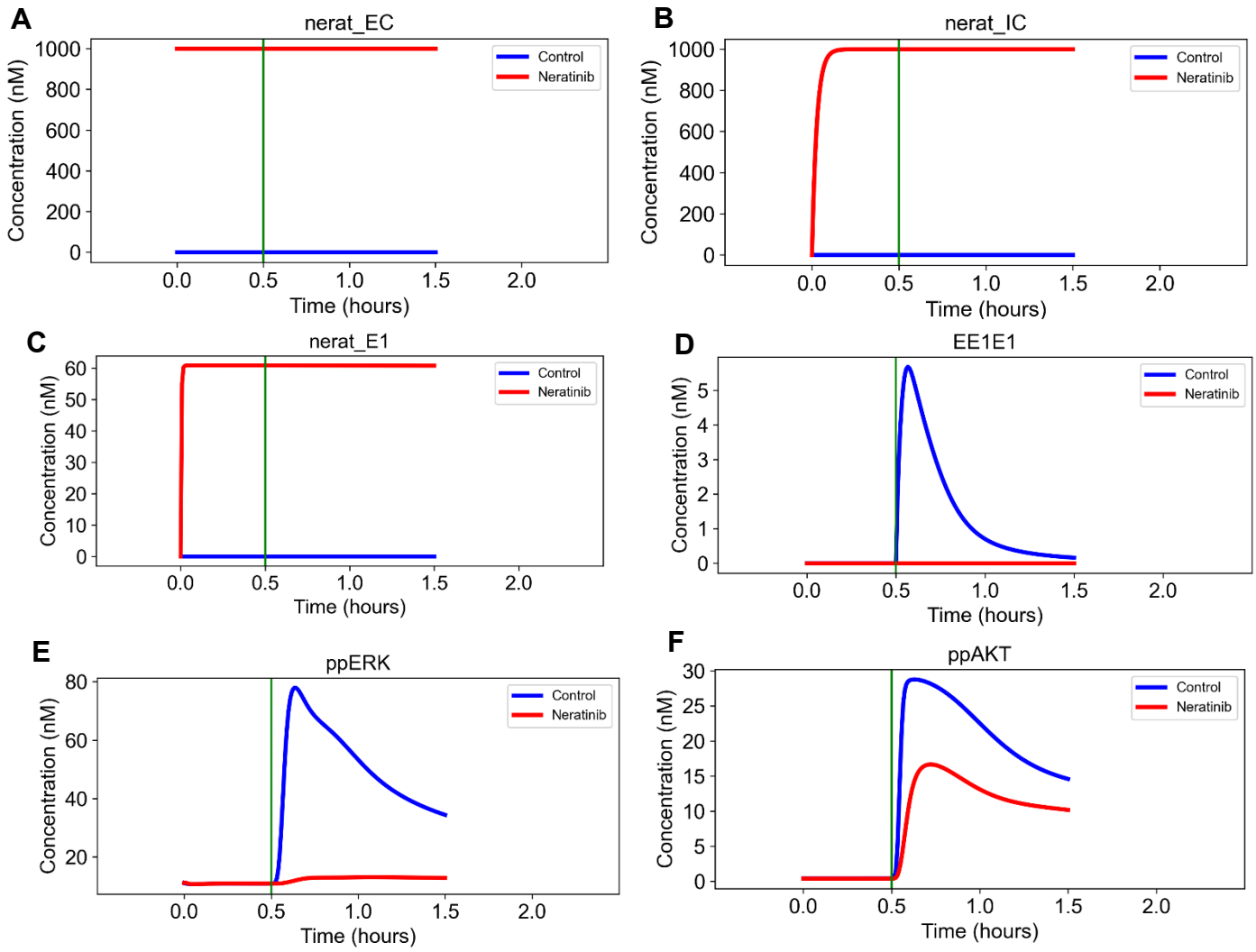


Figure 4.15: Validation simulation results for Neratinib drug action. Results have been generated with single cell deterministic simulations in the MCF10A context. Serum-starved cell has been incubated for 30 minutes with 10 μM Neratinib dose or control condition before addition of 3.3 nM EGF, 0.005 nM HGF and 1721.0 nM insulin. Simulation results confirm extracellular (A) and intracellular (B) distribution of Neratinib dose, binding of Neratinib with one of its modeled target EGFR (C) and the resulting reduction of EGF-EGFR dimer formation (D), which leads to reduction of ERK (E) and AKT (F) activities.

4.4.2 Lineage-Resolved Simulations

Asynchronous population: Cell population simulations are initiated by creating a representation of an asynchronously cycling cell population. The starting cell population is specified by the user. For each starting cell, initial conditions representing an average serum-starved MCF10A cell is used to create a heterogenized cell population[32]. Then, a growth media with doses of EGF (3.3 nM) and insulin (1721 nM) is introduced and stochastic simulations are run for each individual cell for 48 hours. From the generated state trajectories, a timepoint is randomly selected using the NumPy random number generator for each cell to be used as the initial condition of a first generation of cell. For the user specified experiment time (typically 72 hours), single cell simulations are executed for all first generation cells.

Identifying cell division events: Once the single cell simulations are completed, the generated outputs are analyzed to determine cell division events. The cell division events are detected using troughs observed in Cyclin B-CDK1 trajectories. For any individual cell, if a division event is detected, timepoints after the occurrence of cell division events are discarded and the state vector at the time of cell division is selected as the initial condition for two new second generation cells.

Identifying cell death events: As a readout for cell death, we look at the trajectory of DNA repair enzyme PARP, which gets cleaved as a result of apoptotic signaling. For any single cell, if more than half of PARP has been cleaved at any time point, the cell is proclaimed dead at that time.

Subsequent generations: For each generation, we refine the protein level outputs with division and death events. For every cell, we scan the output for the duration of its lifetime to find division events. The division time point is subtracted from the total simulation time to determine required second generation simulation time. The single cell outputs at the time point of division of each mother cell is recorded as initial conditions for next generation of daughter cells. Thus, we define the required simulation time, population, and initial conditions for the next generation of cells. This process is repeated for the subsequent generations of cell populations. In a given generation, if there is no cell division event observed within the simulation time, the population simulation is terminated.

Software: The computation for this simulation is performed using HPC-compatible parallel processing in Python whereby single cell simulations are run in individual CPU threads. The pipeline for the simulation is available in the latest release of the SPARCED git repository (github.com/birtiwistlelab/SPARCED). To run the cell population simulation, a computational environment with an implementation of MPI (Message Passing Interface), such as OpenMPI[190] on Linux and MSMPI on Windows systems needs to be set up in addition to the dependencies of the SPARCED model pipeline. Before the simulation can be performed, the SPARCED model is to be built using the python script under scripts/createModel.py, which creates an executable single cell model based on the specifications in the input files. Once the model build process is complete, MPI can be used to run cell population simulation using the following command:

```
mpirun -np [n_cpu] python cellpop_drs.py --arguments [argument_value]
```

Here, n_cpu is the number of CPU threads that the user may decide to use for parallelization. Also, the following arguments are to be passed to the python script for specification of simulation parameters:

sim_name: An arbitrary string defined by the user to create a directory under sparced/output where simulation outputs will be saved.

cellpop: An integer specifying the number of starting cells for simulation

exp_time: Duration of experiment in hours

drug: String specifying species name for the drug of interest

dose: Applied concentration of the drug in μM

egf: Serum EGF concentration in nM

ins: Serum INS concentration in nM

hgf: Serum HGF concentration in nM

nrg: Serum Heregulin concentration in nM

pdgf: Serum PDGF concentration in nM

igf: Serum IGF concentration in nM

fgf: Serum FGF concentration in nM

Upon completion of simulations, the results are saved to disk as Python “pickle” objects.

4.4.3 Visualization

GR Score Calculation

Dose response from cell population simulation has been calculated using the Hafner-Niepel growth rate inhibition metric (GR)[184]. Dose response simulations were run for 10 dose-levels matching experimental data for each drug and 10 replicates of each dose. Outputs from the cell population simulations were read and analyzed to determine total number of living cells over time for the duration of experiment time. Using the number of living cells at 72 hours, the GR scores were computed for each replicate was computed with the Python script provided as part of GR-metrics git repository.

4.5 Discussion

We developed a cell population simulation framework depicting emergent population level outcome as an aggregate of single cell events. In this work, single cell events are dictated by a pan-cancer driver pathway model incorporating stochastic cell proliferation and death in response to drug actions. This framework enables us to establish a crucial link between the collective understanding of cell signaling biology and data from dose response experiments.

We further evaluated the applicability of the cell population simulations by virtually replicating dose response experiments. The results showed some significant mismatches with experiments for palbociclib and neratinib, which uncovered some important findings about the current knowledge of signaling pathways. For palbociclib, the simulations overpredicted its efficacy, showing very high growth inhibition at moderate doses to complete cytostasis at high doses which underlines the indispensability of the drug target, CDK4/6 as per the current knowledge of cell cycle pathway. However, in experiments, even the highest doses could result in only partial growth inhibition, indicating dependence on CDK4/6 for cell cycle completion is likely to be overstated. It suggests undiscovered mechanisms in the cell cycle pathway that may dispense the reliance on CDK4/6 when it is inhibited.

The role of CDK4/6 in cell cycle regulation is associated with facilitation of traversing the cell cycle restriction point[191]. When the cell is in a senescent state, one of the regulators of restriction point, Rb is bound to a key transcription factor of the cell cycle process, E2F. In presence of a growth stimulus, CDK4/6 is activated when bound to cyclin D. This activated cyclin D-CDK4/6 complex can phosphorylate to inactivate Rb,

which then releases E2F. Subsequently, there is an upregulation of E2F which then mediates S phase entry and progression by activating cyclin E, and cyclin A. Drug action mechanism of CDK4/6 inhibitors such as Palbociclib attempt to induce cytostasis by preventing the inactivation of Rb by CDK4/6[192]. As per the experimental results, MCF10A cells may possess resistance mechanism that can compensate for this effect. One of the known resistance mechanisms of CDK4/6 inhibition is the loss of Rb function[193,194]. However, since MCF10A cells do not harbor such mutations, it is an unlikely explanation in this case. Another reported resistance mechanism in cancer cells is the overexpression of Cyclin E[195,196], which is a regulator of the later stages of cell cycle. If a cell cycling cell is past its restriction point at the time of CDK4/6 inhibition, overexpression of later stage cell cycle regulators may explain why actions of CDK4/6 was no longer necessary to complete the cell cycle. Even though this mechanism was observed in tumor cells, an intrinsic feed forward control between CDK4/6 and cyclin E could potentially explain resistance of CDK4/6 inhibition in a non-tumorigenic context such as MCF10A cells. Another possibility is a lack of details in the modeled function of CDK4/6. In the SPARCED single cell model, the cell cycle submodel was adopted from an earlier work of Gerard and Goldbeter[27]. Even though this model describes the functionality of activated cyclin D-CDK4/6 and its action on Rb, it does not explicitly account for the expression of CDK4/6. Therefore, if a possible resistance mechanism exists which relies on transcriptional regulation of CDK4/6, it will not be possible to capture its effect without a major revision of the cell cycle model.

For neratinib, the simulations underpredicted its efficacy, showing weak inhibition for moderate to high doses whereas the experiments showed significant growth

inhibition to complete cytostasis within this range. Deeper analysis of simulation results showed cell cycle occurring in absence of ligand induced ERK signaling which explained why the modeled drug action based on inhibition of ligand-receptor interaction was unable to replicate the growth inhibition observed in experiments. MCF10A cells are known to be unable to proliferate without the presence of EGF in growth media[188,189], hence, ideally, the model should incorporate a more improved balance between basal and ligand induced signaling for stratification of cell proliferation events. In our previous work, this limitation did not affect the behavior of single cells where stochastic single cell simulations initiated from a representation of a serum-starved MCF10A cell which did not enter cell cycle without the presence of EGF in growth media[32]. However, for cell population simulations, single cells are subject to randomized sampling for induction of an asynchronously cycling population which more closely resembles the experimental conditions whereby drug treatment is applied after growth media is introduced to the cells. Under such conditions, the effects of the limitation become more apparent at the population level.

A potential solution to be considered for our future work is to implement a more robust computational pipeline to determine specific rate constants in the single cell model such that observed phenotypical outcomes satisfy biologically justified constraints. In our previous work an attempt was made to do this by means of a process that we call “initialization”. In initialization, certain model parameters and initial conditions are determined for a specific cell-line context using a set of focused parameter estimation operations which aim to tune parameters based on constraints placed on model observables. It is a computationally intensive process whereby each

parameter estimation step performs iterative execution of deterministic model simulations. The SPARCED model is composed of a gene expression and a protein biochemistry module which are executed simultaneously. However, communication bottlenecks between the modules caused the computation time to be impractical for the purpose of initialization. Hence, the preliminary initialization protocol was limited to the protein biochemistry module, precluding a robust estimation of the basal ERK and AKT pathway activities. Recently, we were able to solve the communication bottleneck problem which speeded up the deterministic execution by over 200-fold[197]. This drastic increase in computation speed for deterministic simulations will allow a more exhaustive exploration of the model parameters essential for defining a more robust initialization protocol.

As a mechanistic single cell model, SPARCED is one of the largest models of mammalian cell signaling. Still, it falls short of encapsulating the entirety of intracellular biomolecular networks on a genomic scale. Nevertheless, considering the intricate nature of cell biology, a more comprehensive approach towards single cell pharmacodynamic modeling can be developed by implementing incremental expansion with additional cellular pathways. In this regard, the cell population simulation framework may help monitor the emergent outcome of those incremental changes at the cell population level, ensuring that they align closely with observed biological phenomena and prevent substantial deviations.

Chapter 5

A STRETEGY FOR OMICS-INFORMED PHARMACODYNAMIC MODELING OF CANCER CELL LINES

5.1 Introduction

Innovations in genome technologies such as whole exome sequencing and whole genome sequencing, along with comprehensive characteristics of expression context with RNAseq, as well as mass-spectrometry based proteomics have provided a foundation to develop a more holistic understanding on the pathophysiology and heterogeneity of cancer[14,33]. The advent of large-scale multi-omics datasets has unveiled new challenges in effectively integrating these data for quantitative analysis. Consequently, there is an emerging effort towards devising data-driven mathematical and computational methods to analyze high-dimensional datasets obtained from high throughput multi-omics techniques.

Recent advancements in omics-informed statistical and data-driven approaches have enabled identification of clinically relevant disease subtypes, identification of putative biomarkers for diagnostics and driver genes for tumor progression[198–203]. However, deriving mechanistic insights about dynamics of disease progression, drug response and resistance mechanisms remains a significant challenge. Mechanistic models, founded on the principles of biophysical processes that drive phenotypical functions and higher-level physiological activities have the potential to generate novel insights about tumor progression and possible efficacious therapeutic strategies.

Integration of multi-omics datasets into such models may enable deeper understanding of the causality and dynamics of disease progression and resistance.

Depending on the size and complexity of mechanistic models, integrating high dimensional muti-omics dataset in a manner that does not destabilize the model structure can be a daunting prospect. We recently built one of the largest models of single cell proliferation and death signaling capable of dynamically describing stochastic cell fate in response to growth stimulus and drug actions[32,174]. The granularity of the model enables inclusion of genomic, transcriptomic and proteomic data to define its context. After being built with the MCF10A context, its adaptability to a new cell line, U87 was demonstrated by integrating its genomic and transcriptomic data with a computational process called “initialization”. In this chapter, we present a revised and more robust initialization procedure that can be applied to one of the largest multi-omics datasets of cancer cell lines, the Cancer Cell Line Encyclopedia (CCLE)[33]. We retrieved the omics data for 251 cancer cell lines and successfully applied the initialization procedure to 59 of those cell lines to create context-specific single cell models of cancer cell lines. Furthermore, we applied a mechanistic cell population simulation framework on these cell line specific models and created representations of dynamic cell populations of 28 cell lines which are consistent with their experimentally observed growth rates. We employ these dynamic cell population models to evaluate a strategy for the mechanistic exploration of their experimental drug sensitivity profiles.

5.2 Initialization Overview

The goal of the initialization procedure is to create a single cell model representing a specific cell line, based on its omics data. The SPARCED model consists of several signal transduction pathways, such as, receptor tyrosine kinase (RTK), ERK and AKT pathways for cell proliferation, cell cycle, DNA damage and apoptosis. These pathways are represented within the model as various species and their interactions, primarily, proteins, their modified variants and protein complexes. This network of proteins is supported by a central dogma based foundation of genes and mRNAs. Our design principles dictate that amounts of these species are defined by taking input of genomic, transcriptomic and proteomic data. The first step in defining the context of the model involves compiling the omics input and assigning the gene copy numbers and calculating the mRNA and protein concentrations. At this stage, the appropriate initial conditions for the model species such as nascent proteins, protein variant forms, and protein complexes are not known. Do determine those, we take an iterative approach. At first, we assign the protein amount to the nascent protein forms. Then we run the model to steady state at various stages. During these stages, initial protein levels are reassigned to their variant forms, depending on basal interactions. At the same time, we also ensure that certain biological functionalities of the model are intact even when omics context has been changed. For this purpose, we perform unit tests pertaining to specific biological functions and readjust certain model parameters to ensure that the model with its new context may pass the unit tests. Such biological features include conservation of the overall transcriptomic and proteomic levels, functionality of ERK and AKT proliferative signaling, transcription factor activity, cell cycle, survival signaling,

DNA damage and apoptosis. The end result of the initialization process is the representation of a single cell from a specific cell line, which has been serum starved, and is prepared for growth stimulation.

5.3 Previous Initialization Workflow and Limitations

The initialization procedure was presented along with the original Pan Cancer Driver Pathway model developed at our laboratory[32]. This procedure was used to redefine the biological context of the model by taking inputs of omics data from U87 glioma cells and successfully demonstrated with drug sensitivity prediction for the new cell line context. However, there were technical limitations to this procedure due to the structural properties of the model. Initialization is a computationally intensive process which relies on numerous iterative executions of model simulations to perform certain unit tests focusing on specific biological functionalities of the model. Implementation of gene expression noise and cell to cell variability was accomplished by means of a hybrid structure whereby a stochastic gene expression module and a deterministic protein interaction module were executed simultaneously. Within the computational workflow of a simulation, inter-module communication is the biggest bottleneck. Even though the model is run deterministically during initialization, computation speed of the hybrid structure precluded its application for initialization. For this purpose, the computations of the initialization were kept limited to the protein interaction module, which employs a system of ordinary differential equations (ODE) and could be executed rapidly with the use of ODE solvers. This imposed certain limitations to the capabilities of the initialization procedure since it could only tune functionalities that do not rely on

inter-module communication. As a result, functionalities such as transcriptional regulation, cell cycle initiation, replicative stress and survival signaling could not be tuned robustly to accommodate a wider range of cell line contexts. As part of the revised edition of this model, the SPARCED model[174], we also developed a version of the initialization pipeline compatible with the newer format. The computational workflow of the initialization procedure itself was left unaltered. Recently we were able to overcome this technical limitation by implementing a new ODE based structure which includes all deterministic interactions of the model within a single system of ODEs[197]. This new structure can produce the same result as the hybrid model when executed deterministically while achieving more than 200-fold increase in computation speed. This allowed us to develop a revised pipeline for initialization, which is capable of tuning biological functionalities that the original initialization procedure was missing. In this work, we retrieved genomic, transcriptomic and proteomic data from 251 cell lines in the Cancer Cell Line Encyclopedia (CCLE)[33,204] and used those to test our new initialization pipeline. Out of those, we successfully completed the initialization for 59 cell lines. The computational workflow of the initialization procedure has been described in the following section.

Table 5.1: Comparison of initialization steps in Bouhaddou 2018 model and new initialization pipeline		
Initialization Step	Bouhaddoul 2018 Initialization	New Initialization Pipeline
Translational rate constant	✓	✓
Basal ERK pathway activity	×	✓
Basal AKT pathway	×	✓
Basal cell cycle activity	✓	✓
Transcriptional activators	×	✓
Survival signal	×	✓
Basal apoptosis signal	✓	✓
Basal DNA damage	✓	✓
Replicative stress	×	✓
Apoptosis (time to death)	×	✓

5.4 Initialization Procedure

Step 1 – Translational Rate Constant Tuning

This is the first step in the initialization workflow whereby model parameters are tuned with rigorous and iterative unit testing. The goal of this step is to ensure that steady state levels of proteins and mRNAs in the model outputs match their measurements as per the omics data. The model consists of various forms for proteins, namely, their nascent form, post-translational modifications, and protein complexes. At the start of this step protein levels derived from the proteomics data are assigned to the nascent protein species, i.e. the protein forms that have been immediately translated and lack any post translational modification. Once the model is run to a steady state, the initial protein amounts may accumulate into various other forms due to basal pathway activities. Many such protein form variants have different degradation rates than their nascent counterparts and as a result, total protein levels at steady state may no longer match the proteomics data (Figure 5.1). In this step, gene-specific translation rate constants (k_{TL}) are iteratively adjusted while the model is run to steady state in absence of any growth stimulation during each iteration. The ratio of simulated and measured protein level is calculated and used to optimize the translation rate constant corresponding to the source gene. For each gene product, the rate constant is readjusted until simulated protein levels are within 1% of experimental data. This step is further divided into several sub-step considering the dynamics of modeled interactions. There are certain species and reactions in the model, which if left unrestrained, may alter translation rate globally and hence cause difficulties in attaining a steady state. These include ribosome synthesis, dynamic EIF4E levels and all mRNA transcription

rates. Initially all these aspects are turned off such that they may not affect the translation rate. In subsequent sub-steps of translation rate constant adjustments, these features are enabled gradually to ensure minimal fluctuations in rate constant adjustment. A significant majority of CCLE cell lines that we tested were able to pass this step, only exceptions are cases where certain mRNA levels are missing from the dataset, despite presence of corresponding gene copy number and proteomics data. Examples include, SW1990, NCIH1792, SAOS2, OVCAR8, HCC1395, PC14, HEL9217, HCC70, HT1376, SNU1, COLO678 and MDAMB157.

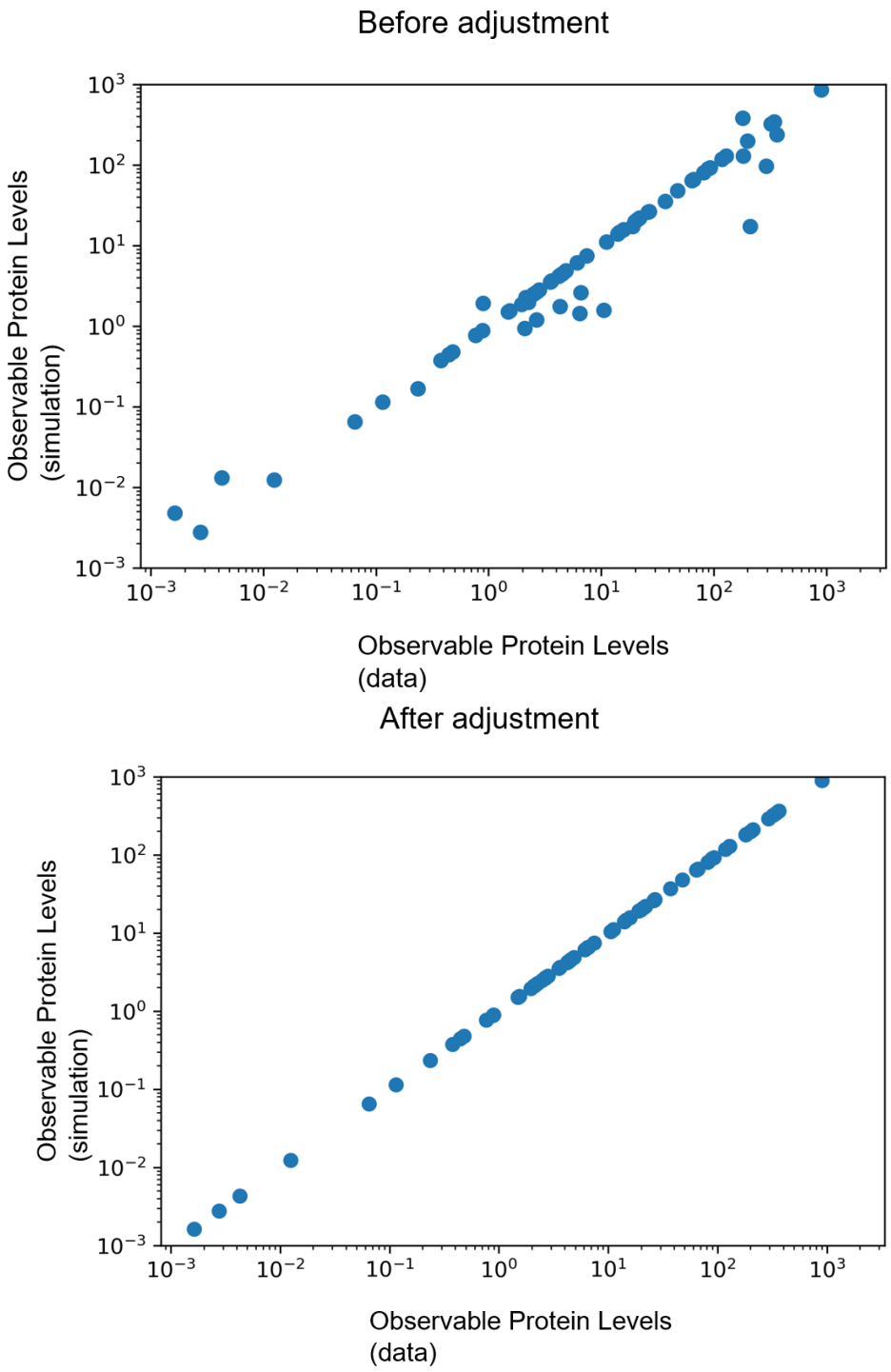


Figure 5.1: Comparison between simulated and measured protein levels before and after the translation rate constant adjustment step for an example cell line (AU565)

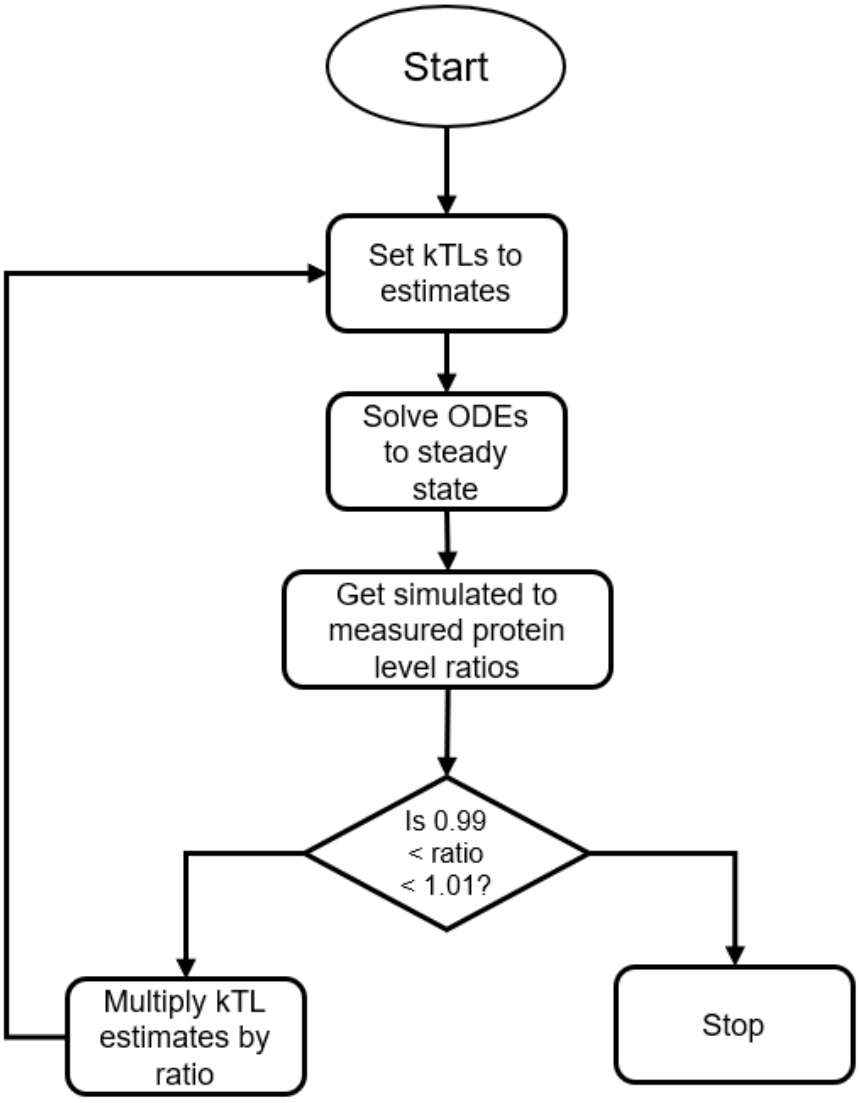


Figure 5.2: Computational workflow of the translation rate constant adjustment.

Step 2 – Basal ERK Pathway Activity Tuning

In the SPARCED model, basal ERK pathway activity is represented by the basal Ras activation and inactivation reactions. Ras is a protein that may exist in a GDP-bound inactive or GTP-bound active state. An inactive RasGDP may exchange its GDP for a GTP to adopt an active state which may happen due to basal or ligand induced proliferative signals. In this step, we tune the basal mode of Ras activation. The goal of initialization is to alter the biological context of the MCF10A cell represented by the original SPARCED model while retaining the biological functionalities of the model. Hence, we create a serum-starved cell for the new context of interest for which basal pathway ERK pathway activity should not result in cell growth. Here we iteratively adjust the basal RasGDP to RasGTP exchange rate in deterministic simulations without growth stimulus. We ensure that the overall ERK pathway activity, observed as a ratio of phosphorylated ERK to total ERK levels, remains close to its value observed in the original (MCF10A) context of the model. Since alteration of this rate parameter may also change steady state levels of other proteins within the model, we also readjust the translation rate constants as described in the previous step as part of this step. A majority of CCLE cell lines that we included in our test passed this step as the desired ppERK/ERK ratio was attainable by adjusting the basal RasGDP to RasGTP exchange rate (Figure 5.4 A,B). However, for several other cell lines this ratio was not attainable and this step failed (Figure 5.4 C,D). This could be attributed to the limitations of SPARCED model and biomolecular pathway activities that are currently out of its scope.

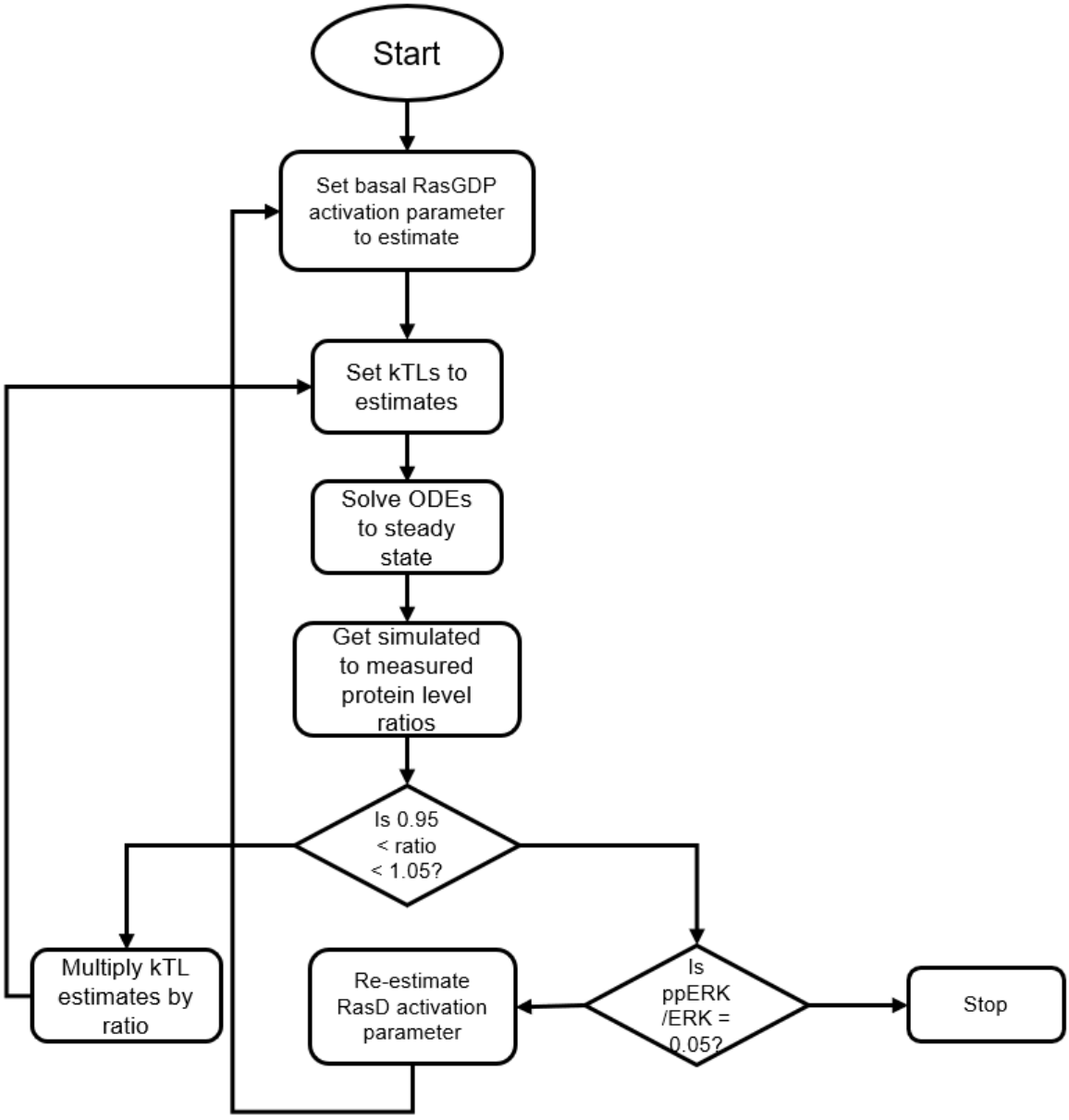


Figure 5.3: Computational workflow of the basal ERK pathway activity tuning step.

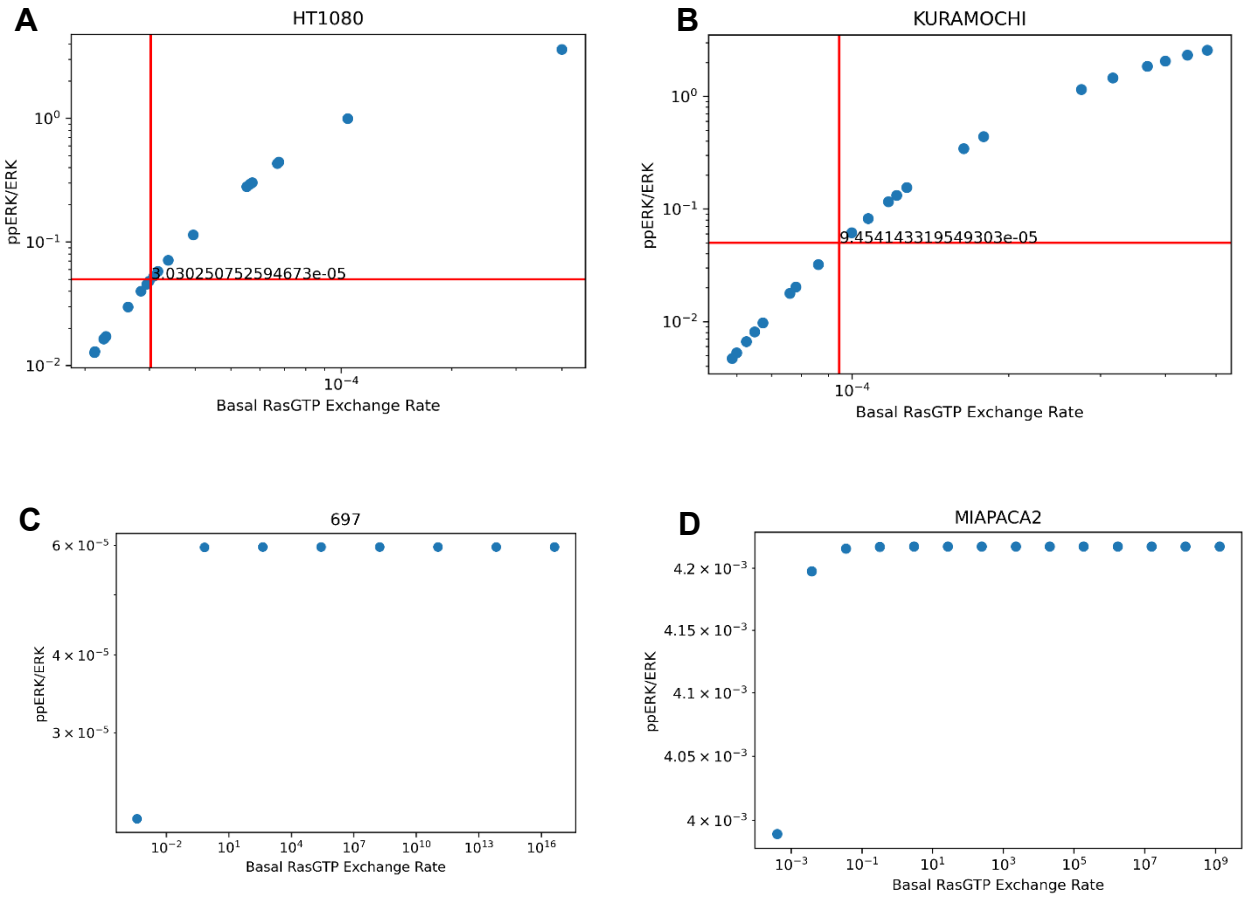


Figure 5.4: Examples of parameter screening performed during the basal ERK pathway activity tuning step. (A,B) Example cell lines HT1080 and KURAMOCHI for which this step was successful, as evident by the attained ppERK/ERK ratio. (C,D) Example cell lines 697 and MIAPACA2 for which this step failed as the ppERK/ERK ratio was not attainable by tuning of basal RasGDP to RasGTP exchange rate.

Step 3 – Basal AKT Pathway Activity Tuning

This step is conceptually similar to the previous basal ERK pathway tuning step. In this case, we focus on the AKT pathway. The basal activities in AKT pathway are represented with basal PIP2 phosphorylation and dephosphorylation reactions. In this step, we iteratively adjust the basal PIP2 phosphorylation rate such that the ratio of phosphorylated AKT to total AKT remains close to the value observed in the original MCF10A context. At the same time we readjust any translation rate constant preventing deviation of simulated protein levels from their experimentally measured levels. This ensures that the basal AKT pathway activities will not result in proliferative outcomes when the cell is at a serum starved state. This procedure was adequate in tuning the basal AKT pathway activity to desired level for all cell lines and we did not encounter any failure at this step.

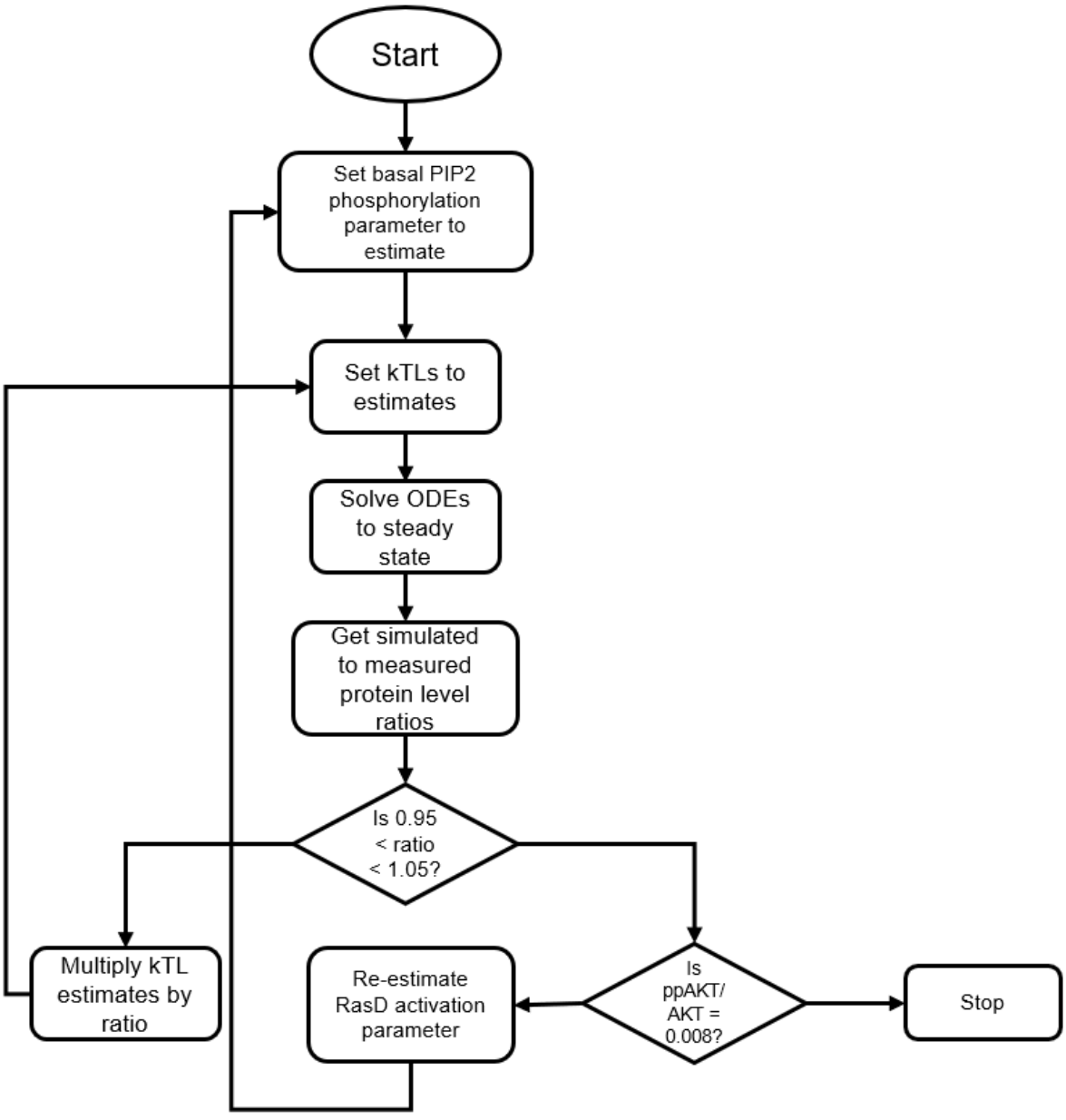


Figure 5.5: Computational workflow of the basal AKT pathway activity tuning step.

Step 4 – Basal Cell Cycle Pathway Activity

This initialization step estimates two rate parameters associated with the cell cycle process, namely, the basal cyclin D synthesis rate and basal p21 degradation rate. Most of the species in the cell cycle pathway are kept to their initial concentrations as per the original model. However, cyclin D and p21 levels are derived from the proteomics data which is used to specify their basal levels in the model. The basal cyclin D synthesis rate and basal p21 degradation rates need to be tuned accordingly such that the model can maintain those levels at steady state. In this step, we tune those rates iteratively. This ensures that basal cyclin D and p21 levels are maintained and they do not initiate cell cycle in absence of a growth stimulus. Once this tuning is complete, we readjust the translation rate constants once again to ensure steady state protein levels match proteomics data.

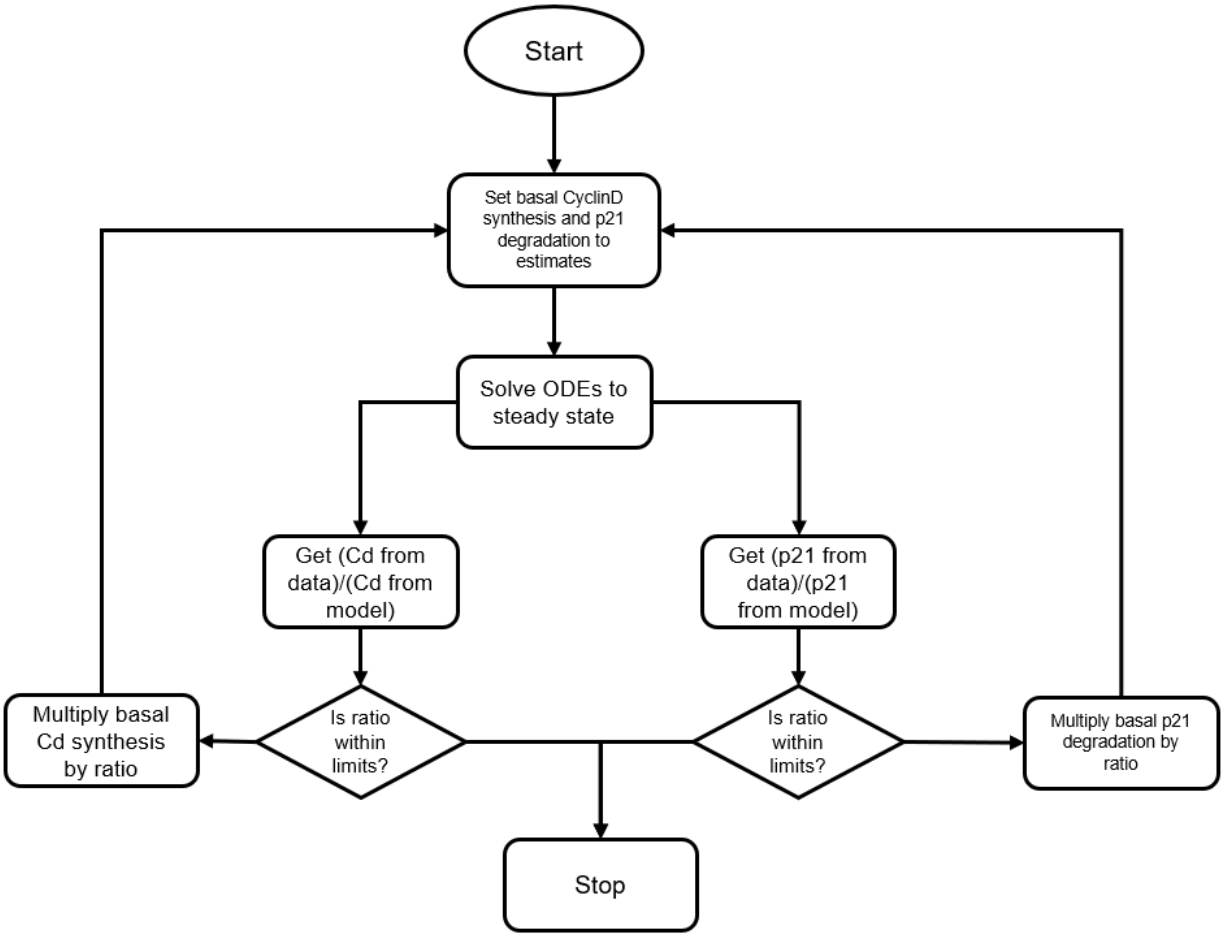


Figure 5.6: Computational workflow of the basal cell cycle pathway activity tuning step.

Step 5 – Transcriptional Activators

This initialization step ensures the capability of SPARCED model to initiate cell cycle in presence of growth stimulus. The cell cycle is initiated by the upregulation of cyclin D by the transcription factors AP1 and MYC. Hence AP1 and MYC serve as primary proliferative inputs into the cell cycle model. These transcription factors are upregulated as a result of elevated ERK and AKT activities respectively. This in turn drives the synthesis of cyclin D and production of cyclin D above a certain threshold pushes the system beyond the restriction point. This initiates the cell cycle followed by oscillations in characteristic cell cycle species, namely, the activate cyclin and cyclin-dependent protein kinase complexes. Transcription of cyclin D is governed by the following rate law:

$$v_{bm} = k_{leak} \cdot g^* + k_{max} \cdot \left(\frac{\left(\frac{[TA_{AP1}]}{kA_{50-AP1}}\right)^{na_{AP1}}}{1+\left(\frac{[TA_{AP1}]}{kA_{50-AP1}}\right)^{na_{AP1}}} \right) \cdot \left(\frac{\left(\frac{[TA_{MYC}]}{kA_{50-MYC}}\right)^{na_{MYC}}}{1+\left(\frac{[TA_{MYC}]}{kA_{50-MYC}}\right)^{na_{MYC}}} \right) \cdot g^*$$

Here, the parameters kA_{50} represent the half maximal effective concentrations for the transcriptional activators AP1 and MYC on the upregulation of cyclin D. Since the steady state levels of these transcription factors are informed by the omics data, these are expected to be different for different cell line contexts. In order for the cell cycle submodel to be functional across contexts, these kA_{50} parameters need to be readjusted based on new proteomics input. In this initialization step, we run deterministic simulations in presence of growth stimulus and iteratively adjust these parameters until cell cycle may be observed. We use oscillations in active cyclin B/CDK1 complex levels as confirmation of cell cycle. We select the lowest value for each parameter for which at lease 3 peaks of active cyclin B/CDK1 complex may be

observed. Successful completion of the basal cell cycle pathway activity and transcriptional activator adjustment ensures the functionality of the cell cycle submodel such that cell cycle is absent in cells without growth stimulus, but as soon as the growth factors are added, cell cycle may be initiated. This is one of the steps that a significant number of CCLE cell lines (60) failed to complete. The reason for this failure is likely to be elevated levels of p21 (figure 5.9), which is an inhibitor of cell cycle. P21 is one of the proteins that were missing from label-free quantification of CCLE proteomics dataset. In absence of this data, p21 levels were estimated using the mRNA levels and the protein to mRNA ratio observed in MCF10A context. This limitation stems from poor-quality of data and could potentially be resolved with the availability of proteomics data with higher accuracy.

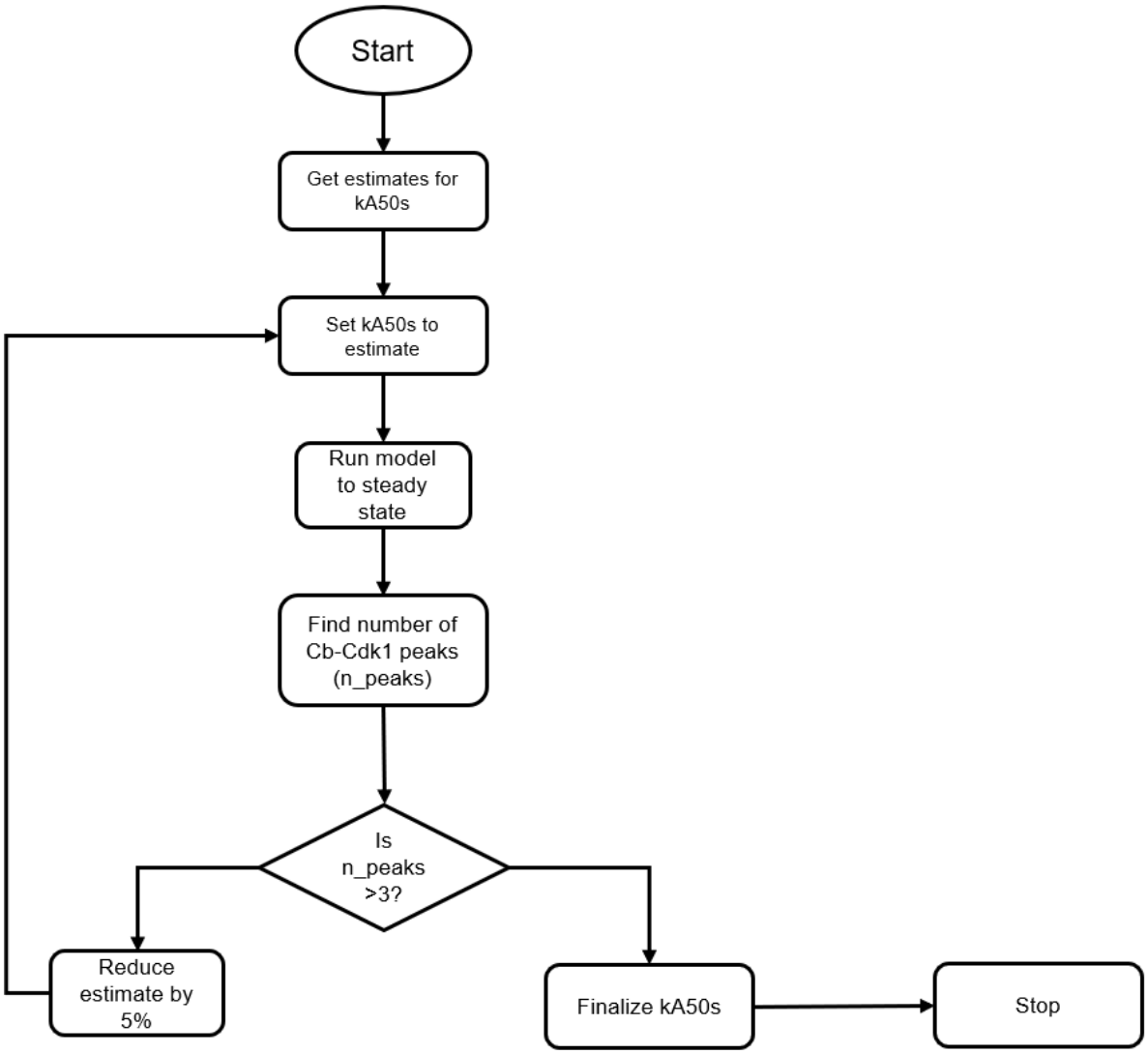


Figure 5.7: Computational workflow of the transcriptional activator tuning step.

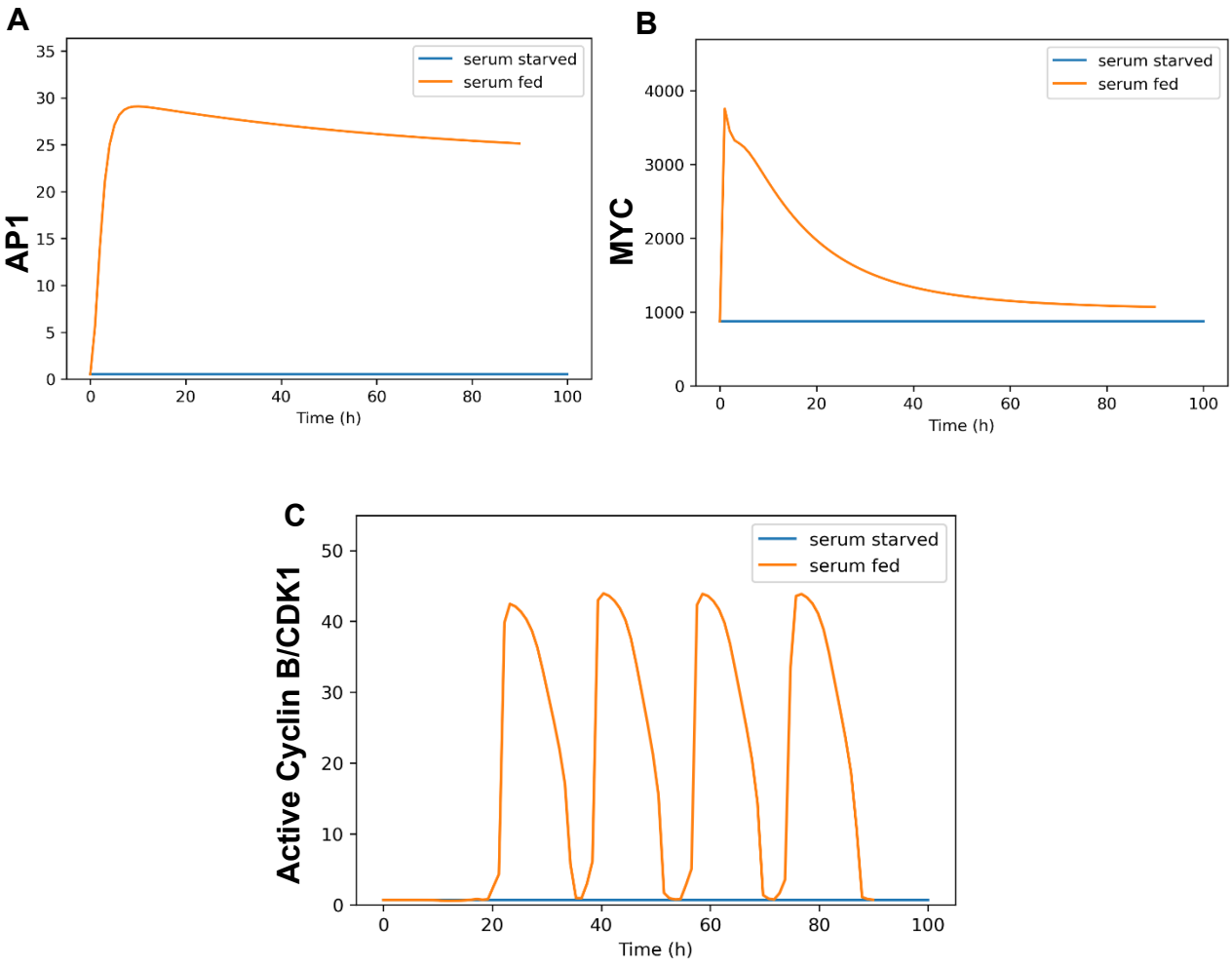


Figure 5.8: Visual confirmation of the successful completion of basal cell cycle activity and transcriptional activator tuning with protein levels from deterministic simulation in AU565 cells. In absence of serum, transcriptional activators AP1, MYC and cyclins remain at steady state. When serum is fed, the growth factor presence upregulates the transcriptional activators and cell cycle can be observed with peaks in active cyclin B/CDK1 levels.

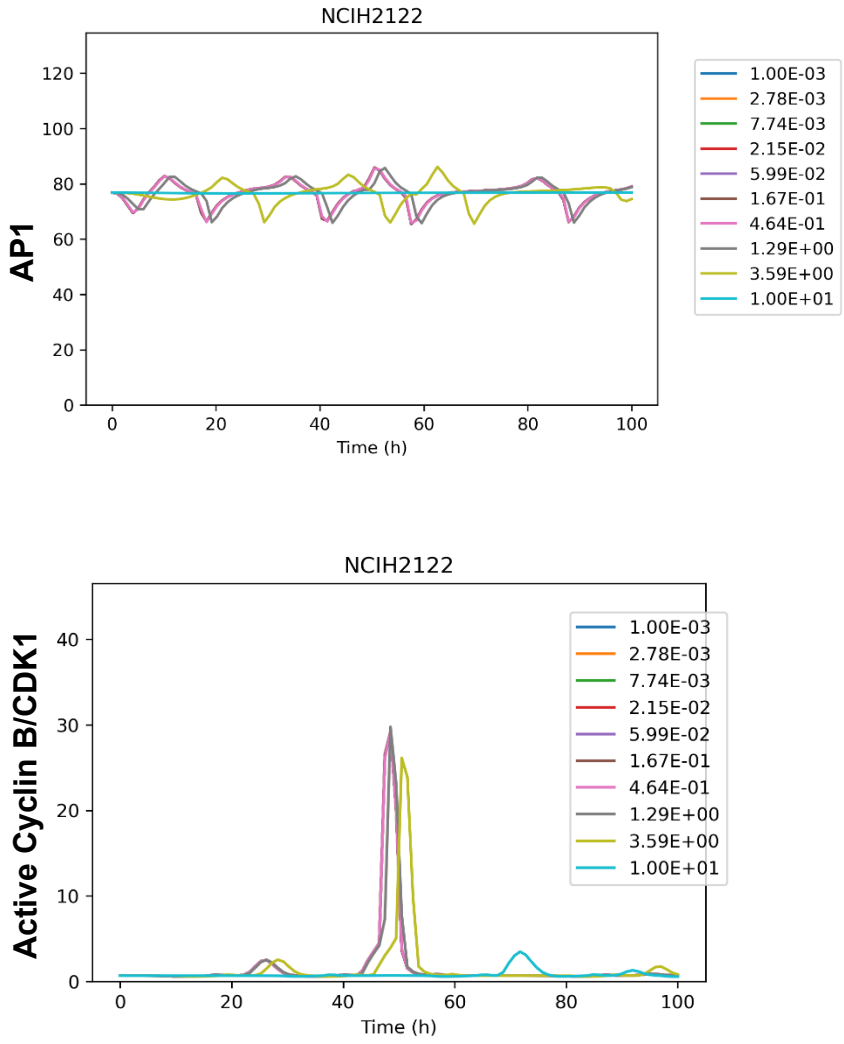


Figure 5.9: An example cell line (NCHI2122) failing to demonstrate persistent Cyclin B/CDK1 peaks during transcriptional activator tuning for a wide range of values of half-maximal kA50 rates. Elevated levels of p21 was also observed in this cell line.

Step 6 – Survival Signal Tuning

Survival signaling consists of the activities that negate the effects of pro-apoptotic signaling in presence of growth stimulus. In the SPARCED model, one of the primary mechanisms of survival signaling is the upregulation of AKT activity. When activated, AKT may phosphorylate FOXO, which is a pro-apoptotic transcriptional activator. In absence of elevated AKT activity, FOXO may translocate to the nucleus and upregulate transcription of the pro-apoptotic protein BIM. Phosphorylation by AKT prevents the nuclear localization of FOXO and as a result fails to deliver pro-apoptotic signal to BIM. The rate law that governs the transcription of BIM is as follows:

$$v_{bm} = k_{leak} \cdot g^* + k_{max} \cdot \left(\frac{\left(\frac{[TA_{FOXO}]}{kA_{50-FOXO}}\right)^{na_{FOXO}}}{1+\left(\frac{[TA_{FOXO}]}{kA_{50-FOXO}}\right)^{na_{FOXO}}} \right) \cdot g^*$$

This rate law includes a “leak” or constitutively active term, which represents the mRNA synthesis mechanism the model does not include explicitly, followed by an “induction” term which represents the effects of transcriptional activators. In this initialization step we iteratively adjust the k_{leak} and k_{max} parameter to ensure that 90% of BIM mRNA level is maintained by its transcriptional activator. As a result, phosphorylation of FOXO leads to decrease in BIM levels, which is the intended survival signaling mechanism.

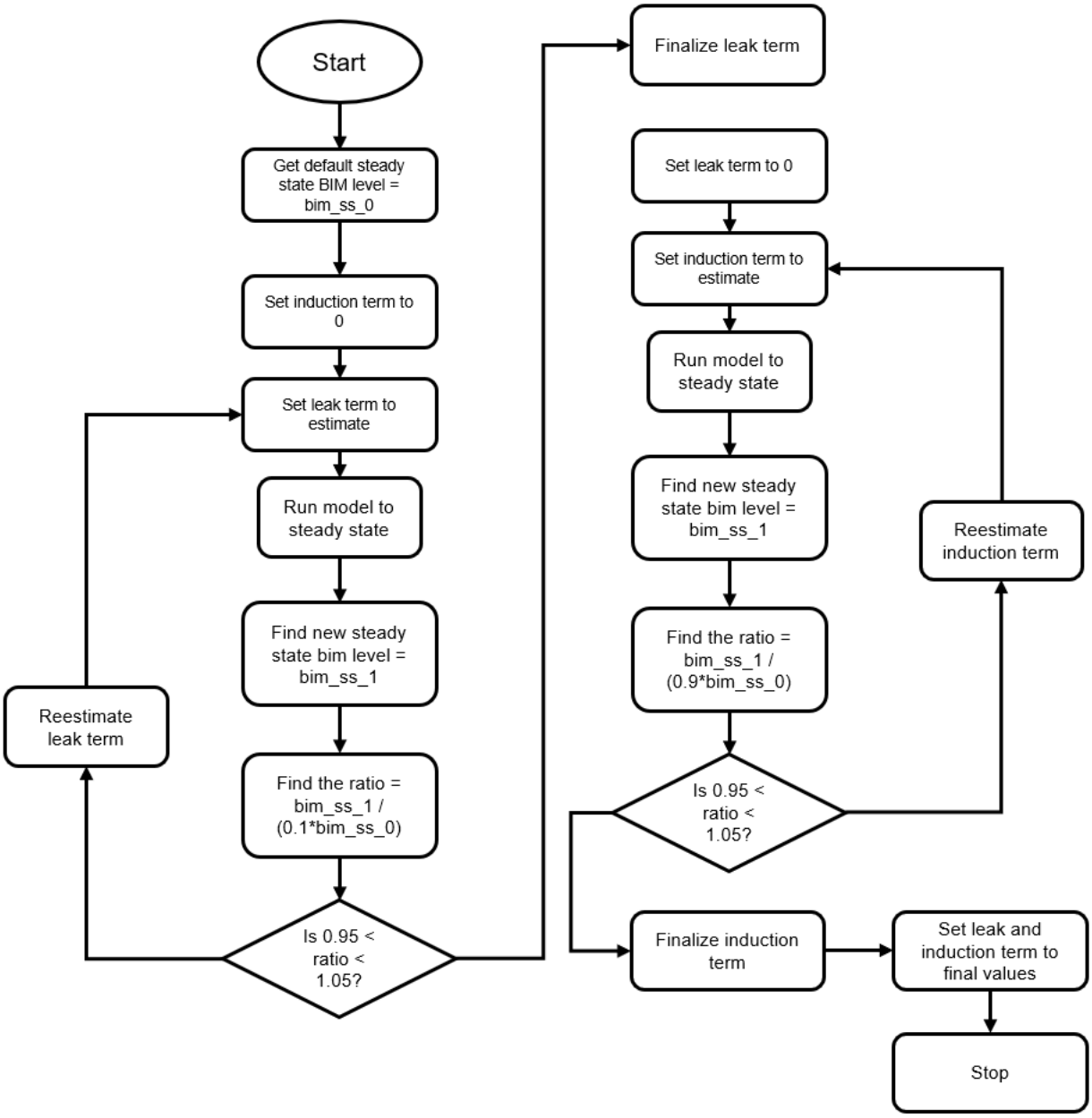


Figure 5.10: Computational workflow of the survival signal tuning step.

Step 7 – Basal Apoptosis Signal Tuning

The apoptosis submodel consists of the activities of initiator caspases (caspase 8 and caspase 10) leading to the activation of executioner caspases (caspase 3 and caspase 7). Executioner caspases drive the digestion of many critical cellular components, which results in cell death. In the SPARCED model, this mechanism is represented by cleavage of PARP, a DNA repair enzyme. A single cell is pronounced dead once the majority of PARP has been cleaved. The model also represents internal pro-apoptotic stimulus by means of PUMA/NOXA upregulation by p53. This requires a negative regulation of anti-apoptotic proteins, as well as a basal flux of death signaling through apoptosis pathways. Basal apoptosis signaling is the result of basal levels of caspases. The chosen mechanism for basal apoptosis signaling in the model is a first-order reaction for caspase 8 cleavage. Until this point in initialization, this reaction rate parameter is switched off to ensure unintended activation of apoptosis pathway. In this initialization step, basal caspases levels are allowed to equilibrate while the basal caspase 8 cleavage rate is incrementally increased. As soon as apoptosis occurs within a 500-hour simulation period, the rate is set to its previous increment. After this, the translation rates are adjusted once again to ensure that simulated protein levels match the proteomic data. This is one of the initialization steps which a large number of CCLE cell lines (36) failed to complete. A likely explanation is the proteomic levels of one or more mediators of apoptosis signaling is incompatible with the stable operation of apoptosis submodel, such that inclusion of proteomic data from those cell lines result in constitutive activation of apoptosis. It indicates a limitation in the SPARCED model,

which lacks the mechanism to prevent excessive pro-apoptotic signaling in certain cell line contexts.

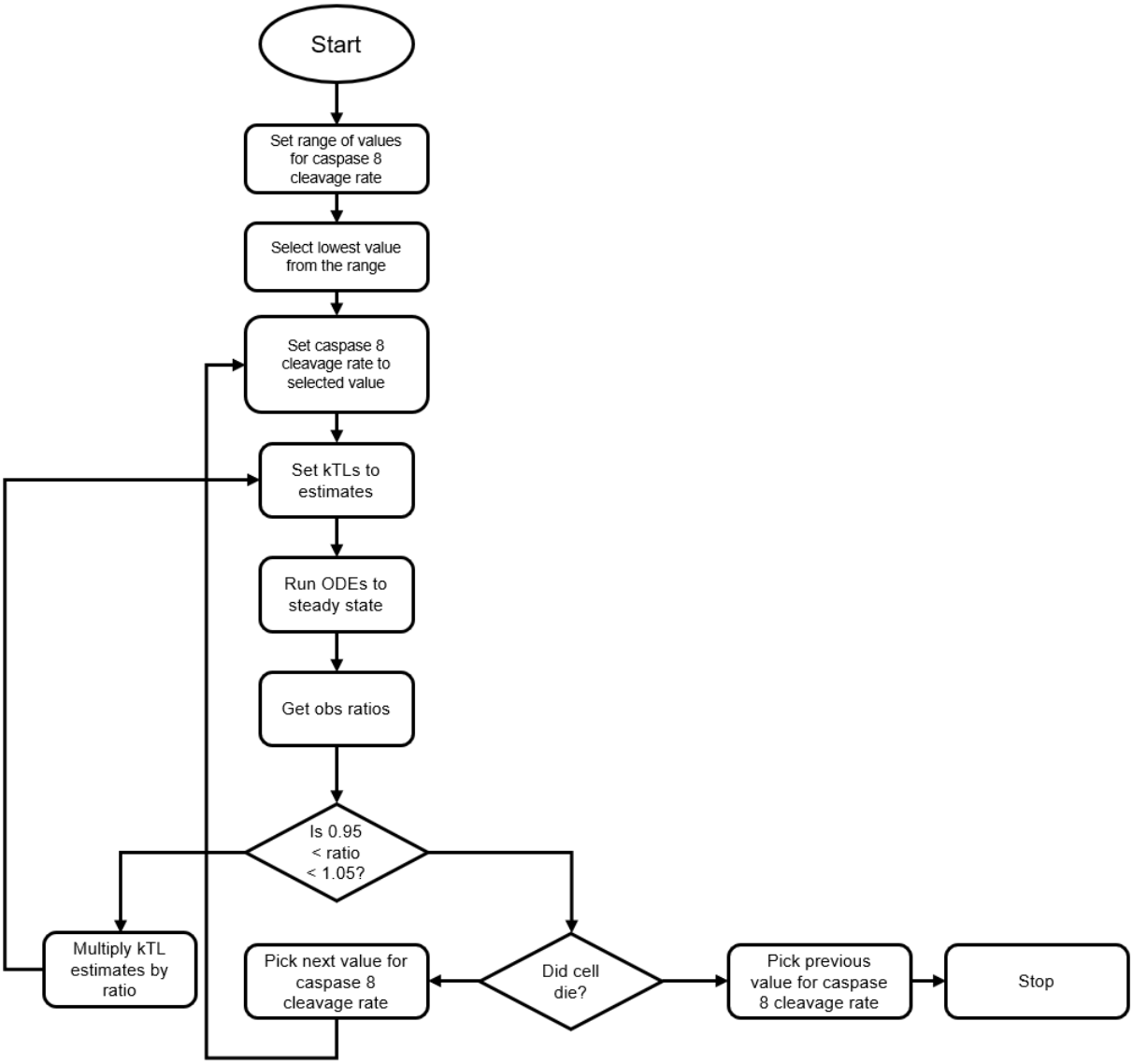


Figure 5.11: Computational workflow of the basal apoptosis signal tuning step.

Step 8 – Basal DNA Damage Tuning

The DNA damage submodel describes the activation of p53 due to double and single stranded breaks. Due to external and well as internal stressors, living cells tend to possess a basal amount of DNA damage. However, this basal level of DNA damage should not be high enough to induce a p53 response. To represent this effect, we included a first order reaction inducing double strand breaks in the model. In this initialization step, we estimate this reaction rate. It is accomplished by parameter sensitivity analysis considering the effects of this rate on the overall active p53 levels in deterministic simulations without any growth stimulation. From this sensitivity analysis, the bifurcation point is detected and the basal DNA damage rate is set to a value 10 times lower than this bifurcation point.

Step 9 – Replicative Stress Tuning

During cell cycle, the DNA undergoes replication. During S-phase, the DNA is uncoiled from histones which makes it vulnerable to insult. Because of the physical relocation of DNA, breaks are more likely to occur. This replicative stress is represented in the model with a reaction that induces DNA damage based on the levels of Cyclin E and Cyclin A which are elevated during cell cycle. Similar to the basal DNA damage, this replicative stress should not result in full activity of p53. This initialization step estimates the rate parameter governing replicative stress towards this goal. This includes a parameter sensitivity analysis similar to the previous step, only in this case the deterministic simulations are run in presence of growth stimulation.

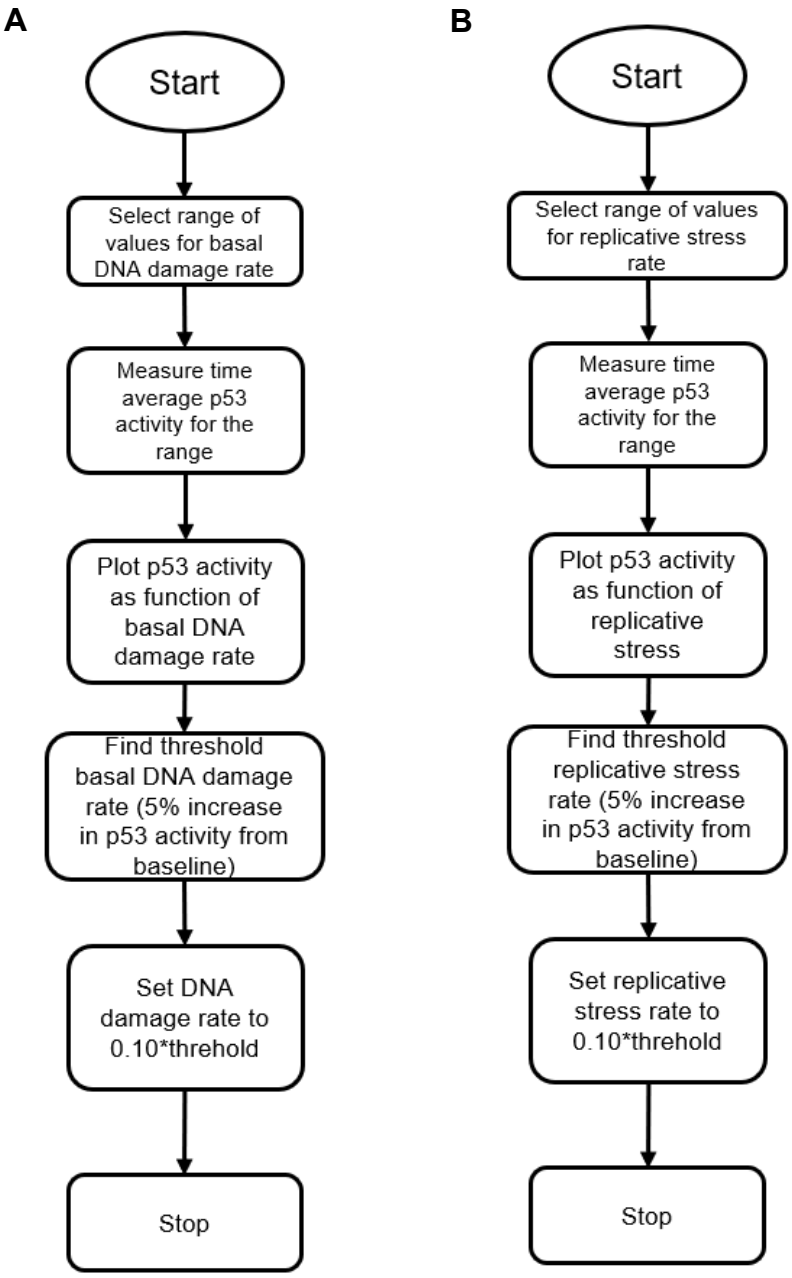


Figure 5.12: Computational workflow of the basal DNA damage (A) and replicative stress tuning (B) steps.

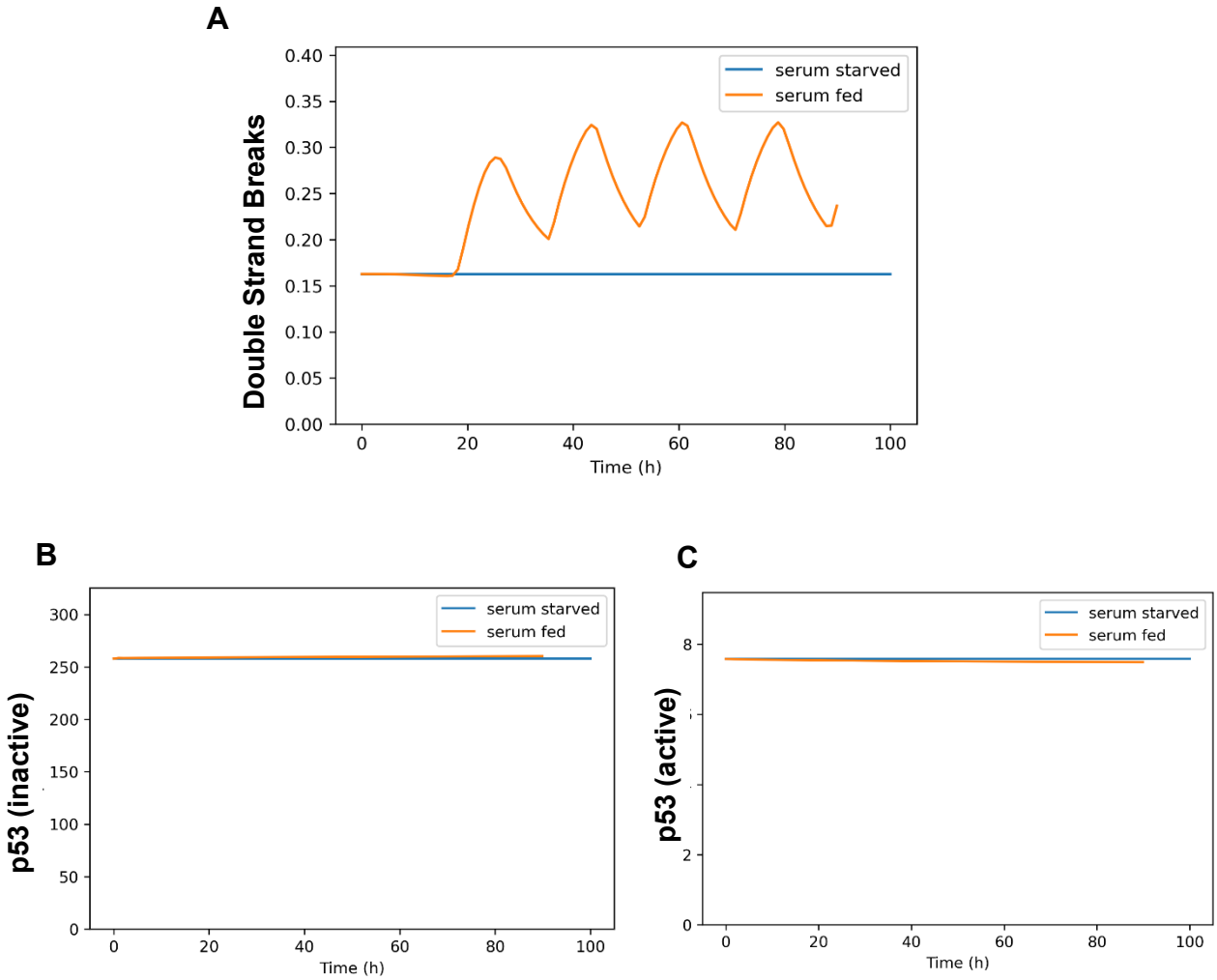


Figure 5.13: Visual confirmation of successful completion of steps 7 and 8 demonstrated with deterministic simulations for AU565 cells. Basal DNA damage and replicative stress exist with and without serum presence respectively (A). However, p53 remains mostly inactive (B,C)

Step 10 – Apoptosis Tuning

Tuning of basal apoptosis signaling was introduced in step 7 of initialization, which ensures basal levels of caspases may exist at steady state. However, this tuning step alone is insufficient to ensure apoptosis may occur in presence of explicit pro-apoptotic signals. To accomplish this, we ensure that the cell has equal likelihood of living and dying at the end of 72 hours in a serum-starved state. When cells are tuned to die at 72 hours in a deterministic simulation, it will have equal changes of survival and death in stochastic simulations. In this step, we start with the previously estimated value of basal caspase 8 cleavage rate and then iteratively adjust it until apoptosis may occur at the end of 72 hours in a serum starved state. For this purpose, we observe the level of PARP, at any time point where more than 50% of its initial level is cleaved, we consider that cell to be dead. There are certain CCLE cell lines which failed to pass this step. For those cell lines, inclusion of their proteomics input results in lack of pro-apoptotic signals strong enough to finish apoptosis. It indicates a limitations in the SPARCED model which may lack underlying mechanisms to compensate for weaker apoptotic signaling in those cell line contexts.

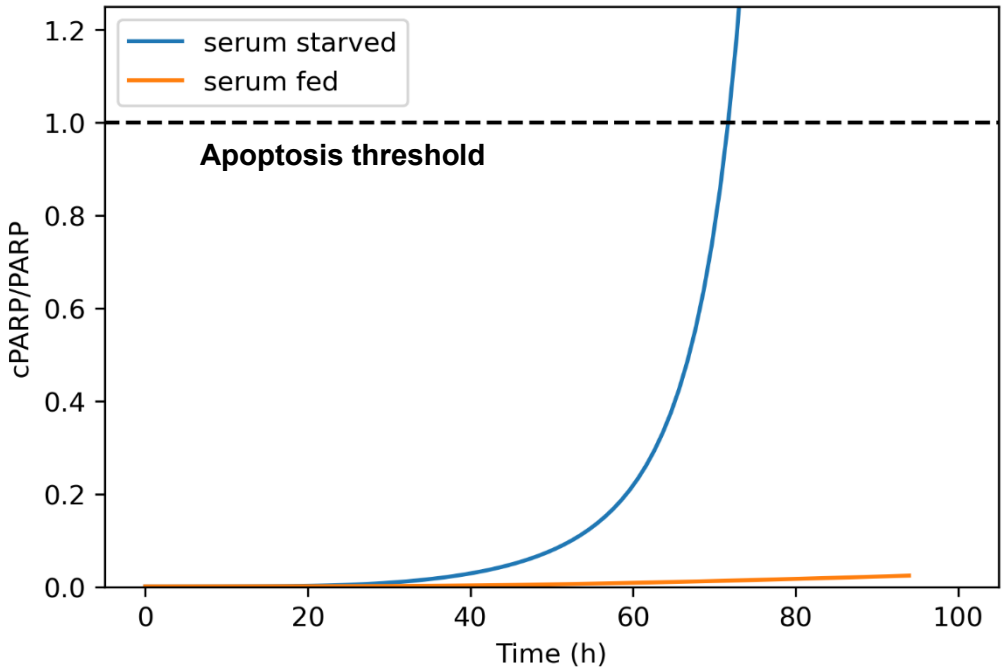


Figure 5.14: Visual confirmation of successful completion of apoptosis and survival signal tuning demonstrated with deterministic simulations for AU565 cells. Here we can observe the dynamic ratio of cleaved to intact PARP levels with and without serum presence. In absence of serum, the cell dies at the end of 72 hours, as per the apoptosis threshold represented with the dashed horizontal line. In presence of serum, survival signaling may rescue the cell from apoptosis.

5.5 Results I – Applying the Initialization Pipeline to Omics Datasets in the Cancer Cell Line Encyclopedia

The revised initialization pipeline described in the previous section allows us to redefine the context of SPARCED single cell model and help represent new cell lines by taking inputs of genomic, transcriptomic and proteomic datasets for each cell line. To evaluate its applicability, we chose the Cancer Cell Line Encyclopedia (CCLE). This is one of the largest and most comprehensive datasets of cancer cell lines originating from a wide range of cancer types. In addition, it also provides drug sensitivity profiles for a large number of cancer cell lines. In our previous work, we have built one of the largest single cell models of proliferation and death signaling which describes how the coordinated dynamics of multiple biological pathways and drug actions stochastically drive phenotypical outcomes and drug dose response. We intend to use this as a foundation towards a more comprehensive single cell pharmacodynamic modeling encompassing a wider range of biological context. This will likely require expansion of the model structure to include biological pathways that were not included in previous work as well as effects of various genomic aberrations such as point and indel mutations which have implication in tumor pathophysiology. For this purpose, large scale pharmacogenomic datasets such as the CCLE may serve as a repository of information about perturbations of biological systems and help make decisions for further model development. In this work, we have attempted to combine our initialization pipeline with the cell population simulation framework to build a system for mechanistic exploration of such datasets. At first, we retrieved the genomic, transcriptomic and proteomic datasets available in CCLE and used them as a test case for our initialization

pipeline. We refined the list of available cells based on their availability in the individual datasets. To test the initialization pipeline, we required cell lines which are present in all omics datasets as well as the drug sensitivity profiles. After this consideration, we were left with a total of 251 cell lines. We applied the initialization pipeline on all these cell lines and a significant number of cell lines (59) the initialization process was completed successfully, which meant at the end of this process we had a unique version of the SPARCED single cell model representing each cell line. A complete list of cell lines that were subject to the initialization process is described in Table 5.2, indicating their results and details of failure.

Table 5.2: Initialization Results for CCLE Cell Lines

Cell line	Result	Comments (failure step)
697	Failed	Basal ERK
22RV1	Failed	Transcriptional activators
769P	Failed	Transcriptional activators
786O	Passed	N/A
8305C	Failed	Basal ERK
8505C	Failed	Transcriptional activators
A172	Failed	Basal ERK
A204	Passed	N/A
A2058	Passed	N/A
A2780	Failed	basal CC equilibration
A375	Failed	Transcriptional activators
A549	Failed	Basal ERK
A673	Passed	N/A
ASPC1	Failed	Apoptosis
AU565	Passed	N/A
BT20	Passed	N/A
BT549	Failed	Survival signaling
BXPC3	Failed	Transcriptional activators
C32	Failed	Basal ERK
CAKI2	Failed	Transcriptional activators

Cell line	Result	Comments (failure step)
CAL27	Failed	Basal apoptosis
CAL851	Passed	N/A
CALU1	Failed	Transcriptional activators
CALU6	Passed	N/A
CAMA1	Passed	N/A
CCK81	Failed	Apoptosis
COLO320	Failed	Basal ERK
COLO678	Failed	Translational rate
COLO679	Failed	Transcriptional activators
COLO741	Failed	Apoptosis
CORL105	Failed	Basal ERK
CORL23	Passed	N/A
DAOY	Failed	Basal apoptosis
DETROIT562	Failed	Basal apoptosis
DMS114	Passed	N/A
DU145	Failed	Apoptosis
DV90	Failed	Transcriptional activators
EBC1	Failed	Transcriptional activators
EFM19	Failed	Apoptosis
F36P	Passed	N/A
FADU	Passed	N/A
FUOV1	Failed	Basal apoptosis

Cell line	Result	Comments (failure step)
G401	Failed	Transcriptional activators
G402	Failed	Transcriptional activators
GAMG	Failed	Transcriptional activators
GB1	Passed	N/A
HCC1187	Failed	Transcriptional activators
HCC1395	Failed	Translational rate
HCC15	Failed	Apoptosis
HCC1806	Passed	N/A
HCC1954	Failed	Basal apoptosis
HCC44	Failed	Basal ERK
HCC56	Failed	Survival signaling
HCC70	Failed	Translational rate
HCC827	Failed	Basal ERK
HCT116	Failed	Transcriptional activators
HCT15	Failed	Basal ERK
HDMYZ	Failed	Apoptosis
HDQP1	Failed	Apoptosis
HEC1A	Failed	Translational rate
HEC251	Failed	Apoptosis
HEC265	Failed	Transcriptional activators
HEC59	Failed	Basal apoptosis
HEC6	Failed	Transcriptional activators

Cell line	Result	Comments (failure step)
HEL9217	Failed	Translational rate
HEP3B217	Failed	Basal apoptosis
HEPG2	Failed	Basal apoptosis
HEYA8	Passed	N/A
HGC27	Failed	Basal ERK
HLF	Passed	N/A
HS294T	Failed	Transcriptional activators
HS695T	Failed	Transcriptional activators
HS944T	Failed	Transcriptional activators
HT1080	Failed	Transcriptional activators
HT1197	Failed	Transcriptional activators
HT1376	Failed	Translational rate
HT29	Passed	N/A
HUH1	Failed	Transcriptional activators
HUPT3	Failed	Basal apoptosis
HUPT4	Failed	Apoptosis
IALM	Failed	Basal apoptosis
IGR37	Passed	N/A
IGR39	Failed	Basal ERK
IGROV1	Failed	Basal ERK
IM95	Failed	Transcriptional activators
IPC298	Passed	N/A

Cell line	Result	Comments (failure step)
ISHIKAWAHERAKLIO02ER	Passed	N/A
J82	Failed	Basal ERK
JHH4	Failed	Basal apoptosis
JHH5	Passed	N/A
JHH6	Failed	Transcriptional activators
JHH7	Passed	N/A
JHOS2	Failed	Apoptosis
JHUEM2	Failed	Transcriptional activators
JM1	Failed	basal CC equilibration
JMSU1	Failed	Basal apoptosis
JURKAT	Failed	Basal ERK
K029AX	Failed	Basal ERK
KARPAS299	Failed	Apoptosis
KARPAS422	Failed	Survival signaling
KASUMI2	Failed	Transcriptional activators
KMRC1	Failed	Transcriptional activators
KMS11	Failed	Survival signaling
KMS12BM	Failed	Survival signaling
KNS42	Passed	N/A
KNS81	Failed	Transcriptional activators
KP2	Failed	Transcriptional activators
KP4	Failed	Survival signaling

Cell line	Result	Comments (failure step)
KURAMOCHI	Failed	Apoptosis
KYM1	Failed	Transcriptional activators
KYSE150	Failed	Transcriptional activators
KYSE180	Passed	N/A
KYSE30	Failed	Transcriptional activators
KYSE410	Failed	Basal apoptosis
KYSE450	Passed	N/A
KYSE510	Failed	Apoptosis
KYSE70	Failed	Transcriptional activators
L33	Failed	Apoptosis
L428	Failed	Apoptosis
LCLC103H	Passed	N/A
LN18	Failed	Basal apoptosis
LN229	Failed	Transcriptional activators
LOXIMVI	Failed	Basal ERK
LS411N	Failed	Basal apoptosis
LS513	Failed	Transcriptional activators
LUDLU1	Failed	Basal apoptosis
MCF7	Failed	Transcriptional activators
MDAMB157	Failed	Translational rate
MDAMB436	Passed	N/A
MDAMB453	Passed	N/A

Cell line	Result	Comments (failure step)
MDAMB468	Passed	N/A
MEWO	Failed	Apoptosis
MFE280	Failed	Basal apoptosis
MFE296	Failed	Basal apoptosis
MFE319	Failed	Apoptosis
MIAPACA2	Failed	Basal ERK
MKN45	Failed	Apoptosis
MKN7	Failed	Apoptosis
MONOMAC1	Passed	N/A
MSTO211H	Failed	Transcriptional activators
NCIH1048	Failed	Basal apoptosis
NCIH1155	Passed	N/A
NCIH1299	Passed	N/A
NCIH1355	Passed	N/A
NCIH1568	Failed	Apoptosis
NCIH1573	Failed	Basal apoptosis
NCIH1581	Failed	Apoptosis
NCIH1650	Failed	Basal ERK
NCIH1666	Failed	Transcriptional activators
NCIH1693	Passed	N/A
NCIH1703	Passed	N/A
NCIH1792	Failed	Translational rate

Cell line	Result	Comments (failure step)
NCIH1793	Failed	Apoptosis
NCIH1944	Failed	Transcriptional activators
NCIH1975	Failed	Apoptosis
NCIH2009	Failed	Basal apoptosis
NCIH2030	Passed	N/A
NCIH2052	Failed	Transcriptional activators
NCIH2122	Failed	Transcriptional activators
NCIH2170	Passed	N/A
NCIH2172	Failed	Basal apoptosis
NCIH2228	Failed	Apoptosis
NCIH226	Failed	Transcriptional activators
NCIH2286	Failed	Transcriptional activators
NCIH23	Passed	N/A
NCIH3255	Failed	Transcriptional activators
NCIH358	Failed	Apoptosis
NCIH441	Passed	N/A
NCIH460	Failed	Apoptosis
NCIH520	Failed	Basal apoptosis
NCIH522	Failed	Apoptosis
NCIH647	Failed	Apoptosis
NCIH650	Failed	Basal ERK
NCIH661	Failed	Apoptosis

Cell line	Result	Comments (failure step)
NCIH747	Failed	Basal apoptosis
NCIN87	Passed	N/A
NIHOVCAR3	Passed	N/A
NUGC3	Passed	N/A
OCIAML5	Failed	Transcriptional activators
OCUM1	Failed	Transcriptional activators
OE33	Failed	Basal ERK
OPM2	Failed	Basal ERK
OV90	Passed	N/A
OVCAR4	Failed	Apoptosis
OVCAR8	Failed	Translational rate
OVSAHO	Failed	Apoptosis
PANC0203	Passed	N/A
PANC0403	Passed	N/A
PC14	Failed	Translational rate
PC3	Failed	Basal ERK
QGP1	Failed	Apoptosis
RD	Failed	Basal apoptosis
REH	Failed	Basal ERK
RERFLCMS	Passed	N/A
RKO	Failed	Transcriptional activators
RPMI7951	Failed	Basal ERK

Cell line	Result	Comments (failure step)
RT112	Failed	Basal apoptosis
RT4	Failed	Basal apoptosis
RVH421	Failed	Basal ERK
SAOS2	Failed	Translational rate
SBC5	Failed	Apoptosis
SCC25	Failed	Apoptosis
SF295	Passed	N/A
SHP77	Failed	Basal apoptosis
SJSA1	Failed	Transcriptional activators
SKCO1	Passed	N/A
SKES1	Failed	Apoptosis
SKHEP1	Failed	Transcriptional activators
SKLU1	Failed	Apoptosis
SKMEL30	Failed	Apoptosis
SKMEL5	Failed	Transcriptional activators
SKNAS	Failed	Basal apoptosis
SNGM	Failed	Transcriptional activators
SNU1	Failed	Translational rate
SNU423	Passed	N/A
SNU449	Failed	Survival signaling
SNUC2A	Failed	Apoptosis
SQ1	Failed	Basal apoptosis

Cell line	Result	Comments (failure step)
SU8686	Failed	Transcriptional activators
SUDHL4	Failed	Basal ERK
SUDHL6	Failed	Translational rate
SUIT2	Failed	Basal ERK
SW1271	Passed	N/A
SW1417	Passed	N/A
SW1573	Failed	Apoptosis
SW1990	Failed	Translational rate
SW403	Passed	N/A
SW48	Failed	Transcriptional activators
SW480	Passed	N/A
SW620	Passed	N/A
T24	Failed	Basal apoptosis
T47D	Failed	Basal apoptosis
TC71	Passed	N/A
TCCSUP	Failed	Transcriptional activators
TE1	Failed	Transcriptional activators
TE11	Failed	Transcriptional activators
TYKNU	Failed	Basal apoptosis
U118MG	Failed	Transcriptional activators
U2OS	Failed	Basal apoptosis
U87MG	Failed	Transcriptional activators

Cell line	Result	Comments (failure step)
U937	Failed	Basal apoptosis
UACC257	Failed	Transcriptional activators
UACC62	Failed	Transcriptional activators
UMUC3	Failed	Basal ERK
VMRCRCW	Failed	Apoptosis
WM115	Failed	Transcriptional activators
WM1799	Failed	Basal ERK
WM2664	Failed	Transcriptional activators
WM793	Failed	Transcriptional activators
WM88	Failed	Transcriptional activators
ZR751	Failed	Basal apoptosis

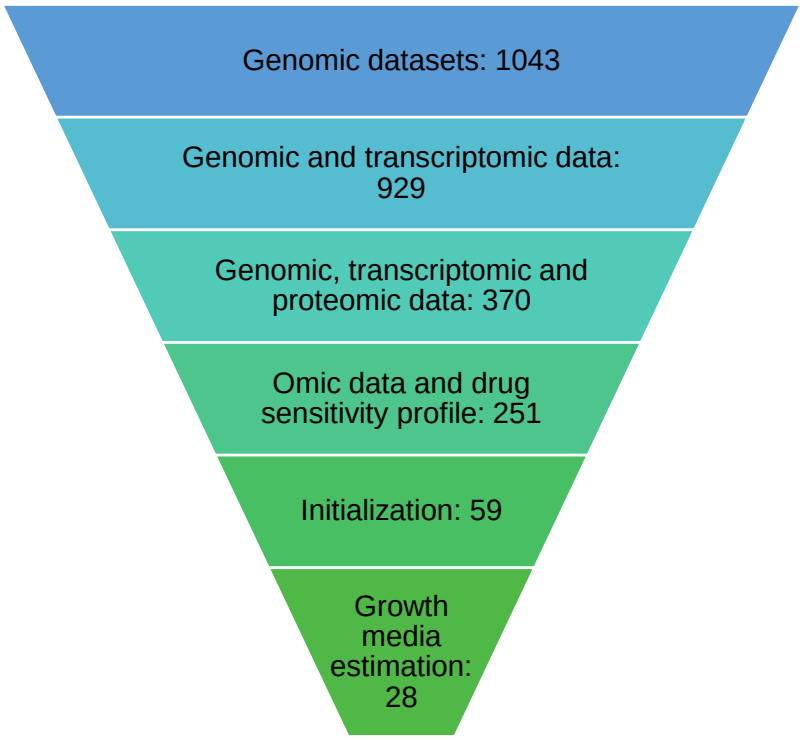


Figure 5.15: Refinement of the CCLE cell lines through various stages of processing. Each level of the inverted pyramid diagram indicates the number of cell lines available at that stage.

5.6 Results II - Growth Media Estimation for Initialized Cell Lines

In the previous chapter, we presented the cell population simulation framework which allows us to execute in silico replications of dose response experiments. Population dynamics retrieved from these virtual dose response experiments allow us to measure drug response as a function of its dose in a manner analogous to experimental results. A direct comparison between experimental and simulation results helps us validate the causality of drug response in terms of the underlying signaling mechanisms in cases where the match. In other cases when we see mismatches, it helps us identify crucial knowledge gaps as well as help develop hypotheses in addressing those. Since initialization procedure allows expansion of the biological context of the model beyond its MCF10A foundation, it also enables the usage of drug sensitivity datasets available for the new cell line contexts. Before dose response simulations can be conducted on the new cell lines, questions may arise whether a dynamic cell population may be represented with the use of initialized cell line models. It is important to ensure their population growth dynamics are in qualitative agreement with their experimentally observed behavior. In our simulation framework, growth stimulus is provided by means of various doses for growth factor which serve as an input to the modeled protein signaling network. However, there is a lack of clarity between the details with which growth conditions are required to be defined for simulations and current methods of cell culture in experiments. One of the key components in cell culture media is fetal bovine serum (FBS). The components in FBS have not been fully identified, and their effects on the composition and cell culture may vary depending on the individual fetus. FBS is prepared after removing clotting factors and blood corpuscles from cattle blood and its

composition may vary depending on the diet and environment of the cattle, making it difficult to accurately state the ingredients and contents[205]. However, it is known to contain growth factors, hormones, proteins, cofactors, and minerals that can affect cell culture. The lack of reliable characteristic information and stringent standards for growth media imposes certain difficulties when producing simulated growth conditions analogous to experiments. To account for this uncertainty, we assume a baseline growth media as an array of several growth factors included in the model, namely, EGF, heregulin, HGF, PDGF, FGF, IGF, and insulin. In the baseline array, concentration for each growth factor was set to its binding affinity for the canonical receptor. We then considered a series of seven log-spaced multipliers, ranging from 10^{-3} to 10^{3} for the baseline growth media and ran cell population simulations for each iteration across all 59 cell lines that passed the initialization process. In these simulations, we observed context specific differential growth rates for all cell lines. Hence, definition of cellular context by means of genomic, transcriptomic and proteomic data has enabled us to include effects of omics context in cell line specific growth and proliferation rates. Furthermore, for many of the cell lines, simulated growth rate within the range of growth media of varying strength closely was also within the range of their experimentally reported doubling time. As per this observation, we classified the cell lines into three groups:

- Group 1 – Cell lines for which experimentally reported doubling time was within the range of simulated growth (28 in total),
- Group 2 – Cell lines for which experimentally reported doubling time was outside the range of simulated growth (15 in total), and

- Group 3 – Cell lines for which growth was insufficient (16 in total).

For the group 3 cell lines, an overall decline in population was observed regardless of the applied strength of growth media. A likely explanation is that inclusion of omics context for these cell lines resulted in aggressive apoptotic signaling, causing population wide mass apoptosis. It indicates a knowledge gap in the current SPARCED model with regards to possible survival signaling mechanism employed in the group 3 cell lines. For the group 2 cell lines, even though population growth may be observed, the experimental doubling time was outside the range of simulated doubling time. It may indicate inadequacies in the underlying cell proliferation mechanism as well as the uncertainty surrounding the characteristics of growth media. We chose to exclude groups 2 and 3 from further analysis and selected the media strength for the group 1 cells which brings it closest to its experimental doubling time. These are the growth conditions that we selected for conducting virtual drug dose response simulations with group 1 cells, of which there were 28.

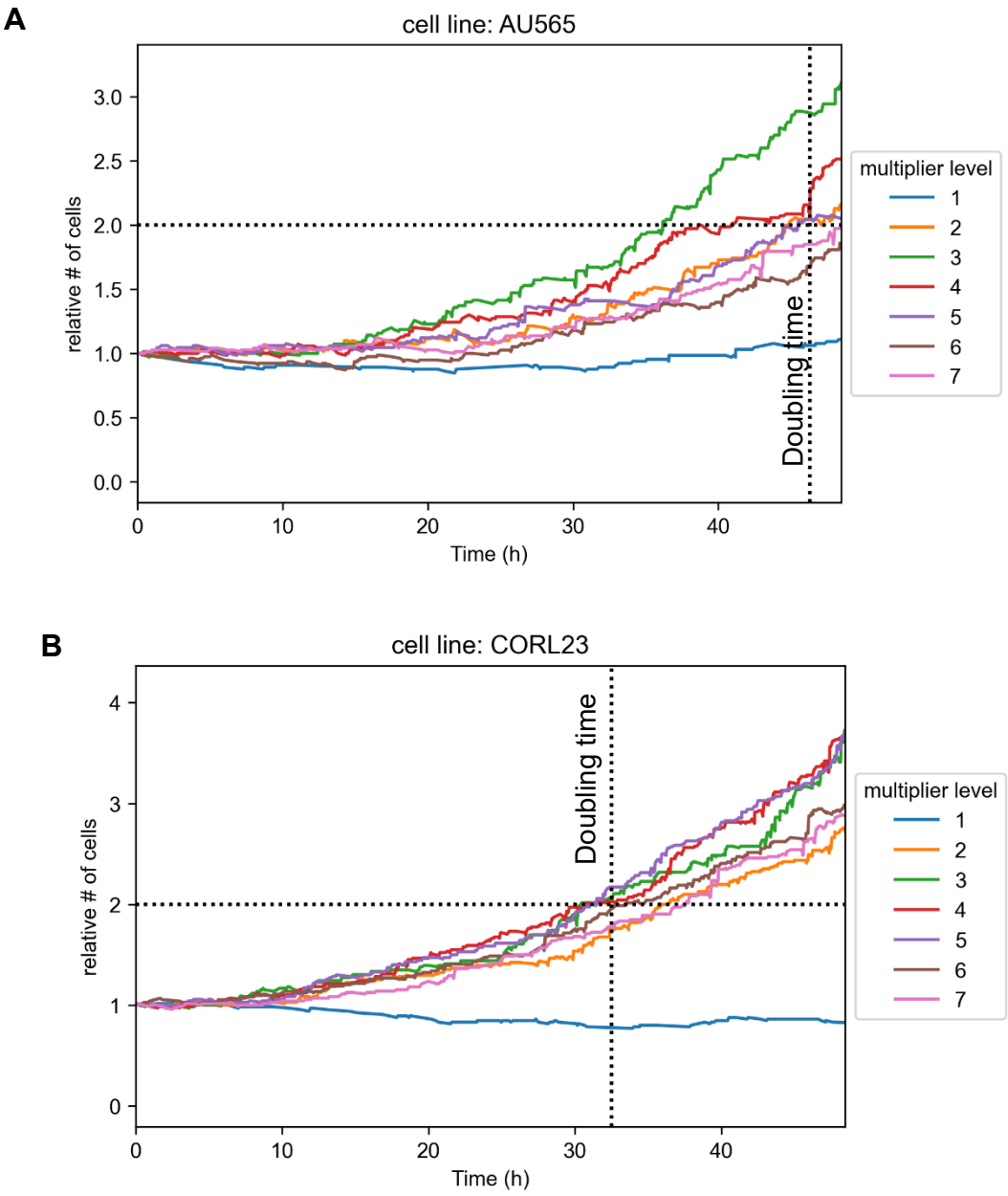


Figure 5.16: Simulated population dynamics of examples of cell lines categorized into group 1 as per simulated growth behavior, (A) AU565 and (B) CORL23. The multiplier levels indicate the ascending order of log spaced multipliers from 10^{-3} to 10^{3} representing growth media of varying strength applied to each cell population simulation. Vertical dashed line represents experimentally observed doubling time.

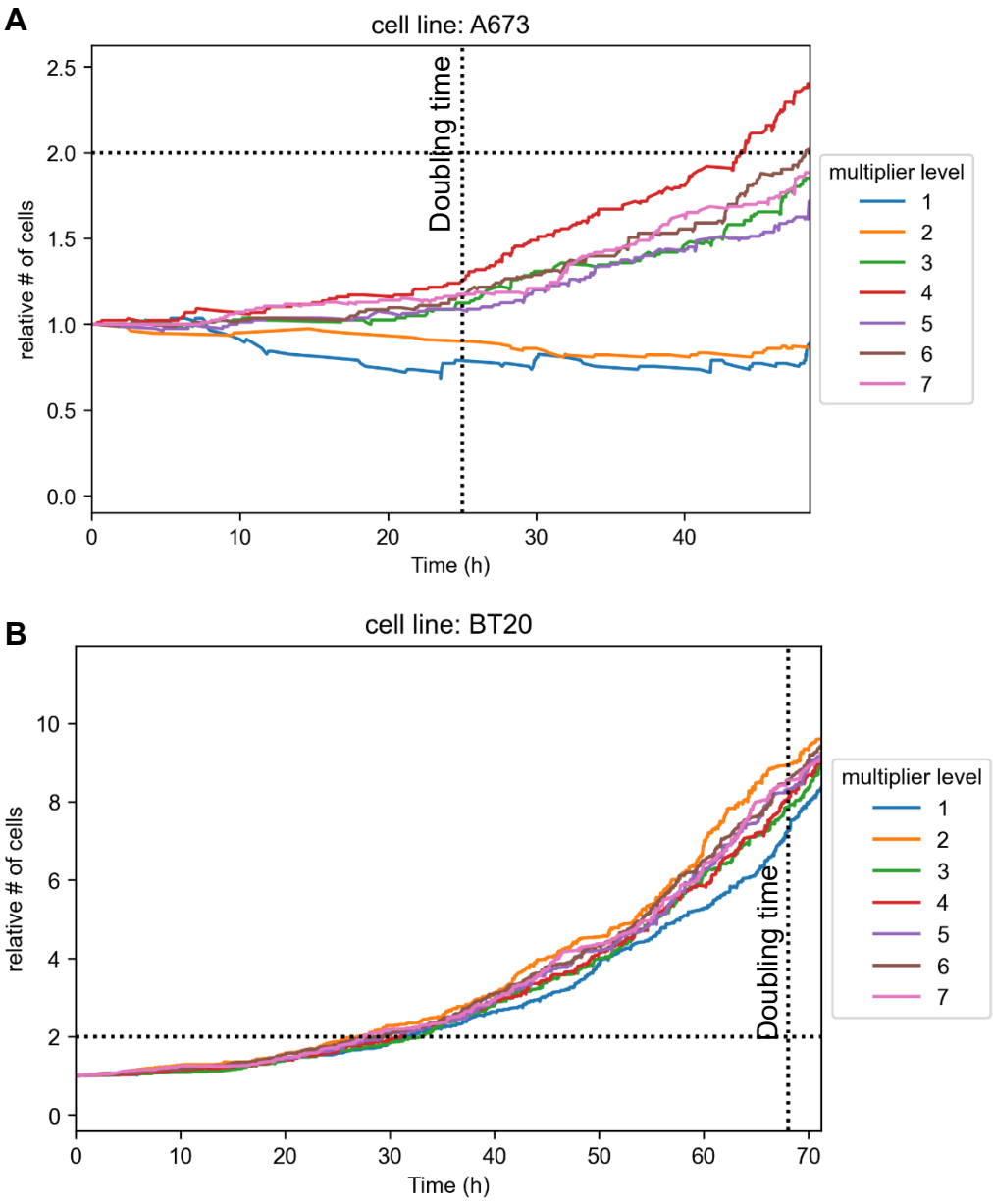


Figure 5.17: Simulated population dynamics of examples of cell lines categorized into group 2 as per simulated growth behavior, (A) A673 and (B) BT20. The multiplier levels indicate the ascending order of log spaced multipliers from 10^{-3} to 10^{3} representing growth media of varying strength applied to each cell population simulation. Vertical dashed line represents experimentally observed doubling time.

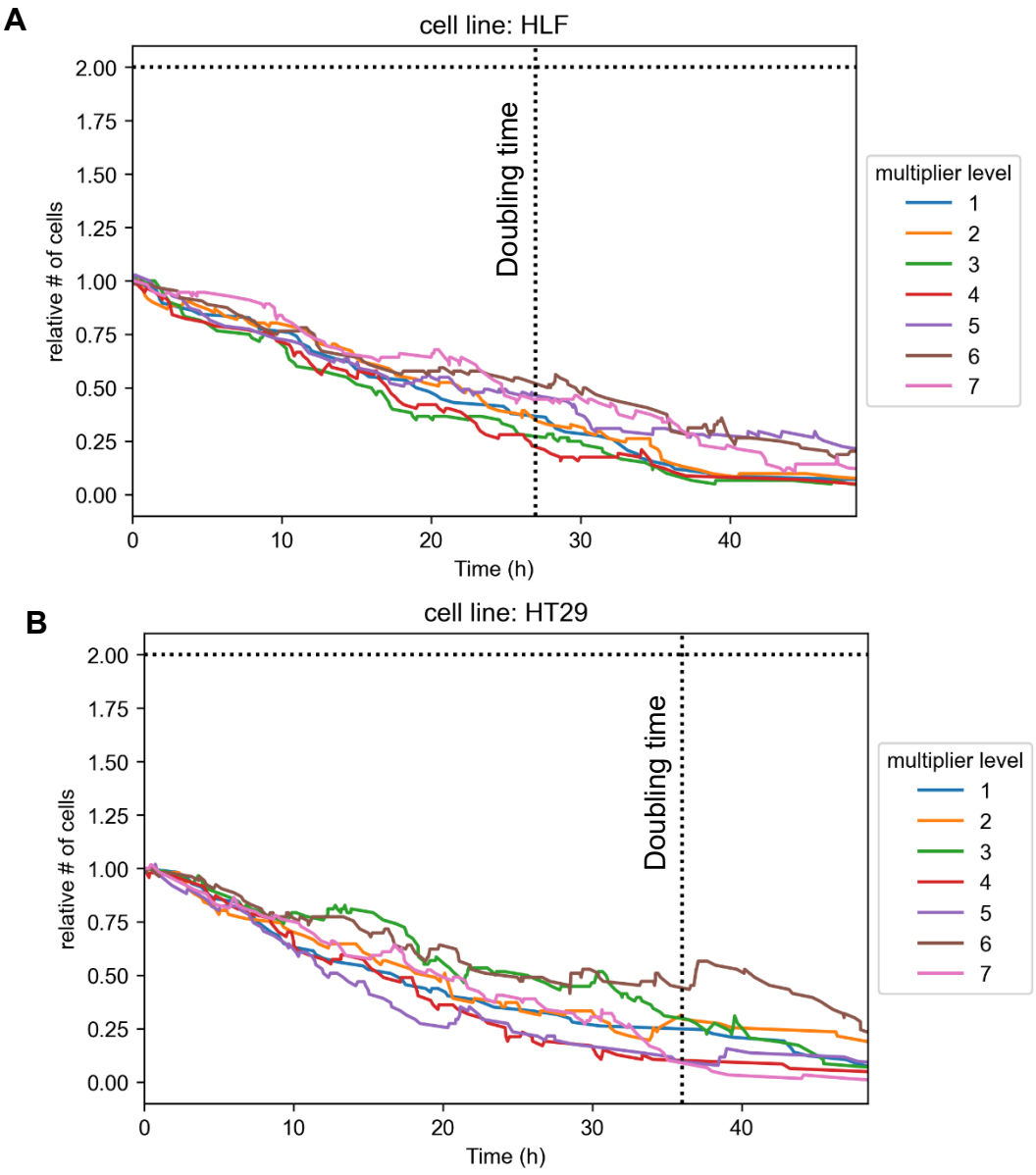


Figure 5.18: Simulated population dynamics of examples of cell lines categorized into group 3 as per simulated growth behavior, (A) HLF and (B) HT29. The multiplier levels indicate the ascending order of log spaced multipliers from 10^{-3} to 10^{3} representing growth media of varying strength applied to each cell population simulation. Vertical dashed line represents experimentally observed doubling time.

5.7 Results III – Dose Response Simulations Across Cell Lines

Thus far we have created cell line specific variants of the SPARCED model with initialization. Furthermore, we have identified 28 cell lines which may be used to create representations of dynamic cell populations consistent with experimental observations. Drug sensitivity profiles for all these cell lines across 24 anticancer drugs have been published in the Cancer Cell Line Encyclopedia[33]. In this section we explore whether our mechanistic cell population simulation framework can explain the drug sensitivity and resistance mechanisms across multiple cell line contexts. For this purpose, we selected 17 of those 24 anticancer drugs which are targeted therapeutic agents and had protein targets already included as species within the SPARCED model. We then retrieved the KINOMEscan datasets for these drugs from the HMS LINCS database[206]. These datasets contain a summary of binding affinity of each drug to a wide range of subcellular target proteins. For each drug, we identified the target proteins which are part of our modeled signal transduction network, and included the corresponding drug actions based on the reported binding affinity. In this manner, we included drug actions of all 17 selected drugs.

In this section, we focus on Mirdametinib, which is an inhibitor of the MEK kinase. As per the drug sensitivity profile reported in CCLE, Mirdametinib possesses a broad range of response for our panel of 28 initialized cell lines (Figure 5.19) with at half of those resistant to it and at least 11 cell lines strongly sensitive to it. To what extent can our mechanistic cell population simulations recapitulate the observed sensitivity and resistance across cell lines contexts? To answer this question, we conducted dose

response simulations for Mirdametinib doses ranging from 0.008 uM to 8 uM with our selected cell lines.

Interestingly, the dose response simulations also generated a spectrum of sensitivity to Mirdametinib dose for the selected cell lines. Here as an example we present the cell lines PANC0403 and AU565 (Fig. 5.20), out of which PANC0403 was observed to be sensitive to Mirdametinib in both simulations and experiments where AU565 was insensitive in both cases. In CCLE, the metric for drug sensitivity profiles is percentage of cell viability which is not directly comparable to the GR-score for simulations. However, we may estimate the nature of sensitivity based on certain characteristic features of the dose response curves. Specifically, slope, and area under the curve (AUC). Here, a cell line that is more sensitive will tend to have a more negative slope and lower AUC in its dose response curve.

The dose response simulations for one of the cell lines, MDAMB436 generated inconclusive results, presumably due to its low growth rate and high variability. We considered the Mirdametinib dose response curves for the remaining 27 cell lines and generated the slope vs AUC scatter plot for both simulation results and experimental dataset (Fig. 5.21). The spread of datapoints on these plots is representative of the wide range of dose responses observed in experiments which to an extent simulations were also able to capture.

If we were to obtain a binary classification of sensitive and insensitive cell lines, we may divide the slope vs AUC space with a line with negative slope. Here, ambiguity may arise from the uncertainty associated with discretization of a continuum. For this purpose, the line was drawn based on the visual observation of the dose response

curves of cell lines near the boundary line such that cell lines of the same class may occupy the same side of the line. Based on this classification, the experimental drug sensitivity profile had 15 sensitive and 12 insensitive cell lines. On the other hand, the simulated drug sensitivity profiles had 6 sensitive and 21 insensitive cell lines. Interestingly, no false positives were predicted in simulations. Hence, all 6 cell lines predicted to be sensitive in simulations were also observed as such in experimental data. Consequently, the 12 insensitive cell lines as per experimental data were also predicted to be insensitive in simulations. The final results for this classification have been summarized in Table 5.3.

One of the standard methods of evaluating binary classification is the Matthew's Correlation Coefficient (MCC)[207]. For this, we constructed the confusion matrix by filling in the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) as described in Fig. 5.22. Hence, we may calculate the value of MCC using the following formula:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} = 0.478$$

Here, the resulting value of 0.478 indicates that our method has moderate predictive capability and has performed significantly better than any method that might generate random classifications, despite the numerous sources of uncertainties in our approach. This result instills confidence in our ability to relate drug sensitivity to omics informed single cell signal transduction pathway activities. We anticipate that by resolving the uncertainties and knowledge gaps we have discussed, the overall prediction outcome will improve.

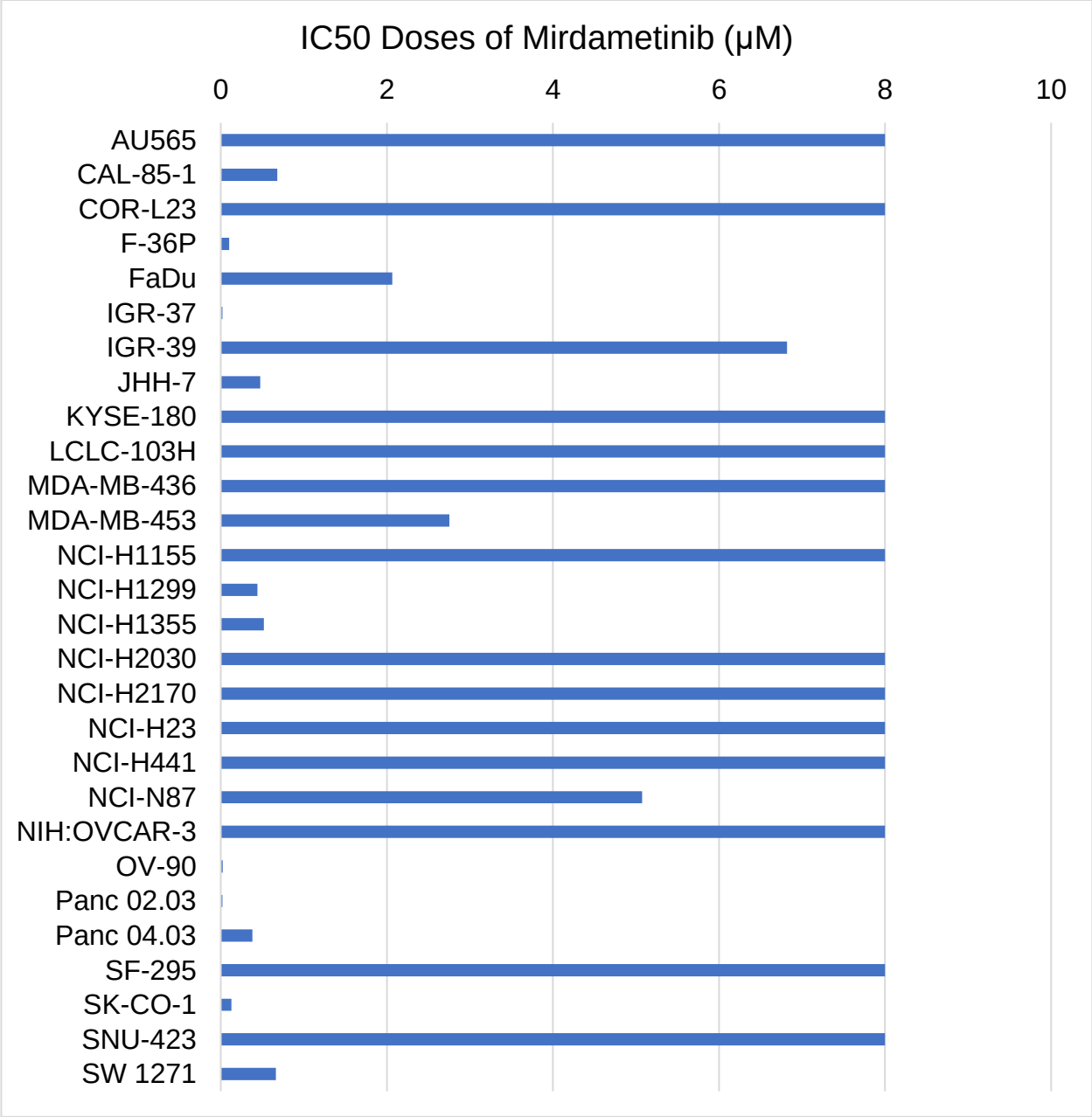


Figure 5.19: Sensitivity profile of Mirdametinib across our panel of initialized cell lines.

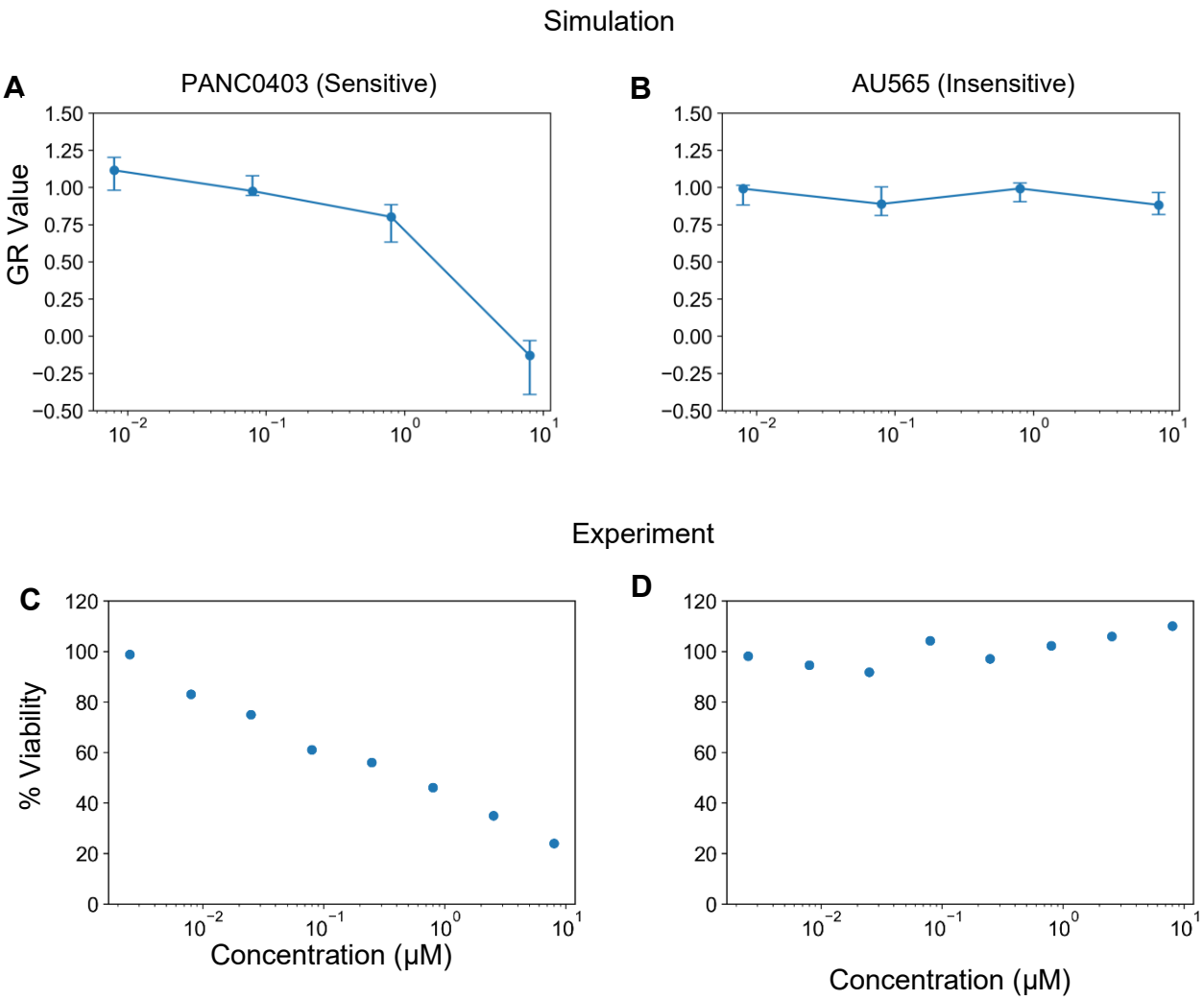


Figure 5.20: (A,B) Mirdametinib dose response simulation results for cell lines (A) PANC0403, and (B) AU565 expressed growth rate inhibition scale (GR-score)

(C,D) Experimentally reported % cell viability as a function of Mirdametinib concentration for (C) PANC0403 and (D) AU565

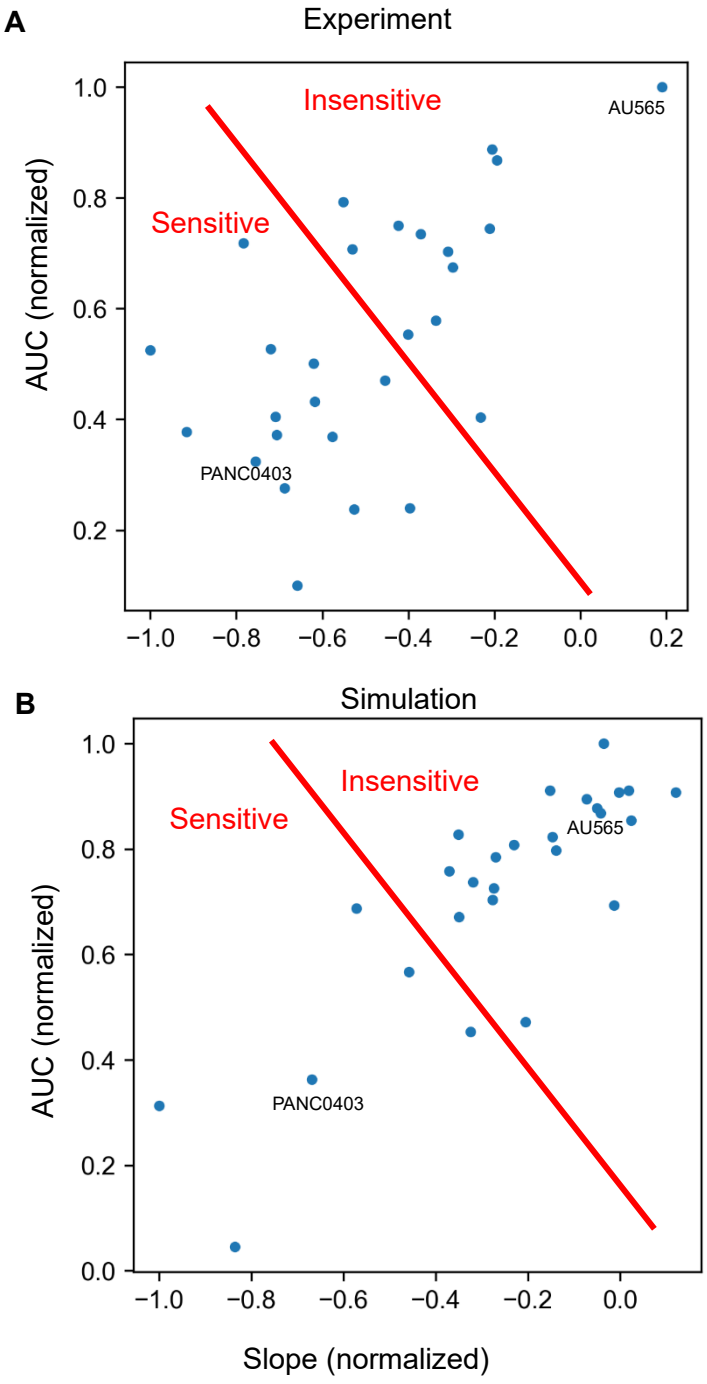


Figure 5.21: Comparison of slope vs area under the curve (AUC) for experimental (A) and simulated (B) dose response curves of Mirdametinib across various cell lines. A decision line with a negative slop has been drawn to classify cell lines into sensitive and insensitive categories.

Table 5.3: Binary classification of cell lines as sensitive and insensitive to Mirdametinib according to experimental data and simulation results.

Cell line	Classification for experimental data	Classification for simulation results	Match/mismatch
AU565	Insensitive	Insensitive	Match
CAL851	Sensitive	Insensitive	Mismatch
CORL23	Insensitive	Insensitive	Match
F36P	Insensitive	Insensitive	Match
FADU	Sensitive	Sensitive	Match
IGR37	Sensitive	Insensitive	Mismatch
IGR39	Sensitive	Insensitive	Mismatch
JHH7	Sensitive	Insensitive	Mismatch
KYSE180	Insensitive	Insensitive	Match
LCLC103H	Insensitive	Insensitive	Match
MDAMB453	Sensitive	Sensitive	Match
NCIH1155	Insensitive	Insensitive	Match
NCIH1299	Sensitive	Insensitive	Mismatch
NCIH1355	Sensitive	Sensitive	Match
NCIH2030	Insensitive	Insensitive	Match
NCIH2170	Insensitive	Insensitive	Match
NCIH23	Sensitive	Insensitive	Mismatch
NCIH441	Insensitive	Insensitive	Match

Cell line	Classification for experimental data	Classification for simulation results	Match/mismatch
NCIN87	Sensitive	Sensitive	Match
NIHOVCAR3	Insensitive	Insensitive	Match
OV90	Sensitive	Insensitive	Mismatch
PANC0203	Sensitive	Insensitive	Mismatch
PANC0403	Sensitive	Sensitive	Match
SF295	Insensitive	Insensitive	Match
SKCO1	Sensitive	Insensitive	Mismatch
SNU423	Insensitive	Insensitive	Match
SW1271	Sensitive	Sensitive	Match

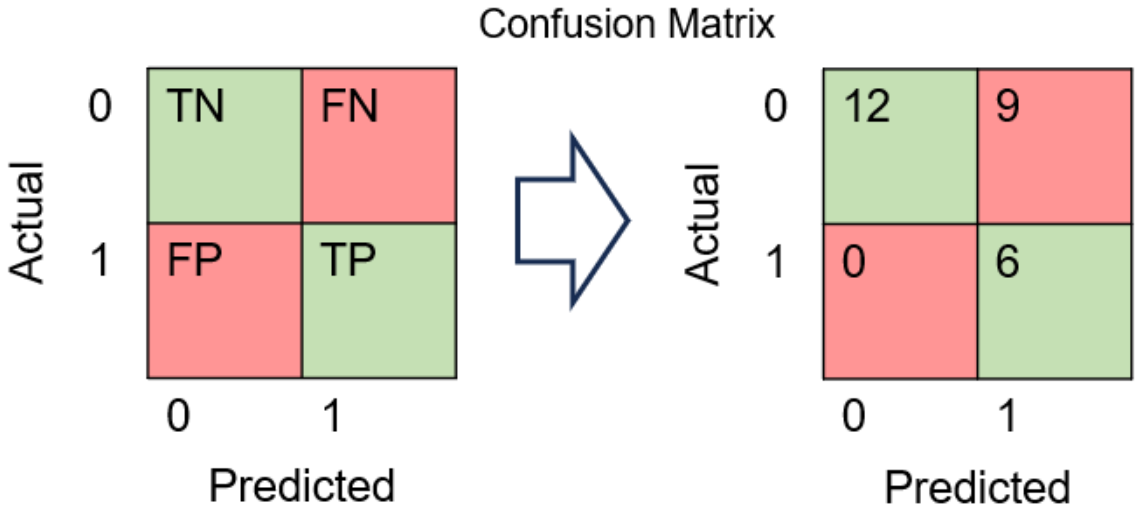


Figure 5.22: Evaluation of binary classification of Mirdametinib sensitivity prediction for selected cancer cell line. The confusion matrix is constructed by filling in the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

5.8 Methods – Pharmacodynamic Modeling

We represented the pharmacodynamics of each drug by including its interactions with modeled target species according to the binding affinity reported in the LINCS KINOMEScan datasets. For all drugs, we assume they are transported through the cell by diffusion to bind reversibly with their intracellular targets. We further assume that binding of drug with its target prevents further interactions and post translational modifications of the target. A summary of included drugs and their targets have been provided in Table 5.3. As an example we describe the drug actions of Mirdametinib below. Reactions representing Mirdametinib drug actions are:

Mirdametinib + MEK => Mirdametinib bound MEK

Mirdametinib bound MEK => Mirdametinib + MEK

Mirdametinib bound MEK => Ø

As a validation for the modeled drug action, we performed deterministic simulation of the SPARCED model in the MCF10A context. Serum-starved cells incubated for 30 minutes were added with 3.3 nM EGF, 0.005 nM HGF, 1721.0 nM insulin and 1000 nM Mirdametinib. Simulated trajectory of the drug, drug bound target and downstream target activity for drug dose and control conditions (Figure 5.20) confirm the intended outcome of drug action at the single cell level as inhibition of MEK results in significant reduction of its activity.

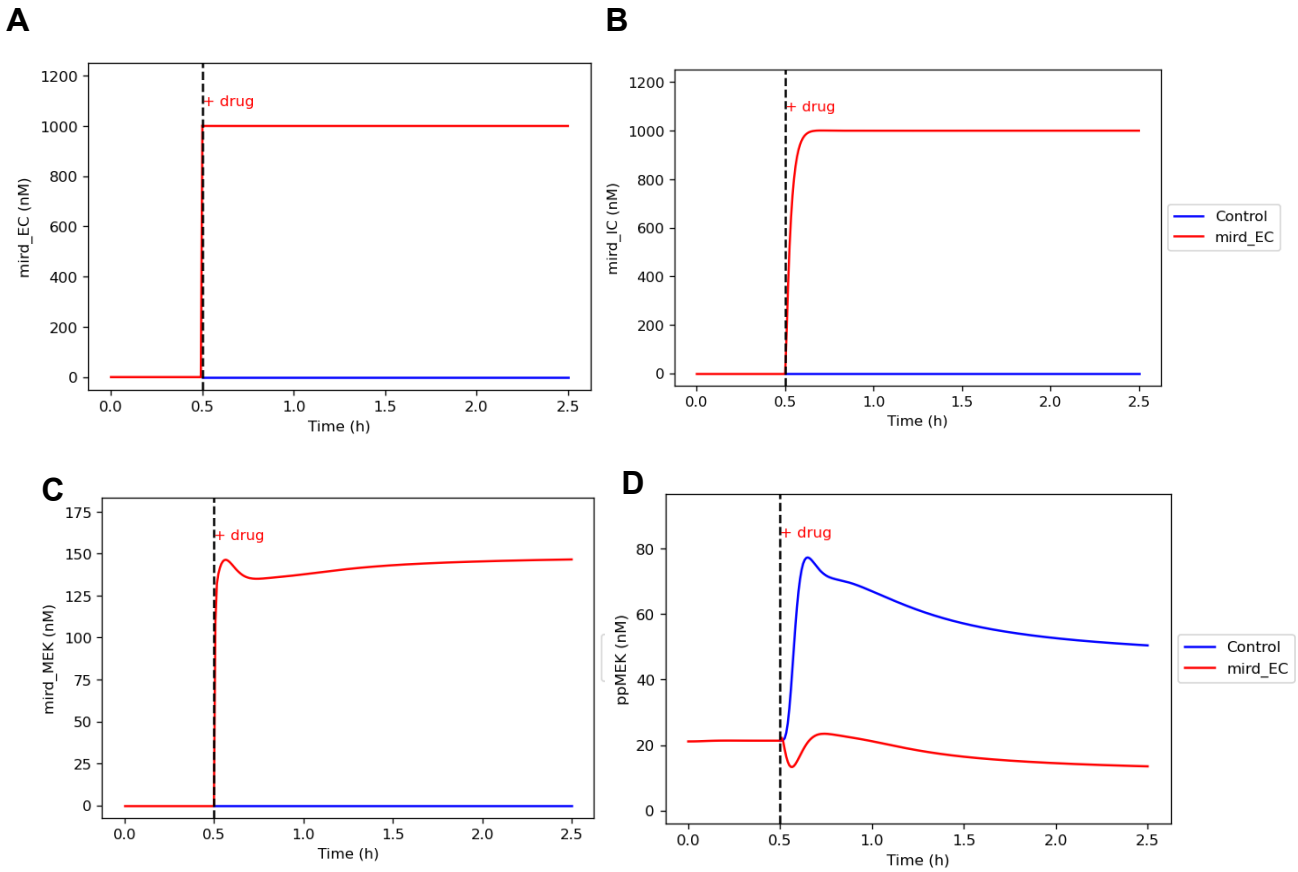


Figure 5.23: Deterministic simulation results for Mirdametinib drug action validation. Simulated trajectory of extracellular (A) and intracellular (B) Mirdametinib, as well as drug bound target (MEK) (C) and downstream target activity (ppMEK) (D) for drug dose and control conditions confirm the effect of drug at single cell level.

Table 5.4: Summary of included drug actions

Drug	Alternative name(s)	Species name	Targets in the model
Erlotinib		erlot_EC	EGFR (E1), ERBB2 (E2), ERBB3 (E3), MET (Met)
Lapatinib		lapat_EC	EGFR (E1), ERBB2 (E2), ERBB4 (E4)
PHA665752		METi_EC	MET (Met), INSR(Isr), FGFR2(Fr), CDK4/6 (Md), MAP2K1/2 (MEK), PDGFRA/B (Pr), S6K1 (S6K)
Crizotinib	PF-2341066	criz_EC	IGF1R (Ir), INSR (Isr), MET (Met), S6K1 (S6K)
TAE684		Tae_EC	CDK2(Ma, Me), CHEK1 (Chk1), EGFR (E1), ERBB2 (E2), ERBB4 (E4), FGFR1/2 (Fr), IGF1R (Ir), INSR (Isr), MET (Met), PDGFRA/B (Pr), RSK1/2/3/4 (RSK), WEE1 (Wee1), CDK4/6 (Md), MAP2K1/2 (MEK), S6K1 (S6K)
Vandetanib	Zactima	vandet_EC	EGFR (E1), ERBB2 (E2), ERBB3 (E3), ERBB4 (E4), MET (Met), FGFR1/FGFR2 (Fr)
Nilotinib	Tasigna	nilot_EC	BRAF (BRaf), PDGFRA/B (Pr), RAF1 (CRaf)
Saracatinib	AZD0530	sarac_EC	EGFR(E1), PDGFRB(Pr)
Sorafenib	BAY-439006, Nexavar	soraf_EC	BRAF (BRaf), CDK2 (Ma, Me) , FGFR1/FGFR2 (Fr), RAF1 (CRaf)

Drug	Alternative name(s)	Species name	Targets in the model
Dovitinib	TKI258	dovit_EC	CDK4 (Md), CHEK1 (Chk1), FGFR1/FGFR2 (Fr), GSK3B (GSK3b), WEE1 (Wee1)
Palbociclib	PD-0332991	palbo_EC	CDK4/6 (Md)
AEW541		aew_EC	IGF1R (Ir)
RAF265	CHIR265	RAFi_EC	BRAF (BRaf), RAF1 (CRaf), Fr (FGFR1), MET (Met), PDGFRA/B (Pr)
PLX4720		plx_EC	BRAF (BRaf), RAF1 (CRaf), FGFR1/2 (Fr), MP2K1/2 (MEK)
Mirdametinib	PD0325901	mird_EC	MAP2K1/MAP2K2 (MEK)
Selumetinib	AZD6244	selum_EC	EGFR (E1), MAP2K1/MAP2K2 (MEK)
Nutlin3		nutlin_EC	TP53 (p53inac), MDM2, MDM4

5.9 Discussion

An overarching goal of our work is to develop predictive understanding of anticancer drug response at the single cell level while accounting for the biomolecular mechanisms that drive such outcome. To a great extent such capability relies on how well our single cell model may describe the biological context of different tumor types. Application of high throughput omics techniques is currently one of the most prevalent methods for characterizing and understanding the molecular properties of cancer cell lines and tumor samples. In this chapter, we present a strategy for the omics informed context definition in single cell anti-cancer pharmacodynamic modeling. We start our work with the SPARCED model, a single cell model of stochastic cell proliferation and death signaling. The initial version of the model introduced the “initialization” procedure, which is used to determine certain parameter values and initial conditions required to represent a certain cell line based on its genomic, transcriptomic and proteomic data. It is a computationally intensive procedure whereby the model is subject to stepwise iterative unit testing focused on specific biological functionalities, such as protein and level conservation, cell cycle, DNA damage, and apoptosis. Even though the first initialization pipeline was successfully applied to the U87 glioma cell line, the multi-module hybrid nature of the simulation algorithm imposed certain limitations in the manner intensive computations could be performed on the model. As a result, certain important features of the model excluded from the initialization process, such as basal ERK and AKT signaling, transcriptional activation, mRNA level conservation, and survival signaling. We later made modifications to the model structure and simulation

algorithm because of which we were able to achieve more than 200-fold increase in deterministic simulation computation speed.

This significant modification helped us overcome the previous technical limitations which allowed us to revise the initialization procedure and build a more robust pipeline. This revised initialization procedure was applied to omics data retrieved from 251 cell lines from the Cancer Cell Line Encyclopedia, 59 of which passed the initialization procedure successfully. Next we attempted to evaluate the applicability of cell-line specific single cell models in dose response prediction. As a first step, we created representations of a dynamic cell population using these models. For this, we utilized the cell population simulation framework developed in Chapter 3. For each cell lines, we defined growth conditions using varying doses of growth factors included in the model. In cell population simulations, we observed context specific differential growth rates across our panel of 59 cell lines. However, for 28 of these cell lines, the experimentally observed growth rates were within the range of simulated growth rates. Since these cell line models matched a functional outcome of experimental observations, we selected these for context-specific drug dose response simulations.

The Cancer Cell Line Encyclopedia contains drug sensitivity profiles for the selected cell lines across 24 anticancer drugs. For 17 of these drugs, the known drug-target interactions could be represented in our model by binding of drugs with one or more of modeled species. To evaluate the context specific drug response predictions, we chose Mirdametinib, which is an inhibitor of the MEK kinase, and demonstrates a wide range of sensitivity profiles for our selected cell lines.

The initialization procedure that we devised has several notable limitations since it fails to function for a majority of our selected cell lines. It highlights the need for additional mechanisms in the model and corresponding initialization steps to tune them. One of the initialization steps where a large number of cell lines failed is the basal ERK signaling tuning step. For these cell lines, it was not possible to maintain even a minimal level of ERK activity after integrating their proteomic levels, due to low levels of certain upstream signaling proteins. A possible explanation is that many of these cell lines might possess genomic aberrations resulting in constitutive ERK activity which does not rely on any upstream protein level. Since our model does not account for any such mechanism, the initialization procedure falls short. Future iterations of this work may need to include mechanisms that can potentially explain this behavior to successfully integrate these cell lines.

Another initialization step where we experienced failures is the tuning of transcriptional activators, AP1 and MYC, which regulate cyclin D overexpression for the initiation of cell cycle. One of the key factors in this regard is p21, an inhibitor of the cell cycle process. For the cell lines that failed this step, p21 was found to be significantly high compared to the original MCF10A context. It is also a protein for which we were unable to obtain proteomics values due to lack of reliable data. To account for this lapse in information, we estimated the protein level based on the protein to mRNA ratio observed in the MCF10A context. A possible explanation for failure in this step could be an undiscovered transcriptional regulatory mechanism acting on p21 such that cells may retain cell cycle functionality.

Apoptosis is another functionality of the model which was not compatible with the omics context of all cell lines. For these cell lines, integration of omics dataset resulted in either too aggressive apoptosis signaling, resulting in immediate cell death, or a complete lack of apoptosis. A potential solution could be introduction of more detailed apoptosis signaling mechanisms with apoptotic regulators that are not currently included in the model.

A significant aspect about the molecular characteristics of cancer cells is the effect of gain of function or loss of function mutations which are yet to be integrated into the model. In many cases we observed mismatches between experimental observations and simulation results which highlights the need for further investigation into the extent to which such mutations may affect the biomolecular network. A possible strategy to include effects of these mechanism may include modification of the model structure to add mutant variant of species and their resulting interactions. Consequently, initialization procedure will require modifications to accommodate these mechanisms.

Chapter 6

A PRELIMINARY REVISION OF THE CELL CYCLE SUBMODEL

6.1 Introduction

The cell cycle pathway is a series of events that occur in a eukaryotic cell leading to its division and duplication. It is one of the most essential cellular functions that ensures growth and maintenance of living tissues. In normal cells, it is a manifestation of proliferative stimulation in a cell. It is also one of the pathways which may be subject to dysregulation due to genomic aberration in cancer cells leading to uncontrolled cell growth. Hence, it is also an ideal candidate for targeted cancer therapy whereby drugs are designed to inhibit specific components of the pathway[208]. Currently, inclusion of the cell cycle submodel in the SPARCED models allows us to observe the stochastic fate of cells undergoing division. Current cell cycle submodel in the SPARCED single cell model have been adopted from an earlier work of Gérard and Goldbeter[27]. However, there are certain technical limitations within this submodel precluding its integration with the gene expression module. Previously, inclusion of gene expression noise within the cell cycle submodel resulted in instability and irregular behavior[32]. The genome is the prime foundation of biological systems. Gene expression is an inherently stochastic process[82] which gives rise to cell-to-cell variability in the mRNA and protein levels and their dynamics. The highly conserved nature of biological pathway implies the existence of noise dampening mechanism in the network modalities. The behavior of the cell cycle submodel in presence of gene expression noise implies the presence of undiscovered

mechanism that provides the cell cycle with natural robustness against gene expression noise. Hence, even though the cell cycle submodel describes a manifestation of proliferative signaling in cells, it is partially inconsistent with their molecular biology. The results of Palbociclib dose response simulations from Chapter 4 also implies that the cell cycle submodel is missing key regulatory mechanisms which may help explain drug response under certain conditions. To cope with the instability observed in cell cycle submodel, mRNA levels of most species associated with cell cycle were kept at a constant. This caused a disconnect between the integration of omics data and gene expression noise and the current cell cycle submodel. This arrangement is inconsistent with the design principles that were followed in other parts of the model. As a result, we are unable to account for the effects of omics context on matters concerning the cell cycle. In this chapter, we discuss development of a revised cell cycle submodel as a potential solution. We built a preliminary version of this submodel capable of representing the biomolecular interactions for initiation of cell cycle in response to growth stimulation. In this submodel, the involved species are informed by their omics levels in the MCF10A context. Application of growth stimulus in deterministic simulation captures the characteristic molecular signatures of cell cycle initiation and S-phase entry. Further work is needed such that this submodel may capture progression of cell cycle, G2/M phase transition and cell cycle completion.

6.2 Current Cell Cycle Submodel and Limitations

Current cell cycle submodel has been derived from an earlier work of Gérard Goldbeter[27]. It describes the initiation of cell cycle by the transcriptional upregulation of cyclin D by the transcription factors AP1 and cMyc in G0 phase of cell cycle. Active

cyclin D then phosphorylates Rb to de-repress E2F transcription. Furthermore, E2F upregulates cyclin E and cyclin A, representing the beginning and progression of S-phase and transition into G2 phase. It is followed by the activation of cyclin B/CDK1, marking the beginning of mitosis, which completes the cell cycle and returns the cell to G1 phase. The submodel correctly maintains the order and timing of cyclin expressions corresponding to distinct phases. However, a major portion of the submodel is not consistent with the design principles of the rest of the SPARCED model as it is not connected to the gene expression module. It was discovered that inclusion of gene expression noise resulted in irregular and uncontrollable behavior of the cell cycle, such as spontaneous cycling in absence of cyclin D induction, disorderly expression and upregulation of various cyclins, and a lack of regular frequency, amplitude and duration of cyclin peaks. It implies the existence of undiscovered cellular mechanisms that provide robustness of cell cycle to gene expression noise. Since such a mechanism was not included at the time of publication, the cell cycle submodel was restricted from the gene expression module. To maintain stability of the cell cycle functionality, the mRNA levels of most of the species involved in the process were kept to a constant level. As a result, the omics input corresponding to individual cell lines do not inform the cell cycle specific mechanism. This is a significant limitation of the SPARCED single cell model, as it precludes derivation of cell line context specific insights related to cell cycle functionality, anti-cancer drug actions that target proteins in this pathway and resistance mechanisms that may help explain drug response. Further work is needed to ensure that the SPARCED single cell model can describe the activity of cell cycle without compromising the integrity of gene expression functions. We believe a revision of the

cell cycle submodel by inclusion of species and interactions that could potentially explain the natural robustness of cell cycle from gene expression noise may help complete this work. Possible mechanisms could include more recently discovered inhibitory regulators of the cyclin dependent kinases (CDKs) such as the Ink4 and the Cip/Kip family of regulators and transcriptional repressors[191]. Furthermore, several regulatory mechanisms currently included in the cell cycle submodel are not consistent with the contemporary knowledge of the cell cycle pathway. For example, Rb has been included as a direct negative regulator of cyclin D, cyclin E and cyclin A, even though its regulatory function is known to rely on sequestering of the transcriptional activator E2F[191]. Currently, E2F is only included as a transcriptional activator and a positive regulator of cell cycle. But more recently, the extended family of E2F regulators have been discovered with functionalities in both positive and negative regulation of the cell cycle[209]. We believe a revised cell cycle submodel with inclusion of these updated mechanisms could describe activation and completion of cell cycle in presence of growth stimulus while maintaining consistency with the gene expression module.

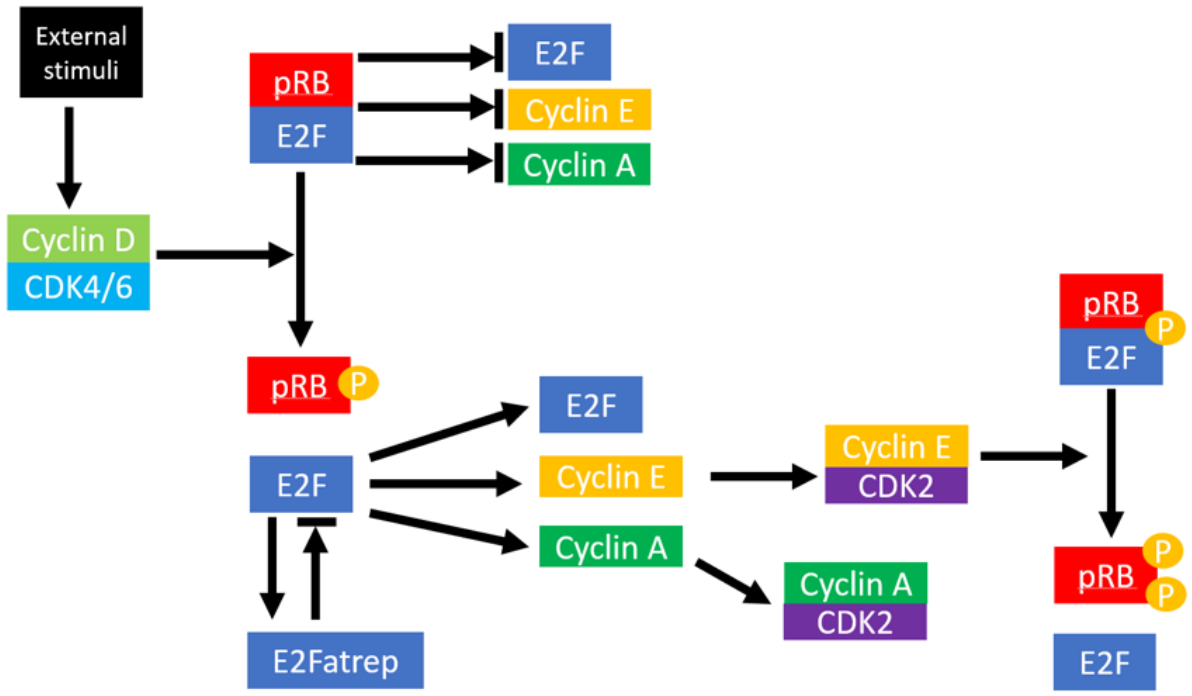


Figure 6.1: Kinetic scheme of the preliminary revision of cell cycle submodel. The included mechanism represents initiation of cell cycle, restriction point crossing and progression through S phase.

6.3 A Preliminary Revision of the Cell Cycle Submodel

We started with a modified version of the SPARCED model, from which the current cell cycle submodel had been removed. We reviewed current literature sources on the regulation of cell cycle pathway and its components and decided to include a set of species and interactions which are consistent with more recent knowledge of the pathway. Each new species is represented at the level of genes, mRNAs and proteins. We included the amount of each species based on their gene copy number, mRNA levels and protein concentrations as derived from the omics data of the MCF10A cell line. The reactions have been modeled according to the law of mass action. The series of included interactions and involvement of the species have been described below.

6.3.1 Initiation of Cell Cycle

Cell cycle is the result of proliferative signals being propagated throughout various signaling mechanisms, starting from the interaction between RTK family of receptors and their cognate ligands. This leads to phosphorylation of the tyrosine kinase residues of the receptor and subsequent binding of adaptor proteins. Consequently, this results in activation of the Ras/ERK and PI3K/AKT signaling pathways. Further downstream, the transcription factors AP1 and cMyc are upregulated. The signaling mechanisms in the SPARCED model until this point were left unchanged. The next step is upregulation of Cyclin D by AP1 and cMyc, which is one of the first characteristic molecular signatures of the cell cycle.

Cell cycle is characterized by oscillatory upregulation and downregulation of the cyclin family of proteins. In mammalian cells, according to their order of activation, these proteins include cyclin D, cyclin E, cyclin A and cyclin B. The cell cycle progression is positively regulated by a family of protein kinases, called the cyclin-dependent kinases (CDKs). CDKs are specific to the type of cyclin they bind with. Cyclins binding to the CDKs leads to their activation. The order in which each CDK is activated is associated with the progression of cell cycle. The cell cycle is conceptually divided into four distinct phases. The main phases include the S phase, or synthesis phase and M phase, or mitosis phase. S and M phases are separated by two 'gap' phases. The gap phase between M and S phases is called G1, while the gap between S and M phases is called G2. When the cell exits the cell cycle due to absence of growth stimulus, it is known to be in G0 or quiescent state. In presence of growth factor, cyclin D is upregulated and binds to CDK4/6. During further progression of the cell cycle, cyclin E binds to CDK2 at the transition between G1 and S phase. During the S phase, cyclin A binds to CDK2. cyclin A also binds to CDK1 during the G2 phase, with cyclin B binding to CDK1 marking the M phase and completion of cell cycle. In the preliminary cell cycle submodel, we have included these cyclins and their corresponding CDKs. The binding and activation of cyclins and CDKs in the G1 phase, transition between G1 and S phases and progression of S phase has been included.

6.3.2 Transcriptional Regulation by E2F Transcription Factors

Another important group of regulators of the cell cycle process is the E2F family of transcriptional factors. Eight different types of E2F proteins (E2F1-8) have been known to be encoded by mammalian cells, with varying roles and actions[209,210]. Among

these, E2F1-3 are known to be activators of S phase cyclins, E and A. Hence, E2F1-3 play critical roles in progression of S phase. In contrast, E2F4-8 are known to be repressors of transcription. Their actions are known to regulate the activators E2F1-3. Among these, E2F4-6 are known as “canonical repressors” which are constitutively expressed, but rely on activation mechanisms by other proteins. E2F7-8 are known as “atypical repressors”, of which the levels are known to peak during the late S phase. In the preliminary cell cycle submodel, we included all these E2F gene products as E2F activator, E2F repressor and E2F atypical repressor proteins. Transcriptional regulation mechanisms of these regulators, both activation and repression have been included in the gene regulation module of the SPARCED model. Here, we have a self-regulatory activation mechanism of E2F activator on itself which serves as a positive feedback loop to initiate cell cycle. E2F activator also has transcriptional activation mechanism on cyclin E and cyclin A. Meanwhile, E2F repressor, along with their activators p107 and p130 serve as transcriptional repressor of E2F activator, cyclin E and cyclin A.

6.3.3 Regulation of CDKs by Inhibitors

There are small inhibitory proteins expressed in mammalian cells which can bind directly onto CDKs and prevent activation of Cyclin-CDK complexes[211]. There are two major groups of these inhibitors. The first group is the Ink4 family, which consists of p16, p15, p18 and p19. These are specific to CDK4 and CDK6. The other group is the Cip/Kip family which consists of p21, p27 and p57. These proteins follow a more generalized mode of action and can bind to a wide range of CDKs. The key function of these inhibitors is the maintenance of quiescent state in absence of growth stimulus. Out of these, p21 has a specific role of halting cell cycle progression in the event of

excessive DNA damage. These inhibitors have and their actions on CDKs have been included in the preliminary cell cycle submodel.

6.3.4 Regulation by Pocket Proteins

Another important group of proteins is the Rb family, also known as “pocket proteins”. It includes the proteins, Rb, p107 and p130[212]. Out of these, p107 and p130 are known activators of repressor E2F, with a role to suppress activation of activator E2F-dependent transcriptions. Rb is a protein that maintains quiescent status by binding on to activator E2F. At the start of cell cycle, when cyclin D is upregulated, cyclin D-CDK4/6 complex enters an active state. This complex is known to phosphorylate E2F-bound Rb. Phosphorylation of Rb leads to its inactivation, ultimately freeing activator E2F. Once freed, the activator E2F may initiate its role in S-phase progression. This is a mechanism that acts as restriction point control for cell cycle. The Rb group of proteins, along with their role in repression has been included in the preliminary cell cycle submodel.

6.4 Submodel Validation

To evaluate the validity of included species and reactions in the preliminary cell cycle submodel, we ran deterministic simulations with and without growth stimulus in the MCF10A context. The simulation results (Figure 6.2) confirm a maintenance of quiescent status in absence of growth stimulus. With the addition of 100 nM of EGF, hergulin, HGF, PDGF, FGF, and IGF each and 1721.0 nM insulin, we can observe and upregulation of ERK and ATK activity. This in turn leads to the initiation of cell cycle as we can observe the point where Rb is hyperphosphorylated and E2F is upregulated, indicating that the cell has gone past the cell cycle restriction point. Finally, periodic

upregulation and downregulation of cyclin E and cyclin A, indicating progression into S phase can also be observed.

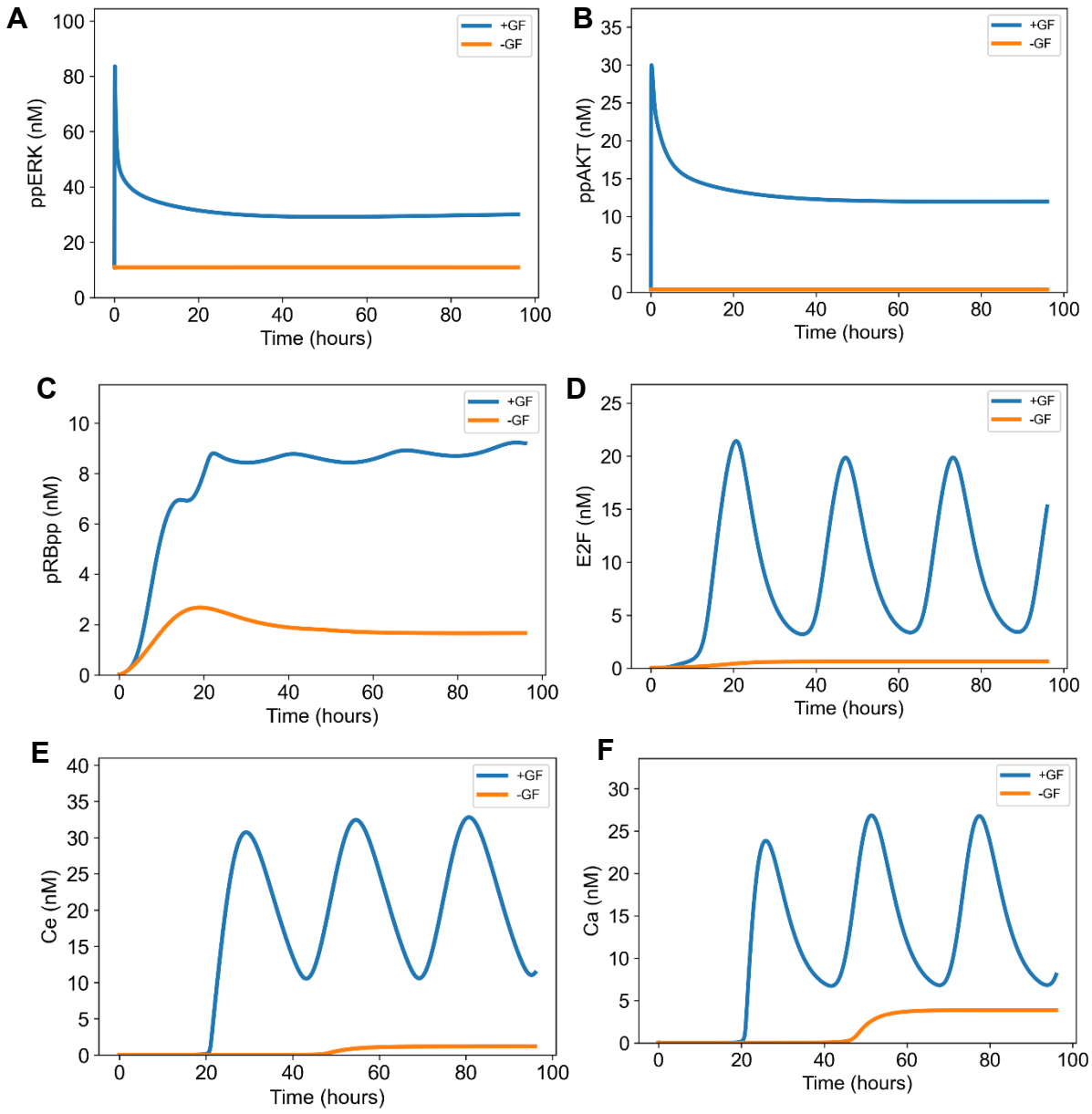


Figure 6.2: Results from deterministic simulation in MCF10A context showing ERK (A) and AKT (B) pathway activation in presence of growth factor stimulation. Pathway activity results in initiation of cell cycle whereby restriction point is traversed, as evident by the hyperphosphorylation of pRB (C) and upregulation of activator E2F (D). Furthermore, periodic upregulation and downregulation of cyclin E (E) and cyclin A (F), indicating S phase entry and progression can be observed.

Table 6.1: Included species in the revised cell cycle submodel

Species ID	Description	HGNC Gene Symbol
pRB	Rb (free)	RB1
pRBp	Rb (hypophosphorylated)	RB1
pRBpp	Rb (hyperphosphorylated)	RB1
E2F	Activator E2F	E2F1, E2F2, E2F3
Cd	Cyclin D	CCND1, CCND2, CCND3
Cd_Cdk46	Cyclin D/Cdk4-6 complex	CCND1, CCND2, CCND3, CDK4, CDK6
Cd_Cdk46_p27	Cyclin D/Cdk4-6 complex bound to p27	CCND1, CCND2, CCND3, CDK4, CDK6, CDKN1B
Ce	Cyclin E	CCNE1, CCNE2
Ce_Cdk2	Cyclin E/Cdk2 complex	CCNE1, CCNE2, CDK2
Skp2	Skp2	SKP2ï¿½
Ce_Cdk2_p27	Cyclin E/Cdk2 complex bound to p27	CCNE1, CCNE2, CDK2, CDKN1B
Pe	CDC25A Phosphatase	CDC25A
Pai	CDC25B Phosphatase	CDC25B
Pei	CDC25A Phosphatase (inactivated)	CDC25A

Species ID	Description	HGNC Gene Symbol
Pbi	CDC25C Phosphatase (inactivated)	CDC25C
Ca	Cyclin A	CCNA2
Ca_Cdk2	Cyclin A/Cdk2 complex	CCNA2, CDK2
Ca_Cdk2_p27	Cyclin A/Cdk2 complex bound to p27	CCNA2, CDK2, CDKN1B
p27	p27 inhibitor	CDKN1B
Cdh1i	Ubiquitin ligase Cdh1 (inactivated)	CDH1
Cdh1a	Ubiquitin ligase Cdh1 (activated)	CDH1
E2Fp	Phosphorylated activator E2F	E2F1, E2F2, E2F3
p27p	Phosphorylated p27 inhibitor	CDKN1B
Pa	CDC25B phosphatase	CDC25B
Cb	Cyclin B	CCNB1
Cb_Cdk1	Cyclin B/Cdk1 complex	CCNB1, CDK1
Cdc20i	Ubiquitin ligase Cdc20 (inactivated)	CDC20
Cdc20a	Ubiquitin ligase Cdc20 (activated)	CDC20
Pb	CDC25C Phosphatase	CDC25C

Species ID	Description	HGNC Gene Symbol
Wee1	Cell cycle checkpoint protein Wee1	WEE1
Wee1p	Cell cycle checkpoint protein Wee1 (phosphorylated)	WEE1
Cb_Cdk1_p27	Cyclin B/Cdk1 complex bound to p27	CCNB1, CDK1, CDKN1B
Chk1	Cell cycle checkpoint protein Chk1	CHEK1
pRB_E2F	Non phosphorylated Rb bound to E2F	RB1, E2F1, E2F2, E2F3
pRBp_E2F	Hypophosphorylated Rb bound to E2F	RB1, E2F1, E2F2, E2F3
p21	p21 inhibitor	CDKN1A
Cd_Cdk46_p21	Cyclin D/Cdk4-6 complex bound to p21	CCND1, CCND2, CCND3, CDK4, CDK6, CDKN1A
Ce_Cdk2_p21	Cyclin E/Cdk2 complex bound to p21	CCNE1, CCNE2, CDK2, CDKN1A
Ca_Cdk2_p21	Cyclin A/Cdk2 complex bound to p21	CCNA2, CDK2, CDKN1A
Cb_Cdk1_p21	Cyclin B/Cdk1 complex bound to p21	CCNB1, CDK1, CDKN1A
Cdk1	Cyclin dependent kinase 1	CDK1

Species ID	Description	HGNC Gene Symbol
Cdk2	Cyclin dependent kinase 2	CDK2
Cdk46	Cyclin dependent kinase 4/6	CDK4, CDK6
pRBpp_E2F	Hyperphosphorylated Rb bound to E2F	RB1, E2F1, E2F2, E2F3
Cd_Cdk46_pRB	Cyclin D/Cdk4-6 complex bound to Rb	CCND1, CCND2, CCND3, CDK4, CDK6, RB1
Cd_Cdk46_pRB_E2F	Cyclin D/Cdk4-6 complex bound to Rb/E2F complex	CCND1, CCND2, CCND3, CDK4, CDK6, RB1, E2F1, E2F2, E2F3
Ce_Cdk2_pRBp	Cyclin D/Cdk4-6 complex bound to hypophosphorylated Rb	CCND1, CCND2, CCND3, CDK4, CDK6, RB1
Ce_Cdk2_pRBp_E2F	Cyclin D/Cdk4-6 complex bound to hyperphosphorylated Rb/E2F complex	CCND1, CCND2, CCND3, CDK4, CDK6, RB1, E2F1, E2F2, E2F3
p18	p18 inhibitor	CDKN2C
p19	p19 inhibitor	CDKN2D
p57	p57 inhibitor	CDKN1C
Cd_Cdk46_p18	Cyclin D/Cdk4-6 complex bound to p18	CCND1, CCND2, CCND3, CDK4, CDK6, CDKN2C

Species ID	Description	HGNC Gene Symbol
Cd_Cdk46_p19	Cyclin D/Cdk4-6 complex bound to p19	CCND1, CCND2, CCND3, CDK4, CDK6, CDKN2D
Cd_Cdk46_p57	Cyclin D/Cdk4-6 complex bound to p57	CCND1, CCND2, CCND3, CDK4, CDK6, CDKN1C
Ce_Cdk2_p57	Cyclin E/Cdk2 complex bound to p57	CCNE1, CCNE2, CDK2, CDKN1C
Ca_Cdk2_p57	Cyclin A/Cdk2 complex bound to p57	CCNA2, CDK2, CDKN1C
Cb_Cdk1_p57	Cyclin B/Cdk1 complex bound to p57	CCNB1, CDK1, CDKN1C
E2Frep	Repressor E2F	E2F4, E2F5, E2F6
E2Fatrep	Atypical repressor E2F	E2F7, E2F8
p107	Rb family protein p107	RBL1
p130	Rb family protein p130	RBL2
p107_E2Frep	Repressor E2F/p107 complex	RBL1, E2F4, E2F5, E2F6
p130_E2Frep	Repressor E2F/p130 complex	RBL2, E2F4, E2F5, E2F6

Table 6.2: Included reactions in the revised cell cycle submodel

Reaction ID	Reaction equation	Rate law
vC1	pRB + E2F => pRB_E2F	k337 * pRB * E2F
vC2	pRB_E2F => pRB + E2F	k338 * pRB_E2F
vC3	Cd + Cdk46 => Cd_Cdk46	k339 * Cd * Cdk46
vC4	Cd_Cdk46 => Cd + Cdk46	k340 * Cd_Cdk46
vC5	pRB + Cd_Cdk46 => Cd_Cdk46_pRB	k341 * pRB * Cd_Cdk46
vC6	Cd_Cdk46_pRB => pRB + Cd_Cdk46	k342 * Cd_Cdk46_pRB
vC7	Cd_Cdk46 + pRB_E2F => Cd_Cdk46_pRB_E2F	k343 * Cd_Cdk46 * pRB_E2F
vC8	Cd_Cdk46_pRB_E2F => Cd_Cdk46 + pRB_E2F	k344 * Cd_Cdk46_pRB_E2F
vC9	Cd_Cdk46_pRB => pRBp + Cd_Cdk46	k345 * Cd_Cdk46_pRB
vC10	pRBp + E2F => pRBp_E2F	k346 * pRBp * E2F
vC11	pRBp_E2F => pRBp + E2F	k347 * pRBp_E2F
vC12	Cd_Cdk46_pRB_E2F => Cd_Cdk46 + pRBp_E2F	k348 * Cd_Cdk46_pRB_E2F
vC13	Ce + Cdk2 => Ce_Cdk2	k349 * Ce * Cdk2
vC14	Ce_Cdk2 => Ce + Cdk2	k350 * Ce_Cdk2

Reaction ID	Reaction equation	Rate law
vC15	pRBp + Ce_Cdk2 => Ce_Cdk2_pRBp	k351 * pRBp * Ce_Cdk2
vC16	Ce_Cdk2_pRBp => pRBp + Ce_Cdk2	k352 * Ce_Cdk2_pRBp
vC17	Ce_Cdk2_pRBp => pRBpp + Ce_Cdk2	k353 * Ce_Cdk2_pRBp
vC18	Ce_Cdk2 + pRBp_E2F => Ce_Cdk2_pRBp_E2F	k354 * Ce_Cdk2 * pRBp_E2F
vC19	Ce_Cdk2_pRBp_E2F => Ce_Cdk2 + pRBp_E2F	k355 * Ce_Cdk2_pRBp_E2F
vC20	Ce_Cdk2_pRBp_E2F => Ce_Cdk2 + pRBpp_E2F	k356 * Ce_Cdk2_pRBp_E2F
vC21	pRBpp_E2F => pRBpp + E2F	k357 * pRBpp_E2F
vC22	Cd_Cdk46 + p18 => Cd_Cdk46_p18	k358 * Cd_Cdk46 * p18
vC23	Cd_Cdk46_p18 => Cd_Cdk46 + p18	k359 * Cd_Cdk46_p18
vC24	Cd_Cdk46 + p19 => Cd_Cdk46_p19	k360 * Cd_Cdk46 * p19
vC25	Cd_Cdk46_p19 => Cd_Cdk46 + p19	k361 * Cd_Cdk46_p19

Reaction ID	Reaction equation	Rate law
vC26	Cd_Cdk46 + p21 => Cd_Cdk46_p21	k362 * Cd_Cdk46 * p21
vC27	Cd_Cdk46_p21 => Cd_Cdk46 + p21	k363 * Cd_Cdk46_p21
vC28	Ce_Cdk2 + p21 => Ce_Cdk2_p21	k364 * Ce_Cdk2 * p21
vC29	Ce_Cdk2_p21 => Ce_Cdk2 + p21	k365 * Ce_Cdk2_p21
vC30	Ca_Cdk2 + p21 => Ca_Cdk2_p21	k366 * Ca_Cdk2 * p21
vC31	Ca_Cdk2_p21 => Ca_Cdk2 + p21	k367 * Ca_Cdk2_p21
vC32	Cd_Cdk46 + p27 => Cd_Cdk46_p27	k368 * Cd_Cdk46 * p27
vC33	Cd_Cdk46_p27 => Cd_Cdk46 + p27	k369 * Cd_Cdk46_p27
vC34	Ce_Cdk2 + p27 => Ce_Cdk2_p27	k370 * Ce_Cdk2 * p27
vC35	Ce_Cdk2_p27 => Ce_Cdk2 + p27	k371 * Ce_Cdk2_p27
vC36	Ca_Cdk2 + p27 => Ca_Cdk2_p27	k372 * Ca_Cdk2 * p27

Reaction ID	Reaction equation	Rate law
vC37	Ca_Cdk2_p27 => Ca_Cdk2 + p27	k373 * Ca_Cdk2_p27
vC38	p27 + Cb_Cdk1 => Cb_Cdk1_p27	k374 * p27 * Cb_Cdk1
vC39	Cb_Cdk1_p27 => p27 + Cb_Cdk1	k375 * Cb_Cdk1_p27
vC40	Cd_Cdk46 + p57 => Cd_Cdk46_p57	k376 * Cd_Cdk46 * p57
vC41	Cd_Cdk46_p57 => Cd_Cdk46 + p57	k377 * Cd_Cdk46_p57
vC42	Ce_Cdk2 + p57 => Ce_Cdk2_p57	k378 * Ce_Cdk2 * p57
vC43	Ce_Cdk2_p57 => Ce_Cdk2 + p57	k379 * Ce_Cdk2_p57
vC44	Ca_Cdk2 + p57 => Ca_Cdk2_p57	k380 * Ca_Cdk2 * p57
vC45	Ca_Cdk2_p57 => Ca_Cdk2 + p57	k381 * Ca_Cdk2_p57
vC46	Cb_Cdk1 + p57 => Cb_Cdk1_p57	k382 * Cb_Cdk1 * p57
vC47	Cb_Cdk1_p57 => Cb_Cdk1 + p57	k383 * Cb_Cdk1_p57

Reaction ID	Reaction equation	Rate law
vC48	E2Frep + p107 => p107_E2Frep	k384 * E2Frep * p107
vC49	p107_E2Frep => E2Frep + p107	k385 * p107_E2Frep
vC50	E2Frep + p130 => p130_E2Frep	k386 * E2Frep * p130
vC51	p130_E2Frep => E2Frep + p130	k387 * p130_E2Frep
vC52	Ca + Cdk2 => Ca_Cdk2	k388 * Ca * Cdk2
vC53	Ca_Cdk2 => Ca + Cdk2	k389 * Ca_Cdk2
vC54	Cb + Cdk1 => Cb_Cdk1	k390 * Cb * Cdk1
vC55	Cb_Cdk1 => Cb + Cdk1	k391 * Cb_Cdk1
vCd1	Cd_Cdk46 => ∅	k392 * Cd_Cdk46
vCd3	Cd_Cdk46_p27 => ∅	k393 * Cd_Cdk46_p27
vCd4	Ce_Cdk2 => ∅	k394 * Ce_Cdk2
vCd6	Ce_Cdk2_p27 => ∅	k395 * Ce_Cdk2_p27
vCd7	Pe => ∅	k396 * Pe
vCd8	Ca_Cdk2 => ∅	k397 * Ca_Cdk2
vCd10	Ca_Cdk2_p27 => ∅	k398 * Ca_Cdk2_p27
vCd11	Cdh1i => ∅	k399 * Cdh1i
vCd12	E2Fp => ∅	k400 * E2Fp
vCd13	p27p => ∅	k401 * p27p
vCd14	Pa => ∅	k402 * Pa
vCd15	Cb_Cdk1 => ∅	k403 * Cb_Cdk1
vCd17	Cdc20a => ∅	k404 * Cdc20a

Reaction ID	Reaction equation	Rate law
vCd18	Pb => ∅	k405 * Pb
vCd19	Wee1p => ∅	k406 * Wee1p
vCd20	Cb_Cdk1_p27 => ∅	k407 * Cb_Cdk1_p27
vCd21	pRB_E2F => ∅	k408 * pRB_E2F
vCd22	pRBp_E2F => ∅	k409 * pRBp_E2F
vCd23	Cd_Cdk46_p21 => ∅	k410 * Cd_Cdk46_p21
vCd24	Ce_Cdk2_p21 => ∅	k411 * Ce_Cdk2_p21
vCd25	Ca_Cdk2_p21 => ∅	k412 * Ca_Cdk2_p21
vCd26	Cb_Cdk1_p21 => ∅	k413 * Cb_Cdk1_p21
vCd27	pRBpp_E2F => ∅	k414 * pRBpp_E2F
vCd28	Cd_Cdk46_pRB => ∅	k415 * Cd_Cdk46_pRB
vCd29	Cd_Cdk46_pRB_E2F => ∅	k416 * Cd_Cdk46_pRB_E2F
vCd30	Ce_Cdk2_pRBp => ∅	k417 * Ce_Cdk2_pRBp
vCd31	Ce_Cdk2_pRBp_E2F => ∅	k418 * Ce_Cdk2_pRBp_E2F
vCd36	pRBp => ∅	k419 * pRBp
vCd37	pRBpp => ∅	k420 * pRBpp
vCd38	Cd_Cdk46_p18 => ∅	k421 * Cd_Cdk46_p18
vCd39	Cd_Cdk46_p19 => ∅	k422 * Cd_Cdk46_p19
vCd40	Cd_Cdk46_p57 => ∅	k423 * Cd_Cdk46_p57
vCd41	Ce_Cdk2_p57 => ∅	k424 * Ce_Cdk2_p57
vCd42	Ca_Cdk2_p57 => ∅	k425 * Ca_Cdk2_p57

Reaction ID	Reaction equation	Rate law
vCd43	Cb_Cdk1_p57 => ∅	k426 * Cb_Cdk1_p57
vCd44	p107_E2Frep => ∅	k427 * p107_E2Frep
vCd45	p130_E2Frep => ∅	k428 * p130_E2Frep

6.5 Discussion

We envision the future application of the SPARCED model as a predictive tool for cancer therapy which may be generalizable across multiple tumor types originating from different tissues. Our current strategy towards this development includes testing the applicability of this model in various biological cell line contexts. For this purpose, we attempt to tailor the context of the model with omics input from different cell lines. Hence it is important to be able to take the effects of omics context into account for all aspects of the model. However, the status of the current cell cycle submodel presents a significant challenge to this approach. The gene expression module of the SPARCED submodel describes the activation and inactivation of model genes, and resulting transcription events leading to the synthesis of mRNAs. These events are simulated stochastically, giving rise to gene expression noise, which propagates into the protein interaction module creating cell to cell variability. Limitations of the cell cycle submodel prevent it from withstanding gene expression noise. To retain the cell cycle functionality of the SPARCED model, the cell cycle submodel was kept isolated from the processes that describe activation of genes and transcription. This is contrary to the biological function of cells and a significant knowledge gap our single cell model.

Developing a completely new cell cycle submodel, constructed from the ground up and aligned with contemporary knowledge from the literature on the cell cycle pathway, could potentially address this issue. In this chapter we discuss a potential solution for the revision of the cell cycle submodel. In a preliminary version of the revised cell cycle submodel, we included activities of cyclins and cyclin dependent kinases, regulatory activities of the E2F and Rb family of proteins and inhibitory

activities of some known Ink4 and Cip/Kip group of CDK inhibitors. This preliminary cell cycle submodel can currently represent initiation of cell cycle by upregulation of cyclin D, restriction point crossing by inactivating phosphorylation of Rb, subsequent release and upregulation of E2F, and entry and progression into S phase by upregulation of cyclin E and cyclin A. This is only a partial description of the possible solution. Further work is needed to describe upregulation of cyclin B and completion of cell cycle. We believe optimization of the activity of included inhibitors could potentially explain the robustness of cell cycle components from gene expression noise. A successful completion of this submodel will enable a more accurate representation of omics context in the cell cycle functionality. It has the potential to enhance the compatibility of the SPARCED model for cell line contexts in which our previously described initialization procedure.

Chapter 7

CONCLUSION

7.1 Conclusion

This dissertation was motivated by difficulties in cancer treatment and surrounding complexities that challenge innovations in enhancing treatment efficacy. The foundation for this work is a previously developed single cell mechanistic model[32] of stochastic proliferation and death signaling predictive cell fate outcomes which attempts to formalize the complex molecular physiology of human cells. In this dissertation, we evaluated the application of single cell pharmacodynamic modeling as a means to develop a bottom-up understanding of anticancer drug response. We focused on several aspects that challenge its enhancement of drug response prediction accuracy across biological context of different tumor types, namely, (1) enhancement of its accessibility, modularity and computational efficiency; (2) introducing methods for validation of modeled biomolecular processes and identification of associated knowledge gaps by comparison with experimental dose response data; and (3) enhancement of its adaptability to represent new biological context by integration of genomic, transcriptomic and proteomic data.

In Chapter 2, we introduced a revised version of our single cell model, SPARCED, featuring a modular and scalable pipeline for model construction and simulation. The Bouhaddou2018 model consists of 1197 species in total, which includes genes, mRNAs, lipids, proteins, post translationally modified proteins and protein complexes. Furthermore, there are more than 2400 reactions detailing their dynamic

interactions. The software pipeline for the construction and simulation of this model had a complex structure of dependencies and hardcoded scripts. The sheer magnitude of the model and lack of organization in its software pipeline introduced increasing levels of difficulties associated with modification of the model structure for future studies. In this chapter, the revised model construction and simulation pipeline streamlines the procedure for the modification of model structure. This enabled efficient expansion of the model with pharmacodynamics for a broader range of drugs which we discuss in the subsequent chapters.

In Chapter 3, we focus on improving the computational efficiency of the SPARCED model. The structure of the SPARCED model consists of two modules, one that captures gene expression and another that captures protein level interactions. It can be simulated in a hybrid stochastic/deterministic mode, where gene expression dynamics follow Poisson-like process, giving rise to gene expression noise and the resulting cell to cell variability in mRNAs and protein levels. It can also be simulated in a fully deterministic mode to demonstrate average cell behavior. Regardless of the mode of operation, computation speed was a major concern. Continuation of our work with the SPARCED model greatly relies on our ability to perform increasingly complex and resource-intensive computation, such as model initialization, parameter estimation and sensitivity analysis. Insufficient computation speed poses a challenge in this regard. In this chapter, we performed extensive performance benchmarking of our simulation algorithm which enabled us to improve inter-module communication for hybrid simulations, achieving at least a 4-fold increase in computation speed. Furthermore, we sought to eliminate the need for inter-module communication by integrating both

modules into a single system of differential equations which helped us achieve more than 200-fold increase in computation speed for deterministic simulations.

In Chapter 4, we describe the development of a mechanistic cell population simulation framework, and demonstrate its application in reconciliation of experimental and simulation dose response results. Dose response assays in general measure drug sensitivity or resistance by capturing cell population characteristics, such as viable cell counts at specific durations of treatment. In order to accomplish expansion and enhancement of single cell models based on experimental dose response data, a linkage needs to be established between dynamic interactions at the cellular pathway level and their emergent outcomes at the cell population level. Utilizing the functionality of the SPARCED model in describing biomolecular events at the single cell level, we developed a mechanistically informed cell population simulation framework. This framework combines detailed mechanistic descriptions of anti-cancer drug action with lineage-tracking based recording of individual division and death events to construct simulation outputs that are directly comparable to drug dose viability response assay. We simulated dose responses to multiple drugs, namely, Alpelisib (PI-3K inhibitor), Trametinib (MEK inhibitor), Palbociclib (CDK4/6 inhibitor) and Neratinib (EGFR inhibitor). The results show agreement with experimental results for strong growth inhibition by Trametinib and overall lack of efficacy for Alpelisib, but substantial discrepancy for Palbociclib and Neratinib. Deeper analyses investigating the reasons for these differences suggests that (i) contemporary belief in the importance of CDK4/6 for driving cell cycle completion is likely to be overestimated, and (ii) the cellular balance between basal (tonic) and ligand-induced ERK signaling is a critical determinant of

response to irreversible EGFR inhibitors. This work lays a foundation for mechanistic analysis of experimental drug dose viability response data sets.

In Chapter 5, we explore a strategy to implement omics-informed context definition in the SPARCED model. Cancer cell lines originating from a wide range of tumor types have been extensively characterized by high throughput omics technologies, such as genomics, transcriptomics and proteomics. The granularity of the SPARCED model enables integration of these omics datasets to define cell line specific contexts. However, inclusion of a new set of omics data requires that we adjust certain model parameters to ensure the integrity of the biological functionalities featured in the model. Previously, the initialization procedure was introduced for this purpose, with which the context of the original model was redefined to represent U87 cell lines. Initialization is a computationally intensive procedure whereby the model is subject to multi-step iterative unit testing with each step focused towards specific biological functionality, such as conservation of protein levels, basal cell cycle signaling, basal apoptosis signaling and basal DNA damage. The structural properties of the SPARCED model and its simulation algorithm imposed certain technical restrictions in this regard. For this reason, the previous initialization procedure was unable to tune parameters related to basal ERK and AKT pathways, transcriptional activation, survival signaling and replicative stress. This limitation challenged the adaptability of the SPARCED model across a wider range of cell line context available in large scale pharmacogenomic datasets, such as the Cancer Cell Line Encyclopedia (CCLE). However, this limitation was overcome after the improvement of computational efficiency as described in Chapter 3. This enabled us to revise the initialization procedure to develop a more

robust pipeline for initialization which can tune the previously excluded functionalities and was successfully applied to 59 cell lines from CCLE. Using the cell line specific variants of the SPARCED model and the cell population simulation framework developed in Chapter 4, we generated representations of dynamic cell populations of 28 of these cell lines which are consistent with their experimentally observed growth rates. We employed these dynamic cell population models to evaluate a strategy for the mechanistic exploration of their experimental drug sensitivity profiles.

In Chapter 6, we discuss the current state of the cell cycle submodel and its limitations. The cell cycle submodel describes the initiation of cell cycle due to growth stimulus by cyclin D upregulation, and subsequent oscillatory activation and inactivation of regulatory cell cycle proteins such as cyclins and cyclin dependent kinases which is the characteristic molecular signature of cell cycle. However, inclusion of gene expression noise within the species that regulate cell cycle resulted in unexpected and irregular behavior. It implies the existence of key regulatory mechanisms in the cell cycle pathway which provides natural robustness to the cell cycle process against gene expression noise. We discuss possible a solution by means of a revision of the cell cycle submodel consistent with more recently discovered regulatory mechanisms such as, E2F family of regulators, Rb protein group and Cip/Kip group of inhibitors. We developed a preliminary version of this revision which can describe cell cycle initiation, S phase entry and progression in presence of growth stimulus. Further work is needed to optimize the interaction between these species to ensure the description of cell cycle completion and natural gene expression noise suppression.

7.2 Future Directions

We envision single cell pharmacodynamic modeling as a promising approach towards developing predictive capabilities in cancer treatment efficacy. It may help us formalize the complexities observed in molecular pathophysiology with a view to understanding and predicting the extent to which these processes may affect treatment outcomes. The single-cell mechanistic model presented in this work aims to predict dynamic outcomes of signal transduction processes modulating essential cellular processes such as proliferation and apoptosis and the manner in which drug actions may perturb these phenomena. The results presented in this work show several promising avenues for future work.

1. Composition of a cell cycle pathway model capable of withstanding natural gene expression noise: The genome makes up the fundamental foundation of biological systems. Gene expression is an inherently stochastic process giving rise to cell-to-cell variability in mRNA and protein levels. Despite the noisy nature of gene expression, the highly conserved nature of biological pathways implies built-in regulatory mechanisms in their network modalities that can provide robustness. Our current model possess certain limitations in this regard, since inclusion of gene expression noise within the cell cycle submodel results in unpredictable, and irregular behavior. It implies that presence of key regulatory mechanisms in the cell cycle process yet to be included in the model. It is a significant challenge to the predictive capability of the model for drug actions targeting the cell cycle pathway and a challenge to the adaptive capability of the model to new cell line contexts. An important

future direction in this regard could be a revision of the cell cycle pathway submodel. To aid this, several essential regulatory mechanisms that have been discovered more recently and not included in the original model could be included. Such as a more detailed mechanism of transcriptional regulation by the E2F family of cell cycle regulators, Cip/kip group of inhibitors, pocket protein regulators. This revision should aim at being able to describe the characteristic functionalities of the cell cycle process without compromising the integrity of the gene expression processes.

2. Expansion of biological signaling pathways: The SPARCED single cell model currently encompasses several biological signaling pathways known to be implicated in oncogenic transformation. An overarching goal of our approach is to have predictive capability which is generalizable across tumor types. However, certain pathways with known alterations in various tumor types, not currently included in our model, may still impact treatment outcomes. Hence, enhancement of predictive capability and adaptability across tumor types is likely to require inclusion of additional pathways. In a recent study of tumor samples across 33 cancer types characterized by the Cancer Genome Atlas[15], somatic driver mutations were identified in ten canonical pathways, namely, cell cycle, Hippo, Myc, Notch, Nrf2, PI3-Kinase/AKT, RTK-RAS, TGFβ, p53 and β-catenin/wnt, five of which are currently not included in the SPARCED model. Furthermore, 89% of tumors had at least one driver alteration in these pathways with 57% of tumors having at lease one alteration potentially targetable by currently available drugs. Novel therapeutic

drugs targeting alterations in these pathways are also under investigation[213–219]. A viable strategy in this regard may be prioritizing the currently excluded pathway based on the occurrence of their alterations in cell lines in the Cancer Cell Line Encyclopedia. Published models of these individual pathways may be investigated for inclusion into the SPARCED model[220–223].

3. Integrating effects of mutations: A key concern about oncogenic transformation and disease progression is the dysregulation of biological pathway activities. Many such aberrations occur as a result of copy number variations, such as overexpression or copy number loss. Another important group of genomic aberrations is gain of function or loss of function mutations which are not directly captured by the expression data. Nevertheless, they may have functional impacts within the biomolecular networks, such as the KRAS, BRAF or EGFR mutations. A feasible approach for selecting mutations to include in the SPARCED model could include a deeper investigation of the cell line specific dose response results from Chapter 5 and prioritizing the mutation based on the frequency with which they occur in the cell lines with mismatched results. To formalize the effects of each mutation, species with representing mutated gene products could be included whose interaction rates are different from those of wild-type variants, depending on the nature of mutations. For example, if a mutation causes constitutive activation of a protein, the corresponding interactions could be changed such that the mutant variant does not rely on upstream signal.

4. Predicting the effects of drug combinations: The multivariate complexity of tumor progression often necessitates the use of drug combinations for effective treatment. However, rationalizing the choice of combinations for specific tumor types can be challenging since their synergy and antagonism cannot be easily predicted. Moreover, experimental screening of such combinations can be daunting because of the intractable number of combinations that may require consideration. However, reliably modeling effects of individual drug perturbation in biological networks could potentially explain the resulting synergy or antagonism when multiple drugs are combined. For this purpose, one could shortlist cell lines and drugs where our simulations can accurately capture sensitivity. Within individual context, deterministic simulations with combinations of varying doses of drug pairs can be generated and their effects on certain pathway level biomarkers such as ERK and AKT activities may be quantified. Thus, the resulting synergy and antagonism can be quantified and prioritized for combination dose response simulations as well as experimental validations.

In conclusion, this dissertation lays the groundwork for the use of computational models in building a bottom up understanding of the pathophysiology of cancer. Beginning with a single-cell model of signal transduction mechanisms driving phenotypic outcomes, we illustrate the potential for creating a predictive tool for anticancer drug response at the cellular level that may be generalized across tumor types. I believe endeavors such as this will help us build a deeper quantitative

understanding of the molecular intricacies of cancer which may one day help optimize treatment outcome in a patient specific manner.

References:

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. 2009;458(7239):719-724. doi:10.1038/nature07943

2. Stratton MR. Exploring the genomes of cancer cells: progress and promise. *Science*. 2011;331(6024):1553-1558. doi:10.1126/science.1204040

3. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209-249. doi:10.3322/caac.21660

4. DeVita VT Jr, Chu E. A History of Cancer Chemotherapy. *Cancer Research*. 2008;68(21):8643-8653. doi:10.1158/0008-5472.CAN-07-6611

5. Moses MA, Brem H, Langer R. Advancing the field of drug delivery: Taking aim at cancer. *Cancer Cell*. 2003;4(5):337-341. doi:10.1016/S1535-6108(03)00276-9

6. Shapira A, Livney YD, Broxterman HJ, Assaraf YG. Nanomedicine for targeted cancer therapy: Towards the overcoming of drug resistance. *Drug Resistance Updates*. 2011;14(3):150-163. doi:10.1016/j.drup.2011.01.003

7. Quintás-Cardama A, Kantarjian H, Cortes J. Imatinib and beyond—exploring the full potential of targeted therapy for CML. *Nat Rev Clin Oncol*. 2009;6(9):535-543. doi:10.1038/nrclinonc.2009.112

8. Motzer RJ, Michaelson MD, Rosenberg J, et al. Sunitinib Efficacy Against Advanced Renal Cell Carcinoma. *The Journal of Urology*. 2007;178(5):1883-1887. doi:10.1016/j.juro.2007.07.030

9. Valabrega G, Montemurro F, Aglietta M. Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. *Annals of Oncology*. 2007;18(6):977-984. doi:10.1093/annonc/mdl475

10. Meric-Bernstam F, Brusco L, Shaw K, et al. Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials. *J Clin Oncol*. 2015;33(25):2753-2762. doi:10.1200/JCO.2014.60.4165

11. Sabnis AJ, Bivona TG. Principles of resistance to targeted cancer therapy: lessons from basic and translational cancer biology. *Trends Mol Med*. 2019;25(3):185-197. doi:10.1016/j.molmed.2018.12.009

12. Li Q, Li Z, Luo T, Shi H. Targeting the PI3K/AKT/mTOR and RAF/MEK/ERK pathways for cancer therapy. *Mol Biomed*. 2022;3(1):47. doi:10.1186/s43556-022-00110-2

13. Wan X, Harkavy B, Shen N, Grohar P, Helman LJ. Rapamycin induces feedback activation of Akt signaling through an IGF-1R-dependent mechanism. *Oncogene*. 2007;26(13):1932-1940. doi:10.1038/sj.onc.1209990

14. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113-1120. doi:10.1038/ng.2764

15. Sanchez-Vega F, Mina M, Armenia J, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*. 2018;173(2):321-337.e10. doi:10.1016/j.cell.2018.03.035

16. Ganini C, Amelio I, Bertolo R, et al. Global mapping of cancers: The Cancer Genome Atlas and beyond. *Molecular Oncology*. 2021;15(11):2823-2840. doi:10.1002/1878-0261.13056

17. Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: predictions and reality. *Trends in Molecular Medicine*. 2023;29(7):554-566. doi:10.1016/j.molmed.2023.03.007

18. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*. 2018;15(2):81-94. doi:10.1038/nrclinonc.2017.166

19. Boshuizen J, Peeper DS. Rational Cancer Treatment Combinations: An Urgent Clinical Need. *Mol Cell*. 2020;78(6):1002-1018. doi:10.1016/j.molcel.2020.05.031

20. Unger JM, Cook E, Tai E, Bleyer A. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. *Am Soc Clin Oncol Educ Book*. 2016;(36):185-198. doi:10.1200/EDBK_156686

21. Roskoski R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2023 update. *Pharmacological Research*. 2023;187:106552. doi:10.1016/j.phrs.2022.106552

22. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*. 2013;501(7467):338-345. doi:10.1038/nature12625

23. Johnson BE, Mazor T, Hong C, et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science*. 2014;343(6167):189-193. doi:10.1126/science.1239947

24. Sottoriva A, Spiteri I, Piccirillo SGM, et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *PNAS*. 2013;110(10):4009-4014. doi:10.1073/pnas.1219747110

25. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning. *Nat Med*. 2018;24(10):1559-1567. doi:10.1038/s41591-018-0177-5

26. Birtwistle MR, Hatakeyama M, Yumoto N, Ogunnaike BA, Hoek JB, Kholodenko BN. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Mol Syst Biol*. 2007;3:144. doi:10.1038/msb4100188

27. Gerard C, Goldbeter A. Temporal self-organization of the cyclin/Cdk network driving the mammalian cell cycle. *Proceedings of the National Academy of Sciences*. 2009;106(51):21643-21648. doi:10.1073/pnas.0903827106

28. Batchelor E, Loewer A, Mock C, Lahav G. Stimulus-dependent dynamics of p53 in single cells. *Mol Syst Biol*. 2011;7:488. doi:10.1038/msb.2011.20

29. Albeck JG, Burke JM, Spencer SL, Lauffenburger DA, Sorger PK. Modeling a Snap-Action, Variable-Delay Switch Controlling Extrinsic Cell Death. *PLOS Biology*. 2008;6(12):e299. doi:10.1371/journal.pbio.0060299

30. Barrette AM, Bouhaddou M, Birtwistle MR. Integrating Transcriptomic Data with Mechanistic Systems Pharmacology Models for Virtual Drug Combination Trials. *ACS Chem Neurosci*. 2018;9(1):118-129. doi:10.1021/acschemneuro.7b00197

31. Zhang XY, Birtwistle MR, Gallo JM. A General Network Pharmacodynamic Model–Based Design Pipeline for Customized Cancer Therapy Applied to the VEGFR Pathway. *CPT Pharmacometrics Syst Pharmacol*. 2014;3(1):e92. doi:10.1038/psp.2013.65

32. Bouhaddou M, Barrette AM, Stern AD, et al. A mechanistic pan-cancer pathway model informed by multi-omics data interprets stochastic cell fate responses to drugs and mitogens. *PLOS Computational Biology*. 2018;14(3):e1005985. doi:10.1371/journal.pcbi.1005985

33. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603-607. doi:10.1038/nature11003

34. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. 2013;41(Database issue):D955-D961. doi:10.1093/nar/gks1111

35. Rees MG, Seashore-Ludlow B, Cheah JH, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol*. 2016;12(2):109-116. doi:10.1038/nchembio.1986

36. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):160018. doi:10.1038/sdata.2016.18

37. Ma'ayan A, Rouillard AD, Clark NR, Wang Z, Duan Q, Kou Y. Lean Big Data integration in systems biology and systems pharmacology. *Trends in Pharmacological Sciences*. 2014;35(9):450-460. doi:10.1016/j.tips.2014.07.001

38. Gomez-Cabrero D, Abugessaisa I, Maier D, et al. Data integration in the era of omics: current and future challenges. *BMC Systems Biology*. 2014;8(2):I1. doi:10.1186/1752-0509-8-S2-I1

39. Stites EC, Aziz M, Creamer MS, Von Hoff DD, Posner RG, Hlavacek WS. Use of Mechanistic Models to Integrate and Analyze Multiple Proteomic Datasets. *Biophysical Journal*. 2015;108(7):1819-1829. doi:10.1016/j.bpj.2015.02.030

40. Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes (Basel)*. 2019;10(2):87. doi:10.3390/genes10020087

41. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*. 2017;8. Accessed February 11, 2024. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2017.00084

42. Zeng ISL, Lumley T. Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science). *Bioinform Biol Insights*. 2018;12:1177932218759292. doi:10.1177/1177932218759292

43. Jensen KJ, Janes KA. Modeling the latent dimensions of multivariate signaling datasets. *Phys Biol*. 2012;9(4):045004. doi:10.1088/1478-3975/9/4/045004

44. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precis Onc*. 2020;4(1):1-10. doi:10.1038/s41698-020-0122-1

45. Ianevski A, Giri AK, Gautam P, et al. Prediction of drug combination effects with a minimal set of experiments. *Nat Mach Intell*. 2019;1(12):568-577. doi:10.1038/s42256-019-0122-4

46. Liu H, Zhang W, Nie L, Ding X, Luo J, Zou L. Predicting effective drug combinations using gradient tree boosting based on features extracted from drug-protein heterogeneous network. *BMC Bioinformatics*. 2019;20(1):645. doi:10.1186/s12859-019-3288-1

47. Wong D, Yip S. Machine learning classifies cancer. *Nature*. 2018;555(7697):446-447. doi:10.1038/d41586-018-02881-7

48. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

49. Kleppe A, Albregtsen F, Vlatkovic L, et al. Chromatin organisation and cancer prognosis: a pan-cancer study. *The Lancet Oncology*. 2018;19(3):356-369. doi:10.1016/S1470-2045(17)30899-9

50. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

51. Yu MK, Ma J, Fisher J, Kreisberg JF, Raphael BJ, Ideker T. Visible Machine Learning for Biomedicine. *Cell*. 2018;173(7):1562-1565. doi:10.1016/j.cell.2018.05.056

52. Baker RE, Peña JM, Jayamohan J, Jérusalem A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*. 2018;14(5):20170660. doi:10.1098/rsbl.2017.0660

53. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. *Journal of Big Data*. 2015;2(1):1. doi:10.1186/s40537-014-0007-7

54. Wang F, Casalino LP, Khullar D. Deep Learning in Medicine—Promise, Progress, and Challenges. *JAMA Internal Medicine*. 2019;179(3):293-294. doi:10.1001/jamainternmed.2018.7117

55. Yang JH, Wright SN, Hamblin M, et al. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. *Cell*. 2019;177(6):1649-1661.e9. doi:10.1016/j.cell.2019.04.016

56. Kholodenko BN, Hancock JF, Kolch W. Signalling ballet in space and time. *Nat Rev Mol Cell Biol*. 2010;11(6):414-426. doi:10.1038/nrm2901

57. Carrera J, Covert MW. Why Build Whole-Cell Models? *Trends Cell Biol*. 2015;25(12):719-722. doi:10.1016/j.tcb.2015.09.004

58. Karr JR, Sanghvi JC, Macklin DN, et al. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*. 2012;150(2):389-401. doi:10.1016/j.cell.2012.05.044

59. Carrera J, Elena SF, Jaramillo A. Computational design of genomic transcriptional networks with adaptation to varying environments. *Proc Natl Acad Sci U S A*. 2012;109(38):15277-15282. doi:10.1073/pnas.1200030109

60. Münzner U, Klipp E, Krantz M. A comprehensive, mechanistically detailed, and executable model of the cell division cycle in Saccharomyces cerevisiae. *Nat Commun*. 2019;10(1):1308. doi:10.1038/s41467-019-08903-w

61. Saez-Rodriguez J, Blüthgen N. Personalized signaling models for personalized treatments. *Mol Syst Biol*. 2020;16(1):e9042. doi:10.15252/msb.20199042

62. Halasz M, Kholodenko BN, Kolch W, Santra T. Integrating network reconstruction with mechanistic modeling to predict cancer therapies. *Sci Signal*. 2016;9(455):ra114. doi:10.1126/scisignal.aae0535

63. Macklin DN, Ahn-Horst TA, Choi H, et al. Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. *Science*. 2020;369(6502):eaav3751. doi:10.1126/science.aav3751

64. Santos SDM, Verveer PJ, Bastiaens PIH. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nat Cell Biol*. 2007;9(3):324-330. doi:10.1038/ncb1543

65. Kholodenko BN, Demin OV, Moehren G, Hoek JB. Quantification of short term signaling by the epidermal growth factor receptor. *J Biol Chem*. 1999;274(42):30169-30181. doi:10.1074/jbc.274.42.30169

66. Tyson JJ. Modeling the cell division cycle: cdc2 and cyclin interactions. *Proc Natl Acad Sci U S A*. 1991;88(16):7328-7332. doi:10.1073/pnas.88.16.7328

67. Nyman E, Fagerholm S, Jullesson D, Strålfors P, Cedersund G. Mechanistic explanations for counter-intuitive phosphorylation dynamics of the insulin receptor and insulin receptor substrate-1 in response to insulin in murine adipocytes. *FEBS J*. 2012;279(6):987-999. doi:10.1111/j.1742-4658.2012.08488.x

68. Schmierer B, Tournier AL, Bates PA, Hill CS. Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *Proceedings of the National Academy of Sciences*. 2008;105(18):6608-6613. doi:10.1073/pnas.0710134105

69. Vilar JMG, Guet CC, Leibler S. Modeling network dynamics. *J Cell Biol*. 2003;161(3):471-476. doi:10.1083/jcb.200301125

70. Kofahl B, Klipp E. Modelling the dynamics of the yeast pheromone pathway. *Yeast*. 2004;21(10):831-850. doi:10.1002/yea.1122

71. Tyson JJ, Chen K, Novak B. Network dynamics and cell physiology. *Nat Rev Mol Cell Biol*. 2001;2(12):908-916. doi:10.1038/35103078

72. Puszyński K, Hat B, Lipniacki T. Oscillations and bistability in the stochastic model of p53 regulation. *J Theor Biol*. 2008;254(2):452-465. doi:10.1016/j.jtbi.2008.05.039

73. Sedaghat AR, Sherman A, Quon MJ. A mathematical model of metabolic insulin signaling pathways. *Am J Physiol Endocrinol Metab*. 2002;283(5):E1084-1101. doi:10.1152/ajpendo.00571.2001

74. Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. *Mol Syst Biol*. 2014;10(7):735. doi:10.15252/msb.20145108

75. Fröhlich F, Kessler T, Weindl D, et al. Efficient Parameter Estimation Enables the Prediction of Drug Response Using a Mechanistic Pan-Cancer Pathway Model. *Cell Systems*. 2018;7(6):567-579.e6. doi:10.1016/j.cels.2018.10.013

76. Dalle Pezze P, Sonntag AG, Thien A, et al. A dynamic network model of mTOR signaling reveals TSC-independent mTORC2 regulation. *Sci Signal*. 2012;5(217):ra25. doi:10.1126/scisignal.2002469

77. Capuani F, Conte A, Argenzio E, et al. Quantitative analysis reveals how EGFR activation and downregulation are coupled in normal but not in cancer cells. *Nat Commun*. 2015;6(1):7999. doi:10.1038/ncomms8999

78. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol*. 2010;28(3):245-248. doi:10.1038/nbt.1614

79. Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of metabolomics. *Brief Bioinform*. 2006;7(2):140-150. doi:10.1093/bib/bbl007

80. Sherman MS, Cohen BA. A Computational Framework for Analyzing Stochasticity in Gene Expression. *PLOS Computational Biology*. 2014;10(5):e1003596. doi:10.1371/journal.pcbi.1003596

81. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology*. 2006;4(10):e309. doi:10.1371/journal.pbio.0040309

82. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*. 2008;135(2):216-226. doi:10.1016/j.cell.2008.09.050

83. Faeder JR, Blinov ML, Goldstein B, Hlavacek WS. Rule-based modeling of biochemical networks. *Complexity*. 2005;10(4):22-41. doi:10.1002/cplx.20074

84. Harris LA, Hogg JS, Tapia JJ, et al. BioNetGen 2.2: advances in rule-based modeling. *Bioinformatics*. 2016;32(21):3366-3368. doi:10.1093/bioinformatics/btw469

85. Xu W, Smith AM, Faeder JR, Marai GE. RuleBender: a visual interface for rule-based modeling. *Bioinformatics*. 2011;27(12):1721-1722. doi:10.1093/bioinformatics/btr197

86. Boutillier P, Maasha M, Li X, et al. The Kappa platform for rule-based modeling. *Bioinformatics*. 2018;34(13):i583-i592. doi:10.1093/bioinformatics/bty272

87. Lopez CF, Muhlich JL, Bachman JA, Sorger PK. Programming biological models in Python using PySB. *Mol Syst Biol*. 2013;9:646. doi:10.1038/msb.2013.1

88. Sneddon MW, Faeder JR, Emonet T. Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. *Nat Methods*. 2011;8(2):177-183. doi:10.1038/nmeth.1546

89. Hogg JS, Harris LA, Stover LJ, Nair NS, Faeder JR. Exact Hybrid Particle/Population Simulation of Rule-Based Models of Biochemical Systems. *PLOS Computational Biology*. 2014;10(4):e1003544. doi:10.1371/journal.pcbi.1003544

90. Goldberg AP, Szigeti B, Chew YH, Sekar JAP, Roth YD, Karr JR. Emerging whole-cell modeling principles and methods. *Curr Opin Biotechnol*. 2018;51:97-102. doi:10.1016/j.copbio.2017.12.013

91. Porubsky VL, Goldberg AP, Rampadarath AK, Nickerson DP, Karr JR, Sauro HM. Best Practices for Making Reproducible Biochemical Models. *Cell Syst*. 2020;11(2):109-120. doi:10.1016/j.cels.2020.06.012

92. Azeloglu EU, Iyengar R. Good practices for building dynamical models in systems biology. *Sci Signal*. 2015;8(371):fs8. doi:10.1126/scisignal.aab0880

93. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*. 2003;19(4):524-531. doi:10.1093/bioinformatics/btg015

94. Keating SM, Waltemath D, König M, et al. SBML Level 3: an extensible format for the exchange and reuse of biological models. *Mol Syst Biol*. 2020;16(8):e9110. doi:10.15252/msb.20199110

95. Hoops S, Sahle S, Gauges R, et al. COPASI—a COmplex PAthway SImulator. *Bioinformatics*. 2006;22(24):3067-3074. doi:10.1093/bioinformatics/btl485

96. Loew LM, Schaff JC. The Virtual Cell: a software environment for computational cell biology. *Trends in Biotechnology*. 2001;19(10):401-406. doi:10.1016/S0167-7799(01)01740-1

97. Smith LP, Bergmann FT, Chandran D, Sauro HM. Antimony: a modular model definition language. *Bioinformatics*. 2009;25(18):2452-2454. doi:10.1093/bioinformatics/btp401

98. Rensin DK. *Kubernetes - Scheduling the Future at Cloud Scale*.; 2015. http://www.oreilly.com/webops-perf/free/kubernetes.csp

99. Thurgood B, Lennon RG. Cloud Computing With Kubernetes Cluster Elastic Scaling. In: *Proceedings of the 3rd International Conference on Future Networks and Distributed Systems*. ICFNDS '19. Association for Computing Machinery; 2019:1-7. doi:10.1145/3341325.3341995

100. Smarr L, Crittenden C, DeFanti T, et al. The Pacific Research Platform: Making High-Speed Networking a Reality for the Scientist. In: ; 2018:1-8. doi:10.1145/3219104.3219108

101. Kluyver T, Ragan-Kelley B, P&#233, et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press; 2016:87-90. doi:10.3233/978-1-61499-649-1-87

102. Fröhlich F, Kaltenbacher B, Theis FJ, Hasenauer J. Scalable Parameter Estimation for Genome-Scale Biochemical Reaction Networks. *PLOS Computational Biology*. 2017;13(1):e1005331. doi:10.1371/journal.pcbi.1005331

103. Fröhlich F, Theis FJ, Rädler JO, Hasenauer J. Parameter estimation for dynamical systems with discrete events and logical operations. *Bioinformatics*. 2017;33(7):1049-1056. doi:10.1093/bioinformatics/btw764

104. Nakakuki T, Birtwistle MR, Saeki Y, et al. Ligand-specific c-Fos expression emerges from the spatiotemporal control of ErbB network dynamics. *Cell*. 2010;141(5):884-896. doi:10.1016/j.cell.2010.03.054

105. von Kriegsheim A, Baiocchi D, Birtwistle M, et al. Cell fate decisions are specified by the dynamic ERK interactome. *Nat Cell Biol*. 2009;11(12):1458-1464. doi:10.1038/ncb1994

106. Fröhlich F, Weindl D, Schälte Y, et al. AMICI: high-performance sensitivity analysis for large ordinary differential equation models. *Bioinformatics*. 2021;37(20):3676-3677. doi:10.1093/bioinformatics/btab227

107. Crudu A, Debussche A, Radulescu O. Hybrid stochastic simplifications for multiscale gene networks. *BMC Systems Biology*. 2009;3(1):89. doi:10.1186/1752-0509-3-89

108. Gibson MA, Bruck J. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J Phys Chem A*. 2000;104(9):1876-1889. doi:10.1021/jp993732q

109. Yeom JS, Georgouli K, Blake R, Navid A. Towards dynamic simulation of a whole cell model. In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '21. Association for Computing Machinery; 2021:1-10. doi:10.1145/3459930.3471161

110. Neal ML, Gennari JH, Waltemath D, Nickerson DP, König M. Open modeling and exchange (OMEX) metadata specification version 1.0. *J Integr Bioinform*. 2020;17(2-3):20200020. doi:10.1515/jib-2020-0020

111. Krause F, Uhlendorf J, Lubitz T, Schulz M, Klipp E, Liebermeister W. Annotation and merging of SBML models with semanticSBML. *Bioinformatics*. 2010;26(3):421-422. doi:10.1093/bioinformatics/btp642

112. Neal ML, Thompson CT, Kim KG, et al. SemGen: a tool for semantics-based annotation and composition of biosimulation models. *Bioinformatics*. 2019;35(9):1600-1602. doi:10.1093/bioinformatics/bty829

113. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013

114. Ghaffarizadeh A, Heiland R, Friedman SH, Mumenthaler SM, Macklin P. PhysiCell: An open source physics-based cell simulator for 3-D multicellular systems. *PLOS Computational Biology*. 2018;14(2):e1005991. doi:10.1371/journal.pcbi.1005991

115. Swat MH, Thomas GL, Belmonte JM, Shirinifard A, Hmeljak D, Glazier JA. Multi-Scale Modeling of Tissues Using CompuCell3D. *Methods Cell Biol*. 2012;110:325-366. doi:10.1016/B978-0-12-388403-9.00013-8

116. Szigeti B, Roth YD, Sekar JAP, Goldberg AP, Pochiraju SC, Karr JR. A blueprint for human whole-cell modeling. *Curr Opin Syst Biol*. 2018;7:8-15. doi:10.1016/j.coisb.2017.10.005

117. Hindmarsh AC, Brown PN, Grant KE, et al. SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Trans Math Softw*. 2005;31(3):363-396. doi:10.1145/1089014.1089020

118. Yates AD, Achuthan P, Akanni W, et al. Ensembl 2020. *Nucleic Acids Research*. 2020;48(D1):D682-D688. doi:10.1093/nar/gkz966

119. Braschi B, Denny P, Gray K, et al. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res*. 2019;47(D1):D786-D792. doi:10.1093/nar/gky930

120. Ahn-Horst TA, Mille LS, Sun G, Morrison JH, Covert MW. An expanded whole-cell model of E. coli links cellular physiology with mechanisms of growth rate control. *npj Syst Biol Appl*. 2022;8(1):1-21. doi:10.1038/s41540-022-00242-9

121. Mardinoglu A, Bjornson E, Zhang C, et al. Personal model-assisted identification of NAD+ and glutathione metabolism as intervention target in NAFLD. *Mol Syst Biol*. 2017;13(3):916. doi:10.15252/msb.20167422

122. Spann R, Roca C, Kold D, Eliasson Lantz A, Gernaey KV, Sin G. A probabilistic model-based soft sensor to monitor lactic acid bacteria fermentations. *Biochemical Engineering Journal*. 2018;135:49-60. doi:10.1016/j.bej.2018.03.016

123. Uhlen M, Zhang C, Lee S, et al. A pathology atlas of the human cancer transcriptome. *Science*. 2017;357(6352):eaan2507. doi:10.1126/science.aan2507

124. Purcell O, Jain B, Karr JR, Covert MW, Lu TK. Towards a whole-cell modeling approach for synthetic biology. *Chaos*. 2013;23(2):025112. doi:10.1063/1.4811182

125. Macklin DN, Ruggero NA, Covert MW. The future of whole-cell modeling. *Current Opinion in Biotechnology*. 2014;28:111-115. doi:10.1016/j.copbio.2014.01.012

126. Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences*. 2005;102(8):2685-2689. doi:10.1073/pnas.0406811102

127. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*. 2010;5(1):93-121. doi:10.1038/nprot.2009.203

128. Thornburg ZR, Bianchi DM, Brier TA, et al. Fundamental behaviors emerge from simulations of a living minimal cell. *Cell*. 2022;185(2):345-360.e28. doi:10.1016/j.cell.2021.12.025

129. Faeder JR, Blinov ML, Hlavacek WS. Rule-based modeling of biochemical systems with BioNetGen. In: *Systems Biology*. Springer; 2009:113-167.

130. Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK. From word models to executable models of signaling networks using automated assembly. *Mol Syst Biol*. 2017;13(11):954. doi:10.15252/msb.20177651

131. Somogyi ET, Bouteiller JM, Glazier JA, et al. libRoadRunner: a high performance SBML simulation and analysis library. *Bioinformatics*. 2015;31(20):3315-3321. doi:10.1093/bioinformatics/btv363

132. Stapor P, Fröhlich F, Hasenauer J. Optimization and profile calculation of ODE models using second order adjoint sensitivity analysis. *Bioinformatics*. 2018;34(13):i151-i159. doi:10.1093/bioinformatics/bty230

133. Schmiester L, Schälte Y, Bergmann FT, et al. PEtab—Interoperable specification of parameter estimation problems in systems biology. *PLOS Computational Biology*. 2021;17(1):e1008646. doi:10.1371/journal.pcbi.1008646

134. Roth YD, Lian Z, Pochiraju S, Shaikh B, Karr JR. Datanator: an integrated database of molecular data for quantitatively modeling cellular behavior. *Nucleic Acids Research*. 2021;49(D1):D516-D522. doi:10.1093/nar/gkaa1008

135. Lang PF, Chebaro Y, Zheng X, et al. BpForms and BcForms: a toolkit for concretely describing non-canonical polymers and complexes to facilitate global biochemical networks. *Genome Biology*. 2020;21(1):117. doi:10.1186/s13059-020-02025-z

136. Smith LP, Hucka M, Hoops S, et al. SBML Level 3 package: Hierarchical Model Composition, Version 1 Release 3. *Journal of Integrative Bioinformatics*. 2015;12(2):603-659. doi:10.1515/jib-2015-268

137. Erdem C, Bensman EM, Mutsuddy A, et al. A Simple and Efficient Pipeline for Construction, Merging, Expansion, and Simulation of Large-Scale, Single-Cell Mechanistic Models. *bioRxiv*. Published online November 10, 2020:2020.11.09.373407. doi:10.1101/2020.11.09.373407

138. Hughey JJ, Lee TK, Covert MW. Computational modeling of mammalian signaling networks. *WIREs Systems Biology and Medicine*. 2010;2(2):194-209. doi:10.1002/wsbm.52

139. Singhania R, Sramkoski RM, Jacobberger JW, Tyson JJ. A Hybrid Model of Mammalian Cell Cycle Regulation. *PLOS Computational Biology*. 2011;7(2):e1001077. doi:10.1371/journal.pcbi.1001077

140. Herz AVM, Gollisch T, Machens CK, Jaeger D. Modeling Single-Neuron Dynamics and Computations: A Balance of Detail and Abstraction. *Science*. 2006;314(5796):80-85. doi:10.1126/science.1127240

141. Bachmann J, Raue A, Schilling M, et al. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol Syst Biol*. 2011;7:516. doi:10.1038/msb.2011.50

142. Schoeberl B, Pace EA, Fitzgerald JB, et al. Therapeutically Targeting ErbB3: A Key Node in Ligand-Induced Activation of the ErbB Receptor–PI3K Axis. *Science Signaling*. 2009;2(77):ra31-ra31. doi:10.1126/scisignal.2000352

143. Eduati F, Doldàn-Martelli V, Klinger B, et al. Drug Resistance Mechanisms in Colorectal Cancer Dissected with Cell Type-Specific Dynamic Logic Models. *Cancer Res*. 2017;77(12):3364-3375. doi:10.1158/0008-5472.CAN-17-0078

144. Fey D, Halasz M, Dreidax D, et al. Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Science Signaling*. 2015;8(408):ra130-ra130. doi:10.1126/scisignal.aab0990

145. Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res*. 2002;12(1):9-18. doi:10.1038/sj.cr.7290105

146. Chen WW, Schoeberl B, Jasper PJ, et al. Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular Systems Biology*. 2009;5(1):239. doi:10.1038/msb.2008.74

147. Altan-Bonnet G, Germain RN. Modeling T Cell Antigen Discrimination Based on Feedback Control of Digital ERK Responses. *PLoS Biol*. 2005;3(11):e356. doi:10.1371/journal.pbio.0030356

148. Hoffmann A, Levchenko A, Scott ML, Baltimore D. The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science*. 2002;298(5596):1241-1245. doi:10.1126/science.1071914

149. Lee E, Salic A, Krüger R, Heinrich R, Kirschner MW. The Roles of APC and Axin Derived from Experimental and Theoretical Analysis of the Wnt Pathway. *PLoS Biol*. 2003;1(1):e10. doi:10.1371/journal.pbio.0000010

150. Park CS, Schneider IC, Haugh JM. Kinetic analysis of platelet-derived growth factor receptor/phosphoinositide 3-kinase/Akt signaling in fibroblasts. *J Biol Chem*. 2003;278(39):37064-37072. doi:10.1074/jbc.M304968200

151. Swainston N, Smallbone K, Hefzi H, et al. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*. 2016;12:109. doi:10.1007/s11306-016-1051-4

152. Hass H, Masson K, Wohlgemuth S, et al. Predicting ligand-dependent tumors from multi-dimensional signaling features. *NPJ Syst Biol Appl*. 2017;3:27. doi:10.1038/s41540-017-0030-3

153. Bajikar SS, Janes KA. Multiscale Models of Cell Signaling. *Ann Biomed Eng*. 2012;40(11):2319-2327. doi:10.1007/s10439-012-0560-1

154. Qutub AA, Popel AS. Elongation, proliferation & migration differentiate endothelial cell phenotypes and determine capillary sprouting. *BMC Syst Biol*. 2009;3:13. doi:10.1186/1752-0509-3-13

155. Ruscone M, Montagud A, Chavrier P, et al. Multiscale model of the different modes of cancer cell invasion. *Bioinformatics*. 2023;39(6):btad374. doi:10.1093/bioinformatics/btad374

156. Groß A, Kracher B, Kraus JM, et al. Representing dynamic biological networks with multi-scale probabilistic models. *Commun Biol*. 2019;2(1):1-12. doi:10.1038/s42003-018-0268-3

157. Wertheim KY, Puniya BL, Fleur AL, Shah AR, Barberis M, Helikar T. A multi-approach and multi-scale platform to model CD4+ T cells responding to infections. *PLOS Computational Biology*. 2021;17(8):e1009209. doi:10.1371/journal.pcbi.1009209

158. Aghamiri SS, Puniya BL, Amin R, Helikar T. A multiscale mechanistic model of human dendritic cells for in-silico investigation of immune responses and novel therapeutics discovery. *Frontiers in Immunology*. 2023;14. Accessed November 14, 2023. https://www.frontiersin.org/articles/10.3389/fimmu.2023.1112985

159. Tripathi S, Park JH, Pudakalakatti S, Bhattacharya PK, Kaipparettu BA, Levine H. A mechanistic modeling framework reveals the key principles underlying tumor

metabolism. *PLOS Computational Biology*. 2022;18(2):e1009841. doi:10.1371/journal.pcbi.1009841

160. Hormuth DA, Jarrett AM, Lima EABF, McKenna MT, Fuentes DT, Yankeelov TE. Mechanism-Based Modeling of Tumor Growth and Treatment Response Constrained by Multiparametric Imaging Data. *JCO Clinical Cancer Informatics*. 2019;(3):1-10. doi:10.1200/CCI.18.00055

161. Azer K, Kaddi CD, Barrett JS, et al. History and Future Perspectives on the Discipline of Quantitative Systems Pharmacology Modeling and Its Applications. *Frontiers in Physiology*. 2021;12. Accessed November 14, 2023. https://www.frontiersin.org/articles/10.3389/fphys.2021.637999

162. Aghamiri SS, Amin R, Helikar T. Recent applications of quantitative systems pharmacology and machine learning models across diseases. *J Pharmacokinet Pharmacodyn*. 2022;49(1):19-37. doi:10.1007/s10928-021-09790-9

163. Chung D, Bakshi S, van der Graaf PH. A Review of Quantitative Systems Pharmacology Models of the Coagulation Cascade: Opportunities for Improved Usability. *Pharmaceutics*. 2023;15(3):918. doi:10.3390/pharmaceutics15030918

164. Gonçalves E, Bucher J, Ryll A, et al. Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. *Molecular BioSystems*. 2013;9(7):1576-1583. doi:10.1039/C3MB25489E

165. Malik-Sheriff RS, Glont M, Nguyen TVN, et al. BioModels—15 years of sharing computational models in life science. *Nucleic Acids Research*. 2020;48(D1):D407-D415. doi:10.1093/nar/gkz1055

166. Heiser LM, Sadanandam A, Kuo WL, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci U S A*. 2012;109(8):2724-2729. doi:10.1073/pnas.1018854108

167. Schenone M, Dančík V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol*. 2013;9(4):232-240. doi:10.1038/nchembio.1199

168. Wang D, Hensman J, Kutkaite G, et al. A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates. *eLife*. 9:e60352. doi:10.7554/eLife.60352

169. Jackson R, Bayrak ES, Wang T, Coufal M, Undey C, Cinar A. High Performance Agent-Based Modeling to Simulate Mammalian Cell Culture Bioreactor. In: Eden MR, Ierapetritou MG, Towler GP, eds. *Computer Aided Chemical Engineering*. Vol 44. 13 International Symposium on Process Systems Engineering (PSE 2018). Elsevier; 2018:1453-1458. doi:10.1016/B978-0-444-64241-7.50237-8

170. Yu JS, Bagheri N. Agent-Based Models Predict Emergent Behavior of Heterogeneous Cell Populations in Dynamic Microenvironments. *Frontiers in Bioengineering and Biotechnology*. 2020;8. Accessed October 3, 2023. https://www.frontiersin.org/articles/10.3389/fbioe.2020.00249

171. Gregg RW, Shabnam F, Shoemaker JE. Agent-based modeling reveals benefits of heterogeneous and stochastic cell populations during cGAS-mediated IFNβ production. *Bioinformatics*. 2021;37(10):1428-1434. doi:10.1093/bioinformatics/btaa969

172. Bonabeau E. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*. 2002;99(suppl_3):7280-7287. doi:10.1073/pnas.082080899

173. Gonzalez-de-Aledo P, Vladimirov A, Manca M, et al. An optimization approach for agent-based computational models of biological development. *Advances in Engineering Software*. 2018;121:262-275. doi:10.1016/j.advengsoft.2018.03.010

174. Erdem C, Mutsuddy A, Bensman EM, et al. A scalable, open-source implementation of a large-scale mechanistic model for single cell proliferation and death signaling. *Nat Commun*. 2022;13(1):3555. doi:10.1038/s41467-022-31138-1

175. Single A, Beetham H, Telford BJ, Guilford P, Chen A. A Comparison of Real-Time and Endpoint Cell Viability Assays for Improved Synthetic Lethal Drug Validation. *J Biomol Screen*. 2015;20(10):1286-1293. doi:10.1177/1087057115605765

176. Bessette DC, Tilch E, Seidens T, et al. Using the MCF10A/MCF10CA1a Breast Cancer Progression Cell Line Model to Investigate the Effect of Active, Mutant Forms of EGFR in Breast Cancer Development and Treatment Using Gefitinib. *PLOS ONE*. 2015;10(5):e0125232. doi:10.1371/journal.pone.0125232

177. Warleta F, Campos M, Allouche Y, et al. Squalene protects against oxidative DNA damage in MCF10A human mammary epithelial cells but not in MCF7 and MDA-MB-231 human breast cancer cells. *Food and Chemical Toxicology*. 2010;48(4):1092-1100. doi:10.1016/j.fct.2010.01.031

178. Martins MM, Zhou AY, Corella A, et al. Linking tumor mutations to drug responses via a quantitative chemical-genetic interaction map. *Cancer Discov*. 2015;5(2):154-167. doi:10.1158/2159-8290.CD-14-0552

179. Niepel M, Hafner M, Mills CE, et al. A Multi-center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. *Cell Systems*. 2019;9(1):35-48.e5. doi:10.1016/j.cels.2019.06.005

180. Fritsch C, Huang A, Chatenay-Rivauday C, et al. Characterization of the Novel and Specific PI3Kα Inhibitor NVP-BYL719 and Development of the Patient Stratification Strategy for Clinical Trials. *Molecular Cancer Therapeutics*. 2014;13(5):1117-1129. doi:10.1158/1535-7163.MCT-13-0865

181. Kim S, Tiedt R, Loo A, et al. The potent and selective cyclin-dependent kinases 4 and 6 inhibitor ribociclib (LEE011) is a versatile combination partner in preclinical cancer models. *Oncotarget*. 2018;9(81):35226-35240. doi:10.18632/oncotarget.26215

182. Yoshida T, Kakegawa J, Yamaguchi T, et al. Identification and Characterization of a Novel Chemotype MEK Inhibitor Able to Alter the Phosphorylation State of MEK1/2. *Oncotarget*. 2012;3(12):1533-1545.

183. PubChem. Neratinib. Accessed October 27, 2023. https://pubchem.ncbi.nlm.nih.gov/compound/9915743

184. Hafner M, Niepel M, Chung M, Sorger PK. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat Methods*. 2016;13(6):521-527. doi:10.1038/nmeth.3853

185. Takeuchi K, Ito F. EGF receptor in relation to tumor development: molecular basis of responsiveness of cancer cells to EGFR-targeting tyrosine kinase inhibitors. *The FEBS Journal*. 2010;277(2):316-326. doi:10.1111/j.1742-4658.2009.07450.x

186. Seshacharyulu P, Ponnusamy MP, Haridas D, Jain M, Ganti AK, Batra SK. Targeting the EGFR signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets*. 2012;16(1):15-31. doi:10.1517/14728222.2011.648617

187. Wee P, Wang Z. Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers*. 2017;9(5):52. doi:10.3390/cancers9050052

188. Chou J, Fan Z, DeBlasio T, Koff A, Rosen N, Mendelsohn J. Constitutive overexpression of cyclin D1 in human breast epithelial cells does not prevent G1 arrest induced by deprivation of epidermal growth factor. *Breast Cancer Res Treat*. 1999;55(3):267-283. doi:10.1023/A:1006217413089

189. Scaling AL, Prossnitz ER, Hathaway HJ. GPER Mediates Estrogen-Induced Signaling and Proliferation in Human Breast Epithelial Cells and Normal and Malignant Breast. *HORM CANC*. 2014;5(3):146-160. doi:10.1007/s12672-014-0174-1

190. Gabriel E, Fagg GE, Bosilca G, et al. Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation. In: Kranzlmüller D, Kacsuk P, Dongarra J, eds. *Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Lecture Notes in Computer Science. Springer; 2004:97-104. doi:10.1007/978-3-540-30218-6_19

191. Li Y, Barbash O, Diehl JA. Regulation of the Cell Cycle. In: *The Molecular Basis of Cancer*. Elsevier; 2015:165-178.e2. doi:10.1016/B978-1-4557-4066-6.00011-1

192. Schmidt M, Sebastian M. Palbociclib—The First of a New Class of Cell Cycle Inhibitors. In: Martens UM, ed. *Small Molecules in Oncology*. Recent Results in

Cancer Research. Springer International Publishing; 2018:153-175. doi:10.1007/978-3-319-91442-8_11

193. Lukas J, Parry D, Aagaard L, et al. Retinoblastoma-protein-dependent cell-cycle inhibition by the tumour suppressor p16. *Nature*. 1995;375(6531):503-506. doi:10.1038/375503a0

194. Finn RS, Dering J, Conklin D, et al. PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. *Breast Cancer Res*. 2009;11(5):R77. doi:10.1186/bcr2419

195. Knudsen KE, Weber E, Arden KC, Cavenee WK, Feramisco JR, Knudsen ES. The retinoblastoma tumor suppressor inhibits cellular proliferation through two distinct mechanisms: inhibition of cell cycle progression and induction of cell death. *Oncogene*. 1999;18(37):5239-5245. doi:10.1038/sj.onc.1202910

196. Caldon CE, Sergio CM, Kang J, et al. Cyclin E2 overexpression is associated with endocrine resistance but not insensitivity to CDK2 inhibition in human breast cancer cells. *Mol Cancer Ther*. 2012;11(7):1488-1499. doi:10.1158/1535-7163.MCT-11-0963

197. Mutsuddy A, Erdem C, Huggins JR, et al. Computational speed-up of large-scale, single-cell model simulations via a fully integrated SBML-based format. *Bioinformatics Advances*. 2023;3(1):vbad039. doi:10.1093/bioadv/vbad039

198. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*. 2012;28(24):3290-3297. doi:10.1093/bioinformatics/bts595

199. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics*. 2013;29(20):2610-2616. doi:10.1093/bioinformatics/btt425

200. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26(12):i237-i245. doi:10.1093/bioinformatics/btq182

201. Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics*. 2015;16:1022. doi:10.1186/s12864-015-2223-8

202. Yuan Y, Savage RS, Markowetz F. Patient-Specific Data Fusion Defines Prognostic Cancer Subtypes. *PLoS Comput Biol*. 2011;7(10):e1002227. doi:10.1371/journal.pcbi.1002227

203. Nguyen H, Shrestha S, Draghici S, Nguyen T. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*. 2019;35(16):2843-2846. doi:10.1093/bioinformatics/bty1049

204. Nusinow DP, Szpyt J, Ghandi M, et al. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell*. 2020;180(2):387-402.e16. doi:10.1016/j.cell.2019.12.023

205. Lee DY, Lee SY, Yun SH, et al. Review of the Current Research on Fetal Bovine Serum and the Development of Cultured Meat. *Food Sci Anim Resour*. 2022;42(5):775-799. doi:10.5851/kosfa.2022.e46

206. Xie Z, Kropiwnicki E, Wojciechowicz ML, et al. Getting Started with LINCS Datasets and Tools. *Current Protocols*. 2022;2(7):e487. doi:10.1002/cpz1.487

207. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6. doi:10.1186/s12864-019-6413-7

208. Suski JM, Braun M, Strmiska V, Sicinski P. Targeting Cell-cycle Machinery in Cancer. *Cancer Cell*. 2021;39(6):759-778. doi:10.1016/j.ccell.2021.03.010

209. Kent LN, Leone G. The broken cycle: E2F dysfunction in cancer. *Nat Rev Cancer*. 2019;19(6):326-338. doi:10.1038/s41568-019-0143-7

210. Chen HZ, Tsai SY, Leone G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat Rev Cancer*. 2009;9(11):785-797. doi:10.1038/nrc2696

211. Besson A, Dowdy SF, Roberts JM. CDK Inhibitors: Cell Cycle Regulators and Beyond. *Developmental Cell*. 2008;14(2):159-169. doi:10.1016/j.devcel.2008.01.013

212. Farkas T, Hansen K, Holm K, Lukas J, Bartek J. Distinct Phosphorylation Events Regulate p130- and p107-mediated Repression of E2F-4. *Journal of Biological Chemistry*. 2002;277(30):26741-26752. doi:10.1074/jbc.M200381200

213. Whitfield JR, Beaulieu ME, Soucek L. Strategies to Inhibit Myc and Their Clinical Applicability. *Frontiers in Cell and Developmental Biology*. 2017;5. Accessed March 4, 2024. https://www.frontiersin.org/articles/10.3389/fcell.2017.00010

214. Park HW, Guan KL. Regulation of the Hippo pathway and implications for anticancer drug development. *Trends in Pharmacological Sciences*. 2013;34(10):581-589. doi:10.1016/j.tips.2013.08.006

215. Aster JC, Blacklow SC. Targeting the Notch Pathway: Twists and Turns on the Road to Rational Therapeutics. *JCO*. 2012;30(19):2418-2420. doi:10.1200/JCO.2012.42.0992

216. Takebe N, Nguyen D, Yang SX. Targeting Notch signaling pathway in cancer: Clinical development advances and challenges. *Pharmacology & Therapeutics*. 2014;141(2):140-149. doi:10.1016/j.pharmthera.2013.09.005

217. Buijs JT, Stayrook KR, Guise TA. The role of TGF-β in bone metastasis: novel therapeutic perspectives. *Bonekey Rep*. 2012;1:96. doi:10.1038/bonekey.2012.96

218. Sheen YY, Kim MJ, Park SA, Park SY, Nam JS. Targeting the Transforming Growth Factor-β Signaling in Cancer Therapy. *Biomol Ther (Seoul)*. 2013;21(5):323-331. doi:10.4062/biomolther.2013.072

219. Pai SG, Carneiro BA, Mota JM, et al. Wnt/beta-catenin pathway: modulating anticancer immune response. *J Hematol Oncol*. 2017;10(1):101. doi:10.1186/s13045-017-0471-6

220. Cellière G, Fengos G, Hervé M, Iber D. plasticity of TGF-β signaling. *BMC Syst Biol*. 2011;5:184. doi:10.1186/1752-0509-5-184

221. Shin SY, Nguyen LK. Unveiling Hidden Dynamics of Hippo Signalling: A Systems Analysis. *Genes (Basel)*. 2016;7(8):44. doi:10.3390/genes7080044

222. Padala RR, Karnawat R, Viswanathan SB, Thakkar AV, Das AB. Cancerous perturbations within the ERK, PI3K/Akt, and Wnt/β-catenin signaling network constitutively activate inter-pathway positive feedback loops. *Mol BioSyst*. 2017;13(5):830-840. doi:10.1039/C6MB00786D

223. Sivakumar KC, Dhanesh SB, Shobana S, James J, Mundayoor S. A Systems Biology Approach to Model Neural Stem Cell Regulation by Notch, Shh, Wnt, and EGF Signaling Pathways. *OMICS: A Journal of Integrative Biology*. 2011;15(10):729-737. doi:10.1089/omi.2011.0011