



OPEN

# A biochemical network modeling of a whole-cell

Paulo E. P. Burke<sup>1✉</sup>, Claudia B. de L. Campos<sup>2</sup>, Luciano da F. Costa<sup>3</sup> & Marcos G. Quiles<sup>2</sup>

All cellular processes can be ultimately understood in terms of respective fundamental biochemical interactions between molecules, which can be modeled as networks. Very often, these molecules are shared by more than one process, therefore interconnecting them. Despite this effect, cellular processes are usually described by separate networks with heterogeneous levels of detail, such as metabolic, protein–protein interaction, and transcription regulation networks. Aiming at obtaining a unified representation of cellular processes, we describe in this work an integrative framework that draws concepts from rule-based modeling. In order to probe the capabilities of the framework, we used an organism-specific database and genomic information to model the whole-cell biochemical network of the *Mycoplasma genitalium* organism. This modeling accounted for 15 cellular processes and resulted in a single component network, indicating that all processes are somehow interconnected. The topological analysis of the network showed structural consistency with biological networks in the literature. In order to validate the network, we estimated gene essentiality by simulating gene deletions and compared the results with experimental data available in the literature. We could classify 212 genes as essential, being 95% of them consistent with experimental results. Although we adopted a relatively simple organism as a case study, we suggest that the presented framework has the potential for paving the way to more integrated studies of whole organisms leading to a systemic analysis of cells on a broader scale. The modeling of other organisms using this framework could provide useful large-scale models for different fields of research such as bioengineering, network biology, and synthetic biology, and also provide novel tools for medical and industrial applications.

Cells are mainly composed of water, proteins, nucleic acids, metabolites, and an enveloping lipid membrane. However, what makes a cell alive are the interactions between these components. Chemical reactions interconnect molecules into intricate biochemical networks in order to perform particular tasks. The whole set of possible chemical interactions is meticulously regulated in order to maintain cellular function, growth, and replication<sup>1</sup>.

Biochemical interactions with related functions have been traditionally grouped into cellular processes<sup>2,3</sup>. Among these processes, metabolism<sup>4,5</sup>, signaling<sup>6–8</sup>, and transcription regulation are those most frequently described and modeled as networks<sup>9–12</sup>. Though cellular processes are often described in terms of separate networks<sup>13,14</sup>, they are neither physically nor functionally independent<sup>15</sup>. The simple fact that molecular species are shared between them makes their dynamics dependent on each other. For instance, the intracellular concentration of the energetic molecule ATP affects and is affected by, several processes simultaneously.

The ever-growing data availability regarding cellular biology has implied in ever more comprehensive computational models of cells<sup>16,17</sup>. Having pursued metabolic, signaling, and gene regulation networks toward the limit, we are now interconnecting these data<sup>18–21</sup>. Bolder approaches are aiming at complete representations of unicellular organisms<sup>22–24</sup>, taking into account all known cellular processes, whether it is network-oriented or not. For example, Karr *et al.*<sup>23</sup> performed whole-cell simulations of the organism *Mycoplasma genitalium* using a hybrid simulation approach. Briefly, cellular processes are simulated independently using diverse simulation approaches, and information is interchanged only from time to time. Such whole-cell approaches are paving the way to enhancing the pace of advances toward the understanding of cellular behavior, favored by their capacity to predict cell's phenotype from genotype with good accuracy<sup>25</sup>.

Despite the promising achievements of the above mentioned whole-cell approaches, much remains to be done towards achieving reference modeling frameworks<sup>26</sup>. Current models are tailored for respective organisms,

<sup>1</sup>University of São Paulo, Bioinformatics Graduate Program, São Carlos, SP, Brazil. <sup>2</sup>Institute of Science and Technology, Federal University of São Paulo, São José dos Campos, SP, Brazil. <sup>3</sup>São Carlos Institute of Physics, University of São Paulo, São Carlos, SP, Brazil. ✉email: pauloepburke@gmail.com

leaving no straightforward means to adapt them for other species. The hybrid approach to modeling and simulation also has its limitations. Additionally to integration issues in simulations, which are out of scope for this work, it is difficult to have a broader view of the biological system of interest once relationships between molecules are encapsulated within several distinct mathematical and computational modules. At the same time, it is not an easy task to describe such modules using community standards such as SBML<sup>27</sup>.

Assuming that all cellular processes can be approached by their underlying biochemical interactions, we can ultimately integrate them by using reactions that share the same cellular substrate, such as given proteins, DNA, and metabolites. Thus, there should be an approach allowing all cellular components and known interactions to be accommodated into a single biochemical network. Such a framework could then provide a whole picture of the possible molecular interactions inside a cell at any time of a cellular cycle. Moreover, the role of given molecules could then be addressed in a systemic context, revealing interfaces between cellular processes.

In this work, we aim at modeling and integrating several cellular processes into a single biochemical network, hence providing a more homogeneous framework to model whole organisms. We propose a rule-based modeling approach to represent a diverse variety of interactions between molecular species that can be found within a cell. The set of molecules and reactions composes a biochemical network where molecules are linked to reactions according to their role as reactants or products. In addition, we explicitly link to reactions the molecules that act as their catalysts or regulators. For the sake of generalization, we hence call this kind of relationship as *modifiers*, following the nomenclature already implemented in SBML<sup>28</sup>. The stoichiometry of the interactions is described as weights associated with the links. To model cellular processes that are not usually represented as networks, we built sets of reactions (templates) that can be replicated for different substrates such as the transcription of several genes, or the translation of several mRNAs.

As a study case, we modeled biochemical and genomic data about all the known processes of the bacterium *Mycoplasma genitalium* resulting in what we call a “whole-cell biochemical network” of the organism. We choose this organism because of its relative simplicity, corresponding to the smallest known genome, and for having respective integrated data at the whole-cell level deposited in the WholeCellKB<sup>16</sup>. The network contains information about 15 cellular processes, naturally integrated into a single component. To validate the model, we perform cascading failure analysis to predict essential genes of *M. genitalium* and compare the obtained results with experimental data from global transposon mutagenesis gene disruption<sup>29</sup>.

We believe that the presented approach can pave the way to more scalable and adaptable whole-cell models while providing a more homogeneous basis for whole-cell analysis and simulations. Other interesting prospects could be explored in the context of constraint-based methods<sup>30,31</sup>, synthetic biology<sup>32,33</sup>, and network science<sup>34–36</sup>.

## Modeling cellular processes through reactions

All cellular processes can be understood by means of their underlying biochemical reactions. When trying to do so, one can face two related problems: the lack of sufficiently detailed information and the highly combinatorial nature of biomolecular interactions. By far, the most detailed cellular process in terms of specific biochemical reactions is the metabolism<sup>37</sup>. For many organisms, we already have very detailed and maybe complete metabolic maps<sup>38</sup>. Even though, metabolism is just a small set of reactions among many others that composes other cellular processes. One step further from metabolic interactions we can find their proteic catalysts, known as enzymes. Proteins such as enzymes can often be modified in specific sites, therefore modulating their activity as catalysts and also their interaction with other proteins. The number of different states for each protein in the organism and the complexes they can form grows very fast thus bringing us to the combinatorial problem.

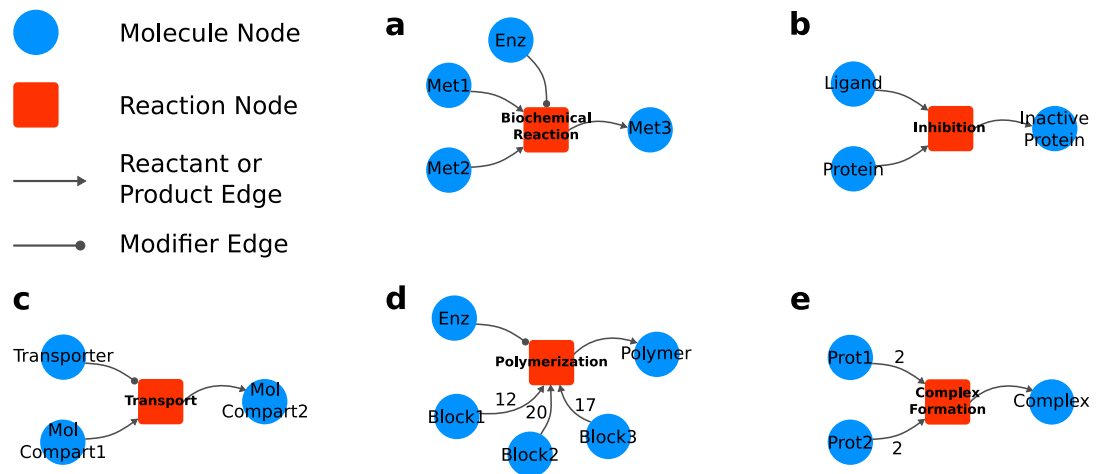
One approach to model the high number of states and interactions of proteins is the rule-based modeling<sup>39</sup>. In brief, this method makes use of generic rules representing molecular interactions that can be applied to many substrates. By doing so, there is no need for the modeler to manually write all the possible interactions among substrates, which can be now automatically generated by the so specified rules. Rule-based modeling is mainly applied to study the signaling process in cells<sup>40,41</sup>. The high number of protein modifications and complex formations lead us to the first-mentioned problem: the lack of detailed information regarding which of these interactions really occurs.

While the interaction of metabolism and protein-protein interactions have already been tackled in the literature<sup>42</sup>, other cellular processes such as protein and RNA synthesis, DNA replication, protein and RNA degradation, and cell cycle, are rarely integrated to these models at the biochemical level. Thus, in the next sections, we will make use of the rule-based modeling principles allied to whole-cell scale databases to propose representations of the most diverse type of molecules and interactions possible of given organisms. Our approach aims at expanding the rule-based modeling to encompass all known cellular processes integrating them into a single reaction network.

**Reaction modeling framework.** To better explain our modeling approach, a simple graphical notation will be used to depict molecules and interactions in the form of a network. Thus, using network science nomenclatures, representations of molecules and reactions will be called as *nodes*. Relationships between molecules and reactions will be called as *edges*. Figure 1 shows the graphical notation adopted in this work.

In our framework, molecule nodes can represent any physical entity within a cell such as metabolites, DNA regions, proteins, and RNAs. Each different state of a molecule (e.g., the active and inactive form of a protein) will be represented by different molecule nodes. Similarly, molecules in different cellular compartment locations (e.g., intra and extracellular glucose) are also represented by distinct nodes.

We can represent any interaction between molecules by a reaction node. In addition to biochemical reactions, such as in metabolism, more complex interactions such as gene transcription, protein synthesis, transport, protein complex formation, chromosome replication, and cell division can be incorporated in the model.



**Figure 1.** Graphical representation of the reaction modeling framework. Molecule nodes are represented by blue circles. Red squares represent reaction nodes. Arrow-ended edges indicate reactant or product relationships depending on direction. Circle-ended edges indicate modifiers to the reaction. Some modeling examples of distinct kinds of biochemical interactions are depicted as follows. **(a)** A biochemical reaction  $\text{Met1} + \text{Met2} \xrightarrow{\text{Enz}} \text{Met3}$ ; **(b)** inhibition of a protein by a ligand; **(c)** molecular transport from compartment 1 to compartment 2 by a transporter; **(d)** a polymerization reaction; **(e)** a tetrameric protein complex formation. All detailed network visualizations in this and further Figures were generated using Cytoscape<sup>43</sup>.

Reactions can be regulated and we can explicitly link the “modifiers” to their respective reaction nodes. To do so, we use a distinct type of edge, called *modifier edge*. In Fig. 1a, the modifier edge is drawn as a circle-ended line connecting molecule “Enz” to the reaction. This connection means that the Enzyme is needed so that the reaction can occur, but it is neither consumed nor produced in the reaction. For example, other molecules such as transcription factors, genes, and mRNAs can act as modifiers, since their concentration does not change in some reactions.

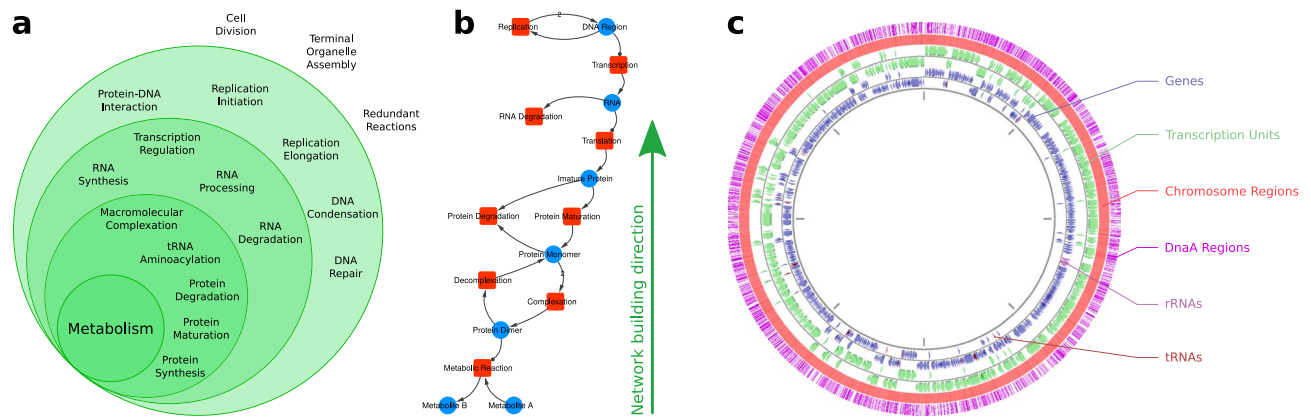
To illustrate the modeling of some molecular interactions, Fig. 1 shows some use cases derived from the generalizations we adopted. Example (a) depicts a biochemical reaction where an enzyme combines Met1 and Met2 into Met3. In (b), a given protein interacts with a ligand (e.g., in an allosteric site) producing the protein’s inhibited form. Transport through a cellular membrane can be approached as in example (c), where a given molecule is carried from extra to the intracellular environment by a transporter protein. Example (d) illustrates a polymerization process catalyzed by an enzyme where different stoichiometries of three basic building blocks are combined into a single polymer, such as in protein synthesis reaction catalyzed by a ribosome. Protein complexation of four subunits (e.g.,  $\alpha$  and  $\beta$  hemoglobin subunits) is represented in example (e).

**Process templates.** Certain cellular processes make use of the same molecular machinery to produce different outputs given different inputs. It is the case of the synthesis of macromolecules such as proteins, RNAs, and DNAs. For instance, the synthesis of proteins involves a set of reactions that repeats for each mRNA given as an input. Most of the molecules involved in the process are the same for different mRNAs, sometimes only changing the stoichiometry of amino acids. For such processes, we created templates which are sets of reactions that can be replicated and adapted for different substrates in the same sense of rule-based modeling. Particularities of each reaction can be incorporated if data is available. It is important to notice that the templates generated in this work are made specifically to the organism used as an example and the modeling assumptions can vary from modeler to modeler.

### *Mycoplasma genitalium* case study

Recently, a public database called WholeCellKB was implemented aiming at gathering complete biological information about specific organisms<sup>16</sup>. The first organism deposited in this database was the pathogenic bacterium *Mycoplasma genitalium*. This organism yields the smallest genome known, with 580 kb and 525 genes. Because of its relative simplicity, *M. genitalium* has served as the model organism for breakthrough advances in synthetic biology<sup>29</sup> and whole-cell simulations<sup>23</sup>. The simplicity of this organism, allied to the structured data provided by the database, makes *M. genitalium* a particularly suitable model for the comprehensive integration of cell-scale biochemical interaction into a whole-cell biochemical network.

Using the proposed framework, we modeled almost all the biochemical interaction information contained in WholeCellKB about *M. genitalium*. The database accounted for interactions for the following processes: DNA Replication Initiation and Elongation, DNA Damage, DNA Repair, Cellular Division, Transcription, RNA Processing, tRNA Aminoacylation, Translation, Protein Modification and Complexation, Metabolism, Transmembrane Transport, Host Interaction, and RNA and Protein Decay. The only process not taken into account was the DNA Damage, which we could not find an appropriate means to represent in our model because of its high combinatorial nature. Also, RNA modification reactions were not incorporated in the model due to systematic inconsistencies found in the WholeCellKB regarding these reactions. Some processes described in Karr’s work<sup>23</sup>



**Figure 2.** The hierarchical construction of the *Mycoplasma genitalium*'s whole-cell biochemical network. Starting from the metabolism several cycles of introducing necessary molecules to the reaction and their synthesis/degradation reactions. A simplified version of a full cycle of network building starting from a single metabolic reaction is depicted in **b**. When the network construction reach the DNA level, the chromosome is divided in regions in order to better represent locus-specific protein interactions. Figure **c** was partially generated using CGView<sup>44</sup>.

were encompassed by major processes in our model. For example, processes like “Protein Folding” and “Protein Processing” were incorporated by “Translation”. The “Ribosome assembly”, and “FtsZ ring polymerization” were incorporated by “Macromolecular Complexation”. The “Chromosome condensation” was incorporated by “Protein-DNA Interaction”. The “Chromosome segregation”, and “Cytokinesis” were incorporated by “Cellular Division”.

**Network building process.** To start building the biochemical network of the *M. genitalium* organism, we first queried all metabolic reactions from WholeCellKB and included them as reaction nodes in the model. Then, all metabolites which act as reactant or product, as well as enzymes, were incorporated in the network as molecule nodes and properly linked to their respective reactions. From this point, a recursive process starts by adding biosynthesis and degradation reactions for each molecule already in the network. For example, protein complexes are biosynthesized by macromolecular complexation reactions. Protein monomers are produced by translation and protein modification reactions. They also are degraded by proteolysis. For these newly incorporated reactions, the necessary molecules are added as molecule nodes and linked to them. This process repeats itself until all molecules have their respective biosynthesis and degradation reactions. At the end of this recursive process, it is expected the network to have all reactions from metabolism to DNA Replication, passing through all the central dogma of biology (Fig. 2a). A simplified example of a full pathway from DNA to metabolite is depicted in Fig. 2b a homodimeric enzyme that catalyzes a given metabolic reaction has its biosynthesis pathway built upwards to the DNA level passing through protein complexation, protein maturation, protein synthesis, RNA synthesis, and DNA duplication. Degradation reactions for proteins and RNA are also shown.

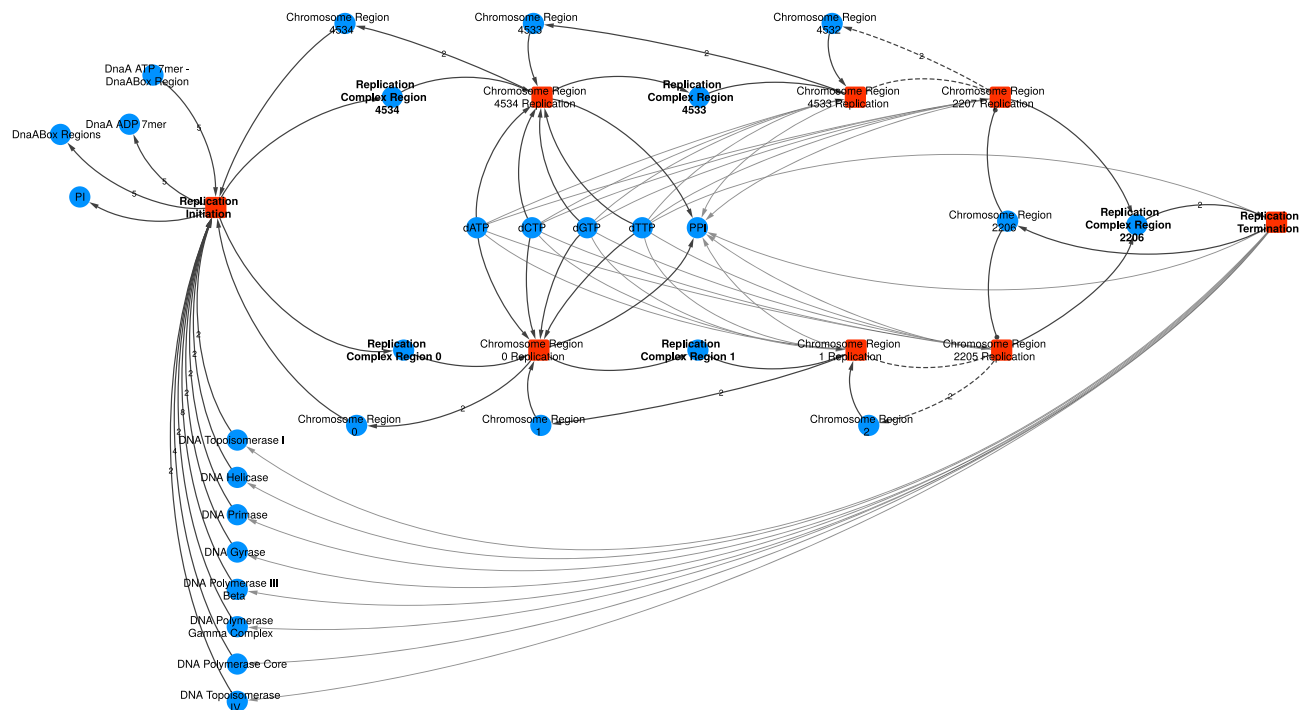
On top of the so obtained network, we queried the WholeCellKB for reactions that are still not included in the model, such as redundant reactions, and added them. To finalize the network, we manually included the “Cell Division Reaction” to which all necessary proteic complexes, such as FtsZ Ring, Chromosome Segregation proteins, and the duplicated Chromosome, are linked as modifiers. A detailed description can be found in the Supplementary Information.

In order to guarantee the coherence of the reactions in the network, we calculate the mass-balance for all reactions. Following the principle of mass conservation, the difference between the mass of reactants and products, weighted by their respective stoichiometry, should be zero. Although this approach can assure the mass-balance, literature evidence is still needed to ensure their correctness.

**Chromosome representation.** Despite being the smallest known chromosome, *M. genitalium*'s still a large and lengthy circular molecule having 580kb. In order to better represent locus-specific protein interactions, we divided the chromosome into regions, each one being represented by a molecule node.

We used as reference the *M. genitalium* G37 genome<sup>45</sup>, available at the NCBI database (NC\_000908.2). In Fig. 2c we can observe the circular chromosome representation as well as the genes distributed along with it. The transcription units (TUs) are the regions that are transcribed in RNA. One TU can encompass one or more genes, the last case also being called “Polycistronic RNAs”. The RNAs from TUs with more than one gene can be further cleaved into separated RNA molecules, which is the case of tRNAs, or left as one molecule. In any case, each RNA molecule, polycistronic or not, is represented by a single molecule node.

Although intuitive approaches would be to divide the chromosome regions according to the TUs or even to split into regions of the same length, the more granular the division, more details can be incorporated in the model. The DnaA protein interacts with small 8 nucleotides length sequences repeated all over the chromosome. Figure 2c depicts the distribution of DnaA binding sites annotated in the WholeCellKB. The binding and polymerization of this protein at specific nucleotide sequences in the chromosome are the main mechanisms



**Figure 3.** The template for Replication Initiation and Replication Elongation processes. The network depicted shows three steps of Replication Elongation going both chromosome directions. The dashed lines mask the rest of the process and go straight to the final reaction, where the terminus region (2206) is replicated and the Replisomes released. The actual molecules' IDs in the network are listed in the Supplementary Information. Although some stoichiometries are shown, most of them are hidden for better visualization.

to control cellular replication and the binding sites present a more granular division of the chromosome. Thus, we adopted the DnaA binding sites as division points to define the chromosome regions, with the addition of the replication origin and terminus sites. More specifically, each DnaA binding site is defined as a chromosome region, and each region between DnaA binding sites is also defined as a chromosome region.

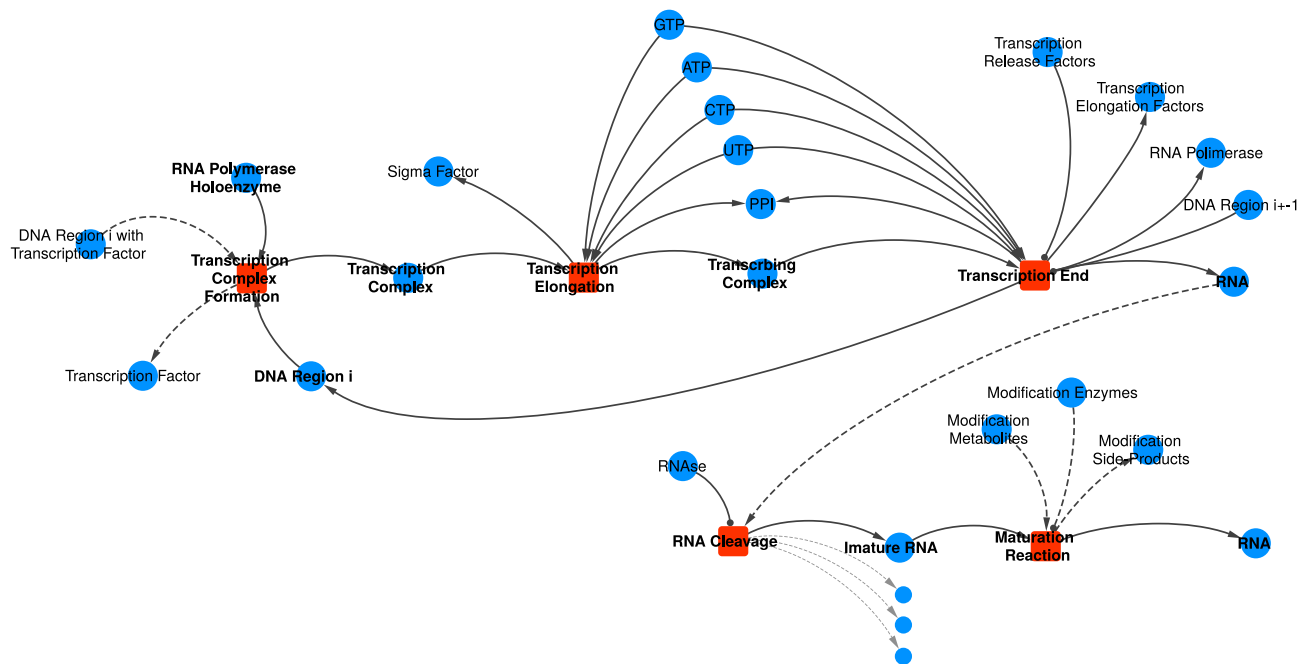
**Modeling canonical processes.** Although some processes, such as metabolism, have a straightforward modeling transition from WholeCellKB to the proposed framework, other processes require more attention. For instance, it is not usual to describe chromosome replication, gene transcription, protein synthesis, and some other processes as networks. Thus, we manually created templates based on literature for these processes. Particularities in the synthesis process of individual proteins, RNAs, and DNA are incorporated accordingly to data availability in WholeCellKB.

**Chromosome replication.** The chromosome's replication starts when the DnaA protein polymerizes in five specific DnaA binding regions near the replication origin and recruits all necessary molecular machinery to replicate the DNA. It is the formation of the two Replisomes at the origin of replication in the chromosome, which we call the Replication Complexes.

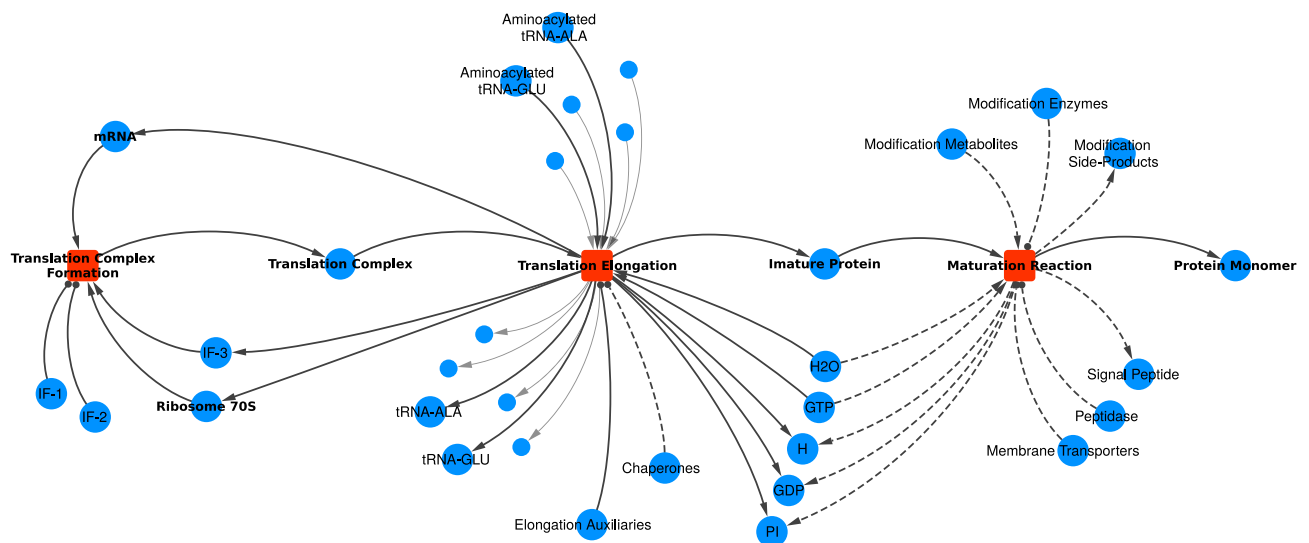
Given that the chromosome is divided into regions, the formation of the Replication Complex also takes as a reactant a chromosome region. The two Replisomes, bound to the Chromosome Regions 0 and 4534, undergo their respective replication reactions where the free deoxy-nucleotides are consumed according to the regions' sequence. Each replication reaction produces two copies of the current Chromosome Region and consumes the next region, making the Replisome move through the DNA molecule (Fig. 3). The two Replication Complexes move in opposite directions until they reach the replication terminus region, where the replication completes and the Replisomes' subunits are released. The collision of the Replisome with other Protein-DNA complexes, such as DnaA and RNA Polymerases, are handled separately.

**RNA synthesis.** The RNA Synthesis is the process in which an RNA Polymerase makes an RNA molecule based on a Transcription Unit (TU), a region of the chromosome that may contain one or more genes. Given that a TU can be fragmented in several Chromosome Regions, as observed in Fig. 2c, the transcription process follows a similar approach to the Chromosome Replication, once the Transcription Complex moves through the DNA and its template is shown in Fig. 4.

The binding of the RNA Polymerase Holoenzyme to the beginning of a TU in the chromosome might require a Transcription Factor already bound to the respective Chromosome Region, as indicated with dashed links to the "Transcription Complex Formation" reaction in Fig. 4. Then, given that the TU begins at the Chromosome Region  $i$ , the Transcription Complex, the example in Fig. 4 depicts a TU divided into only two Chromosome



**Figure 4.** The template for the Translation process. Dashed lines indicate relationships that are not always required. For example, not all RNAs goes through RNA Cleavage and Maturation reactions, such as tRNAs. Most of the RNA modifications available in WholeCellKB were not included in the model due to systemic inconsistencies found in the database regarding the positions of the modifications.

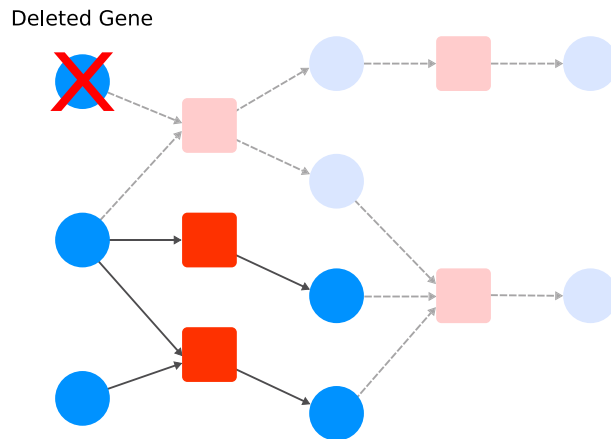


**Figure 5.** The template for Translation and Protein Maturation processes. Dashed lines indicate relationships that are not always required. For example, not all proteins require Chaperones. Membrane Transporters, Peptidase, and the production of the Signal Peptide are only required for secreted proteins. Only two of the twelve amino acids are shown to better visualization. The small blue circles illustrate the other amino acids and associated tRNAs.

Regions, being the second one at the position  $i + 1$  or  $i - 1$  depending on which strand of the DNA the TU is found. TUs that are polycistronic and need to be cleaved to produce the individual functional molecules, the so transcribed RNA goes through further cleavage and maturation reactions.

**Protein synthesis.** The template for the translation process is shown in Fig. 5, including the translation complex formation, translation elongation, and protein maturation.

The translation process starts with the complexation of the Ribosome 70S, Initiation Factor (IF) 3, and the mRNA with IF-1 and IF-2 as auxiliary molecules. Then, the translation complex proceeds to the elongation stage



**Figure 6.** Example of cascading failure. The nodes with shaded color and dashed edges were removed after cascading failure by removing the initial gene molecule node.

where aminoacylated tRNAs and energy molecules (GTPs) are consumed according to the mRNA sequence while the respective tRNAs without amino acids and the IF-3 are released. Elongation auxiliary proteins act as modifiers and chaperones might also be linked depending on the protein's annotations in WholeCellKB. Similarly, post-translational modifications can be incorporated in the Maturation Reaction, transforming an immature form of the protein into the functional one. Proteins that are secreted to the external environment have also linked to the Maturation Reaction the necessary membrane transporters, the peptidase to cleave the Signal Peptide, and the Signal Peptide itself.

A more detailed description of the translation's modeling, including exact labels and nomenclature adopted for nodes, as well as the modeling templates for other processes, namely Transcription Stall, Translation Stall, Protein and RNA Decay, and Cellular Division, can be found in the Supplementary Information. Also, one should notice that all the following analysis refer to the network built using our modeling assumptions. Different modeling assumptions might lead to different structures, and therefore, to different conclusions.

### Cascading failure analysis

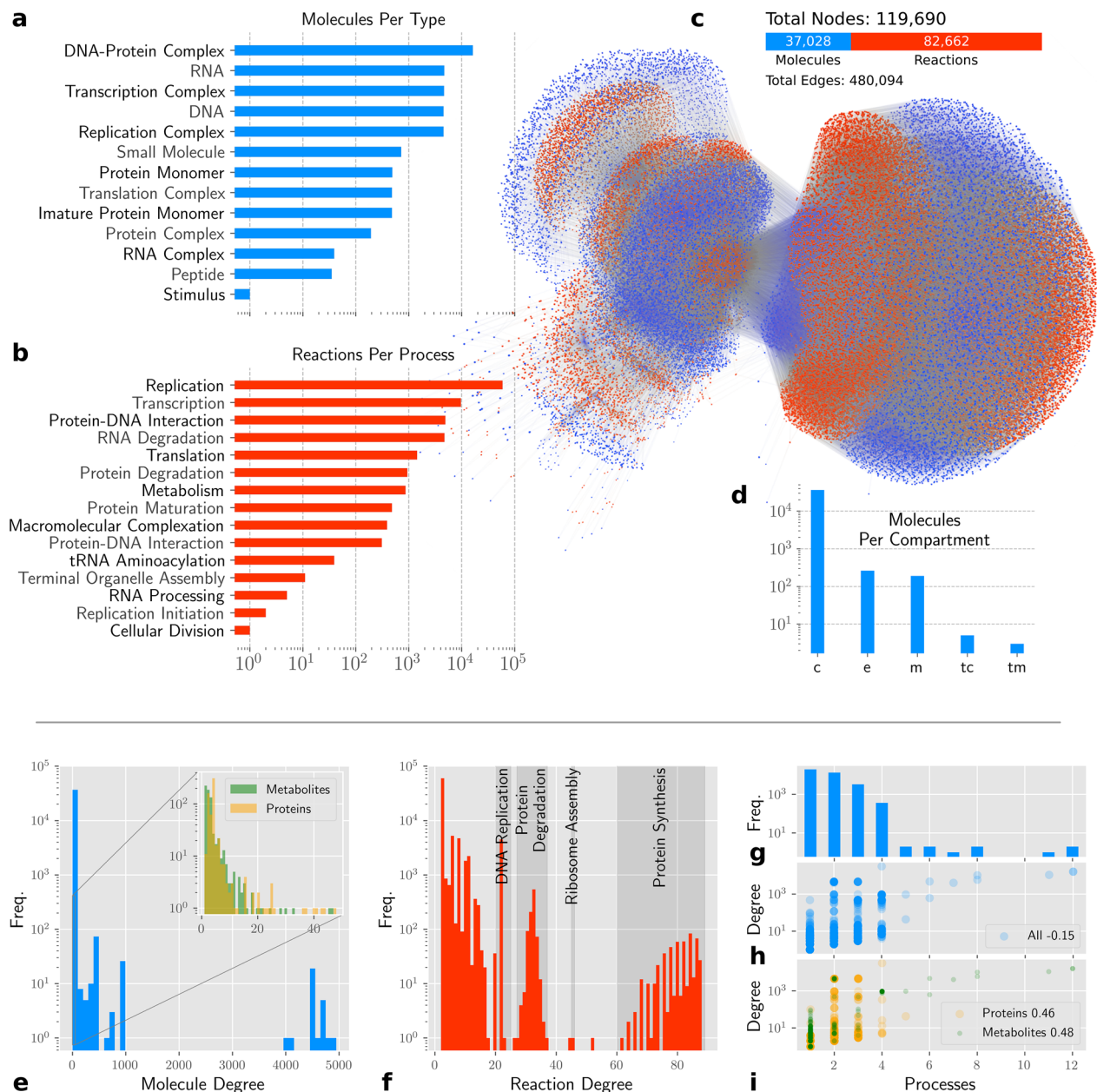
Some specific gene deletions can trigger a deadly cascade of failures in cells. Other deletions might not cause such an impact, characterizing cellular robustness<sup>46</sup>. Cascading failure analysis has been used to evaluate robustness on several network-based systems<sup>47–50</sup>. Analogously, it has been successfully applied for the estimation of essential genes in metabolic networks. They removed reactions regulated or catalyzed by a given gene product, and the impact of cascade failure propagation on the network structure revealed a correlation with gene knock-out lethality<sup>51,52</sup>.

The proposed framework and metabolic networks share many structural characteristics. Because of their similarity, the same cascading failure approach can be used to estimate the impact of gene deletion on a whole-cell biochemical network.

We adopted the same algorithm described by Mombach *et al.*<sup>51</sup> to perform the cascading failure analysis. As demonstrated in their work, cascade failures were started by removing reactions regulated by a given enzyme in order to quantify its essentiality. However, enzymes and other regulators are explicitly connected to their respective reactions in our framework. Thus, we use molecule nodes which represent the functional molecule of each gene as starting points of the cascade failure dynamics, instead of reaction nodes, as shown in Fig. 6. For genes that are further translated into proteins, we selected the molecule nodes, which represent its protein monomers. For tRNAs and ribosomal RNAs, the selected nodes were the ones representing the RNA molecules themselves.

In order to observe the effect of gene deletion in the network, we checked if a specific critical reaction, the cellular division, is still present in the network after the gene deletion. In other words, we remove the node representing the respective gene and perform the cascading failure analysis. Then, the gene is classified as essential if the critical reaction is absent in the network. Otherwise, the gene is classified as non-essential. Additionally to single gene deletions, we performed double gene deletions to analyze relationships between pairs of genes.

**Network rewiring.** Randomization of network topologies is one of the most commonly applied methods to investigate how much information is encoded within a model<sup>53–55</sup>. Here we employ a randomization process that changes the source and target of each edge with a probability  $p$ . For each edge, a random number between 0 and 1 is generated from a uniform distribution. If this number is smaller or equal than the probability  $p$ , the source and target of this edge are assigned to new nodes randomly chosen. The new nodes are chosen among ones with the same type to keep the bipartite characteristic of the network.



**Figure 7.** The *M. genitalium* whole-cell biochemical network and topological analysis. **a** number of molecules per functional group in logarithmic scale. **b** Number of reactions per cellular process in logarithmic scale. **c** Graphical visualization of the *M. genitalium* whole-cell network. Edges with high associated stoichiometry (over 100) are hidden for better visualization. The big circular group on the right is mainly composed by DNA-Protein complexes and their respective formation reactions. **d** Number of molecules per compartment in logarithmic scale. **e** Degree distribution of molecule nodes. The subgraph in the figure is the same degree distribution but showing only the protein and metabolite subgroups of nodes. **f** Degree distribution for reaction nodes where the degree is the number of connections solely. **g** Distribution of number of processes a given molecule node participates. **h** Spearman correlation between node degree and number of processes for all molecule nodes. **i** Spearman correlation between node degree and number of processes for only proteins and metabolites.

## Results

Based on WholeCellKB's<sup>16</sup> and genomic information, we built the whole-cell biochemical network of *M. genitalium*, comprising the molecular types and cellular processes indicated in Fig. 7a,b.

**Topological analysis.** The generated model is a bipartite, weighted, and directed network containing a single connected component with 119,690 nodes and 480,094 links (Fig. 7c). The nodes comprise 37,028 molecules and 82,662 reactions (Fig. 7c). The network is available in GML and SBML formats in a Github repository indicated in section "Availability of Data and Materials". We annotated molecule nodes with their functional

|                | Match  | False positive | False negative |
|----------------|--------|----------------|----------------|
| Single         | 0.5409 | 0.0171         | 0.442          |
| Double         | 0.56   | 0.0190         | 0.421          |
| Only essential | 0.9524 | 0.0476         | 0.0            |

**Table 1.** Validation of gene essentiality predictions against experimental data.

group. Figure 7a shows its distribution among the nodes. We also annotated reactions according to the cellular process to which they belong. As shown in Fig. 7b, almost all known cellular processes can be found among the reaction nodes.

Figure 7d depicts the distribution of nodes between three considered locations: cytosol (c), cellular membrane (m), Terminal Organelle Cytosol (tc), Terminal Organelle Membrane (tm), and extracellular environment (e). The 249 molecules located in the extracellular environment account for nutrients, side-products expelled from metabolism as well as secreted proteins.

Considering the *degree* of a node as the number of connections it owns, we analyzed the degree distributions for molecule and reaction nodes. For both cases, multi-modal distributions were obtained. It is known that metabolic and protein-protein interaction networks often have power-law-like degree distributions<sup>2,4,56,57</sup>. Thus, we analyzed the distribution of proteins and metabolites separately (Fig. 7e). We found that their distributions corroborate literature, but this assumption is not true for the whole system. The degree distribution for reactions showed different well-separated distributions that were found to be related to different processes (Fig. 7f). For instance, Protein Degradation reactions were found grouped in a Gaussian-like region. Other processes have reactions with a signature degree, such as DNA Replication and Ribosome Assembly, the former being accounted into the Macromolecular Complexation process.

Regarding the network completeness, 20% of the proteins monomers and protein complexes (a total of 137 molecules) have no interactions described in the WholeCellKB. These molecules are connected only to their biosynthesis and degradation reactions, participating in no other interactions. Many of these molecules with unknown function are putative membrane proteins. Because there is still no signaling pathway reported in *M. genitalium*, these putative membrane proteins can be a starting point for the elucidation of signaling processes in this organism.

**Interface between processes.** Molecules shared by reactions from different processes are interfaces between them. It means that the concentration of such molecules can affect, and be affected by, the dynamics of more than one cellular process. We found that 46.8% of the molecule nodes participate in at least two cellular processes, making the bridges between them. Parting from the hypothesis that the more connected a molecule is in the network, the higher is the chance that it connects cellular processes, a positive correlation is expected between the node's degree and the number of processes that it participates. Figure 7h shows that it is not the case, where we obtained a low negative correlation between those measures. Nevertheless, we found a significant positive correlation when analyzing only proteins and metabolites (Fig. 7i).

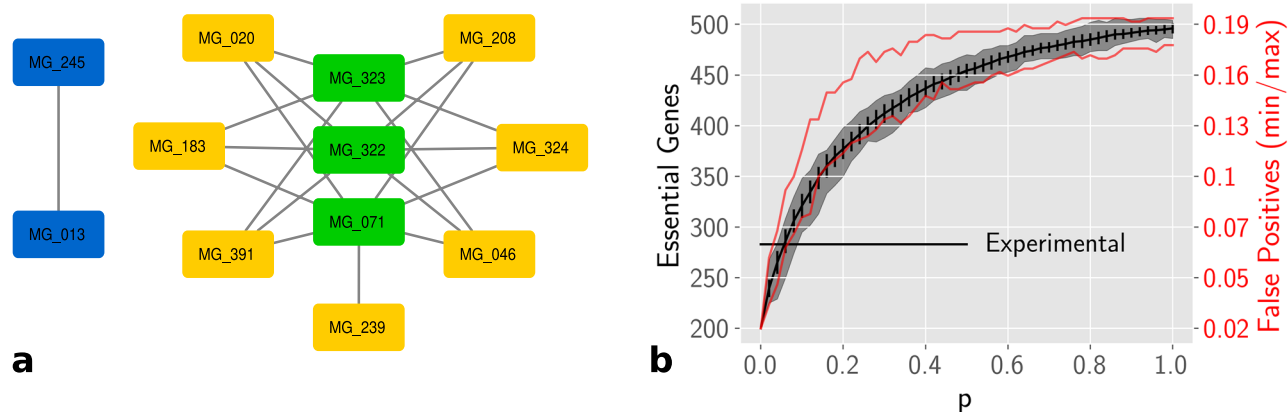
**Gene essentiality prediction.** In order to validate the *M. genitalium* whole-cell biochemical network generated in this work, we simulated gene deletions by performing the removal of nodes that represent each gene product, followed by cascade failure, and then analyzed the damage caused on the network. The simulation results regarding the 525 genes of *M. genitalium* were compared to experimental gene essentiality classification by global transposon mutagenesis gene disruption available in the literature<sup>29</sup>. The experimental data classified 382 genes out of 525 as being essential to the organism so it can replicate itself.

We performed single and double gene deletion experiments. For single-gene deletions, each gene was removed following by its cascading failure. Table 1 shows the comparison with experimental data achieving 54% of exact matches with 1.7% of false positives.

For double gene deletion, we simultaneously removed each pair of genes and performed the cascading failure. The double gene deletion slightly enhanced the classification, which resulted in 56% of correct matches. Among the genes classified as essential only in the double deletions, three distinct groups could be outlined as depicted in Fig. 8a. The first is composed of three genes (MG071, MG322, and MG323) responsible for ion transport across the membrane. The second group is composed of seven genes (MG020, MG046, MG183, MG208, MG239, MG324, and MG391) involved in the protein degradation process. The deletion of almost any combination of genes from these two groups, except for MG239 that is only essential with MG071, indicates a non-viability of the cell. The third group is composed of two genes (MG013 and MG245) that are involved in the folate metabolism and are only classified as essential if both are simultaneously removed.

It is important to observe that genes with unknown functions, which compose 22.24% of all genes, were included in this analysis. These genes might have a considerable impact on the matches rate, once all were classified as non-essential. Nonetheless, considering only genes classified as essential by the combined approach, the correct matches increase to 95%, indicating the high accuracy of the model.

In order to check the statistical relevance of the *M. genitalium* whole-cell biochemical network, we generated randomly rewired networks based on the original one by reconnecting each edge, preserving directionality, with a probability  $p$ . A total of 50 replicates were generated for each probability with  $p$  ranging from 0 to 1 with a 0.02



**Figure 8.** **a** Genes classified as essential in double deletions. The connected pairs means that their conjunct deletion resulted in the deletion of the cell division reaction. They are grouped by colors respective to their functional groups: cross-membrane ion transporters (green), protein degradation (yellow), and folate metabolism (blue). **b** Gene essentiality prediction of randomly rewired networks with probability  $p$  from the original *M. genitalium* whole-cell biochemical network. Each rewiring rate accounts for 50 replicates.

step. In Fig. 8b we can observe that the higher the network randomization level, the higher is the number of false positives in the essentiality classification.

## Discussion

***Mycoplasma genitalium* whole-cell biochemical network.** We choose the pathogenic bacterium *M. genitalium* organism as a case study to illustrate the capabilities of the proposed framework because of its simplicity and data availability in the WholeCellKB database. Despite its simplicity, the so obtained network comprises thousands of biochemical interactions from which several cellular processes emerge.

This approach provides a means such as that all cellular processes are explicitly described as a unified network of biochemical interactions. To support this affirmation, we mention the fact that the whole-cell biochemical network of *M. genitalium* has only one component, which means that all biochemical interactions are in some level interconnected. Moreover, 46.8% of the molecules participate in reactions respective to more than one process, acting as the intrinsic interfaces between them. Despite the analyses performed in this work, we believe that more information about how processes are integrated could be extracted from this model.

Indications about the reliability of the obtained reactions can be derived from the WholeCellKB, where each reaction is assigned to its respective literature evidence. The reactions in the network that were not explicitly described in the database, such as replication, transcription, and translation reactions, were derived from basic biological knowledge about these processes. Some of these reactions also incorporated particular molecules such as transcriptional factors and chaperones, which therefore have their literature evidence indicated in the database as well.

**Network topology.** Network models of natural systems usually share some topological characteristics. Considering the *degree* of a node as the number of connections it owns, the degree distribution through the nodes of a network is one of these shared features<sup>58</sup>. Somehow, it is believed that the processes that create these natural systems, particularly the biological systems, resulted in patterns of interaction between its elements which follows a power-law distribution<sup>2,4,56,57</sup>. In other words, these systems tend to have several elements with few interactions and some few elements with a high number of interactions. This feature is one of the current explanations for the robustness of living systems<sup>46,59,60</sup>. For example, if we consider a random failure among its elements, we will find that it is more likely the failure to affect elements with few connections, which tend to have more secondary than central roles in the system.

We found evidence that the degree distribution in our network resembles a power-law but only for selected molecule groups. It is the case of proteins and metabolites. The degree distribution when considering all the nodes is not trivial. Whilst having found such power-law-like structures inside the network, they represent a small fraction of all molecular entities represented in the model. The assumption of a cell's robustness explained by network topologies seems more limited in this perspective, even though the so obtained distributions rely on our modeling assumptions. Nonetheless, the properties of the power-law-like subsets could propagate to the rest of the network by cascading failures.

**Essential genes.** As a validation of the so obtained model, we simulated gene deletions by removing nodes that represent gene products, followed by their respective cascading failure, to analyze the impact of their removal on the network. High impact removals were associated with gene essentiality and compared to experimental data. Considering the static nature of the network, and the fact that 22% of the genes have unknown function, a reasonable result was obtained, achieving 56% of correctness in gene essentiality prediction. The high rate of false negatives can be a consequence of several reasons. Firstly, not all genes have a known function.

Secondly, even if the deletion of a gene does not cause significant damage to the network structure, the cell may become nonviable for dynamical reasons, such as flux bottlenecks in metabolic pathways that cannot be captured in our analysis. It is supported by the fact that almost no genes directly related to metabolism were classified as essential. Also, a given deleted gene classified as non-essential could perhaps provide the cell with the ability to grow in specific environmental conditions considered in the experimental analysis.

An interesting observation is that no significant improvement, approximately 2%, was obtained from single to double gene deletions, suggesting little redundancy in the genes with at least one known function. Therefore, we may outline the following hypothesis while considering that *M. genitalium* has a restricted genome size: (1) the organism achieves robustness through other means; (2) there are genes which have currently unknown function which could provide such redundancy; or (3) genes with at least one known function might have more roles in the cell.

Despite little improvement in the prediction, double deletions provided useful insights into *M. genitalium*'s biology. For instance, the fact that the single deletion of trans-membrane ion transporters does not have an impact in the cell division reaction, suggests that the cell has other means to obtain those ions. This other mean could be inferred by the double deletion of transporters together with genes related to protein degradation. In other words, the degradation of proteins inside the cell can provide recycled ions to the cell's cytoplasm.

Although we could not classify all of the 382 genes experimentally identified as essential, 95% of the genes we did classify as essential were correct. We can then assume that we cannot rely on the correctness of our approach if a gene is classified as non-essential, however, if a gene deletion imposes significant damage to the network, it has 95% chance of being essential to the organism indeed. Furthermore, our approach provided a computationally inexpensive solution to identify primordial genes to the cell's reproduction, being able to be computed within a few minutes on a regular computer.

Additionally, rewiring simulations showed that little changes in the network's topology increase the number of genes classified as essential but also the rate of false positives, suggesting that the results of gene essentiality prediction are strongly related to the network structure.

**Related modeling methodologies.** Network-based models have been extensively used to represent cellular processes. Besides to metabolic and signaling networks, which was already tackled in this work, we can give as examples: (a) genetic regulation network, representing positive and negative relationships between gene expressions<sup>61,62</sup>; and (b) protein-protein interaction (PPI) networks, indicating physical interactions between proteins<sup>63–65</sup>. Some of these networks, namely metabolic and signaling networks, are already representing their processes based on their underlying biochemical reactions. On the other hand, genetic regulation and PPI networks represent processes that are composed sometimes by several intermediate biochemical reactions, therefore providing a high-level representation.

If we consider a set of networks representing processes of the same organism by using the network models above mentioned, there is currently no means for a straightforward integration between them once they have very different modeling assumptions. Studies addressing the integration of cellular processes often maintain the respective underlying networks in separated layers<sup>19,20,66,67</sup>.

No quantitative comparison can be established between the so constructed *M. genitalium* network and any other network model for two reasons: (a) there is no other network model for this organism; (b) no modeling approach have integrated cellular processes that could be topologically addressed as a single network. Nevertheless, the closest approach available in the literature is the one provided by *Reactome.org*<sup>17</sup>. They aim at representing several different cellular processes through biochemical reactions. However, biochemical pathways are organized in a nested form, thus not providing a single integrated network. Also, processive processes, such as transcription and translation, are not represented. Even so, *Reactome.org* represents an accessible, curated, and annotated biochemical content which can be further transcribed, with no much effort, into single whole-cell networks using our approach.

Regarding the network's construction process, our approach could further benefit from related methodologies for metabolic networks. For instance, orthology-based network reconstructions<sup>68,69</sup> could be directly applied using the *M. genitalium* whole-cell biochemical network and other networks of model organisms that can emerge. The development of databases containing whole-cell biochemical networks would be increasingly useful for such reconstructions as new models are provided, also benefiting the study of non-model organisms.

## Conclusions

In this work, a framework was presented for integrating cellular processes at a whole-cell scale, using rule-based modeling in a broader context. Besides the incorporation of network-modeled processes, such as metabolism, we were also capable to model processes that are not usually represented as networks, such as replication, transcription, and translation. We applied the framework directives to model whole-cell scale information about the *Mycoplasma genitalium* organism stored in specialized databases aiming to probe the capabilities of the framework. The obtained whole-cell biochemical network accounted for a great variety of molecules and cellular structures, as well as the interactions between them covering almost all processes known in the organism.

Many are the applications of whole-cell biochemical networks. In bioengineering, for example, it could provide more extensive biochemical interaction maps of given organisms, serving as a tool to better manipulate them. Also, the overlapping between whole-cell biochemical networks of interacting organisms could provide insights on their relationship at the molecular level. Whole-cell biochemical networks can also pave the way to more comprehensive complex networks-based investigations, where we could study whole organisms through the topology of their biochemical interactions in a broader sense.

Whole-cell biochemical networks could also provide plenty of information for constraint-based approaches such as Flux Balance Analysis (FBA). If not considering modifier edges, the whole-cell metabolic network has the same structure of metabolic networks, therefore being able to be represented by a stoichiometric matrix. Regulation relationships indicated by modifier edges can be incorporated in FBA as additional constraints to their respective flux's boundaries. Networks generated with the proposed framework can also serve as the underlying model to deriving dynamic simulations using approaches such as the reaction rate equations and the stochastic simulation algorithms.

Despite the promising applications of the framework, the current scripts are limited to read data only from the WholeCellKB. Though impressive amounts of biological data continue to be generated, they tend to flow into relatively specific analyses and databases. A next step would be to develop software capable of searching these databases, integrate information from different sources, and therefore provide a more comprehensive and automated approach to whole-cell modeling. Some initiatives are already heading in the direction of aggregating available data<sup>70</sup>, making it usable for simulation purposes, and performing community-based development and validation of whole-cell models<sup>71</sup>. In this sense, our framework could pave the way for community-based modeling, where experts in different cellular processes can speak the same modeling language.

At any extent, the *M. genitalium* organism remains a suitable model for whole-cell simulations, and its study has a medical interest as a consequence of its pathogenic nature. Therefore, the so obtained whole-cell biochemical network can provide useful information for the previously mentioned research fields, as can be further enhanced with the emergence of new data.

## Data availability

The scripts developed for parsing the WholeCellKB and genomic data into the *M. genitalium* whole-cell network, as well as the network files, are freely available in the Github repository <https://github.com/pauloburke/whole-cell-network>.

Received: 20 August 2019; Accepted: 23 July 2020

Published online: 06 August 2020

## References

- Han, J.-D.J. Understanding biological functions through molecular networks. *Cell Res.* **18**, 224–237. <https://doi.org/10.1038/cr.2008.16> (2008).
- Oltvai, Z. N. & Barabási, A.-L. Life's complexity pyramid. *Science* **298**, 763–764. <https://doi.org/10.1126/science.1078563> (2002).
- Slak Rupnik, M. *et al.* Network science of biological systems at different scales: A review. *Phys. Life Rev.* **24**, 118–135. <https://doi.org/10.1016/j.plev.2017.11.003> (2017).
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654. <https://doi.org/10.1038/35036627> (2000).
- Reed, J. L., Famili, I., Thiele, I. & Palsson, B. O. Towards multidimensional genome annotation. *Nat. Rev. Genet.* **7**, 130–141. <https://doi.org/10.1038/nrg1769> (2006).
- Tyson, J. J., Chen, K. C. & Novak, B. Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.* **15**, 221–231. [https://doi.org/10.1016/S0955-0674\(03\)00017-6](https://doi.org/10.1016/S0955-0674(03)00017-6) (2003).
- Sobie, E. A. Bistability in biochemical signaling models. *Sci. Signal.* **4**, tr10. <https://doi.org/10.1126/scisignal.2001964> (2011).
- Yugi, K. *et al.* Reconstruction of insulin signal flow from phosphoproteome and metabolome data. *Cell Rep.* **8**, 1171–1183. <https://doi.org/10.1016/j.celrep.2014.07.021> (2014).
- De La Fuente, A., Brazhnik, P. & Mendes, P. Linking the genes: Inferring quantitative gene networks from microarray data. *Trends Genet.* **18**, 395–398. [https://doi.org/10.1016/S0168-9525\(02\)02692-6](https://doi.org/10.1016/S0168-9525(02)02692-6) (2002).
- Materna, S. C. & Oliveri, P. A protocol for unraveling gene regulatory networks. *Nat. Protoc.* **3**, 1876–1887. <https://doi.org/10.1038/nprot.2008.187> (2008).
- Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **8**, 1–9. <https://doi.org/10.1038/msb.2011.99> (2012).
- Herbach, U., Bonnaffoux, A., Espinasse, T. & Gandrillon, O. Inferring gene regulatory networks from single-cell data: A mechanistic approach. *BMC Syst. Biol.* **11**, 105. <https://doi.org/10.1186/s12918-017-0487-0> (2017).
- Arkin, A. P. & Schaffer, D. V. Network news: Innovations in 21st century systems biology. *Cell* **144**, 844–849. <https://doi.org/10.1016/j.cell.2011.03.008> (2011).
- Yu, D., Kim, M., Xiao, G. & Hwang, T. H. Review of biological network data and its applications. *Genom. Inform.* **11**, 200. <https://doi.org/10.5808/GI.2013.11.4.200> (2013).
- Shapiro, J. A. Revisiting the Central Dogma in the 21st Century. *Ann. N. Y. Acad. Sci.* **1178** VN -, 6–28. <https://doi.org/10.1111/1749-6632.2009.04990.x> (2009).
- Karr, J. R., Sanghvi, J. C., Macklin, D. N., Arora, A. & Covert, M. W. WholeCellKB: Model organism databases for comprehensive whole-cell models. *Nucleic Acids Res.* **41**, D787–D792. <https://doi.org/10.1093/nar/gks1108> (2013).
- Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655. <https://doi.org/10.1093/nar/gkx1132> (2018).
- Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**, 129 (2009).
- Wu, Y. *et al.* Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell* **158**, 1415–1430 (2014).
- Yugi, K., Kubota, H., Hatano, A. & Kuroda, S. Trans-omics: how to reconstruct biochemical networks across multiple-omic layers. *Trends Biotechnol.* **34**, 276–290 (2016).
- Malod-Dognin, N. *et al.* Towards a data-integrated cell. *Nat. Commun.* **10**, 805. <https://doi.org/10.1038/s41467-019-08797-8> (2019).
- Tomita, M. *et al.* E-CELL: Software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84. <https://doi.org/10.1093/bioinformatics/15.1.72> (1999).
- Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401. <https://doi.org/10.1016/j.cell.2012.05.044> (2012).
- Muenzner, U., Klipp, E. & Krantz, M. A comprehensive, mechanistically detailed, and executable model of the Cell Division Cycle in *Saccharomyces cerevisiae*. *Nat. Commun.* **10**, 1308 (2019).
- Sanghvi, J. C. *et al.* Accelerated discovery via a whole-cell model. *Nat. Methods* **10**, 1192–5. <https://doi.org/10.1038/nmeth.2724> (2013).

26. Balaji, N. G. B. S. Whole-cell modeling and simulation: A brief survey. *New Gen. Comput.* **38**, 259–281. <https://doi.org/10.1007/s00354-019-00066-y> (2020).
27. Waltemath, D. *et al.* Toward community standards and software for whole-cell modeling. *IEEE Trans. Biomed. Eng.* **63**, 2007–2014. <https://doi.org/10.1109/TBME.2016.2560762> (2016).
28. Hucka, M. *et al.* The Systems Biology Markup Language (SBML): Language specification for level 3 version 1 core. *Nat. Prec.* <https://doi.org/10.1038/npre.2010.4959.1> (2010).
29. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proc. Nat. Acad. Sci.* **103**, 425–430. <https://doi.org/10.1073/pnas.0510013103> (2006).
30. Chen, Q., Wang, Z. & Wei, D. Progress in the applications of flux analysis of metabolic networks. *Chin. Sci. Bull.* **55**, 2315–2322. <https://doi.org/10.1007/s11434-010-3022-x> (2010).
31. Orth, J. D., Thiele, I. & Palsson, B. O. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248. <https://doi.org/10.1038/nbt.1614> (2010).
32. Purcell, O., Jain, B., Karr, J. R., Covert, M. W. & Lu, T. K. Towards a whole-cell modeling approach for synthetic biology. *Chaos* **23**. <https://doi.org/10.1063/1.4811182> (2013).
33. Rees-Garbutt, J. *et al.* Designing minimal genomes using whole-cell models. *Nat. Commun.* **11**, 836. <https://doi.org/10.1038/s41467-020-14545-0> (2020).
34. Di Ventura, B., Lemerle, C., Michalodimitrakis, K. & Serrano, L. From in vivo to in silico biology and back. *Nature* **443**, 527–533. <https://doi.org/10.1038/nature05127> (2006).
35. Kholodenko, B., Yaffe, M. B. & Kolch, W. Computational approaches for analyzing information flow in biological networks. *Sci. Signaling* **5**. <https://doi.org/10.1126/scisignal.2002961> (2012).
36. Nacher, J. C. & Akutsu, T. Structural controllability of unidirectional bipartite networks. *Sci. Rep.* **3**, 1647. <https://doi.org/10.1038/srep01647> (2013).
37. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Nat. Acad. Sci.* **104**, 1777–1782 (2007).
38. Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinform.* **11**, 213. <https://doi.org/10.1186/1471-2105-11-213> (2010).
39. Faeder, J. R., Blinov, M. L., Goldstein, B. & Hlavacek, W. S. Rule-based modeling of biochemical networks. *Complexity* **10**, 22–41 (2005).
40. Hlavacek, W. S. & Faeder, J. R. The complexity of cell signaling and the need for a new mechanics. *Sci. Signal.* **2**, 1–4 (2009).
41. Chylek, L. A. *et al.* Rule-based modeling: A computational approach for studying biomolecular site dynamics in cell signaling systems. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **6**, 13–36 (2014).
42. Machado, D., Herrgård, M. J. & Rocha, I. Stoichiometric representation of gene-protein-reaction associations leverages constraint-based analysis from reaction to gene-level phenotype prediction. *PLoS Comput. Biol.* **12**, 1–24 (2016).
43. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
44. Stothard, P., Wishart, D. S. Circular genome visualization and exploration using CGView *Bioinformatics* **21**(4), 537–539. <https://doi.org/10.1093/bioinformatics/bti054> (2005).
45. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–404 (1995).
46. Silva-Rocha, R. & de Lorenzo, V. Noise and robustness in prokaryotic regulatory networks. *Annu. Rev. Microbiol.* **64**, 257–275 (2010).
47. Crucitti, P., Latora, V. & Marchiori, M. Model for cascading failures in complex networks. *Phys. Rev. E* **69**, 045104 (2004).
48. Wang, W.-X. & Chen, G. Universal robustness characteristic of weighted networks against cascading failure. *Phys. Rev. E* **77**, 026101 (2008).
49. Smart, A. G., Amaral, L. A. N. & Ottino, J. M. Cascading failure and robustness in metabolic networks. *Proc. Nat. Acad. Sci. U.S.A.* **105**, 13223–13228. <https://doi.org/10.1073/pnas.0803571105> (2008).
50. Huang, X., Vodenska, I., Havlin, S. & Eugene Stanley, H. Cascading failures in bi-partite graphs: Model for systemic risk propagation. *Sci. Rep.* **3**, 13. <https://doi.org/10.1038/srep01219> (2013).
51. Lemke, N., Herédia, F., Barcellos, C. K., dos Reis, A. N. & Mombach, J. C. M. Essentiality and damage in metabolic networks. *Bioinformatics* **20**, 115–119 (2004).
52. Wunderlich, Z. & Mirny, L. A. Using the topology of metabolic networks to predict viability of mutant strains. *Biophys. J.* **91**, 2304–2311. <https://doi.org/10.1529/biophysj.105.080572> (2006).
53. Fraser, H. B. Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
54. Takahashi, D. Y., Sato, J. R., Ferreira, C. E. & Fujita, A. Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PLoS ONE* **7**. <https://doi.org/10.1371/journal.pone.0049949> (2012).
55. Huang, B. *et al.* Interrogating the topological robustness of gene regulatory circuits by randomization. *PLoS Comput. Biol.* **13**, 1–21 (2017).
56. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947–4957. <https://doi.org/10.1242/jcs.02714> (2005).
57. Khanin, R. & Wit, E. How scale-free are biological networks. *J. Comput. Biol.* **13**, 810–818 (2006).
58. Barabási, A.-L. & Bonabeau, E. Scale-free networks. *Sci. Am.* **288**, 60–69 (2003).
59. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664. <https://doi.org/10.1126/science.1069492> (2002).
60. Bruggeman, F. J. & Westerhoff, H. V. The nature of systems biology. *Trends Microbiol.* **15**, 45–50 (2007).
61. Schlitt, T. & Brazma, A. Current approaches to gene regulatory network modelling. *BMC Bioinform.* **8**, S9 (2007).
62. Le Novère, N. Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.* **16**, 146 (2015).
63. Martha, V.-S. *et al.* Constructing a robust protein-protein interaction network by integrating multiple public databases. *BMC Bioinform.* **12**, S7 (2011).
64. Taghipour, S., Zarrineh, P., Ganjtabesh, M. & Nowzari-Dalini, A. Improving protein complex prediction by reconstructing a high-confidence protein-protein interaction network of *Escherichia coli* from different physical interaction data sources. *BMC Bioinform.* **18**, 10 (2017).
65. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
66. Covert, M. W., Schilling, C. H. & Palsson, B. Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* **213**, 73–88. <https://doi.org/10.1006/jtbi.2001.2405> (2001).
67. Oliveira, A. P. *et al.* Regulation of yeast central metabolism by enzyme phosphorylation. *Mol. Syst. Biol.* **8** (2012).
68. Hamilton, J. J. & Reed, J. L. Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environ. Microbiol.* **16**, 49–59 (2014).
69. Notebaart, R. A., Van Enckevort, F. H. J., Francke, C., Siezen, R. J. & Teusink, B. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinform.* **7**, 296 (2006).
70. Weaver, D. S., Keseler, I. M., Mackie, A., Paulsen, I. T. & Karp, P. D. A genome-scale metabolic flux model of *Escherichia coli* K-12 derived from the EcoCyc database. *BMC Syst. Biol.* **8**, 79 (2014).
71. Goldberg, A. P. *et al.* Emerging whole-cell modeling principles and methods. *Curr. Opin. Biotechnol.* **51**, 97–102 (2018).

## Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. MGQ thanks FAPESP (Grant No. 2015/50122-0) and CNPq (Grant No. 313426/2018-0). Luciano da F. Costa thanks CNPq (Grant No. 307085/2018-0) for sponsorship. This work has also benefited from FAPESP Grant No. 15/22308-2.

## Author contributions

P.E.P.B. conceived the idea. P.E.P.B. and M.G.Q. established the network modeling framework. P.E.P.B. and C.B.L.C. modeled the cellular processes using the framework. P.E.P.B. and Ld.F.C. carried the analysis of the model and wrote the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-70145-4>.

**Correspondence** and requests for materials should be addressed to P.E.P.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020