Life Sciences Use Case: Initial Requirements and Scenario Definitions Work Package 1 Task 1.1-1.2 Deliverable 1.1

Authors

Cirillo, Davide – BSC
Montagud, Arnau – BSC
Ponce de León, Miguel – BSC
Valencia, Alfonso – BSC
Monti, Michele – CRG
Tartaglia, Gian Gaetano – CRG

Distribution list:

Group:	Others:
WP Leader: BSC Task Leader: BSC	Internal Reviewer: NCSR Demokritos (NCSR) INFORE Management Team INFORE Project Officer

Document history:

Revision	Date	Section	Page	Modification
0.5	28/02/2019	All	All	Creation.
0.6	12/03/2019	All	All	Revisions from ATHENA incorporated.
0.7	15/03/2019	-	-	Internal Review Comments Received
0.8	23/03/2019	All	All	Version after Internal Review
1.0	26/03/2019	All	All	Revisions from NCSR incorporated. Sections expanded.

Approvals:

First Author: Davide Cirillo (BSC) Date: 28/02/2019

Internal Reviewer: Elias Alevizos (NCSR) Date: 15/03/2019

Coordinator: Antonios Deligiannakis (Athena) Date: 26/03/2019

Table of contents:

1 Executive Summary

This deliverable describes the requirements and scenario definitions of the biological use case in accordance with expert users' questionnaires. By using a Boolean model of carcinoma cell lines (AGS) and a lattice-free physics-based cell simulator for 3D multicellular modelling, PhysiBoSS, we aim to replicate experimental data of growth profiles of cancer cells treated with different drug regimes. The ultimate goal of this use case is to provide a *"virtual laboratory"* for studying cancer growth and evolution by using multi-scale models of tumour systems. The development of such a framework will facilitate the design, test, and optimization of cancer treatments based on combinations of different drugs and dose scheduling strategies.

In our simulations, we will explore two alternative geometries for cell arrangement: i) one-cell-thick 2D monolayers, representing cells growing in a plate such as those observed in *in-vitro* cell cultures; and ii) 3D spheroids, representing cells growing in three-dimensional space, an arrangement that resembles cancer cells growing *in-vivo*. These simulations will be scaled up by several orders of magnitude using the Barcelona Supercomputing Center (BSC) MareNostrum4 allowing the incorporation of forecasting techniques for various events of interest. Moreover, interactive learning techniques will be used to assist the calibration of such *in-silico* models.

In particular, we aim to scale up simulations of cancer cell 3D spheroids from 5,000 up to 500,000 cells using high-performance computing. This scenario will allow the design of different set-ups that tally cancer tumour growth conditions with increased number of cells, altered microenvironmental physical properties, different cell types, as well as to study the interaction between cancer cells and the immune system. In addition, this will allow the consideration of different simulations' set-ups beyond cancer tumour growth such as *in-vitro* streak experiments, microfluidic designs, among others.

2 Introduction

Critical to the development of new cancer treatments is understanding the mechanisms underlying the emergence of cells resistant to different drug treatments, as well as discovering synergistic combinations of drugs that reduce the chance of this event. Disentangling the **emergence of drug resistance** is challenged by the inherent complexity of biological systems, which are characterized by a deep interplay among processes that occur at different scales [21]. Indeed, the emergence of drug resistance is determined by complex molecular mechanisms that ultimately lead to a reduction of the effectiveness of a particular drug, as well as various types of dynamic processes concerning large populations of cells [7]. Consequently, multicellular systems' dynamics such as tumour growth and evolution can only be understood by studying how the individual cells grow, divide and die, and their interactions at the population level [2]. **The *in-silico* simulation of tumour growth and evolution requires modelling multi-scale processes**, bridging the gap between different levels of description, and connecting events that occur at different scales [1]. These simulations can address the modelling of a **monolayer of cells**, such as the ones observed in *in-vitro* cell growth in Petri dishes; the modelling of a sphere made up of 3000 cells surrounded by matrigel, called **spheroid**; and the modelling of **organoids** of cancer cells, such as the ones present in xenografts, with their associated array of immune cells, blood vessels and a detailed extracellular matrix. In spite of great advances in the field, in order to address organoids' simulation, effective tools that scale-up their simulations outputs and include such complex biological and physical set-ups are still much needed.

In this project, the Life Sciences use case will provide a *"virtual laboratory"* for studying cancer growth and evolution by using multi-scale models of tumour systems [31]. The goal of this use case is to facilitate the design, test, and optimization of cancer treatments based on combinations of different drugs and dose scheduling strategies. This study will be key to understanding the mechanisms underlying the emergence of cancer resistance to target therapies and, ultimately, extend the life expectancy of patients [20]. Although it will not suppress the need for experimental research, *in-silico* modelling can help refine the experimental programs aiming to reduce costs and increase research efficiency [31]. For instance, *in-silico* screening of drug effects have been used to predict synergistic effects between pairs of drugs [12].

Given the biophysical, biochemical, and biomechanical factors included, **multi-scale models (MSM) can help identify the factors that drive a treatment to success or failure**. However, due to the uncertainties regarding the underlying biology, MSMs can be cast in many alternative ways, and entail a high number of parameters [25]. Furthermore, simulating MSMs implies the resolution of many different numerical problems (*e.g.* large systems of partial differential equations (PDE), boolean equations), thus representing a computationally demanding task. High-performance computing (HPC) systems, such as the Barcelona Supercomputing Center (BSC) MareNostrum 4 supercomputer, represent amenable environments for intensive simulation of MSMs that produce extreme-scale data streams as outputs. In summary, the exploitation of such models is challenged by three main factors: i) the uncertainties of the underlying biological assumptions of the model; ii) the high-dimensional parameter space to be explored; and iii) the processing of the extreme-scale data streams generated by a single simulation.

Herein, to improve the scope and predictive power of MSM simulations, we plan to integrate existing technologies that provide a computational means for high-throughput hypotheses testing and parameter selection. First, we will extend and deploy the existent framework PhysiBoSS [22] in the BSC MareNostrum 4 supercomputer, enabling multi-scale simulations of cancer growth in a HPC environment. Then, we will integrate the forecasting approaches developed within the INFORE project to improve the parameter space exploration. To fit the model parameters, we will use experimental data of cell lines growing under different drug dosages. **Forecasting techniques will be used to conduct an online monitorization of the simulations to interactively learn which parameters are critical to reproduce the experimental results.** In particular, complex event forecasting methods will be used for the online monitoring of the simulation, aiming to discard those treatments strategies that exhibit poor performance in terms of reducing or stabilizing tumour growth. Finally, the calibrated models will be used to predict novel treatments strategies.

3 Description of the Multi-scale Model

To study the process of resistance acquisition by cancer cells through *in-silico* simulations, we will use a MSM, *i.e.* **a model that integrates cellular processes taking place at multiple time and space scales** (*e.g.* molecular level, population level). MSM simulations represent a powerful approach to test alternative hypotheses about phenomena observed in cancer, enabling the prioritization of optimal drug treatments [19]. By considering multiple scales of description in a single model, it is possible to find putative causal connections between processes acting at subcellular level (*e.g.* mutations or drug affecting the regulatory networks of a cell), and processes that occur at a population level (*e.g.* competition for resources, tumour heterogeneity). For instance, to study the relationship between tumour heterogeneity and treatment response, a population level description in which each individual cell exhibits certain variability in its molecular machinery, is required [21]. At a lower scale of description, processes taking place inside each individual cell are modelled through specific systems biology approaches, *e.g.* Boolean Logic is used to simulate signalling networks.

In this study, we will use PhysiBoSS, a tool that merges the agent-based simulator PhysiCell [14] with the stochastic Boolean model simulator MaBoSS [27][28]. **The higher scale of description is the population level, which is modelled using the agent-based model part of PhysiBoSS**. This flexible framework is used to simulate population dynamics, while also considering spatial and physical constraints. In a PhysiBoSS simulation, each individual agent represents a single cell that can grow, divide, or die (from necrosis or apoptosis), according to a set of physical rules and constraints, as well as inputs coming from the lower levels of description. Moreover, this agent-based framework is also able to model cell-cell and cell-environment interactions as well as the environment itself. The environment is represented in terms of concentration or densities of different molecules such as nutrients (e.g. O_2, glucose), waste products of the cells (CO_2, lactate), drugs, and signalling molecules. In PhysiBoSS, **the diffusion of the different molecules is governed by a system of PDEs.** Moreover, each individual cell is represented by a discrete agent; the physical behaviour of the agents (e.g. movement, cell-cell contacts) is ruled by mechanical equations.

At the lower scale of description (individual cells level), three main processes are modelled within the agent-based model part of PhysiBoSS (*i.e.* PhysiCell): i) **nutrient consumption** and cell maintenance; ii) the **cell cycle** which includes cellular growth and division; and iii) the **signal transduction** network of the cell, as well as the underlying decision making process where the cell adopts alternative phenotypes or phases (*e.g.* proliferation, apoptosis) to cope with the external conditions and signals. As for nutrient consumption and cell maintenance, we will use the default growth model integrated in PhysiCell[1]. On the other hand, as for the cases of signal transduction network and cell cycle, we are developing model improvements explained in Sections 3.1 and 3.3, respectively.

3.1 The Signalling Network Component

The signal transduction machinery of a cell is composed of a network of molecular components (*e.g.* protein complexes, small molecules), which allows the cell to decode different signals and adjust its internal state to respond to different stimuli [32]. The interaction between the different molecular components can result in activation or inhibition and be wired in such a way that these pathways can include feedback and/or feedforward loops. Thus, **a signalling pathway consists of sets of proteins and small compounds used by the cell to process different signals and modify its state in response to certain stimuli.**

[1] For further details visit: https://github.com/MathCancer/PhysiCell/blob/master/documentation/User_Guide.pdf

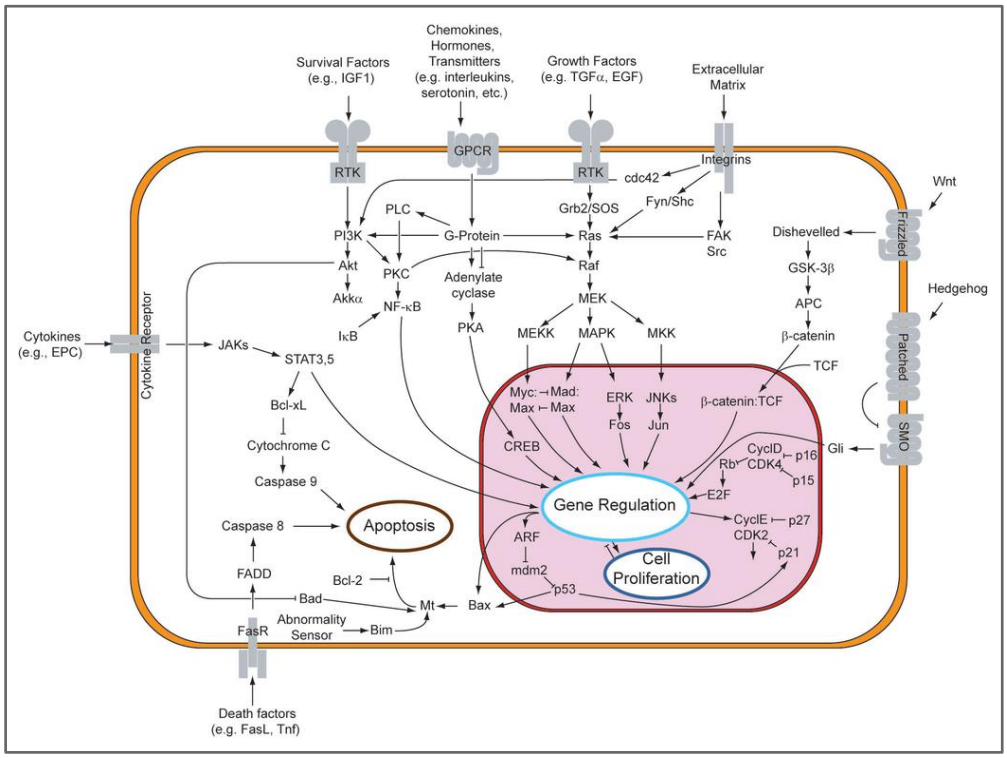

Figure 1. Schematic representation of a signalling network which includes different signalling pathways found in most human tissues. Downstream the signal transduction pathways, different effectors change the cell phase or phenotype (e.g. proliferation, apoptosis) to respond to different stimuli. Orange and red rounded rectangles represent the cell membrane and nucleus, respectively. Labels indicate protein complexes whereas arrows and T-bar lines represent activation and inhibitory interactions, respectively. Source [2]

In Figure 1, a schematic representation of two signalling pathways involved in prostate cancer is depicted. Distinct signalling pathways are meant to process different classes of external and internal signals, such as the availability of nutrients and the space to proliferate, the level of DNA damage, and others. The transduction of a signal will induce the cell to respond by changing its internal state or the phenotype of the cell.

Due to molecular noise, among other factors, living systems tend to accumulate mutations in the DNA during their lifespan. While some mutations may not alter the molecular functions of the cell, others can cause the malfunction or dysregulation of particular cellular processes. For instance, mutations affecting pathways involved in the cell cycle can lead to undesired cell phenotypes, such as uncontrolled cell growth, *i.e.* cancer [16]. For this reason, **targeting dysfunctional signalling pathways is one of the main focus in the development of novel target therapies.** Signalling networks are complex systems that exhibit non-trivial behaviour. Thus, mathematical and computational models are needed to rationalize and understand their functioning and predict novel therapies.

From a modeling point of view, a signaling network is a dynamical systems governed by a set of differential equations. The structure of systems is described as a network where the nodes, representing proteins and complexes, are connected through directed signed interactions, indicating activation or inhibition [9]. Additionally, each node can be found in one of two states (active or inactive). The state at a given time point is computed through an activation function which integrates the different inputs of the node. The activation function operates as a logical gate that integrates the incoming inputs of a node (from a previous time point) and gives the updated node state. **It is important to clarify the difference between the signalling network model and the approach used to simulate its behaviour: the signalling network can change in different cell types, while the same signalling model can be simulated with alternative approaches.**

[2] https://en.wikipedia.org/wiki/File:Signal_transduction_pathways.png

Different mathematical approaches have been developed to simulate and analyse signalling network models. Most common approaches are based on ordinary differential equations [34], Boolean logic [8], or stochastic approaches such as MaBoSS [27][28]. Such simulation frameworks are used to study the dynamics of signalling networks and, in the case of cancer research, to predict the effect of single drugs and combined therapies. In this project, **we will simulate and analyse signalling models using the stochastic boolean approach implemented in PhysiBoSS**, as it allows to compute the probabilities associated to the proliferative and apoptotic phenotypes.

The signalling network is subject to the physiological role performed by the cell, thus cells belonging to different tissues exhibit differences in their signalling pathways. Differences can also be found between individual cells, and even between cells belonging to the same tumour. For this reason, models of signalling networks should be calibrated or adjusted to represent a particular cell type [6]. In this project, **we will use a curated signalling model of the adenocarcinoma cancer cell line AGS** [12]. The model depicted in Figure 2 accounts for most of the signalling pathways known to play a relevant role in this type of cancer (*e.g.* MAPK, PI3K/AKT/mTOR, Wnt/β-catenin and NF-κB). This model was successfully used to predict synergistic effects of different pairs of drugs that were validated experimentally [12].

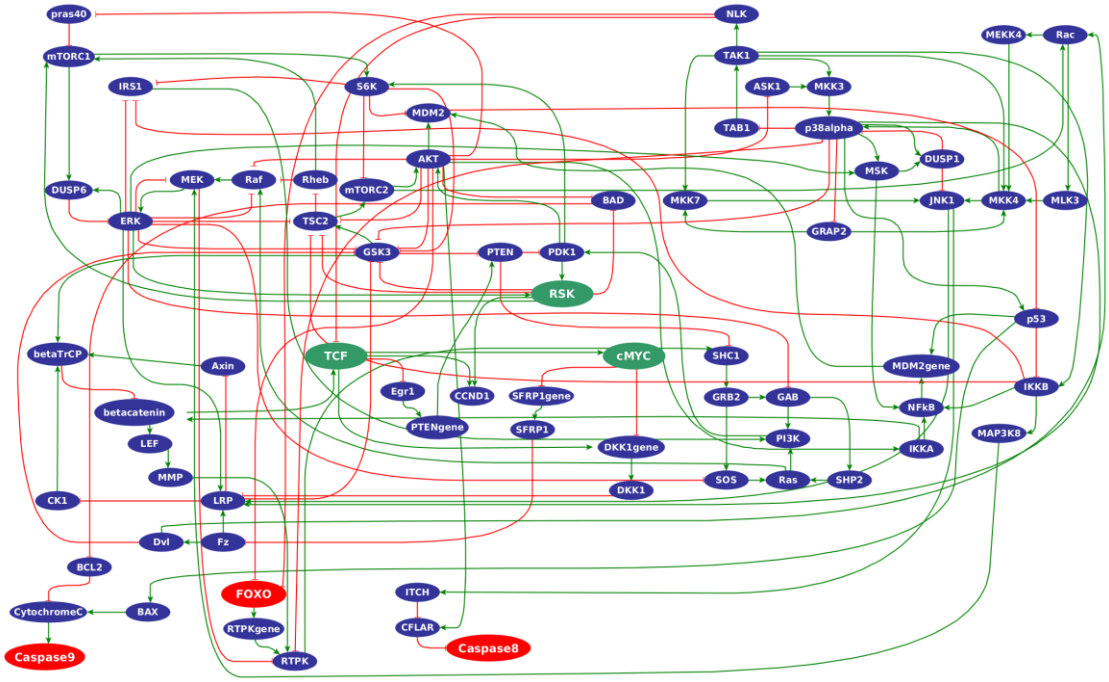

Figure 2. Network model of the AGS cell line signalling. Nodes correspond to proteins or complexes of proteins. Red and green arrows correspond to inhibitory and activation interactions, respectively. Red and green nodes indicate those nodes associated with anti-survival and pro-survival phenotypes, respectively.

3.2 The Agent-Based Component

Population dynamics are simulated using the agent-based modelling part of PhysiBoSS, that comes from PhysiCell (https://github.com/MathCancer/PhysiCell), a powerful lattice-free physics-based agent-based cell simulator for 2-D and 3-D multicellular systems [14]. The PhysiCell framework allows studying different genetic variations (mutants, patient- or cell-line-specific boolean models), variations to physical properties (cell-cell adhesions, cell-matrix adhesions) under different microenvironmental conditions (presence of oxygen, signalling molecules or extracellular matrix). On the other hand, MaBoSS is used to simulate the signalling network of each individual agent. The integration of PhysiCell and MaBoSS was done by collaborators at Curie Institute in a new tool termed PhysiBoSS (https://github.com/sysbio-curie/PhysiBoSS/wiki). Thus PhysiBoSS consists of a stochastic boolean model

simulator (MaBoSS) that predicts cell fates embedded in a flexible agent-based model (PhysiCell) that simulates multicellular systems [22].

Figure 3 represents the outcome of a typical agent-based simulation. In this representation each single sphere corresponds to a cell agent and the colour code indicates the cellular phenotype. For instance, the brown-coloured cell at the center of the structure correspond to necrotic cells, *i.e.* cells that die as a consequence of the lack of nutrients and oxygen. On the right-hand side of Figure 3, some properties of the individual agents are listed. In particular, notice that each individual agent holds its own signalling network, which is used to update the phenotype of the cell at each time step.

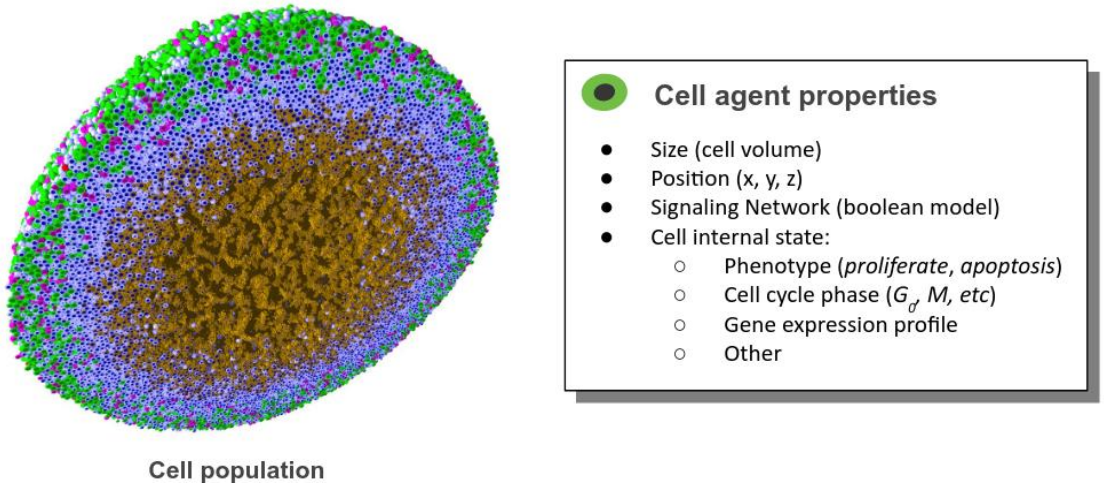

Figure 3. Visual representation of a multi-scale simulation of a population of cells. On the left side a population of cells arranged in a 3-D spheroid are depicted. Cell colours indicate cells exhibiting alternative phenotypes. For instance, the brown core at the center of the spheroid corresponds to necrotic cells that die because of the lack of nutrients. On the right-side panel, the different attributes or properties of each individual cell agent are shown.

The coupling between the different scales is conducted by using output of one model as input of the other. Specifically, the availability of nutrients or the presence of drugs around a cell agent are used as inputs of the signalling model of that cell. Subsequently, the output of the signalling model is used to update the behaviour of the cell agent. For example, during a simulation, a drug pulse is introduced in the environment; when the drug reaches a particular agent, it will affect the function of its protein target, which corresponds to a node within the cell agent signalling network; then, the state of the signalling network is updated to assess the effect of the drug perturbation by running the boolean stochastic simulator of PhysiBoSS. The new state of the signalling network is used to update the phenotype of the cell agent, which for instance might enter into apoptosis due to the drug effect, closing the loop.

Therefore, **we will use PhysiBoSS as the main framework to perform MSM simulations**. Nonetheless, we plan to develop further extensions to this framework to account for other biological processes such as metabolism. Finally, an example of a typical simulation result is reported in Appendix B.

3.3 The Cell Cycle Neural Network Component

BSC and CRG are collaborating to integrate a dynamical cell cycle model in the Agent-Based-Model framework (Physicell/PhysiBoss). To provide a quantitative computational architecture to simulate and describe the cell cycle, a neural network that reproduces gene expression time course will be implemented. The rationale for this approach is: i) gene expression levels can be modelled as normalized continuous variables that evolve in time (Figure 3); ii) a deep neural network approach suits well the proposed large-scale framework.

Building on CRG expertise on neural networks [10][29], a new method will be proposed to predict the cell cycle state at time $t+\Delta t$ based on the expression levels of a set of genes measured at times $\leq t$. **The method will include state-of-the-art Deep Learning methodologies** including Long Short-Term Memory (LSTM) [17] and attention mechanism (AM) [4]. We will use a 'dual attention' network: the first level AM will be responsible for selecting the genes out of the pool that are master regulators of the cell cycle state; and the second AM will be applied on the selected genes of the first level and will be responsible for focusing on important time steps of their expression. These two attention models will be integrated within an LSTM-based recurrent neural network (RNN) and will be jointly trained using standard backpropagation. In this way, the Dual Attention – RNN will adaptively select the most important time steps of most relevant inputs and capture the long-term temporal dependencies.

The aim of reproducing a cell cycle behaviour, and being able to connect it with the growth rate of the cell will enable to incorporate such model to the multi-scale simulation of cell growth. In a preliminary study, we explored the feasibility of connecting this RNN model describing the cell cycle state with the activity of the signalling proteins present in the signalling boolean model.

4 Scenario Definition

The Life Sciences use case will explore different mechanistic explanations to help to understand how cancer cells acquire resistance to a particular treatment [18]. In this context, a mechanistic explanation will be defined by a set of assumptions and a set of model parameters, which need to be validated by contrasting the simulation outputs against experimental results. For instance, there are alternative models to explain the emergence of a resistant cell [26]; resistant cells can be present in a population before being exposed to the drug and when the drug is provided the resistant cells are selected. In an alternative model, the stress induced to cells by the introduction of the drug triggers cells' responses by changing the expression of the genes to find some way to avoid the drug effect. In addition, the acquisition of resistance to a particular drug can be reversible or irreversible, meaning that it can occur at the level of the phenotype or the genotype. The alternative models will be implemented and tested in the multi-scale framework described in Section 3, to test whether any model reproduces better the experimental data. On the other hand, MSM simulations require a high number of parameters to be tuned. In addition to the exploration of alternative resistance acquisition models, we will study how the emergence of resistant cells can be mitigated by using combinatorial therapies based on pairs of drugs exhibiting a synergistic effect [12]. On the other hand, we will explore different drug scheduling schemes, such as intermittent pulses and drug alternation [5], to test which strategies show better results.

In Figure 4, a schematic representation of the workflow to run MSM simulation is depicted. In the figure-panes a-b, the components of the multi-scale framework described in Section 3 are used to run a simulation. In Figure 4-c the simulation outcome for several *in-silico* experiments performed by Letort *et* al. [22], are shown: in each of the three experiments a different drug dose schedule was evaluated. Strikingly, the authors found that the drug dose provided in pulses at intervals of 300 minutes proved to be the best strategy [22]. PhisyBoss software allows to perform powerful simulations and to explore in more detail complex processes that can play a critical role in the emergence of resistant cells. Moreover, by considering the intracellular cell processes it is also possible to model phenomena that take place at the molecular level, such as the presence of synergy between different drug combinations. However, due to its complexity, the simulations with this software are very computationally demanding. Consequently, the possibility to perform an online monitoring of the simulation is of great interest in the detection of patterns, which allow anticipating or predicting future events of interest, without the need to wait until the end of the simulation.

We plan to have a first scenario in PhysiBoSS that employs MSM to replicate experimental results obtained from the growth of AGS cell lines treated with different drug regimes. Through these simulations, we will explore two alternative geometries for cell arrangement: one-cell-thick 2D monolayers and 3D spheroids. The 2D monolayer is used to simulate cells growing in a plate, which correspond to the setting in which drug experiments where performed. On the other hand, the 3D spheroids are more suitable to simulate the growth of an *in-vivo* tumour and will be used to simulate real tumours and to explore different treatment strategies.

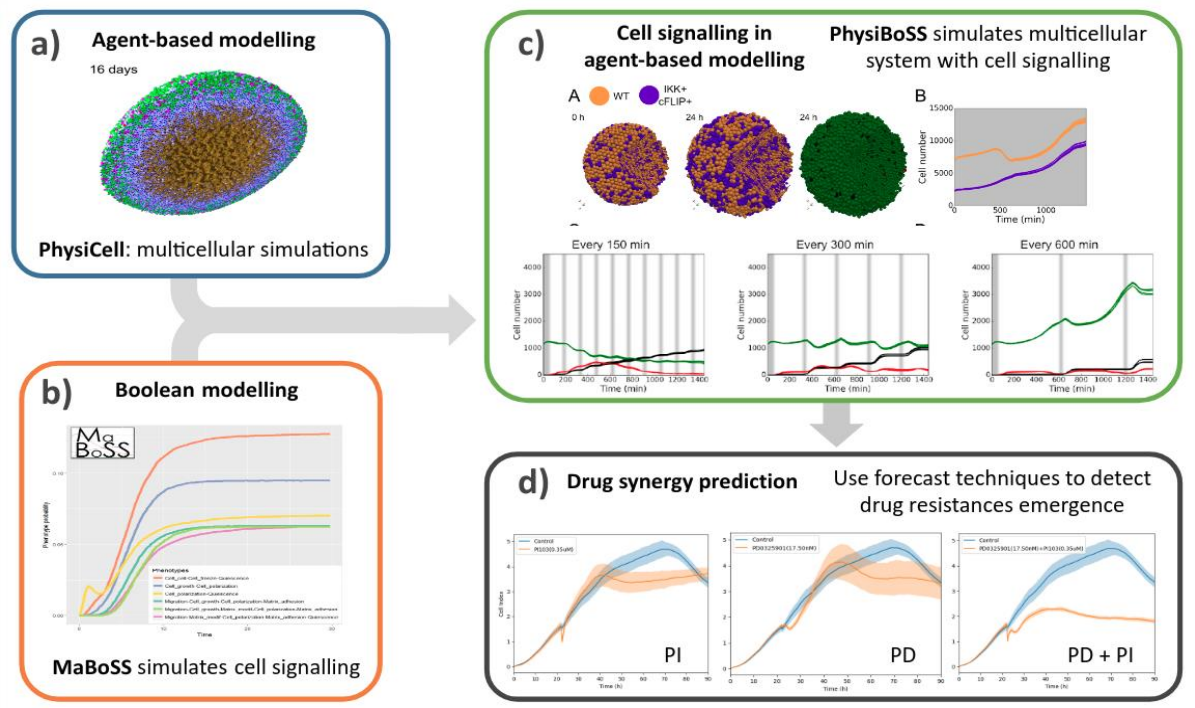

Figure 4. Representation of the MSM framework. MSM combines simulations of cell populations and the environment using an agent-based model approach (a), and the internal cells' state using a stochastic Boolean simulator (b). Results of the application of PhysiBoSS to investigate drug regime treatments are shown (c) (Letort et al, 2018). Experimental growth curves of cells treated with different drugs dosages (PI or PD, as examples of drugs) or combinations of drugs (PD+PI) in orange are shown (d) (untreated cells are shown in blue).

We will start with two AGS boolean models based on Flobak *et al.*[12] (Figure 2), the complete and reduced signalling models, and later on we will consider using an expanded model which include more nodes participating in the transduction of signal at the level of individual cells. Together with consortium partners, we will study how to include techniques to forecast various events of interest that occur during simulations, as well as interactive learning techniques to assist the calibration of the parameters of *in-silico* models. Specifically, we are interested in detecting patterns that can be broadly grouped into two classes of event: 1) **detection of cases of undesired simulations outcomes**, when performing large-scale parameter screenings; and 2) **detection of events of biological interest**, for example identifying combinations of small changes in physiological or molecular markers that can anticipate the appearance of resistant clones.

The first class of events will be used to optimize parameter and model exploration while fitting experimental data. In this context, possible undesired simulations outcomes include: 1.1) cases which are unfeasible from a biological point of view; 1.2) cases where there are no longer living cell-agents; 1.3) cases of drug effect saturation; and 1.4) cases in which the simulation trajectories do not fit the experimental growth curves. On the other hand, events of biological interest will include: 2.1) non-trivial patterns that can be used to predict the early emergence of resistant cells; and 2.2) discovery of drug dosage and combination strategies that help reduce or stabilize the tumour growth. An example of simulation's preliminary results can be found in Appendix B.

As a second scenario, we will address bigger simulation set-ups with cancer cell spheroids of more than 5,000 cells and aim to scale this up to 500,000 cells using high-performance computing at BSC's MareNostrum4. In this scenario, a reliable model of cell cycle regulation is crucial to deliver accurate simulations of cell growth [11][33]. We plan to integrate dynamic cell cycle models from partners at CRG (see Section 3.3). The PhysiBoSS framework is flexible enough to be able to work with other signalling network models with minimal code additions. In fact, we

are planning to integrate alternative signalling network models that are focused in other cancers such as breast cancer [15] and cell fates phenotypes [8], as well as other generic cancer signalling models [13]. Furthermore, we plan to extend the PhysiBoSS framework to take into account the cellular metabolism within each agent.

As a third scenario, we will consider more complex simulation set-ups such as considering microenvironmental physical properties (extracellular matrix stiffness, accessibility), as well the interaction between cancer cells and other cell types, to study combined therapies [30]. In particular, we can explore the use of different cell types by implementing specific boolean models for each one of them, being distinct cancer-associated cells such as immune cells, stroma cells, etc. Finally, as a perspective, the simulation framework could be potentially applied to study other biological scenarios such as streak experiments, microfluidic designs, microenvironmental heterogeneities, etc.

4.1 Key Performance Indicators and Case Scenario

The design and implementation of the Life Sciences use case will allow us to reach the Key Performance Indicators (KPIs) described in the Grant Agreement. For instance, the optimal use of high-performance computing (HPC) at BSC's MareNostrum4 will allow scaling-up the simulations from thousands of cells up to a billion of cells. In this regard, our modelling framework already allows to use OpenMP (https://www.openmp.org/), permitting us to fully exploit the multicore architecture of each node of MareNostrum4. In the project, we are designing different ways to further optimize the parallelization process to multiple nodes using MPI and other parallel frameworks such as COMPs [3], developed at BSC. We estimate that the optimized parallelization techniques will scale the volume of the data stream from 100 MB/min to 100 GB/min by increasing the number of agents and small molecules considered, as well as the time resolution at which the simulation outputs results.

In addition to the scale of simulations, we will create a benchmark scenario to test the forecasting techniques using an a signalling model we previously used in our work (*i.e.* the AGS cell line). In INFORE Grant Agreement, the prototype on this use case aimed to forecast up to 5 key events. We will use our framework to detect undesired simulation outcomes, when performing large-scale parameter screenings; and to detect events of biological interest, for example identifying combinations of small changes in physiological and molecular markers that can anticipate the appearance of resistant clones.

Preliminary key events to be forecasted are grouped in desired and undesired behaviours. The undesired simulations outcomes include cases which are unfeasible from a biological point of view as well as cases of drug effect saturation. For instance, a preliminary proposal of forecast event is the detection of uninformative simulations, *i.e.* simulations that have read-outs that greatly differ from relevant biological data. The forecasting of these events will reduce uninformative simulations and avoid wasting computational resources while addressing highly demanding simulations such as those using billions of cell agents. On the other hand, events of biological interest include non-trivial patterns that can be used to predict the early emergence of resistant cells or the discovery of drug dosages and drug regimes that help reducing or stabilizing the tumour growth. A more detailed description of these 5 key events was discussed in the introduction of Section 4 ("Scenario Definition").

In Deliverable D1.2, we will provide a definitive list of key events to forecast by the INFORE prototype.

4.2 Set of Applicable Computational Techniques

- **Boolean model solver:** To study Boolean asymptotic solutions, we will use GINsim software [24] and BoolNet R package [23].
- **Stochastic Boolean solver:** To study the transitions of the model when reaching asymptotic solutions, we will use MaBoSS [27][28]. This tool, based on Monte-Carlo kinetic algorithm, performs stochastic simulations on logical models. This strategy allows to calculate probabilities of each phenotype enabling a semi-quantitative approach to study the model outcomes and their response to different perturbations such as drugs. MaBoSS results are analysed using tailored scripts in Python and R.
- **Multi-scale model:** Our MSM merges the stochastic boolean model of MaBoSS with the agent-based PhysiCell software [14] into a tool named PhysiBoSS [22]. PhysiCell as well as PhysiBoSS can be used

with OpenMP and thus are suitable to be deployed in multi-thread computers. We are planning coordinated activities with PhysiCell developers to design an implementation of an MPI-based PhysiBoSS code.

- **Visualisation of simulation results:** We generally use Paraview software (https://www.paraview.org/) to visualise PhysiBoSS results. In addition, other software is used for general data analyses such as R or Python.

5 Experimental Data for Model Calibration

The use of drugs and mutants to study cell growth is an established strategy to uncover mechanistic associations in the molecular biology of cell growth, as drugs and mutants can hamper the functioning of a vital pathway and cell viability to grow is diminished. If the target of the drug or the function of the gene is known, researchers can study the robustness of the different signalling pathways against a given perturbation. In our modelling framework these drugs' hampering effects can be simulated by inhibiting the activity of a given node that correspond to a protein or pathway. This same strategy can be followed to study the effects of different mutants on cell growth, such as knock-outs, over-expressed genes or knock-downs.

We plan to use growth experiments on a gastric cell line (AGS) to calibrate and test the predictions of the MSM simulations in 2D monolayer and 3D spheroids. AGS is a cell line derived from gastric adenocarcinoma and is a model system to study this type of cancer. We have gathered experimental datasets for cells growing under different treatments conditions, which include single drug experiments at different doses as well as studies of pairs of drugs (see Table A.1 in Appendix A for further details regarding drug and dose used). These works have been published by collaborators at the Norwegian University of Science and Technology (NTNU) [12].

Cell growth measurements were done in real-time using the xCELLigence RTCA SP growth assay (Roche Applied Science)[3]. Experiments were done in four replicates and controls (untreated cells) were included. This system uses culture plates with gold electrode arrays at the bottom of each well. The electrodes are used to perform real-time measurements of the impedance across the gold arrays. These measurements are reported in the dimensionless unit of cell index, which is taken to correspond to the number of cells (for further details see Flobak *et al.* [12]).

Examples of growth under a single drug, combination of two drugs and their controls are depicted in Figure 4d. Cells grow exponentially until nutrients are consumed and cells die of starvation in the control experiment (Figure 4d, blue lines), while the population stops growing and, in some cases, decreases upon treatment with drugs (Figure 4d, orange lines), i.e., drugs hamper the functioning of a vital pathway and cells grow less. We plan to use part of these experiments to calibrate parameters of the MSM model in 2D monolayer and 3D spheroids – to use the same experimental conditions to simulate the model and reproduce the experimental results. Once this is done, we will use the calibrated model to propose other experiments not included during the parameter calibration (or learning) step and predict their outcome.

Additionally, we have another dataset of unpublished gene expression RNA-seq data for the same AGS cell line under five drug treatments from our collaborators at NTNU. We plan to use these gene expression data to, first, constraint the boolean AGS signalling network and, afterwards, to use it to refine the cell cycle model from CRG that uses the actual cell cycle state (t) of a cell to infer the gene expression profile at time t+1 (see Section 3.2 for further details on the cell cycle model implementation).

Finally, we also have access to drug treatment experiments on other cell lines, which include 19 different drugs and their combinations on eight different cell lines of different cancer types (see Appendix A). For this dataset, we have access to the cell-line-specific IC50[4] values for the different drugs. On a later phase of the project, we plan to use the calibrated multi-scale model to predict these IC50 values.

Additional datasets can be retrieved from public repositories of cell-line-specific drug sensitivities data such as Genomics of Drug Sensitivity in Cancer (GDSC)[5]. This database covers the effect of over 300 drugs on over 500 cell lines, AGS being among them.

[3] Experimental system device to monitor cells in real-time. It allows a non-invasive measurement of cell viability

[4] The IC50 is a measure of the potency of a substance in inhibiting a specific biological or biochemical function. The quantitative measure how much of a particular drug or other substance (inhibitor) is needed to inhibit a given biological process (source: https://en.wikipedia.org/wiki/IC50)

[5] https://www.cancerrxgene.org/

6 Future developments: Multi-Scale Model HPC Implementation

We are currently working in the deployment of the MSM simulation framework PhysiBoSS in BSC's supercomputer, the MareNostrum 4. This involves the parallelization process discussed in Section 4.1 using different technologies such as OpenMP, MPI or COMPs to scale-up our simulations in terms of numbers of cells considered.

To complement this scaling-up, we plan to execute another test that follows previously published cases on treatment regime studies [14][22]. For instance, we plan to reproduce the in-silico experiments of the Tumour Necrotic Factor (TNF) effect on spheroids growth and to expand the experiment including other regimes. This would allow testing the effect of microenvironmental changes in chemicals presence and microenvironment set-ups in 3D spheroids that AGS drug experiments did not address. Due to the high dimensionality of the parameter space, we plan to run the simulations following the high throughput computing (HTC) framework that integrates a mechanistic multi-scale simulator (PhysiBoSS) with an extreme-scale model exploration platform [25].

These advances will allow addressing, bigger simulations setups with cancer cell spheroids of more than 5,000 cells and aiming to scale this up to 500,000 cells optimally using high-performance computing. As a third scenario, we could consider more complex simulations set-ups such as streak experiments, microfluidic designs, etc.

Another future development will be the inclusion of metabolic modelling in PhysiBoSS. In this way, we will integrate a metabolic scale to the multi-scale simulator to study the interactions of environment, cell signalling and metabolic alterations in real-sized tumours.

7 Expert Users and their Questionnaires

We contacted researchers from public European research centres and high-education to be our Life Science's use case expert users. They are both experienced researchers that cover different aspects in Systems Biology, modelling, omics data processing and pathway-based data analyses. At the stage of the project, the two expert users are established scientists.

Both researchers were contacted and consented to participate in our questionnaire (Appendix C) and their anonymised responses, accounted for in our previous discussion, can be found in Appendix D. From the interviews and questionnaires, we extracted workflows of interest and analysed their needs and wills to build concrete user scenarios and analytical tools. All expert users agreed on the importance of the present project and the impact its outcomes would have on their current workflows. The models and know-how of these experts are a reflection of the state of the art on the field of biological modelling and their opinions are crucial as they will be the ones in charge of exploring and incorporating in their future developments real-time data and forecasting techniques developed in the project. In fact, according to their questionnaires and interviews they were willing to incorporate these tools and data on their day-to-day work and were very positive on the impact these would have in their fields. They had high interest on being able to have a framework that could facilitate the development of forecasting techniques on real-time data.

8 Incorporating Complex Event Forecasting within the Simulation Process

The study of complex biological systems relies on the execution of heuristic methods that require very large numbers of simulations, a process Ozik and collaborations termed as model exploration (ME) (Ozik *et al*, [25]). Applying ME to MSMs involves an iterative workflow where simulations are run across a high dimensional parameter space and changing initial conditions to explore alternative simulation outcomes.

Figure 5 schematically represents the workflow of ME, where simulations' outputs from a set of in-silico experiments are evaluated against some predetermined metric, which informs the next iteration of simulation experiments. In our case, we plan to use the growth experiment to calibrate the model. Moreover, the ME process can be enhanced with the use of complex event forecasting techniques, which can be used to improve the parameter space exploration. Here, we will study how to include techniques for forecasting various events of interest that occur during the simulation, as well as interactive learning techniques to assist the calibration of the *in-silico* models. An example of simulation's preliminary results can be seen in Appendix B.

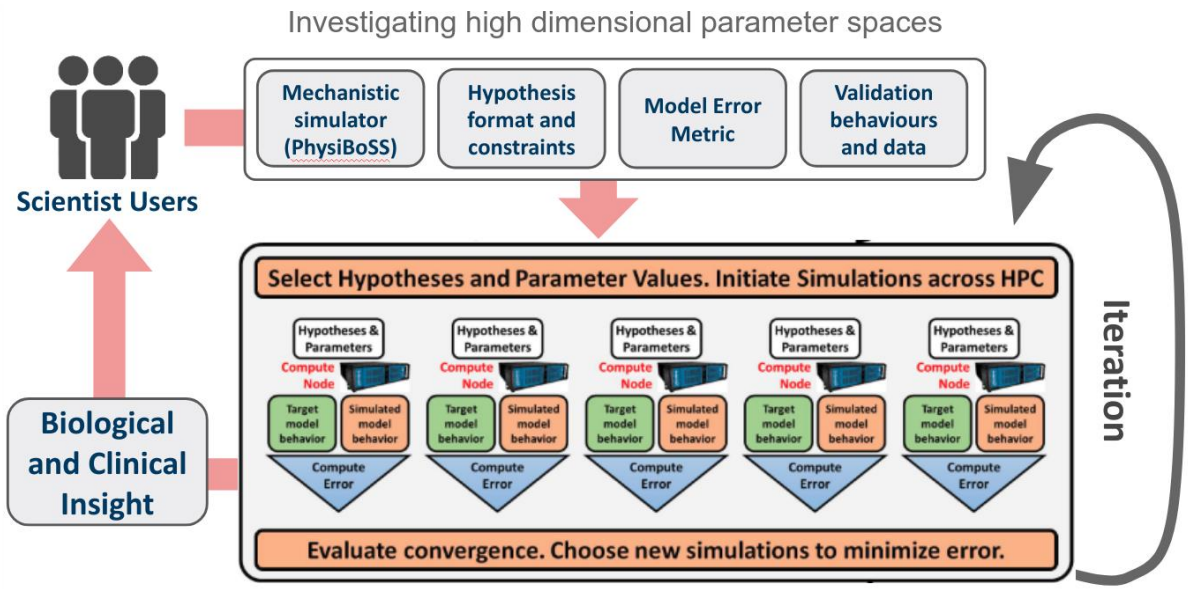

Figure 5. Model Exploration workflow. Adapted from (Ozik et al, 2018 [25]).

9 Conclusions and Future Work

In this deliverable, we described the key steps in achieving Work Package 1 objectives and included results of the first collection of questionnaires for the INFORE Life Sciences user scenario. We have detailed the Life Sciences use case: the different scales of description, the different parts and how its design and implementation allows us to reach the KPIs described in the Description of Work (DoW). The collection of the expert users' questionnaires, the description of the use case as well as the documentation process has already allowed discussions about the functionalities, terminologies, and dependencies within the consortium. As scheduled in the DoW, this initial use case scenario definition will be further studied and discussed in the following months, which will lead to an extended and revised version of this deliverable (D1.3 on Month 18). D1.3 will detail the actual analysis algorithms to be used for the reduction of uninteresting simulations among other forecasted events and how these algorithms interact with the data coming from the use case. The description of the user scenarios and requisites will guide the work until the completion of the project. In addition, results in the form of INFORE's prototypes will undergo further expert user evaluation according to our work plan.

Furthermore, we have detailed the different components of our modelling framework, how they are connected and how each one of them is necessary to have simulations of billions of cells that can be experimentally validated. We are currently working on the introduction of the cell cycle component in the boolean component to be able to capture transient effects that might not be covered by the boolean component. In a first development phase, we aim to scale up simulations of cancer cell 3D spheroids from 5,000 up to 500,000 cells using optimized high-performance computing. This scenario will allow the design of different set-ups that tally cancer tumour growth conditions with increased number of cells, altered microenvironmental physical properties, different cell types, as well as to study the interaction between cancer cells and the immune system. In future work, this will allow the consideration of different simulations' set-ups beyond cancer tumour growth such as *in-vitro* streak experiments, microfluidic designs, among others.

10 References

[1] An G (2010) Closing the Scientific Loop: Bridging Correlation and Causality in the Petaflop Age. *Sci. Transl. Med.* **2:** 41ps34-41ps34

[2] Anderson ARA, Weaver AM, Cummings PT & Quaranta V (2006) tumour Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. *Cell* **127:** 905–915

[3] Badia RM, Conejero J, Diaz C, Ejarque J, Lezzi D, Lordan F, Ramon-Cortes C & Sirvent R (2015) COMP Superscalar, an interoperable programming framework. *SoftwareX* **3**–**4:** 32–36

[4] Bahdanau D, Cho K & Bengio Y (2014) Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv14090473 Cs Stat* Available at: http://arxiv.org/abs/1409.0473 [Accessed March 26, 2019]

[5] Barros de Andrade e Sousa LC, Kühn C, Tyc KM & Klipp E (2016) Dosage and dose schedule screening of drug combinations in agent-based models reveals hidden synergies. *Front. Physiol.* **6:** 398

[6] Béal, J., Montagud, A., Traynard, P., Barillot, E. & Calzone, L. Personalization of Logical Models With Multi-Omics Data Allows Clinical Stratification of Patients. Front. Physiol. 9, (2019).

[7] Brady SW, McQuerry JA, Qiao Y, Piccolo SR, Shrestha G, Jenkins DF, Layer RM, Pedersen BS, Miller RH, Esch A, Selitsky SR, Parker JS, Anderson LA, Dalley BK, Factor RE, Reddy CB, Boltax JP, Li DY, Moos PJ, Gray JW, et al (2017) Combating subclonal evolution of resistant cancer phenotypes. *Nat. Commun.* **8:** 1231

[8] Calzone L, Tournier L, Fourquet S, Thieffry D, Zhivotovsky B, Barillot E & Zinovyev A (2010) Mathematical modeling of cell-fate decision in response to death receptor engagement. *PLoS Comput. Biol.* **6:** e1000702

[9] Calzone, L., Barillot, E. & Zinovyev, A. Logical versus kinetic modeling of biological networks: applications in cancer research. Current Opinion in Chemical Engineering 21, 22–31 (2018).

[10] Delli Ponti R, Marti S, Armaos A & Tartaglia GG (2017) A high-throughput approach to profile RNA structure. *Nucleic Acids Res.* **45:** e35

[11] Ferrell JE, Tsai TY-C & Yang Q (2011) Modeling the Cell Cycle: Why Do Certain Circuits Oscillate? *Cell* **144:** 874–885

[12] Flobak Å, Baudot A, Remy E, Thommesen L, Thieffry D, Kuiper M & Lægreid A (2015) Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling. *PLoS Comput. Biol.* **11:** e1004426

[13] Fumiã HF & Martins ML (2013) Boolean network model for cancer pathways: predicting carcinogenesis and targeted therapy outcomes. *PloS One* **8:** e69008

[14] Ghaffarizadeh A, Heiland R, Friedman SH, Mumenthaler SM & Macklin P (2018) PhysiCell: An open source physics-based cell simulator for 3-D multicellular systems. *PLOS Comput. Biol.* **14:** e1005991

[15] Gómez Tejeda Zañudo J, Scaltriti M & Albert R (2017) A network modeling approach to elucidate drug resistance mechanisms and predict combinatorial drug treatments in breast cancer. *Cancer Converg.* **1:** 5

[16] Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. Cell 100, 57–70 (2000).

[17] Hochreiter S & Schmidhuber J (1997) Long Short-Term Memory. *Neural Comput.* **9:** 1735–1780

[18] Holohan C, Van Schaeybroeck S, Longley DB & Johnston PG (2013) Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* **13:** 714–726

[19] Ji Z, Su J, Wu D, Peng H, Zhao W, Zhao BN, Zhou X, Ji Z, Su J, Wu D, Peng H, Zhao W, Nlong Zhao B & Zhou X (2017) Predicting the impact of combined therapies on myeloma cell growth using a hybrid multi-scale agent-based model. *Oncotarget* **8:** 7647–7665

[20] Kessler DA, Austin RH & Levine H (2014) Resistance to chemotherapy: Patient variability and cellular heterogeneity. *Cancer Res.* **74:** 4663–4670

[21] Kim E, Kim J-Y, Smith MA, Haura EB & Anderson ARA (2018) Cell signaling heterogeneity is modulated by both cell-intrinsic and -extrinsic mechanisms: An integrated approach to understanding targeted therapy. *PLOS Biol.* **16:** e2002930

[22] Letort G, Montagud A, Stoll G, Heiland R, Barillot E, Macklin P, Zinovyev A & Calzone L (2018) PhysiBoSS: a multi-scale agent-based modeling framework integrating physical dimension and cell signalling. *Bioinformatics* Available at: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty766/5087713 [Accessed February 27, 2019]

[23] Müssel C, Hopfensitz M & Kestler HA (2010) BoolNet -- an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics 26: 1378–1380

[24] Naldi A, Hernandez C, Abou-Jaoudé W, Monteiro PT, Chaouiya C & Thieffry D (2018) Logical Modeling and Analysis of Cellular Regulatory Networks With GINsim 3.0. Front. Physiol. 9: 646

[25] Ozik J, Collier N, Wozniak JM, Macal C, Cockrell C, Friedman SH, Ghaffarizadeh A, Heiland R, An G &

Macklin P (2018) High-throughput cancer hypothesis testing with an integrated PhysiCell-EMEWS workflow. *BMC Bioinformatics* **19:** 483
[26] Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, Beqiri M, Sproesser K, Brafford PA, Xiao M, Eggan E, Anastopoulos IN, Vargas-Garcia CA, Singh A, Nathanson KL, Herlyn M & Raj A (2017) Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546:** 431–435
[27] Stoll G, Viara E, Barillot E & Calzone L (2012) Continuous time Boolean modeling for biological signaling: application of Gillespie algorithm. BMC Syst. Biol. 6: 116
[28] Stoll G, Caron B, Viara E, Dugourd A, Zinovyev A, Naldi A, Kroemer G, Barillot E & Calzone L (2017) MaBoSS 2.0: an environment for stochastic Boolean modeling. Bioinformatics 33: 2226–2228
[29] Tartaglia GG, Pechmann S, Dobson CM & Vendruscolo M (2009) A relationship between mRNA expression levels and protein solubility in E. coli. *J. Mol. Biol.* **388:** 381–389
[30] Thomas F, Donnadieu E, Charriere GM, Jacqueline C, Tasiemski A, Pujol P, Renaud F, Roche B, Hamede R, Brown J, Gatenby R & Ujvari B (2018) Is adaptive therapy natural? *PLOS Biol.* **16:** e2007066
[31] Trisilowati & Mallet DG (2012) In Silico Experimental Modeling of Cancer Treatment. *ISRN Oncol.* **2012:** Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3302021/ [Accessed March 22, 2019]
[32] Tyson JJ, Chen KC & Novak B (2003) Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.* **15:** 221–231
[33] Tyson JJ & Novák B (2015) Models in biology: lessons from modeling regulation of the eukaryotic cell cycle. *BMC Biol.* **13:** 46
[34] Wittmann DM, Krumsiek J, Saez-Rodriguez J, Lauffenburger DA, Klamt S & Theis FJ (2009) Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. *BMC Syst. Biol.* **3:** 98

11 Glossary

- **AGS:** is the name of a cell line derived from gastric adenocarcinoma used as a laboratory model to study the aforementioned disease (https://web.expasy.org/cellosaurus/CVCL_0139).
- **Apoptosis:** or programmed cell death that occur in multicellular organisms.
- **Cell cycle:** or cell-division cycle, is the series of events that take place in a cell leading to duplication of its DNA (DNA replication) and division of cytoplasm and organelles to produce two daughter cells[6].
- **Drug synergy:** positive interaction between two or more drugs that causes the total effect of the drugs to be greater than the sum of the individual effects of each drug.
- **Necrosis:** traumatic cell death caused by acute cellular injury or starvation (lack of nutrients).
- **Resistant cell:** cell that become immune to a specific drug or treatment.
- **Signalling pathway:** is a set of proteins in a cell that work together to translate external signals and to control one or more cell functions, such as cell division or cell death[7]. The emergence of cancer cell is closely related to the deregulation of malfunctioning of specific signalling pathway.

[6] https://en.wikipedia.org/wiki/Cell_cycle
[7] https://en.wikipedia.org/wiki/Cell_signaling#Signaling_pathways

Appendices

Appendix A: Experimental Data Description

We have experimental data for single and combined drug screening in the AGS cell lines. Table A.1. describes the names of the different drugs as well their molecular target. For those drugs indicated with (*) we have time course experiments, i.e. the total number of cells at each time point (before and after the drug supply). For the other drugs we have experimental measurements of the IC50 for different concentrations of a single drug as well as combinations of two of drugs. While the time course experiments were only conducted on the AGS cell lines, the IC50 experiments were conducted in many different cell lines, listed in Table A.2.

Table A.1. Relation between drugs used in the experiments and the biological entity (protein/complex) target.

Abbreviation	Drug Name	Drug Effect	Protein/Complex Target
5Z	*(5Z)-7-oxozeaenol	inhibits	MAP3K7
AK	* AKTi-1,2	inhibits	AKT_f
BI	BIRB0796	inhibits	MAPK14
CT	CT99021	inhibits	GSK3_f
PD	*PD0325901	inhibits	MEK_f
PI	*PI103	inhibits	PIK3CA
PK	PKF118-310	inhibits	CTNNB1
JN	JNK Inhibitor XVI, JNK-IN-8	inhibits	JNK_f
D1	BI-D1870	inhibits	RSK_f
60	BI605906 (BIX02514)	inhibits	IKBKB
SB	SB-505124	inhibits	TGFBR1, ACVR1
RU	Ruxolitinib (INCB18424)	inhibits	JAK_f
D4	D4476	inhibits	CK1_f
F4	10058-F4	inhibits	MYC
SF	SF1670	inhibits	PTEN
ST	Stattic	inhibits	STAT3
G2	GSK2334470	inhibits	PDPK1
G4	GSK-429286	inhibits	ROCK1
P5	P 505-15 (PRT 062607)	inhibits	SYK

Table A.2. Cell line descriptors in which drug experiments were conducted and for which we have IC50 values available.

Cell line name	Cell line descriptor	Link
A498	Renal cell carcinoma	CVCL_0139
AGS	Gastric adenocarcinoma	CVCL_1056
COLO205	Colon adenocarcinoma	CVCL_0218
DU145	Prostate carcinoma	CVCL_0105
MDA-MB-468	Breast adenocarcinoma	CVCL_0419
SF295	Colon adenocarcinoma	CVCL_1690
SW620	Melanoma	CVCL_0547
UACC62	Melanoma	CVCL_1780

For the experiments performed on the AGS cell lines we also have gene expression level measured before (control) and after the drug/s supply. Figure A.1 shows a heatmap representing the expression level of those protein coding

genes that are included in the AGS signalling network model. The clustering on the left-hand side indicates the replicates of the experiments are coherent. The information on gene expression can be used to further calibrate the MSM and, in particular, it can be used to inform the cell cycle model.

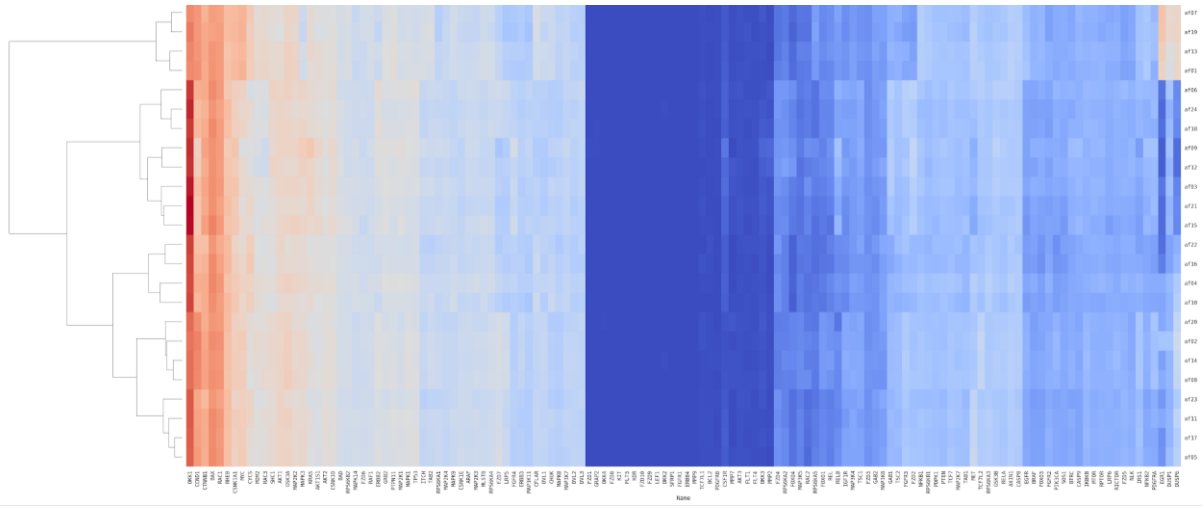

Figure A.1. Expression profile for those genes included in the signalling model corresponding to five different drug treatments and the control. Experiments were done in four replicates. Samples were clustered using a standard Hierarchical Clustering algorithm to check that they are clustered by their replicates (dendrogram on the left side of the plot).

Appendix B: Example of Simulation Results

For each run, PhysiBoSS builds one file per time point with the results of each of the studied variables. This semicolon-separated file has a header row with the names of the variables and one row for each agent (cells, in this case).

Example of output file for cells at time 0:

Time;ID;x;y;z;radius;volume_total;radius_nuclear;contact_ECM;freezer;polarized_fraction;motility;cell_line;Cell_cell;phase;Cycle;NFkB

0;0;-46.758;-10.7294;-85.806;10.0174;4210.69;6.02332;0;0;0.1;0.01;0;2.65594;0;0;-1

0;1;-46.5751;7.86895;-82.8054;9.81311;3958.3;5.90048;0;0;0.1;0.01;0;3.66865;0;0;-1

0;2;-31.2033;-37.4872;-84.2829;9.5;3591.36;5.71221;0;0;0.1;0.01;0;2.99975;1;0;-1

0;3;-30.9612;-19.2273;-82.8856;10.6359;5039.78;6.39521;0;0;0.1;0.01;0;4.65177;0;0;-1

0;4;-33.8193;-0.642819;-82.4852;9.78174;3920.46;5.88162;0;0;0.1;0.01;0;5.55423;0;0;-1

…

0;945;39.239;-19.5593;86.5458;10.4203;4739.46;6.26558;0;0;0.1;0.01;0;4.97028;0;0;-1

0;946;42.6192;-2.5138;88.2562;11.4447;6279.16;6.88153;0;0;0.1;0.01;0;4.00144;0;0;-1

0;947;42.7765;18.3806;88.1489;9.64008;3752.58;5.79644;0;0;0.1;0.01;0;2.5922;0;0;-1

Similarly, time-specific semicolon-separated files are built for the different densities (free-roaming molecules on the extracellular space, such as oxygen, signalling molecules, microenvironment density, etc). This are named after the density they represent.

An Example of an output file for microenvironment density at time 0 is given below. Note that there is no header row. The first three columns correspond to spatial coordinates and the fourth to the value of the density:

-417.5;-492.5;-492.5;0.0630239

-357.5;-492.5;-492.5;0.0630185

-267.5;-492.5;-492.5;0.0630086

-252.5;-492.5;-492.5;0.0630072

-237.5;-492.5;-492.5;0.0630059

-117.5;-492.5;-492.5;0.0630041

Appendix C: Expert User Questionnaires, Life Sciences Use Case: Original questions

1. User background information

- Company/organisation
- Professional position [and years of experience]
- Domain of expertise
- Background studies (university degree major, etc.)
- What are your main job tasks?
- To whom are you responsible for performing these tasks?

2. Existing workflow

- Please describe the different kinds of data sources that you use in your day-to-day tasks and the tools that you use to process them

Kind of data sources	Format	Volume of data (approx. MB, GB)	Purpose (task involved that generated the data)	Tools used to process the data*	Data analysis is automatic/ manual/ semi-automatic	Data gathering is historical/ real-time (online)	If real-time, update frequency

*If custom programs are used to process the data, please mention the programming language.

- Which is the aim of the analyses you perform (what kind of insights do you try to find)?
- What data processing challenges do you experience in your day-to-day tasks (e.g., fusion of heterogeneous sources, performance, analytics, scale of simulations, volume of simulation outputs, etc)?
- Provide examples of use/case studies/projects in which you have used these data.
- What problems do you run into in your day-to-day work when performing your data analysis?
    - Why is this a problem?
    - How do you currently solve the problem?
    - Is there a standard way of solving the problem or do you have a workaround?
    - How would you ideally like to solve the problem?
- Is any of the tools, mentioned in the table above, a must (one that no alternative execution on other tools/platforms would be allowed) for the use/case studies/projects that you describe?
- Are you able to program/set up a new/custom data processing workflow?
- How difficult is to program/set up a new data processing workflow?
- Are you capable of optimizing your data analysis operations?

- On what kind of infrastructure do you usually run the analysis (e.g. personal laptop, high spec workstation, cloud server, cluster, HPC, etc)

3. Specific aspects of the Life Sciences use case

- Do you work with real-time data?
- If not currently working with real-time data, what would you need to be able to work with such data?
- Would you be interested in using a data processing workflow that would allow using real-time data?
    - Would using this data allow you to address different problems than the ones you are currently addressing?
    - Would you be interested in applying real-time data processing workflows, such as those implemented in INFORE, to your projects/work?
- Have you ever worked with biological models? With multi-scale agent-based models?
    - To what extent do multi-scale agent-based simulations of tumoural cell growth relate to your projects/work?
    - Would you be interested in applying multi-scale agent-based simulations of tumoural cell growth, such as those implemented in INFORE, to your projects/work?
    - Rate the following components of the INFORE Life Sciences use case according to your interest on the foreseen results (1: Not useful, 2: Of some use, 3: Average Use, 4: Quite useful, 5: Very useful)
        - multi-scale models
        - Boolean models
        - cell cycle dynamics
        - drug synergies inference
        - real size tumour simulations
        - real-time/online data processing

4. Expected benefits from using INFORE

A. Please mention more data sources that would you like to use and why.

B. Which new information would you like to extract from all of the aforementioned data sources?

C. Are there specific events that you would like to forecast in real-time, which you currently cannot forecast?

D. Would you find it an acceptable trade-off to significantly speed up your data analysis tasks, if the provided output was a fairly accurate approximation of the correct result?

E. Rate the following objectives of INFORE, based on how useful they may be at YOUR data analysis (1: Not useful, 2: Of some use, 3: Average Use, 4: Quite useful, 5: Very useful)

- Ability to design data processing workflows with no code required
- Ability to change algorithm parameters graphically

- ○ Ability to receive quick approximate answers instead of 100% accurate, but long running queries
- ○ Ability to interactively explore the data in order to detect patterns/features of interest
- ○ Ability to accurately forecast events of interest
- ○ Ability to automatically optimize your data analysis task over different data processing platforms (HPC, Big Data Platforms, etc).

F. Rate the following objectives of INFORE, based on how useful they may be at the data analysis of OTHER data types in your organization (1: Not useful, 2: Of some use, 3: Average Use, 4: Quite useful, 5: Very useful)

- ○ Ability to design data processing workflows with no code required
- ○ Ability to change algorithm parameters graphically
- ○ Ability to receive quick approximate answers instead of 100% accurate, but long running queries
- ○ Ability to interactively explore the data in order to detect patterns/features of interest
- ○ Ability to accurately forecast (defined or currently unknown) events of interest
- ○ Ability to automatically optimize your data analysis task over different data processing platforms (HPC, Big Data Platforms, etc).

G. What are your expectations regarding the system usability?

H. What is the expected added value from INFORE for you and your corporation?

Appendix D: Expert User Questionnaires , Life Sciences Use Case: Anonymised answers

1. User background information

- Company/organisation:
  Expert user 1: Research Center, private and publicly funded
  Expert user 2: Public university
- Professional position [and years of experience]:
  Expert user 1: Research engineer [>15 years]
  Expert user 2: Professor [>20 years]
- Domain of expertise:
  Expert user 1: Systems Biology / Mathematical modelling
  Expert user 2: Systems biology/medicine, functional genomics, molecular biology, cancer biology
- Background studies (university degree major, etc.):
  Expert user 1: PhD in Biology
  Expert user 2: PhD in Biochemistry
- What are your main job tasks?
  Expert user 1: Construction of mathematical models
  Expert user 1: Analyses of models
  Expert user 1: Participation to the development of appropriate tools
  Expert user 1: Writing grants
  Expert user 1: Writing articles
  Expert user 2: Research group leader
  Expert user 2: University teacher
- To whom are you responsible for performing these tasks?
  Expert user 1: No answer
  Expert user 2: The university

2. Existing workflow

- Please describe the different kinds of data sources that you use in your day-to-day tasks and the tools that you use to process them

Kind of data sources	Format	Volume of data (approx. MB, GB)	Purpose (task involved that generated the data)	Tools used to process the data*	Data analysis is automatic/ manual/ semi-automatic	Data gathering is historical/ real-time (online)	If real-time, update frequency
Outputs of simulations	CSV files	MB	Model analysis / Mutant simulations / Drug effect predictions	Logical modeling (GINsim, MaBoSS, PhysiBoSS)	Manual	historical	
In lab and public data on cancer cell	Numeri cal (e.g.	In the high MBs	Observe cancer cell	Statistical tools (SPSS,	semi-automatic	historical	

line cellular and molecular (proteomic, transcriptomic) responses to drug treatment (single and combination); Knowledge bases with data on molecules, interaction, and function (e.g. Uniprot, Intact, reactome, SIGNOR, Gene ontology annotation; etc)	cell viability), qualitative (nucleotide sequence, protein/ peptide composition)		drug response for testing of predictions from simulations with computational models	Matlab, R-tools) Functional overrepresentation (e.g. BINGO, DAVID) PARADIGM (infer protein state from transcriptomics and genomics)			
Cellular responses are mainly cell viability and are measured mainly in robotic high throughput systems; Molecular data are both large scale (NGS, MS) and small scale (e.g. western blot)	Numerical (e.g. cell viability), qualitative (nucleotide sequence, protein/ peptide composition)	In the high MBs	Molecular data: biomarkers for configuring computational models; testing of molecular hypotheses generated by models	Statistical tools (SPSS, Matlab, R-tools) Functional overrepresentation (e.g. BINGO, DAVID) PARADIGM (infer protein state from transcriptomics and genomics)	semi-automatic	historical	

*If custom programs are used to process the data, please mention the programming language.

- Which is the aim of the analyses you perform (what kind of insights do you try to find)?
  Expert user 1: Reproduce experimental data
  Expert user 1: Use model outputs to stratify cancer patients
  Expert user 1: Predict outcomes of drug treatments
  Expert user 2: Insights that can help us build computational models predictive of individual cancer patient drug response and resistance

- What data processing challenges do you experience in your day-to-day tasks (e.g., fusion of heterogeneous sources, performance, analytics, scale of simulations, volume of simulation outputs, etc)?
  Expert user 1: Performance: current computation time spent on laptop is too long
  Expert user 1: Parameter fitting is an important aspect of our research and the current tools do not allow a comprehensive search of parameter search
  Expert user 2: Integration data from heterogeneous sources
  Expert user 2: Inference of state of one type of molecular entities in the cell (proteins) from other types of molecular entities in the cells (transcripts, genome)

- Provide examples of use/case studies/projects in which you have used these data.
  Expert user 1: All publications we have recently worked on use MaBoSS or PhysiBoSS: PMID: 30733688;

PMID:30169736; PMID:29237040
Expert user 2: Predict response of cancer cell lines (colorectal, breast, prostate) to targeted anti-cancer drugs (single and in combination)

- What problems do you run into in your day-to-day work when performing your data analysis?
  Expert user 1: Time for each simulation
  Expert user 2: Difficult to A) integrate large scale molecular data from heterogeneous sources; B)infer protein state from transcriptomics and genomics
    - Why is this a problem?
      Expert user 1: The model we use are bigger and bigger (with high number of nodes, which requires more time for simulations).
      Expert user 2: A) standardization of data format is not optimal across data sources; B) science lacks sufficient insight into the relationship between protein-transcript-gene and problem B) may be impossible to solve without additional data.
    - How do you currently solve the problem?
      Expert user 1: We use the clusters made available by our research center.
      Expert user 2: A) extra work (partially manual) to align data; B) available computational tool (PARADIGM).
    - Is there a standard way of solving the problem or do you have a workaround?
      Expert user 1: No. I usually wait for the simulations to be done on my computer.
      Expert user 2: No answer.
    - How would you ideally like to solve the problem?
      Expert user 1: Not sure
      Expert user 2: Standardized data format and interconversions interlinks between data and knowledge sources; improved computational tools for inferring protein states from transcriptomics and genomics, possibly by taking additional data and knowledge into consideration.
- Is any of the tools, mentioned in the table above, a must (one that no alternative execution on other tools/platforms would be allowed) for the use/case studies/projects that you describe?
  Expert user 1: There are other tools but all will meet the same issues of model sizes at some point.
  Expert user 2: Yes, all are a must.
- Are you able to program/set up a new/custom data processing workflow?
  Expert user 1: Yes, when we meet new problems, we create the appropriate tools or try to optimize the ones we use.
  Expert user 2: Yes, group members are able to.
- How difficult is to program/set up a new data processing workflow?
  Expert user 1: It is mainly time consuming
  Expert user 2: No answer.
- Are you capable of optimizing your data analysis operations?
  Expert user 1: Yes, we try as often as possible, but again, this is time consuming.
  Expert user 2: Yes, group members are able to.
- On what kind of infrastructure do you usually run the analysis (e.g. personal laptop, high spec workstation, cloud server, cluster, HPC, etc)
  Expert user 1: Laptop / workstation / cluster
  Expert user 2: Personal laptops, high specs workstations, work servers.

3. Specific aspects of the Life Sciences use case

- Do you work with real-time data?
  Expert user 1: No.
  Expert user 2: No.
- If not currently working with real-time data, what would you need to be able to work with such data?
  Expert user 1: The technology is not available yet. No real way to keep track of protein concentration as a function of time.
  Expert user 2: More funding.
- Would you be interested in using a data processing workflow that would allow using real-time data?
  Expert user 1: If technology would be there to gather real-time biological data, definitely yes. Currently, we could only use simulated real-time data and, to do this, we would need tools that are not currently in the market.
  Expert user 2: Yes.
    - Would using this data allow you to address different problems than the ones you are currently addressing?
      Expert user 1: Yes, definitely. Even more if these would be biological data (protein variations in time, etc).
      Expert user 2: Maybe different, definitely better address current problems.
    - Would you be interested in applying real-time data processing workflows, such as those implemented in INFORE, to your projects/work?
      Expert user 1: Yes.
      Expert user 2: Yes.
- Have you ever worked with biological models? With multi-scale agent-based models?
  Expert user 1: Yes
  Expert user 2: We work with logical, Boolean models of cancer cell signalling underlying cancer cell growth; multiscale in the sense that they predict cellular outcome (viability) based on molecular mechanisms and information; not agent-based.
    - To what extent do multi-scale agent-based simulations of tumoural cell growth relate to your projects/work?
      Expert user 1: Completely in line with what we do
      Expert user 2: Very relevant and necessary.
    - Would you be interested in applying multi-scale agent-based simulations of tumoural cell growth, such as those implemented in INFORE, to your projects/work?
      Expert user 1: Yes.
      Expert user 2: Yes.
    - Rate the following components of the INFORE Life Sciences use case according to your interest on the foreseen results (1: Not useful, 2: Of some use, 3: Average Use, 4: Quite useful, 5: Very useful) Expert user 1; Expert user 2
        - multi-scale models: 5; 5
        - Boolean models: 5; 5
        - cell cycle dynamics: 5; 5
        - drug synergies inference: 5; 5

- real size tumour simulations: 4; 5
- real-time/online data processing: 3; 5

4. Expected benefits from using INFORE

A. Please mention more data sources that would you like to use and why.
Expert user 1: Other types of biological data, such as CRISPR data that we would use to calibrate the parameters for our models
Expert user 2: Real-time cancer cell growth in vitro and in vivo (xenografts) for improved capturing of drug response.
Expert user 2: High coverage, high quality (specific, dynamic, sensitive) proteomics data including post.

B. Which new information would you like to extract from all of the aforementioned data sources?
Expert user 1: Constraints for modeling tasks.
Expert user 2: Time-dependent cancer cell responses to drugs.
Expert user 2: Accurate understanding of proteome in biological condition of interest.

C. Are there specific events that you would like to forecast in real-time, which you currently cannot forecast?
Expert user 1: Evolution of protein concentration over time in real tissues.
Expert user 2: In vitro and in vivo (xenograft) cancer cell viability/growth and response to drugs (single and combinations)

D. Would you find it an acceptable trade-off to significantly speed up your data analysis tasks, if the provided output was a fairly accurate approximation of the correct result?
Expert user 1: Yes
Expert user 2: Yes.

E. Rate the following objectives of INFORE, based on how useful they may be at YOUR data analysis (1: Not useful, 2: Of some use, 3: Average Use, 4: Quite useful, 5: Very useful) Expert user 1; Expert user 2

- Ability to design data processing workflows with no code required: 5; 5
- Ability to change algorithm parameters graphically: 4; 5
- Ability to receive quick approximate answers instead of 100% accurate, but long running queries: 4; 5
- Ability to interactively explore the data in order to detect patterns/features of interest: 4; 5
- Ability to accurately forecast events of interest: 4; 5
- Ability to automatically optimize your data analysis task over different data processing platforms (HPC, Big Data Platforms, etc): 4; 5

F. Rate the following objectives of INFORE, based on how useful they may be at the data analysis of OTHER data types in your organization (1: Not useful, 2: Of some use, 3: Average Use, 4: Quite useful, 5: Very useful) Expert user 1; Expert user 2

- Ability to design data processing workflows with no code required: 5; 5
- Ability to change algorithm parameters graphically: 4; 5
- Ability to receive quick approximate answers instead of 100% accurate, but long running queries: 4; 5
- Ability to interactively explore the data in order to detect patterns/features of interest: 4; 5

- Ability to accurately forecast (defined or currently unknown) events of interest: 4; 5

- Ability to automatically optimize your data analysis task over different data processing platforms (HPC, Big Data Platforms, etc): 4; 5

G. What are your expectations regarding the system usability?
Expert user 1: If we were to use it, it should be visual and user-friendly.
Expert user 2: No answer.

H. What is the expected added value from INFORE for you and your corporation?
Expert user 1: Speed up computation time.
Expert user 2: Increased insight from in vitro and in vivo cancer models that would improve functionality of predictive models also for tumour patients.