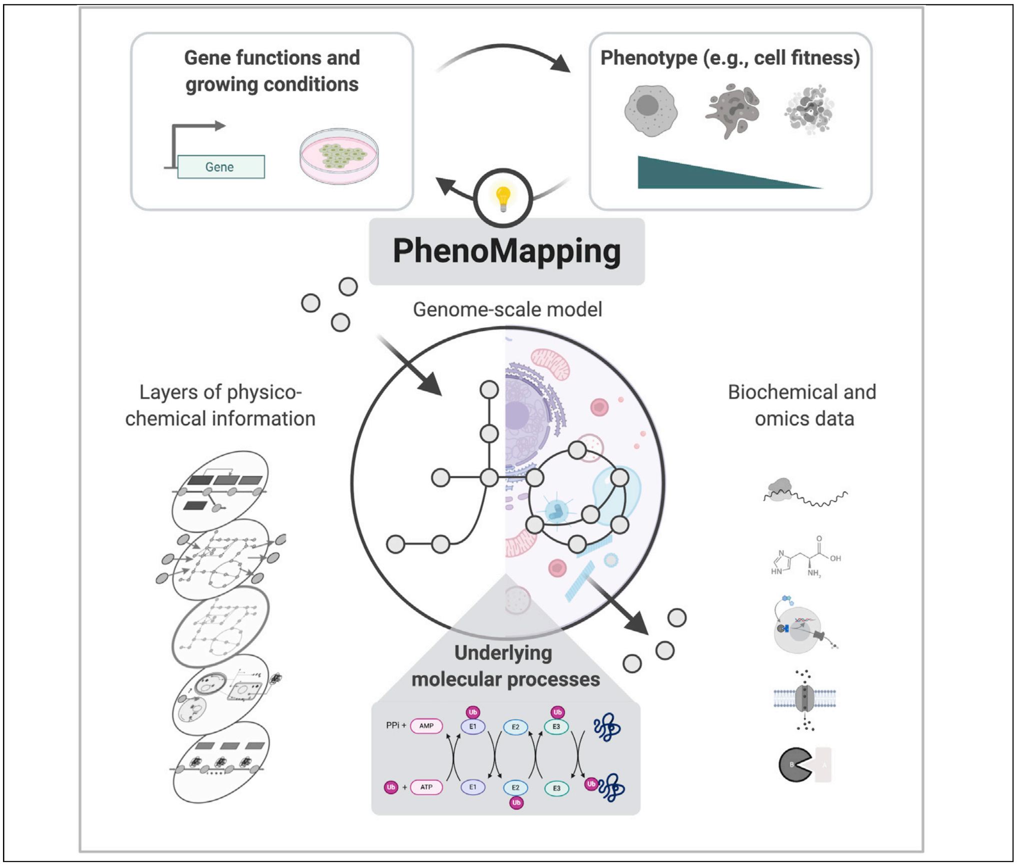
STAR Protocols

CellPress
OPEN ACCESS

Protocol

PhenoMapping: a protocol to map cellular phenotypes to metabolic bottlenecks, identify conditional essentiality, and curate metabolic models


Gene functions and growing conditions
Gene
Phenotype (e.g., cell fitness)
PhenoMapping
Genome-scale model
Layers of physico-chemical information
Biochemical and omics data
Underlying molecular processes
PPi +
AMP
E1
E2
E3
Ub
ATP

Targeted identification of cellular processes responsible for a phenotype is of major importance in guiding efforts in bioengineering and medicine. Genome-scale metabolic models (GEMs) are widely used to integrate various types of omics data and study the cellular physiology under different conditions. Here, we present PhenoMapping, a protocol that uses GEMs, omics, and phenotypic data to map cellular processes and observed phenotypes. PhenoMapping also classifies genes as conditionally and unconditionally essential and guides a comprehensive curation of GEMs.

Anush Chiappino-Pepe, Vassily Hatzimanikatis

anush.chiappinopepe@alumni.epfl.ch

HIGHLIGHTS

Systematic identification of cellular processes causing phenotypes

Decoding nutrient usage from genetic screens, as shown in two parasites

Curation of two genome-scale models leads to 80% accuracy in essentiality predictions

Classification of conditional essentiality will guide drug targeting strategies

Chiappino-Pepe & Hatzimanikatis, STAR Protocols 2, 100280
March 19, 2021 © 2020 The Author(s).
https://doi.org/10.1016/j.xpro.2020.100280

CellPress
OPEN ACCESS

Protocol

PhenoMapping: a protocol to map cellular phenotypes to metabolic bottlenecks, identify conditional essentiality, and curate metabolic models

Anush Chiappino-Pepe[1,2,3,4,5,*] and Vassily Hatzimanikatis[1]

[1]Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, EPFL, Lausanne, Switzerland

[2]Present address: Department of Genetics, Harvard Medical School, Boston, MA, USA

[3]Present address: Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[4]Technical contact

[5]Lead contact

*Correspondence: anush.chiappinopepe@alumni.epfl.ch

https://doi.org/10.1016/j.xpro.2020.100280

SUMMARY

Targeted identification of cellular processes responsible for a phenotype is of major importance in guiding efforts in bioengineering and medicine. Genome-scale metabolic models (GEMs) are widely used to integrate various types of omics data and study the cellular physiology under different conditions. Here, we present PhenoMapping, a protocol that uses GEMs, omics, and phenotypic data to map cellular processes and observed phenotypes. PhenoMapping also classifies genes as conditionally and unconditionally essential and guides a comprehensive curation of GEMs.

For complete details on the use and execution of this protocol, please refer to Stanway et al. (2019) and Krishnan et al. (2020).

BEFORE YOU BEGIN

In this first section, we present a brief relation of PhenoMapping to prior art and the preparatory steps to perform a PhenoMapping analysis (Figure 1). We discuss a set of decisions to make and how they impact the subsequent PhenoMapping analysis. In addition, we introduce the input data needed and how to set up the GEM and software. To adapt to all ranges of expertise in metabolic modeling, we provide links to the troubleshooting section, where we describe technical details to perform the related steps. We present as an example the application of these preparatory steps to *Plasmodium berghei* in the section Expected outcomes. These steps were applied similarly in *Toxoplasma gondii* and are generalizable to any organism and study case. For examples of studies validating the results and insights obtained following the PhenoMapping protocol, please refer to Stanway et al., 2019 and Krishnan et al., 2020.

Relation of method to prior art

Identifying cellular processes responsible for a phenotype is especially complex and relevant in biological systems. Genome-scale models (GEMs) are widely used to integrate all available biochemical information of an organism and various types of omics data to study the metabolic function at different conditions. The protocol described here builds on three decades of method development to construct and analyze GEMs. It complements available protocols (Thiele and Palsson, 2010) and tools (Agren et al., 2013; Devoid et al., 2013; Heirendt et al., 2019, 2019; Lieven et al., 2020; Machado et al., 2018; Salvy et al., 2018; Wang et al., 2018) for high-quality reconstruction and analysis of GEMs. This protocol provides a systematic guideline to identify cellular bottlenecks underlying

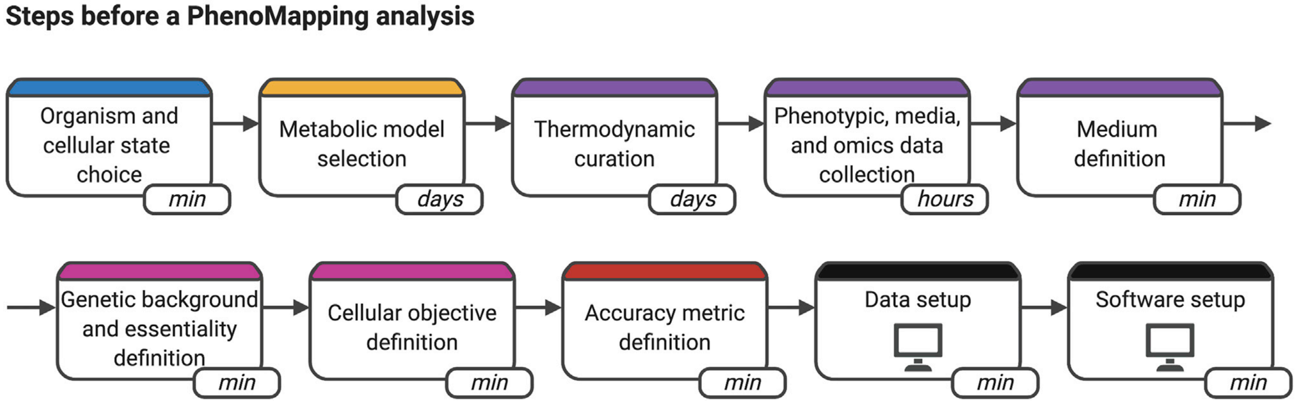

Steps before a PhenoMapping analysis
Organism and cellular state choice
min
Metabolic model selection
days
Thermodynamic curation
days
Phenotypic, media, and omics data collection
hours
Medium definition
min
Genetic background and essentiality definition
min
Cellular objective definition
min
Accuracy metric definition
min
Data setup
min
Software setup
min

Figure 1. Preparatory steps for a PhenoMapping analysis

Color code is consistent with related steps in the main PhenoMapping workflow. We include an approximate assessment of the timing each step takes.

phenotypes. It also describes how to use the knowledge about metabolic bottlenecks toward the understanding of conditional essentiality and curation of GEMs, as recently shown (Krishnan et al., 2020; Stanway et al., 2019).

Alternative methods to increase the predictive accuracy of GEMs include automatized approaches like AMMEDEUS (Medlock and Papin, 2020), GrowMatch (Kumar and Maranas, 2009) or GlobalFit (Hartleb et al., 2016), and others less automatized like RING (Sohn et al., 2012). The solutions provided by these methods may remain limited to the physico-chemical constraints integrated into the GEM. This protocol suggests a systematic classification and evaluation of such physico-chemical constraints for the identification and curation of a broader range of knowledge gaps in GEMs. Moreover, through the systematic classification and analysis of bottlenecks defined in this protocol, one gains important biological insights like substrates linked to conditional gene essentiality that had remained rather unexplored *in silico* so far. Currently, tools like COBRA (Heirendt et al., 2019), RAVEN (Wang et al., 2018), modelSEED (Devoid et al., 2013), KBase (US Department of Energy Systems Biology Knowledgebase, http://kbase.us), CarveMe (Machado et al., 2018), TFA (Salvy et al., 2018) etc. are widely used to construct and analyze GEMs. PhenoMapping, as defined in this protocol and accompanying GitHub repository (www.github.com/EPFL-LCSB/phenomapping), can be applied in combination with any of those tools. This protocol provides rigorous details for a PhenoMapping study design, integrative analysis using omics and phenotypic data, and comprehensive evaluation of results.

Organism and cellular state choice

Timing: 1–10 min

1. Select an organism and strain or cell line of interest.
2. Select the *cellular state(s)* of interest.
   a. Select a life-stage (if applicable).
   b. Select a specific time point in the life-stage.

⚠ CRITICAL: the cellular state selected will determine the metabolic state of your organism or cell, which further restricts the gathering of data (see section on Phenotypic, media, and omics data collection) and selection of a cellular objective (see section Cellular objective definition) in the PhenoMapping analysis. We recommend the ‘‘safe and easy’’ selection of a highly metabolically active state for which the cellular objective can be represented as a ‘‘desire to maximize growth.’’

Metabolic model selection

⏲ Timing: ~1 day

Note: the time spent to select a GEM varies dramatically depending both on the experience of the user with the organism of study and analysis of GEMs, and on the availability and quality of the GEMs.

3. Search in databases like BiGG (King et al., 2015), modelSEED (Henry et al., 2010), KBase (US Department of Energy Systems Biology Knowledgebase, http://kbase.us), LCSB database (LCSB, 2020), etc. or publications for a GEM of your organism of interest.

⚠ CRITICAL: if there exists no GEM for your organism and strain, you may want to follow standard protocols to construct a GEM; Troubleshooting 1. If there exist multiple GEMs, you need to select one for the subsequent analysis; Troubleshooting 2. Any additional analysis and evaluation of the GEMs prior to PhenoMapping are an alternative with a consequent time extension.

4. Select a GEM for your organism and strain of interest.

Note: in the subsequent steps the GEM will be thermodynamically curated, prepared, and initialized to be ready for a PhenoMapping analysis.

Thermodynamic curation

⏲ Timing: ~1 day

Note: the time spent to thermodynamically curate a GEM varies considerably depending on the tools used to perform this task. In addition, the time will vary based on the extent to which the user wants to *a posteriori* evaluate and curate the performance of the GEM under thermodynamic constraints. If no systematic tool is used to curate the GEM thermodynamically, variables like GEM size will affect the timing too.

5. Include all thermodynamic data in the GEM as necessary within the TFA framework (Henry et al., 2006, 2007; Jankowski et al., 2008; Salvy et al., 2018) to perform a thermodynamically consistent flux balance analysis; Troubleshooting 3.
6. Verify the GEM has all fields required to be thermodynamically curated in a systematic fashion following the steps defined in the section Software setup.

Phenotypic, media, and omics data collection

⏲ Timing: 5–8 h

Note: the time spent in a literature search and mapping of the data to the GEM greatly varies depending on the organism of study, amount of data available, and automatization of the mapping.

7. Get phenotypic data for the organism and cellular state to study.
   a. List all genes in the GEM and map the collected data of *in vivo* phenotypes (e.g., essential or dispensable upon single gene knockout).
8. Gather information about the media composition at the cellular state.
   a. List all extracellular metabolites in the GEM and map the available information about the availability of each metabolite at the cellular state.

9. Assemble available metabolomics data for the organism and cellular state to study.
   a. List the set of metabolites included in the GEM and map the corresponding concentration ranges (minimum and maximum absolute values).
10. Assemble available RNA-seq or proteomics datasets for the organism and cellular state to study.
    a. List all genes in the GEM and map the corresponding and unique RNA or protein level.

Medium definition

⏲ Timing: 1–10 min

11. Select a media composition for the PhenoMapping analyses.

⚠ CRITICAL: we recommended to select a rich medium at this point. We define a rich medium when a broad range of metabolites (if possible, all extracellular) are allowed to be taken up by the GEM. Selecting a rich medium at this stage will allow PhenoMapping to map substrate availability to essentiality (conditional essentiality). PhenoMapping will only be able to map conditionally essential genes and the responsible substrates for the genes' essentiality, if those substrates are made available (can be taken up) at this step (see section Metabolic model contextualization).

12. Define the medium composition in the GEM with the desired maximum uptake rates allowed; Troubleshooting 4.

Genetic background and essentiality definition

⏲ Timing: 1–5 min

13. Select the type of essentiality analysis to perform.
    a. Single gene knockout.
    b. Single reaction knockout.
    c. Multiple gene knockout, or multiple reaction knockout.

Note: by default, PhenoMapping performs single gene knockout. This is because the phenotypic data available are normally single gene knockout data. If a single gene knockout analysis is selected, PhenoMapping will map bottlenecks to individual essential genes. A single gene knockout analysis is normally preferred to a multiple knockout analysis because it is faster (given the current available methods to unbiasedly perform double or multiple gene knockout analysis *in silico*). In this protocol, we will refer to single essential genes. And, the same concepts apply to any set of essential genes or reactions identified *in silico* as decided at this step.

14. Decide the genetic background of the *in silico* organism (GEM) on which the PhenoMapping analysis will be performed.
    a. Keep the wild-type genetic background to perform PhenoMapping analyses of single essential genes.
    b. Define *in silico* deletion strains, i.e., with a deleted gene or reaction or multiple deleted genes or reactions (even if that does not match the genetic background of the organism chosen) to efficiently identify bottlenecks of sets of redundant genes.

Note: a scenario with a knockout background might be desired when one knows that a gene is part of a synthetic lethal pair and one desires to map bottlenecks to the synthetic lethal pair.

This strategy will not require a double knockout analysis within PhenoMapping, which is more time consuming and computationally expensive.

15. Define an *essentiality threshold* or a percentage of the optimal value of the objective function. The knockout that renders a value of objective function below this threshold is considered as lethal (see Essentiality prediction section).
    a. PhenoMapping uses by default an essentiality threshold of 0.1, which indicates that every gene whose knockout leads to a growth reduction of 90% or more with respect to a reference value (normally the wild-type growth) is considered essential.

Cellular objective definition

Timing: 1–5 min

Note: the time spent to define the objective function varies considerably depending on three main factors: the type of objective function chosen; the feasibility of the GEM for the given objective function under the defined conditions; and the experience of the user to accurately formulate the desired objective function. Difficulties in any of these points can expand the timing to days and weeks.

16. Select and define in the GEM an objective function that represents the cellular objective at the state of study (Schuetz et al., 2007). By default, PhenoMapping uses maximization of growth as the objective function (Feist and Palsson, 2010); Troubleshooting 5.
17. Verify that it is possible to obtain a solution for the objective function selected in the medium and genetic background defined; Troubleshooting 6.

Accuracy metric definition

Timing: 1–5 min

18. Familiarize with the description of knockouts based on the predicted outcome using the GEM:
    a. Positives: the GEM predicts little or no effect on wild-type growth (positive growth) upon knockout of the gene
    b. Negatives: the GEM predicts a negative effect on wild-type growth upon knockout of the gene
19. Familiarize with a contingency matrix for the comparison of predictions and data, which includes the following definitions:
    a. True positives (TP): dispensable both *in silico* and *in vivo.*
    b. True negatives (TN): essential both *in silico* and *in vivo.*
    c. False positives (FP): dispensable *in silico* and essential *in vivo.*
    d. False negatives (FN): essential *in silico* and dispensable *in vivo.*

20. Select a metric to assess the accuracy of your GEM in the essentiality prediction. These metrics can be systematically computed within PhenoMapping (see section Expected outcomes):
    a. Matthew correlation coefficient (MCC):

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

    b. Overall accuracy:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

c. Negative prediction rate (NPR):

$$NPR = \frac{TN}{TN + FN}$$

d. Positive prediction rate (PPR):

$$PPR = \frac{TP}{TP + FP}$$

e. Sensitivity:

$$Sensitivity = \frac{TP}{TP + FN}$$

f. Specificity:

$$Specificity = \frac{TN}{TN + FP}$$

Note: the MCC and overall accuracy tend to be used as the main metrics of accuracy assessment. However, they assume that incorrect predictions, i.e., FPs and FNs, are equally "bad," and this is arguably not the case. Given a reliable and high-quality dataset of phenotypes, one would normally prefer to work with a GEM that has less FNs than FPs. Such GEM is not over-constrained, so it does not incorrectly predict essential genes, but it lacks a set of constraints to increase the number of true essentiality predictions.

For example, when one uses a highly curated life-stage agnostic GEM, one expects that it already contains all biochemical information available about the organism. To make it life-stage specific, we should *a priori* only add physico-chemical constraints that are context-specific. These constraints should increase the number of correctly identified conditionally essential genes, while not increasing the incorrect essentiality predictions (FNs). This means, to generate context-specific GEMs we may try to increase the positive prediction rate while keeping the negative prediction rate constant.

Data setup

⏲ Timing: 1 h

This section includes a suggestion on the setup of data within the PhenoMapping repository. The data were set up in this format for the example scripts and PhenoMapping analyses of the *Plasmodium berghei* metabolic model (iPbe) (Stanway et al., 2019) and *Toxoplasma gondii* metabolic model (iTgo) (Krishnan et al., 2020).

21. GEM setup
    a. Save the GEM in MATLAB (.mat) format in the folder "models" of the PhenoMapping directory.
    b. Generate a folder with your model name in the PhenoMapping subfolder "tests/ref."
22. Phenotypic data setup
    a. Format the phenotypic data in a 2-column csv file: genes names (column 1) and observed phenotype upon knockout (column 2).
    b. Store the csv file with the phenotypic data in the PhenoMapping subfolder tests/ref/model-name.

⚠ CRITICAL: verify gene identifiers match those included in the model in the field "genes." If not available in the GEM, a warning will state that these genes are not in the GEM and the data will not be considered for further analysis.

23. Media data setup
    a. Format the media data in a 2-column csv file: metabolite names or identifiers (column 1) and information about their availability or uptake (column 2).
    b. Store the csv file with media data in the PhenoMapping subfolder "tests/ref/modelname."

⚠ CRITICAL: verify metabolite identifiers match those included in the model in the field "mets" or "metNames." If not available in the GEM, a warning will state that these metabolites are not present in the GEM and the data will not be integrated or considered for further analysis.

24. Metabolomics data setup
    a. Format the metabolomics data as a 3-column csv file: names of metabolites for which concentration data are available (column 1) and concentration data formatted as explained below in columns 2 and 3.
    b. Compute the minimum (average minus standard deviation) and maximum (average plus standard deviation) values of concentration measured.
    c. Convert the units of the metabolomics data into mol/L_{cell}.
    d. Add the concentration values to the csv file: minimum concentration values (column 2) and maximum concentration values (column 3).
    e. Store the csv file with metabolomics data in the PhenoMapping subfolder tests/ref/modelname.

⚠ CRITICAL: verify metabolite identifiers match those included in the model in the field "mets" or "metNames." If not available in the GEM, a warning will state that these metabolites are not present in the GEM and the data will not be integrated or considered for further analysis.

25. Transcriptomics or proteomics data setup
    a. Format the transcriptomics or proteomics data in a 2-column csv file: genes names (column 1) and a unique value of measured RNA or protein level (column 2).

Note: the units of the RNA-seq or proteomics measurements are not relevant at this point as long as all RNAs or proteins measured share these units. This is because the GEMs used in PhenoMapping do not integrate the concentration of RNAs or proteins as variables. PhenoMapping will evaluate the distribution of RNA or protein levels across all genes in the GEM and will discretize these distributions in three groups using TEX-FBA (Pandey et al., 2019); lowly expressed, medium expression, and highly expressed (see the section describing the transcriptomics analysis and TEX-FBA parameters definition).

    b. Store the csv file with transcriptomics or proteomics data in the PhenoMapping subfolder tests/ref/modelname.

Software setup

⏱ Timing: 5–30 min

This section includes a suggestion on the setup of paths and preprocessing of the GEM for a PhenoMapping analysis using a sample script. There are many alternatives, and some are discussed in more detail in the tutorials script within the PhenoMapping repository and in this protocol in the Troubleshooting section.

Note: PhenoMapping requires MATLAB, CPLEX, and the GitHub repositories matTFA, TEX-FBA, and PhenoMapping. Links to these have been included in the Materials and equipment section.

26. Prepare a settings script using as reference the templates provided in the PhenoMapping subfolder tests, i.e., settings_ipbeblood.m, settings_ipbeliver.m, settings_itgo.m.
    a. Provide paths, file names, and variable names for the GEM and data to be used in PhenoMapping.
    b. Select whether the GEM should be thermodynamically curated. This is an input of the initTestPhenoMappingModel function.
27. Run the settings script to (1) verify that all paths to matTFA, TEX-FBA, and CPLEX are found, (2) check that all data files are found, and (3) preprocess the GEM for PhenoMapping analysis.

⚠ CRITICAL: this step will highlight any problem to add paths to CPLEX, matTFA, or TEX-FBA. This step will also spot any missing information or field in the GEM as required for PhenoMapping; Troubleshooting 7.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
P. berghei relative growth rate phenotypes in blood stages	Bushell et al., 2017	https://doi.org/10.1016/j.cell.2017.06.030
P. berghei relative growth rate phenotypes in liver stages	Stanway et al., 2019	https://doi.org/10.1016/j.cell.2019.10.030
P. berghei RNA-seq data in blood stages	Otto et al., 2014	https://doi.org/10.1186/s12915-014-0086-0
P. berghei RNA-seq data in liver stages	Caldelari et al., 2019	https://doi.org/10.1186/s12936-019-2968-7
Compiled metabolomics dataset from *P. falciparum*	Chiappino-Pepe et al., 2017	https://doi.org/10.1371/journal.pcbi.1005397
GEM of *P. berghei iPbe*	Stanway et al., 2019	https://doi.org/10.1016/j.cell.2019.10.030
T. gondii relative growth rate phenotypes in tachyzoites	(Sidik et al., 2016)	https://doi.org/10.1016/j.cell.2016.08.019
T. gondii tachyzoite RNA-seq data	(Hehl et al., 2015); ToxoDB v45	www.toxodb.org
GEM of *T. gondii iTgo*	Krishnan et al., 2020	https://doi.org/10.1016/j.chom.2020.01.002
Software and algorithms		
PhenoMapping	www.github.com/EPFL-LCSB/phenomapping	1.0
TEX-FBA	www.github.com/EPFL-LCSB/texfba	1.0
matTFA	www.github.com/EPFL-LCSB/matTFA	1.0
MATLAB	Mathworks (https://www.mathworks.com/products/matlab.html)	R2016a - R2019a
CPLEX	https://www.ibm.com/analytics/cplex-optimizer	12.8
COBRA Toolbox (used updated version within matTFA)	www.github.com/EPFL-LCSB/matTFA	1.0

MATERIALS AND EQUIPMENT

Software

MATLAB (MathWorks: https://www.mathworks.com/products/matlab.html)

Alternatives: While the current implementation of PhenoMapping is in MATLAB, the PhenoMapping rationale and workflow is extendable to any other programming language like python.

CPLEX (IBM: https://www.ibm.com/analytics/cplex-optimizer.html)

Alternatives: While the current implementation of PhenoMapping uses CPLEX, other available solvers like gurobi could be implemented and used.

matTFA (GitHub: www.github.com/EPFL-LCSB/matTFA)

Alternatives: An implementation of matTFA in python exists and it is called pyTFA (Salvy et al., 2018).

TEX-FBA (GitHub: www.github.com/EPFL-LCSB/texfba)

Alternatives: While the current implementation of TEX-FBA is in MATLAB, the TEX-FBA formulation is extendable to any other programming language like python.

PhenoMapping (GitHub: www.github.com/EPFL-LCSB/phenomapping)

⚠ CRITICAL: PhenoMapping only requires CPLEX, matTFA, and TEX-FBA to be on the MATLAB path for optimization analysis. Further instructions on how to optimally handle paths in PhenoMapping are available in Troubleshooting 7.

Data

The data used in PhenoMapping are collected from separate studies. Here, we summarize data types for which a PhenoMapping analysis is currently implemented. We define two types of data depending on their use in PhenoMapping: data type 1 involves phenotypic data used for a comparison with the GEM predictions to assess the accuracy of the GEM; and data type 2 involves other datasets like omics data and media data integrated into the GEM to contextualize it. None of the datasets but the GEM is truly essential since one can perform a purely *in silico* analysis of phenotypes and bottlenecks with a metabolic model in PhenoMapping. However, broader biological insights are achieved when experimental data are integrated into the PhenoMapping pipeline. We define how essential each dataset is (++++, necessary; +++, strong; ++, medium; +, low) for a PhenoMapping analysis, and the suggested labels or values for the data.

Data (PhenoMapping data type)	Degree of requirement	Suggested labels
GEM	++++	MATLAB format with standard fields defined in constraint-based modeling
Phenotypic data (data type 1)	+++	essential; non-essential; (slow[a])
Media data (data type 2)	+	available; non-available; (unknown)
Thermodynamic data (data type 2)	++	Thermodynamically curated GEM (Salvy et al., 2018)
Metabolomics data (data type 2)	++	Absolute values (mol/L_{cell}). Minimum and maximum measured or allowed concentration values per metabolite
RNA-seq data (data type 2)	++	TPMs or absolute values. Unique RNA level per gene

[a] Some experimental datasets like those obtained for the blood and liver stages of the *Plasmodium* development (Bushell et al., 2017; Stanway et al., 2019) may include "slow" phenotypes. These genes might be considered as essential or dispensable in PhenoMapping depending on the GEM context, as explained in the next sections.

⚠ CRITICAL: PhenoMapping identifies metabolic bottlenecks responsible for the essentiality of a gene. PhenoMapping can map bottlenecks to *in silico* essential genes but stronger biological insights can be obtained when experimentally observed essential genes or phenotypic data (data type 1) are used in the pipeline.

Alternatives: To study context-specific functions of the cell, PhenoMapping integrates omics and media data (data type 2) into a GEM. If no omics data are available, PhenoMapping will only map context-specific essential genes to substrate availability. If omics data are available, PhenoMapping will identify which minimum alternative sets of concentration levels (as measured in the omics datasets) can explain an observed gene essentiality or phenotype (data type 1).

STEP-BY-STEP METHOD DETAILS

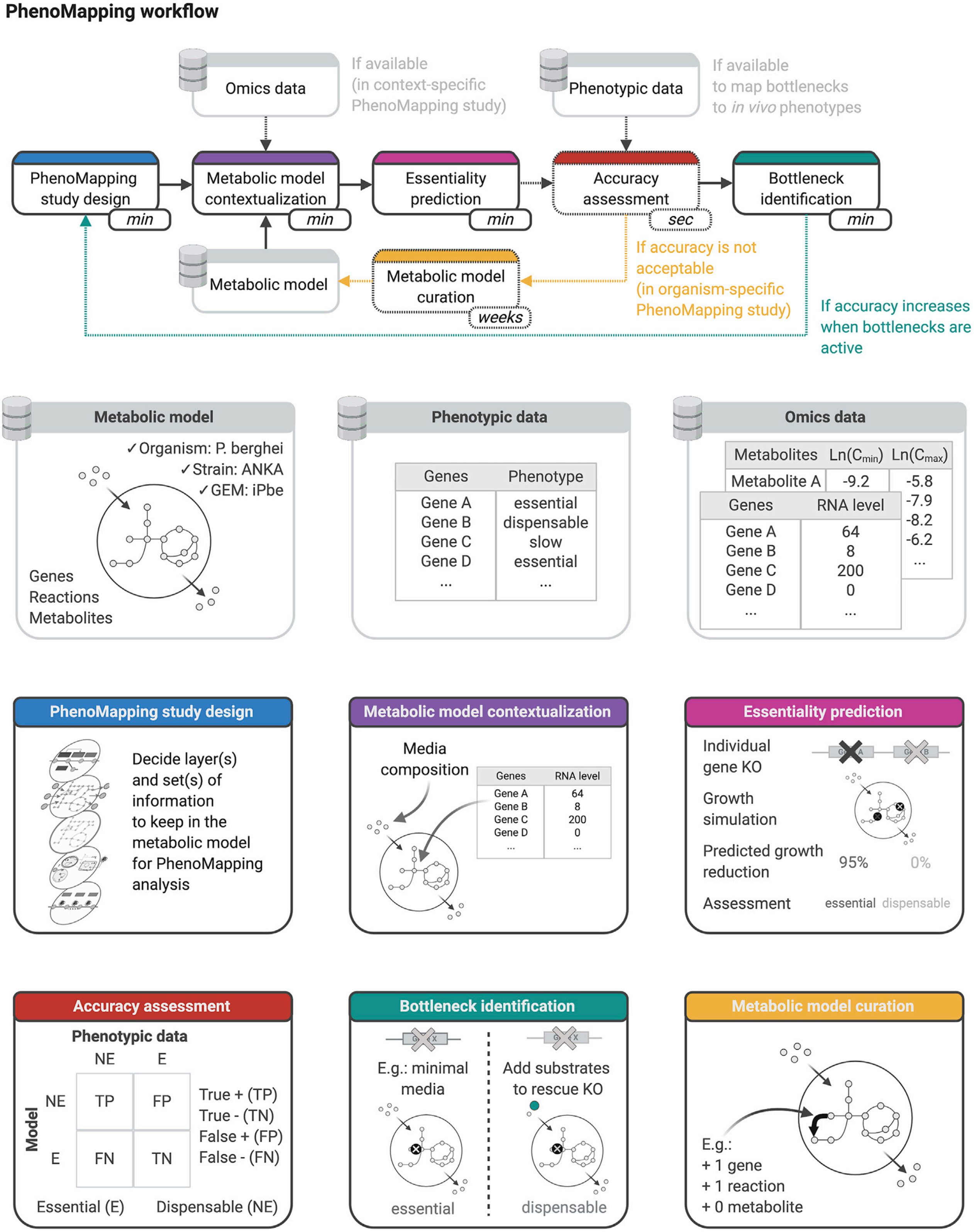
To enable identification of cellular processes underlying phenotypes, PhenoMapping leverages all available biochemical information of an organism as integrated into a GEM, as well as omics (e.g., metabolomic, transcriptomic) and phenotypic data in one or more conditions or life stages. These measurements are used along with the metabolic model of the organism of interest to study context-specific metabolic function and essentiality, identify sets of conditions that explain phenotypes, and if necessary further curate the GEM. Comparison between essentiality predictions and phenotypic data allows to assess accuracy of the GEM. Some of the steps have been extensively described previously, and some were recently first introduced (Chiappino-Pepe et al., 2017; Stanway et al., 2019). Here, we present a comprehensive protocol describing the proper and practical integration of all relevant PhenoMapping steps, as well as advice on checks and troubleshooting, to allow efficient and accurate analysis of origin of phenotypes and curation of GEMs (Figure 2). We define both **setup** and **analysis steps**. In a **setup step**, we conceptualize a study or perform changes in the GEM that do not involve any analysis. These steps are GEM- and case-specific and require mental or manual work. In an **analysis step**, we perform actual analysis on the GEM. All **analysis steps** are automatized within PhenoMapping. We emphasize the applications of this protocol to study eukaryotic pathogens like malaria (*Plasmodium*) and toxoplasma (*Toxoplasma*) parasites for which there is a higher uncertainty in the metabolism and growing conditions. This protocol can be easily adjusted for other complex eukaryotic organisms like human cells and also prokaryotic systems.

PhenoMapping study design

Timing: 1–30 min

The PhenoMapping workflow is summarized in Figure 2. The first step is a **setup step** and involves the design of a PhenoMapping analysis. PhenoMapping classifies the information included in a GEM in two classes: organism-specific and context-specific information. Each class has also its layers of information that correspond to constraints in a GEM with different types of physico-chemical meaning and hierarchy (Figure 3). In the PhenoMapping study design, we use the knowledge of these layers to select pieces of information and data (among all datasets collected in the Before you begin section) to integrate into a GEM for the next analyses.

Note: the hierarchy of the organism-specific layers in PhenoMapping makes it possible to distinguish between two classes of incorrect essentiality predictions or false negatives: those arising due to a lack of information in the model, like missing genome annotations, and those arising due to an incorrectly defined pre-assumed transport/reaction directionality or enzymatic irreversibility (*ad hoc* constraints). PhenoMapping suggests adding first all possible missing gene annotations and metabolite transports, and later introducing *ad hoc* irreversibility constraints when needed. The hierarchy of the context-specific layers in PhenoMapping is suggested based on the uncertainty of the data and methodology to integrate such data into the GEM. For example, data on media composition tend to be more reliable than



Genes	Phenotype
Gene A	essential
Gene B	dispensable
Gene C	slow
Gene D	essential
...	...

Metabolites	$Ln(C_{min})$	$Ln(C_{max})$
Metabolite A	-9.2	-5.8
		-7.9
		-8.2
		-6.2
		...

Genes	RNA level
Gene A	64
Gene B	8
Gene C	200
Gene D	0
...	...

		NE	E
Model	NE	TP	FP
	E	FN	TN

Figure 2. The PhenoMapping workflow showing steps (colored boxes) and input data (boxes marked with databases)
A GEM is a necessary input (solid arrow) to the workflow and phenotypic and omics data are optional inputs (dashed arrows). When phenotypic data are not available a purely *in silico* analysis of predicted phenotypes and bottlenecks will be performed. The PhenoMapping workflow involves five steps to map phenotypes to bottlenecks: PhenoMapping study design (blue), metabolic model contextualization (lila), essentiality prediction (pink), accuracy assessment (red), and bottleneck identification (green). An additional step can be added to curate the metabolic model if needed (yellow). The PhenoMapping workflow is often iterative (feedback loop). We include an approximate assessment of the timing each step takes.

measures of RNA levels. In addition, simulating the effect of a lack of substrate in the medium is more straightforward than simulating the effect of an RNA level on the cellular physiology using a GEM.

1. Select the type of PhenoMapping analysis to perform:
   a. *Organism-specific* PhenoMapping analysis to identify unconditionally essential genes and curate a generic GEM. This analysis also maps phenotypes to the following layers of information: (1) metabolic functions annotated to the genome, (2) enzyme localization, (3) transportability of metabolites between intracellular compartments, (4) enzymatic irreversibility, and (5) a set of metabolic tasks related to biomass production (Figure 3).
   b. *Context-specific* PhenoMapping analysis to study conditional essentiality and generate a context-specific GEM. This analysis also maps phenotypes to the following layers of information: the (6) media composition or uptakes, (7) thermodynamic feasibility at some given intracellular conditions including metabolite concentrations, (8) gene expression, and (9) transcriptional regulation or regulation of expression between isoenzymes (Figure 3).

2. Within an organism-specific or context-specific analysis, select the layer(s) of information to keep within the GEM:
   a. *Biochemistry layer.* Analysis of the biochemistry layer serves to identify essential genes as defined uniquely by the genome annotation and metabolic capabilities included in the GEM. This analysis will not account for any physiological constraint in cellular metabolism and hence allows to identify metabolic gaps purely due to missing functional annotations.
   b. *Localization layer* (in eukaryotes). A comparative analysis between the biochemistry and localization layer will identify genes that become essential due to compartmentalization of enzymes and metabolic pathways. The localization layer includes (on the top of the biochemistry layer) localization of enzymes and metabolites, and allows transport of all metabolites without a phosphate, acyl-carrier protein (ACP), and CoA moiety between cytosol and other compartments. If there is experimental evidence that there exists a transporter for a metabolite with a phosphate, acyl-carrier protein, and CoA moiety, this should be allowed.

   Note: there exists arguably some uncertainty in transport mechanisms and annotated transporters in cells and cellular organisms. In PhenoMapping, as done before (Chiappino-Pepe et al., 2017; Krishnan et al., 2020; Stanway et al., 2019; Tymoshenko et al., 2015), we assume that any metabolite that contains a phosphate, acyl-carrier protein (ACP), and CoA moiety might not easily diffuse (by simple diffusion) through lipid bilayer membranes and requires a specialized transporter or transport mechanism. The exception is free phosphate that cells normally take up.

   c. *Intracellular transportability layer* (in eukaryotes). A comparative analysis between the localization and intracellular transportability layers will highlight genes that become essential when there exist constraints on intracellular transportability. Beside the localization information, this analysis will integrate *ad hoc* directionalities for transporters or blocked transports.
   d. *Enzymatic irreversibility layer.* Analysis of the enzymatic irreversibility layer can be compared to the biochemistry (in prokaryotes) or intracellular transportability (in eukaryotes). Such comparison suggests genes that become essential due to irreversible biotransformations (Ataman and Hatzimanikatis, 2015). This layer includes *ad hoc* and pre-assigned reaction directionalities that should be applicable in every growing context and life-stage.

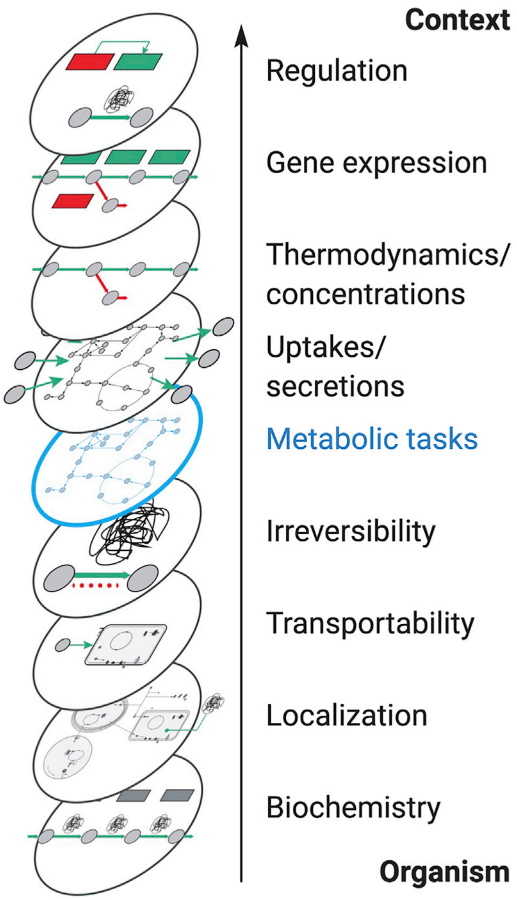


Figure 3. Layers of information or physico-chemical constraint types in a GEM and their hierarchy as suggested within PhenoMapping

Note: many GEMs tend to include pre-assumed reaction directionalities as *ad hoc* reaction bounds. While this approach might increase the prediction accuracy at a specific growth condition, it can also limit the usage of such GEMs in a different scenario and the identification of actual bottlenecks responsible for a phenotype. For example, there might be a set of metabolites whose concentrations are responsible for those reaction directionalities (see Metabolomics layer); we would not identify these bottleneck metabolites when the source of the reaction directionality is not TFA but *ad hoc* reaction bounds. We recommend defining generic GEMs with the minimum information on a context such that they serve as platforms for integration of media composition and omics data and generation of context-specific GEMs. Such strategy was followed before (Krishnan et al., 2020; Stanway et al., 2019) with the generation of generic GEMs like iPbe and iTgo and context-specific GEMs like iPbe-blood, iPbe-liver, and iTgo-tachy.

e. *Metabolic tasks.* Metabolic tasks serve to evaluate in a modular way how metabolism works, as described before (Agren et al., 2013; Carey et al., 2017; Chiappino-Pepe et al., 2017; Richelle et al., 2019; Tymoshenko et al., 2015; Wang et al., 2018). In an analysis of metabolic tasks, we define a set of input molecules (extracellular nutrients or intracellular precursors) and expected output molecules (biomass precursors or expected end products of a metabolic pathway). Next, we evaluate whether the task is feasible. A task is feasible when it is possible to produce all output molecules using the input molecules. We next evaluate what metabolic pathway was used and which genes are essential to fulfill a task.

f. *Media layer.* Analysis of *in silico* minimal media allows to study nutritional requirements, evaluate substrate substitutability, and identify genes that become essential upon substrate inaccessibility. The minimum sets of substrates that rescue essentiality when added to an *in silico* minimal medium are bottleneck substrates (Stanway et al., 2019). The media layer study comprises a systematic analysis of *in silico* minimal media, essentiality at each minimal medium, and identification of bottleneck substrates.

Note: media analysis with PhenoMapping is especially useful in organisms for which the growing conditions (media composition) and nutritional requirements is uncertain. This is the case in intracellular parasites (Chiappino-Pepe et al., 2017; Krishnan et al., 2020; Stanway et al., 2019; Tymoshenko et al., 2015).

g. *Metabolomics layer.* Thermodynamics-based flux analysis will pinpoint genes that become essential due to a set of reaction directionalities imposed by thermodynamic constraints. It is possible to identify sets of metabolites whose concentration ranges determine such reaction directionalities and these are called bottleneck metabolites (Chiappino-Pepe et al., 2017). The metabolomics layer study involves a systematic integration of metabolomics data within the TFA framework (Salvy et al., 2018), thermodynamically consistent essentiality analysis with or without metabolomics, and identification of bottleneck metabolites.

h. *Transcriptomics layer.* Integrative analysis of RNA-seq data helps to identify genes that become essential due to gene expression constraints. TEX-FBA (Pandey et al., 2019) will try to maximize consistency between RNA levels and metabolic reaction fluxes. There will be three classes of genes: highly, medium, and lowly expressed (defined by TEX-FBA parameters; Troubleshooting 8). For reactions linked to highly expressed genes, TEX-FBA tries to increase metabolic flux. For reactions uniquely linked to lowly expressed genes, TEX-FBA tries to minimize metabolic flux. The maximum number of such type of agreements counts for a consistency score. Reaction fluxes linked to genes with medium expression are free to vary. PhenoMapping uses TEX-FBA to integrate RNA-seq data and calculate a maximum consistency score. It next performs essentiality analysis at the maximum consistency score and identifies the metabolic fluxes that are responsible for a gene essentiality, also called bottleneck reactions (Stanway et al., 2019).

i. *Regulation layer.* Transcriptomics data analysis also allows identifying isoenzymes that become essential due to lack of transcriptional regulation of counterpart isoenzymes or bottleneck isoenzymes. PhenoMapping regulation analysis include a systematic integration of transcriptomics data within the TEX-FBA framework (Pandey et al., 2019), essentiality analysis with transcriptomics considering lack of regulation between isoenzymes, and identification of bottleneck isoenzymes.

Metabolic model contextualization

⏲ Timing: 10–30 min

The second step in the PhenoMapping workflow (Figure 2) is a **setup step** and involves the contextualization of the GEM as designed in the first step. Predictions from a GEM are the consequence of the biochemical information and physico-chemical constraints integrated into the GEM (Figure 3). Here, we define the protocol to define each layer of information in the GEM. We assume that the initial GEM already includes all the corresponding organism-specific information and putatively context-specific information like a medium composition.

3. Select one step between the followings to perform a PhenoMapping analysis at each iteration.
   a. Generate a GEM with the ground *biochemistry layer.*
      - ■ Remove all constraints related to omics data integrated into the GEM.
      - ■ Define a rich medium and allow uptake and secretion of all metabolites in the medium.
      - ■ Remove *ad hoc* reaction (intracellular reactions and transports) directionalities in the GEM.

- If applicable (eukaryotic organism), remove compartmentalization. This is done by defining all reactions in the cytosol or by allowing all metabolites to be transported and present in all intracellular compartments.

b. Generate a GEM with *localization layer* (in eukaryotes).
- Remove all constraints related to omics data integrated into the GEM.
- Define a rich medium and allow uptake and secretion of all metabolites in the medium.
- Remove *ad hoc* reaction (intracellular reactions and transports) directionalities in the GEM.
- If applicable (eukaryotic organism), allow all metabolites without a phosphate, acyl-carrier protein (ACP), and CoA moiety to be transported between cytosol and other compartments. If there is experimental evidence about a transporter for a metabolite with a phosphate, acyl-carrier protein, and CoA moiety, this should be allowed.

c. Generate a GEM with *intracellular transportability layer* (in eukaryotes).
- Remove all constraints related to omics data integrated into the GEM.
- Define a rich medium and allow uptake and secretion of all metabolites in the medium.
- Remove *ad hoc* directionalities only for intracellular reactions in the GEM.

Note: do not modify the directionalities of intracellular metabolite transportability from the initial GEM.

d. Generate a GEM with *enzymatic irreversibility layer*.
- Remove all constraints related to omics data integrated into the GEM.
- Define a rich medium and allow uptake and secretion of all metabolites in the medium.

e. Prepare a GEM for *metabolic tasks* analysis.
- Decide whether an analysis with or without thermodynamic constraints should be performed.
- Select the sets of metabolites whose production you want to test. By default, this will be all biomass building blocks.
- Define an essentiality threshold (see section Essentiality definition).

f. Generate a *thermodynamically curated GEM* for subsequent context-specific PhenoMapping analysis.
- Remove all constraints related to omics data integrated into the GEM.
- Define a rich medium and allow uptake and secretion of all metabolites in the medium.
- Define thermodynamically relevant information for each intracellular compartment: pH, generic metabolite concentrations (minimum and maximum allowed values), generic ionic strength (unique value), membrane potential.
- Curate the GEM thermodynamically within the TFA framework.

g. Generate a GEM for a *media* analysis.
- Select whether analysis of uptakes or secretions should be performed.
- Select the sets of transports of extracellular metabolites among which the minimal uptake or secretion analysis will be performed; Troubleshooting 4. By default, all substrates in the media will be selected. We indicate below two types of analysis (targeted or untargeted) that one can perform depending on the substrates made available and the uptakes selected for media analysis.

⚠ CRITICAL: it is important to note that the algorithm will not unblock uptakes or secretions in the GEM. If uptakes and secretions were blocked in the GEM input to the PhenoMapping media analysis, they will remain blocked in the media analysis. Hence, it is important to properly define a media composition before one begins the PhenoMapping analysis (see section Medium definition).

Recommended: For an untargeted analysis of *in silico* minimal media, one should have defined a rich medium in the GEM (see the Medium definition section; Troubleshooting 4). At this stage, all uptakes should be selected for the media analysis. Following this setup, one will identify all alternative minimal sets of molecules required for *in silico* growth in the correct combination. It was shown before (Chiappino-Pepe et al., 2017) that such an analysis provides further understanding of the molecular substructures or backbone moieties that a cell needs to scavenge. The requirement to scavenge such moieties occurs when the biochemical information and further physio-chemical constraints defined in the GEM (that probably represent the metabolic function of the organism) do not allow the biosynthesis of such backbone moieties.

Alternatives: For a targeted analysis of *in silico* minimal media, one should have defined a specific medium in the GEM before beginning the PhenoMapping analysis. In this medium the GEM should be feasible. At this stage, one defines the subset of substrates of interest for the media analysis. This analysis identifies within the subset of substrates, the minimum number of substrates required to achieve a minimum value of the objective (e.g., growth).

- Perform the *in silico* minimal medium analysis.
- Define in the GEM a combined minimal medium comprising all substrates identified across all alternative *in silico* minimal media. We use minimal media alternatives that contain the same number of substrates.

⚠ CRITICAL: defining here a combined minimal medium simplifies the process to infer the medium of a context-specific GEM. Check the section Quantification and statistical analysis (Results of a PhenoMapping analysis of bottleneck substrates) for more details on the importance of this last step to optimally guide the definition of the media in the context-specific GEM.

h. Generate a GEM for a *metabolomics* analysis.
   - Integrate the metabolomics dataset into the GEM.
   - Verify that the GEM is feasible within TFA when metabolomics data are integrated; Troubleshooting 9.

i. Generate a GEM for a *transcriptomics* and *regulation* analysis.
   - Decide whether or not to plot the distribution of gene expression values.
   - Select the TEX-FBA parameters defining the percentile of lowly and highly expressed genes in the distribution of gene expression values; Troubleshooting 8.
   - Select the TEX-FBA parameters defining the bounds assigned to lowly and highly expressed reactions; Troubleshooting 8.
   - Select the reactions for which gene expression constraints should not be defined
   - Decide which transcriptomics profile an output GEM should include.

Note: upon integration of transcriptomics data, TEX-FBA will identify all alternative transcriptomic profiles that render a maximum consistency score between gene and reaction levels. One can select one specific transcriptomic profile for the subsequent analysis. Alternatively, one can also select a combined expression profile, which will account uniquely for the expression constraints common to all transcriptomic profiles.

   - Integrate the transcriptomics dataset into the GEM.

j. Define the following inputs common to any context-specific PhenoMapping analysis.
   - Remove all constraints related to omics data integrated into the GEM. This is not necessary if one uses a generic GEM.
   - Define the expected value of the selected objective function (normally growth) at the conditions to study.
   - Define the selected essentiality threshold (see section Genetic background and essentiality definition).

Note: a value of the selected objective function and essentiality threshold will be used to identify a minimum required objective value. For example, in the media analysis, we first identify the *in silico* minimal media or the minimum number of substrates required to achieve at least a required value of the objective. Such value is given by the input values of the objective function and essentiality threshold.

- Select a time limit (in seconds) for the CPLEX solver. By default, none.
- Define whether one wants to identify uniquely alternatives for the optimal value of the objective function (preferred). Alternatively, one can look for suboptimal solutions. For example, one can find an *in silico* minimal media with 19 substrates and identify all alternative combinations of 19 substrates that allow growth. One can also identify alternative combinations with 20 or more substrates.
- Select a maximum number of alternative solutions to obtain. This is applicable every time a mixed integer formulation is defined. For example, for the identification of alternative *in silico* minimal media and alternative bottleneck substrates.

Note: it is preferred to select a high number of alternatives like 5,000; check Troubleshooting 10 for suggestions when the optimization crashes or the number of alternatives selected was not enough.

Essentiality prediction

Timing: 1–10 min

The third step in an **analysis step** and involves a prediction of gene essentiality with the contextualized GEM. If one performs a context-specific PhenoMapping analysis, the set of essential genes will be compared with the unconditionally essential genes or genes predicted as essential with the general GEM input to the PhenoMapping workflow. If one performs an organism-specific PhenoMapping analysis, one might compare the set of essential genes with the ones obtained in the immediate previous layer of information. Predictions of gene essentiality are expected to vary between the contextualized GEMs. Here, we define the suggested steps to identify essential genes in any GEM within PhenoMapping.

Note: check Troubleshooting 11 if infeasibilities arise when calculating essentialities.

Optional: In a transcriptomic analysis, one can perform two types of essentiality analysis consistent with transcriptomics: (a) fix a unique transcriptomic profile; this is done by fixing the integer variables linked to all up and down reaction levels; (b) fix an assembly of transcriptomic profiles that satisfy a maximum consistency score. The maximum consistency score is a variable within TEX-FBA and defines the number of reactions that can carry low or high fluxes and are consistent with the classification of lowly and highly expressed genes, respectively. There may exist multiple alternative transcriptomic profiles that share the same maximum consistency score. These alternatives are an assembly of transcriptomic profiles. To fix such an assembly, we define the lower bound of the maximum consistency score with a value that is some decimals below the optimal objective value. This is to avoid problems with the precision of the solver; Troubleshooting 12. To provide some flexibility around the transcriptomic profiles, one can further relax the lower bound of the maximum consistency score by some integers. This option-b only differs from option-a if there is more than one alternative transcriptomic profile for the maximum consistency score within TEX-FBA.

4. Define the objective function chosen for the GEM; Troubleshooting 5.
5. Double check that the contextualized GEM is feasible; Troubleshooting 6.
6. Perform the *in silico* essentiality study.

Optional: An essentiality analysis per growth associated metabolic task can be performed. This will identify which biomass building block is responsible for the observed essentiality (Chiappino-Pepe et al., 2017) (Figure 4).

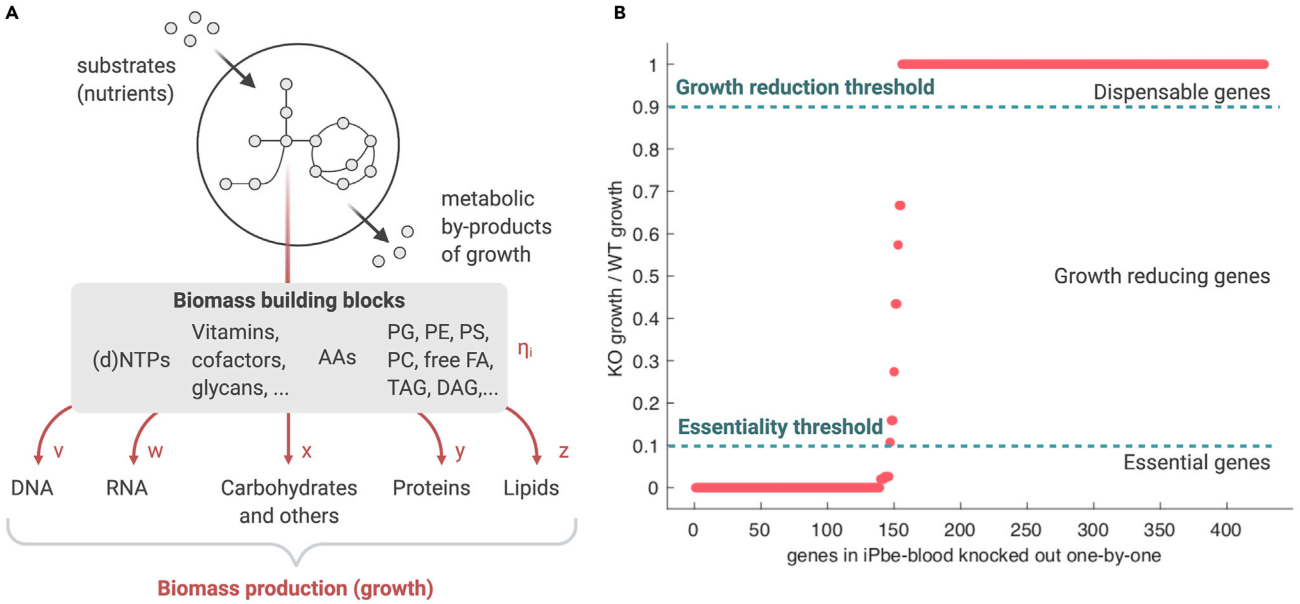


Figure 4. Schema of growth simulation using a GEM

(A)The GEM uses a set of substrates or nutrients to produce molecules required for growth or biomass building blocks in the stoichiometrically required amounts (η_i). Biomass building blocks are monomers of macromolecules required for the cellular function. The stoichiometric coefficients of the biomass building blocks satisfy the concentration of macromolecules in the cell (v, w, x, y, z).

(B) Predicted growth upon single knockout of each gene in the blood-stage-specific *P. berghei* GEM (iPbe-blood). We classify genes based on the essentiality threshold (dashed line bottom) and growth reduction threshold (dashed line top) into essential, growth reducing, and dispensable. The essentiality threshold (here 10%) and growth reducing threshold (here 90%) define which genes are essential and growth reducing, respectively, based on the predicted growth upon knockout (KO growth) compared to the predicted wild-type growth (WT growth). iPbe-blood predicts 146 essential genes, 9 growth reducing genes, and 273 dispensable genes (with solver version CPLEX 12.8.1 or above). Acronyms: AAs, amino acids; PG, phosphatidylglycerol; PE, phosphatidylethanolamine; PS, phosphatidylserine; PC, phosphatidylcholine; free FA, free fatty acids; TAG, triacylglycerol; DAG, diacylglycerol.

7. Select an output to identify essential genes:
   a. Ratio of optimal value of the objective function between the input (normally wild-type) GEM and the single gene knockout GEM.
   b. Absolute value of the objective function in the single gene knockout GEM.

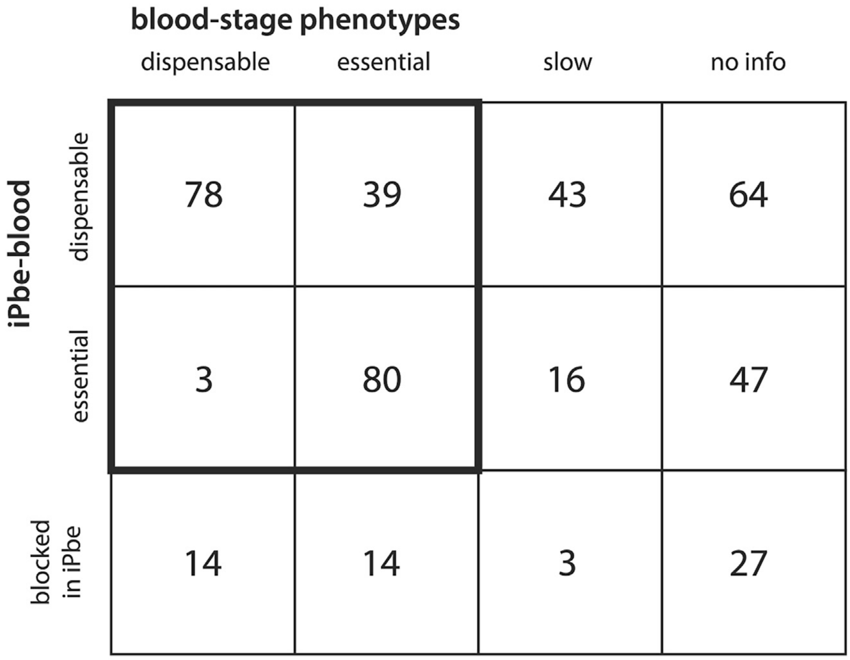
   Note: both approaches if compared with the corresponding reference values (as described in the next point) ultimately result in the same set of essential genes. However, selecting the absolute value of the objective function to identify essential genes allows to define an arbitrary value of the objective function as reference. This latter option might be more appropriate in a situation of high uptake rates (very unconstrained model) and high growth.

8. Identify all knockouts rendering a value below the essentiality threshold chosen (Figure 4)
   a. If the ratio is below the essentiality threshold the gene is considered as essential.
   b. If the predicted value of the objective function is below the product of the essentiality threshold and the initial value of the objective function the gene is considered as essential.

   Note: by default, in PhenoMapping all infeasible solutions (NaN) upon a gene knockout consider the gene essential. However, a lack of convergence in the optimization and problems with the solver might also render infeasible solutions; Troubleshooting 12.

Accuracy assessment

Timing: 1–5 min



		blood-stage phenotypes			
		dispensable	essential	slow	no info
iPbe-blood	dispensable	78	39	43	64
	essential	3	80	16	47
	blocked in iPbe	14	14	3	27

Figure 5. Contingency matrix for the blood-stage-specific *P. berghei* GEM (iPbe-blood) compared to the blood-stage-specific PlasmoGEM phenotypes

The accuracy values for this contingency matrix are: MCC = 0.63, overall accuracy = 0.79, NPR = 0.96, PPR = 0.67, sensitivity = 0.96, and specificity = 0.67.

The fourth step is an **analysis step** and involves an accuracy assessment of the gene essentiality prediction with the contextualized GEM. This step is possible when there are phenotypic data available in the PhenoMapping workflow (Figure 2). In this step, a contingency matrix is generated (Figure 5) and the set of correct and incorrect predictions is identified.

9. Compare the list of *in silico* essential and non-essential genes with the experimentally observed (*in vivo*) phenotypes.
10. Classify all compared genes in the GEM in four groups: TPs, TNs, FPs, and FNs.

Note: if "slow" phenotypes are available, one should decide how to treat them, i.e., as essential, or dispensable. This decision might be determined by the layer of information analyzed within PhenoMapping. For example, during a PhenoMapping analysis of the biochemistry layer, slow phenotypes might be better considered as dispensable. This is because one expects that slow phenotypes arise due to the presence of a redundant and non-optimal function that can partially compensate for the loss of the slow-phenotype gene. However, during a PhenoMapping analysis of the transcriptomics layer, slow phenotypes might be well considered as essential. This is because a GEM with transcriptomics data integrated identifies genes that are essential to maintain the defined (optimal) transcriptomic state. Hence, knocking out such a gene might render a transition to a different (suboptimal) transcriptomic or physiological state.

Optional: one can also add a classification for genes without data, blocked, or with slow phenotypes (Figure 5). The genes classified as blocked and without data are not considered for the computation of the accuracy. The treatment of the slow phenotypes may vary depending on the context of the GEM.

Note: Blocked genes are genes linked to reactions that cannot carry any flux in the reference conditions, also called blocked reactions. This occurs when any of the metabolites participating in the linked reactions cannot be mass balanced. Blocked reactions are identified with a flux variability analysis (Mahadevan and Schilling, 2003). We recommend performing a flux variability analysis in the generic GEM (in a rich medium and without any data integrated) and without any growth requirement. That way, there are no conditional or context-specific constraints leading to the non-function of the gene.

11. Generate a contingency matrix by defining the number of TPs, TNs, FPs, and FNs.
12. Compute the selected metric to assess the accuracy using the numbers defined in the contingency matrix.

Note: there might be situations in which the number of available *in vivo* phenotypes is very low compared to the number of genes in the GEM or with *in silico* phenotypes. Such cases decrease the confidence on the accuracy of the model. Although there is little that a user can do to improve such a situation regarding the availability of phenotypes, the user can choose how to evaluate the FPs and FNs to better assess the accuracy of the GEM's predictions. FNs arise when the model misses biochemical information or gene annotations or has incorrectly defined constraints. FPs arise when the model misses the definition of a context or constraints. As previously mentioned in this protocol, one can argue that (if the data are fully trustable) having more FNs than FPs in a GEM is worse than having more FPs than FNs. This is because one does not want to include false constraints into the GEM. In many situations a single constraint in the GEM (as identified with PhenoMapping) can be responsible for the essentiality of a FN and a TN. In such a case, we recommend not blindly integrating such constraint to increase the TNs (since that will also increase the FNs). We recommend first adding missing information into the GEM like new biochemistry or gene annotations such that later the named constraint becomes responsible uniquely for the essentiality of the TN. This means we recommend focusing first on correcting or reducing FNs (increasing TPs) and then on reducing FPs (increasing TNs) using PhenoMapping. See the follow-up discussion in the section Quantification and statistical analysis. The fact that PhenoMapping maps phenotypes to constraints may also increase the confidence on the prediction of genes without phenotype. If a constraint is responsible for one or more TNs and a gene for which no *in vivo* phenotype is available, we might feel more confident on the essentiality of the gene – primarily if the TNs and the gene without phenotype share metabolic pathways or tasks.

Bottleneck identification

Timing: 5–60 min

The fifth and last step of the PhenoMapping workflow is an **analysis step** and involves the mapping of bottlenecks to phenotypes. After new genes are identified as essential in the contextualized GEM, PhenoMapping will identify the bottlenecks or underlying cellular processes responsible for that essentiality. This is done by performing one-by-one a knockout of the essential genes in the contextualized GEM and identifying the conditions that rescue growth (Figure 6). Here, we define the steps to identify bottlenecks as followed within the systematic bottleneck analysis of each layer of information.

13. Knockout the essential gene in the contextualized GEM.
14. Identify all alternative bottlenecks or the minimum set of information (e.g., substrates, metabolite concentrations, reaction levels) that should be relaxed to rescue the gene knockout (Figure 6).

Note: bottleneck substrates are those that can rescue essentiality of the gene when added to the *in silico* minimal medium. Bottleneck metabolites are those whose concentrations ranges

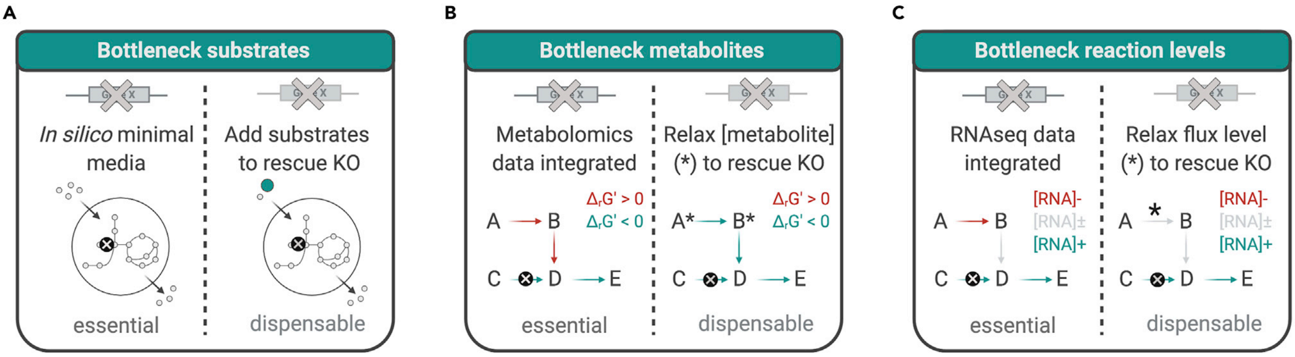


Figure 6. Representation of bottlenecks studies in PhenoMapping

(A–C) (A) Bottleneck substrates, (B) bottleneck metabolites, and (C) bottleneck reaction levels. PhenoMapping first simulates with the GEM some conditions (here: (A) *in silico* minimal media, (B) metabolomics data integrated, and (C) transcriptomics data integrated) and identifies *in silico* a phenotype (here single gene essentiality for growth). Next, PhenoMapping looks for the bottlenecks responsible for the predicted phenotype (here: (A) missing substrates in the media, (B) sets of metabolite concentration ranges, and (C) sets of levels of reaction fluxes and their corresponding RNA levels). The color code is consistent with the related step in the main PhenoMapping workflow (Figure 2).

should be relaxed (with respect to the experimentally measured concentration ranges) to rescue essentiality of the gene. Bottleneck reactions are those whose levels (considered to be high or low within the feasible flux range) should be relaxed to rescue essentiality of the gene.

Metabolic model curation

Timing: 1–60 days

Note: the timing to curate a GEM varies radically depending on the available GEM and data and the experience and endurance of the user. Many aspects of the GEM can require curation. The problems in a GEM can range from being badly elementally balanced to missing a considerable amount of gene annotations and associated reactions; see Troubleshooting 2 to identify all elements that an ideal GEM may include. This section aims to define a pipeline to spot and solve those problems faster.

This is a step that combines both **setup** and **analysis steps**. We curate a metabolic model when we change the biological and biochemical information it contains. Such information involves genes, gene functions, protein associations (protein complexes or isoenzymes), biochemical reactions, transporters, and biomass building blocks. Since GEMs are normally constructed following a bottom-up approach, it is more likely that a metabolic model curation involves adding missing information. However, curation of a GEM might also involve removing incorrectly defined *ad hoc* constraints.

The curation of the GEM is an optional step in the iterative PhenoMapping workflow (Figure 2). The information achieved by mapping *in silico* bottlenecks to phenotypes facilitates and accelerates the identification of missing biological and biochemical information in the GEM, as well as incorrectly defined *ad hoc* constraints.

A suggested workflow to curate the GEM using PhenoMapping is defined in Figure 7. This workflow is primarily manual. Here, we suggest conceptually how to perform a GEM curation in combination with the main PhenoMapping workflow (Figure 2). We propose analyzing one-by-one the incorrect gene predictions: first the FNs and then the FPs. One may follow the steps below in the order defined and select the steps depending on the type of inconsistency (FN or FP) for the gene of study. If the

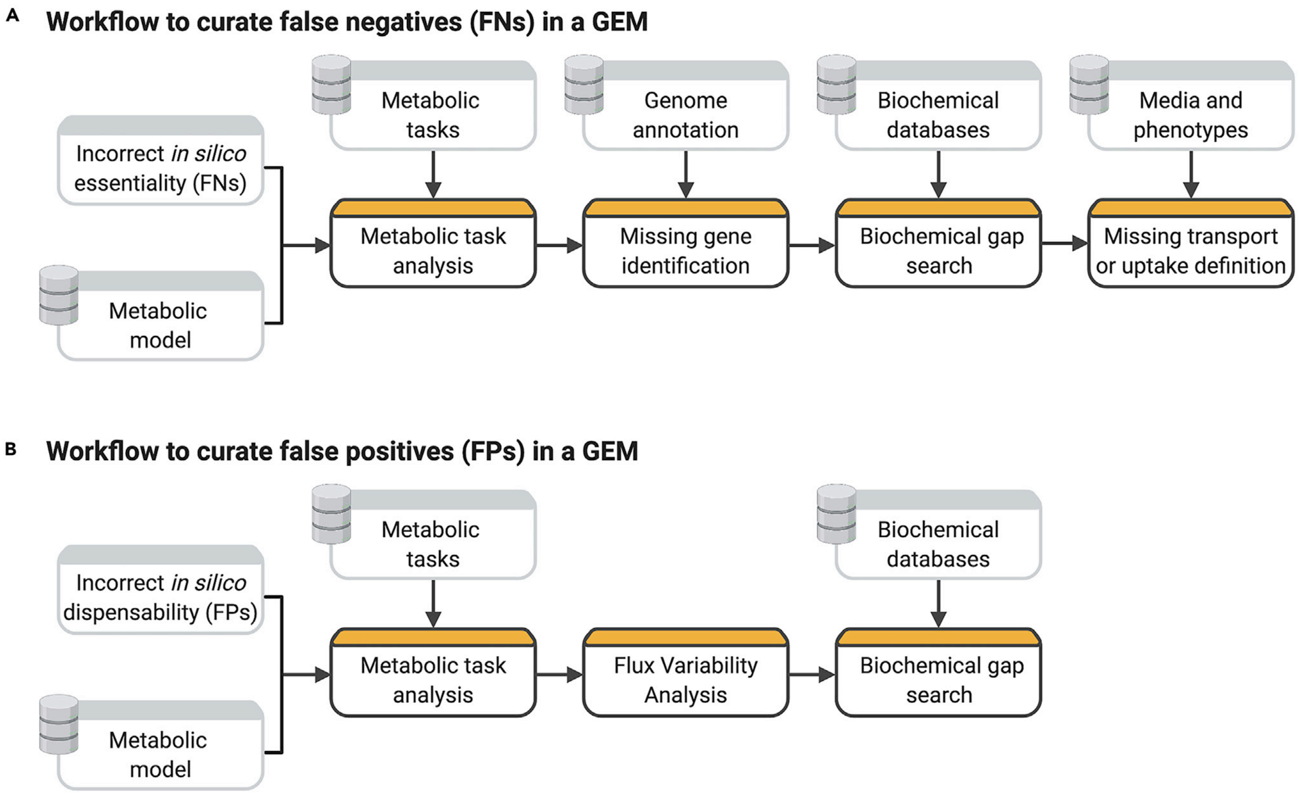

A
Workflow to curate false negatives (FNs) in a GEM
Incorrect in silico essentiality (FNs)
Metabolic model
Metabolic tasks
Genome annotation
Biochemical databases
Media and phenotypes
Metabolic task analysis
Missing gene identification
Biochemical gap search
Missing transport or uptake definition
B
Workflow to curate false positives (FPs) in a GEM
Incorrect in silico dispensability (FPs)
Metabolic model
Metabolic tasks
Biochemical databases
Metabolic task analysis
Flux Variability Analysis
Biochemical gap search

Figure 7. Suggested workflow to curate a metabolic model in combination with the PhenoMapping workflow

This workflow identifies missing biological and biochemical information in a GEM.

(A) The workflow to curate false negatives (FNs) involves four steps.

(B) The workflow to curate false positives (FPs) includes three steps. The curation of a GEM requires collecting and using different types of datasets. The color code is consistent with the related step in the main PhenoMapping workflow (Figure 2).

GEM is modified, one may perform a new essentiality analysis and accuracy assessment to evaluate the impact of the curation on the GEM performance.

Note: Adding information around true predictions may prevent mismatches in a more constrained scenario. For instance, one needs to identify isoenzymes linked to a reaction (even if the reaction is true positive) to prevent it from being false negative in a constrained scenario. When one supposedly has all information mapped to the GEM, we may wait until a false prediction arises to introduce corrective measures. False predictions arise normally in the PhenoMapping analysis with layers of information that are hierarchically higher. Alternatively, we may perform an unbiased integration of alternative information around true predictions and screen the performance of the model in a more constrained scenario for selection of the best corrective measure. This later option is not discussed here.

In this section, we do not consider the integration of context-specific information (like definition of uptake rates and integration of metabolomics and transcriptomics data) as part of the metabolic model curation. We consider that the integration of context-specific information is part of the metabolic model contextualization. The section Quantification and statistical analysis explains how to contextualize a GEM based on the bottleneck information from a context-specific PhenoMapping analysis.

Note: curating a metabolic model can be a daunting and time-consuming task. Be patient, do sports, eat healthy, and talk with friends and family to remain mentally ok.

15. (FN) Perform an essentiality analysis per metabolic task.
    a. If there is a metabolic task uniquely responsible for a set of FNs and no true prediction, remove the biomass building block from the biomass reaction.
    b. If a biomass building block was removed, update the stoichiometric coefficients of the remaining biomass building blocks accordingly (Chan et al., 2017).
16. (FN) Perform a reannotation of the genome defining more relaxed parameters, e.g., E-values, or look in databases for genes with the same function that are not part of the GEM.
    a. If there exists a potential gene with the same function, add the gene to the GEM with an OR relation in the gene-protein-reaction association.

17. (FN) Perform a gap-filling with the gene knocked out to identify missing alternative biochemistry in the GEM.
    a. Select a proper database to look for alternative biochemistry. We distinguish three classes of databases: (1) GEMs of closely related organisms, which can be found in databases for GEMs like BiGG (King et al., 2015), modelSEED (Devoid et al., 2013), KBase (US Department of Energy Systems Biology Knowledgebase, http://kbase.us), publications, etc.; (2) databases of biological reactions like KEGG (Kyoto University, 1995), MetaCyc (Caspi et al., 2018), BRENDA (Jeske et al., 2019), etc.; (3) the upper bound of biochemistry with hypothetical biochemical reactions between known compounds based on known enzyme reaction rules, i.e., the ATLAS of Biochemistry (Hadadi et al., 2016; Hafner et al., 2020).

    ⚠ CRITICAL: the compatibility of metabolite identifiers between the GEM and the database plays a critical role in the selection of the database. Metabolite identifiers need to match to assure the proper connectivity of the metabolic networks of the GEM and database. It is also important to consider which version of the database to use. We would recommend working with the latest version, but that might create conflicts with metabolite identifiers or other identifiers like genes. For this reason, the user might consider working with an earlier version.

    b. Identify a gap-filler that suits the GEM, database, computational power available, and desired gap-filling strategy.
        ■ There exist multiple examples of gap-fillers, as summarized before (Pan and Reed, 2018). Some recent examples are: gapseq (Zimmermann et al., 2020) or OptFill (Schroeder and Saha, 2020).
    c. If there exists an alternative biochemistry that rescues the knockout, integrate it into the GEM.

18. (FN) Investigate the possibility of a metabolite in the GEM being scavenged to rescue the KO.
    a. If there is evidence that the metabolite selected might be available at the cellular state studied, and the transport of such metabolite is possible (by any transport mechanism), define the transport in the GEM.
19. (FP) Search for a missing metabolic task downstream of the FP gene.
    a. If there is a downstream product that could be a biomass precursor and its definition as a metabolic task does not create inconsistencies, add it to the biomass reaction.
    b. If a biomass building block was added, update the stoichiometric coefficients of the remaining biomass building blocks accordingly (Chan et al., 2017).
20. (FP) Perform a flux variability analysis with a non-zero lower bound for the objective function.
    a. If the reactions linked to the gene cannot carry flux, perform a gap-filling (next step).
21. (FP) Perform a gap-filling with the objective function redefined to require flux through the FP gene.
    a. Select a proper database to search for missing biochemistry.
    b. Identify a gap-filler that suits the GEM, database, computational power available, and desired gap-filling strategy.

- There exist multiple examples of gap-fillers, as summarized before (Pan and Reed, 2018). Some recent examples are: gapseq (Chan et al., 2017) or OptFill (Schroeder and Saha, 2020).

c. If there exist biochemical steps that can connect the metabolic network defined in the GEM with the FP gene, define it.

Note: The order in which these steps are applied affects the definition of the GEM and later the identification of bottlenecks. We recommend following the order of steps defined in this section. The steps to curate FNs and FPs are defined in this order to identify first issues on which the user has more confidence. For example, to curate FNs we first check if there is an error in the objective function (a fact). Then we evaluate whether a gene is missing from the GEM (an E-value defines the confidence of the annotation). If no gene is found, we perform a gap-filling (a hypothetical non-annotated biochemical function). If no gap-filling reaction is found, we allow uptake of a metabolite (GEMs show the highest uncertainty in the definition of metabolite transports).

The PhenoMapping workflow is often iterative

Timing: variable

One would perform as many passages through the PhenoMapping workflow as layers of information one desires to analyze (feedback in Figure 2). The layers of information can be analyzed independently or in a cumulative fashion. An independent analysis is recommended for the first PhenoMapping iterations and in a generic GEM. A cumulative analysis is recommended after an independent analysis and in a context-specific GEM. The order of the analysis would be hierarchical as suggested in Figure 3 and argued in the section PhenoMapping study design. In addition, more iterations through the PhenoMapping workflow might be required when a curation of the GEM is selected.

22. Perform an independent, individual, and separate analysis of each layer of context-specific information to identify individual bottlenecks responsible for phenotypes. In an independent analysis of constraint types, a generic GEM is used to integrate each dataset independently, identify new essential genes, and map gene essentiality to bottlenecks.
23. Perform also a cumulative and hierarchical integration of constraint types (Figure 3). In such a cumulative analysis, new essentialities might be identified at each integration step and can be mapped to sets of constraints within the last layer of information considered. The independent analysis of constraint types may limit the set of predicted phenotypes. A cellular phenotype is the product of a cumulative rather than individual effect of physico-chemical constraints. Hence, a cumulative analysis is recommended to analyze a context-specific GEM.

EXPECTED OUTCOMES

Here, we present an example to aid in the definition of a PhenoMapping analysis and understanding of its output. We use the example of blood-stage and liver-stage *P. berghei* to illustrate the iterative workflow of PhenoMapping (Figure 8). The same analysis was performed for tachyzoite *T. gondii*. The metabolic models of *P. berghei* (iPbe) and *T. gondii* (iTgo), the *P. berghei* blood- and liver-stage relative growth phenotypes, the tachizoyte *T. gondii* genome-wide screen, and all integrated omics data are available in www.github.com/EPFL-LCSB/phenomapping. The scripts with all input values used for the analyses are also available in the GitHub repository. These files are enough to enable the user to follow along and repeat all computational steps of these examples with PhenoMapping. We present the results obtained for both examples. Additional analyses of the outputs were done as explained in the following section.

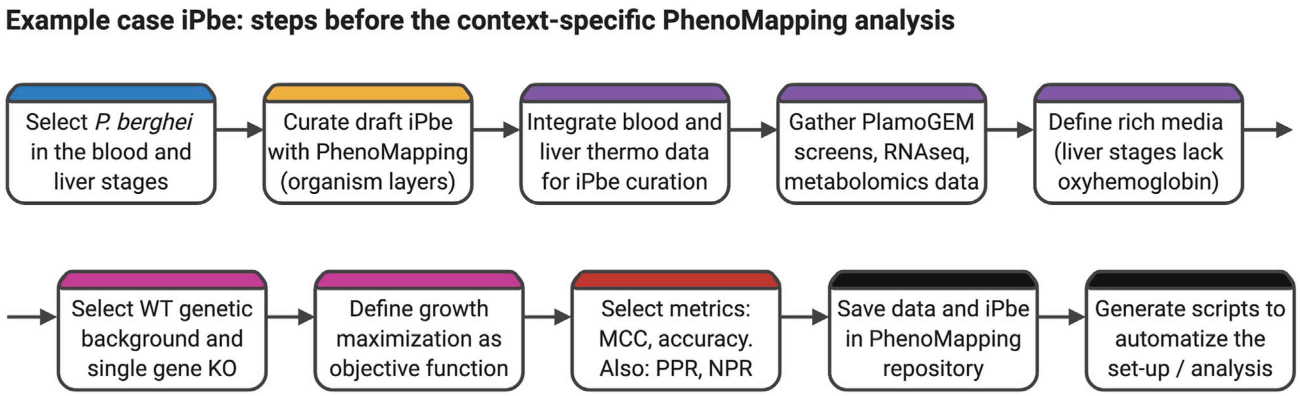

Example case iPbe: steps before the context-specific PhenoMapping analysis
Select *P. berghei* in the blood and liver stages
Curate draft iPbe with PhenoMapping (organism layers)
Integrate blood and liver thermo data for iPbe curation
Gather PlamoGEM screens, RNAseq, metabolomics data
Define rich media (liver stages lack oxyhemoglobin)
Select WT genetic background and single gene KO
Define growth maximization as objective function
Select metrics: MCC, accuracy. Also: PPR, NPR
Save data and iPbe in PhenoMapping repository
Generate scripts to automatize the set-up / analysis

Figure 8. Schema of preparatory steps as applied to study blood and liver-stage phenotypes with iPbe and PhenoMapping

The preparatory steps are shown in Figure 1. Color code is consistent with related steps in the main PhenoMapping workflow (Figure 2).

Preparatory steps for a PhenoMapping analysis of genome-wide blood and liver-stage phenotypes in *P. berghei*

All preparatory steps for the PhenoMapping analysis of iPbe were performed as follows (Figure 8), and the result (models and data) are available in www.github.com/EPFL-LCSB/phenomapping.

Organism and cellular state choice

In the PhenoMapping analysis of *P. berghei* two conditions were selected for study and comparison, i.e., the blood and liver stages of the malaria life cycle. The time points at each life-stage, i.e., 24 h in the blood stages and 48 h in the liver stages, were selected to represent the highest metabolically active state during each developmental stage. Similarly, in the PhenoMapping analysis of *T. gondii* the highly metabolically active state, i.e., tachyzoite, was selected.

Metabolic model selection

There was no GEM available for *P. berghei*, and we constructed a GEM as explained in detail before (Stanway et al., 2019). We used as reference a previously developed GEM for *P. falciparum* (iPfa) (Chiappino-Pepe et al., 2017) and applied the PhenoMapping principles discussed here and multiple iterations through the Metabolic model curation loop (Stanway et al., 2019). This construction rendered the life-stage agnostic GEM of *P. berghei* called iPbe.

Thermodynamic curation

We collected thermodynamic data for iPbe specific to the liver and blood stages (Table S3 in (Stanway et al., 2019)). We added such data in the "*CompartmentData*" field in iPbe.

We mapped metabolite SEED IDs where available to all metabolites, as required in the current version of matTFA (Salvy et al., 2018) for a thermodynamic curation of the GEM. We saved such data in the "metSEEDID" field in iPbe.

> *Note:* in the current PhenoMapping repository, a thermodynamic curation and setup of the GEM for PhenoMapping analysis can be done with the function *initTestPhenoMappingModel*. In such function, one should define the input *tagThermo* as 1 or true. An example is provided in the "*settings*" scripts (see section Software setup).

Phenotypic, media, and omics data collection

Phenotypic data: we used the values of relative growth rate upon single gene knockout obtained with the PlasmoGEM technology in the blood stages (Bushell et al., 2017) and liver stages (Stanway et al., 2019).

Metabolomics data: as an *in silico* study case, we used the values of metabolite concentration ranges obtained in *P. falciparum* (Teng et al., 2009, 2014; Vo Duy et al., 2012). However, we did not integrate such values in the final blood- and liver-stage specific GEMs.

Transcriptomics data: we used previously reported RNA-seq values for the blood-stage (Otto et al., 2014) and liver-stage (Caldelari et al., 2019) *P. berghei*. We computed the average of the replicates at the time point of highly active metabolic state.

Medium definition

For the blood-stage-specific PhenoMapping analyses, we allowed the uptake of all 248 metabolites present in the extracellular space of iPbe.

For the liver-stage-specific PhenoMapping analyses, we allowed the uptake of 247 metabolites, which includes all 248 extracellular metabolites in iPbe but oxyhemoglobin.

Genetic background and essentiality definition

We preserved the wild-type genetic background (without any gene knockout) for all context-specific PhenoMapping analyses. We selected a single gene knockout essentiality analysis.

Cellular objective definition

Since we selected a highly proliferative and metabolically active state both in the blood and liver stages, we defined as cellular objective the desire to maximize growth.

Accuracy metric definition

We used the MCC and overall accuracy as the final metrics of accuracy assessment. To generate iPbe-blood and iPbe-liver, we tried to increase the positive prediction rate while keeping the negative prediction rate constant. This rationality was followed in the selection of constraints to integrate into iPbe after each passage through the PhenoMapping workflow.

> *Note:* in the current PhenoMapping repository, an accuracy assessment can be done with the function *accuracyAssessment*. An example is provided in *test_accuracyAssessment*.

Data setup

GEM setup: we saved the two versions of iPbe, i.e., with thermodynamic and medium data for the blood and liver stages, in the "models" folder of the GitHub repository as "tipbe4blood.mat" and "tipbe4liver.mat."

We generated a "pbe" folder in the directory "tests/ref", where we saved all data in the format defined.

Software setup

We adapted the settings script to iPbe for blood and liver-stage analyses, i.e., *settings_ipbeblood*, *settings_ipbeliver*. These are saved in the "*tests*" folder. We next run them to add all relevant directories to the path and prepare iPbe for PhenoMapping.

PhenoMapping workflow applied to study genome-wide blood and liver-stage phenotypes in *P. berghei*

These steps and results could be generated automatically by running the *tutorial_basics* script in www.github.com/EPFL-LCSB/phenomapping v1.0 (Figure 9, entry point 1). One can also work with the individual test scripts for an analysis of each layer of information within PhenoMapping.

Running the "**analysis steps**" of the Step-by-step section (using the *tutorial_basics* script) for the liver-stage PhenoMapping analysis of iPbe takes approximately 4 h in a MacBook Pro 2.2 GHz Intel

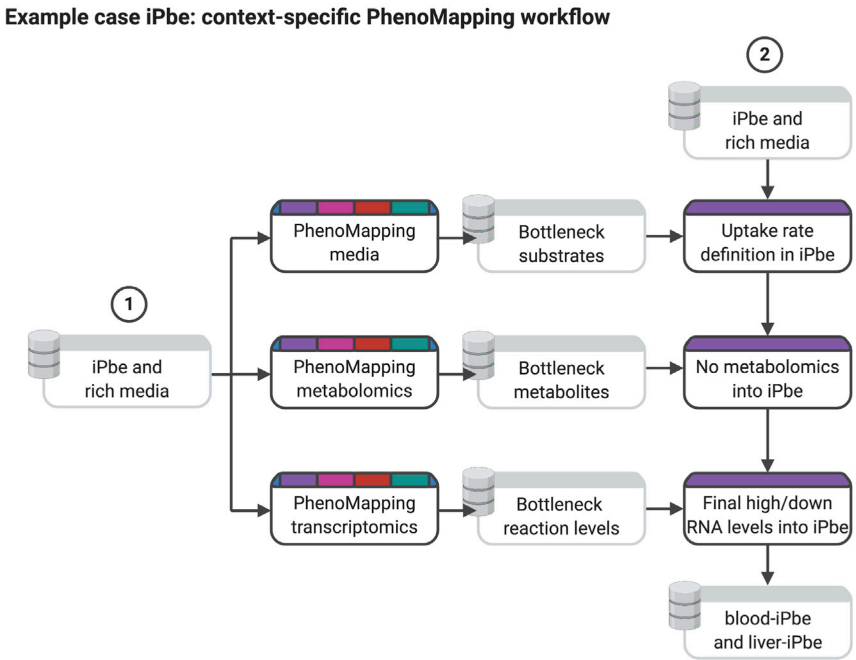

Example case iPbe: context-specific PhenoMapping workflow
2
iPbe and rich media
Uptake rate definition in iPbe
PhenoMapping media
Bottleneck substrates
1
iPbe and rich media
PhenoMapping metabolomics
Bottleneck metabolites
No metabolomics into iPbe
PhenoMapping transcriptomics
Bottleneck reaction levels
Final high/down RNA levels into iPbe
blood-iPbe and liver-iPbe

Figure 9. Schema of the PhenoMapping workflow

Schema of the PhenoMapping workflow applied to (1) analyze genome-wide blood and liver-stage phenotypes with iPbe and (2) generate the blood-stage-specific iPbe (blood-iPbe) and liver-stage-specific iPbe (liver-iPbe). We first performed an independent PhenoMapping media, metabolomics, and transcriptomics analysis to identify bottleneck substrates, metabolites, and reaction levels, respectively. The mapping of phenotypes to bottlenecks guided the definition of the media composition and consideration of highly and lowly expressed genes in the final life-stage specific models, as explained in the section Quantification and statistical analysis.

Core i7 16 GB 1,600 MHz DDR3. The time it takes to perform a PhenoMapping analysis is GEM-specific and it does not necessarily correlate with the size of the metabolic network but rather with the inherent complexity or connectivity of the network.

This script will save all results in a temporary result folder called "*tmpresults*." The script *test_io* can be used to extract all data from the temporary results folder and save it in a text format.

Generic considerations for all context-specific PhenoMapping analyses

The life-stage agnostic model iPbe does not contain any constraint related to omics data. Hence, we did not have to remove them.

We selected a value of 0.12 h^{-1} and 0.35 h^{-1} of optimal growth for the blood and liver analyses, respectively.

> *Note:* we calculated the growth value assuming that 16 and 30,000 merozoites of *P. berghei* are exponentially formed from a single cell in 24 and 30 h in the blood and liver stages, respectively.

$$gr = \frac{\ln \frac{N_t}{N_0}}{t}$$

, where N_t and N_o are the initial and final number of cells and *gr* is the growth rate.

The essentiality threshold was 10% or 0.1 across all analyses.

We did not select a CPLEX time limit.

We looked only for alternative solutions of optimal value (tagMin=1) when analyzing *in silico* minimal media. The analysis of *in silico* minimal media is a mixed integer linear programming (MILP) problem. This means we identified all alternative *in silico* minimal media with the same number of substrates. Such number is the minimal possible to achieve at least 10% (value of the essentiality threshold) of the optimal growth.

The bottleneck analyses in PhenoMapping are also MILP formulations and we looked for *all* alternative bottleneck sets (of optimal and suboptimal integer count). This is done by default in the identification of bottleneck substrates and reaction levels, and we specified it (tagMax=0) in the analysis of bottleneck metabolites.

We selected a maximum value of 5,000 alternatives.

Results of a single gene knockout analysis in the life-stage agnostic iPbe

A single gene knockout in iPbe with TFA yields 109 *in silico* essential genes, as obtained with the *test_core_essentiality* script. These genes are considered unconditionally essential.

> ***Note:*** in the current PhenoMapping repository, an essentiality analysis can be done with the function *thermoSingleGeneDeletion*. An example is provided in the *test_core_essentiality* script. Each PhenoMapping layer calls this function automatically within its own set of functions. The *thermoSingleGeneDeletion* function performs a prediction of essential genes using the FastSL formulation (Pratapa et al., 2015) and additionally accounting for thermodynamic constraints.

Results of a PhenoMapping analysis of bottleneck substrates

The PhenoMapping results for a bottleneck substrate analysis are saved in the form of tables, as presented in Table 1. This table is a subset of all bottleneck results obtained with the *test_core_substrates* and *test_core_substrates_joint* scripts. In the first column, there is a list of genes identified as essential at *in silico* minimal media. The second and third column map the experimentally observed phenotype to the *in silico* essential gene. The subsequent columns are alternative solutions of bottleneck substrates.

Bottleneck substrates are those that can rescue the essentiality of the gene (column 1) when added to the *in silico* minimal medium. Hence, the absence of all of those alternative sets of substrates in the minimal medium is responsible for the essentiality of the gene.

Table 1. Subset of PhenoMapping result of bottleneck substrates

Essential gene at IMM[a]	Blood phenotype[b]	Liver phenotype[c]	Bottleneck substrates, Alt 1	Bottleneck substrates, Alt 2	Bottleneck substrates, Alt 3
PBANKA_0516900	essential	no info	spermine	S-adenosyl-methioninamine| N-methyl-putrescine	aminopropyl-cadaverine| N-methyl-putrescine
PBANKA_0522400	Dispensable	essential	(9Z)-octadecenoic acid		
PBANKA_0608000	Dispensable	slow	coproporphyri-nogen III	heme	protoporphy-rinogen IX
PBANKA_0613600	Dispensable	dispensable	pantetheine	N-((R)-pantothe-noyl)-L-cysteine	
PBANKA_0621800	Slow	dispensable	14-dihydroxy-2-naphthoate	2-succinyl-benzoate	

[a] Essential gene at *in silico* minimal medium.

[b] PlasmoGEM phenotypes for blood-stage development of *P. berghei*.

[c] PlasmoGEM phenotypes for liver-stage development of *P. berghei*.

Note: an example of PhenoMapping analysis of the media is provided in the scripts *test_core_substrates* and *test_core_substrates_joint*. These scripts include functions for analysis of *in silico* minimal media (function: *analysisIMM*), generation of alternative solutions (function: *findDPMax*), identification of essential genes at each *in silico* minimal medium (function: *getEssGeneIMM*), and identification of bottleneck substrates (function: *linkEssGeneIMM2Subs*). All substeps defined in the section Metabolic model contextualization are inputs to these functions. These inputs are also defined in the scripts to facilitate their identification.

Note: all results are available in Table S3.4 from (Stanway et al., 2019).

Results of a PhenoMapping analysis of bottleneck metabolites

The PhenoMapping results for a bottleneck metabolite analysis look as presented in Table 2. In the first column, there is a list of genes identified as essential when metabolomics data are integrated. The second and third column map the experimentally observed phenotype to the essential gene. The subsequent columns are alternative solutions of bottleneck metabolites.

Bottleneck metabolites are those whose concentrations ranges should be relaxed (with respect to the experimentally measured concentration ranges) to rescue the essentiality of the gene (column 1). Hence, any of those alternative sets of metabolite concentrations is responsible for the essentiality of the gene.

Note: an example of PhenoMapping metabolomics analysis is provided in the script *test_core_metabolomics*. This script includes functions for integration of metabolomics data (function: *prepMetabCons*), identification of essential genes when metabolomics data are integrated (function: *thermoSingleGeneDeletion*), and identification of bottleneck metabolites (function: *linkEssGeneMetab2Mets*). All substeps defined in the section Metabolic model contextualization are inputs to these functions. These inputs are defined in the *test_core_metabolomics* script to facilitate their identification.

Note: all results are available in Table S3.7 from (Stanway et al., 2019).

Results of a PhenoMapping analysis of bottleneck reactions

The PhenoMapping results for a bottleneck reaction analysis look as presented in Table 3. In the first column, there is a list of genes identified as essential when transcriptomics data are integrated. The

Table 2. Subset of PhenoMapping result of bottleneck metabolites

Essential gene [met][a]	Blood phenotype[b]	Liver phenotype[c]	Bottleneck metabolites, Alt 1	Bottleneck metabolites, Alt 2	Bottleneck metabolites, Alt 3
PBANKA_0507400	slow	dispensable	UDP-N-acetyl-D-glucosamine[c]|UDP-N-acetyl-D-glucosamine[r]|UMP[c]	UDP-N-acetyl-D-glucosamine[c]|UMP[c]|UMP[r]	UDP-N-acetyl-D-glucosamine[c]|UDP-N-acetyl-D-glucosamine[r]|UMP[r]
PBANKA_0918200	slow	essential	UDP-N-acetyl-D-glucosamine[c]|UDP-N-acetyl-D-glucosamine[r]|UMP[r]	UDP-N-acetyl-D-glucosamine[c]|UMP[c]|UMP[r]	UDP-N-acetyl-D-glucosamine[c]|UDP-N-acetyl-D-glucosamine[r]|UMP[c]
PBANKA_1112400	no info	no info	UDP-N-acetyl-D-glucosamine[c]|UDP-N-acetyl-D-glucosamine[r]|UMP[r]	UDP-N-acetyl-D-glucosamine[c]|UDP-N-acetyl-D-glucosamine[r]|UMP[c]	UDP-N-acetyl-D-glucosamine[c]|UMP[c]|UMP[r]
PBANKA_1232300	dispensable	dispensable	UDP-glucose[c]|UMP[r]	UDP-glucose[c]|UDP-glucose[r]|UMP[c]	

[a] Essential gene when metabolomics data are integrated.
[b] PlasmoGEM phenotypes for blood-stage development of *P. berghei*.
[c] PlasmoGEM phenotypes for liver-stage development of *P. berghei*.

Table 3. Subset of PhenoMapping result of bottleneck reaction levels

Essential gene RNA-seq[a]	Blood phenotype[b]	Liver phenotype[c]	Bottleneck reactions, Alt 1
PBANKA_0107400	Dispensable	dispensable	UP_R00667_c
PBANKA_0516900	Essential	no info	UP_R00670_c|UP_R00178_c
PBANKA_1346500	Dispensable	essential	UP_R07761_r

[a] Essential gene when transcriptomics data are integrated.
[b] PlasmoGEM phenotypes for blood-stage development of *P. berghei*.
[c] PlasmoGEM phenotypes for liver-stage development of *P. berghei*.

second and third column map the experimentally observed phenotype to the essential gene. The subsequent columns are alternative solutions of bottleneck reaction.

Bottleneck reactions are those whose levels (considered to be UP or DOWN within the feasible flux range) should be relaxed to rescue the essentiality of the gene (column 1). Hence, any of those alternative sets of reactions is responsible for the essentiality of the gene.

> ***Note:*** an example of a PhenoMapping transcriptomics and regulation analysis is provided in the scripts *test_core_transcriptomics* and *test_core_transcriptomics_noregulation*, respectively. This script includes functions for integration of transcriptomics data (function: *integrateGeneExp*), identification of essential genes when transcriptomics data are integrated (functions: *getEssGeneExp* and *getEssReg*), and identification of bottleneck reactions due to RNA levels (function: *linkEssGene2Exp*). All substeps defined in the section Metabolic model contextualization are inputs to these functions. These inputs are defined in the *test_core_transcriptomics* script to facilitate their identification.

> ***Note:*** all results are available in Table S3.8 from (Stanway et al., 2019).

QUANTIFICATION AND STATISTICAL ANALYSIS

Here, we discuss a methodology to handle PhenoMapping results (Figure 9, entry point 2), i.e., phenotype-bottleneck mapping. This interpretation can aid in the understanding of context-specific physiology and the generation of a context-specific GEM. Following, we analyze the results presented in the section Expected outcomes (Tables 1, 2, and 3).

Results of a PhenoMapping analysis of bottleneck substrates

Here, we explain how to use the bottleneck substrate information from PhenoMapping and phenotypic data to infer the composition of the media and substrate availability in the context of study.

The bottleneck substrate analysis links genes identified as essential at an *in silico* minimal medium (here referred to as *IMM-genes*) and substrates (Table 1). The absence of all alternative sets of bottleneck substrates from the *in silico* minimal medium is responsible for the essentiality of the IMM-gene. This means, we should avoid the presence of any of the mapped combinations of bottleneck substrates in the *in silico* medium if the IMM-gene is experimentally defined as essential. When an alternative contains more than one bottleneck substrate, it is enough to remove one of the substrates in the set from the medium to avoid the presence of the combined set. Otherwise, if the IMM-gene is experimentally observed to be dispensable, we should add any alternative combination of bottleneck substrates to the *in silico* minimal medium to rescue the IMM-gene essentiality.

1. Define a minimal medium composition in the GEM – to be consistent with the medium defined in the section Metabolic model contextualization. This medium is also the one on which we performed the analysis of bottleneck substrates.
   a. If we did not apply the last optional step in the media analysis, we select and define one alternative *in silico* minimal medium.

   b. If we applied the last optional step in the media analysis (preferred), we define the combined minimal media. As described before, such media comprises all substrates identified across all alternative *in silico* minimal media. We combine substrates of all alternative *in silico* minimal media that contain the same total number of substrates.

Note: what should be the starting *in silico* minimal medium composition given that there are many alternatives?

To avoid the need to select one of the alternative *in silico* minimal medium and be biased toward a selected minimal medium composition, we suggest to use a joint *in silico* minimal medium. This medium comprises all substrates identified in all alternative *in silico* minimal media. One can perform first the bottleneck substrate analysis at each *in silico* minimal medium (*test_core_substrates* script) and later perform a bottleneck substrate analysis at a joint *in silico* minimal medium (*test_core_substrates_joint* script). This is part of the last optional step in the media analysis included in the section Metabolic model contextualization.

2. Identify genes with essential phenotype at *in silico* minimal media or IMM-genes.

Example: in Table 1, an IMM-gene is PBANKA_0516900.

3. Identify each alternative set of bottleneck substrates linked to the IMM-gene.

Example: the IMM-gene PBANKA_0516900 is mapped to three sets of bottleneck substrates (Table 1): 1) spermine, 2) S-Adenosyl-methioninamine and N-Methyl-putrescine, and 3) Aminopropyl-cadaverine and N-Methyl-putrescine.

4. Identify the experimental phenotype of the IMM-gene.

Example: in the blood stages, PBANKA_0516900 is essential. There is no available phenotype for the liver stages (Table 1).

5. Decide whether the alternative sets of substrates should be added or not to the (joint) *in silico* minimal media based on the experimental phenotype information.
   a. If the IMM-gene is essential based on experiments, do not add any of the alternative sets of substrates in its full definition to the (joint) *in silico* minimal medium.

   Note: an alternative set of bottleneck substrates may contain more than one substrate (see Table 1 alternative solutions 2 and 3 for IMM-gene PBANKA_0516900). This result means that the combined and simultaneous presence of all substrates within the set rescues the *in silico* essentiality of the IMM-gene. In other words, one can rescue the *in silico* essentiality of the IMM-gene when one adds simultaneously all substrates within a set to the *in silico* minimal medium. Hence, these substrates should not be present simultaneously in the medium of the final context-specific model to assure the *in silico* conditional essentiality of the gene.
   b. If the IMM-gene is dispensable based on experiments, we should add at least one of the alternative sets of substrates to the *in silico* minimal medium.
   c. If there is no available experimental phenotype for the IMM-gene, decide about the availability of the mapped bottleneck substrates using information from the other IMM-genes.

Note: what happens if a set of bottleneck substrates is linked both to a dispensable and essential experimental phenotype?

Based on our definition of the accuracy metric, we prioritize to avoid a new FN rather than to correct a FP. Hence, we will add bottleneck substrates to the medium when they correct FPs without generating FNs. Bottleneck substrates linked both to FPs and FNs might not be real metabolic bottlenecks underlying the observed phenotypes. However, such scenarios might suggest more careful consideration and further curation of the GEM is necessary.

6. Decide which bottleneck substrates to add to the *in silico* minimal medium.
   a. If the IMM-gene is essential or based on experiments, we can add as many bottleneck substrates as desired as long as they do not combine as identified in any of the alternative solutions of bottleneck substrates.
   b. If the IMM-gene is dispensable or growth reducing based on experiments, we select one or more alternative sets of bottleneck substrates. These sets will be added to the medium.
   c. If the IMM-gene is growth reducing based on experiments, we may limit the uptake rate of such bottleneck substrates.
7. Define the media as selected in the previous step
   a. Select a set of bottleneck substrates linked to dispensable and growth reducing IMM-genes (based on experimental phenotypes). Any alternative set of bottleneck substrates linked to an IMM-gene is equally valid. Hence, when possible, select sets of bottleneck substrates that are not mapped to IMM-genes that are experimentally essential.
   b. Assemble a media composition that includes the joint minimal media and the selected set of bottleneck substrates linked to dispensable and growth reducing genes.
   c. Define that media composition in the GEM.
   d. Verify that the GEM identifies now new context-specific essentiality predictions (decreasing FPs) while not increasing FNs. If this is not the case, one might want to reevaluate the medium definition to include those bottleneck substrates that are responsible for the FNs.

Note: can we gain information about the function of the metabolic pathways from a bottleneck substrates analysis?

Yes, from the bottleneck substrates analysis one can infer flux coupling between genes/reactions. Coupled genes/reactions might be those sharing alternative combinations of bottleneck substrates.

Example: PhenoMapping suggests that the essentiality of PBANKA_0516900 is due to the simultaneous absence in the media of all combinations of bottleneck substrates. Hence, the media for iPbe-blood should not contain any such combinations of bottleneck substrates. We should eliminate spermine from the medium, and we can select any substrate from the alternatives two and three.

There is no available phenotype for the liver stages, so we should decide first what is the media composition based on the remaining genes. If no gene with available phenotype is linked to these bottleneck substrates, we may allow them or not to be taken up in iPbe-liver.

The gene PBANKA_0522400 is mapped to a single bottleneck substrate (Table 1), i.e., (9Z)-Octadecenoic acid. In the blood stages, PBANKA_0522400 is dispensable. Based on PhenoMapping, (9Z)-Octadecenoic acid should be available in the blood stages to avoid a false essentiality prediction. However, in the liver stages, PBANKA_0522400 is essential. Therefore, based on PhenoMapping, (9Z)-Octadecenoic acid should be inaccessible in the liver stages to correctly predict the IMM-gene as essential.

The gene PBANKA_0613600 is dispensable both in the blood and liver stages (Table 1). Hence any set of the bottleneck substrates should be accessible in both stages. We decide to keep both alternative sets in the medium, i.e., N-((R)-Pantothe-noyl)-L-cysteine and Pantetheine.

Results of a PhenoMapping analysis of bottleneck metabolites

Here, we explain how to use the bottleneck metabolites information from PhenoMapping together with metabolomic and phenotypic data to identify sets of metabolites whose concentrations are key determinants of the phenotype in the context of study. We did not integrate this dataset into iPbe to generate iPbe-blood or iPbe-liver. We only used this dataset as a study case to identify putative bottleneck metabolites.

The bottleneck metabolite analysis links genes identified as essential when metabolomics data are integrated (here referred to as *Met-gene*) and intracellular metabolite concentrations (Table 2). The presence of any alternative set of bottleneck metabolite concentrations is responsible for the essentiality of the Met-gene. This means, we should keep any of the mapped combinations of bottleneck metabolite concentrations if the Met-gene is experimentally defined as essential. Otherwise, if the Met-gene is experimentally observed to be dispensable, we should prohibit the simultaneous definition of bottleneck metabolite concentrations identified in every alternative. When an alternative contains more than one bottleneck metabolite, it is enough to remove one metabolite in each alternative set from the whole metabolomics dataset to avoid the Met-gene essentiality prediction.

8. Identify essential genes when metabolomics data are integrated into the GEM or Met-gene.

Example: in Table 2, an example of Met-gene is PBANKA_1232300.

9. Identify each alternative set of bottleneck metabolites linked to the Met-gene.

Example: the bottleneck metabolites mapped to the Met-gene PBANKA_1232300 are: 1) a set of two metabolites, i.e., UDP-glucose in the cytosol and UMP in the endoplasmic reticulum, and 2) a set of three metabolites, i.e., UDP-glucose in the cytosol and in the endoplasmic reticulum, together with UMP in the cytosol (Table 2).

10. Identify the experimental phenotype of the Met-gene.

Example: the gene PBANKA_1232300 is dispensable both in the liver and blood stages of the malaria infection based on the experimental PlasmoGEM screen data (Table 2).

11. Decide whether the alternative sets of metabolites should be kept or not in the metabolomics dataset based on the experimental phenotype information.
    a. If the Met-gene is essential based on experiments, integrate at least one of the alternative sets of metabolite concentrations into the GEM.
    b. If the Met-gene is dispensable based on experiments, do not add any of the alternative sets of metabolites in its full definition to the metabolomics dataset.
    c. If there is no available experimental phenotype for the Met-gene, preferably keep the bottleneck metabolite concentration data in the metabolomics dataset unless it creates conflict with other Met-genes.

Example: only the simultaneous presence of the concentrations of metabolites within a set renders the Met-gene essential. Hence, cytosolic UDP-glucose is the metabolite that should be removed from the whole metabolomics dataset to render PBANKA_1232300's knockout dispensable *in silico*. This is because, UDP-glucose in the cytosol is the only metabolite that participates in all alternatives (Table 2).

PBANKA_0918200 is dispensable in the blood and essential in the liver stages (Table 2). There are three sets of bottleneck metabolites mapped to it. *A priori*, these concentration ranges should be absent in the blood-stage GEM and present in the liver-stage GEM. However, these bottleneck metabolite concentrations are also mapped to dispensable genes in the liver stages, e.g., PBANKA_0507400.

Since we want to prevent the increase in false rate prediction, we would not integrate these concentrations into the liver-stage GEM. This is an example of a bottleneck that individually but not holistically can explain a phenotype. Hence, it is probable this is not the correct biological marker of the gene essentiality.

Results of a PhenoMapping analysis of bottleneck reactions

Here, we explain how to use the bottleneck reactions information from PhenoMapping together with transcriptomic and phenotypic data to identify sets of reaction levels whose concentrations are key determinants of the phenotype in the context of study.

The bottleneck reaction analysis links genes identified as essential when transcriptomics data are integrated (here referred to as *RNA-gene*) and reaction flux levels (Table 3). The presence of any alternative set of bottleneck reactions is responsible for the essentiality of the RNA-gene. This means, we should keep any of the mapped combinations of bottleneck reaction levels if the RNA-gene is experimentally defined as essential. Otherwise, if the RNA-gene is experimentally observed to be dispensable, we should prohibit the simultaneous definition of bottleneck reaction levels identified in every alternative. When an alternative contains more than one bottleneck reaction, it is enough to remove one reaction in each alternative set from the whole transcriptomics dataset to avoid the RNA-gene essentiality prediction.

12. Identify essential genes in the GEM when transcriptomics data are integrated or RNA-gene.

Example: in Table 3, an example of RNA-gene is PBANKA_0107400.

13. Identify each alternative set of bottleneck reaction levels linked to the RNA-gene.

Example: PhenoMapping maps the essentiality of the gene to a high flux through cytosolic reaction R00667 (Table 3).

14. Identify the experimental phenotype of the RNA-gene.

Example: the RNA-gene PBANKA_0107400 is dispensable both in the blood and liver stages (Table 3).

15. Decide whether the alternative sets of reaction levels should be kept or not in the transcriptomics dataset based on the experimental phenotype information.
    a. If the RNA-gene is essential based on experiments, integrate at least one of the alternative sets of reaction levels to the GEM.
    b. If the RNA-gene is dispensable based on experiments, do not add any of the alternative sets of reaction levels in its full definition to the transcriptomics dataset.
    c. If there is no available experimental phenotype for the RNA-gene, preferably keep the bottleneck reactions data in the transcriptomics dataset unless it creates conflict with other RNA-genes.
16. Integrate the updated RNA-seq data into the GEM.
    a. List the set of bottleneck reaction levels for which we will not consider a direct correlation of gene level to reaction level.
    b. Define this set of reactions as input for the integration of RNA-seq data with TEX-FBA. This is the input called *rxnLevelOUT* of the TEX-FBA function *integrateGeneExp*. This input is a cell of reaction identifiers.

Example: we should not keep the high gene expression level to high flux correlation assumption through the bottleneck reaction R00667 in the GEM (neither iPbe-blood or iPbe-liver). We can infer the bottleneck reaction does not necessarily carry a high flux in the blood and liver stages in *Plasmodium*.

PBANKA_0516900 is essential in the blood stages (Table 3). PhenoMapping maps the essentiality to a combined pair of cytosolic reactions, i.e., R00670 and R00178, that carry high flux. *A priori*, we can keep these two high reaction fluxes in iPbe-blood to predict PBANKA_0516900 as essential. However, the presence of R00670 will increase the incorrect prediction of essentiality for PBANKA_0107400. We hence decide not to allow these constraints in iPbe-blood.

PBANKA_1346500 is dispensable in the blood and essential in the liver stages (Table 3). The responsible reaction is R07761 in the endoplasmic reticulum. R07761's high flux should not be imposed in the blood stages and should be present in the liver stages.

LIMITATIONS

We enumerate below some limitations of PhenoMapping. We explain the source of these limitations, their putative effect on the predictions, and the possibility to overcome these limitations with future work.

Limited information about regulation

PhenoMapping only considers the possibility of lack of regulation of gene expression between isoenzymes. Other regulatory processes that control gene expression and, with it, the cell functions available for the phenotype of study are not part of GEMs yet. This lack of information about regulatory processes, determines that PhenoMapping misses conditional essentiality (conditional TNs). When our knowledge about the regulation of gene expression increases, we will be able to include these interactions in mathematical models of the cell function. Next, PhenoMapping can be expanded to map phenotypes to those regulatory processes.

Limited biochemical information

GEMs are databases of the metabolic function. However, we still lack knowledge about gene functions. The limited biochemical information available determines incorrect essentiality predictions (both FNs and FPs). Computational predictions of novel biochemistry (as available in the ATLAS of Biochemistry (Hadadi et al., 2016; Hafner et al., 2020)) and experimental characterization of protein functions are invaluable to expand our knowledge of biochemistry and better define GEMs. PhenoMapping can be combined with gap-fillers and databases of novel biochemistry to reduce the number of FNs and FPs (in preparation). The databases and GEMs should be updated as we increase our biochemical knowledge.

Limited information about metabolite transportability

The information about metabolite transportability is limited, mostly in less characterized organisms. Metabolite transportability is relevant in the prediction of essential functions in GEMs. Missing transporters might lead to incorrect essentiality predictions (both FNs and FPs). To handle this situation, in PhenoMapping we suggest allowing the simple diffusion of metabolites without phosphate, CoA, and acyl-carrier protein (ACP) moieties, unless experimentally invalidated. PhenoMapping can later map conditionally essential gene functions to the (non-)function of a subset of transporters. When our knowledge of metabolite transportability increases, we can integrate this information into PhenoMapping to validate or invalidate the predictions.

Limited information about metabolite concentrations

When metabolomics data are not available for a thermodynamically consistent analysis, it is assumed that intracellular metabolite concentrations can vary between 1 μM and 50 mM. This range covers the intracellular concentrations of a wide set of metabolites in various cells and conditions (Bennett et al., 2009; Ishii et al., 2007; Teng et al., 2009, 2014; Vo Duy et al., 2012). Such broad concentration ranges might allow some reactions to work bidirectionally, while in reality the reactions might be unidirectional. In this regard, PhenoMapping might be missing the prediction of essential genes (TNs) and the mapping to the responsible sets of metabolite concentrations. A sensitivity analysis of metabolite concentrations (Kiparissides and Hatzimanikatis, 2017) can help identify sets of metabolites relevant for a phenotype. Ultimately, a broader coverage in metabolomics studies and the integration of these datasets into PhenoMapping will allow better predictions of essentiality and mapping of essentiality to metabolite concentrations.

Assuming correlation between gene expression (mRNA) and reaction flux levels

In the TEX-FBA framework and PhenoMapping, we assume *a priori* that higher and lower mRNA levels correlate with higher and lower reaction fluxes, respectively. TEX-FBA and PhenoMapping do not impose such correlation as other methods do, but rather try to maximize the number of reactions for which this assumption is possible (quantified in the form of a maximum consistency score). It is known that mRNA levels and flux levels do not fully correlate, and hence we might encounter wrong essentiality predictions when transcriptomics data are integrated. PhenoMapping allows to identify those FNs or reactions for which the mRNA-flux correlation is not valid. These reactions can be ignored in a subsequent integration of transcriptomics data within PhenoMapping. A preferred option is to integrate proteomics data into PhenoMapping, since the correlation of protein and flux levels is higher.

Assuming essentiality with omics data integrated represents essentiality for growth

PhenoMapping contextualizes GEMs by integrating omics data and later identifies new conditionally essential genes. We say these genes are conditionally essential for the objective function defined (normally growth). However, they are rather essential to maintain the cellular state defined by the omics data and methodology used. It might happen that upon knockout of such a conditionally essential gene, a cell can rapidly achieve another cellular state that also sustains growth and hence that gene is not truly essential for survival. Non-linear dynamic models will be necessary to predict transitions between metabolic states and changes in metabolite concentrations and reaction fluxes upon perturbations. These dynamic models as well as metabolic control analysis help to rank essential genes based on how sensitive growth is to a perturbation of the enzyme function.

> *Note:* some of these assumptions are inherited from the methodologies used to integrate data into the GEMs, i.e., matTFA (Salvy et al., 2018) and TEX-FBA (Pandey et al., 2019).

TROUBLESHOOTING

Problem 1: There is no GEM available

You do not have a GEM for the organism and strain of interest, but you do have a draft GEM. We define a draft GEM as a metabolic model that has not been extensively curated. Draft GEMs normally show varying degrees of genome annotation, metabolite connectivity, and medium and biomass definition. FBA can be performed to a certain extent in such GEMs. Draft GEMs could be generated automatically with available systematic construction pipelines, as recently shown with paraDIGM (Carey et al., 2019).

Potential solution 1

A detailed explanation on the reconstruction steps of a GEM is out of the scope of this protocol, but we provide here some references that might be of guidance. A protocol for guidance on high-quality reconstruction of GEMs following a bottom-up approach is available (Thiele and Palsson, 2010). Such bottom-up reconstructions could be performed with available toolboxes like COBRA (Becker et al., 2007; Ebrahim et al., 2013; Heirendt et al., 2019; Schellenberger et al., 2011), RAVEN (Agren et al., 2013; Wang et al., 2018), modelSEED (Devoid et al., 2013), KBase (US Department of Energy Systems Biology Knowledgebase, http://kbase.us), etc. GEMs can also be constructed following a top-down approach, as suggested with CarveMe (Machado et al., 2018). GEMs can also be adaptded from other strains, as recently suggested (Norsigian et al., 2020).

Once a draft GEM is available, one should improve its performance iteratively by using available frameworks. Among those we highlight:

1. RAVEN: one can perform analysis of metabolic tasks in the draft GEM. Such analysis identifies gaps in metabolism and guides the definition of reactions that render functional metabolic pathways (Agren et al., 2013; Wang et al., 2018).
2. MEMOTE: one can systematically identify missing standard information and major flaws in the GEM like elemental imbalance or a wrongly defined biomass reaction (Lieven et al., 2020).
3. AMMADEUS: one can use an ensemble of draft GEMs and unsupervised learning to generate models that are consistent with experimental data (Medlock and Papin, 2020).
4. NICegame: one can fill metabolic gaps with metabolic reactions, like hypothetical reactions from the ATLAS of Biochemistry (Hadadi et al., 2016; Hafner et al., 2020), and annotate orphan reactions with BridgIT (Hadadi et al., 2019). The NICegame workflow works with any objective function (in preparation).
5. PhenoMapping: this protocol can also be followed to refine GEMs in a modular fashion.

Problem 2: There are multiple GEMs available for the organism of interest

There are multiple GEMs available for the organism and strain of interest, and you need to choose one.

Potential solution 2

The selection of a GEM is critical for the subsequent analysis.

1. Use available tools like MEMOTE (Lieven et al., 2020) to systematically evaluate a series of benchmarks that assure the GEM is properly elementally balanced and defined following community standards. MEMOTE will generate a report with a score that serves as a quantitative basis for comparison.
2. Compare the GEMs based on the following criteria. These are elements that an ideal GEM may include:
   a. An available history of its development, e.g., a GitHub history as available for Yeast 8 (Lu et al., 2019).
   b. Elementally balanced reactions.
   c. The metabolic pathways of interest.
   d. Compatibility of metabolite identifiers with a desired reference database.
   e. Fully defined metabolites vs generic metabolites with R groups.
   f. The highest genome annotation coverage.
   g. A curated electron transport chain pathway with proper ratios between protons (H^+) pumped, reacting cofactors, and ATP formed.
   h. A curated gene-protein-reaction definition. Especial attention should be given to reactions catalyzed by a protein complex, since these genes should be linked with an "AND" rule.
   i. A broad and non-context-specific set of biomass building blocks with stoichiometric coefficients properly defined to produce 1 g of biomass (Chan et al., 2017).
   j. The broadest set of intracellular compartments.
   k. A curated and unbiased localization of reactions. If there is uncertainty in the localization of an enzyme, a multi-localization might be preferred to avoid biased metabolic flux distributions and false essentiality predictions (FNs).
   l. A curation of reactions occurring at the membrane interphase.
   m. A list of metabolic tasks used to evaluate the performance and coverage of the model.
   n. A minimum set of *ad hoc* and pre-assigned reaction directionalities. Pre-assigned directionalities should aim to define the enzymatic irreversibility of reactions and not a pre-assumed thermodynamically feasible directionality (Ataman and Hatzimanikatis, 2015).
   o. Thermodynamic data for the extracellular environment and intracellular compartments, i.e., pH, as well as membrane potential and ionic strength.
   p. Thermodynamic properties of compounds, i.e., ranges of Gibbs free energy of formation ($\Delta_f G'$) at the corresponding extracellular of intracellular conditions (Jankowski et al., 2008).

q. Thermodynamic properties and curation of reactions, i.e., balancing with protons (H^+) and computation of ranges of Gibbs free energy of reaction (Δ_rG') at the defined conditions (Henry et al., 2006, 2007; Jankowski et al., 2008; Salvy et al., 2018).
r. Definition of metabolic channeling when applicable and verified that separately the individual reactions are not thermodynamically feasible (Chiappino-Pepe et al., 2017).
s. Multiple identifiers and 2D structure information linked to a metabolite to facilitate comparison with other models and databases.
t. Properly mapped metabolic subsystems to reactions.
u. E.C. numbers linked to reactions.
v. Information about the sources and methodology used for the gene-function annotation, including reference genome or protein sequences, software, annotation parameters, and reaction databases.
w. A summary of the orphan reactions (reactions that are not linked to any gene), their source database, and the metabolic tasks that motivated their inclusion in the GEM.
x. The criteria followed to define transports and transport mechanisms, since these constitute the set of reactions with the highest uncertainty in the annotation.
y. Information about the sources and methodology used for the localization: reference protein sequences, software, parameters.

3. Analyze reaction fluxes (when the conditions of interest are defined in the GEMs) and compare the GEMs in terms of:
   a. Blocked reactions.
   b. Disconnected or dead-end metabolites.
   c. Gene and reaction essentiality and comparison with available phenotypes for a pre-assessment of the GEM accuracy. At this point it is recommended that the number of FNs is the lowest possible.
   d. Metabolic tasks fulfilled.
   e. Other analyses that you may be interested in should be included in this list.

Problem 3: The GEM is not thermodynamically curated

You would like to perform a thermodynamically consistent analysis within PhenoMapping, which uses matTFA (Salvy et al., 2018), but the GEM is not thermodynamically curated.

Potential solution 3

1. If you want to perform a PhenoMapping analysis accounting for thermodynamic constraints, you additionally need to assemble and include the following information and fields in the GEM, i.e., the COBRA model structure:
   a. *metSEEDID*: list of SEED IDs mapped to the metabolites. It shares length with metabolites. This field does not include compartment information and hence it might contain duplicated seed identifiers.
   b. *CompartmentData*: includes thermodynamic data related to compartments.
   c. *CompartmentData.compSymbolList*: row cell with one letter as defined for each compartment.
   d. *CompartmentData.compNameList*: row cell with names of compartments.
   e. *CompartmentData.membranePot*: square matrix containing membrane potential information between compartments. It shares length with compartment.
   f. *CompartmentData.pH*: row vector with pH values for each compartment. It shares length with compartment.
   g. *CompartmentData.ionicStr*: row vector with ionic strength values in each compartment. It shares length with compartment.
   h. *CompartmentData.compMaxConc*: row vector with maximum default molar concentration values (M or mol/L_{cell}) allowed for metabolites in each compartment. It shares length with compartment.

i. *CompartmentData.compMinConc*: row vector with minimum default molar concentration values (M or mol/L_{cell}) allowed for metabolites in each compartment. It shares length with compartment.
j. *metCompSymbol*: list of one letter compartment symbols defining localization of each metabolite. It should be of the same length and order as the "*mets*" and corresponding fields. It is not strictly necessary to include this field, since PhenoMapping will attempt to generate it based on the compartmentalization information provided in the metabolite identifiers. PhenoMapping will extract the compartment letter provided after any of the following symbols "_," " [", or "(" appearing among the last three characters of the metabolite identifier in "*mets*." Metabolites without such tags will be automatically assigned to the cytosol. A manual check is recommended to assure such assumption is correct or to define a tag to the metabolites that miss it.

Note: one can automatically generate the metCompSymbol structure within PhenoMapping. The metabolite SEED identifiers and thermodynamic information included in the remaining fields is GEM- and context-specific and should be gathered and mapped by the user.

2. Perform a systematic thermodynamic curation within PhenoMapping.

Alternatives: if you want to provide your own thermodynamically curated GEM (skipping the systematic thermodynamic curation within PhenoMapping), verify it contains the following TFA-specific fields (Salvy et al., 2018):

a. *constraintNames*: list of constraints included in the model.
b. *varNames*: list of variables included in the model.
c. *var_lab*: lower bound of variables. It shares length with variables.
d. *var_ub*: upper bound of variables. It shares length with variables.
e. *A*: matrix of constraints (rows) and variables (columns).
f. *f*: vector defining objective function. It shares length with variables.
g. *rhs*: vector containing values of the right hand side of the linear problem. It shares length with constraints.
h. *constraintType*: type of constraint. It shares length with constraints.
i. *vartypes*: type of variable. It shares length with variables.
j. *objtype*: −1 if it is a maximization problem, and +1 if it is a minimization problem.

Note: for further reference, please check the structure of these fields in the iPbe model available in the PhenoMapping repository.

Problem 4: Unfamiliar with the process to define a medium composition within PhenoMapping

How to define a media composition in a GEM within PhenoMapping.

Potential solution 4

1. Familiarize with the definition of uptakes in a GEM for a PhenoMapping analysis:

Note: for every optimization problem, PhenoMapping (as matTFA) will take into account the lower reaction bounds (*var_lb*) and upper reaction bounds (*var_ub*) defined for each reaction variable. We recommend working with net fluxes of reactions (tagged with NF_ and followed by the reaction name) included in the list of variables of the GEM (*varNames*).

The definition of bounds of net reaction fluxes follows standards defined in the COBRA Toolbox (Heirendt et al., 2019):

For a general reaction defined as A ⇔ B, a negative lower bound of the net reaction flux implies that the enzyme can produce metabolite A by catalyzing the reaction in the backward direction. Analogously, a positive upper bound of the net reaction flux implies that the enzyme can produce metabolite B by catalyzing the reaction in the forward direction. Bounds defined as zero block a specific reaction directionality. These can be known enzymatic irreversibilities and as *ad hoc* / pre-assumed reaction directionalities, which are not desired (Ataman and Hatzimanikatis, 2015).

We suggest defining a medium composition through the definition of bounds of exchange or boundary reactions:

Exchange or boundary reactions should be defined as A ⇔ ∅. For this definition, a negative lower bound for the net flux of the exchange reaction implies that the uptake of the metabolite A is allowed. Uptakes are blocked when the lower bound is zero or positive. A positive upper bound for the net flux of the reaction allows secretion of the metabolite A.

⚠ CRITICAL: PhenoMapping will automatically define as A ⇔ ∅ any exchange or boundary reaction that is inversely defined as ∅ ⇔ A. This is done within the functions *initTestPhenoMappingModel* and *analysisIMM.*

2. Allow uptake of all metabolites in the medium by defining an arbitrary negative lower bound (e.g., −50 mmol/g-DW/h) for the corresponding exchange reactions and transporters in the *lb* field of the GEM.

   Note: beware some GEMs block transports of molecules (from the extracellular space to the cytosol) and not the exchange or boundary reactions of those molecules. Both the transport and exchange of a molecule should be open for molecules to be allowed to be consumed.

   ⚠ CRITICAL: transport of metabolites that follow a specific mechanism different from simple diffusion should be treated carefully. If no thermodynamic constraints are taken into account, it might be better to keep pre-assigned directionalities for such transporters to avoid major flaws. For example, there might exist imports of molecules that consume ATP. Such imports might end up generating ATP through secretion of molecules if their pre-assigned directionalities are relaxed in a pure FBA.

3. Convert the model to a PhenoMapping friendly format using *initTestPhenoMappingModel.*
4. Select the metabolites whose uptake will be allowed.
5. Select the uptake rate (in mmol/g-DW/h) for those metabolites.

   Note: the uptake rate could be obtained from experimental data. This value is normally constrained for the carbon sources defined in the medium. For example, the uptake rate of glucose in the GEM of *Escherichia coli* is normally 10 mmol/g-DW/h. Some molecules like oxygen can easily diffuse through cellular membranes. Others have not been measured. In such cases, one could define a default large value like 50 mmol/g-DW/h.

6. Identify all exchange reactions. This can be done with standard COBRA functions included in matTFA like *findExcRxns.*
7. Block the uptake of all metabolites by defining a lower bound (*var_lb*) of value zero for the net flux of the exchange reactions (in *varNames*).
8. Among all exchange reactions, identify those corresponding to the metabolites for uptake.
9. Allow the uptake of this set of metabolites by defining a negative lower bound (*var_lb*) for the net flux of their exchange reactions (in *varNames*).
10. Refine the lower bounds with the (negative) values of uptake rate selected for the set of metabolites.

11. Optimize and check if the GEM is feasible in the defined medium composition.
12. If the GEM remains infeasible perform the analysis of missing substrates in the medium as described in Troubleshooting 6.

Problem 5: There is no objective function defined in the GEM

There is no objective function defined in the GEM, and you would like to define one.

Potential solution 5

For every optimization problem, PhenoMapping (as matTFA) will optimize the value of the variables (*varNames*) defined as objective in the vector *f*. The problem will be a maximization if the *objtype* is −1 and a minimization if the *objtype* is +1. PhenoMapping will define both maximization (see *findDPMax* function) and minimization (see *findDPMin* function) problems.

1. Verify that the variable(s) (*varNames*) to optimize are the only ones marked with a 1 in the corresponding row of the vector *f*.
2. Definition of maximization of growth as the objective function implies having a 1 in the row of the forward biomass reaction. The forward biomass reaction is defined with the tag *F_* and followed by the biomass reaction name in *varNames*.
3. Verify that the GEM is feasible, else proceed to Troubleshooting 6.

Problem 6: The GEM is not feasible

The GEM is not feasible for the objective function selected.

Potential solution 6

1. List the potential reasons that make the GEM infeasible. We define here two general sources of infeasibility:
   a. The GEM includes context-specific information (constraints) under which the selected objective function is not feasible.
   b. The GEM lacks the biochemistry or metabolite transportability capabilities to perform the selected objective function.
2. Evaluate whether the GEM contains the biochemistry and transport capabilities to render the objective function feasible.
   a. Define the biochemistry layer as defined in the PhenoMapping workflow. This can also be done by removing all *ad hoc* constraints from the GEM and defining a rich medium.
   b. Evaluate the feasibility of the GEM. If the GEM is feasible, proceed to step 3. Otherwise, jump to step 5.
3. If the GEM became feasible in the biochemistry layer, identify the minimum number of initial preassigned directionalities that should be relaxed to render a feasible GEM.
   a. Use as input the GEM with the biochemistry layer.
   b. Define MILP constraints linked to each pre-assigned directionality.
   c. Define a requirement for the objective function.
   d. Identify alternative sets of minimum number of reactions whose pre-assigned directionalities should be relaxed.

Note: the rationality defined in step 3 is followed in the analysis of *in silico* minimal media or minimal secretion. An analysis of *in silico* minimal media identifies the minimum number of substrates that we should add to the medium to allow growth. Similarly, an analysis of *in silico* minimal secretion identifies the minimum number of growth by-products that we should allow to be secreted in the GEM to satisfy mass balances and allow growth. Point 4 defines step-by-step how to perform the *in silico* minimal media analysis. An analogous procedure applies to the *in silico* minimal secretion analysis and all other analyses that follow the rationality defined in step 3.

4. Application of step 3 for the analysis of missing substrates in the medium.
   a. List the set of uptakes that are blocked in the infeasible GEM.
   b. Define an *in silico* rich medium in the infeasible GEM, in which all uptakes are open.
   c. Set a requirement for the objective function (non-zero lower bound).
   d. Define only the set of blocked uptakes as input for the analysis of *in silico* minimal media.
   e. Identify the minimum number of substrates required to render the GEM feasible.
   f. Include the additional set of substrates in the medium of the previously infeasible GEM.
   g. Verify that the GEM is now feasible.
5. If the objective function in the infeasible GEM is growth or growth is set as a requirement (with a fixed non-zero lower bound), evaluate the individual production of biomass building blocks (within the analysis of metabolic tasks).
   a. Using the infeasible GEM, identify the biomass building blocks that cannot be produced individually.
   b. Verify that removing those from the biomass reaction turns the GEM feasible.

   Note: the result of the metabolic task analysis will also allow to connect a non-produced biomass building block with a set of substrates or pre-assigned directionalities. One can verify that the substrates and metabolic pathways identified are meaningful for the production of the biomass building block.

6. We follow this step if the GEM did not become feasible in the biochemistry layer or the set of pre-assigned directionalities identified should not be violated. Such scenario requires a curation of the GEM by adding new biochemistry or transport capabilities.
   a. Use as input the infeasible GEM.
   b. Set a requirement for the objective function (non-zero lower bound).
   c. Identify a database of biochemical reactions.
   d. Merge the infeasible GEM with the database of biochemical reactions.
   e. Identify alternative biochemistry to render the GEM feasible.
   f. Define the hypothetical biochemistry in the GEM.

   Note: NICegame (in preparation) is an example of an approach to follow in this scenario since it fills metabolic gaps with hypothetical reactions and annotates genes to orphan reactions.

7. If no biochemistry was found to render the GEM feasible in step 6, one may:
   a. Use as input the GEM with the biochemistry layer and repeat step 6.
   b. Identify the furthest metabolic precursor of the non-produced biomass building block (as identified in step 5) and define a transport and exchange in the GEM to allow its uptake.
   c. Identify a by-product in the metabolic pathway of the non-produced biomass building block (as identified in step 5) whose secretion is not possible. Define a transport and exchange in the GEM to allow its secretion.
   d. Completely eliminate the non-produced biomass building block from the biomass reaction. This last option implies redefining the objective function by normalizing the stoichiometric coefficients (Chan et al., 2017).

Problem 7: The MATLAB paths to initialize the GEM and PhenoMapping are not found

The paths to initialize the GEM and PhenoMapping are not found.

Potential solution 7

There are multiple alternatives to initialize paths.

We recommend setting up the repositories to capitalize on the systematic identification of paths as predefined in PhenoMapping:

1. Clone the matTFA, TEX-FBA, and PhenoMapping repositories in the same path, i.e., sharing the same root or directory.
2. Identify the directory of the CPLEX folder called CPLEX_StudioVersion. For example, *CPLEX_Studio128*.
3. Run the script *settings*.
4. When the following message arises: "Please provide your cplex path and press enter," paste the whole path to the CPLEX folder of point 2. Press enter.

In the script *tutorial_basics* you can find other suggestions on how to set up the repositories matTFA, TEX-FBA, and PhenoMapping. You could also adapt the function *initPhenoMappingPaths*. For example, one may want to include personalized paths or directories to be added automatically.

Problem 8: The default TEX-FBA parameters for a transcriptomics analysis with PhenoMapping result in many FNs

You would like to adjust the TEX-FBA parameters before integrating transcriptomics data.

Potential solution 8

Alternatives: adjust the percentiles of lowly and highly expressed genes.

1. Plot the distribution of gene expressions for all genes in the GEM.
2. Select the percentiles of lowly and highly expressed genes based on the distribution.

⚠ CRITICAL: the selection of percentiles should be adjusted based on the transcriptomics dataset and GEM. The identification of bottleneck reaction levels can guide the definition of a better set of lowly and highly expressed genes. We have seen those percentile values that cut the tails of the distribution at the inflection point lead to low number of bottleneck reaction levels with incorrect *in silico* essentiality.

Alternatives: adjust the flux limits assigned to lowly and highly expressed genes.

3. To relax flux limits, decrease the lower bound required for highly expressed genes (p_h parameter) and increase the upper bound required for lowly expressed genes (p_l parameter).

Note: TEX-FBA performs a flux variability analysis and defines flux limits within the feasible range of fluxes. Many methodologies can be followed to define flux limits. These require slight adjustments of the TEX-FBA formulation.

Problem 9: The GEM is infeasible when metabolomics data are integrated

You integrated metabolomics data into the GEM within the TFA framework or PhenoMapping, and the GEM is infeasible.

Potential solution 9

This happens because there is a set of metabolite concentrations that define a set of infeasible reaction directionalities. These new reaction directionalities block flux through an essential reaction or set of redundant reactions. Hence, one can identify bottleneck metabolites that render the GEM infeasible. This can be done by linking *a priori* dispensable genes with bottleneck metabolites.

1. Use the function *linkEssGeneMetab2Mets* for which an example implementation is defined in the script *test_core_metabolomics.*
2. Identify all genes that are dispensable in the GEM prior to metabolomics data integration. This set of genes will be the input *essTFAmetab* to the function *linkEssGeneMetab2Mets.*

3. The rest of the inputs to *linkEssGeneMetab2Mets* are defined as in the sample script *test_core_metabolomics*.
4. Obtain the solution of bottleneck metabolites linked to all dispensable genes.
5. Identify the minimum number of metabolites that appear in all solutions of bottleneck metabolites.
6. Delete those bottleneck metabolites from the input metabolomics dataset.
7. Integrate the new metabolomics dataset into the GEM.
8. Verify the GEM with a reduced metabolomics dataset is feasible now. This should always be the case.
9. If the GEM is not feasible, you have probably not selected a complete minimum set of bottleneck metabolites to be removed from the input metabolomics dataset. Return to step 5 from this Potential Solution 9.

Problem 10: The PhenoMapping analysis stopped and you had not generated all alternative MILP solutions

Your PhenoMapping analysis involving a mixed integer linear programming (MILP) formulation stopped. It could have stopped for any reason, e.g., the number of alternatives defined was low and you did not identify all alternative solutions of the optimal size, or you stopped the optimization, or the solver crashed, etc. You would like to continue the optimization without regenerating all previous solutions.

Potential solution 10

To avoid regenerating solutions of a MILP problem one should regenerate the integer cut constraints. PhenoMapping integrates integer cuts into the GEM every time a new solution for a MILP is generated (see function *findDPMax* and *findDPMin*). Those intermediate solutions are saved in the PhenoMapping directory (tmpresults subfolder) and can be an input for the steps below to regenerate the cut constraints.

Here, we show how to restore integer cuts for all MILP maximization problems within PhenoMapping, which use the function *findDPMax* to generate alternative solutions. An example of MILP maximization problem is the media analysis. The media analysis tend to involve many solutions (thousands) and hence this troubleshooting pipeline might become handy.

1. Use the function *recoverModel4DPMax.* This function restores integer cut constraints in the GEM. Integer cut constraints avoid the repeated prediction of a solution to a MILP.
2. Identify the model that you used as input for the function *findDPMax* before it stopped. This will be the first input for *recoverModel4DPMax*.

   Note: PhenoMapping saves in the folder *tmpresults* a matrix with intermediate solutions for every problem. The intermediate MILP solutions are tagged with "*_DPs*," as defined within *findDPMax*. The name of the file before the tag "*_DPs*" is an input to the *findDPMax* function.

   ⚠ CRITICAL: intermediate solutions are deleted at the end of every example script, with the function *elimIntermPMFile*. It is wise not to delete intermediate solutions if the number of alternatives is low and you expect to use those intermediate solutions later.

3. Identify the name of the file where the intermediate solutions are saved.
4. Load the matrix saved in the file. This will be the second and last input for the function *recoverModel4DPMax*.
5. Run the function *recoverModel4DPMax* and the output is the new model with the cut constraints regenerated.

Problem 11: The essentiality analysis fails although the GEM is feasible

Your essentiality analysis shows error but your GEM is feasible.

Potential solution 11

This occurs because PhenoMapping uses a thermodynamically consistent implementation of FastSL (Pratapa et al., 2015) to accelerate the identification of essential genes. This formulation identifies the optimal value of the objective function, e.g., growth, and sets a requirement of the objective value. Next, it minimizes flux through all reactions in the GEM. The first solution achieved (a single solution among all possible alternative flux solutions) is used to identify fluxes with zero value, as done in the parsimonious FBA or pFBA (Lewis et al., 2010). Reactions that do not carry flux are by definition not essential and hence do not need to be tested afterward in an essentiality analysis. A strict definition of the required objective value can render the next problem infeasible. Hence, one might need to relax this value.

⚠ CRITICAL: the required objective value should not be lower than (1 - essentiality threshold). The essentiality threshold should have been previously defined for a PhenoMapping analysis.

Note: the thermodynamically consistent implementation of pFBA is included in the function *optimizeThermoModel*.

1. Identify the requirement of the objective value defined within *optimizeThermoModel.*
2. Reduce the requirement of the objective value but do not define a requirement < (1 - essentiality threshold).
3. You may need to round the required objective value to avoid problems with the precision of the solver.

Problem 12: You obtain solver-related infeasible solutions

You obtain solver-related infeasible solutions. These are infeasible solutions that arise due to problems with the solver or problem-definition rather than the GEM per se. For example, the solver may not converge fast enough to a solution and returns a NaN.

Potential solution 12

Infeasible solutions are unfortunately common when working with optimization problems and solvers. There can be multiple reasons for these infeasibilities, and the resolution is very case-specific. Here, we provide a check list with some common troubleshooting approaches we have used to face solver-related infeasibilities.

1. Consider as essential any gene knockout rendering an infeasible solution in the essentiality analysis with TEX-FBA.
2. Verify that the GEM is feasible by evaluating it as defined in Troubleshooting 6.
3. Tighten the reaction flux bounds with the minimum and maximum feasible solutions as identified with a flux variability analysis (FVA) (Mahadevan and Schilling, 2003). This is done by defining the FVA solution as lower and upper bound of the net fluxes. The FVA should account for thermodynamic constraints if the infeasibility arises within TFA. Round the fluxes to have less than five decimals. These rounded values should fall outside the flux range accounting for all decimals.
4. Tighten the bounds of the Gibbs free energy of reaction ($\Delta_r G'$) with the minimum and maximum feasible solutions as identified with a thermodynamically consistent Variability Analysis for these variables. These rounded values should fall outside the $\Delta_r G'$ range accounting for all decimals.
5. If you defined a time limit for the solver, increase it.
6. Verify that the feasibility tolerance of the solver is low enough. By default, in matTFA and PhenoMapping the feasibility tolerance is 10^{-9}. This value of tolerance for CPLEX is defined in the function *changeToCPLEX_WithOptions* within the matTFA package.

7. Evaluate any other parameter specific to the solver like the time limit, integrality tolerance, emphasis on precision, or scaling. An example of how these parameters are defined and the default values within matTFA is provided in the function *changeToCPLEX_WithOptions* within the matTFA package.
8. Round all reaction bounds to have less than five decimals.
9. Verify that the product of any stoichiometric coefficient (e.g., in the biomass reaction) and the corresponding reaction flux (e.g., growth) is not below the tolerance of the solver.
10. If applicable, round the metabolomics data integrated to have less than five decimals.
11. Round any right-hand-side value (*rhs*) to have less than five decimals.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Anush Chiappino-Pepe (anush.chiappinopepe@alumni.epfl.ch).

Materials availability

Outputs generated with this protocol are provided in previous publications (Krishnan et al., 2020; Stanway et al., 2019). These outputs are reproducible as described here.

Data and code availability

Documented implementation of the PhenoMapping workflow in MATLAB is available on www.github.com/EPFL-LCSB/phenomapping.

PhenoMapping requires the matTFA toolbox (Salvy et al., 2018) for TFA and the TEX-FBA toolbox (Pandey et al., 2019) for the integration of gene expression data.

The metabolic models of *P. berghei* (iPbe) and *T. gondii* (iTgo), the *P. berghei* blood-stage relative growth phenotypes, the tachizoyte *T. gondii* genome-wide screen, and all integrated omics data are available in www.github.com/EPFL-LCSB/phenomapping v1.0. These files are enough to enable the user to follow along and repeat all computational steps of the examples presented in this PhenoMapping protocol.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Aarti Krishnan, Dr. Vikash Pandey, Dr. Noushin Hadadi, Evangelia Vayena, Chris Giuliano, and Raina Thomas for critical discussions of the PhenoMapping repository. The authors would like to thank the reviewers, Dr. Noushin Hadadi, Evangelia Vayena, and Dr. Liliana Ángeles-Martínez for critical feedback of the manuscript. All figures (except Figures 3 and 4B) were created with BioRender.com. A.C.-P. and V.H. were funded by grant 2013/155 (MalarX), a SystemsX.ch grant awarded by the SNSF. The authors would like to thank Dr. Dominique Soldati-Favre, Dr. Oliver Billker, and Dr. Volker T Heussler for motivating the development of PhenoMapping.

AUTHOR CONTRIBUTIONS

Conceptualization, A.C.-P. and V.H.; Methodology, A.C.-P.; Software, A.C.-P.; Formal analysis, A.C.-P.; Investigation, A.C.-P.; Writing – Original Draft, A.C.-P.; Writing – Review & Editing, A.C.-P.; Visualization, A.C.-P.; Supervision, A.C.-P. and V.H.; Project Administration, A.C.-P. and V.H.; Funding Acquisition, V.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The RAVEN Toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. PLoS Comput. Biol. *9*, e1002980.

Ataman, M., and Hatzimanikatis, V. (2015). Heading in the right direction: thermodynamics-based network analysis and pathway engineering. Curr. Opin. Biotechnol. *36*, 176–182.

Becker, S.A., Feist, A.M., Mo, M.L., Hannum, G., Palsson, B.Ø., and Herrgard, M.J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat. Protoc. *2*, 727–738.

Bennett, B.D., Kimball, E.H., Gao, M., Osterhout, R., Van Dien, S.J., and Rabinowitz, J.D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. Nat. Chem. Biol. *5*, 593–599.

Bushell, E., Gomes, A.R., Sanderson, T., Anar, B., Girling, G., Herd, C., Metcalf, T., Modrzynska, K., Schwach, F., Martin, R.E., et al. (2017). Functional profiling of a *Plasmodium* genome reveals an abundance of essential genes. Cell *170*, 260–272.e8.

Caldelari, R., Dogga, S., Schmid, M.W., Franke-Fayard, B., Janse, C.J., Soldati-Favre, D., and Heussler, V. (2019). Transcriptome analysis of *Plasmodium berghei* during exo-erythrocytic development. Malar. J. *18*, 330.

Carey, M.A., Papin, J.A., and Guler, J.L. (2017). Novel *Plasmodium falciparum* metabolic network reconstruction identifies shifts associated with clinical antimalarial resistance. BMC Genomics *18*, 543.

Carey, M.A., Medlock, G.L., Stolarczyk, M., Petri, W.A., Guler, J.L., and Papin, J.A. (2019). Comparative analyses of parasites with a comprehensive database of genome-scale metabolic models. bioRxiv, 772467.

Caspi, R., Billington, R., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P.E., Ong, Q., Ong, W.K., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. *46*, D633–D639.

Chan, S.H.J., Cai, J., Wang, L., Simons-Senftle, M.N., and Maranas, C.D. (2017). Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. Bioinformatics *33*, 3603–3609.

Chiappino-Pepe, A., Tymoshenko, S., Ataman, M., Soldati-Favre, D., and Hatzimanikatis, V. (2017). Bioenergetics-based modeling of *Plasmodium falciparum* metabolism reveals its essential genes, nutritional requirements, and thermodynamic bottlenecks. PLoS Comput. Biol. *13*, e1005397.

Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A.A., and Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. Methods Mol. Biol. *985*, 17–45.

Ebrahim, A., Lerman, J.A., Palsson, B.O., and Hyduke, D.R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst. Biol. *7*, 74.

Feist, A.M., and Palsson, B.O. (2010). The biomass objective function. Curr. Opin. Microbiol. *13*, 344–349.

Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A., and Hatzimanikatis, V. (2016). ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. ACS Synth. Biol. *5*, 1155–1166.

Hadadi, N., MohammadiPeyhani, H., Miskovic, L., Seijo, M., and Hatzimanikatis, V. (2019). Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. Proc. Natl. Acad. Sci. U S A *116*, 7298.

Hafner, J., MohammadiPeyhani, H., Sveshnikova, A., Scheidegger, A., and Hatzimanikatis, V. (2020). Updated ATLAS of biochemistry with new metabolites and improved enzyme prediction power. ACS Synth. Biol. *9*, 1479–1482.

Hartleb, D., Jarre, F., and Lercher, M.J. (2016). Improved metabolic models for *E. coli* and *Mycoplasma genitalium* from globalfit, an algorithm that simultaneously matches growth and non-growth data sets. PLoS Comput. Biol. *12*, e1005036.

Hehl, A.B., Basso, W.U., Lippuner, C., Ramakrishnan, C., Okoniewski, M., Walker, R.A., Grigg, M.E., Smith, N.C., and Deplazes, P. (2015). Asexual expansion of Toxoplasma gondii merozoites is distinct from tachyzoites and entails expression of non-overlapping gene families to attach, invade, and replicate within feline enterocytes. BMC Genomics *16*, 66.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat. Protoc. *14*, 639–702.

Henry, C.S., Jankowski, M.D., Broadbelt, L.J., and Hatzimanikatis, V. (2006). Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. Biophys. J. *90*, 1453–1461.

Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. Biophys. J. *92*, 1792–1805.

Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., and Stevens, R.L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat. Biotechnol. *28*, 977–982.

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., Hirasawa, T., Naba, M., Hirai, K., Hoque, A., et al. (2007). Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. Science *316*, 593.

Jankowski, M.D., Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V. (2008). Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys. J. *95*, 1487–1499.

Jeske, L., Placzek, S., Schomburg, I., Chang, A., and Schomburg, D. (2019). BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. *47*, D542–D549.

King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., and Lewis, N.E. (2015). BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. *44*, D515–D522.

Kiparissides, A., and Hatzimanikatis, V. (2017). Thermodynamics-based metabolite sensitivity analysis in metabolic networks. Metab. Eng. *39*, 117–127.

Krishnan, A., Kloehn, J., Lunghi, M., Chiappino-Pepe, A., Waldman, B.S., Nicolas, D., Varesio, E., Hehl, A., Lourido, S., Hatzimanikatis, V., et al. (2020). Functional and computational genomics reveal unprecedented flexibility in stage-specific *Toxoplasma* metabolism. Cell Host Microbe *27*, 290–306.e11.

Kumar, V.S., and Maranas, C.D. (2009). GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS Comput. Biol. *5*, e1000308.

Kyoto University (1995). KEGG (Kyoto Encyclopedia of Genes and Genomes). https://www.genome.jp/kegg/kegg1.html.

LCSB (2020). Laboratory of Computational Systems Biotechnology (LCSB) database. https://lcsb-databases.epfl.ch/.

Lewis, N.E., Hixson, K.K., Conrad, T.M., Lerman, J.A., Charusanti, P., Polpitiya, A.D., Adkins, J.N., Schramm, G., Purvine, S.O., Lopez-Ferrer, D., et al. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. Mol. Syst. Biol. *6*, 390.

Lieven, C., Beber, M.E., Olivier, B.G., Bergmann, F.T., Ataman, M., Babaei, P., Bartell, J.A., Blank, L.M., Chauhan, S., Correia, K., et al. (2020). MEMOTE for standardized genome-scale metabolic model testing. Nat. Biotechnol. *38*, 272–276.

Lu, H., Li, F., Sánchez, B.J., Zhu, Z., Li, G., Domenzain, I., Marcišauskas, S., Anton, P.M., Lappa, D., Lieven, C., et al. (2019). A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat. Commun. *10*, 3586.

Machado, D., Andrejev, S., Tramontano, M., and Patil, K.R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. *46*, 7542–7553.

Mahadevan, R., and Schilling, C.H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab. Eng. *5*, 264–276.

Medlock, G.L., and Papin, J.A. (2020). Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning. Cell Syst. *10*, 109–119.e3.

Norsigian, C.J., Fang, X., Seif, Y., Monk, J.M., and Palsson, B.O. (2020). A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. Nature Protocols *15*, 1–14.

Otto, T.D., Böhme, U., Jackson, A.P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W.A.M., Religa, A.A., Robertson, L., Sanders, M., Ogun, S.A., et al. (2014). A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biol. *12*, 86.

Pan, S., and Reed, J.L. (2018). Advances in gap-filling genome-scale metabolic models and modeldriven experiments lead to novel metabolic discoveries. Curr. Opin. Biotechnol. *51*, 103–108.

Pandey, V., Gardiol, D.H., Chiappino-Pepe, A., and Hatzimanikatis, V. (2019). TEX-FBA: a constraint-based method for integrating gene expression, thermodynamics, and metabolomics data into genome-scale metabolic models. bioRxiv, 536235.

Pratapa, A., Balachandran, S., and Raman, K. (2015). Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks. Bioinformatics *31*, 3299–3305.

Richelle, A., Chiang, A.W.T., Kuo, C.-C., and Lewis, N.E. (2019). Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions. PLoS Comput. Biol. *15*, e1006867.

Salvy, P., Fengos, G., Ataman, M., Pathier, T., Soh, K.C., and Hatzimanikatis, V. (2018). pyTFA and matTFA: a Python package and a Matlab toolbox for thermodynamics-based flux analysis. Bioinformatics *35*, 167–169.

Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat. Protoc. *6*, 1290–1307.

Schroeder, W.L., and Saha, R. (2020). OptFill: a tool for infeasible cycle-free gapfilling of stoichiometric metabolic models. IScience *23*, 100783.

Schuetz, R., Kuepfer, L., and Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. Mol. Syst. Biol. *3*, 119.

Sidik, S.M., Huet, D., Ganesan, S.M., Huynh, M.-H., Wang, T., Nasamu, A.S., Thiru, P., Saeij, J.P.G., Carruthers, V.B., Niles, J.C., et al. (2016). A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes. Cell *166*, 1423–1435.e12.

Sohn, S.B., Kim, T.Y., Lee, J.H., and Lee, S.Y. (2012). Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of in silico/in vivo mutant growth. BMC Syst. Biol. *6*, 49.

Stanway, R.R., Bushell, E., Chiappino-Pepe, A., Roques, M., Sanderson, T., Franke-Fayard, B., Caldelari, R., Golomingi, M., Nyonda, M., Pandey, V., et al. (2019). Genome-scale identification of essential metabolic processes for targeting the *Plasmodium* liver stage. Cell *179*, 1112–1128.e26.

Teng, R., Junankar, P.R., Bubb, W.A., Rae, C., Mercier, P., and Kirk, K. (2009). Metabolite profiling of the intraerythrocytic malaria parasite *Plasmodium falciparum* by ^{1}H NMR spectroscopy. NMR Biomed. *22*, 292–302.

Teng, R., Lehane, A.M., Winterberg, M., Shafik, S.H., Summers, R.L., Martin, R.E., van Schalkwyk, D.A., Junankar, P.R., and Kirk, K. (2014). ^{1}H-NMR metabolite profiles of different strains of *Plasmodium falciparum*. Biosci. Rep. *34*.

Thiele, I., and Palsson, B.Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. *5*, 93–121.

Tymoshenko, S., Oppenheim, R.D., Agren, R., Nielsen, J., Soldati-Favre, D., and Hatzimanikatis, V. (2015). Metabolic needs and capabilities of *Toxoplasma gondii* through combined computational and experimental analysis. PLOS Comput. Biol. *11*, e1004261.

Vo Duy, S., Besteiro, S., Berry, L., Perigaud, C., Bressolle, F., Vial, H.J., and Lefebvre-Tournier, I. (2012). A quantitative liquid chromatography tandem mass spectrometry method for metabolomic analysis of *Plasmodium falciparum* lipid related metabolites. Anal. Chim. Acta *739*, 47–55.

Wang, H., Marcišauskas, S., Sánchez, B.J., Domenzain, I., Hermansson, D., Agren, R., Nielsen, J., and Kerkhoven, E.J. (2018). RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLoS Comput. Biol. *14*, e1006541.

Zimmermann, J., Kaleta, C., and Waschina, S. (2020). gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. bioRxiv, 2020.03.20.000737.