# ETFL: A formulation for flux balance models accounting for expression, thermodynamics, and resource allocation constraints

Pierre Salvy[1], Vassily Hatzimanikatis[1,*],

**1** Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
* vassily.hatzimanikatis@epfl.ch

## Abstract

Since the introduction of metabolic models and flux balance analysis (FBA) in systems biology, several attempts have been made to add expression data. However, directly accounting for enzyme and mRNA production in the mathematical programming formulation is challenging because of macromolecules, which introduces a bilinear term in the mass-balance equations that become harder to solve than linear formulations like FBA. Furthermore, there have been no attempts to include thermodynamic constraints in these formulations, which would yield an even more complex mixed-integer non-linear problem.

We propose here a new framework, called Expression and Thermodynamics Flux (ETFL), as a new ME-model implementation. ETFL is a top-down model formulation, from metabolism to RNA synthesis, that simulates thermodynamic-compliant intracellular fluxes as well as enzyme and mRNA concentration levels. The formulation results in a mixed-integer linear problem (MILP) that enables both relative and absolute metabolite, protein, and mRNA concentration integration. The proposed formulation is compatible with mainstream MILP solvers and does not require a non-linear solver. It also accounts for growth-dependent parameters, such as relative protein or mRNA content.

We present here the formulation of ETFL along with its validation using results obtained from a well-characterized *E. coli* model. We show that ETFL is able to reproduce proteome-limited growth, which FBA cannot. We also subject it to different analyses, including the prediction of feasible mRNA and enzyme concentrations in the cell, and propose ETFL-based adaptations of other common FBA-based procedures.

The software is available on our public repository at `https://github.com/EPFL-LCSB/etfl`.

## Author summary

Metabolic modeling is a useful tool for biochemists who want to tweak biological networks for the direct expression of key products, such as biofuels, specialty chemicals, or drug candidates. To provide more accurate models, several attempts have been made to account for protein expression and growth-dependent parameters, key components of biological networks, though this is computationally challenging, especially when also attempting to include thermodynamics. To the best of our knowledge, there is no published methods integrating these three types of constraints in one model. We propose here a transparent mathematical formulation to model both expression and

metabolism of a cell, along with a reformulation that allows a computationally tractable inclusion of growth-dependent parameters and thermodynamics. We demonstrate good performance using community-standard software, and propose ways to adapt classical modeling studies to expression-enabled models. The incorporation of thermodynamics and growth-dependent variables provide a finer modeling of expression because they eliminate thermodynamically unfeasible solutions and consider phenotypic differences in different growth regimens, which are key for accurate modeling.

# Introduction

Metabolic modeling, which helps makes sense of the metabolism in a biological network, is an important tool for engineering biocatalysts, with applications in biofuels, drug design, microbial community analysis, and personalized medicine. Model accuracy is instrumental to the success of these applications through an efficient engineering of the host organisms. However, incorporating expression information into metabolic networks poses a significant challenge, and most current models do not even attempt it—effectively excluding an important network in biological systems that can drastically affect results. In metabolic engineering, strains are modified and controlled at the genome level through the transcriptome, and the effects are observed at the fluxome level, which accounts for the range of metabolic reactions in an organism. In between these two levels is the proteome that actually performs the biochemical transformations according to the genetic template, though it is this middle step in the process that cannot yet be robustly and efficiently incorporated into models of metabolic systems. Because of the complex interplay between these different layers of control, understanding expression and incorporating this into future models is key for improving metabolic engineering.

Classically, model-based strain design has relied on a tool that uses the DNA sequence of an organism to detail the network of metabolic reactions that happen inside a cell of that organism, which is called a genome-scale model (GEM). With current technologies and tools like metagenome sequencing [1], it is possible to generate GEMs for hundreds of different species at a time. GEMs are particularly amenable to flux balance analysis (FBA), which models metabolism at the fluxome level using linear optimization techniques. However, plain FBA has been known to predict biochemically unrealistic solutions like free high-flux cycles or thermodynamically infeasible pathways. It also scales growth linearly with carbon uptake, which is not observed at high-uptake fluxes. FBA also fails to capture growth-dependent and protein-level effects, such as enzyme saturation or proteome-related limitations. Hence, several efforts have been made to supplement FBA with additional constraints to improve its predictive power. For example, thermodynamics-based flux analysis (TFA) [2,3] uses thermodynamic constraints to enforce thermodynamically consistent reaction directionalities and to allow the integration of metabolomics. Resource balance models add a total proteome capacity constraint, as formulated in Beg *et al.* [4], to model the proteome-related limitations of the cell, as enzymes have to compete for the constrained total amount of cellular proteins. Frameworks like GECKO [5] further build on this resource balance idea and include flux constraints based on proteomics, such as $v \leq V_{max} = k_{cat}\,[E]$ as well as a constraint on the total proteome mass. Finally, metabolomics and expression models (ME-models) [6,7] were the first to integrate the entirety of the expression mechanisms of the cell from the bottom-up, including mRNA and protein synthesis.

However, simultaneously accounting for all of these constraints is challenging because of the formulation of each method, as TFA models involve integer variables that yield a mixed-integer linear program (MILP), whereas ME-models involve bilinear constraints that require special optimization procedures and a solver [8–10]. Mixing

these methods would require the inclusion of integers in ME-models, which is not straightforward and would lead to more complex mixed-integer non-linear programs (MINLP) that are computationally intensive to solve. Additionally, the amount of RNA and protein, the RNA and protein expression rates, and their stabilities are all growth dependent [11], and including accurate representations of these variables leads to even more complex, non-linear models. Meanwhile, although resource balance models such as GECKO could theoretically be integrated into TFA or ME-models in the current formulations, to the best of our knowledge, no link with TFA or ME-models has been proposed. Therefore, the metabolic engineering community needs a common formulation for these methodologies to build the most accurate models.

We investigated the development of such a framework and propose herein a unified formulation for Expression and Thermodynamics-enabled FLux models (ETFL) that can account for the above integration issues. In ETFL, we address the compatibility of the formulations by expressing the growth rate variable in bilinear products as a piece-wise constant function. This reformulation allows us to transform the problem into a MILP, which we can solve efficiently using widely available open source or commercial solvers. The resulting model is then effectively able to directly integrate thermodynamic constraints as well as expression constraints and growth-dependent parameters. In this model, metabolite, enzyme, and mRNA concentration levels are explicitly defined to enable fast and easy omics integration. Finally, we show an application of this framework to a well characterized *E. coli* model, iJO1366 [12].

# 1 Results and Discussion

## 1.1 Formulation of the expression problem

To transparently account for expression mechanisms and increase the predictive power of our models, we needed to derive the equations that could bridge the biochemistry with the optimization problem that is ETFL. Here we present a summary of these equations, and detail their derivation in the section Materials and Methods. We derived these equations using assumptions similar to those used in the formulation of the GECKO [5] and ME-model [6–8].

This formulation relies on derivations rooted in the biological mechanism of expression and depends on a number of biochemical parameters related to the cell. In particular, the mass balances of the macromolecules are expressed using concentration variables. Each mass balance will yield an equation where the concentrations of the macromolecules will be variables, thus effectively formulating a new constraint of the model and allowing us to calculate concentration values by solving the model.

We can write the quasi-steady state mass balance for macromolecules as follows:

$$v_{\star}^{syn} - v_{\star}^{deg} - \mu * G_{\star} = 0, \tag{1}$$

where $\star$ represents the indexing of the macromolecule, $v_{\star}^{syn}$ is the synthesis term, $v_{\star}^{deg}$ is the degradation term, and $\mu * G_{\star}$ is the dilution term. The detail of the derivation is available in the Materials and Methods.

Using this formalism, for each macromolecule we can define and link together a synthesis flux, a degradation flux, and the macromolecule's concentration. Knowing enzyme concentrations allows us to bound the variables representing metabolic reaction fluxes with their maximum catalytic rate according to the classical equation:

$$v \leq k_{cat} \cdot E, \tag{2}$$

where $k_{cat}$ is the catalytic rate constant of the enzyme $E$ with respect to flux $v$. In this same fashion, we can also constrain the synthesis flux for the peptides, which are

then assembled into enzymes. Peptide synthesis is simply a metabolic reaction that consumes energy (under the form of GTP) and charged tRNAs and produces a peptide and uncharged tRNAs. The catalytic rate of the reaction is proportional to the maximum ribosomal catalytic rate divided by the length of the peptide to be synthesized. The same can be said about mRNA synthesis, which uses nucleoside triphosphates and is catalyzed by the RNA polymerase. The details of these constraints are explained in the Materials and Methods.

The part of the matrix that has been added to the FBA problem to account for expression has been termed the expression problem (EP). Although this initial formulation is bilinear, we will detail in the following section how we cast it to a MILP.

### 1.1.1 Biomass reaction synthesis and mass balance

In FBA, the biomass reaction is an artificial, lumped reaction that represents the consumption of metabolites in proportion to the cell growth rate. This consumption reflects nucleoside triphosphate (NTP) requirements for mRNAs, amino acid requirements for proteins, lipid requirements for the cell wall, or metal ion needs. Biomass reaction inclusiveness depends on the modeling assumptions made during the model curation process and can vary significantly among models of the same species. The consumed amount of each metabolite is usually estimated experimentally by measuring the the amounts of these metabolites in dried cell mass. Because the stoichiometric ratios of metabolites in the biomass reaction are fixed, the abundance of metabolites is the same for all growth rates. This simplifying assumption, necessary in FBA, goes against experimental evidence. Neidhardt and Curtis [11] report for instance that mRNA and protein mass ratios in the cell change with growth rate.

Because ETFL has explicit expression requirements through transcription, translation, and tRNA-charging reactions, it is possible to account for varying ratios of NTPs and amino acids as the growth rate changes, an effect that is captured in experiments [11]. In this context, the approximation made in FBA can be written using ETFL terms:

$$\forall aa_i, \qquad \eta_{aa_i}^{v_{biomass}} \cdot \mu \approx v_{aa_i}^{charging}, \tag{3}$$

$$\forall NTP_i, \qquad \eta_{NTP_i}^{v_{biomass}} \cdot \mu \approx \sum_j v_{NTP_i}^{tcr_j}, \tag{4}$$

where $v^{biomass}$ represents the biomass equation, and $\eta_{m_i}^{v_{biomass}}$ is the expression-related participation of metabolite $m_i$ in the biomass reaction. Hence, to avoid accounting for the expression requirements twice (once through the biomass equation, once through the EP), it is necessary to remove the participation of these metabolites linked to expression from the biomass reaction.

### 1.1.2 Summary

Here we show the formulation of the constraints of ETFL. For clarity, we use different indexing sets, each referring to a specific object in the model, detailed in Table 1. The variables are detailed in Table 2, and the parameters are in Table 3

*Catalytic constraints*

$$v_j^+ - k_{cat}^{j,+} E_j \leq 0 \tag{FC$_j$}$$

$$v_j^- - k_{cat}^{j,-} E_j \leq 0 \tag{BC$_j$}$$

**Table 1.** Indices used in the formulation.

| Index letter | Indexed variables |
|---|---|
| $i$ | Metabolite |
| $j$ | Reaction/Flux/Enzyme |
| $l$ | Gene/Peptide/mRNA |
| $s$ | Binary coefficient for growth discretization |
| $u$ | Binary coefficient for interpolation discretization |

**Table 2.** Variables used in the formulation.

| Symbol | Variable |
|---|---|
| $\mu$ | Growth rate |
| $v_j^{\pm}$ | $j^{th}$ net positive/negative biochemical flux |
| $E_j$ | Concentration of the $j^{th}$ enzyme |
| $F_l$ | Concentration of the $l^{th}$ mRNA |
| $P_l$ | Concentration of the RNA polymerase assigned to the $l^{th}$ mRNA |
| $R_l$ | Concentration of the ribosome assigned to the $l^{th}$ peptide |
| $T_{aa_i}^u$ | Concentration of the $i^{th}$ uncharged tRNA |
| $T_{aa_i}^c$ | Concentration of the $i^{th}$ charged tRNA |
| $v_l^{tsl}$ | Translation rate of the $l^{th}$ gene |
| $v_l^{tcr}$ | Transcription rate of the $l^{th}$ gene |
| $v_j^{asm}$ | Assembly rate of the $j^{th}$ enzyme |
| $v_j^{deg}$ | Degradation rate of the $j^{th}$ enzyme |
| $v_l^{deg}$ | Degradation rate of the $l^{th}$ mRNA |
| $v_{aa_i}^{charging}$ | Charging rate of the $i^{th}$ tRNA |

**Table 3.** Parameters used in the formulation.

| Symbol | Parameter |
|---|---|
| $k_{cat}^{j,\pm}$ | Forward/backward catalytic rate constant of the $j^{th}$ net biochemical flux |
| $k_{deg}^j$ | Degradation rate constant of the $j^{th}$ enzyme |
| $k_{deg}^l$ | Degradation rate constant of the $l^{th}$ mRNA |
| $\eta_l^j$ | Stoichiometry of the $l^{th}$ peptide in the $j^{th}$ enzyme |
| $\eta_{aa_i}^l$ | Stoichiometry of the $i^{th}$ amino acid in the $l^{th}$ peptide |
| $L_l^{aa}$ | Length in amino acids (aa) of the $l^{th}$ peptide |
| $L_l^{nt}$ | Length in nucleotides (nt) of the $l^{th}$ mRNA |
| $L_{rib}^{nt}$ | Ribosome footprint size on mRNA, in nucleotides |
| $\rho$ | Ribosome occupancy |
| $\pi$ | RNA polymerase occupancy |

*Metabolite mass balance*

$$S \cdot v = 0 \tag{FBA}$$

*Expression mass balance*

$$v_l^{tsl} - \sum_j \eta_l^j \cdot v_j^{asm} = 0 \tag{PB$_\mathrm{l}$}$$

$$v_{rRNA_l}^{tcr} - v_{rib}^{asm} = 0 \tag{RB$_\mathrm{rRNA_l}$}$$

$$v_j^{asm} - v_j^{deg} - \mu * E_j = 0 \tag{EB$_\mathrm{j}$}$$

$$v_l^{tcr} - v_l^{deg} - \mu * F_l = 0 \tag{MB$_\mathrm{l}$}$$

$$-v_{aa_i}^{charging} + \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^u = 0 \tag{TB$^\mathrm{u}_\mathrm{aa_i}$}$$

$$v_{aa_i}^{charging} - \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^c = 0 \tag{TB$^\mathrm{c}_\mathrm{aa_i}$}$$

*Degradation fluxes*

$$v_j^{deg} - k_{deg}^j \cdot E_j = 0 \tag{ED$_\mathrm{j}$}$$

$$v_l^{deg} - k_{deg}^l \cdot F_l = 0 \tag{MD$_\mathrm{l}$}$$

*Expression constraints*

$$v_l^{tcr} - \frac{k_{cat}^{RNAP}}{L_l^{nt}} P_l \leq 0 \tag{TR1$_\mathrm{l}$}$$

$$v_l^{tsl} - \frac{k_{cat}^{rib}}{L_l^{aa}} R_l \leq 0 \tag{TR2$_\mathrm{l}$}$$

$$R_l - \frac{L_l^{nt}}{L_{rib}^{nt}} F_l \leq 0 \tag{EX$_\mathrm{l}$}$$

*Total capacity*

$$\sum_l R_l + R_F - E_{rib} = 0 \tag{TC2}$$

$$\sum_l P_l + P_F - E_{RNAP} = 0 \tag{TC1}$$

$$R_F - (1-\rho)\, E_{rib} = 0 \tag{RR}$$

$$P_F - (1-\pi)\, E_{RNAP} = 0 \tag{PR}$$

## 1.2 Reformulation

### 1.2.1 Bilinearity of the problem

The main issue with the EP formulation presented above lies in the continuous bilinear terms that describe the dilution of the macromolecules, $G_\star \in \{E_j\} \cup \{F_l\} \cup \left\{tRNA_{aa_i}^{(un)charged}\right\}$. We use $\star$ as a placeholder for the indexing of $G$. Using previous notations for the synthesis, degradation, and growth rate:

$$v_\star^{syn} - v_\star^{deg} - \mu * G_\star = 0. \tag{5}$$

In this state, the dilution term is bilinear, and the formulation requires a bilinear solver or potentially a mixed-integer bilinear solver if thermodynamics are to be added. The original ME-model formulation has similar terms as we are presenting here [6, 7]. As such, its recent adaptation in Lloyd *et al.* [8] uses the two-level algorithm SolveME [9] that requires a dedicated non-linear solver. We present instead a MILP approximation of the problem that makes it compatible and solvable with mainstream MILP solvers. We achieve this through the discretization and linearization of the bilinear products. This operation can be understood as locally approximating the bilinear problem by several linear subproblems and choosing the best approximation.

### 1.2.2 Approximation of the growth rate

In ETFL, we approximate the growth rate $\mu$ in bilinear products with a piecewise-constant function $\widehat{\mu}$ ($0^{\text{th}}$ order approximation) to transform the continuous bilinear terms into mixed (integer $\times$ continuous) bilinear terms. This simplifies the problem, as these mixed bilinear terms can be linearized in a MILP setting using the Petersen linearization scheme [13], a particular case of the Glover linearization scheme [14] that was previously used in metabolic engineering by Hatzimanikatis *et al.* [15, 16].

Let $\overline{\mu}$ be an upper bound to $\mu$, $(p, N) \in N^2, p \leq N$. We can approximate $\mu$ with the following $0^{\text{th}}$ order approximation:

$$\forall \mu \in [0, \overline{\mu}], \qquad \mu \approx \widehat{\mu} = p \cdot \frac{\overline{\mu}}{N}. \tag{6}$$

.

With this notation, $\frac{\overline{\mu}}{N}$ is, in fact, the resolution of the approximation. We can then perform the binary expansion of $p$:

$$p = \sum_{s=0}^{\lceil \log_2 N \rceil} 2^s \cdot \delta_s, \tag{7}$$

where $\lceil \log_2 N \rceil$ denotes the smallest majoring integer to $\log_2 N$, and $\delta_s \in \{0, 1\}$ is $r^{\text{th}}$ digits from the right of the binary notation of $p$.

As an example, let us consider modeling an organism whose growth rate does not exceed $\mu_{max} = 2.3\,\text{h}^{-1}$. To do this, we can set $\overline{\mu} = 2.5 \geq \mu_{max}$. Let us choose a resolution of $0.25\,\text{h}^{-1}$, which gives $N = 10$. Then, $log_2 N \approx 3.32$, and $\lceil \log_2 N \rceil = 4$. A growth rate $\mu = 1.4$ will be approximated by:

$$\begin{aligned}
\widehat{\mu} &= 1.5 = 6 \cdot \frac{\overline{\mu}}{10}, \\
\widehat{\mu} &= \left(\delta_0 \times 2^0 + \delta_1 \times 2^1 + \delta_2 \times 2^2 + \delta_3 \times 2^3 + \delta_4 \times 2^4\right) \cdot \frac{\overline{\mu}}{10}, \\
\widehat{\mu} &= (0 \times 1 + 1 \times 2 + 1 \times 4 + 0 \times 8 + 0 \times 16) \cdot \frac{\overline{\mu}}{10}.
\end{aligned}$$

This example is illustrated in Fig. 1. Conceptually, a heuristic for solving this problem would be:

1. Solve the FBA for $\mu$
2. Select the corresponding, closest $\widehat{\mu}$
3. Apply it to compute dilution values
4. Solve the EP with fixed dilution

5. Apply the catalytic constraints to the FBA
6. Recalculate the FBA under catalytic constraints
7. If $\mu \notin \left\{\widehat{\mu} \pm \frac{\overline{\mu}}{N}\right\}$, go back to 3, else, end.

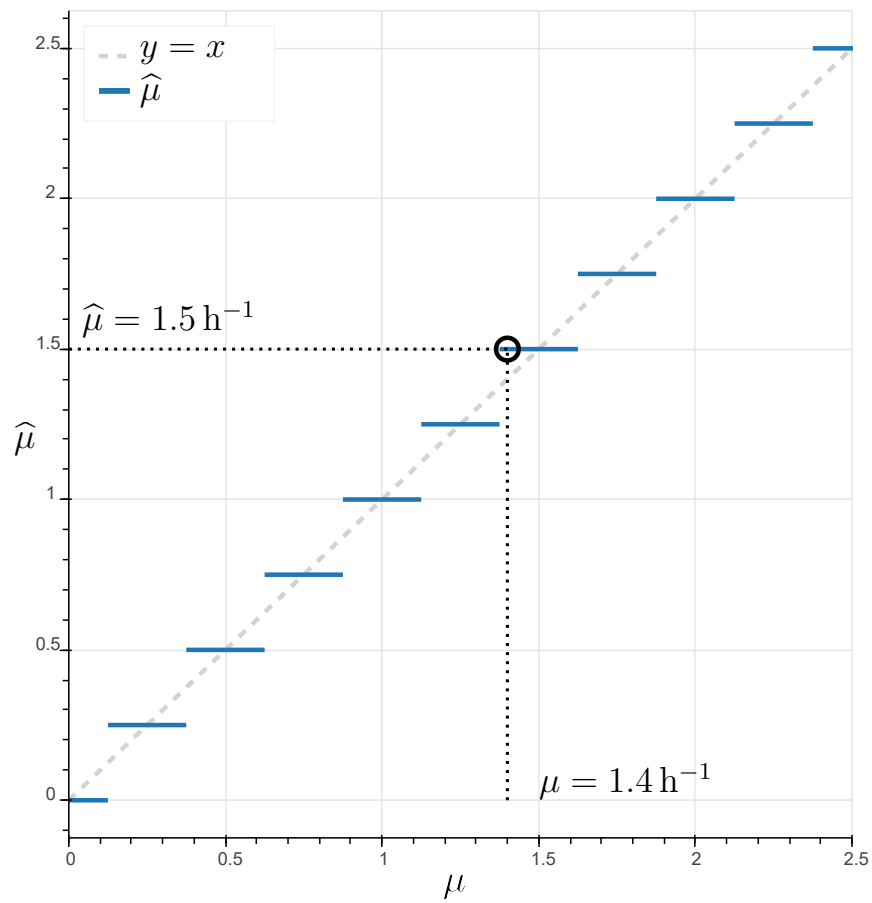


**Fig 1.** Discretization of $\mu$ into $\widehat{\mu}$. The step approximation transforms the continuous interval $[0, 1]$ into the discrete interval $\{0, 0.25, \ldots, 2.5\}$.

### 1.2.3 Linearizing the bilinearity

In the previous derivation, we replaced the growth rate variable by a discrete number of acceptable values. We can approximate the continuous product $\mu * G_\star$, which represents the dilution, as follows:

$$\mu * G_\star \approx \widehat{\mu} * G_\star, \tag{8}$$

$$\widehat{\mu} * G_\star = \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N} \cdot \delta_s * G_\star, \tag{9}$$

The product $\delta_l * G_\star$ is then still bilinear, but one of its variables is binary. Assuming a constant $M > G_\star$, We can use Petersen's linearization theorem [13, 14] to replace the product $\delta_s * G_\star$ with a single nonnegative variable $z_\star^s$, as described in the Materials and Methods.

Because of the binary expansion, the complexity of the model grows only as $\mathcal{O}(\log_2 N) = \mathcal{O}\left(\log_2 \frac{1}{\epsilon}\right)$, where $\epsilon = 1/N$ is proportional to the resolution of the approximation (which is $\frac{\overline{\mu}}{N}$). This means that the linearization part of a model with a resolution of $0.01\,\mathrm{h}^{-1}$ is only around twofold bigger than that of a model with a resolution $0.04\,\mathrm{h}^{-1}$, while resolution has been improved fourfold.

### 1.2.4 Discretization of growth-dependent parameters

**mRNA and enzyme content** Since growth has been discretized, it is now possible to also directly discretize other growth-dependent parameters of the problem, regardless of whether they are in a linear or non-linear relationship with growth. This is a direct consequence of the formulation of ETFL, which allows some flexibility in the modeling assumptions of the user. As an example, we described the relationship between growth and protein and mRNA mass ratios, $P^m$ and $R^m$, in the cell as reported in Neidhardt *et al.* [11] by interpolating and discretizing the protein ratio and mRNA ratio as functions of the growth rate. We used type 1 special ordered set constraints (SOS1) to model these variables with a first-order (piecewise linear) approximation:

$$\sum_j MW_j \cdot E_j - \sum_u \lambda_u \cdot P_u^m = 0, \tag{IC1}$$

$$\sum_l MW_l \cdot F_l - \sum_u \lambda_u \cdot R_u^m = 0, \tag{IC2}$$

$$\sum_u \lambda_u = 1. \tag{SOS1}$$

$P_u^m$ and $R_u^m$ are growth-dependent, interpolated protein and RNA mass ratios (in $\mathrm{g.g}^{-1}$). Given a growth rate, they define the relative mass of the cell that is protein or RNA. $MW_\star$ represents the molar weight of the corresponding enzyme or RNA. The first two constraints enforce equality between the interpolated data and the model production. The last line is the SOS1 constraint that forces only one of the $\lambda_u$ to be active.

Additionally, it is necessary to have the integer index of $\lambda_u$ equal to the index of the growth rate. This is obtained through the constraint:

$$\sum_u u \cdot \lambda_u - \sum_l 2^l \cdot \delta_l = 0. \tag{EQI}$$

The first term represents the growth integer index, and the second represents its binary expansion. However, imposing such mass ratios requires the addition of a dummy mRNA as well as a dummy protein to represent the part of the transcriptome/proteome that is either missing from the expression model or altogether unrelated to metabolic function. We use average amino acid frequencies and GC content to model this. Explicit interpolation functions can also be used, such as the growth-dependent functions given in Pramanik *et al.* [17].

The simultaneous use of catalytic constraints on metabolic reactions (Eq. $\mathrm{FC_j}$, $\mathrm{BC_j}$) and maximal enzyme load (Eq. IC3) effectively implements allocation constraints like in GECKO [5], although in ETFL, the enzyme concentrations are also directly linked to the metabolism. In GECKO, the metabolic cost of building the enzymes is not taken into account.

Fig. 2 shows an example piecewise linear interpolation of the growth-dependent protein mass ratio in *E. coli* according to Neidhardt *et al.* [11]. The reported values (red circles) are interpolated using a piecewise linear function (dashed line), which is then discretized (full line). Using the integer constraints described above, the model can

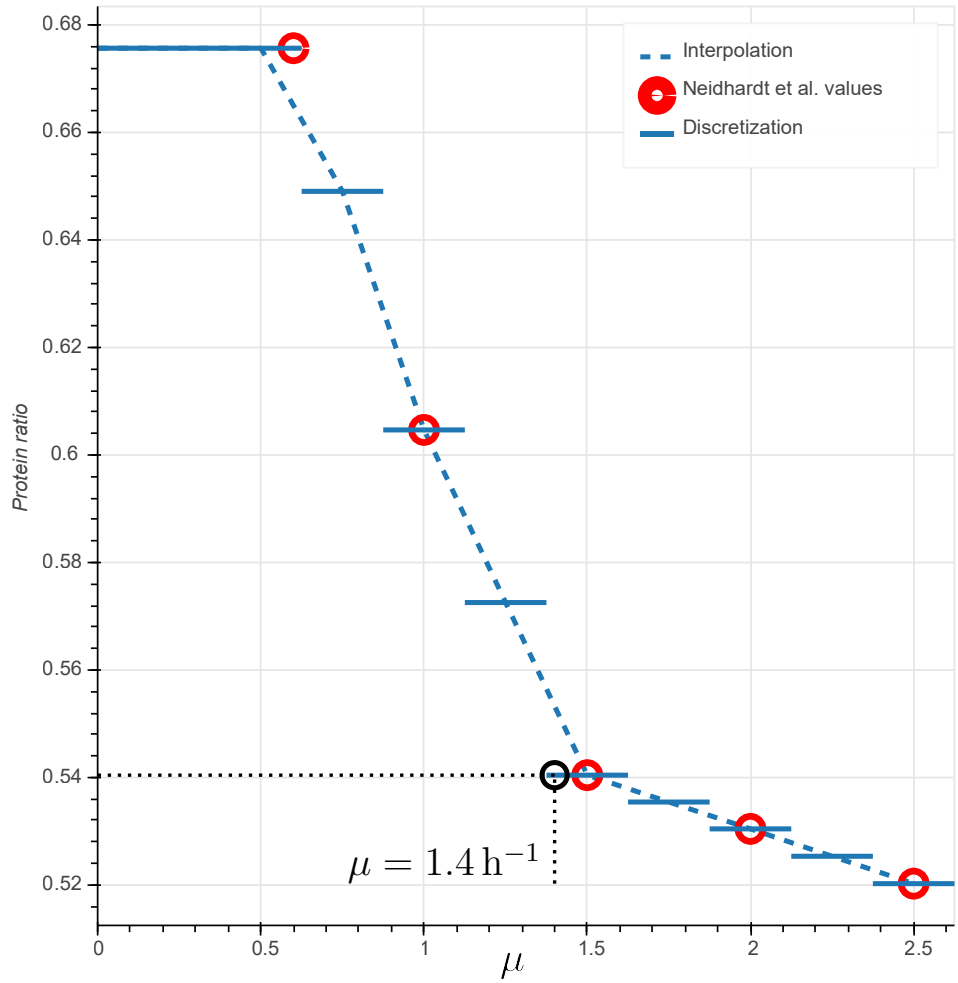


**Fig 2.** Example of piecewise linear interpolation and discretization of the protein mass ratio from Neidhardt *et al.* [11]. Red circles represent the values reported. The dashed line is the piecewise linear interpolation. The solid line is its discretization.

be forced to display a protein content that corresponds to its growth. We apply the same techniques to mRNA and DNA content.

**DNA content** To further increase the scope of macromolecules covered by the model, it is also possible to add growth-dependent DNA content. DNA mass ratios at specific growth rates are reported in Neidhardt *et al.* [11]. We model the DNA reaction synthesis as follows:

$$
\begin{aligned}
v_{DNA}^{synthesis}: \quad & (1-\gamma)\, L_{DNA}^{bp} dATP, \\
& + (1-\gamma)\, L_{DNA}^{bp} dTTP, \\
& + \gamma L_{DNA}^{bp} dGTP, \\
& + \gamma L_{DNA}^{bp} dCTP, \\
& \rightarrow DNA + 2L_{DNA}^{bp} PPi,
\end{aligned}
$$

where $\gamma$ is the GC content of the cell, and $L_{DNA}^{bp}$ is the total length in base pairs of the DNA. As with $mRNA_l$ and $Enz_j$, $DNA$ has a mass-balance equation of the following shape:

$$\frac{d}{dt}[DNA] = 0 = v_{DNA}^{synthesis} - v_{DNA}^{degradation} - v_{DNA}^{dilution}, \tag{10}$$

$$v_{DNA}^{synthesis} - v_{DNA}^{deg} - \mu * DNA = 0. \tag{$DB_{DNA}$}$$

We consider that the DNA does not degrade, meaning the only source of DNA consumption is dilution caused by the growth of the cell and $k_{deg}^{DNA} = 0$. We then define the molar weight of DNA $MW_{DNA}$ and enforce the DNA mass ratio $Dm$ as we did with both proteins and mRNA:

$$\begin{aligned} MW_{DNA} = (1-\gamma)\, L_{DNA}^{bp} \left(MW_{dATP} + MW_{dTTP}\right), \\ +\gamma L_{DNA}^{bp} \left(MW_{dGTP} + MW_{dCTP}\right), \end{aligned} \tag{11}$$

$$MW_{DNA} \cdot DNA - \sum_u \lambda_u \cdot Dm_u = 0. \tag{IC3}$$

### 1.2.5 Scale

A critical issue in the formulation of this problem is that the variables are different orders of magnitude. Fluxes are typically between $10^{-3} - 10^{1}\,\mathrm{mmol.gDW^{-1}.h^{-1}}$, whereas protein concentrations are around $10^{-6} - 10^{-3}\,\mathrm{mmol.gDW^{-1}}$ and mRNA concentrations are $10^{-10} - 10^{-6}\,\mathrm{mmol.gDW^{-1}}$. The relationship between these scales is given by the catalytic rate constant of enzymes and expression machinery, which spans from $10^3 - 10^6 h^{-1}$. In particular, the ribosome rate constant for translation ($\sim 12\,\mathrm{aa.s^{-1}} = 43\,200\,\mathrm{aa.h^{-1}}$) as well as the RNA polymerase rate constant of transcription ($\sim 85\,\mathrm{nt.s^{-1}} = 306\,000\,\mathrm{nt.h^{-1}}$) are responsible for strong differences in the concentrations and fluxes between transcription- and translation-related parts of the problem. Consequently, the constraint matrix becomes ill-conditioned, and the solver has to operate close to, or sometimes beyond, its maximal solving accuracy (usually around $10^{-9}$ for commercial solvers such as ILOG CPLEX or Gurobi).

To circumvent these limitations, we scale the EP, which will reduce the numerical difficulty of the problem, using nondimensionalization. We create nondimensionalized variables by dividing the variables of the initial problem by an estimated upper bound. For example, by definition, macromolecule concentrations cannot exceed $1\,\mathrm{g.gDW^{-1}}$, and the following constrains the transformed macromolecule variables between 0 and 1:

$$\widehat{X} = \frac{X}{\sigma_X}, \sigma_X \geq \sup(X) \implies 0 \leq \widehat{X} \leq 1. \tag{12}$$

In this scheme, $\sigma_X$ is an upper bound to $X$. In particular, if we consider $\sigma_X$ to be the concentration of $1\,\mathrm{g.gDW^{-1}}$:

$$\begin{aligned} \sigma_X &= 1\,\mathrm{g.gDW^{-1}}, \\ &= 1\,\mathrm{g.gDW^{-1}} \times \frac{1}{\mathrm{MW}(X)}\,\mathrm{mmol.g^{-1}}, \\ &= \frac{1}{\mathrm{MW}(X)}\,\mathrm{mmol.gDW^{-1}}, \end{aligned} \tag{13}$$

where $\mathrm{MW}(X)$ denotes the molecular weight of the macromolecule in SI units ($\mathrm{kg.mol}^{-1} \equiv \mathrm{g.mmol}^{-1}$), and $\widehat{X}$ represents the mass fraction of the molecule in the cell. We scale the fluxes using a method derived from this, detailed in the supporting file S1 Nondimensionalization. It is also possible to further refine this upper bound by performing a variation analysis on $X$ and re-generating a model using the newly estimated upper bound.

For the sake of clarity, all problem formulations will be kept in their dimensionalized form in the subsequent equations although the implementation is actually nondimensionalized. The nondimensionalized problem is described further in S1 Nondimensionalization.

### 1.2.6 Recovering the FBA problem

In the ETFL formulation, enzyme synthesis is driven by the coupling between FBA and EP through the catalytic constraints. To carry flux, the cell needs to produce enzymes whose production will also use the metabolic resources of the cell. If allocation constraints are enforced, the amount of protein and mRNA synthesized must meet predefined mass ratios for the problem to be feasible. Hence, the metabolic requirement terms for the expression machinery (amino acids and NTP) have been removed from the biomass reaction and are accounted for in the tRNA charging and transcription reactions. Thus, the FBA solutions can be recovered from the ETFL formulation by the following routine:

- Setting $\forall j, \quad k_{cat}^{j,\pm} = +\infty$,
- Constraining $\forall aa_i, \quad v_{aa_i}^{charging} = \eta_{aa_i}^{v_{biomass}} \cdot \mu$,
- Constraining $\forall NTP_i, \quad \sum_j v_{NTP_i}^{transcription_j} = \eta_{NTP_i}^{v_{biomass}} \cdot \mu$,
- If applicable, relaxing the allocation constraints,
- If applicable, relaxing the thermodynamic coupling constraints.

## 1.3 Application: *E. coli* genome-scale model iJO1366

iJO366 [12] is a well-curated and well-studied GEM of *E. coli* that is closely related to the GEM used in developing both ME-models iOL1650-ME [7] and iJL1678b-ME [8]. Additionally, this model has been extensively applied in the literature and is aligned with a variety of datasets that can be used for data integration. We wanted to subject the model to classical studies that would highlight the power of ETFL, particularly as pertains to proteome-limited growth, macromolecule concentration variability analysis, and gene knock-out studies. We also wanted to assess the sensitivity of the model with respect to the presence of thermodynamic constraints as well as growth-dependent parameters.

Thus, we first experimented with four different models using ETFL with or without thermodynamic constraints and growth-dependent protein/RNA/DNA allocation following Table 2 as reported by Neidhardt *et al.* [11]. The following Table 4 details the nomenclature used to refer to these different models. The features of the most constrained model containing both thermodynamic and growth-dependent parameters, vETFL, are detailed in Table 5. These four models were optimized for maximal growth at increasing glucose uptake rates to assess their behavior with respect to excess substrate, which will show the non-linearity of the relationship between growth and glucose uptake at high uptake rates. A plateau in the growth rate was expected, which indicates a proteome-limited phenotype that cannot be observed with FBA. We also

**Table 4.** Nomenclature of the models used in the study of *E. coli* iJO1366. EFL stands for Expression and FLuxes, ETFL for Expression, Thermodynamics, and FLuxes, and the v- prefix indicates the inclusion of growth-dependent parameters (see 1.2.4)

| | growth-independent parameters | growth-dependent parameters |
|---|---|---|
| (-) thermodynamics | EFL | vEFL |
| (+) thermodynamics | ETFL | vETFL |

**Table 5.** Properties of the vETFL model generated from iJO1366.

| | |
|---|---|
| Growth upper bound $\widehat{\mu}$ | $3.5h^{-1}$ |
| Number of bins $N$ | 128 |
| Resolution $\frac{\widehat{\mu}}{N}$ | $0.0273h^{-1}$ |
| Number of constraints | 68304 |
| Number of variables | 49207 |
| Number of species | 3240 |
| – Metabolites | 1806 |
| – Peptides | 1431 |
| – rRNA | 3 |
| Number of enzymes | 562 |
| Number of reactions | 8023 |
| – Metabolic | 1543 |
| – Transport | 733 |
| – Exchange flux | 330 |
| – Transcription | 1431 |
| – Translation | 1431 |
| – Complexation | 562 |
| – Degradation | 1993 |
| Number of metabolites $\Delta_f G'^o$ | 1737 |
| Number of reactions $\Delta_r G'^o$ | 1787 |
| Percent of metabolites $\Delta_f G'^o$ | 93.9% |
| Percent of reactions $\Delta_r G'^o$ | 68.6% |

subsequently subject vETFL to a variability analysis and gene essentiality analysis, which will respectively show us the flexibility of the model and its accuracy in predicting gene knock-out behavior.

### 1.3.1 Growth rate prediction

To study the behavior of the model at different carbon uptake rates, we simulated growth on a minimal medium with only glucose as a carbon source, unlimited oxygen, and some essential inorganic compounds. This would allow us to show that at a higher carbon uptake, the model would predict a limited growth – unlike FBA that would predict an unlimited linear increase.

Figure 3 shows the predicted growth rate of the different (v)E(T)FL models described in Table 4 with respect to the glucose uptake of the cell. As expected and in contrast to current FBA models, all four models plateau after a certain uptake rate, which indicates a proteome-limited phenotype due to the limited capacity of the cells to make more enzymes to metabolize the glucose. As discussed for the ME-models [7] and GECKO [5] formulations, within the context of models accounting for protein usage, this is caused by (i) the protein burden necessary to metabolize higher fluxes; (ii) the increased demand in protein synthesis at higher growth rates; and (iii) for the models with allocation constraints, the allowed protein and RNA mass ratio. We can see that

models featuring protein, RNA, and DNA allocation constraints (vE[T]FL) consistently predict a lower growth rate than models without allocation constraints. This is expected, as the data we input requires additional proteins and mRNA to account for non-metabolism-related macromolecules. Models featuring thermodynamic constraints ([v]ETFL) also predict a lower growth rate, consistent with the fact that thermodynamic constrain the model to valid solutions whose flux is in the subspace of the FBA feasible space. The most constrained model (vETFL) consequently has the lowest growth rate at any glucose uptake. This is in accordance with published TFA results that eliminated biologically infeasible flux profiles yielding non-realistic higher growth rates [2].

We summarize the constraint matrix of the EP of vETFL in Fig. S2 Example EP constraint matrix, where each line represents a type of constraint and each column represents a type of variable. The blocks of the matrix that are non-zero are colored, and these blocks directly reflect the involvement of the constrained variables.

### 1.3.2 Modeling missing enzymes

Although we initially focused on including only enzymes for which we had all the necessary information (catalytic rate and peptide constitution), we wanted to assess the behavior of our model when the missing enzymes were modeled as well as check our model's sensitivity to changes in the catalytic rate constants. Thus, we additionally built three more models, based on vETFL, with the following properties: (i) all the missing enzymes were estimated by averaging the properties of the known enzymes based on the curation for the vETFL iJO1366 (333 amino acids long, average $k_{cat}^{\pm} = 172\,\mathrm{s}^{-1}$); (ii) all the enzymes (including the missing enzymes) but the ribosome, RNA polymerase, and ATP synthase were assumed to have a catalytic rate constant $k_{cat}^{\pm} = 65\,\mathrm{s}^{-1}$; and, for comparison purposes, (iii) all the *known* enzymes of vETFL except for the ribosome, RNA polymerase, and ATP synthase were assumed to have a catalytic rate constant of $k_{cat}^{\pm} = 65\,\mathrm{s}^{-1}$. $k_{cat} = 65\,\mathrm{h}^{-1}$ is an assumption used in O'Brien *et al.* [7] that we reproduce here as a study on sensitivity of the quality of the $k_{cat}$ values. We observe that this value matches the median value of the catalytic rate constants in the 562 vETFL enzymes. For clarity, we will refer to these models as (i) the estimated mean enzymes model; (ii) the full median model; and (iii) the partial median model. The ribosome, RNA polymerase, and ATP synthase were not modified, as their catalytic rates directly and strongly affect the growth of the organism. Any drastic change in these would make changes related to other enzymes negligible in comparison.

Figure 3b shows a comparison of the growth prediction for the estimated mean enzymes (brown), full median (pink), and partial median (purple) models designed to account for the missing enzymes. For a better comparison, we also reproduce the vETFL results in red on the same graph. The partial median model (purple) shows a higher predicted growth than the original vETFL model (red). This implies that limiting enzymes in the original vETFL model have a $k_{cat}$ parameter lower than the median value. Both models featuring inferred enzymes, the full median model (pink) and estimated mean enzymes model (brown) show a lower growth rate at a given uptake. This is expected as fluxes which previously had no enzymes assigned in vETFL are now subject to catalytic constraints, and thus the models are more constrained. Additionally, we observe that the model with estimated mean enzymes (brown) is also below the full median model (pink). Similarly to vETFL and the partial median model, this shows that the limiting enzymes in the model with estimated mean enzymes have a $k_{cat}$ parameter lower than the median value. Finally, we observe that the differences between these four models only appear at glucose uptake rates higher than $\approx 6\,\mathrm{mmol_{glc}.DW^{-1}.h^{-1}}$, when the problem switches from being stoichiometry-limited to proteome-limited. Thus, this experiment also illustrates the importance of well-curated catalytic rate constants for modeling organisms grown in proteome-limited regimens.

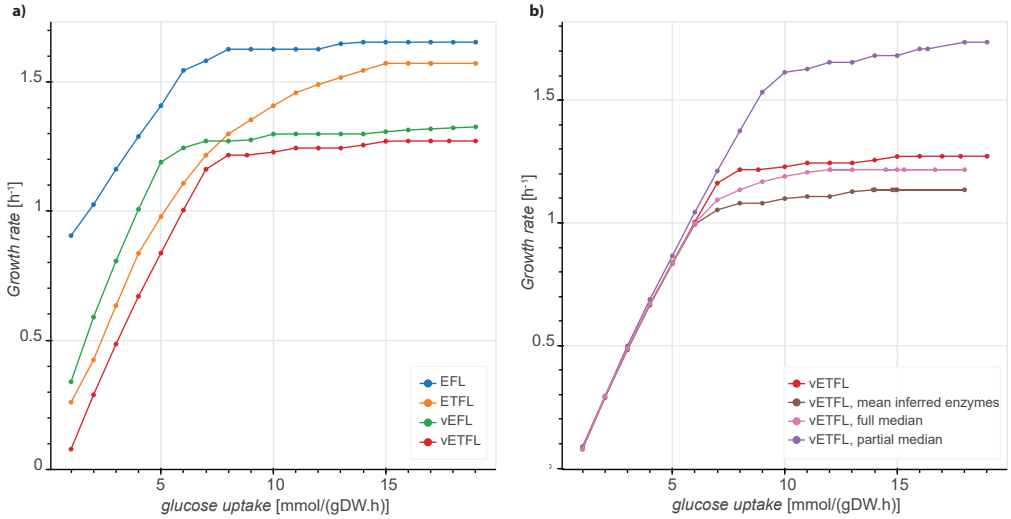


**Fig 3.** Growth rate with respect to glucose uptake for differently constrained models in the ETFL framework. **a.** Growth rate predictions using the [v]E[T]FL models with $k_{cat}$ values aggregated from the reaction database SabioRK [18], **b.** Growth rate predictions accounting for missing enzymes using vETFL and models (i)-(iii) representing different initial enzyme assumptions, with $k_{cat}$ values obtained from SabioRK or $k_{cat} = 65\,\mathrm{h}^{-1}$, and with/without inferred enzymes.

These results demonstrate the capability of ETFL to predict different phenotypes depending on growth rate. ETFL is also amenable to hypothesis testing, as evidenced using the models that estimate the missing enzymes. In particular, we showed with ETFL that an uptake increase does not yield a proportional growth rate increase as with FBA and that ETFL provides a maximal uptake rate that is unmodeled in FBA, thus more effectively modeling growth-dependent biomass yield in *E. coli* . This allows for more realistic predictions for phenotypes that are limited by the expression capabilities of the cell as well as captures the variability of the biomass composition in different growth regimens.

### 1.3.3 Variability analysis

It is also possible to subject the model to a range of variability analyses. These are routinely used in FBA to assess the flexibility of the system and in TFA to find the ranges of allowed metabolite concentrations. We can extend their use in ETFL to explore the allowed proteome and transcriptome. For example, we measured the admissible extreme concentrations of each mRNA in aerobic growth conditions as described in McCloskey *et al.* [19] by performing a variability analysis on the mRNA concentration variables. Fig. 4 depicts the admissible peptide concentration upper and lower bounds, sorted by average, for vETFL with a glucose uptake set to $12.5\,\mathrm{mmol.gDW}^{-1}.\mathrm{h}^{-1}$, which yields a proteome-limited phenotype, according to our results in Fig 3a. It is important to note that all peptides with a non-zero minimal concentration (most of the left of the figure) are, by definition, essential peptides: These are always present at this uptake rate and are hence necessary for the cell to grow at an optimum growth rate. The same study can be performed for enzyme concentrations or even metabolite log-concentrations for models with thermodynamics. This type of study is useful for comparing how the model performs in relation to actual proteomics, transcriptomics, or metabolomics data. The method for running these other types of variability analyses is exactly the same – only the variables subject to the variability

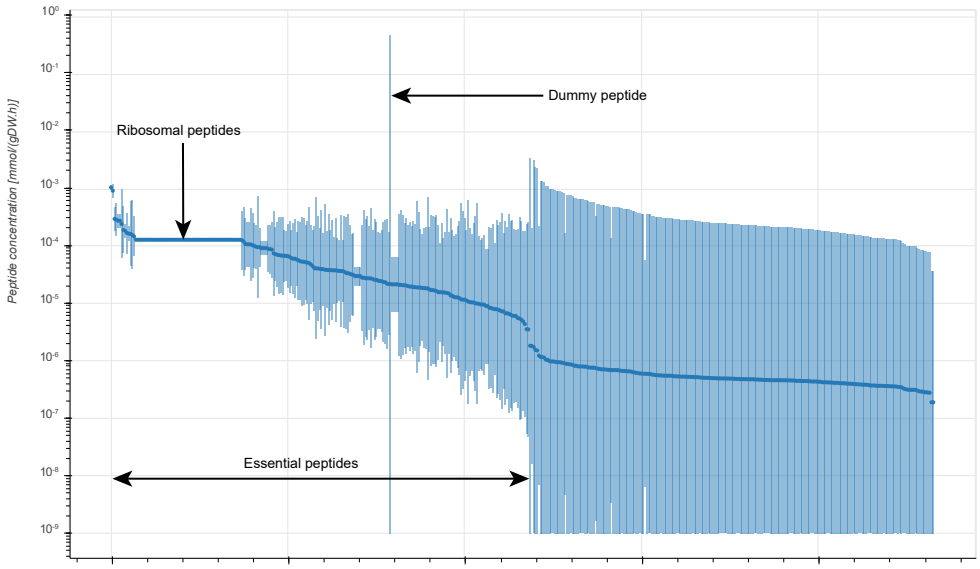


**Fig 4.** Concentration variability of peptide species, sorted by average peptide concentration (darker disc). Lower bounds that were 0 were set to the accuracy of the solver, $10^{-9}$. The horizontal line on the left side of the figure represents ribosomal peptides, which is narrow due to their instrumental role in making the tightly constrained amount of protein in the cell at a given growth rate. The vertical line in the middle represents the dummy peptide, which accounts for unmodeled peptides (non-metabolic proteins and enzymes with missing information) and therefore is used by the solver as a slack.

analysis are changed.

A specific usage of a variability analysis is the study of the allowed proteome (resp. transcriptome) that is done by performing a variability analysis on the enzyme (mRNA) concentration variables. This type of study can, for instance, be compared with transcriptomics to check if the expression profile of an engineered strain corresponds to what is expected in its corresponding model. A way to visualize the average allowed proteome (transcriptome) is to use the average value of the variability of each enzyme (mRNA) concentration as a feasible observation. Due to the convexity of the solution space, it is a solution to the problem. This observation is then plotted on a finite area, which can be done using the online software Proteomaps [20, 21]. This method and software are often used by biologists to represent protein abundances in the cell, and using the data from ETFL, we can generate similar comparative graphs that can help biologists analyze the variability in the different concentration variables using a visualization they are familiar with.

Fig. 5 is an example of such a representation, graphed using the mRNA concentrations corresponding to the solution represented by the dark dots in Fig 4 as an input. In this figure, mRNAs are clustered using KEGG Gene Ontology (GO) annotations. GO annotations form a tree describing the physiological role of genes, ranging from the least specific (e.g. general metabolism) to most specific (e.g. araH gene). The area of each (sub)cluster is proportional to the relative abundance of each (sub)group of mRNAs.

We used the mean of the variability analysis as the observation rather than a single optimal solution because the optimality principle in LP only guarantees a unique global optimum value and not a unique optimal solution. Moreover, solver heuristics give

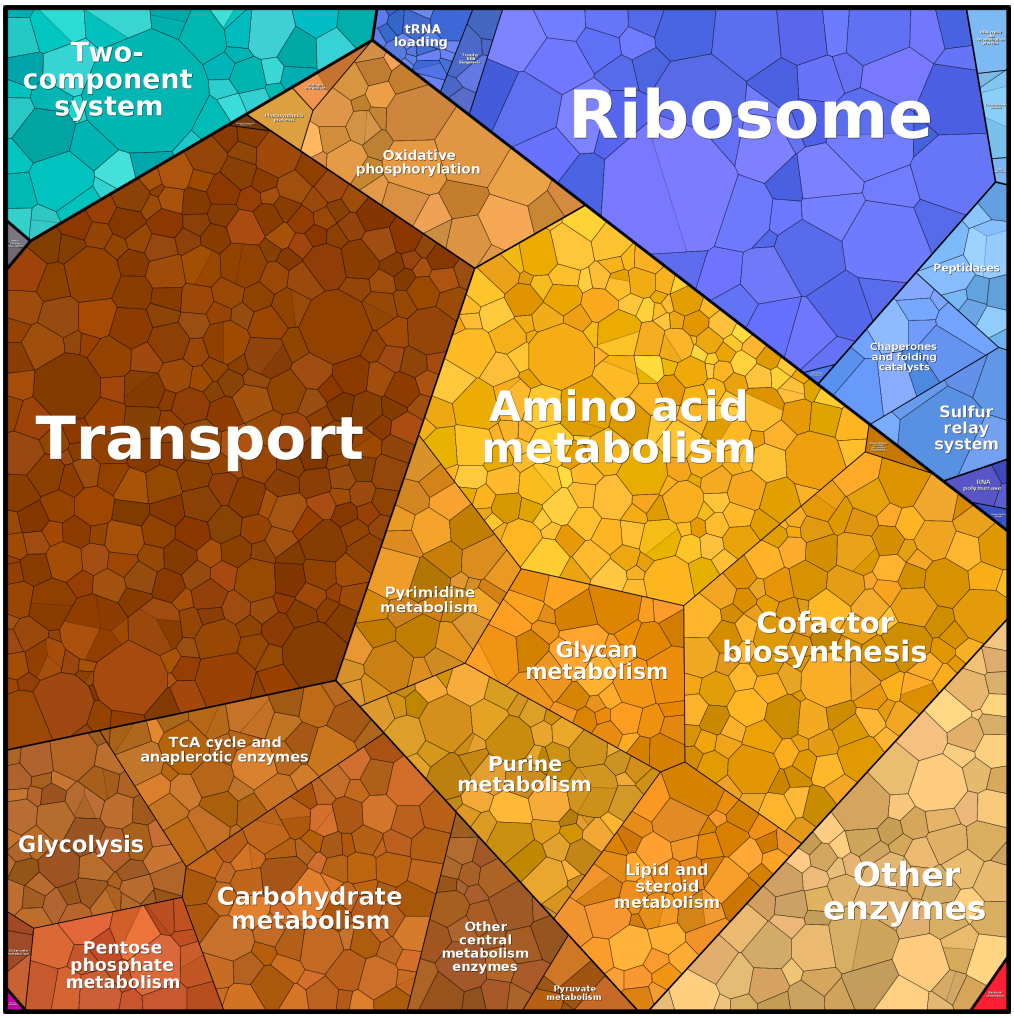


**Fig 5.** Proteomaps of the predicted abundances of mRNAs. Each colored patch represents a different mRNA in proportion to its relative abundance. Genes can be clustered using KEGG Gene Ontology (GO) terms.

sparse and extreme results (corners of the explored simplex), which do not accurately represent the full extent of the considered solution space.

### 1.3.4 Essentiality analysis

The ETFL framework can also analyze the essentiality of specific genes by performing single gene knockouts. The growth of models with knocked-out genes can then be compared to experimental data to assess the quality of the model as a validation. We performed this analysis and compared it to the results reported in the publication of iJO1366 by Orth *et al.* [12]. We use the Matthew's correlation coefficient (MCC) as a metric for the quality of the prediction, which is preferred over accuracy as it is not sensitive to the imbalance between the number of essential genes and non-essential genes. The MCC reads like a usual correlation coefficient, with 1 being a perfect correlation, -1 perfect anti-correlation, and 0 no correlation. We used the essentiality data and conventions given in the supplementary material of Orth *et al.* [12], as explained in Fig. 6-a and Fig. 6-b. Knockouts in ETFL are simulated by setting the corresponding

| a) | | Experiment | |
|---|---|---|---|
| | **Prediction** | Essential | Non-essential |
| | Essential | TN | FN |
| | Non-essential | FP | TP |

| b) | **iJO1366** | Essential | Non-essential |
|---|---|---|---|
| | Essential | 168 | 39 |
| | Non-essential | 80 | 1079 |

| c) | **vETFL** | Essential | Non-essential |
|---|---|---|---|
| | Essential | 136 | 35 |
| | Non-essential | 112 | 1083 |

| d) | **vETFL, kcat = 65 h-1** | Essential | Non-essential |
|---|---|---|---|
| | Essential | 130 | 31 |
| | Non-essential | 118 | 1087 |

**% True Prediction**

| |
|---|
| 100 |
| 90 |
| 80 |
| 70 |
| 60 |
| 50 |
| 40 |
| 30 |
| 20 |
| 10 |
| 0 |

**Fig 6. a.** Conventions from Orth *et al.* [12] for gene essentiality. TN is True Negative. FN is False Negative. FP is False Positive. TP is True Positive. The color shading represents how good the classification is in the experimental class. Perfect classification should have a strict red first diagonal, as shown on this example. **b.** Gene essentiality prediction for FBA model iJO1366, yielding a Matthew's correlation coefficient (MCC) of 0.69. **c.** Gene essentiality prediction for vETFL model, yielding a MCC of 0.60. **d.** Gene essentiality prediction for vETFL model with all $k_{cat} = 65\,\mathrm{h}^{-1}$, yielding a MCC of 0.59

transcription reaction rate to 0 as opposed to in FBA where knockouts derive from Gene Protein association Rules (GPR), which are Boolean rules that directly set metabolic reaction rates to 0 if the knocked-out gene is essential to this reaction. The difference is that in ETFL, the solver actually sees the knock-out as a part of the model, while FBA changes the model to accommodate the knock-out. This feature can be used in strain design strategies to optimize directly for knock-outs. The results are presented in Fig. 6-c and 6-d, respectively for vETFL and the partial median model.

We observe that, compared to iJO1366, (i) ETFL presents more false positives (experimentally essential genes predicted as non-essential); and (ii) ETFL predicts fewer false negatives (experimentally non-essential genes predicted as essential). This first point might indicate that ETFL is less constrained than iJO1366—the cell has more genetic alternatives for growth. This is an artificial effect that likely stems from the missing enzyme data, because if a reaction depends on a missing enzyme, by default there will be no coupled expression and hence no impact if the related gene is knocked out. As more enzyme data is added to the model, the false positive rate should decrease. The second point indicates that ETFL better captures the genes that are not essential for growth than iJO1366. This is an expected effect of explicit expression coupling, from transcription to enzyme concentration, as opposed to Boolean gene reaction rules.

### 1.3.5 Sampling

Sampling the feasible solution space of FBA is a common way to study solution robustness and variability. Since there are often multiple FBA solutions at the optimal objective value, representative solutions are often sought, and sampling is one way to obtain them. However, because ETFL contains integer variables, it is not compatible with traditional sampling methods in its current formulation. It is possible, though, to make the model convex, and hence amenable to sampling, by fixing the integers to their values at a given growth rate and, if applicable, TFA directionality. This will block the flux directions as well as the growth-dependent parameters if TFA is performed. The resulting model is then solely linear, and sampling can be performed with traditional techniques, such as artificially centered hit and run (ACHR) [22], gpSampler [23], or optGpSampler [24]. Once it has converged, a sampling should provide a better representation of the center of the solution space than the mean of the variability analysis.

## 1.4 Performance

ETFL relies on solver-specific MILP algorithms and heuristics, which also means that great variability in performances can be observed depending on the solver parameters. We provide tuned presets for different tasks (gene knock-out, variability analysis, growth maximization) with the package, and recommend that users run their own solver tuning if long run times are observed. We witnessed an up to 10× increase in performance using such tuning.

## 1.5 Adaptation of FBA-based methods to ETFL

The ETFL formulation is amenable to further kinds of analyses. Leveraging both the explicit expression constraints and the MILP nature of the problem, we present several possibilities for future studies using ETFL:

**Growth-dependent parameters** It has been reported that several other parameters, such as the ribosome transcription rate constant $k_{rib}$, are growth dependent [11]. Although such dependency is not taken into account in the presented results, it is possible to account for this by (i) discretizing $k_{rib}$ following the method used in 1.2.4, and (ii) using Petersen's linearization scheme from 1.2.3 on the product $k_{rib} * E_{rib}$. Other parameters that can be transformed in this way include, but are not limited to, the RNAP transcription rate constant $k_{trans}$, free ribosomes, and the RNAP ratios $\rho$ and $\pi$.

**Omics integration** Explicit mRNA and enzyme concentrations allow the direct integration of absolute or relative proteomics and transcriptomics by changing the bounds of the corresponding variables in the EP. An additional gauge constraint will be needed for relative data. Previous transcriptomic integration methods, such as REMI [25], iMAT [26], GIMME [27], or MINEA [28], can also be adequately reformulated for ETFL. Metabolomics can still be integrated using TFA [2, 3].

**Minimization of adjustment** In the original paper, the hypothesis behind the Minimization of Metabolic Adjustment (MOMA) method is that the metabolic fluxes of an organism subject to a gene knock out show a minimal change compared to the metabolic fluxes of the wild-type organism [29]. The underlying hypothesis is that the enzyme distribution and assignments remain the same except for the knocked-out gene. With ETFL, it is possible to directly compute a Minimization of Protein Adjustment

(MOPA) by reformulating the objective function as a Minimization of Expression Adjustment (MOXA):

$$\min \sum_j \left\| E_j - E_j^0 \right\|_p, \quad p \in \{0,1\} \tag{MOPA}$$

where $\|\cdot\|_p$ is either the Manhattan norm ($p = 1$, $\ell_1$-norm) or the Euclidean norm ($p = 2$, $\ell_2$-norm), which will require a MIQP solver. In the same fashion, it is also possible to formulate a (weighted) Minimization of mRNA Adjustment (MORA) or even a Minimization of eXpression Adjustment (MOXA) using the following formulations:

$$\min \sum_l \left\| F_l - F_l^0 \right\|_p, \quad p \in \{0,1\} \tag{MORA}$$

$$\min \theta \cdot \sum_j \left\| E_j - E_j^0 \right\|_p + (1-\theta) \cdot \sum_l \left\| F_l - F_l^0 \right\|_p. \quad p \in \{0,1\}, \theta \in [0,1] \tag{MOXA}$$

**Parsimonious analysis** Parsimonious FBA (pFBA) [23] was developed to address the high fluxes of some of the solutions given by FBA. Although this concern is addressed in ETFL by the combined actions of the EP and thermodynamics, pFBA can be adapted to ETFL to study an organism under parsimonious constraints. For example, it is possible to reformulate it into a parsimonious expression problem to find the minimal expression level required to meet a growth target using objective functions similar to MOPA, MORA, and MOXA. It is also possible to turn the problem around to consider the allowed enzyme amounts under minimal flux constraint obtained by pFBA to assess the metabolic flexibility of an organism.

**Dynamic ETFL (dETFL)** Dynamic FBA (dFBA) [30] is a method that uses FBA to predict the dynamics of a biological system represented with a stoichiometric model. In its original static optimization approach (SOA) formulation, a FBA problem is solved at each time step. The value of boundary fluxes of the FBA problem are updated at each iteration with values produced with a kinetic law, such as Michaelis-Menten glucose uptake and oxygen diffusion. Because ETFL allows direct access to enzyme concentrations, it is possible to use the latter to reformulate dFBA in its SOA. The original SOA approach uses ad-hoc constraints on the absolute flux change at each time step. However, in ETFL, it is possible to bound flux changes indirectly by bounding enzyme and mRNA concentration changes in the EP. Effectively, this approach allows the movement from ad-hoc constraints to physiological constraints.

**Use in kinetic frameworks** Often, kinetic frameworks require a reference flux distribution as an input. ETFL can provide such a distribution, with an increased accuracy as compared to FBA.

# 2 Conclusions

ETFL is a framework can implement expression and thermodynamic formalism using mainstream MILP solvers. This could not be previously accomplished using state-of-the-art ME-models, which use specialized quad-precision solvers and do not support integer variables. The formalism itself is based on the explicit and direct relationship with the underlying biochemistry and provides a way to incorporate growth-dependent variables using MILP linearization techniques. These new growth-dependent variables provide a finer modeling of expression because they consider phenotypic differences in different growth regimens, which is key for accurate modeling.

ETFL can also compute explicit mRNA and enzyme concentrations as well as perform direct -omics data integration. In this, ETFL complements and extends FBA capabilities by using explicit relationships in lieu of the typical assumptions on the relationships between the transcriptome, proteome, and fluxome. This explicit accounting of expression mechanisms provides a finer level of control and a more relevant prediction of gene-editing outcomes. Because of this and its operational similarity with classic FBA-related analyses, ETFL can be efficiently integrated in standard model-based pipelines. For example, metagenome-based genome-scale reconstructions such as published by Magnúsdóttir *et al.* [1] can be directly fed to the framework to generate one models for each of the 773 bacteria they identified. In a more general way, ETFL can assess the allowed expression profiles of any biological system amenable to genome-scale modeling, such as the metabolic engineering of biocatalysts, microbial communities, drug design, or personalized medicine.

# 3 Materials and Methods

## 3.1 Formulation of the expression problem

Several parameter values were taken from the BioNumbers database [31], and when used, their identification number as well as the original source from which the value was reported are specified.

### 3.1.1 Preliminaries, Conventions, and Notations

We will write the mass balances for the macromolecules with included concentration variables. Because we assume the cell is growing at a growth rate $\mu$, we must assume that the volume in which the mass balance is calculated varies.

The time derivative of the concentration of a macromolecule $G$ of concentration $C_G$ in the volume $V$, for a total mass $m_G$ in the cell, produced at a rate $v_G^{syn}$ and degraded at a rate $v_G^{deg}$, will be written:

$$\frac{dm_G}{dt} = C_G \frac{dV}{dt} + V \frac{dC_G}{dt} \tag{14}$$

$$= v_G^{syn} \cdot V - v_G^{deg} \cdot V. \tag{15}$$

We can combine equations 14 and 15 and divide by $V$ (necessarily non-zero) to write the time derivative of the concentration $C_G$:

$$\frac{dC_G}{dt} = v_G^{syn} - v_G^{deg} - \frac{1}{V}\frac{dV}{dt} \cdot C_G. \tag{16}$$

By definition, $\frac{1}{V}\frac{dV}{dt} = \mu$ is the growth rate of the cell, and we call the term $\mu \cdot C_G$ the dilution term, or $v_G^{dil}$, as per Fredrickson's work on formulating growth models [32]. It is a common assumption that the concentrations inside the cell remain time invariant (quasi-steady state assumption), effectively yielding the constraint:

$$v_G^{syn} - v_G^{deg} - \frac{1}{V}\frac{dV}{dt} \cdot C_G = 0. \tag{17}$$

It is also understood from the formulation of the FBA that adding a new reaction to the system, such as:

$$v_j : \qquad \eta_A^j A \rightarrow \eta_B^j B, \tag{18}$$

results in adding terms to the mass balances of $A$ and $B$ :

$$\frac{d\,[A]}{dt} = \ldots - \eta_A^j \cdot v_j, \tag{19}$$

$$\frac{d\,[B]}{dt} = \ldots + \eta_B^j \cdot v_j. \tag{20}$$

The further extension of this to reactions of n reactants to m products is trivial.

Several parameter values are taken from the BioNumbers database [31]. When used, we specify their identification number as well as the original source from which the value was reported. Finally, we will represent products between a parameter value and a variable by the symbol " $\cdot$ " and products between two variables by the symbol " $*$ ".

Hereafter, we propose a detailed top-down approach to formulate the constraints being built for ETFL, starting from the metabolite network and moving down to RNA synthesis.

### 3.1.2 Metabolic reactions

From FBA, the mass-balance relationship for metabolites can be written as:

$$S \cdot v = 0. \tag{FBA}$$

For the rest of the formulation, it is necessary to split the net flux $v$ from each reaction into its forward net component and backward net component:

$$v_j = v_j^+ - v_j^-, \qquad v_j^+, v_j^- \geq 0. \tag{21}$$

Biochemical reactions are catalyzed by enzymes. Each enzyme ($Enz_j$) of concentration $E_j$ can catalyze a flux $v_j$ subject to the enzyme capacity constraint, which is a function of its forward and backward catalytic rate constants $k_{cat}^{j,+}$ and $k_{cat}^{j,-}$:

$$0 \leq v_j^+ \leq k_{cat}^{j,+} E_j, \tag{22}$$

$$0 \leq v_j^- \leq k_{cat}^{j,-} E_j, \tag{23}$$

$$v_j^+ - k_{cat}^{j,+} E_j \leq 0, \tag{FC$_\mathrm{j}$}$$

$$v_j^- - k_{cat}^{j,-} E_j \leq 0. \tag{BC$_\mathrm{j}$}$$

The distinction between the bounds of the forward and backward net fluxes is important, as some enzymes have different catalytic activities depending on the direction of the flux.

### 3.1.3 Enzyme assembly

Each enzyme $Enz_j$ in concentration $E_j$ is subject to mass balance, which can be written:

$$\frac{d}{dt} E_j = v_j^{asm} - v_j^{deg} - v_j^{dil}, \tag{24}$$

which reads under quasi-steady state assumption (QSSA):

$$v_j^{asm} - v_j^{deg} - \mu * E_j = 0, \tag{EB$_\mathrm{j}$}$$

where $v_j^{asm}$ is the formation rate of the enzyme by the assembly of its constituent peptides, $v_j^{deg}$ is the degradation rate, $v_j^{dil}$ is the dilution rate, and $\mu$ is the growth rate

of the cell. The formation rate of the enzyme describes the assembly of free peptides, so it is necessary to add the peptide assembly reaction to the stoichiometric matrix:

$$v_j^{asm}: \qquad \sum_l \eta_l^j \cdot Pep_l \rightarrow Enz_j, \tag{25}$$

where $\eta_l^j$ is the stoichiometric coefficient of peptide $Pep_l$ for the formation of the complex of enzyme $Enz_j$.

### 3.1.4 Peptide synthesis

The synthesis of peptides consumes charged tRNAs, which are subsequently uncharged during the current peptide synthesis by a ribosome. The process consumes 2 GTP and releases 2 GDP and 2 Pi per amino acid:

$$\begin{aligned} v_l^{tsl}: \qquad & \sum_i \eta_{aa_i}^l \cdot tRNA_{aa_i}^{charged} + 2L_l^{aa} \cdot (GTP + H_2O) \\ & \rightarrow Pep_l + \sum_i \eta_{aa_i}^l \cdot tRNA_{aa_i}^{uncharged} + 2L_l^{aa} \cdot \left(GDP + Pi + H^+\right), \end{aligned} \tag{26}$$

where $aa_i$ denotes the $i^{\text{th}}$ amino acid, $\eta_{aa_i}^l$ its relative stoichiometric coefficient (count) in the sequence of $Pep_l$, $tRNA_{aa_i}^*$ the (un)charged tRNAs for each amino acid, and $L_l^{aa} = \sum_i \eta_{aa_i}^l$ is the length of the amino acid sequence of $Pep_l$.

As explained in section 1.1, this reaction adds a supplementary term in the mass balances of the metabolites (GTP, GDP, Pi, $H_2O$, $H^+$), the peptide, and the tRNAs (see 3.1.9 for the latter). This term is what connects the expression requirements to the metabolic network defined in the FBA.

In particular, the peptide concentrations obey the mass-balance equation:

$$\frac{d}{dt} Pep_l = v_l^{tsl} - \sum_j \eta_l^j \cdot v_j^{asm} - v_l^{deg} - v_l^{dil}. \tag{27}$$

We assume in the current model that the assembly rates are much faster than dilution and degradation, and thus simplify this mass balance to:

$$\frac{d}{dt} Pep_l = v_l^{tsl} - \sum_j \eta_l^j \cdot v_j^{asm}, \tag{28}$$

which, under QSSA, can be written:

$$v_l^{tsl} - \sum_j \eta_l^j \cdot v_j^{asm} = 0. \tag{PB$_\text{l}$}$$

In this context, the peptides are treated just like regular metabolites in the system. This assumption in $\text{PB}_\text{l}$ can be relaxed without a loss of generality by introducing a dilution and a degradation term, thus introducing a bilinearity.

Finally, we must take into account the actual degradation reactions. We add the ideal degradation reaction to the system:

$$v_j^{deg}: \quad Enz_j \rightarrow \sum_i \eta_{aa_i}^j \cdot aa_i + L_j^{aa} \cdot H_2O. \tag{29}$$

For this, the degradation rate is known:

$$v_j^{deg} - k_{deg}^j \cdot E_j = 0, \tag{ED$_\text{j}$}$$

where $k_{deg}^j$ is the degradation rate constant of the enzyme.

### 3.1.5 Ribosome synthesis and utilization

The peptides are the product of a translation reaction that is catalyzed by a ribosome. As we did with the catalytic constraints for general biochemistry reactions, we can apply the ribosome maximum catalytic rate as an upper bound to its translation rate $v_l^{tsl}$:

$$v_l^{tsl} - \frac{k_{cat}^{rib}}{L_l^{aa}} R_l \leq 0, \tag{TR2$_l$}$$

where $k_{cat}^{rib}$ is the maximum ribosomal translation rate constant ($10 - 12\,\text{aa.s}^{-1}$ for *E. coli* , BioNumbers ID [BNID] 100059 [33]), $L_l^{aa}$ is the amino acid length of the peptide $l$, and $R_l$ is the quantity of ribosomes assigned to the translation of this peptide. This way, the ratio $R_l/Pep_l$ is effectively the average polysome size translating the peptide $l$.

Like any other enzyme, ribosomes verify the mass balance:

$$v_{rib}^{asm} - v_{rib}^{deg} - \mu * E_{rib} = 0. \tag{$EB_{rib}$}$$

$E_{rib}$ denotes the total quantity of ribosomes in a cell. It accounts for $R_l$, the ribosomes assigned to the translation of $Pep_l$, as well as the free ribosomes in the cell, $R_F$. We can then write the total ribosome capacity constraint:

$$\sum_l R_l + R_F - E_{rib} = 0. \tag{TC2}$$

If we know the ratio $\rho$ of free versus occupied ribosomes, we can enforce it:

$$R_F - (1 - \rho)\, E_{rib} = 0. \tag{RR}$$

Finally, the ribosome differs from other enzymes in that it takes ribosomal peptides $rPep_l$ as well as ribosomal RNA $rRNA_l$ for its assembly. Hence, its assembly reaction is:

$$v_{rib}^{asm}: \qquad \sum_l \eta_{rPep_l}^{rib} \cdot rPep_l + \sum_l \eta_{rRNA_l}^{rib} \cdot rRNA_l \rightarrow Rib. \tag{30}$$

As explained earlier, the stoichiometric coefficients $\eta_\star^{rib}$ will appear in the mass balances of each of the compounds of the reaction.

### 3.1.6 RNA Polymerase

The transcription is catalyzed by RNA polymerase (RNAP). For each transcription of mRNA, we can put an upper bound on the transcription rate $v_l^{tcr}$ in the same way as for translation:

$$v_l^{tcr} - \frac{k_{cat}^{RNAP}}{L_l^{nt}} P_l \leq 0, \tag{TR1$_l$}$$

where $L_l^{nt}$ is the length in nucleotides of the mRNA sequence, $k_{cat}^{RNAP}$ is the catalytic rate constant of RNAP ($85\,\text{nt.s}^{-1}$ for *E. coli* , BNID 100060 [33]), and $P_l$ the concentration of RNAP assigned to the transcription of this mRNA.

RNAP is an enzyme, and hence it also satisfies mass balance:

$$v_{RNAP}^{asm} - v_{RNAP}^{deg} - \mu * E_{RNAP} = 0, \tag{$EB_{RNAP}$}$$

where $E_{RNAP}$ is the total amount of RNAP, which also accounts for free RNAP $P_F$, and follows:

$$\sum_l P_l + P_F - E_{RNAP} = 0. \tag{TC1}$$

As we did with the ribosomes, if we know the ratio of occupied RNAP, $\pi$, we can enforce it:

$$P_F - (1-\pi)\, E_{RNAP} = 0. \tag{PR}$$

### 3.1.7 mRNA synthesis and mass balance

During the translation, an mRNA is read to produce a peptide. mRNAs are subject the same mass-balance constraints:

$$v_l^{tcr} - v_l^{deg} - \mu * F_l = 0, \tag{MB$_l$}$$

where $F_l$ is the total concentration of the $l^{\text{th}}$ mRNA ($mRNA_l$), $v_l^{deg}$ is its degradation rate, and $v_l^{tcr}$ is its transcription (synthesis) rate. $F_l$ is to $mRNA_l$ what $E_j$ is to $Enz_j$ –the concentration variable that represents the macromolecule in the EP. The transcription reaction is modeled as follows:

$$\begin{aligned} v_l^{tcr}: \qquad & \eta_A^l \cdot ATP + \eta_U^l \cdot UTP + \eta_C^l \cdot CTP + \eta_G^l \cdot GTP \\ & \rightarrow \left(\eta_A^l + \eta_U^l + \eta_C^l + \eta_G^l\right) PPi + mRNA_l. \end{aligned} \tag{31}$$

Again, the stoichiometric coefficients will appear in the mass balances of each of the metabolites and macromolecules involved.

We must also take into account the relationship between ribosome assignment and mRNA concentration. On each strand of $mRNA_l$, there can be only a finite number $\rho_l$ of ribosomes translating at the same time. This number is given by the ratio of the footprint size of the ribosome $L_{rib}^{nt}$ and the length of the mRNA strand $L_l^{nt}$. This effectively yields the number of ribosomes that can be present at the same time on a given mRNA strand:

$$\rho_l = \frac{L_l^{nt}}{L_{rib}^{nt}}. \tag{32}$$

For *E. coli* , $L_{rib}^{nt}$ is approximately 20 nm (BNID 102320 [34], 100121 [35]), which amounts to approximately 60 base pairs (the length of a nucleotide is approximately 0.3 nm; BNID 103777 [36]). From there we can get the additional constraint:

$$R_l \leq \rho_l \cdot F_l, \tag{33}$$

$$R_l - \frac{L_l^{nt}}{L_{rib}^{nt}} F_l \leq 0. \tag{EX$_l$}$$

Finally, we must take into account the actual degradation reactions. We consider perfect degradation for mRNAs:

$$v_l^{deg}: \qquad mRNA_l \rightarrow \eta_A^l \cdot AMP + \eta_u^l \cdot UMP + \eta_C^l \cdot CMP + \eta_G^l \cdot GMP. \tag{34}$$

And, again, we know the degradation rates:

$$v_l^{deg} - k_{deg}^l \cdot F_l = 0. \tag{MD$_l$}$$

### 3.1.8 rRNA synthesis and mass balance

rRNAs are used in the ribosome assembly reaction. According to the definition of $v_{rib}^{asm}$ in 3.1.5, their mass balance can be written:

$$\frac{d}{dt}\left[rRNA_l\right] = 0 = v_{rRNA_l}^{tcr} - v_{rib}^{asm} - v_{rRNA_l}^{deg} - v_{rRNA_l}^{dil}. \tag{35}$$

We neglect their dilution and degradation under the hypothesis that free rRNAs are scarce and stable [37]. Thus, their mass balance in the model reads:

$$v_{rRNA_l}^{tcr} - v_{rib}^{asm} = 0. \tag{$\mathrm{RB_{rRNA_l}}$}$$

### 3.1.9 tRNA synthesis and mass balance

tRNAs are produced with a charging reaction and consumed by peptide synthesis. We use the following charging reaction:

$$\begin{aligned} v_{aa_i}^{charging}: \qquad & aa_i + tRNA_{aa_i}^{uncharged} + ATP + 2H_2O \\ & \rightarrow tRNA_{aa_i}^{charged} + AMP + 2H^+. \end{aligned} \tag{36}$$

Once again, the stoichiometric coefficients of each reactant will appear in the stoichiometric matrix in the column corresponding to this reaction.

Since tRNAs are relatively stable molecules [37], we neglect their degradation. Let $T_{aa_i}^u$ (resp. $T_{aa_i}^c$) represent $\left[tRNA_{aa_i}^{uncharged}\right]$ (resp. $\left[tRNA_{aa_i}^{charged}\right]$). Then, we can write the following constraints:

$$\frac{d}{dt}T_{aa_i}^u = 0 = -v_{aa_i}^{charging} + \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^u, \tag{37}$$

$$\frac{d}{dt}T_{aa_i}^c = 0 = v_{aa_i}^{charging} - \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^c. \tag{38}$$

$$-v_{aa_i}^{charging} + \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^u = 0, \tag{$\mathrm{TB^u_{aa_i}}$}$$

$$v_{aa_i}^{charging} - \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^c = 0. \tag{$\mathrm{TB^c_{aa_i}}$}$$

## 3.2 Thermodynamics-based constraints

Thermodynamics flux analysis (TFA) [2,3] imposes constraints on a FBA problem to couple reaction directionality to the standard free energy of reactions and metabolite concentrations. We also introduce constraints that couple the sign of the Gibbs energy of a reaction to its directionality through the use of integer variables and a mixed-integer linear coupling formulation. This framework reduces the feasible flux space and improves the predictive power of FBA by removing thermodynamically invalid flux profiles.

Considering $c_i$ is the concentration of $i^{\mathrm{th}}$ metabolite, we define $C_i$ as its scaled logarithm with respect to $c_0$ so that in standard conditions $c_0 = 1\,\mathrm{M}$:

$$\forall i, \qquad C_i = \ln\left(\frac{c_i}{c_0}\right). \tag{39}$$

We use the $\Delta_f G_i'^{\circ}$ to calculate the $\Delta_r G_j'^{\circ}$ the standard Gibbs energy of formation in solution of the $i^{\text{th}}$ metabolite and Gibbs energy in solution of the $j^{\text{th}}$ reaction, respectively, and we obtain the additional variables:

$$C_i^{min} \leq C_i \leq C_i^{max}, \tag{40}$$
$$\Delta_r G_{j,min}'^{\circ} \leq \Delta_r G_i'^{\circ} \leq \Delta_r G_{j,max}'^{\circ}, \tag{41}$$
$$\Delta_r G_{j,min}' \leq \Delta_r G_i' \leq \Delta_r G_{j,max}'. \tag{42}$$

The concentration variables are bounded by experimental measurements or physiological assumptions, and the standard Gibbs energies are bounded by the measurement or estimation error. Since the net flux of each reaction has already been split between forward flux ($v_j^+$) and backward flux ($v_j^-$), (see Eq. 21), we can directly add the constraints described in [2]:

$$\Delta_r G_j' - \mathrm{RT} \sum_{i=1}^{m} \eta_i^j C_i - \Delta_r G_i'^{\circ} = 0, \tag{43}$$
$$\Delta_r G_j' - K + K \cdot b_j^+ \leq 0, \tag{44}$$
$$-\Delta_r G_j' - K + K \cdot b_j^- \leq 0, \tag{45}$$
$$v_j^{\pm} - K \cdot b_j^{\pm} \leq 0, \tag{46}$$
$$b_j^+ + b_j^- \leq 1. \tag{47}$$

R denotes the ideal gas constant, T is the temperature in Kelvin, and $\eta_i^j$ represents the stoichiometry of the metabolite $i$ in the reaction $j$. $K$ is a big-M constant (bigger than all upper bounds), and $b_j^{\pm}$ are binary variables. Eq. 43 defines the actual Gibbs energy of the reaction as a function of its standard Gibbs energy and the scaled logarithms of metabolite concentrations. Eq. 44 and Eq. 45 ensure that $\Delta_r G_j' \leq 0 \iff b_i^+ = 1$ and $\Delta_r G_j' \geq 0 \iff b_i^- = 1$. These binary variables are used to block flux in Eq. 46 if the thermodynamics do not favor it. Finally, Eq. 47 is added to enforce that only one direction is chosen.

## 3.3 Petersen linearization

After discretization of the growth rate, the dilution term for the macromolecule $G_\star$ will consist of a sum of products of the binary variables $\delta_s$ and the continuous variable $G_\star$. We can use the Petersen linearization scheme [13] to transform this product into an equivalent system of one new variable and three new constraints:

$$\begin{aligned} z_\star^s &= \delta_s * G_\star, \\ &\iff \begin{cases} G_\star + M \cdot \delta_s - M \leq z_\star^s \leq M \cdot \delta_s, \\ z_\star^s \leq G_\star \end{cases}, \\ &\iff \begin{cases} G_\star + M \cdot \delta_s - z_\star^s \leq M, \\ z_\star^s - M \cdot \delta_s \leq 0, \\ z_\star^s - G_\star \leq 0. \end{cases} \end{aligned} \tag{48}$$

With this method, we can directly reformulate generalized mass balances as described in Eq. 5 for mRNAs, enzymes, uncharged tRNAs, and charged tRNAs:

$$v_j^{asm} - v_j^{deg} - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N}\overline{\mu} \cdot z_j^s = 0, \tag{EB$'_{\mathrm{j}}$}$$

$$v_l^{tcr} - v_l^{deg} - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N}\overline{\mu} \cdot z_l^s = 0, \tag{MB$'_{\mathrm{l}}$}$$

$$-v_{aa_i}^{charging} + \sum_i \eta_{aa_i}^l \cdot v_l^{tsl} - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N}\overline{\mu} \cdot z_{aa_i}^{u,s} = 0, \tag{TB$'^{\mathrm{u}}_{\mathrm{aa_i}}$}$$

$$v_{aa_i}^{charging} - \sum_i \eta_{aa_i}^l \cdot v_l^{tsl} - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N}\overline{\mu} \cdot z_{aa_i}^{c,s} = 0. \tag{TB$'^{\mathrm{c}}_{\mathrm{aa_i}}$}$$

And we get the additional linearization constraints:

$$\sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N}\overline{\mu} \leq \overline{\mu}, \tag{$GR$}$$

$$-\frac{\overline{\mu}}{2N} \leq \mu - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N}\overline{\mu} \leq \frac{\overline{\mu}}{2N}. \tag{GC}$$

## 3.4 Data

### 3.4.1 Degradation rate constants $k_{deg}$ (proteins, mRNA)

mRNA degradation rates were taken from Bernstein *et al.* [38]. We converted the reported half lives into rate constants using the classical relationship $k = \frac{log(2)}{t_{1/2}}$. Enzyme degradation rates were approximated to 5% per hour, as per a report indicating this rate was between 2% and 7% per hour in [39].

### 3.4.2 Catalytic rate constants $k_{cat}$ (homomers, other complexes)

Catalytic rate constants $k_{cat}^j$ were obtained from Davidi *et al.* [40] for homomer enzymes. Complex formation reactions for non-homomer enzymes were taken from the supplementary information of O'Brien *et al.* [7] and Lloyd *et al.* [8]. EC numbers were obtained from BiGG [41] and the iJO1366 publication [12]. Their corresponding $k_{cat}$ values were assigned using conservative (max) values from SabioRK.

### 3.4.3 Peptide composition of enzymes

Homomer compositions were obtained from Davidi *et al.* [40]. Other peptide compositions of enzymes were taken from the supplementary information of O'Brien *et al.* [7] and Lloyd *et al.* [8]. Additional information was obtained from the Metacyc/Biocyc database [42, 43] using specialized SmartTables queries [44].

## 3.5 Modeling

### 3.5.1 Model modifications

The initial model was subjected to minor changes to accommodate for ETFL modeling. In particular, we added:

- Selenocysteine as a metabolite.

- Cysteine to selenocysteine conversion as a pseudo reaction.
- tRNA charging reactions.

We also modified the biomass reaction by removing its nucleotide and amino acid components, since they are already taken into account by the expression problem as explained in 1.1.1.

### 3.5.2 Enzyme estimation

Given a reaction in the model, if no enzyme is supplied but the reaction possesses a gene reaction rule, it is possible to infer an enzyme from it. The rule expression is expanded, and each term separated by an `OR` boolean operator is interpreted as an isozyme, while terms separated by an `AND` boolean operator are interpreted as unit peptide stoichiometric requirements. The enzyme is then assigned an average catalytic rate constant and degradation rate constant.

### 3.5.3 Essentiality analysis

For increased performance, the essentiality analysis was cast into a feasibility problem. We put a lower bound on growth equal to 10% of the predicted ETFL growth and set the objective to 0. With this method, essential genes will cause the problem to be infeasible, while non-essential genes will return a feasible solution satisfying at least 10% of the growth. This method achieved up to a 5-fold reduction in solving time on the most complex models.

## 3.6 Computation

### 3.6.1 Implementation

The code has been implemented as a plug-in to pyTFA [45], a Python implementation of the TFA method. It uses COBRApy [46] and Optlang [47] as a backend to ensure compatibility with several open-source (GLPK, scipy, ...) as well as commercial (CPLEX, Gurobi, ...) solvers. The code is freely available under the APACHE 2.0 license at `https://github.com/EPFL-LCSB/etfl`. We rely on the Python package Biopython [48] for transcribing and translating sequences of nucleotides and amino acids.

### 3.6.2 Hardware

Computations were done on a 64-bit Ubuntu 18.04.1 LTS (Bionic Beaver); 2 × Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20 GHz (8 cores, 16 threads per socket); 4 ×16 Go @ 2400 MHz RAM. Code was run on Python 3.6 on Docker (18.09.0) containers based on the official python 3.6-stretch container, available on ETFL GitHub.

# Supporting information

**S1 Nondimensionalization. Derivation and formulation.** Details on the variables and constraint transformations to scale the model.

**S2 Example EP constraint matrix. Representation of the constraint matrix of the EP for a vETFL of iJO1366.** Colored cells represent non-zero blocks. Uncolored cells are zero blocks.

**S3 SOP for creating an ETFL model. Tips and Prerequisites.** List of required and optional inputs to transform a COBRA model into an ETFL model

## Acknowledgements

The authors would like to thank Kaycie Butler for her valuable input on the wording and structure of this manuscript. This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 722287.

## References

1. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature biotechnology. 2017;35(1):81.

2. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-based metabolic flux analysis [Journal Article]. Biophysical Journal. 2007;92(5):1792–1805. Available from: `<GotoISI>://WOS:000244373800034`.

3. Soh KC, Hatzimanikatis V. Constraining the flux space using thermodynamics and integration of metabolomics data [Journal Article]. Methods Mol Biol. 2014;1191:49–63. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/25178783`.

4. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabasi AL, et al. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity [Journal Article]. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(31):12663–12668. Available from: `<GotoISI>://WOS:000248603900015`.

5. Sanchez BJ, Zhang C, Nilsson A, Lahtvee PJ, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints [Journal Article]. Molecular Systems Biology. 2017;13(8). Available from: `<GotoISI>://WOS:000406943100001`.

6. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, et al. In silico method for modelling metabolism and gene product expression at genome scale. Nature communications. 2012;3:929.

7. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction [Journal Article]. Mol Syst Biol. 2013;9:693. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/24084808`.

8. Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, et al. COBRAme: A computational framework for genome-scale models of metabolism and gene expression [Journal Article]. PLoS computational biology. 2018;14(7):e1006302.

9. Yang L, Ma D, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO. solveME: fast and reliable solution of nonlinear ME models [Journal Article]. BMC Bioinformatics. 2016;17(1):391. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/27659412`.

10. Ma D, Yang L, Fleming RM, Thiele I, Palsson BO, Saunders MA. Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. Scientific reports. 2017;7:40863.

11. Neidhardt FC, Curtiss R. Escherichia coli and Salmonella: cellular and molecular biology. vol. 2. ASM press Washington, DC:; 1999.

12. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011 [Journal Article]. Molecular Systems Biology. 2011;7. Available from: `<GotoISI>://WOS:000296652600001`.

13. Petersen CC. A note on transforming the product of variables to linear form in linear programs [Journal Article]. Diskussionspapier, Purdue University. 1971;.

14. Glover F. Improved Linear Integer Programming Formulations of Nonlinear Integer Problems [Journal Article]. Management Science. 1975;22(4):455–460. Available from: `https://pubsonline.informs.org/doi/abs/10.1287/mnsc.22.4.455`.

15. Hatzimanikatis V, Floudas CA, Bailey JE. Analysis and design of metabolic reaction networks via mixed-integer linear optimization [Journal Article]. AIChE Journal. 1996;42(5):1277–1292. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.690420509`.

16. Hatzimanikatis V, Floudas CA, Bailey JE. Optimization of regulatory architectures in metabolic reaction networks. Biotechnology and bioengineering. 1996;52(4):485–500.

17. Pramanik J, Keasling J. Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnology and bioengineering. 1997;56(4):398–421.

18. Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, et al. SABIO-RK—database for biochemical reaction kinetics. Nucleic Acids Research. 2012;40(D1):D790–D796. Available from: `http://dx.doi.org/10.1093/nar/gkr1046`.

19. McCloskey D, Gangoiti JA, King ZA, Naviaux RK, Barshop BA, Palsson BO, et al. A Model-Driven Quantitative Metabolomics Analysis of Aerobic and Anaerobic Metabolism in E. coli K-12 MG1655 That Is Biochemically and Thermodynamically Consistent [Journal Article]. Biotechnology and Bioengineering. 2014;111(4):803–815. Available from: `<GotoISI>://WOS:000332008500017`.

20. Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J, Milo R. Visual account of protein investment in cellular functions [Journal Article]. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(23):8488–8493. Available from: `<GotoISI>://WOS:000336976000054`.

21. Otto A, Bernhardt J, Meyer H, Schaffer M, Herbst FA, Siebourg J, et al. Systems-wide temporal proteomic profiling in glucose-starved Bacillus subtilis [Journal Article]. Nature Communications. 2010;1. Available from: `<GotoISI>://WOS:000288225300005`.

22. Schellenberger J, Palsson BO. Use of randomized sampling for analysis of metabolic networks [Journal Article]. J Biol Chem. 2009;284(9):5457–61. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/18940807`.

23. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0 [Journal Article]. Nature Protocols. 2011;6(9):1290–1307. Available from: `<GotoISI>://WOS:000295362900003`.

24. Megchelenbrink W, Huynen M, Marchiori E. optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks [Journal Article]. PLoS One. 2014;9(2):e86587. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/24551039`.

25. Pandey V, Hadadi N, Hatzimanikatis V. Enhanced flux prediction by integrating relative expression and relative metabolite abundance into thermodynamically consistent metabolic models. bioRxiv. 2018;p. 481499.

26. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26(24):3140–3142.

27. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS computational biology. 2008;4(5):e1000082.

28. Pandey V, Hatzimanikatis V. Investigating the deregulation of metabolic tasks via Minimum Network Enrichment Analysis (MiNEA) as applied to nonalcoholic fatty liver disease using mouse and human omics data. bioRxiv. 2018;p. 402222.

29. Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences. 2002;99(23):15112–15117.

30. Mahadevan R, Edwards JS, Doyle r F J. Dynamic flux balance analysis of diauxic growth in Escherichia coli [Journal Article]. Biophys J. 2002;83(3):1331–40. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/12202358`.

31. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers–the database of key numbers in molecular and cell biology [Journal Article]. Nucleic Acids Res. 2010;38(Database issue):D750–3. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/19854939`.

32. Fredrickson A. Formulation of structured growth models. Biotechnology and bioengineering. 1976;18(10):1481–1486.

33. Bremer H, Dennis PP. Modulation of chemical composition and other parameters of the cell by growth rate [Journal Article]. Escherichia coli and Salmonella: cellular and molecular biology. 1996;2(2):1553–69.

34. Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM, et al. Structures of the bacterial ribosome at 3.5 A resolution [Journal Article]. Science. 2005;310(5749):827–34. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/16272117`.

35. Zhu J, Penczek PA, Schroder R, Frank J. Three-dimensional reconstruction with contrast transfer function correction from energy-filtered cryoelectron micrographs: procedure and application to the 70S Escherichia coli ribosome [Journal Article]. J Struct Biol. 1997;118(3):197–219. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/9169230`.

36. Gilbert R. Physical biology of the cell, by Rob Phillips, Jane Kondev and Julie Theriot. Taylor & Francis; 2009.

37. Neidhardt FC. In: The regulation of RNA synthesis in bacteria. vol. 3. Elsevier; 1964. p. 145–181.

38. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays [Journal Article]. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(15):9697–9702. Available from: `<GotoISI>://WOS:000177042400020`.

39. Nath K, Koch AL. Protein degradation in Escherichia coli II. Strain differences in the degradation of protein and nucleic acid resulting from starvation [Journal Article]. Journal of Biological Chemistry. 1971;246(22):6956–6967.

40. Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, et al. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro k(cat) measurements [Journal Article]. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(12):3401–3406. Available from: `<GotoISI>://WOS:000372488200074`.

41. King ZA, Drager A, Ebrahim A, Sonnenschein N, Lewis NE, Palsson BO. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways [Journal Article]. Plos Computational Biology. 2015;11(8). Available from: `<GotoISI>://WOS:000360824500006`.

42. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic acids research. 2007;36(suppl_1):D623–D631.

43. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, et al. EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic acids research. 2005;33(suppl_1):D334–D337.

44. Travers M, Paley SM, Shrager J, Holland TA, Karp PD. Groups: knowledge spreadsheets for symbolic biocomputing. Database. 2013;2013.

45. Salvy P, Fengos G, Ataman M, Pathier T, Soh KC, Hatzimanikatis V. pyTFA and matTFA: A Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis [Journal Article]. Bioinformatics. 2018;.

46. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python [Journal Article]. Bmc Systems Biology. 2013;7. Available from: `<GotoISI>://WOS:000323121200001`.

47. Jensen K, Cardoso J, Sonnenschein N. Optlang: An algebraic modeling language for mathematical optimization [Journal Article]. Journal of Open Source Software. 2016;.

48. Dalke A, Wilczynski B, Chapman BA, Cox CJ, Kauff F, Friedberg I, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 03;25(11):1422–1423. Available from: `https://dx.doi.org/10.1093/bioinformatics/btp163`.