Inference of metabolic fluxes in nutrient-limited continuous cultures: A Maximum Entropy approach with the minimum information

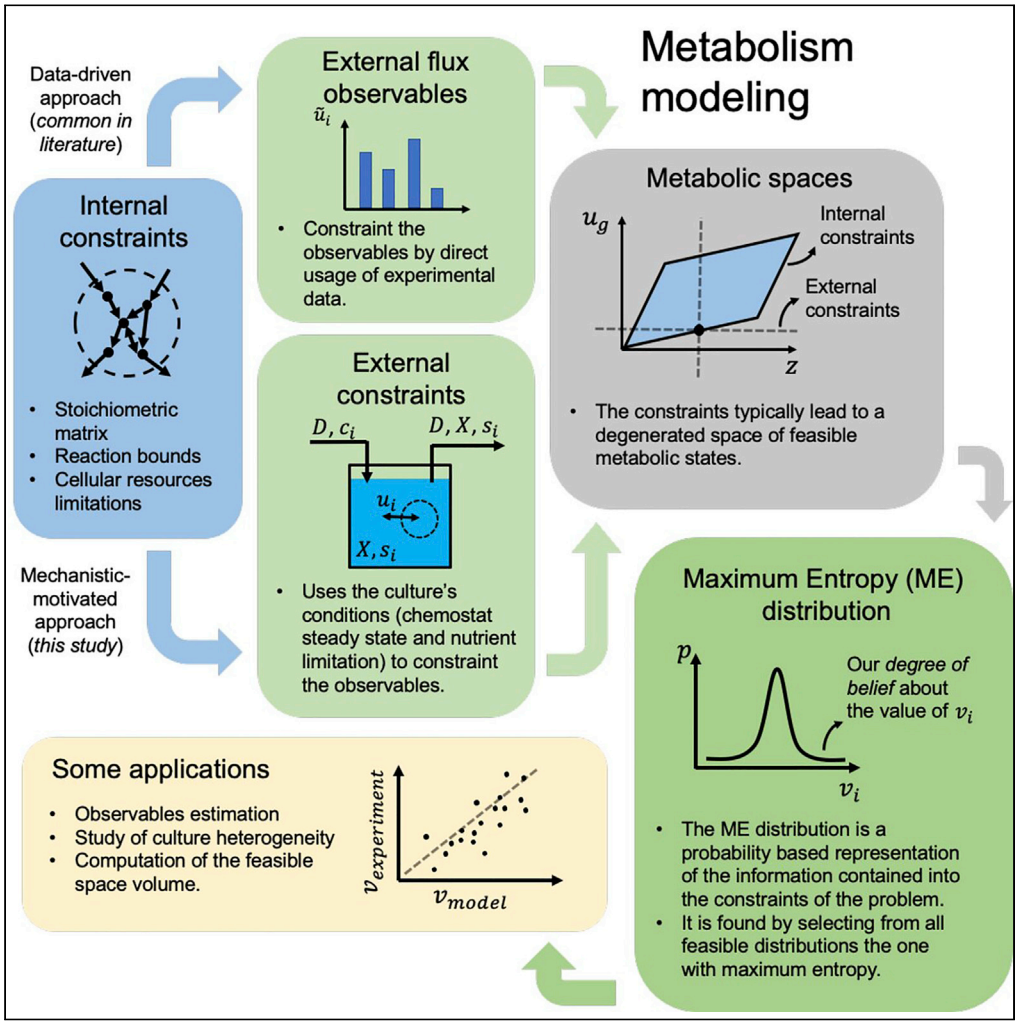
José Antonio Pereiro-Morejon, Jorge Fernandez-De-Cossio-Diaz, Roberto Mulet

iScience

CellPress
OPEN ACCESS

Article

Inference of metabolic fluxes in nutrient-limited continuous cultures: A Maximum Entropy approach with the minimum information



José Antonio Pereiro-Morejón, Jorge Fernandez-de-Cossio-Diaz, Roberto Mulet

mulet@fisica.uh.cu

Highlights

Inference of metabolic fluxes from a minimal set of measurements

Application to *Escherichia coli* experimental data

A dynamical model of the chemostat explains the performance of the method

Pereiro-Morejón et al., iScience *25*, 105450
December 22, 2022
https://doi.org/10.1016/j.isci.2022.105450

iScience

CellPress
OPEN ACCESS

Article

Inference of metabolic fluxes in nutrient-limited continuous cultures: A Maximum Entropy approach with the minimum information

José Antonio Pereiro-Morejón,[1,2] Jorge Fernandez-de-Cossio-Diaz,[3] and Roberto Mulet[1,4,*]

SUMMARY

The study of cellular metabolism is limited by the amount of experimental data available. Formulations able to extract relevant predictions from accessible measurements are needed. Maximum Entropy (ME) inference has been successfully applied to genome-scale models of cellular metabolism, and recent data-driven studies have suggested that in chemostat cultures of *Escherichia coli* (*E. coli*), the growth rate and uptake rates of limiting nutrients are the most informative observables. We propose the thesis that this can be explained by the chemostat dynamics, which typically drives nutrient-limited cultures toward observable metabolic states maximally restricted in the dimensions of those fluxes. A practical consequence is that relevant flux observables can now be replaced by culture parameters usually controlled. We test our model by using simulations, and then we apply it to *E. coli* experimental data where we evaluate the quality of the inference, comparing it to alternative formulations that rest on convex optimization.

INTRODUCTION

The study of cellular metabolism is a research field with a direct impact on the biotechnological industry. Indeed, cell culture-derived products are a major part of a multi-billion market.[1] These products are obtained by exploiting the capabilities of cellular metabolism to produce molecules with a wide range of chemical complexity. Cells are cultured in three common modes: batch, fed-batch, and continuous.[2] In batch, the culture starts with a medium rich in nutrients that are consumed by the cells, often until starvation. Similarly, fed-batch cultures start with a nutrient pool, which is resupplied in discrete time intervals, maintaining the cells alive for longer periods of time. On the other hand, in continuous mode, fresh medium constantly replaces culture fluid at a given rate.[3]

The chemostat is a prototypical continuous cultivation device developed in the 50s.[4,5] Chemostats are often operated at constant volume and in the steady state, which is reached when macroscopic variables of the culture stay constant in time (basically cell and extracellular metabolite concentrations). It is also common to specify which medium component is limiting cell growth. Although the advantages of continuous cell culture have been widely discussed in the literature[6,7,8,9,10] the use of these techniques over batch or fed-batch is hampered by the complexity of continuous systems, *i.e.* culture heterogeneity, hysteresis, multi-stability or sharp transitions between metabolic states[11,12,13,14,15,16,.17] This complexity negatively impacts the yield of bio-processes. In particular, culture heterogeneity is estimated to generate losses of more than 30% in industrial-scale fermentation.[18]

Culture performance is an emergent property derived from the individual metabolic state of each cell,[19] but also the result of interactions between cells. It is fundamental to connect metabolic states at the individual cell level, to macroscopic properties at the culture level. This connection can guide efforts to understand cellular metabolism in a continuous regime and suggest strategies to improve production efficiency.[20]

In this task, the community has been assisted by an increasing number of accurate experimental techniques that generate large amounts of data. In particular, information about cellular metabolism, at the level of individual reactions, has led to the development of genome-scale metabolic networks (GEMs)[21,22,.23] Although at present a full characterization of cellular metabolism is not feasible, Constraint-Based

[1]Group of Complex Systems and Statistical Physics, Physics Faculty, University of Havana, San Lazaro y L, Vedado, La Habana 10400, Cuba

[2]Biology Faculty, University of Havana, San Lazaro y L, Vedado, La Habana 10400, Cuba

[3]Laboratory of Physics of the Ecole Normale Superieure, CNRS UMR 8023, PSL Research, 24 rue Lhomond, 75005 Paris, Ile de France, France

[4]Department of Theoretical Physics, University of Havana, San Lazaro y L, Vedado, La Habana 10400, Cuba

*Correspondence: mulet@fisica.uh.cu

https://doi.org/10.1016/j.isci.2022.105450

Modeling (CBM) approaches help to integrate a variety of data types (*e.g.* stoichiometric, thermodynamic, dynamic, genetic, and so forth) that restrict as much as possible the space of feasible phenotypes that the metabolic network can display.

Constraint-based methods such as Flux Balance Analysis (FBA) have been extensively used to predict a wide range of metabolic observables (*e.g.* culture growth rate, ATP production, and so forth), especially for bacterial batch cultures in the exponential growth phase.[24,25,26,27] FBA can also be exploited in combination with experimental data if the latter provides only partial knowledge about the macroscopic properties of the culture. For example, if the growth rate or metabolic fluxes are known from experimental data, this information can be introduced in the FBA framework to refine predictions about other fluxes in the network.[28] However, as we will discuss in more detail later in discussion, typical FBA formulations can hardly provide any insights about important culture properties such as cellular heterogeneity.

A more general methodology makes use of the Maximum Entropy (ME) Principle.[29] The ME principle has been used, with mixed results, in several fields including Biology[30,31,.32] In the context of metabolic modeling it has been fruitfully combined with constraint-based models.[33,34,35,36] In short, these methods formulate a probabilistic description of the metabolic state of cells by parameterizing an ME distribution that matches a subset of the available culture observables and maximizes the (statistical) entropy.[29] The distribution performance is then tested against the rest of the available experimental data or used to make further analysis. This way it has been shown that ME distributions provide a better fit to measured flux observables than plain FBA models.[33] Also, ME-derived growth rate distributions have been compared to experiments with good results, using single-cell data at different sub-inhibitory antibiotic concentrations.[33] In ref. 37 the authors used ME to exploit all available experimental data and learn the most probable distribution that describes the feasible flux solution space. In turn, the ME distribution can be used to explore the relationship between fitness and heterogeneity of bacteria batch cultures. Moreover, the ME principle is ubiquitous in other fields as well, including physics,[38] ecology[39,40], neuroscience,[41] among others. See[31] for a critical review.

To avoid any confusion with other usages of the term in the literature of metabolic networks, it is important to remark here that we use Entropy purely in the information-theoreetic,[42] statistical, sense, and make no statement about "thermodynamic entropy" or "entropy production" in metabolic reaction networks, which has been studied by other authors.[43] Although deeper connections may be drawn between these concepts[29,38], this is out of scope in the present work.

Most of these methods rely on direct measurements of many external metabolic fluxes in the cell population which is a costly practice. Therefore, an important question that we address in this work, is: What is the minimum number of external observables that we need to know to provide a proper description of the system? In order to tackle this, we begin by carefully re-examining formulations of ME in the context of a chemostat culture. We introduce two metabolic spaces, describing the states of individual cells, and of the population, respectively. Building on previous models coupling cell metabolism with the dynamics of extracellular observables,[17,36] similar also in spirit to the code recently developed in the general context of bacteria communities.[44] We advance the thesis that the chemostat dynamics drives the culture observables toward states maximally constrained in least two dimensions: the observable growth rate and the limiting nutrient uptake. We show that this implies that other uptake observables are essentially redundant and the limiting nutrient uptake can be replaced by information about the cell density and the media composition. By using a minimal set of readily available quantities, our approach greatly extends the possible applications of ME metabolic modeling for continuous cultures. In addition, we show that this mechanism is consistent with the results obtained through data-oriented techniques[45] suggesting that the most informative parameters of an ME distribution are related to the observable cell growth and limiting nutrient uptake rates. Finally, we explore the effects of culture heterogeneity in the flux inference process and highlight some possible sources of bias in common ME formulations.

In short, we support the idea that with the knowledge of only external parameters (*using a minimum set of experimental data*): the chemostat dilution rate, cellular concentration, and the concentration of the limiting metabolite in the feed medium; we can obtain a description of the metabolism of a continuous culture at a steady state equivalent to those which use direct uptake experimental measurements.

The rest of the work is organized as follows. In the next section, we introduce the main concepts of constraint modeling techniques for metabolism and how they can be applied to continuous cultures with limiting nutrients. Then, we introduce Flux Balance Analysis and the Maximum Entropy Principle and explain how they can be used to infer the metabolic fluxes using experimental data from this kind of cultures. Later, we present the results of our work. We first make an analysis of the consequences of imposing different constraints in FBA and ME. Then, we exploit the ME method to infer the metabolic state of a simulated chemostat culture using a simple model of cellular metabolism. We also show the application of our ME formulation on a genome-scale network, inferring a set of literature-available experimental flux observables from glucose-limited *E. coli* chemostat cultures. For completeness, we compare and discuss the results obtained with our methodology (ME) with the solutions obtained through different FBA approximations and previous ME formulations. Finally, we additionally compare our ME formulation with a previous one to evaluate the impact of specific assumptions over the inferred distribution.

CONSTRAINT-BASED METABOLIC MODELS

The formulation of models able to describe, from first principles, the evolution, and properties of biological networks (such as *GEM* s) is in general an open problem. Among the limitations are the complexity of interactions between its many components, the large number of parameters (usually prohibitively large) required to formulate a complete description of the system, and the fact that they are subject to evolution. The latter means that the models need to be continuously updated.[23] Therefore, it is common to study the metabolism considering only the known effect of constraints over the possible physiological states of the system. These constraints can be physicochemical, spatial, topological, environmental, or regulatory in nature. This approach, called Constraint-Based Modeling (*CBM*), leads to the formulation of solution spaces rather than the computation of a single solution.[23]

Metabolic networks and constraint-based modeling

A metabolic network is built by connecting metabolites as described by the stoichiometry of the reactions in the cell. If the network includes a significant portion of the known chemical reactions comprehended in the organism genome, it is called a Genome-Scale Metabolic Network. It constitutes the basis to formulate a constraint-based model of cellular metabolism[46] where the rate of change of the concentration of any metabolite depends on the combined effect of all reactions that involve it.

For a network with N reactions and M metabolites, a balance equation can be written as:

$$\frac{dm_i}{dt} = \sum_j^N S_{ij}\, v_j \qquad \text{(Equation 1)}$$

where $1 \leq i \leq M$, $1 \leq j \leq N$, m_i is the concentration of metabolite i, v_j is the flux value assigned to reaction j, and $S \in \mathbb{R}^{M\times N}$ is the stoichiometric matrix where S_{ij} is the stoichiometric coefficient of metabolite i in reaction j. The common convention is that $S_{ij} = 0$ means that the metabolite does not participate in the reaction, $S_{ij} < 0$ that the metabolite participates as a reactant, and $S_{ij} > 0$ that it participates as a product. The information required to model the time dependency of v_j is not commonly available. Therefore, it is usual to introduce a quasi-steady state assumption for intracellular metabolites,[23] that separates the time scales in the system and allows writing Equation 1 as:

$$0 = \sum_j^N S_{ij} v_j \qquad \text{(Equation 2)}$$

A particular flux configuration is specified by the vector $\boldsymbol{v} \in \mathbb{R}^N$ of all flux values v_j included in the network. In practice, besides the biochemical reactions (*e.g.* catalyzed by enzymes), this vector may contain additional reactions (often artificial) that are included according to a variety of modeling reasons. An example are exchange reactions ($\boldsymbol{u} \in \mathbb{R}^M$), which model the transport of metabolites between the system and its environment. Another important component of $\boldsymbol{v}$ is the biomass reaction ($z \in \mathbb{R}$), which represents the synthesis of new biomass (and secondary products) from a set of precursors. The exchanges and the biomass reaction, represent the boundary of the system. The rest are considered to be internal reactions ($\boldsymbol{r} \in \mathbb{R}^{N-M-1}$), in short $v \equiv (u, r, z)$.

Equation 2 constitutes the first set of constraints that restricts the flux configurations of the network. They form a linear system of equations. Any solution is a vector $\boldsymbol{v}$ that satisfies the balance of mass for each

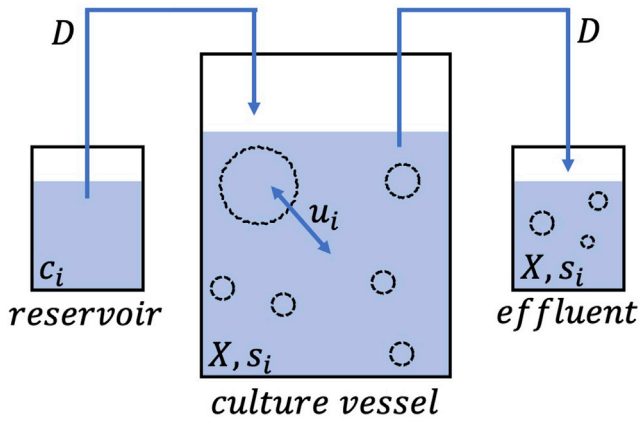


Figure 1. Schema of a chemostat

Fluxes of matter are indicated by an arrow. The most important chemostat parameters are listed: dilution rate D, cell concentration X, exchange of a metabolite between the cells and the medium u_i, and the concentration of a metabolite in the feed medium c_i and in the culture vessel s_i. Adapted from.[17]

metabolite. However, a typical network has more reactions than metabolites ($M < N$), which leads to fewer constraints than variables (fluxes).[23] The system is then under-determined. An infinite set of vectors $\boldsymbol{v}$ satisfies the system of equations.

These constraints lead to unbounded solution space. Therefore, it is common to add a set of inequalities to impose bounds on $\boldsymbol{v}$, such as:

$$\begin{aligned} lb_r \leq r \leq ub_r \\ lb_u \leq u \leq ub_u \\ 0 \leq z \leq ub_z \end{aligned} \qquad \text{(Equation 3)}$$

where $\boldsymbol{lb}_r$ and $\boldsymbol{ub}_r$ are the lower and upper bounds of the internal reactions, which typically contain information about thermodynamic irreversibility and catalytic capacity. On the other hand, $\boldsymbol{lb}_u$ and $\boldsymbol{ub}_u$ are the bounds of the exchange reactions controlling the metabolites that the network can consume or produce, and are linked to properties of the cell membrane (*e.g.* the presence of transporters, ion channels, and so forth). Finally, the biomass reaction can be upper bounded by ub_z, if necessary.

In general, any new information about the culture is integrated by adding balance-like equations (*e.g.* Equation 2) or changing the bounds of the reaction fluxes (*e.g.* Equation 3).[47] For example, we can define a new constraint that accounts for physical and spatial restrictions resulting from the limited resources accessible to the cell (*e.g.* cell volume, membrane area, enzyme solubility, proteome, and so forth)[48,49,50,51,.52] It can be formulated as:

$$\sum_j^N \left(a_j^+ r_j^+ + a_j^- r_j^- \right) \leq 1 \qquad \text{(Equation 4)}$$

where each internal reaction r_j in the network is split into its forward and backward components such that $r_j = r_j^+ - r_j^-$ and $r_j^+, r_j^- \geq 0$ where $a_j^+, a_j^- \geq 0$ are normalized cost coefficients associated with each component of the reaction j respectively. From the mathematical point of view, it is important to note that these constraints define a convex and bounded space of feasible flux configurations.[53]

Constraint-based modeling of the chemostat

In ref. [17,36] we developed a constraint-based model of genome-scale metabolic networks coupled to the dynamical equations governing a chemostat. Here we go a step further, linking culture observables with the constraints that affect the metabolic spaces at the steady state. In a chemostat, a cell culture is maintained in a continuous regime where the fresh medium is pumped into the culture vessel at the same rate that it is extracted, such that the working volume remains constant.[54] In Figure 1 we present a schematic picture of the chemostat. The dynamics of cell and metabolite concentrations in the vessel, X

($g_{CDW} \times l^{-1}$) (*CDW*: cellular dry weight) and s_i (mM), for a classic well-mixed chemostat with a single species, can be expressed as[17]:

$$\frac{dX}{dt} = (\mu - D)X \quad \text{(Equation 5)}$$

$$\frac{ds_i}{dt} = -\overline{u_i}X + (c_i - s_i)D \quad \text{(Equation 6)}$$

where $D(h^{-1})$ is the dilution rate, μ (h^{-1}) is the observable culture growth rate, c_i (mM) and $\overline{u_i}$ ($mmol \times g_{CDW}^{-1} \times h^{-1}$) are the concentration in the fresh medium and the observable exchange rate of metabolite i, respectively. Here and in what follows, we will use an overbar (as in $\overline{u_i}$) to distinguish the average value of flux across all the cells in the culture, from its value in single cells (u_i). Equation 5 says that the rate of change of X is determined by the culture growth rate and its elimination due to medium exchange. Similarly, Equation 6 reflects that the rate of change of any metabolite concentration in the vessel is a balance between its average exchange with the cells (a positive u_i means uptake) and how much of it is being pumped in and out of the vessel.

The culture growth rate (usually the relevant observable) μ, can be modeled to include any metabolic process that impacts the average growth rate of the culture (*e.g.* toxicity, cellular death rate, and so forth). In this work, we only consider the biomass production rate z (h^{-1}), so:

$$\mu = \overline{z}$$

where, as mentioned before, z is just a component of the flux vector $\boldsymbol{v}$ and $\overline{z}$ is its average value on the cell population. It models the flux requirement for cellular division.

The contribution of the negative terms in Equations 5 and 6 (right-hand side) enables the possibility of a steady state regime. One of the main applications of the chemostat is that cultivation can be sustained for a long time in a pseudo-constant environment. This fact is exploited to decouple cellular physiology from extracellular dynamical processes. This is a major difference with batch cultures, where cells are in a constantly changing environment.[56] Therefore, for a chemostat in the steady state, two new constraints can be derived from Equations 5 and 6:

$$\overline{u_i} \leq c_i D / X \quad \text{(Equation 7)}$$

$$\overline{z} = \mu = D \quad \text{(Equation 8)}$$

The first Equation 7 simply states that $s_i \geq 0$ in a steady state.[17]

With this, the full stack of equations needed to define the constraint-based model for a chemostat, in a steady state, is complete. The set of Equations 2-4 reflect the metabolic constraints within each cell, and Equations 7 and 8 constraint the average values of the fluxes in the culture.

Metabolic flux spaces

In the previous sub-sections, we presented two types of constraints. Depending on the source of the information encoded, they can restrict the flux configurations at a single-cell level or at a culture level. For instance, if a metabolic reaction is considered to be thermodynamically irreversible, a constraint enforcing such behavior must be applied to all cells in the culture. On the other hand, data derived from experimental measurements made at a population level (*e.g.* culture growth rate), are interpreted differently: as the average over the configurations of all cells in the culture. But these latter constrains do not necessarily imply that a particular restriction must be fulfilled at a single-cell level. For example, if a culture is considered to be growing at a given rate, this does not imply that all cells are growing at such speed. Some cells might be growing faster, and some slower: it is the population average that defines the measured value.

Therefore, it is convenient to introduce two spaces: $\mathbb{V}$ as the space of all feasible flux configurations $\boldsymbol{v}$ that a particular cell metabolism can display, and $\overline{\mathbb{V}}$ as the space of all feasible average flux configurations $\overline{\boldsymbol{v}}$. If the nature of all constraints acting over the system is linear, as it is in our case, both $\mathbb{V}$ and $\overline{\mathbb{V}}$ are convex polytopes.[53] The convexity of $\mathbb{V}$ implies that $\overline{\mathbb{V}} \subseteq \mathbb{V}$. That is, any observable feasible flux configuration $\overline{\boldsymbol{v}}$ at the

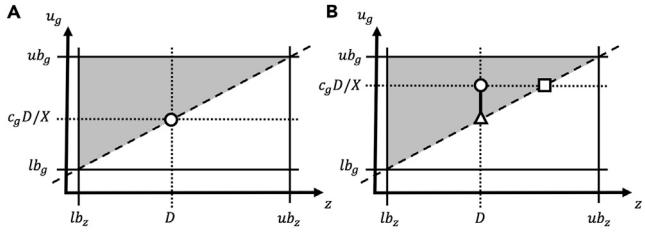


Figure 2. Projections of $\mathbb{V}$ and $\overline{\mathbb{V}}$ on the 2D plane (z, u_g) for a chemostat in steady state

The solid and dashed lines represent the constraints defining $\mathbb{V}$, which is shown in gray. The dotted lines indicate the constraints defining $\overline{\mathbb{V}}$. The left panel (A) shows a situation where the exchange bound is so tight that the projection of $\overline{\mathbb{V}}$ on the (z, u_g) plane is reduced to a single point (white circle). Meanwhile, in the right panel (B), the projection is a line (black solid vertical wide segment). In both cases, the circle marks a point $\overline{\mathbf{v}}$ where the culture is glucose limited. In panel B, the triangle marks the state of maximum yield ($\overline{z}$/ $\overline{u_g}$) and the square marks the unfeasible state which maximizes $\overline{z}$ within the given $\overline{u_g}$ limit.

population level, is a feasible flux configuration for single cells $\boldsymbol{v}$, and no unfeasible single-cell flux configuration can be observed at the population level.

The explicit distinction between both spaces, $\mathbb{V}$ and $\overline{\mathbb{V}}$, plays a key role in the definitions of the inference models in the next sections. In our context, the final formulation for these spaces can be written as:

$$\mathbb{V} = \{\boldsymbol{v} | \mathbb{S}\boldsymbol{v} = \mathbf{0};\ \boldsymbol{v} \in [\boldsymbol{lb_v}, \boldsymbol{ub_v}];$$
$$\left.\sum_j^N \left(a_j^+ r_j^+ + a_j{}^- r_j{}^-\right) \leq 1;\ r_j^+, r_j{}^- \geq 0;\ r_j^+ - r_j{}^- = r_j \in \boldsymbol{v}\right\} \quad \text{(Equation 9)}$$
$$\overline{\mathbb{V}} = \{\overline{\mathbf{v}} \in \mathbb{V} | \overline{u_i} \leq c_i D/X;\ \overline{z} = D;\ \overline{u_i}, \overline{z} \in \overline{\mathbf{v}}\}$$

An important remark about our definition of $\mathbb{V}$ is that it is independent of the chemostat dynamics, it is a property of the cells. The environmental conditions are taken into account only in the definition of $\overline{\mathbb{V}}$ through Equations 7 and 8. Moreover, since usually in a chemostat, D and $\boldsymbol{c}$ (the feed medium composition) are controlled by the researcher, X alone encodes all the information dependent on the chemostat dynamics.

Nutrient-limited cultures

As we already mentioned, we will focus our attention on chemostat cultures with a known limiting nutrient.[56] In practice, this means that from all the exchange constraints defined in Equation 7, only one, associated with this limiting nutrient, is constraining the network. In all the experiments presented here, glucose is the limiting nutrient. That is, the only constraints affecting $\overline{\mathbb{V}}$ are $\overline{z} = D$ and $\overline{u_g} \leq c_g D/X$, where $\overline{u_g}$ is the observable uptake rate of glucose and c_g is its concentration in the feed medium.

The rest of the constraints over $\overline{\mathbb{V}}$ are considered to be non-restrictive and therefore do not influence the culture. In this context, we can build a simpler definition of $\overline{\mathbb{V}}$:

$$\overline{\mathbb{V}} = \left\{\overline{\mathbf{v}} \in \mathbb{V} \middle| \overline{u_g} \leq c_g D/X;\ \overline{z} = D;\ \overline{u_g}, \overline{z} \in \overline{\mathbf{v}}\right\}$$

where, for simplicity, we are not showing the trivial constraints $\overline{u_i} \leq 0$ for the uptakes of metabolites missing in the feed medium ($c_i = 0$).

Now, the chemostat constraints are directly affecting $\overline{\mathbb{V}}$ only in the $(\overline{z}, \overline{u_g})$ subspace. Figure 2 shows a schematic representation of this subspace and the two typical scenarios that can occur in a nutrient-limited culture. In the horizontal axis we plot the growth rate of the cell and in the vertical axis the glucose consumption rate. The shadowed area represents the projection of $\mathbb{V}$, the space of feasible flux configurations, on this plane: cells do not display a (z, u_g) pair outside this area. Moreover, since in a chemostat at a steady state the average growth rate of the culture is set by the dilution rate D, all the possible solutions should be consistent with distributions where $\overline{z} = D$ (see vertical dotted lines on both panels of the figure). This reduces the possible degeneracy of $\overline{\mathbb{V}}$ in this subspace only to the u_g dimension, where the average consumption rate should be lower than $c_g D/X$ (*i.e.* the system is restricted to those distributions where $\overline{u_g}$ rests below the horizontal dotted line in the panels).

Such a combination of constraints leads to two typical scenarios. One is represented in Panel A, where the size of $\overline{\mathbb{V}}$ is reduced to the minimum volume allowed by the environmental constraints (Equations 7 and 8). There, $\overline{\mathbb{V}}$ projection is reduced to a single point (white circle in the figure). In these conditions, the culture is growing with the maximum possible $\bar{z}$/ $\overline{u_g}$ yield and larger values of X are not feasible given the nutrient feed rate ($c_g D$) and the definition of $\mathbb{V}$. We stress that, although $\overline{\mathbb{V}}$ is determined in the (z, u_g) plane in this example, that does not imply that $\overline{\mathbb{V}}$ is not degenerate in other dimensions. The other scenario is represented in Panel B of Figure 2. In this case X is not optimal, and we have a degenerate $\overline{\mathbb{V}}$ even in the (z, u_g) subspace (continuous vertical line). A major difference between both scenarios is that in Panel A, the culture reaches the full carrying capacity of the medium at the given dilution rate, while in Panel B it does not.

Inference of the metabolic state

Most constraint-based frameworks consist of two stages: I) the specification of the constraints and the definition of the feasible spaces, and II) the methods to formulate a description of metabolism from them.[57] In the previous section we already discussed point I, here we focus the attention on the second step, presenting two standard approaches.

Flux balance analysis

Flux Balance Analysis (FBA) is a widely used methodology that addresses the typical degeneration of the metabolic solution space by choosing an objective function (f) (or a stack of them) that the cell metabolism "optimizes." This assumption is not necessarily based on experimental data, but it is an educated guess about the evolutionary pressures to which the biological system is exposed.[47]

FBA has been applied, for several decades now, to model cell cultures at optimal growth conditions with remarkable results.[58,59,60,61] A popular formulation for bacterial cultures is to set the objective function equal to the biomass production rate, z. This is justified in rich medium batch cultures, during the exponential growth phase, where the fastest growing cells end up dominating the population. In these circumstances, FBA models have proven to predict the growth rate of *E. coli* cultures[24] for single carbon source conditions, and even the priority of nutrient intake for more complex media.[48]

From a computational point of view, FBA has the advantage that, if the proposed objective function is formulated as a linear function over a convex space, the optimum flux configurations can be found efficiently using Linear Programming.[62] Typically, FBA formulations[24,57,26,63] do not make an explicit distinction between population and single-cell level metabolic spaces. All constraints are applied over a unique space. Using our notation, this is formally equivalent to making $\mathbb{V} \equiv \overline{\mathbb{V}}$, which typically conceals an implicit culture homogeneity assumption. For FBA formulations this can be justified because generally, the goal is just to infer an observable flux configuration $\bar{\mathbf{v}}$ which optimizes the objective function and the convex set of constraints let to $\overline{\mathbb{V}} \subseteq \mathbb{V}$. The problem to solve can then be stated as:

$$\begin{array}{l} \text{optimize} \arg_{\bar{\mathbf{v}}} f(\bar{\mathbf{v}}) \\ \text{subject to}: \ \bar{\mathbf{v}} \in \overline{\mathbb{V}} \end{array} \qquad \text{(Equation 10)}$$

However, in general, the correct objective function is unknown and determining it can be a focus of study by itself.[26] In more complex scenarios, like cancer or tissues, the problem is particularly challenging and constitutes a severe limitation for the application of FBA models. This is aggravated because these complex scenarios are in fact the norm in nature, while optimal growth conditions are the exception. To make matters more complicated, it may well be the case that the evolution of metabolism leads to overall robustness across many conditions rather than a single condition-specific objective.[64]

Maximum entropy principle and metabolism

As mentioned before, FBA models extract from the available information (the constraints defining the metabolic spaces) a candidate flux configuration based on a given objective function. However, an important limitation is that they provide little insights into other features such as cultural heterogeneity. Additionally, the FBA solution is only affected by constraints that are directly involved in defining the optimal objective value. The remaining constraints are irrelevant, and the information encoded in them is not used.

A more general framework can be conceived using the Maximum Entropy Principle (ME)[29] to develop a probabilistic representation of the metabolic state of the culture. Here, the state of the system is not represented by a flux configuration vector, but by its distribution $\mathcal{P}(\boldsymbol{v})$ over the space of all possible

configurations. This framework has been recently used to model the phenotypic distribution of cells in culture for several growth conditions and cultivation regimes.[35,33] In the context of our constraint-based model, the Maximum Entropy Principle may be formulated as follows:

$$\begin{aligned} &\text{maximizearg}_{\mathcal{P}} \left[-\int_{\mathbb{V}} \mathcal{P}(\boldsymbol{v}) \ln(\mathcal{P}(\boldsymbol{v})) d\boldsymbol{v} \right] \\ &\text{subject to:} \\ &\boldsymbol{v} \in \mathbb{V} \\ &\overline{\boldsymbol{v}} = \int_{\mathbb{V}} \boldsymbol{v} \mathcal{P}(\boldsymbol{v}) d\boldsymbol{v} \in \overline{\mathbb{V}} \end{aligned} \qquad \text{(Equation 11)}$$

which means that from all the feasible distributions $\mathcal{P}$ we must find a distribution, which we call $\mathcal{P}_{\mathcal{ME}}$, that maximizes the (statistical) entropy subject to specific constraints. Following Jayne's[29] interpretation of the principle, ME is the least biased distribution encoding all the information that we have about the system. In other words, in absence of a mechanistic model, we consider $\mathcal{P}_{\mathcal{ME}}$ to be our best guest of the real $\mathcal{P}$, given the available information.

If $\mathbb{V}$ is bounded and the constraints applied over $\overline{\mathbb{V}}$ have the simple forms $\overline{\boldsymbol{v}}_i \leq a_i$ or $\overline{\boldsymbol{v}}_i = a_i$, where $\boldsymbol{a} \in \mathbb{R}^N$ is a constant vector (such as Constraints 7 and 8), it can be proved that $\mathcal{P}_{\mathcal{ME}}$ belongs to the exponential family[33,65]:

$$\mathcal{P}_{\mathcal{ME}}(\boldsymbol{v}) \propto e^{\boldsymbol{\beta}^T \boldsymbol{v}} \qquad \text{(Equation 12)}$$

where $\boldsymbol{\beta} \in \mathbb{R}^N$ is a vector ($\boldsymbol{\beta}^T$ is its transpose) of Lagrange multipliers, where each β_j is associated with the j^{th} reaction, used to select the appropriate $\mathcal{P}_{\mathcal{ME}}$. See Section 9.3.4 for a more formal discussion.

An important point to notice is that the ME probabilistic description of the culture metabolism allows, just like in FBA, the inference of a representative flux configuration $\overline{\boldsymbol{v}}$. More precisely, having found $\mathcal{P}(\boldsymbol{v})$ from 11, we can compute the predicted average values as $\overline{\boldsymbol{v}} = \int_{\mathbb{V}} \boldsymbol{v} \mathcal{P}(\boldsymbol{v}) d\boldsymbol{v}$. However, although both methods use the same input data (the metabolic spaces), ME can be additionally queried about other metabolic features like *e.g.* cell-to-cell growth variability, flux correlations, information variation (*e.g.* due to regulation), and so forth.[33,66] This is a major advantage of ME over FBA, the former exploits better the information available in the constraints.

In part, this is possible because we can effectively decouple the constraints defining the different spaces. Constraints at the single-cell level are used in the definition of $\mathbb{V}$ and constitute the support of the distributions in $\mathcal{P}$. Constraints at the population level (which define $\overline{\mathbb{V}}$) reduce the set of feasible distributions $\mathcal{P}$, from which the one that maximizes the entropy is selected. It is also important to remark that the ME methodology is not limited to the codification of constraints over flux averages. Other types of population constraints can be included (*e.g.* constraints over flux variances). A review of the utilization of ME as a general inference technique in biological problems can be found in.[34]

A downside of Maximum Entropy methods, in comparison with constraint-based modeling such as Flux Balance Analysis, is that solving ME models can be computationally challenging. Considerable efforts have been dedicated to develop methods capable of tackling this and related problems[67,35], but some difficulty remains, specially in comparison with the fast and robust Linear Programming methods that are needed to solve FBA models. As we will see later in discussion, we overcome this challenge by employing a variant of Expectation Propagation.[68]

RESULTS

This section contains the main results of our work. It is divided into four sub-sections. We first present a minimalist model where we explicitly discuss the difference between $\mathbb{V}$ and $\overline{\mathbb{V}}$ and the impact of its definitions on the ME solution. Then, we introduce a toy model of the metabolism and connect it with the dynamics of a chemostat considering a heterogeneous culture. The numerical data about the macroscopic quantities obtained with this model will allow us to finally expose how the understanding of the chemostat dynamics, and its influence over the cellular metabolism, can explain the data-driven findings of.[45] Additionally, the artificial data generated will be used to feed FBA and ME to evaluate the effects of heterogeneity and the metabolic spaces definitions over the inference. In this controlled scenario, we will compare the outputs of these two approaches to clarify the differences between both of them. Finally, we will make a

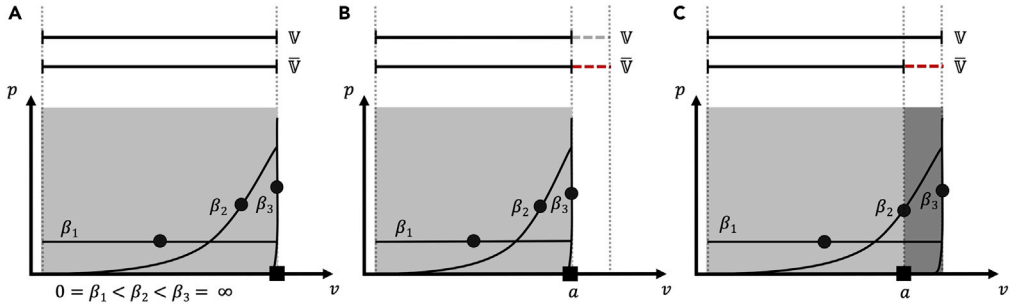


Figure 3. Schemes of three different formulations of a one-dimensional toy model

In each panel, the segments (solid black lines) at the top, represent the definitions of $\mathbb{V} \in \mathbb{R}$ and $\overline{\mathbb{V}} \in \mathbb{R}$ respectively. Those spaces are projected into the *x*-axis of the graphs (following the dotted lines). Each graph contains three $\mathcal{P}_{\mathcal{ME}} \propto e^{\beta v}$ distributions, mapped over $\mathbb{V}$, labeled by its $\beta \in \mathbb{R}$ values: $\beta_1 = 0$, $\beta_2 \in (0, +\infty)$ and $\beta_3 \to +\infty$. The solid black circles mark the mean of the distributions (only its *x*-axis coordinates have meaning) and the solid square mark the value of an optimum FBA solution. Panel A shows a formulation where $\overline{\mathbb{V}} = \mathbb{V}$, whereas Panels B and C show two different formulations for encoding a real constraint over $\overline{\mathbb{V}}$, $\overline{\mathbf{v}} \leq a$ (dashed lines in the segments).

similar comparison with data obtained from real continuous cultures experiments for *E. coli*. With this, we test the feasibility of using ME for genome-scale metabolic networks and further support the analysis done in the toy model.

The minimum picture

In order to gain insights into a few basic aspects of the ME formulation, we will use a model with a single free reaction, $v \in \mathbb{R}$ (see Figure 3), which is affected by only one constraint ($lb_v \leq v \leq ub_v$). This greatly simplifies the problem because the encoding of some equations (such as (2) and (4)) is not necessary.

For this model, the distribution (12) has the exact form:

$$\mathcal{P}_{\mathcal{ME}}(v) = e^{\beta v}\psi(v)/Z$$

where $Z = \int \mathcal{P}_{\mathcal{ME}}(v)dv$ is the normalization constant, β is a scalar, and $\psi(v)$ is an indicator function that returns one when $v \in \mathbb{V}$ and zero otherwise.

In Figure 3 we show a schematic representation of three different formulations of $\mathbb{V}$ and $\overline{\mathbb{V}}$ for this one-dimensional model. Each panel contains a graph with the characteristic $\mathcal{P}_{\mathcal{ME}}$ distributions for three distinctive β values. When $\beta = \beta_1 = 0$, $\mathcal{P}_{\mathcal{ME}}$ is the homogeneous distribution over $\mathbb{V}$. Essentially, the exponential plays no role at all and each flux value v is equally likely. This is the regime of the largest entropy and maximum heterogeneity. On the other hand, at $\beta = \beta_3 \to +\infty$, $\mathcal{P}_{\mathcal{ME}}$ becomes a Dirac's delta, which concentrates all the biomass density at the upper extreme of $\mathbb{V}$. An analogous situation is found, but at the lower extreme, if the sign of β_3 is negated. A Dirac's delta has the lowest possible entropy, and the system is fully determined. Finally, for $\beta = \beta_2 \in (0, +\infty)$, any intermediate average flux value can be achieved by finding the appropriate β value.

With these concepts clear, we can now start to dissect the differences between the three panels and their implications in the inference results for ME and FBA.

First, let's look at panel A. There we do not impose any extra constraints on the average fluxes and so $\overline{\mathbb{V}} \equiv \mathbb{V}$. This resembles the common constraint-based modeling (CBM) formulations for describing exponential growth phases of batch cultures in rich media, where the culture (once defined which metabolites are available) is restricted only by the intrinsic capabilities of the cell metabolism (single-cell level constraints).[58] In this scenario, each of the $\mathcal{P}_{\mathcal{ME}}$ distributions (one for each β value) are feasible and all its mean values (black circles) fall inside the feasible space (shadow area). If we consider an FBA formulation which maximizes $\overline{v}$ over $\overline{\mathbb{V}}$, its solution (black square) is recovered by ME at $\beta = \beta_3 \to +\infty$. This distribution is a Dirac's delta, so our model is describing the culture as a homogeneous system (*i.e.* all cells display the same metabolic state). It is important to remark that at $\beta_i \to \pm\infty$ ME will always find a mean flux vector that optimizes the flux i in $\mathbb{V}$, so it can be viewed as a generalization of FBA.[35]

The situation becomes more subtle when we introduce more restrictions (*e.g.* provided by data obtained from measurements in the culture). Consider for example the new bound $\overline{v} \leq a$ (where a is a constant, resembling Equation 7) that is supposed to be only applicable over $\overline{\mathbb{V}}$. Panels B and C in Figure 3 show two alternative ways to introduce such constraints and their consequences. A first approach does not distinguish between the spaces ($\overline{\mathbb{V}} \equiv \mathbb{V}$). In our two spaces formalism, this is equivalent to apply all constraints always over $\mathbb{V}$ (gray dashed line in Panel B). Although $\overline{\mathbb{V}}$ is reduced accordingly, to affect $\mathbb{V}$ this way is unjustified because this reduction does not follow the rationality leading to the constraint (*i.e* $\overline{v} \leq a \not\Rightarrow v \leq a$). On the other hand, Panel C shows the alternative scenario where the distinction is made and $\mathbb{V}$ is unaffected. Note that in both cases we have the same $\overline{\mathbb{V}}$, therefore, any inference method that relies only on this space produces the same results (*e.g.* FBA).

The comparison of the ME solutions over these two panels illustrates the consequences of adding an unjustified assumption into the space's formulation. Although $\overline{\mathbb{V}}$ is the same in both cases, and the average flux value reported by FBA and ME solutions are not affected, other features of the solution (such as heterogeneity) do differ. For instance, in panel B, the FBA's solution is reached at $\beta = \beta_3 \rightarrow +\infty$, whereas in the right panel, $\mathcal{P}_{\mathcal{ME}}$ achieved the same mean at a finite β value, $\beta = \beta_2$. Therefore, the $\mathcal{P}_{\mathcal{ME}}$ at these β values are quite different. Although in the case of Panel B the system is fully determined, the more rigorous definition used in panel C shows that it is impossible to completely determine the system with the available constraints. FBA formulations usually use the more simple Panel B's scenario because its goal is to infer an optimum flux configuration, which hides the inconsistency, but ME formulations must take extra care if further analysis is intended. In the particular scenario of a nutrient-limited chemostat culture at the steady state, it is not difficult to incur such biased ME formulations. As we already discussed in Section 2.4, Equations 7 and 8 impose strong constraints over $\overline{\mathbb{V}}$, generally leading to a situation where $\overline{\mathbb{V}} \subset \mathbb{V}$.

However, in the literature, it is not usual to find ME formulations making explicit distinction between multiple spaces. For instance, in,[69] the growth rate and the glucose uptake are directly encoded into $\mathbb{V}$. Both reaction's upper and lower bounds in Equation 3 are set to be equal to the reported experimental measurement. Another example can be found at.[36] There, a similar chemostat model is used, but only the observable growth rate ($\overline{z} = D$) is encoded accordingly into $\overline{\mathbb{V}}$. The model uses a single scalar β parameter, and all the constraints over the exchanges are enforced by restricting $\mathbb{V}$ directly. Precisely, a further simplification $u_i \leq c_i D/X$ is made (note that the originally derived from the dynamic model is $\overline{u_i} \leq c_i D/X$), which might lead to a situation analogous to the one represented in Panel B of Figure 3. In Section 4.5 we discuss some consequences of such simplifications.

With this understanding, we reformulate the ME model at.[36] We respect the original form of the constraints over $\overline{\mathbb{V}}$ for the observable growth rate, but we also add the corresponding constraint over the uptake of the limiting nutrient. The formulation will have two Lagrange multipliersβ's (i.e., two non-zero components in the $\boldsymbol{\beta}$ vector of Equation 12). One to enforce the biomass constraint ($\overline{z} = D$) and the other to enforce the mean glucose uptake constraint ($\overline{u_g} \leq c_g D/X$), see Section 9.3.4 for more details. Given that most continuous cultures are nutrient-limited, this has the advantages of potentially avoiding all biases related to the uptakes by adding only an extra parameter. Additionally, it leads us halfway from reconciling our formulation with the data-driven results from.[45] It correctly suggests that the relevant β parameters of the distribution are related to the growth rate and the limiting nutrient uptake rate, but the latter is still free to vary. Next, we explicitly show, using a simple model that in fact, it can be further constrained.

A simple mechanistic model accounting for the chemostat dynamics

In order to gain a deeper insight into the effects of the chemostat dynamics on the form of the metabolic spaces in a controlled system, we introduce a simple model to mimic the metabolism of the cell. In this model, the cell metabolism is reduced to a small size network (see Section 9.3.3) resembling the core (fermentation/respiratory/pentose phosphate) metabolic pathways. The model has 3° of freedom, which we choose to call z, representing the growth rate of the cell, u_g the uptake of a nutrient (glucose), and u_o the uptake of oxygen (a non-limiting nutrient).

To consider the presence of metabolic noise (heterogeneity) we reformulate Equation 5. Such reformulation will highlight the key advantage of ME over FBA-based models. In the new formulation we account for

the time evolution of the biomass associated with each feasible flux configuration $\mathbf{v}$, and introduce a source of heterogeneity $\varepsilon \in (0, 1]$. This defines in an idealized manner, a stochastic biomass redistribution over $\mathbb{V}$, accounting for, *e.g.*, metabolic switches,[49] proteome re-allocations,[52] mutations.[70] The final, non-discretized version, of the dynamic equations for the chemostat are:

$$\begin{aligned}\frac{dX(z,u_g,u_o)}{dt} &= (1-\varepsilon)zX(z,u_g,u_o)\\ &+\frac{\varepsilon}{|\mathbb{V}|}\int_{\mathbb{V}} z'X\left(z',u_g',u_o'\right)dz'du_g'du_o'\\ &-DX(z,u_g,u_o)\end{aligned} \qquad \text{(Equation 13)}$$

$$\frac{ds_g}{dt} = -\int_{\mathbb{V}} u_g X(z,u_g,u_o)dzdu_gdu_o + (c_g - s_g)D \qquad \text{(Equation 14)}$$

where $X(z, u_g, u_o)$ is the biomass concentration (more precisely the biomass concentration density) associated with the given flux configuration and $|\mathbb{V}| = \int_{\mathbb{V}} dzdu_gdu_o$ is the volume of $\mathbb{V}$. Equations for the evolution of other metabolites can be stated analogous to (14), but we will focus our attention only on the limiting nutrient. For computational purposes, the metabolic space $\mathbb{V}$ was discretized (details at Section 9.3.3).

Equation 13 expresses that the rate of change of the biomass concentration associated with a given flux configuration depends on the balance between cellular growth (first two terms) and the extraction of biomass due to the chemostat dilution rate (last term). The two growth terms differ in that the first is related to the growth potential associated with the given flux configuration (local) while the second depends on the growth capacity of the whole culture (global). The global term is just the average of all local terms scaled by ε. At any particular time, it contributes equally to the growth of the biomass associated with each flux configuration. The diffusion parameter ε is used to control how much of the $X(z, u_g, u_o)$ growth is due to its local capacity or because of the relocation of biomass from the rest of the culture. In the extreme $\varepsilon = 1$, all flux configurations have the same growth potential irrespective of its own z value, which leads, if feasible, to the larger heterogeneity of the system. In the opposite case, when $\varepsilon \to 0$, $X(z, u_g, u_o)$ evolves depending exclusively on its local growth potential (z), and there is negligible biomass reallocation.

The biomass associated with each flux configuration allows the computation of the biomass distribution $\mathcal{P}$ at every time step of the simulation, by defining $\mathcal{P}$ as:

$$\mathcal{P}(z,u_g,u_o) = X(z,u_g,u_o)/X \qquad \text{(Equation 15)}$$

where $X = \int_{\mathbb{V}} X(z,u_g,u_o)dzdu_gdu_o$ is the total biomass concentration of the culture at a given time.

Notice also that any constraint acting upon an observable restricts the set of feasible biomass distributions and influences all other observables. For instance, Equation 14 is affected by the implicit constraint $s_g \geq 0$ and since, D, c_g and $\mathbb{V}$ are time independent, such constraint can only be implemented by dynamically transforming $\mathcal{P}$ to guarantee that $\overline{u_g} \leq c_g D/X$ when $s_g \to 0$. Then, to enforce the constraints over $\overline{\mathbb{V}}$, we must add an explicit transformation over $\mathcal{P}$ that links (13) and (14) together.

The transformation that we use is explained in detail in Section 9.3.3, but in practice it reduces to: at any instant of time in which $s_g \approx 0$ and $\overline{u_g} > c_g D/X$ we enforce equality $\overline{u_g} = c_g D/X$ to be true by re-scaling $\mathcal{P}$. With this, we link both (13) and (14) together, keeping X bounded while the culture reaches non-trivial steady states in all the feasible regions of the model. It is important to remark that modeling a more realistic normalization procedure for $\mathcal{P}$ is out of the scope of this work. Our only requirement is that the system, at a steady state, should be subject to the defined constraints, while avoiding any additional unjustified restrictions on $\mathbb{V}$ or $\overline{\mathbb{V}}$. Thus, although this term models the dependency of an observable metabolite uptake rate with its external concentration[71] and its interplay with many complex intracellular processes that contribute to metabolic heterogeneity in cells, such as metabolic shifts[72,52] or mutations,[70], we here select the simplest (though highly idealized) rule that serves this purpose.

Using these dynamics for the chemostat we performed extensive simulations of Equations 13 and 14 for different values of D and ε keeping a constant c_g, and initial X, s_g and $\mathcal{P}$ (a uniform distribution over $\mathbb{V}$).

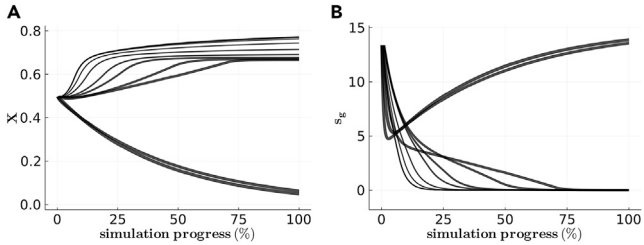


Figure 4. Dynamic simulation of the chemostat

Panels (A) and (B) show the time series from the dynamic simulations of the total cell concentration (X) and the nutrient concentration in the vessel (s_g) respectively for a given D and different ε values. The width of the lines is proportional to $\varepsilon \in [0.001, 1]$.

As an example of the output, Panels A and B of Figure 4 show the time series for the total cell and glucose concentration (the limiting nutrient) in the chemostat, obtained at a constant D for differente's. When the simulations reach a non-trivial steady state we computed the flux distributions and the observables needed to characterize both, the metabolism of the culture and the chemostat environment.

In Figure 5, Panel A, we show a heatmap that represents the volume of $\overline{\mathbb{V}}$ in the $(\overline{z}, \overline{u_g})$ subspace as a function of the constraint bounds D and $c_g\, D/X$. Darker areas in the map mean that the culture steady state configuration is near to the maximal restrictive power of the environmental constraints (and so the minimal $\overline{\mathbb{V}}$ volume). The markers in the figure represent the location of the chemostat parameters of a set of simulations at a steady state. The size of the markers are proportional to $\varepsilon \in [0.001, 1]$.

As can be appreciated, for a particular D value, the larger the ε used in the simulation the larger the volume of $\overline{\mathbb{V}}$ at a steady state. This can be explained by the combination of two factors: i) The tendency of the culture to reach a steady state when the limiting nutrient is depleted, and ii) the redistribution of biomass over $\mathbb{V}$ due to heterogeneity. The first factor can be explained by the feasible dynamic's tendency to increment X indefinitely in a nutrient-unlimited condition (see Section 4.6). The culture will only stop growing (and so the steady state reached) when the limiting nutrient is almost depleted. For the simulations, this implies that at the steady state, the culture will be consuming glucose at a rate close to the upper limit ($c_g D/\, X$). Panel B of Figure 5 shows a correlation that directly supports this claim. Note that this behavior is independent of the metabolic noise, it is a consequence of the nutrient-limited condition.

The second factor can be explained if we study the relation between X and $\overline{u_g}$ at a steady state. From the uptake constraint ($\overline{u_g} \leq c_g D/X$) we can see that at a given glucose feed rate ($c_g D$), X reaches a maximum when $\overline{u_g}$ is minimal. This is equivalent to say that X will be maximal when all cells are consuming glucose at its maximum feasible $\overline{z}/\overline{u_g}$ yield (dashed line in Figure 5). Such necessary homogeneity directly links the culture's heterogeneity with X at a steady state. If the stochastic redistribution of biomass is not null ($\varepsilon > 0$), metabolic states with lower yields will be occupied by the cells (see global term in Equation 13). In those cases the culture will still tend to maximize X, but the heterogeneity will prevent it from reaching the optimum value (and so the minimum $\overline{\mathbb{V}}$ volume) at a steady state. Panel C of Figure 5 shows such a tendency by correlating the results from the simulations with the theoretical maximum X. This is estimated by computing the minimum u_g value compatible with the given growth rate $z = D$ and using the glucose-limited uptake bound ($\max(X) = c_g D/\min(u_g)$).

Given those results, if we revisit Figure 2, all glucose-limited steady states will be located inside $\overline{\mathbb{V}}$ at the circle markers ($\overline{u_g} \approx c_g D/X$). Additionally, a culture with minimal heterogeneity ($\varepsilon \to 0$) will display a $\overline{\mathbb{V}}$ configuration at a steady state as represented in Panel A. Any other $(\overline{z}, \overline{u_g})$ pair is disallowed due to the $\overline{z} = D$ constraints and the maximization of X. If significant heterogeneity is introduced ($\varepsilon \gg 0$), the steady state will be configured as represented in Panel B. Note that the culture $(\overline{z}, \overline{u_g})$ will be far from the optimum $\overline{z}/\overline{u_g}$ yield (marked as a solid triangle in Figure 2). It is important to remark that no constraints (other than the ones defining $\mathbb{V}$) are being formulated to control the distribution into other free dimensions of $\overline{\mathbb{V}}$.

These ideas match the data-driven results reported at.[45] In their formulation, they exploited all the experimental observable fluxes for learning the relevant parameters of their model and showed that only two of these were actually relevant to minimize the inference error, one related to the growth rate and another

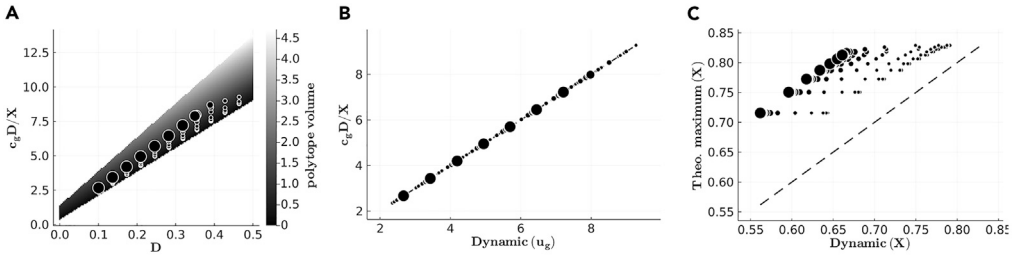


Figure 5. : Study of the toy model dynamic simulation steady state

Panel (A) shows a heatmap measuring the volume of $\overline{\mathbb{V}}$ in the ($\overline{z}$, $\overline{u_g}$) subspace (length of the vertical short solid line in Figure 2 Panel B) as a function of the steady state parameters D and $\overline{u_g}$ upper bound. Darker regions mean smaller volumes (log scale). Over the map, the locations of the steady state parameters of a set of simulations are represented by markers. Panel (B) shows a correlation between the dynamic $\overline{u_g}$ at the steady state and the glucose uptake upper bound value c_gD/X. Panel C shows a correlation between the dynamic X at steady state and the theoretical maximal X given the constraints bound values. The size of the markers are proportional to $\varepsilon \in [0.001, 1]$.

with the limited nutrient observable uptake. Alternatively, we just showed why the dynamics of the chemostat leads typically to a steady state where $\overline{u_g} \approx c_gD/X$ and $\mu = D$. Now, the relevant parameters (β) can be determined using chemostat controlled information (D, c_g) and the cell concentration (X) as the only culture-dependent magnitude. It implies that no other external experimental values are required for parameterizing the distributions and that the data-driven approach has been explained with a mechanistic-motivated model. Actually, in a culture with minimal heterogeneity, X will approach the theoretical maximum which is computable *a priori* from the network and the known nutrient supply using FBA.

Metabolic noise, degeneration, and flux inference

Inspired by these results we now exploit the simulations to additionally evaluate the impact of the definition of the spaces and the heterogeneity on the inference of observable fluxes. Moreover, in addition to ME, we test five different formulations of FBA. The first *FBA*'s objective function we use is the common maximization of biomass. Note that in this case, we do not force the constraint over the growth mean ($\overline{z} = D$). The other four FBA stacks of objective functions account for each one of the $\overline{\mathbb{V}}$'s vertices. Because of the simplicity of the toy model and the chemostat constraints at steady state, $\overline{\mathbb{V}}$ has only four vertices. This means that FBA (as formulated in 10) must yield a single one of these four possible optima (one for each vertex[47]) for any objective function.

Figure 6 shows the correlations between the artificial data and the models for all simulations that reached a non-trivial steady state condition. The mean value for each free flux (z, u_g, u_o) is computed from the dynamic biomass distribution (15) at the steady state, the inferred ME distribution (11), and the solution of the *FBA* optimization (10). In this case, each $\mathcal{P}_{\mathcal{ME}}$ is inferred by finding the two beta parameters that made the distribution fulfill both observable constraints ($\overline{z} = D$ and $\overline{u_g} \leq c_gD/X$) and maximize the entropy (see Section 9.3.4). Each row corresponds to a different inference technique and each column to a free flux.

The first row of the figure presents the results for $\text{FBA}^{(M00)}$ formulation (see Figure 6 caption for notation details), which maximizes the biomass rate. This formulation consistently overestimates z (recall that $\overline{z} = D$ was not enforced this time), but correctly predicts the glucose uptake u_g. The maximization of $\overline{u_g}$ and the additional overestimation of u_o in $\text{FBA}^{(M00)}$ is consistent with the structure of the network and the maximization of z. For instance, the consumption rate of glucose and oxygen is proportional to the *ATP* production rate, which is a reactant in the biomass equation (see Section 9.3.3). As a reference, in Figure 2 Panel B, this solution is located in the squared marker. In our model, for such a solution to be feasible (and the objective function to be valid), either $\varepsilon \to 0$ or $D = ub_z$.

From the second to the fifth row of Figure 6 the results of the rest of FBA's formulations are shown. These formulations respect the chemostat constraint over $\overline{z}$, as is trivially appreciated in the correlations of the first column. Here, the main result is that although the formulations maximize $\overline{u_g}$ ($\text{FBA}^{(fMm)}$ and $FBA^{(fMM)}$) reproduce two of the three free fluxes of the simulations, in general, FBA was unable of capturing the whole metabolic state of the culture. In particular, the formulations were unable to infer $\overline{u_o}$. Although for $\text{FBA}^{(fmM)}$ and $\text{FBA}^{(fmm)}$ the error is less significant given that the minimization of $\overline{u_g}$ also reduces the $\overline{u_o}$

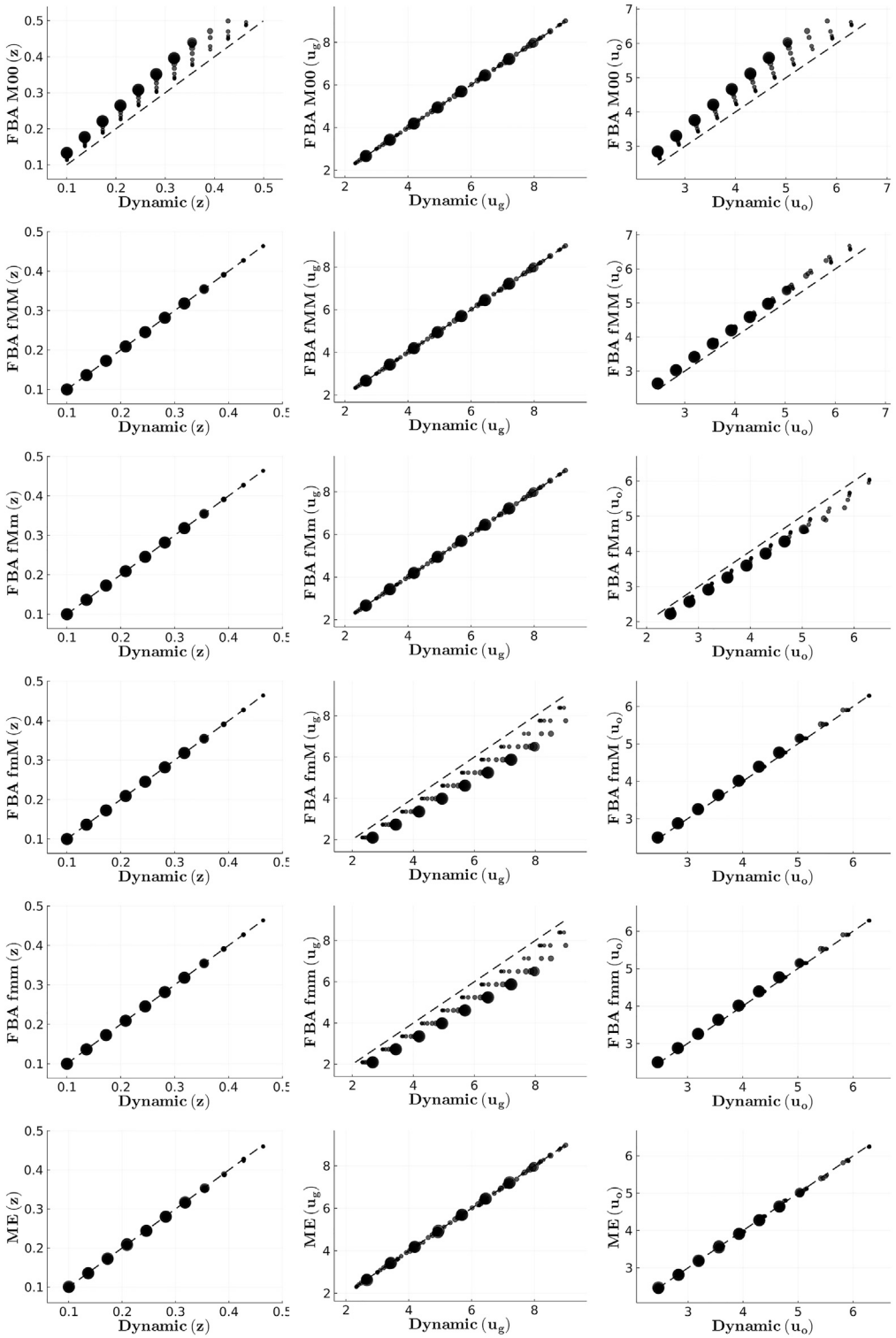


Figure 6. Inference of observable flux configurations for the toy model dynamic simulations

Correlations between the dynamic (x axis) and inferred mean values (y axis) for the free fluxes z, u_g and u_o of the toy network. Each row shows the results of one inference method, and each column of one free flux. In the case of the FBA formulations, we specify the sequence of objective functions required to determine a solution by using a

Figure 6. *Continued*
character triple where: "m" means minimization, "M" maximization, "f" that the flux was fixed to a given value, and "0" that no further action was required. The position of the character expresses the action over z, u_g or u_o respectively (Ex: "M00" means that the maximization of z lead to a single solution). The size of the markers encodes the value of $\varepsilon \in [0.001, 1]$.

feasible range. A possible explanation is that as stated before, the chemostat steady state and the nutrient-limiting condition are only constraining $\overline{\mathbb{V}}$ in the $(\overline{z}, \overline{u_g})$ subspace. If $\overline{\mathbb{V}}$ is degenerate in other dimensions, the observed value is not enforced to be an optimum. Additionally, the error induced increases with ε (in Figure 6, the value of ε is proportional to the size of the markers). Stochasticity leads to more degeneracy and this influences negatively the performance of FBA. This is especially significant given that, as we mentioned, these formulations exhaust the space of possible solutions that linear objective functions can yield for this simple model. That is, it is not the ignorance of the correct FBA's optimization function that is causing these results, it is the fact that the culture is not well described by any optimal metabolic state.

On the other hands, in the last row of Figure 6 we present the results of ME. The panels show that, ME reproduces all the observables independently of the stochasticity of the metabolism. Notice however, that it does so, without explicit inputs about any optimization function followed by the cell. Such good inference results support the idea that the simulation observables were affected significantly only by the constraints used in the definition of the metabolic spaces included in ME. More importantly, it suggests that adding further assumptions will likely bias the ME's solution rather than improve it.

One of the advantages of ME over FBA is that its solution is a full probabilistic description of the metabolic state of the culture. Figure 7 shows the marginal distributions for the free fluxes at different values of ε for simulations at a fixed D. The upper row presents the "real" distributions produced by the dynamical simulation, and the lower row the ones inferred using ME. As we already showed before, ME infers correctly the mean values (dotted lines) of the distribution, but although we can see that it also describes quite well the real shape of the distributions, there are differences.

This gives us an important insight: even if we encode correctly the constraints that are defining $\mathbb{V}$ and $\overline{\mathbb{V}}$, and if these definitions really describe the boundary of the experiments (the simulations), our ME formulation does not include information about constraints acting over higher order moments of the distributions. In this particular case, the arbitrary $\mathcal{P}$ transformation introduced on the dynamic for enforcing the moment constraints, although not affecting $\mathbb{V}$ or $\overline{\mathbb{V}}$, is probably generating a non-uniform effect that differentiates the distribution computed directly from the simulations from the one inferred through ME, $\mathcal{P}_{\mathcal{ME}}$. If those extra constraints were encoded in the model, ME is expected to recover the distributions completely.[65]

Another interesting feature of the results is that even at a very low metabolic noise ($\varepsilon \to 0$), the distributions show a high variance. This is a direct consequence of the definition of the metabolic spaces and the culture parameters. In the typical case, where $\overline{\mathbb{V}} \subset \mathbb{V}$, (see Panel C in Figure 2) the feasible (non-trivial) distribution with minimum entropy is still degenerated (the optimum is reached at $\beta \neq +\infty$). This can be reduced further only by adding more restrictions. The stochasticity (ε) just adds more entropy on top of this threshold.

Observable flux configuration inference in *E. coli*

In this section, we reproduce an analysis similar to the one made previously, but using a genome-scale metabolic network[55] and a set of real experimental observations obtained during *E. coli* glucose-limited continuous cultures[73,74,.75]

In our model, $\overline{\mathbb{V}}$ is constrained only in the (z, u_g) subspace (see Section 2.4). Aiming to elucidate how much the observable space is restricted in real glucose-limited cultures, we contextualized the genome-scale metabolic network according to the experimental conditions (see Section 9.3.1 for details). Later, we compute the volume of the $\overline{\mathbb{V}}$ (z, u_g) subspace. Figure 8, panel A, shows a heatmap that illustrates such volume as a function of the parameters of the chemostat steady state (analogous to the one in Figure 5). The area outside the heatmap (shadow area) represents the unfeasible fluxes. As can be noticed, all experiments (triangular markers) are very close to the limit of the feasible space at the darkest region of the map. In this case, we say that the set of steady state parameters approaches the restrictive limit imposed by the chemostat's constraints (minimum $\overline{\mathbb{V}}$ volume), similarly to the scenario described in panel A of Figure 2. The heatmap shows the results for one dataset, but the others displayed a similar behavior.

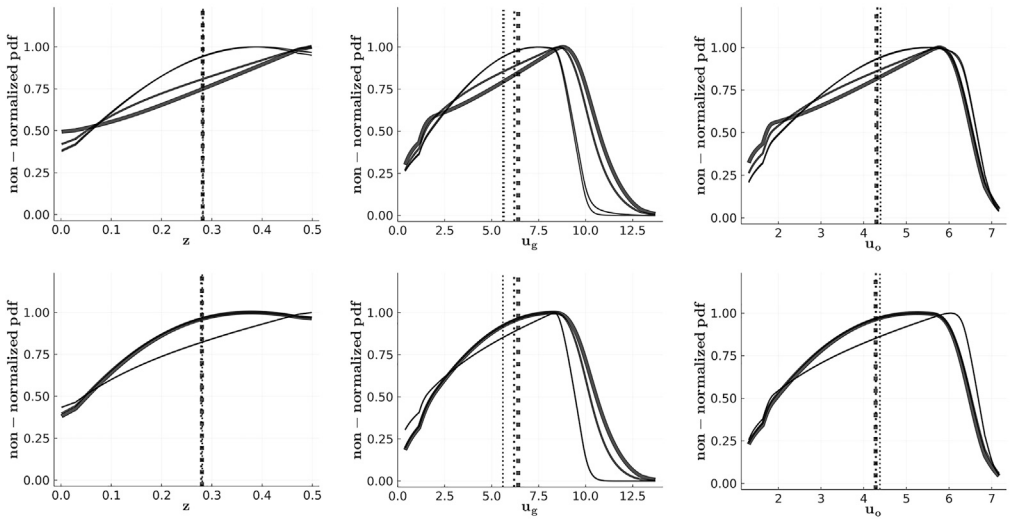


Figure 7. Toy model steady state marginal distributions

Steady state marginal distributions for z (left column), u_g (center column) and u_o (right column) from the toy model dynamic (top row) and ME (bottom row). All results are at a fixed D value, while distributions at different ε are shown. The dotted lines mark the mean for each distribution. The width of the lines is proportional to $\varepsilon \in [0.001, 1]$.

Although, in principle, the experiments can be located at any point within the feasible space, they all sit at the border of the feasible/unfeasible transition. As mentioned before, this transition occurs with the maximum theoretical value expected for X. Panel B of Figure 8 supports this idea: the cultures are close to the theoretical maximum X derived from the experimental conditions and the used metabolic network. Again, the maximum is computed by finding the minimum u_g compatible with the given growth rate $z = D$ and deriving it from the glucose-limited uptake bound ($\max(X) = c_g D / \min(u_g)$).

As stated before, such maximization of X is only possible if the culture is consuming glucose at a rate that nearly matches the nutrient input feed rate ($c_g D$), which implies that most of the residual glucose in the vessel is depleted. For completeness, in Panel C of the figure, the reported concentration of residual glucose in the vessel s_g relative to the feed concentration c_g is shown. As can be observed, all experiments had imperceptible or small amounts of glucose at the steady state (for the higher values, it is less than 4% of the feed concentration).

Finally, we used ME to infer a biomass distribution $\mathcal{P}_{\mathcal{ME}}$ for each experimental condition. In order to do that, and in analogy with the simple model above, we computed the two free components on the $\boldsymbol{\beta}$ vector that allow us to enforce the moment constraints ($\overline{z} = D$ and $\overline{u_g} \leq c_g D / X$) and maximize the entropy (see Section 9.3.4). Due to the large number of variables involved in this case, the $\mathcal{P}_{\mathcal{ME}}$ functional becomes intractable. To overcome this difficulty we use the *Expectation Propagation* (EP)[68,76,37] algorithm in order to approximate these distributions (see Section 9.3.5 for details). Additionally, we used a set of four FBA formulations as a reference to compare the performance of ME. Although this time, for a genome-scale network, $\overline{\mathbb{V}}$ is not as simple as in the toy model and the full set of possible lineal FBA solutions becomes intractable. We first introduced two objective functions common in the literature. It has been found that for chemostat cultures, the maximization of *atp* or biomass yield (the later equivalent to the minimization of $\overline{u_g}$) objectives provide better results approximating experimental data than other tested functions.[26] In addition to those two, we defined the maximization of the glucose uptake $\overline{u_g}$ (motivated by the glucose-limited condition) and the traditional maximization of biomass rate as objective functions to be tested.

In Figure 9 we show the correlations between a set of inferred fluxes (exchanges and internals) and the one experimentally reported in different experimental conditions. The bottom row shows the results for all conditions together. The first four columns represent the different versions of FBA and the later results obtained using ME.

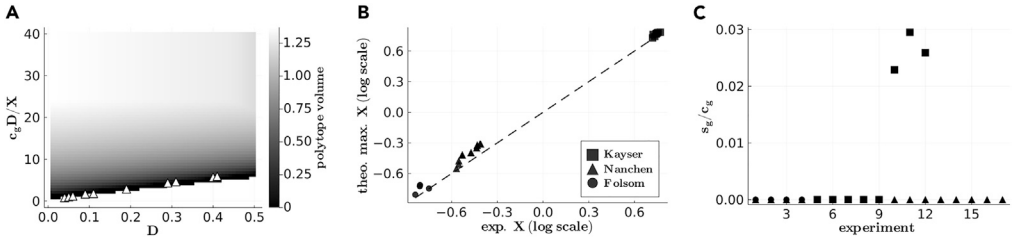


Figure 8. Study of *E. coli* chemostat culture steady state

The left panel (A) shows a heatmap of the polytope (z, u_g) projection box volume (log scale) as a function of the steady state parameters, D and c_gD/X. The triangles show the experiment location in such space. Dark regions correspond with the scenario described in Figure 8 Panel A. Data shown only for *Nanchen*74, but the rest of the datasets displayed similar behavior. On the central panel (B), it is shown a correlation (log scale) of the theoretical maximum X as computed using the metabolic network with respect to the experimentally reported. Finally, the right panel (C) shows the residual glucose in the culture (s_g) relative to its concentration in the feed medium (c_g) for all experiments. Marker shape denotes the experimental data source.

The first thing to notice is that there are significant differences between the results of the FBA formulations. This highlights the sensitivity of FBA to the definition of the objective functions in the presence of a degenerate solution space. Interestingly, biomass optimization appears to be the function with better performance. But, since it lacks the $\overline{z} = D$ constraint, this formulation significantly overestimates the biomass production rate at a steady state (there are similar findings at[37]). This can be corroborated in Figure 10 where the Absolute Relative Error (*ARE*) for all experiments is shown. Another promising FBA approach is the maximization of u_g. In this case, although the growth rate is constrained to match D, the formulations overestimate $\overline{u_g}$. Given glucose was the limiting nutrient affecting the growth rate of the network, a mirror effect between these two formulations is expected.

On the other hand, the remaining two FBA formulations show major deviations from the experimental values. As shown in Figure 10, the minimization of u_g has the worst correlations for the reactions of the Krebs cycle and is not that good for the pentose phosphate pathway. In the case of the maximization of *ATP*, it has at the same time, the best performance for the Krebs cycle but the worst for the pentose phosphate pathway from all formulations.

Furthermore, the FBA model that maximizes $\overline{u_g}$ and ME inferred non-zero acetate production rates (see Figure 9) for all datasets, although the ME ill-prediction is less pronounced. For all the studied cultures the dilution rate was smaller than 0.5 h^{-1} (below the acetate switch[52]) and therefore, experiments do not report acetate production. Additionally, all methods show poor correlations for the produced *CO2* (blue square markers). A sustained overproduction of *CO2* is predicted consistently by the network. However, the data source reporting gas exchanges come from a culture where the experimental carbon recovery was satisfactory (>93%), and no carbon-rich byproduct, other than *CO2* and biomass itself, was produced.[73] This suggests that the inferred overproduction of carbon-rich byproducts can be related to an underestimation of the carbon requirements in the biomass equation, which can also affect the ill-prediction of acetate mentioned before. Moreover, only two of the reported internal fluxes were predicted significantly wrong and consistently by all methodologies: the flux through the glyoxylate cycle (pink markers) and the pyruvate kinase *PYK* (dark green markers) (see Figure 10). However, notice that ME always predicted a non-zero flux, whereas FBA generally assigned an exact zero value through them. The glyoxylate cycle, in particular, is notoriously known to be difficult to predict by linear optimization formulations[33,69], which frequently assign a zero flux to it.

Finally, contrasting with the lack of generality of the evaluated FBA formulations, ME performs among the best in all the studied subsystems (see Figure 10). It is comparable with the best FBA formulation (biomass optimization), but it also resolves satisfactorily $\overline{z}$ and $\overline{u_g}$. This is an important result if we recall that ME uses less information than FBA (lack of objective function). Such a situation resembles the results obtained in the toy model section, suggesting again that the metabolic state of such cultures might be not well described by a polytope vertex (an optimum).

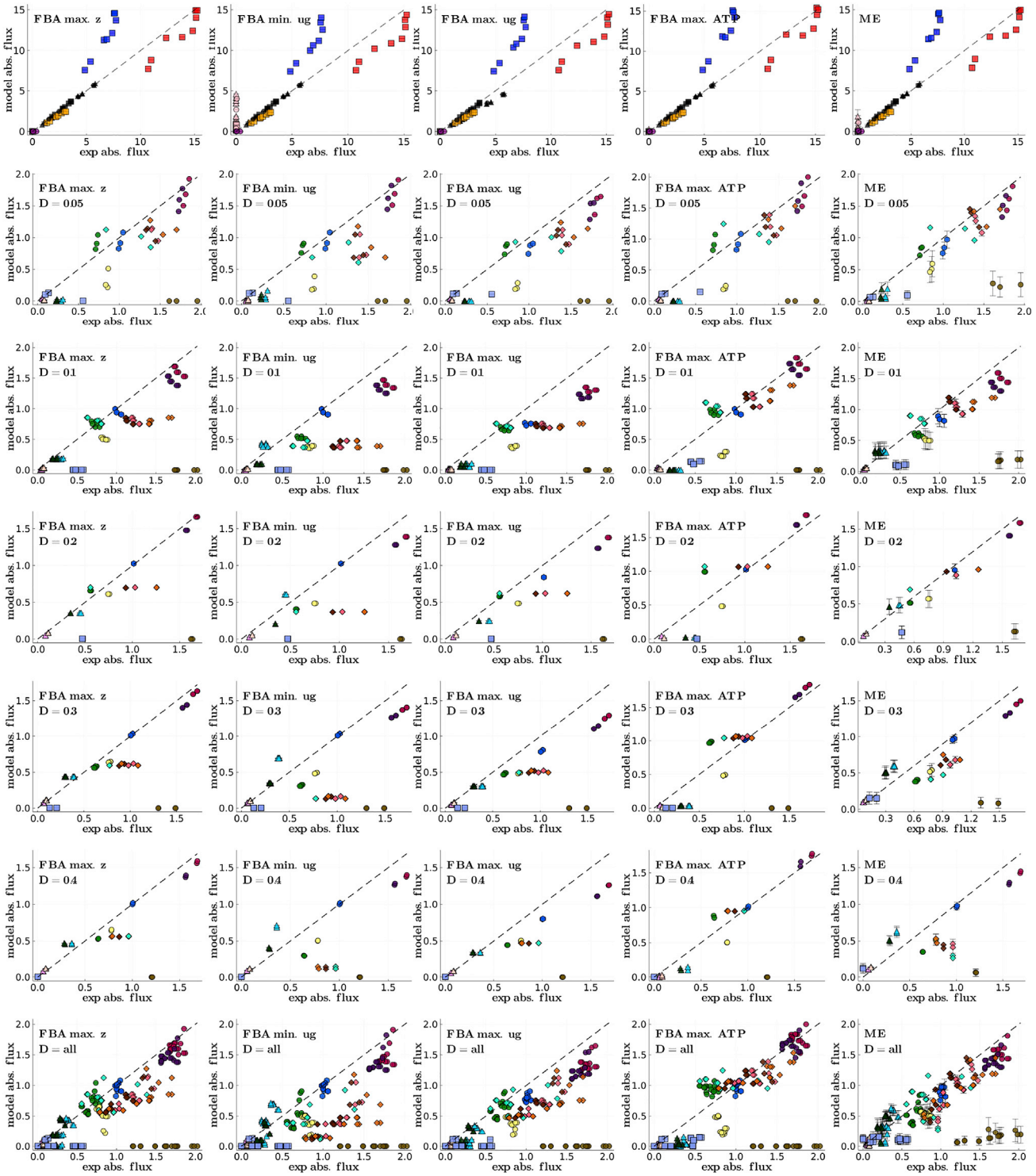


Figure 9. Inference of observable flux configurations for *E. coli* chemostat cultures

Experimental (x axis) *v.s.* model predicted (y axis) absolute relative fluxes, for a set of FBA formulations and ME. All fluxes are normalized by the experimental glucose uptake. The first row shows exchange fluxes reported in[73,74,75] The rest of the rows show some inner fluxes reported in.[74] Each row corresponds to a different dilution rate (h^{-1}). The last row includes all internal flux correlations. Different subsystems are signalized by the shape of the marker, meanwhile different colors denote individual reactions. The legend is as follows: acetate exchange (gray), CO2 exchange (blue square), glyoxylate cycle (pink), pyruvate kinase *PYK* (dark green), Krebs cycle (diamond shaped), glycolysis (circle shaped) and pentose phosphate pathway (triangle shaped).

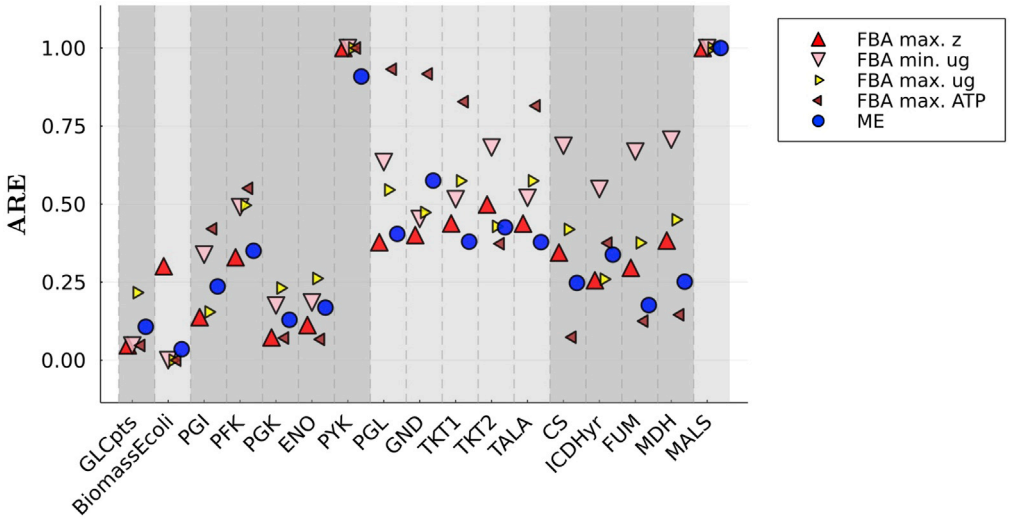


Figure 10. Inference error of observable flux configurations for *E. coli* chemostat cultures

Inference quality measured by the Absolute Relative Error (*ARE*) of internal fluxes, $\overline{u_g}$ and $\overline{z}$ for all experimental conditions. The reactions are grouped by subsystems, from left to right: glucose transport, biomass, glycolysis, pentose phosphate pathway, Krebs cycle, and glyoxylate cycle. Reaction acronyms are taken from.[55] $ARE_i = K^{-1}\sum_k^K \left|1 - v_{ik}^{model}/v^{exp_{ik}}\right|$, is the error of flux i across K samples ($K > 10$). Large ARE values are cut to 1.

Study of additional biases

As discussed in Section 3.2, an advantage of ME over FBA is that it uses more effectively all the information contained in the constraints, which allows inferring properties of the culture metabolic state other than the observable flux configuration. But, it also makes ME more sensible to the introduction of unnoticed biases. To gain a deeper insight, we replicate the same analysis over the *E. coli* experimental data using the ME formulation described at.[36] The new model (called in this section ME^1) uses a single beta parameter (the super index accounts for the number of non-zero β parameters). The only difference with our model (called ME^2 in this section) is that the nutrient limiting constraint is simplified from $\overline{u_g} \leq c_g D/X$ to $u_g \leq c_g D/X$.

Figure 11 presents the comparison between both formulations. In the left column of the figure, we show the marginal distributions (for one experimental condition) for three selected fluxes. The first marginal (top-left) is the one corresponding to the uptake of glucose. As mentioned before, the codification of the chemostat constraint of this flux is the only difference between the two formulations. As can be seen, both marginals differ substantially. This subtle difference, to consider that the knowledge of an observable restricts $\mathbb{V}$, is sufficient to produce a major difference in the solution of ME (*e.g.* the heterogeneity of the culture). Additionally, because the network imposes a structural constraint that is reflected in a correlation between the fluxes, this discrepancy is propagated to others. This can be noticed in the marginal of the biomass reaction (bottom-left). In both cases, the reduction of $\mathbb{V}$ in ME^1 formulation resulted in distributions with smaller degeneracy. A large-scale study of such an effect is shown in the right column of the same figure. There, we show a comparison between both formulation averages (top-right) and variances (bottom-right) for all the fluxes in all experimental conditions. As can be seen, the averages are not particularly affected, but the variances (which are shown in a log scale) are consistently smaller for the ME^1 formulation.

Although ME^2 is more rigorous, this formulation might be not totally free from biases associated with the exchanges. For instance, Equation 7 shows that a metabolite not present in the feed medium should have a negative or zero average exchange rate, which means that the culture can only potentially produce it, not consume it. Even though this is an observable constraint, we made the assumption that ($c_i = 0 \Rightarrow u_i \leq 0$). As mentioned before, the correct formal methodology for encoding such observable constraints is by moving its corresponding components in the $\boldsymbol{\beta}$ vector so that the selected $\mathcal{P}_{\mathcal{ME}}$ does satisfy with the restrictions and $\mathbb{V}$ stays properly unaffected. Because this needs to be done for all the metabolites that the network might produce, it would increase the number of free non-zero $\boldsymbol{\beta}$

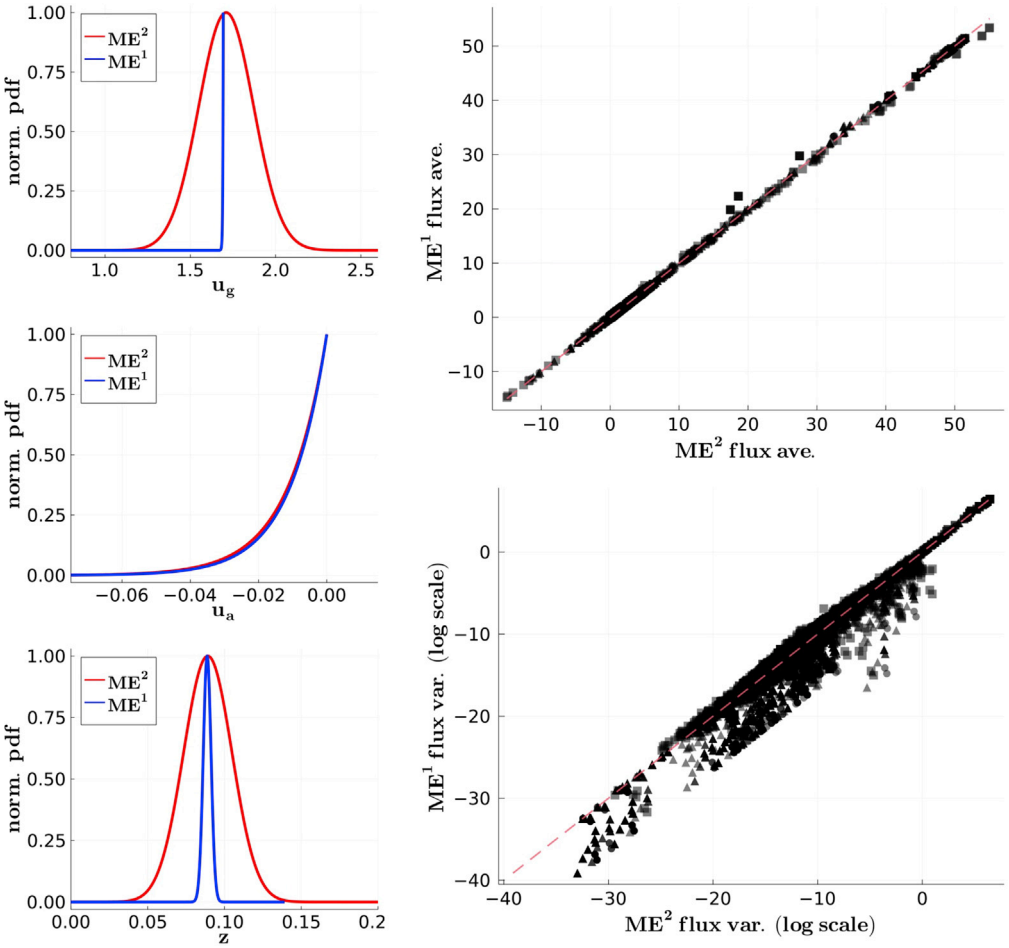


Figure 11. Study of the effect of the different ME formulations on the inferred flux distributions

The left column shows selected marginals of the glucose uptake flux, u_g, the acetate production, u_a, and the biomass production rate, z, for both ME formulations in an experiment (rep. 4) from . *Nanchen*[74] In the right column, it is shown the correlations of all flux averages (top) and all flux variances (log scale) between both formulations for all data sources.

components that need to be tuned for inferring the $\mathcal{P}_{\mathcal{ME}}$ distribution, which would make its computation more challenging.

An example of such phenomena can be appreciated for the acetate exchange rate, whose marginal is shown in the left column of Figure 11 (middle panel). The acetate exchange marginal is abruptly cut at zero. This provokes that, in case of any degeneration, its average gets a value greater than zero. This depends on the assumption that cells can not consume acetate, which is not directly derived from any constraint imposed by the chemostat. As discussed before, ME predicts a wrong non-zero acetate production rate (see Figure 10 top-right panel). This might be another possible cause of discrepancy with the experiments.

Unlimited culture dynamic

In this section, we study the dynamic of the chemostat when Equations 13 and 14 are decoupled. This can be done by running simulations in a nutrient-unlimited environment ($c_g, c_o = +\infty$). In this case, no feedback is produced between the observables and the only significant environmental parameter will be the dilution rate. Figure 12 shows a heatmap with the X value at the end of such simulations as a function of D and ε. The simulations were stopped if either a non-trivial steady state condition was hit ($dX/dt \to 0$ and $X > 0$), the culture X grew forever ($dX/dt > 0$ and $X > 10^6$) or the culture died ($dX/dt < 0$ and $X < 10^{-6}$). From the figure, we can observe that the nutrient-unlimited dynamic typically leads to either an unbounded

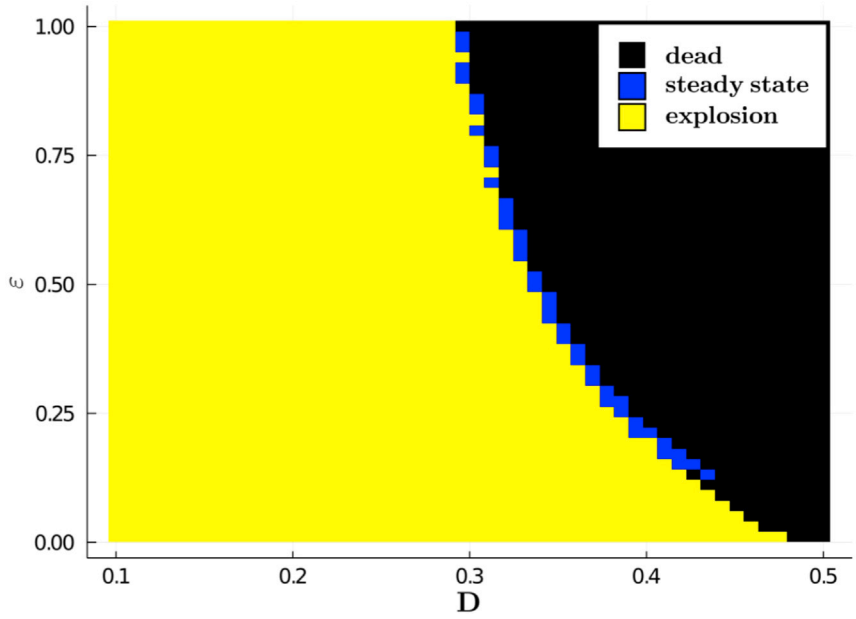


Figure 12. Study of the dynamics of a nutrient-unlimited chemostat

Heat map that represents the evolution of X in a nutrient-unlimited simulation ($c_g = +\infty$) as a function of D and ε.

growth or a wash-out. Non-trivial steady states are only possible at the interface between the two regions. But, this interface represents an unstable regime. Small perturbations on either D or ε will make such steady state unfeasible. It is this tendency to increment X which leads the culture to deplete all the nutrients once the limited condition is reestablished, which implies that the culture nutrient uptake at a steady state will be close to the input rate ($\overline{u_g} \approx c_g D/X$).

DISCUSSION

In the context of metabolic models, the data necessary for formulating environmental constraints are commonly available (*e.g.* medium composition, cellular concentration, culture observables, and so forth). However, although progress in this area has been substantial, metabolic models generally lack the information needed for a complete formulation of internal constraints (*e.g.* kinetic parameters of enzymes, influence of the regulatory network, completeness of the stoichiometric network, and so forth).[23] As a consequence, the relative relevance of both types of constraints in a particular experimental condition determines the effectiveness of the inference method.

For example, in a batch culture grown in a rich medium, the environmental constraints are not too strong. The cells are growing in a context without limiting nutrients, and therefore they can potentially display a wide range of phenotypic behaviors ($\overline{\mathbb{V}} \equiv \mathbb{V}$). In this case, the unknown non-environmental constraints (*e.g.*regulatory or kinetic) are defining the behavior of the culture. If this is the case, an ME formulation that lacks such decisive constraints must lead to a solution that poorly describes the observed phenotypic state of the culture. Traditionally, this issue has been addressed by introducing further constraints based on the available experimental data (*e.g.* fixing a fraction of the fluxes[45,69,37]) or in the case of FBA by defining an objective function (or a stack of them).[59] In the latter case, the objective function tries to encode these (unknown) non-environmental constraints that are driving the system to a specific state inside the very degenerate feasible solution space. On the other hand, in nutrient-limited chemostat culture at steady state, the known environmental constraints (as defined in our model) typically lead to a restricted observable space ($\overline{\mathbb{V}} \subset \mathbb{V}$). So, the unknown constraints have less room to significantly affect the observables. In this case, it is natural to assume that an ME formulation that includes such information should be a better descriptor of the culture metabolic state.

Such a scenario was modeled into the chemostat simulation presented in Section 4.2 where, by construction, $\overline{\mathbb{V}}$ was restricted only by known constraints. As a consequence, ME described accurately the steady

state $\bar{\mathbf{v}}$ of the system (see results in Figure 10). In this case, the chemostat constraints potentially determined $\overline{\mathbb{V}}$ in only two of three free dimensions, so $\overline{\mathbb{V}}$ was generally degenerated. Such a situation had catastrophic consequences for FBA. Due to the degeneracy, the system will eventually occupy all the feasible states due to the introduced stochasticity (which resembles the natural noisy phenomena characteristics of cellular cultures[77,78,79,80,81]). There is no reason to justify that the third dimension needs to be optimum. In other words, no extra assumptions (optimization) need to be formulated for describing the properties of the system. All the required information was already contained in the metabolic spaces formulations.

Furthermore, by increasing the stochasticity in the simulations (beyond minimal space degeneration), we study the more general case where $\overline{\mathbb{V}}$ is degenerate even in the ($\bar{z}$, $\overline{u_g}$) subspace. Such stochasticity affects two important features of the steady state: I) its feasibility, and II) how large can be X (see results in Figure 5). The study revealed a link between the heterogeneity and the size of $\overline{\mathbb{V}}$ at steady state. As mentioned, the optimum X can only be reached by a culture displaying a maximum $\bar{z}/\overline{u_g}$ yield (for a glucose-limited case). The stochasticity prevents that from happening by forcing the culture to allocate biomass at suboptimal states. We stress that this result might suggest a relation between an accessible experimental magnitude (X), and a culture property that is difficult to evaluate (heterogeneity).[19] Finally, this model provides a mechanistic explanation on how the steady state constraints become so relevant, supporting the results of ref. [45]. The simulations demonstrated that the system, when feasible, displays a typical tendency to use the full carrying capacity of the medium (accumulating X until the limiting nutrient is depleted). Although we provided a very simple dynamical model to test this idea, the results can be generalized to more realistic scenarios. For example, by introducing a biomass rate maximization constraint (a popular regulatory constraint for bacteria[64]), the dynamics of the system will change, but the space of observables is still determined by the same environmental constraints ($\bar{z} = D$ and $\overline{u_g} \leq c_g D/X$). This means that unknown and complex constraints could be driving the dynamic phase of the culture, but at the steady state, its consequences over $\overline{\mathbb{V}}$ are ultimately summarized in the value of X. This is because, as mentioned before, $\mathbb{V}$, D and $\mathbf{c}$ are usually considered constant during the culture, and X is the only variable dependent on the dynamics that influence the chemostat constraints. A similar picture was discussed in,[17] where a related model is studied. There, the authors established that the ratio between cell concentration and dilution rate is the control parameter fixing the steady state properties of the chemostat. The conclusion was also extended to more complex scenarios such as multi-stable regimes and perfusion.

We may extrapolate some insights gained from studying the simple model to the interpretation of the results obtained using a genome-scale metabolic network of *E. coli* and the experimental data. The first noticeable result was that the location of the experiments appeared close to the maximal theoretical X, as defined by the metabolic network and the culture conditions (see Figure 7). In the simulation, the culture's heterogeneity was inversely proportional to X (see Figure 5). So, its maximization in the experiments suggests that the *E. coli* cultures are close to the minimal possible heterogeneity (in terms of $\bar{z}/\overline{u_g}$ yield) as a result of the restrictions imposed by the chemostat constraints. This also means that the experimentally feasible $\overline{\mathbb{V}}$ is minimum, *i.e.* it is the more informative state yield by the environmental constraints.[65] Is this enough information for describing the culture observables? As mentioned before, a positive answer would imply that ME must be able to recover such observables. The correlations result in Figure 10, although not conclusive due to the noted limitations of our model (see Section 5.1), point into this direction. *We might be in the desirable situation where the most significant restrictions are the known environmental constraints.* An extra detail related to the experimental conditions, which might support such rationale, is that the studied cultures were run at small dilution rates ($D < 0.5\ h^{-1}$). This locates the cultures in a regime of slow growth rate (wild *E. coli* can growth at $> 2.2\ h^{-1}$ [58]) and below the acetate switch.[52] This is relevant because the lower the growth rate, the less pressure is expected over the cellular resources, and thus, internal regulations such as enzyme cost constraint (see Equation 4) might lose significance[48,52,.82]

Notice also that our framework makes a distinction between constraints at the single-cell level and constraints at the population level. In practice, although our works focus the attention on the difficulties of interpreting the measurements made at the population level, the framework used here can be used for single-cell measurements. In this case, the population level constraints ($\overline{\mathbb{V}}$) are not present, but still one must interpret the heterogeneity in the single-cell metabolism, as a temporal property, provided that the system is ergodic. It is certainly a direction of future interest, also in the context of increasing interest in the

distinction of bulk and single-cell measurements[83,84] and promising new single-cell experimental techniques such as Nanoscale secondary ion mass spectrometry (NanoSIMS).[85]

Limitation of the study

In previous sections, we highlighted the relevance of codifying the different constraints into their corresponding spaces. Although we properly handle the constraints related to the chemostat dynamics, we introduced several simplifications to the definition of $\mathbb{V}$. We take information that is actually based on macroscopic measurements to be representative of each cell. This hides a culture homogeneity assumption. The most significant is related to the definition of the biomass equation, which is determined by experimentally measuring the average cellular composition.[64] In principle, this constraint (dashed line in Figure 2) should be considered to be affecting only $\overline{\mathbb{V}}$. But we made the widely adopted simplification[17,78,35,45,37] of taking it as a hard constraint over $\mathbb{V}$. In the current formulation, ME is only capable to encode average constraints over the reaction bounds (like Equations 7 and 8), not balance constraints such as the biomass equation. In the particular case of a limiting-nutrient chemostat culture, this might be a fundamental source of bias, given the tendency of the culture to maximize $\overline{z}/\overline{u_g}$ (see results in Figure 7 Panel A). At this point, the degeneracy of $\overline{\mathbb{V}}$ is minimal, and so, the variability lost by the biomass simplification might be significant. We leave this question open for future studies. A similar situation occurs in the formulation of the cost constraints in Equation 4. In particular, the definition of each cost weight a_j^+ and a_j^- depends on the total observable protein mass fraction of the cells.[48] This is another balance constraint that can not be encoded into $\overline{\mathbb{V}}$ using the current ME formulation.

Regarding the evaluated experimental data, although the available fluxes are representative of important metabolic pathways, the system is still heavily under-determined. We only have access to approximately 10^1 experimental fluxes in a network with more than 10^3 reactions (rank 10^2). Additionally, our methods depend on how well the used network is a good representation of the metabolism of *E. coli*. Although we use all available experimental data for a better contextualization of the network, many of the parameters used were the generic defaults shipped with the model. The presence of relevant outliers in the correlations actually suggest that there is information missing or bias. All that prevents us from having a conclusive argument on the inference quality for both ME and FBA and the comparison between them. For another conceptual comparison between optimization-based and ME methods, we refer to also to.[69]

Finally, it is important to notice that there are several artifacts that might affect a reconstructed network, even if it was manually curated. One affecting both FBA and ME are thermodynamically infeasible cycles (TICs)[86,69]. Those are sets of reactions that can carry arbitrary flux without breaking the common stoichiometric/boundary constraints. They will increase the overall entropy of an ME distribution, and more significant for our work, the marginal moments of the involved reactions. Such a phenomenon, however, is not a limitation of our approach, but a limitation of the constraints we are imposing (the data). When TIC's are important, it just means that we are lacking relevant constraints. In practice, our methodology is compatible with any further lineal constraints that might be added for addressing TICs. In our model, we used enzymatic cost constraints that automatically penalize these cycles (see (4)). Another well-studied limitation of FBA models is their failure to properly explain an overflow metabolism phenotype observed in bacteria (and other organisms) at higher growth-rates, where rapid carbon intake from glucose consumption is diverted from biomass and CO2 into the production of lactate or ethanol.[27,48,82,49] It has been argued that accounting for enzymatic costs necessary to sustain metabolic fluxes can explain the observed switch,[82,49] which we include in our model in Equation 4.

Conclusions

To conclude, in this work, we exploit the Maximum Entropy Principle to provide a probabilistic description of the culture metabolism that can be used to infer the set of observable average fluxes, as well as a description of the heterogeneity. We showed, exploiting a simple mechanistic model, that at steady state and in limiting nutrient conditions, two external parameters are enough to capture the same information as the boundary flux observables. These parameters correspond to two important constraints of the chemostat environment: one derived from the biomass mass balance, and the other from the limiting nutrient mass balance. The explanation is consistent with data-driven results found in the literature for the studied conditions. The technique was applied to a dynamical model of the chemostat, where the external conditions of the culture were linked with the internal cellular metabolism. Also, it was applied to a genome-scale

metabolic network and tested against experimental data from *E. coli* cultures. We compare the results of our techniques with different variants of FBA. Although the quality of the data makes it difficult to define which techniques provide a better inference of the fluxes, the results suggest that ME is more robust than the different variants of FBA, which support the hypothesis that the metabolism is not necessarily well described by an optimum state. Finally, by relying on a readily available set of minimal experimental quantities describing the system (feed media composition, dilution rate, and steady state cell density), we think that our results may enlarge the space of applications of *ME* based metabolic modeling.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHODS DETAILS
  - *E. coli* continuous cultivation experimental data
  - *E. coli* metabolic network
  - Chemostat dynamic simulation
  - ME algorithm
  - Expectation Propagation

ACKNOWLEDGMENTS

We are indebted to A. de Martino for useful discussions and with A. Muntoni for providing help with the implementation of the EP algorithm. The work was supported by the Horizon 2020 Marie Skłodowska-Curie Action-Research and Innovation Staff Exchange (MSCA-RISE) 2016 grant agreement 734439 (INFERNET: New algorithms for inference and optimization from large-scale biological data). It was also partially funded by the CITMA Project of the Republic of Cuba, PNCB-Statistical Mechanics of Metabolic Interactions-PN223LH010-015. The authors declare no conflict of interest

AUTHOR CONTRIBUTIONS

J.A.P., R.M., and J.F. contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 11, 2022
Revised: October 9, 2022
Accepted: October 23, 2022
Published: December 22, 2022

REFERENCES

1. Weng, Z., Jin, J., Shao, C., and Li, H. (2020). Reduction of charge variants by cho cell culture process optimization. Cytotechnology *72*, 259–269.

2. Xu, S., Gavin, J., Jiang, R., and Chen, H. (2017). Bioreactor productivity and media cost comparison for different intensified cell culture processes. Biotechnol. Prog. *33*, 867–878.

3. Ozturk, S.S. (1996). Engineering challenges in high density cell culture systems. Cytotechnology *22*, 3–16. https://doi.org/10.1007/BF00353919.

4. Monod, J. (1949). The growth of bacterial cultures. Annu. Rev. Microbiol. *3*, 371–394. https://doi.org/10.1146/annurev.mi.03.100149.002103.

5. Novick, A., and Szilard, L. (1950). Description of the chemostat. Science *112*, 715–716. https://doi.org/10.1126/science.112.2920.715.

6. Werner, R.G., Walz, F., Noé, W., and Konrad, A. (1992). Safety and economic aspects of continuous mammalian cell culture. J. Biotechnol. *22*, 51–68. https://doi.org/10.1016/0168-1656(92)90132-s.

7. Griffiths, J.B. (1992). Animal cell culture processes - batch or continuous? J. Biotechnol. *22*, 21–30. https://doi.org/10.1016/0168-1656(92)90129-W.

8. Kadouri, A., and Spier, R.E. (1997). Some myths and messages concerning the batch and continuous culture of animal cells. Cytotechnology *24*, 89–98. https://doi.org/10.1023/A:1007932614011.

9. Werner, R.G., and Noe, W. (1998). Letter to the editor. Cytotechnology *26*, 81–82. https://doi.org/10.1023/A:1007985828899.

10. Croughan, M.S., Konstantinov, K.B., and Cooney, C. (2015). The future of industrial bioprocessing: batch or continuous? Biotechnol. Bioeng. *112*, 648–651. https://doi.org/10.1002/bit.25529.

11. Mulukutla, B.C., Yongky, A., Grimm, S., Daoutidis, P., and Hu, W.-S. (2015). Multiplicity of steady states in glycolysis and shift of metabolic state in cultured mammalian cells. PLoS One *10*, e0121561. https://doi.org/10.1371/journal.pone.0121561.

12. Europa, A.F., Gambhir, A., Fu, P.-C., and Hu, W.-S. (2000). Multiple steady states with distinct cellular metabolism in continuous culture of mammalian cells. Biotechnol. Bioeng. *67*, 25–34. https://doi.org/10.1002/(SICI)1097-0290(20000105)67:1<25::AID-BIT4>3.0.CO;2-K.

13. Altamirano, C., Illanes, A., Casablancas, A., Gamez, X., Cairó, J.J., and Godia, C. (2001). Analysis of cho cells metabolic redistribution in a glutamate-based defined medium in continuous culture. Biotechnol. Prog. *17*, 1032–1041.

14. Hayter, P.M., Curling, E.M., Baines, A.J., Jenkins, N., Salmon, I., Strange, P.G., Tong, J.M., and Bull, A.T. (1992). Glucose-limited chemostat culture of Chinese hamster ovary cells producing recombinant human interferon-γ. Biotechnol. Bioeng. *39*, 327–335. https://doi.org/10.1002/bit.260390311.

15. Gambhir, A., Korke, R., Lee, J., Fu, P.-C., Europa, A., and Hu, W.-S. (2003). Analysis of cellular metabolism of hybridoma cells at distinct physiological states. J. Biosci. Bioeng. *95*, 317–327. https://doi.org/10.1016/S1389-1723(03)80062-2.

16. Follstad, B.D., Balcarcel, R.R., Stephanopoulos, G., and Wang, D.I.C. (1999). Metabolic flux analysis of hybridoma continuous culture steady state multiplicity. Biotechnol. Bioeng. *63*, 675–683. https://doi.org/10.1002/(SICI)1097-0290(19990620)63:6<675::AID-BIT5>3.0.CO;2-R.

17. Fernandez-de-Cossio-Diaz, J., Leon, K., and Mulet, R. (2017). Characterizing steady states of genome-scale metabolic networks in continuous cell cultures. PLoS Comput. Biol. *13*, e10058355–e11005922. https://doi.org/10.1371/journal.pcbi.1005835.

18. Fernandes, R.L., Nierychlo, M., Lundin, L., Pedersen, A.E., Puentes Tellez, P.E., Dutta, A., Carlquist, M., Bolic, A., Schäpper, D., Brunetti, A.C., et al. (2011). Experimental methods and modeling techniques for description of cell population heterogeneity. Biotechnol. Adv. *29*, 575–599. https://doi.org/10.1016/j.biotechadv.2011.03.007.

19. González-Cabaleiro, R., Mitchell, A.M., Smith, W., Wipat, A., and Ofiţeru, I.D. (2017). Heterogeneity in pure microbial systems: experimental measurements and modeling. Front. Microbiol. *8*, 1813. https://doi.org/10.3389/fmicb.2017.01813.

20. Pérez-Fernández, B.A., Fernandez-de Cossio-Diaz, J., Boggiano, T., León, K., and Mulet, R. (2021). In-silico media optimization for continuous cultures using genome scale metabolic networks: the case of CHO-K1. Biotechnol. Bioeng. *118*, 1884–1897.

21. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. *42*, 199–205. https://doi.org/10.1093/nar/gkt1076.

22. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., et al. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. *44*, D471–D480. https://doi.org/10.1093/nar/gkv1164.

23. Palsson, B.Ø. (2015). System Biology - Constraint-Based Reconstruction and Analysis (Cambridge University Press).

24. Ibarra, R.U., Edwards, J.S., and Palsson, B.O. (2002). Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature *420*, 186–189. https://doi.org/10.1038/nature01149.

25. Palsson, B.Ø. (2006). System Biology - Propereties of Reconstructed Networks (Cambridge University Press).

26. 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. https://doi.org/10.1038/msb4100162.

27. Zeng, H., and Yang, A. (2019). Modelling overflow metabolism in Escherichia coli with flux balance analysis incorporating differential proteomic efficiencies of energy pathways. BMC Syst. Biol. *13*, 3–18. https://doi.org/10.1186/s12918-018-0677-4.

28. Robinson, J.L., Kocabaş, P., Wang, H., Cholley, P.-E., Cook, D., Nilsson, A., Anton, M., Ferreira, R., Domenzain, I., Billa, V., et al. (2020). An atlas of human metabolism. Sci. Signal. *13*, eaaz1482. https://doi.org/10.1126/scisignal.aaz1482.

29. Jaynes, E.T. (1957). Information theory and statistical mechanics. Phys. Rev. *106*, 620–630. https://doi.org/10.1103/PhysRev.106.620.

30. Mora, T., Walczak, A.M., Bialek, W., and Callan, C.G. (2010). Maximum entropy models for antibody diversity. Proc. Natl. Acad. Sci. USA *107*, 5405–5410. https://doi.org/10.1073/pnas.1001705107.

31. Roudi, Y., Nirenberg, S., and Latham, P.E. (2009). Pairwise maximum entropy models for studying large biological systems : when they can work and when they can 't. PLoS Comput. Biol. *5*, e1000380. https://doi.org/10.1371/journal.pcbi.1000380.

32. Tovo, A., Suweis, S., Formentin, M., Favretti, M., Volkov, I., Banavar, J.R., Azaele, S., and Maritan, A. (2017). Upscaling species richness and abundances in tropical forests. Sci. Adv. *3*, e1701438. https://doi.org/10.1126/sciadv.1701438.

33. De Martino, D., MC Andersson, A., Bergmiller, T., Guet, C.C., and Tkačik, G. (2018). Statistical mechanics for metabolic networks during steady state growth. Nat. Commun. *9*, 2988. https://doi.org/10.1038/s41467-018-05417-9.

34. De Martino, A., and De Martino, D. (2018). An introduction to the maximum entropy approach and its application to inference problems in biology. Heliyon *4*, e00596. https://doi.org/10.1016/j.heliyon.2018.e00596.

35. Martino, D.D., Capuani, F., and Martino, A.D. (2016). Growth against entropy in bacterial metabolism: the phenotypic trade-off behind empirical growth rate distributions in E. coli. Phys. Biol. *13*, 036005. https://doi.org/10.1088/1478-3975/13/3/036005.

36. Fernandez-de-Cossio-Diaz, J., and Mulet, R. (2019). Maximum entropy and population heterogeneity in continuous cell cultures. PLoS Comput. Biol. *15*, e1006823. https://doi.org/10.1371/journal.pcbi.1006823.

37. Muntoni, A.P., Braunstein, A., Pagnani, A., De Martino, D., and De Martino, A. (2022). Relationship between fitness and heterogeneity in exponentially growing microbial populations. Preprint at arXiv. https://doi.org/10.48550/arXiv.2104.02594.

38. Pressé, S., Ghosh, K., Lee, J., and Dill, K.A. (2013). Principles of maximum entropy and maximum caliber in statistical physics. Rev. Mod. Phys. *85*, 1115–1141.

39. White, E.P., Thibault, K.M., and Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. Ecology *93*, 1772–1778.

40. Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., and Walczak, A.M. (2012). Statistical mechanics for natural flocks of birds. Proc. Natl. Acad. Sci. USA *109*, 4786–4791. https://doi.org/10.1073/pnas.1118633109.

41. Schneidman, E., Berry, M.J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. Nature *440*, 1007–1012.

42. Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal *27*, 379–423.

43. Qian, H., and Beard, D.A. (2005). Thermodynamics of stoichiometric biochemical networks in living systems far from equilibrium. Biophys. Chem. *114*, 213–220.

44. Dukovski, I., Bajić, D., Chacón, J.M., Quintin, M., Vila, J.C.C., Sulheim, S., Pacheco, A.R., Bernstein, D.B., Riehl, W.J., Korolev, K.S., et al. (2021). A metabolic modeling platform for the computation of microbial ecosystems in time and space (COMETS). Nat. Protoc. *16*, 5030–5082. https://doi.org/10.1038/s41596-021-00593-3.

45. De Martino, D., and De Martino, A. (2017). Constraint-based inverse modeling of metabolic networks: a proof of concept.

Preprint at arXiv. https://doi.org/10.48550/arXiv.1704.08087.

46. Gu, C., Kim, G.B., Kim, W.J., Kim, H.U., and Lee, S.Y. (2019). Current status and applications of genome-scale metabolic models. Genome Biol. *20*, 121. https://doi.org/10.1186/s13059-019-1730-3.

47. Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248. https://doi.org/10.1038/nbt.1614.

48. Beg, Q.K., Vazquez, A., Ernst, J., de Menezes, M.A., Bar-Joseph, Z., Barabási, A.L., and Oltvai, Z.N. (2007). Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proc. Natl. Acad. Sci. USA *104*, 12663–12668. https://doi.org/10.1073/pnas.0609845104.

49. Fernandez-de Cossio-Diaz, J., and Vazquez, A. (2017). Limits of aerobic metabolism in cancer cells. Sci. Rep. *7*, 13488.

50. Fernandez-de Cossio-Diaz, J., and Vazquez, A. (2018). A physical model of cell metabolism. Sci. Rep. *8*, 1–13.

51. Scott, M., Gunderson, C.W., Mateescu, E.M., Zhang, Z., and Hwa, T. (2010). Interdependence of cell growth and gene expression: origins and consequences. Science *330*, 1099–1102.

52. Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J.R., and Hwa, T. (2015). Overflow metabolism in Escherichia coli results from efficient proteome allocation. Nature *528*, 99–104. https://doi.org/10.1038/nature15765.

53. Boyd, S., Boyd, S.P., and Vandenberghe, L. (2004). Convex Optimization (Cambridge University Press).

54. Ben Yahia, B., Malphettes, L., and Heinzle, E. (2015). Macroscopic modeling of mammalian cell growth and metabolism. Appl. Microbiol. Biotechnol. *99*, 7009–7024. https://doi.org/10.1007/s00253-015-6743-6.

55. Reed, J.L., Vo, T.D., Schilling, C.H., and Palsson, B.O. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. *4*, R54. https://doi.org/10.1186/gb-2003-4-9-r54.

56. Smith, H.L., and Waltman, P. (1995). The theory of the chemostat: dynamics of microbial competition. Cambridge Studies in Mathematical Biology (Cambridge University Press). https://doi.org/10.1017/CBO9780511530043.

57. Bordbar, A., Monk, J.M., King, Z.A., and Palsson, B.O. (2014). Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. *15*, 107–120. https://doi.org/10.1038/nrg3643.

58. Varma, A., and Palsson, B.O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl. Environ. Microbiol. *60*, 3724–3731.

59. García Sánchez, C.E., and Torres Sáez, R.G. (2014). Comparison and analysis of objective functions in flux balance analysis. Biotechnol. Prog. *30*, 985–991. https://doi.org/10.1002/btpr.1949.

60. 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. https://doi.org/10.1038/msb.2010.47.

61. Morales, Y., Tortajada, M., Picó, J., Vehí, J., and Llaneras, F. (2014). Validation of an FBA model for Pichia pastoris in chemostat cultures. BMC Syst. Biol. *8*, 142. https://doi.org/10.1186/s12918-014-0142-y.

62. Lloyd, C.J., Ebrahim, A., Yang, L., King, Z.A., Catoiu, E., O'Brien, E.J., Liu, J.K., and Palsson, B.O. (2018). COBRAme: a computational framework for genome-scale models of metabolism and gene expression. PLoS Comput. Biol. *14*, e1006302. https://doi.org/10.1371/journal.pcbi.1006302.

63. Herrmann, H.A., Dyson, B.C., Vass, L., Johnson, G.N., and Schwartz, J.-M. (2019). Flux sampling is a powerful tool to study metabolism under changing environmental conditions. NPJ Syst. Biol. Appl. *5*, 32–38. https://doi.org/10.1038/s41540-019-0109-0.

64. Feist, A.M., and Palsson, B.O. (2010). The biomass objective function. Curr. Opin. Microbiol. *13*, 344–349. https://doi.org/10.1016/j.mib.2010.03.003.

65. Jaynes, E.T. (2003). Probability Theory: The Logic of Science (Cambridge University Press).

66. Tourigny, D.S. (2020). Dynamic metabolic resource allocation based on the maximum entropy principle. J. Math. Biol. *80*, 2395–2430. https://doi.org/10.1007/s00285-020-01499-6.

67. Fernandez-de Cossio-Diaz, J., and Mulet, R. (2016). Fast inference of ill-posed problems within a convex space. J. Stat. Mech. *2016*, 073207.

68. Minka, T.P. (2013). Expectation propagation for approximate bayesian inference. Preprint at arXiv. https://doi.org/10.48550/arXiv.1301.2294.

69. Rivas-Astroza, M., and Conejeros, R. (2020). Metabolic flux configuration determination using information entropy. PLoS One *15*, e0243067. https://doi.org/10.1371/journal.pone.0243067.

70. Segrè, D., Vitkup, D., and Church, G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. USA *99*, 15112–15117. https://doi.org/10.1073/pnas.232349399.

71. Senn, H., Lendenmann, U., Snozzi, M., Hamer, G., and Egli, T. (1994). The growth of Escherichia coli in glucose-limited chemostat cultures: a re-examination of the kinetics. Biochim. Biophys. Acta *1201*, 424–436. https://doi.org/10.1016/0304-4165(94)90072-8.

72. Molenaar, D., van Berlo, R., de Ridder, D., and Teusink, B. (2009). Shifts in growth strategies reflect tradeoffs in cellular economics. Mol. Syst. Biol. *5*, 323. https://doi.org/10.1038/msb.2009.82.

73. Kayser, A., Weber, J., Hecht, V., and Rinas, U. (2005). Metabolic flux analysis of Escherichia coli in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state. Microbiology *151*, 693–706. https://doi.org/10.1099/mic.0.27481-0.

74. Nanchen, A., Schicker, A., and Sauer, U. (2006). Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli. Appl. Environ. Microbiol. *72*, 1164–1172. https://doi.org/10.1128/AEM.72.2.1164-1172.2006.

75. Folsom, J.P., Parker, A.E., and Carlson, R.P. (2014). Physiological and proteomic analysis of Escherichia coli iron-limited chemostat growth. J. Bacteriol. *196*, 2748–2761. https://doi.org/10.1128/JB.01606-14.

76. Braunstein, A., Muntoni, A.P., and Pagnani, A. (2017). An analytic approximation of the feasible space of metabolic networks. Nat. Commun. *8*, 14915. https://doi.org/10.1038/ncomms14915.

77. Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186. https://doi.org/10.1126/science.1070919.

78. Fernandez-de-Cossio-Diaz, J., Mulet, R., and Vazquez, A. (2019). Cell population heterogeneity driven by stochastic partition and growth optimality. Sci. Rep. *9*, 9406–9407. https://doi.org/10.1038/s41598-019-45882-w.

79. Huh, D., and Paulsson, J. (2011). Random partitioning of molecules at cell division. Proc. Natl. Acad. Sci. USA *108*, 15004–15009. https://doi.org/10.1073/pnas.1013171108.

80. Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D.I.S., Zairis, S., Abate, F., Liu, Z., Elliott, O., Shin, Y.J., et al. (2016). Clonal evolution of glioblastoma under therapy. Nat. Genet. *48*, 768–776. https://doi.org/10.1038/ng.3590.

81. Tzur, A., Kafri, R., LeBleu, V.S., Lahav, G., and Kirschner, M.W. (2009). Cell growth and size homeostasis in proliferating animal cells. Science *325*, 167–171. https://doi.org/10.1126/science.1174294.

82. Vazquez, A., and Oltvai, Z.N. (2016). Macromolecular crowding explains overflow metabolism in cells. Sci. Rep. *6*, 31007. https://doi.org/10.1038/srep31007.

83. Soifer, I., Robert, L., and Amir, A. (2016). Single-cell analysis of growth in budding yeast and bacteria reveals a common

size regulation strategy. Curr. Biol. *26*, 356–361.

84. Kennard, A.S., Osella, M., Javer, A., Grilli, J., Nghe, P., Tans, S.J., Cicuta, P., and Cosentino Lagomarsino, M. (2016). Individuality and universality in the growth-division laws of single e. coli cells. Phys. Rev. E *93*, 012408.

85. Gao, D., Huang, X., and Tao, Y. (2016). A critical review of nanosims in analysis of microbial metabolic activities at single-cell level. Crit. Rev. Biotechnol. *36*, 884–890.

86. Schroeder, W.L., and Saha, R. (2020). Optfill: a tool for infeasible cycle-free gapfilling of stoichiometric metabolic models. iScience *23*, 100783. https://doi.org/10.1016/j.isci.2019.100783.

87. Bezanson, J., Edelman, A., Karpinski, S., and Shah, V.B. (2017). Julia: a fresh approach to numerical computing. SIAM Rev. Soc. Ind. Appl. Math. *59*, 65–98. https://doi.org/10.1137/141000671.

88. Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., and Palsson, B.Ø. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. *3*, 121. https://doi.org/10.1038/msb4100155.

STAR★METHODS

KEY RESOURCES TABLE

REGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
iJR904 GEM	Kayser et al.,[55]	http://bigg.ucsd.edu/static/models/iJR904.mat
Chemostat culture data	Nanchen et al.,; Folsom et al.,; Braunstein et al.,[73,74,75]	N/A
Software and Algorithms		
Chemostat_EColi.jl	this work	https://doi.org/10.5281/zenodo.7186870
julia (v1.7.3)	Bezanson et al.,[87]	https://julialang.org

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Roberto Mulet (mulet@fisica.uh.cu).

Materials availability

This study did not generate new unique data.

Data and code availability

FBA was implemented using traditional linear programming and ME distributions were approximated using an adaptation of the *ExpectationPropagation* algorithm reported in.[36] All necessary raw data and implementation code (written in Julia[87]) can be found at GitHub (https://github.com/josePereiro/Chemostat_EColi.jl) or Zenodo (https://doi.org/10.5281/zenodo.7186870). Follow the README.md instructions for a complete reproduction of the results of this work.

METHODS DETAILS

E. coli continuous cultivation experimental data

In order to test the predictive power of the different formulations, data of *E. coli* glucose-limited continuous cultures were taken from literature. Three different data sources were used *Kayser*[73], *Nanchen*[74]: and . *Folsom*[75]

E. coli metabolic network

The metabolism of *E. coli* was modeled using the metabolic network iJR904.[73] The network was downloaded from http://bigg.ucsd.edu/static/models/iJR904.mat but it is included as raw data in our code package (see Section 9.2.3). The metabolic network was appropriately contextualized using the available experiment-specific data in the source publications. If specific biomass composition data was available, the generic biomass equation in the metabolic network was also updated. Additional enzymatic constraints were added according to.[48] For defining a bounded $\mathbb{V}$ space, a few exchanges limits were added according to the largest values found at.[58] Those bounds were non-limiting for all studied experimental conditions (with respect to the observables) and played the numerical role of "infinity".

Chemostat dynamic simulation

The toy network used on the dynamics comprehends the follow reactions:

1. $glyc: (-1.0)Glc \rightarrow (2.0)Atp + (4.0)NADH + (1.0)AcCoa$
2. $ppp: (-1.0)Glc \rightarrow (2.0)NADH + (1.0)AcCoa$
3. $resp: (-2.0)NADH + (-1.0)Oxy \rightarrow (5.0)Atp$
4. $tac: (-1.0)AcCoa \rightarrow (1.0)Atp + (4.0)NADH$

5. *ferm:* $(-1.0)AcCoa \rightarrow (1.0)Atp + (1.0)Ac$
6. *ua:* $\leftarrow (1.0)Ac$
7. *ug:* $\rightarrow (1.0)Glc$
8. *uo:* $\rightarrow (1.0)Oxy$
9. *atpm:* $(-8.4)Atp \rightarrow$
10. *z:* $(-14.7)Glc + (-59.8)Atp \rightarrow$

The biomass requirement was defined as $(Y_{X/Glc})Glc + (GAM)Atp$ where $Y_{X/Glc} = 14.7(mmol \times g_{CDW}^{-1})$ is the biomass/glucose yield[58] and $GAM = 59.8\ (mmol \times g_{CDW}^{-1})$ is the growth associated ATP maintenance demand.[88] The non-growth associated ATP maintenance demand $(NGAM = 8.4\ (mmol \times g_{CDW}^{-1}))$[88] was modeled at the *atpm* reaction. All reactions are irreversible (as indicated by the single arrows) and open, except *atpm* which both bounds were fixed to one. The only limiting bound was at the glucose exchange (u_g) where , $ub_g = 20\ (mmol \times g_{CDW}^{-1} \times h^{-1})$[58] the rest was set to an arbitrary large number. The model has 3° of freedom that we choose to be z, u_g and u_o.

We performed a dynamic simulation of the chemostat following Equations 13 and 14. For computation, we discretize $\mathbb{V}$ (and so $\overline{\mathbb{V}}$) using a quantum δ so:

$$z \in \{0, 1\delta, 2\delta, \ldots\}$$
$$u_g \in \{0, 1\delta, 2\delta, \ldots\}$$
$$u_o \in \{0, 1\delta, 2\delta, \ldots\}$$

Refactoring Equation 13, so we include flux and time discretization we have:

$$\frac{\Delta X(z, u_g, u_o)}{\Delta t} = (1 - \varepsilon) z X(z, u_g, u_o) + \frac{\varepsilon}{|\mathbb{V}|_\delta} \sum_{z', u_g', u_o'}^{\mathbb{V}} z' X\left(z', u_g', u_o'\right) - D X(z, u_g, u_o) \quad \text{(Equation 16)}$$

where $|\mathbb{V}|_\delta \in \mathbb{N}$ is the total number of discrete regions contained at $\mathbb{V}$.

Making a similar analysis, we can determine how any external metabolite in the vessel evolves. Here the case for glucose:

$$\frac{\Delta s_g}{\Delta t} = -\sum_{z, u_g, u_o}^{\mathbb{V}} u_g X(z, u_g, u_o) + (c_g - s_g) D \quad \text{(Equation 17)}$$

Additionally, giving the values of $X(z, u_g, u_o)$ we can compute a probability mass function:

$$P(z, u_g, u_o) = X(z, u_g, u_o)/X$$

where $X = \sum_{z, u_g, u_o}^{\mathbb{V}} X(z, u_g, u_o)$.

As stated on Section 4.2, Equations 16 and 17 are not sufficiently connected so the simulation respects the implicit restriction of $s_g \geq 0$ (on the other hand, $s_o \geq 0$ is guaranty by construction given c_o sufficiently large, that is, non-limiting). In order to enforce the constraint, we introduce a transformation over P such:

$$P'(z, u_g, u_o) = \frac{P(z, u_g, u_o)(\gamma - u_g/U_g)}{\sum_{z', u_g', u_o'}^{\mathbb{V}} P\left(z', u_g', u_o'\right)\left(\gamma - u_g'/U_g\right)} \quad \text{(Equation 18)}$$

where U_g is u_g global maximum and $\gamma \in \mathbb{R}, \gamma > 1$ is a parameter that ensures $\sum_{z, u_g, u_o}^{\mathbb{V}} u_g P'(z, u_g, u_o) \approx c_g D/X$ at constant X. Such transformation is applied over P at every step of the simulation, where $s_g \approx 0$ and $\sum_{z, u_g, u_o}^{\mathbb{V}} u_g P(z, u_g, u_o) > c_g D/X$ (when the moment inequality constraint is about to be broken).

ME algorithm

The complete set of constraints defining $\mathbb{V}$ (see Equations 2, 3 and 4) can be expressed as:

$$\mathbb{S}\boldsymbol{v} = \boldsymbol{b}$$
$$\boldsymbol{lb} \leq \boldsymbol{v} \leq \boldsymbol{ub}$$

where $\boldsymbol{v} \in \mathbb{R}^N$, $\boldsymbol{lb} \in \mathbb{R}^N$, $\boldsymbol{ub} \in \mathbb{R}^N$, $\mathbb{S} \in \mathbb{R}^{M \times N}$ and $\boldsymbol{b} \in \mathbb{R}^M$.

A uniform distribution mapped over $\mathbb{V}$ can be written as[76]:

$$U(\boldsymbol{v}) \propto \delta(\mathbb{S}\boldsymbol{v} - \boldsymbol{b}) \prod_n^N \psi(v_n)$$

where $\delta(\mathbb{S}\boldsymbol{v} - \boldsymbol{b})$ is a Dirac's delta with a non-zero value when $\boldsymbol{v}$ solves the linear system (encoding the exact constraints), and $\psi(v_n)$ is an indicator which equals one if $lb_n \leq v_n \leq ub_n$ and zero otherwise (encoding the relaxed constraints).

The extra constraints which define $\overline{\mathbb{V}}$ (see Equations 7 and 8) can be written as:

$$\overline{\boldsymbol{v}} \leq \boldsymbol{h}$$

where $\overline{\boldsymbol{v}} \in \overline{\mathbb{V}}$ and $\boldsymbol{h} \in \mathbb{R}^N$ is a constant.

Given that $\mathbb{V} \subset \mathbb{R}^N$ is a convex space and the constraints over $\overline{\mathbb{V}}$ are linear, it can be proven that the distribution over $\mathbb{V}$ which maximizes the entropy belong to the exponential family.[29] Such exponential take the following form[36]:

$$\mathcal{P}(\boldsymbol{v}|\boldsymbol{\beta}) \propto e^{(\boldsymbol{\beta}^T \boldsymbol{v})} U(\boldsymbol{v}) \qquad \text{(Equation 19)}$$

where the vector $\boldsymbol{\beta} \in \mathbb{R}^N$ contains the selection coefficients of each reaction flux in the network so $\overline{\boldsymbol{v}} \leq \boldsymbol{h}$ (where $\overline{\boldsymbol{v}} = \int_{\mathbb{V}} \boldsymbol{v}\mathcal{P}(\boldsymbol{v}|\boldsymbol{\beta})d\boldsymbol{v}$) is meet, and the entropy is maximized. A remark worth making is that the functional form of (19) is generally intractable, so in this work, an approximated distribution obtained by Expectation Propagation (EP) is used instead.[76] Such procedure is explained in details in the next section, and it is transparent for the current analysis.

In this work, we are trying to enforce two constraints over the mean values, and so, the model has two free parameters (two non-zero components in the $\boldsymbol{\beta}$ vector on 19). One, (β_z), is used to restrict the average growth rate to equal the dilution rate $(\overline{z} = D)$, and the other, (β_{u_g}), is used to restrict the average uptake of glucose in accordance with the glucose supply rate $(\overline{u_g} \leq c_g D/X)$. Equation 19 can be rewritten to make this more explicit:

$$\mathcal{P}\left(\boldsymbol{v}\middle|\beta_z, \beta_{u_g}\right) \propto e^{(\beta_z z)} e^{\left(\beta_{u_g} u_g\right)} U(\boldsymbol{v})$$

Both moments $\overline{z}$ and $\overline{u_g}$ depend on the selected values of (β_z, β_{u_g}). If the corresponding constraint is fulfilled, the beta is called valid, β_z^v or $\beta_{u_g}^v$ respectively. Our goal is to find a pair of valid beta values, so the entropy is also maximal. In order to do that, we use the following algorithm:

Algorithm 1. ME Algorithm

1: procedure MAXENT2D
2: Init β_z and β_{u_g} at zero
3: Compute $\overline{z}$ and $\overline{u_g}$ using $\mathcal{P}(\boldsymbol{v}|0,0)$
4: **if** Constraints $(\overline{z} \approx D)$ and $(\overline{u_g} \lesssim c_g D/X)$ are fulfilled **then**
5: return β_z and β_{u_g}
6:
7: Update (grad. descend) β_z so constraint $(\overline{z} \approx D)$ is fulfilled
8: Compute $\overline{z}$ and $\overline{u_g}$ using $\mathcal{P}(\boldsymbol{v}|\beta_z,0)$
9: **if** Constraints $(\overline{z} \approx D)$ and $(\overline{u_g} \lesssim c_g D/X)$ are fulfilled **then**
10: return β_z and β_{u_g}
11:
12: **while** Constraints $(\overline{z} \approx D)$ and $(\overline{u_g} \approx c_g D/X)$ are NOT fulfilled **do**
13: Update (grad. descend) β_z so constraint $(\overline{z} \approx D)$ is fulfilled
14: Update (grad. descend) β_{u_g} so constraint $(\overline{u_g} \approx c_g D/X)$ is fulfilled
15: Compute $\overline{z}$ and $\overline{u_g}$ using $\mathcal{P}\left(\boldsymbol{v}\middle|\beta_z, \beta_{u_g}\right)$
16: return β_z and β_{u_g}

where each beta update was performed using a simple gradient descent until the given target was approximated.

As can be noticed, the entropy is not explicitly maximized in any of the gradient descents. But, the algorithm ensures that each returned pair $(\beta_z^v, \beta_{u_g}^v)$ does specify the distribution with the maximal entropy from all valid ones. Indeed, the above algorithm is nothing but standard maximization of entropy (following[78]), with the only peculiarity that we must also deal with inequality constraints on average values of the distribution, such as $\overline{v}_i \leq h_i$ for some flux i. The above algorithm is based on the idea that if this constraint is not satisfied automatically when one solves the ME problem without including it, then the optimal solution (when considering also this constraint), will satisfy instead the equality constraint $\overline{v}_i = h_i$. The proof states as follows:

Proof. Let $\mathcal{P}(\boldsymbol{v})$ be a distribution over fluxes $\boldsymbol{v} \in \mathbb{V}$. The entropy:

$$S[\mathcal{P}] = -\int_{\mathbb{V}} \ln(\mathcal{P}(\boldsymbol{v}))\mathcal{P}(\boldsymbol{v})d\boldsymbol{v}$$

is a concave functional of $\mathcal{P}$. Let $\mathbb{P}$ be any convex space of probability distributions. For instance, $\mathbb{P}$ can be the space of probability distributions with support $\mathbb{V}$. We are interested in finding the solution of a ME problem, of the form:

$$\begin{array}{l}\text{maximizearg}_{\mathcal{P}\in\mathbb{P}}\ S[\mathcal{P}]\\ \text{subject to:}\\ \overline{v}_i \leq h_i\end{array}$$

We denote by $\mathcal{P}^c$ and $\overline{\boldsymbol{v}}^c$ (c stand for constrained) the resulting distribution and its average vector. Additionally, we define $\mathcal{P}^g$ and $\overline{\boldsymbol{v}}^g$ to be the solution of the problem if we ignore the average constraint. Clearly, $S[\mathcal{P}^g] \geq S[\mathcal{P}^c]$ because $S[\mathcal{P}^c]$ has the additional inequality constraint (g stand for global maximum). If $\overline{v}_i^g \leq h_i$, both problems have the same solution, that is, $\mathcal{P}^c = \mathcal{P}^g$ and $\overline{\boldsymbol{v}}^c = \overline{\boldsymbol{v}}^g$ (which is the case on lines 3 and 6 on the algorithm 1). If $\overline{v}_i^g > h_i$, the two solutions necessarily differ. We show that in this case $\overline{v}_i^c = h_i$ necessarily. Suppose, to the contrary, that $\overline{v}_i^c < h_i$. This means that $\mathcal{P}^c$ is a local optimum of the entropy within $\mathbb{P}$. However, since the entropy is concave and $\mathbb{P}$ is a convex space, then $\mathcal{P}^c$ must also be a global optimum, that is, $\mathcal{P}^c = \mathcal{P}^g$. But then we have a contradiction, $h_i < \overline{v}_i^g = \overline{\boldsymbol{v}}_i^c < h_i$. Therefore, $\overline{v}_i^c < h_i$ is impossible, and we must have $\overline{v}_i^c = h_i$, as stated (which is the case for the line 16 on the algorithm 1).

Expectation Propagation

As stated in the last section, an ME distribution directly derived from the definition of $\mathbb{V}$ and $\overline{\mathbb{V}}$ has the form:

$$\mathcal{P}_\psi(\boldsymbol{v}) \propto e^{(\beta^T v)}\delta(\mathbb{S}\boldsymbol{v} - \boldsymbol{b})\prod_n^N \psi(v_n)$$

Through Gaussian elimination, we can transform the matrix $\mathbb{S}$ to a row echelon form:

$$\mathbb{S} \equiv [\mathbb{I}|\mathbb{G}]$$

where $\mathbb{I} \in \mathbb{R}^{M\times M}$ is an identity matrix and $\mathbb{G} \in \mathbb{R}^{M\times(N-M)}$.

The structure of the linear constraint induced by the row echelon representation suggests splitting the $\boldsymbol{v}$ variable vector into two sets of variables: the first M variables (dependent) and a second set of N-M variables (independent). To do so, we define:

$$\boldsymbol{v} \equiv \left(\boldsymbol{v}^{(d)}, \boldsymbol{v}^{(i)}\right)$$

where, as we said, $\boldsymbol{v}^{(d)} \in \mathbb{R}^M$ and $\boldsymbol{v}^{(i)} \in \mathbb{R}^{N-M}$ and

$$\boldsymbol{v}^{(d)} = \boldsymbol{b}' - \mathbb{G}\boldsymbol{v}^{(i)}$$

where $\boldsymbol{b}' \in \mathbb{R}^M$ is the transformed (after G. elimination) version of $\boldsymbol{b}$.

We rewrite the probability density function in terms of the new variable definitions:

$$\mathcal{P}_\psi\left(\boldsymbol{v}^{(d)}, \boldsymbol{v}^{(i)}\right) \propto e^{\left(\boldsymbol{\beta}^{(i)^T}\boldsymbol{v}^{(i)}\right)} e^{\left(\boldsymbol{\beta}^{(d)^T}\boldsymbol{v}^{(d)}\right)} \times \delta\left(\mathbb{I}\boldsymbol{v}^{(d)} + \mathbb{G}\boldsymbol{v}^{(i)} - \boldsymbol{b}'\right) \times \prod_m^M \psi\left(\boldsymbol{v}_m^{(d)}\right) \prod_n^{N-M} \psi\left(\boldsymbol{v}_n^{(i)}\right)$$

We now can compute the $\boldsymbol{v}^{(i)}$ marginal as:

$$\mathcal{P}_\psi\left(\boldsymbol{v}^{(i)}\right) \propto \int \left[e^{\left(\boldsymbol{\beta}^{(i)^T}\boldsymbol{v}^{(i)}\right)} e^{\left(\boldsymbol{\beta}^{(d)^T}\boldsymbol{v}^{(d)}\right)} \times \delta\left(\mathbb{I}\boldsymbol{v}^{(d)} + \mathbb{G}\boldsymbol{v}^{(i)} - \boldsymbol{b}'\right) \times \prod_m^M \psi\left(\boldsymbol{v}_m^{(d)}\right) \prod_n^{N-M} \psi\left(\boldsymbol{v}_n^{(i)}\right)\right] d\boldsymbol{v}^{(d)}$$

Note that the delta makes this integral to have a single non-zero contribution at $\boldsymbol{v}^{(d)} = \boldsymbol{b}' - \mathbb{G}\boldsymbol{v}^{(i)}$, so it solves to:

$$\mathcal{P}_\psi\left(\boldsymbol{v}^{(i)}\right) \propto e^{\left(\boldsymbol{\beta}^{(i)^T}\boldsymbol{v}^{(i)}\right)} e^{\left(\boldsymbol{\beta}^{(d)^T}\left(\boldsymbol{b}' - \mathbb{G}\boldsymbol{v}^{(i)}\right)\right)} \times \prod_m^M \psi\left(\boldsymbol{b}'_m - \left[\mathbb{G}\boldsymbol{v}^{(i)}\right]_m\right) \prod_n^{N-M} \psi\left(\boldsymbol{v}_n^{(i)}\right) \quad \text{(Equation 20)}$$

The indicators priors ψ makes the marginals of this distribution hard to compute, so we instead use the approximate multivariate Gaussian $\phi(\boldsymbol{v}; \boldsymbol{a}, \boldsymbol{d})$ with mean vector $\boldsymbol{a} \equiv (\boldsymbol{a}^{(d)}, \boldsymbol{a}^{(i)})$ and variance vector $\boldsymbol{d} \equiv (\boldsymbol{d}^{(d)}, \boldsymbol{d}^{(i)})$ to formulate an approximated join distribution:

$$\mathcal{P}_\phi\left(\boldsymbol{v}^{(i)}\right) \propto e^{\left(\boldsymbol{\beta}^{(i)^T}\boldsymbol{v}^{(i)}\right)} e^{\left(\boldsymbol{\beta}^{(d)^T}\left(\boldsymbol{b}' - \mathbb{G}\boldsymbol{v}^{(i)}\right)\right)} \times \phi\left(\boldsymbol{b}' - \mathbb{G}\boldsymbol{v}^{(i)}; \boldsymbol{a}^{(d)}, \boldsymbol{d}^{(d)}\right) \phi\left(\boldsymbol{v}^{(i)}; \boldsymbol{a}^{(i)}, \boldsymbol{d}^{(i)}\right) \quad \text{(Equation 21)}$$

which is a multivariate Gaussian distribution that can be expressed in standard form as:

$$\begin{aligned} \mathcal{P}_\phi\left(\boldsymbol{v}^{(i)}\right) &\propto \exp\left[\left(\boldsymbol{v}^{(i)} - \overline{\boldsymbol{v}}^{(i)}\right)^T \Sigma^{(i)^{-1}} \left(\boldsymbol{v}^{(i)} - \overline{\boldsymbol{v}}^{(i)}\right)\right] \\ \Sigma^{(i)} &= \left(\mathbb{D}^{(i)} + \mathbb{G}^T \mathbb{D}^{(d)} \mathbb{G}\right)^{-1} \\ \overline{\boldsymbol{v}}(i) &= \Sigma^{(i)} \left(\mathbb{G}^T \mathbb{D}^{(d)} \left(\boldsymbol{b}' - \boldsymbol{a}^{(d)}\right) + \mathbb{D}^{(i)} \boldsymbol{a}^{(i)} - \mathbb{G}^T \boldsymbol{\beta}^{(d)} + \boldsymbol{\beta}^{(i)}\right) \end{aligned}$$

where $\mathbb{D}^{(d)} \in \mathbb{R}^{M \times M}$ and $\mathbb{D}^{(i)} \in \mathbb{R}^{(N-M) \times (N-M)}$ are matrices where all entries are zero and the diagonals equals $1/\boldsymbol{d}^{(d)}$ and $1/\boldsymbol{d}^{(i)}$ receptively.

The parameters of the dependent variables are easily derived from the independents as:

$$\begin{aligned} \Sigma^{(d)} &= \mathbb{G} \Sigma^{(i)} \mathbb{G}^T \\ \overline{\boldsymbol{v}}^{(d)} &= \boldsymbol{b}' - \mathbb{G} \overline{\boldsymbol{v}}^{(i)} \end{aligned}$$

Now, we are in conditions to apply *Expectation Propagation* as described in[76] to find the parameters $\boldsymbol{a}$ and $\boldsymbol{d}$ of the Gaussian priors that better approximate (21) to (20).