A hybrid of metabolic flux analysis and Bayesian factor modeling for multi-omic temporal pathway activation

Claudio Angione,[†,‡] Naruemon Pratanwanich,[†,‡] and Pietro Lió[*,†]

Computer Laboratory - University of Cambridge, UK

E-mail: pl219@cam.ac.uk

Abstract

The growing availability of multi-omic data provides a highly comprehensive view of cellular processes at the levels of mRNA, proteins, metabolites, and reaction fluxes. However, due to probabilistic interactions between components depending on the environment and on the time course, casual, sometimes rare interactions may cause important effects in the cellular physiology. To date, interactions at the pathway level cannot be measured directly, and methodologies to predict pathway cross-correlations from reaction fluxes are still missing. Here, we develop a multi-omic approach of flux-balance analysis combined with Bayesian factor modeling with the aim of detecting pathway cross-correlations and predicting metabolic pathway activation profiles. Starting from gene expression profiles measured in various environmental conditions, we associate a flux rate profile with each condition. We then infer pathway cross-correlations and identify the degrees of pathway activation with respect to the conditions and time

[*] To whom correspondence should be addressed

[†] Computer Laboratory - University of Cambridge, UK

[‡] Joint first authors

course using Bayesian factor modeling. We test our framework on the most recent metabolic reconstruction of *Escherichia coli* in both static and dynamic environments, thus predicting the functionality of particular groups of reactions and how it varies over time. In a dynamic environment, our method can be readily used to characterize the temporal progression of pathway activation in response to given stimuli.

Keywords

Multi-omics, Flux Balance Analysis, Escherichia coli, Bayesian factor modeling, Pathway correlation, Temporal pathway activation

1 Introduction

The recent availability of high-throughput data regarding multiple layers of biological organization ("omics") allows mapping cellular processes at the levels of mRNA, proteins, and metabolites. Analogously, the growing number of defined pathways, where reactions are classified into groups, allows us to better understand the particular functionality achieved by a series of reactions. To date, the study of interactions between pathways taken as single entities has been already applied to genes through gene expression analysis (*1*). Such interactions, also known as *pathway cross-correlations*, are important to produce appropriate response to external stimuli, and are assumed to be the underlying mechanism describing the response to dynamic environments. This suggests that studying biological systems requires a holistic approach that takes the concerted activities of molecules into account (*2*). Previous studies inferred cross-correlations from gene expression data (*3–5*) and others from protein-protein networks (*6*). Another recent study applied a Bayesian network inference to identify causal relationships among the most influential reactions (*7*). However, methods to infer cross-correlations between pathways from reaction fluxes, and therefore making use of the metabolism and its reaction-pathway associations, are still missing.

In the last 25 years, high-quality genome-scale reconstructions of metabolic networks have been combined with constraint-based optimization in order to analyze microorganisms at steady state. To improve the predicting capability of a metabolic model, one can include multiple ‘omic’ layers, such as gene expression, codon usage, protein abundance, and the interaction between these layers. The interdependence among gene expression levels, protein production and growth rate has been analyzed thoroughly by Scott et al. (*8*), highlighting a linear relation between the RNA-protein ratio and the growth rate of the bacterium. Other methodologies regarding how to improve the model predictions by means of gene expression have been recently proposed (*9*).

Arguably, flux-balance analysis (FBA) is the most widely used constraint-based technique to predict flux distributions and network capabilities in large biochemical networks (*10*). FBA has proved useful thanks to its ability to handle large networks: it requires information about biochemical reactions and stoichiometric coefficients, but does not involve kinetic parameters. This makes it well suited to studies that enumerate and characterize perturbations such as different substrates or genetic interventions (e.g. knockouts) leading to obligatory coupling between the growth rate and the production of a desired metabolite (*11*). Recently, more than 1000 prokaryotic genomes have been fully sequenced, thus allowing FBA models to incorporate also information on enzymes and genome, including the relationships among genes, proteins and reactions (*GPR mapping*). To date, more than 90 genome-wide metabolic reconstructions have been published (*12*).

In this study, we propose a methodology to predict the cellular response to environmental conditions from a pathway-based perspective. The classification of metabolic reactions into pathways allows us to understand or predict the functionality of particular groups of reactions under given growth conditions. However, since the interactions at the pathway level cannot be measured directly, we propose to apply a hierarchical Bayesian framework, which supports latent variable models in order to take pathway information into account (*5*). We focus on the cellular activity of *Escherichia coli* on the genomic, fluxomic and pathway levels in different environmental settings by integrating an augmented metabolic model with a machine learning technique applied at the fluxomic level. Integrating a FBA model and a Bayesian factor model leads to determining the degree of metabolic pathway responsiveness and to detecting pathway cross-correlations, starting from gene expression profiles (Figure 1).

The aim of this paper is to combine a Bayesian machine learning technique and a multi-omic flux model augmented with gene expression profiles in order to integrate and analyze data representing heterogeneous biological levels of organization. Our method highlights complex interactions between components of the model at different layers. We start with the investigation of the genome scale model by using metabolic flux analysis in a bilevel setting, namely the maximization of growth rate and acetate production. Through a Bayesian factor approach, we detect pathway cross-correlations that are assumed to be a static, intrinsic property of *E. coli* underlying its response behavior. Furthermore, we infer pathway activation profiles as a bacterial response to an ensemble of environmental conditions. Finally, we use time series of gene expression profiles combined with our hybrid approach in order to investigate changing pathway responsiveness.

2 Flux-balance analysis with continuous gene expression

Flux-balance analysis (FBA) is a linear programming technique that models the steady state condition in a chemical reaction network (*13*). The combination of flux-balance constraints and capacity constraints on the metabolic fluxes is a system of linear homogeneous equations and inequalities, thus its solution space is a convex polyhedral cone representing the feasible flux distributions. The flux-balance constraint is represented by the equation $dX/dt = Sv = 0$, where X represents the vector of the concentrations of all metabolites of the network, S its stoichiometric matrix, and v the vector of flux rates. This constraint can be thought

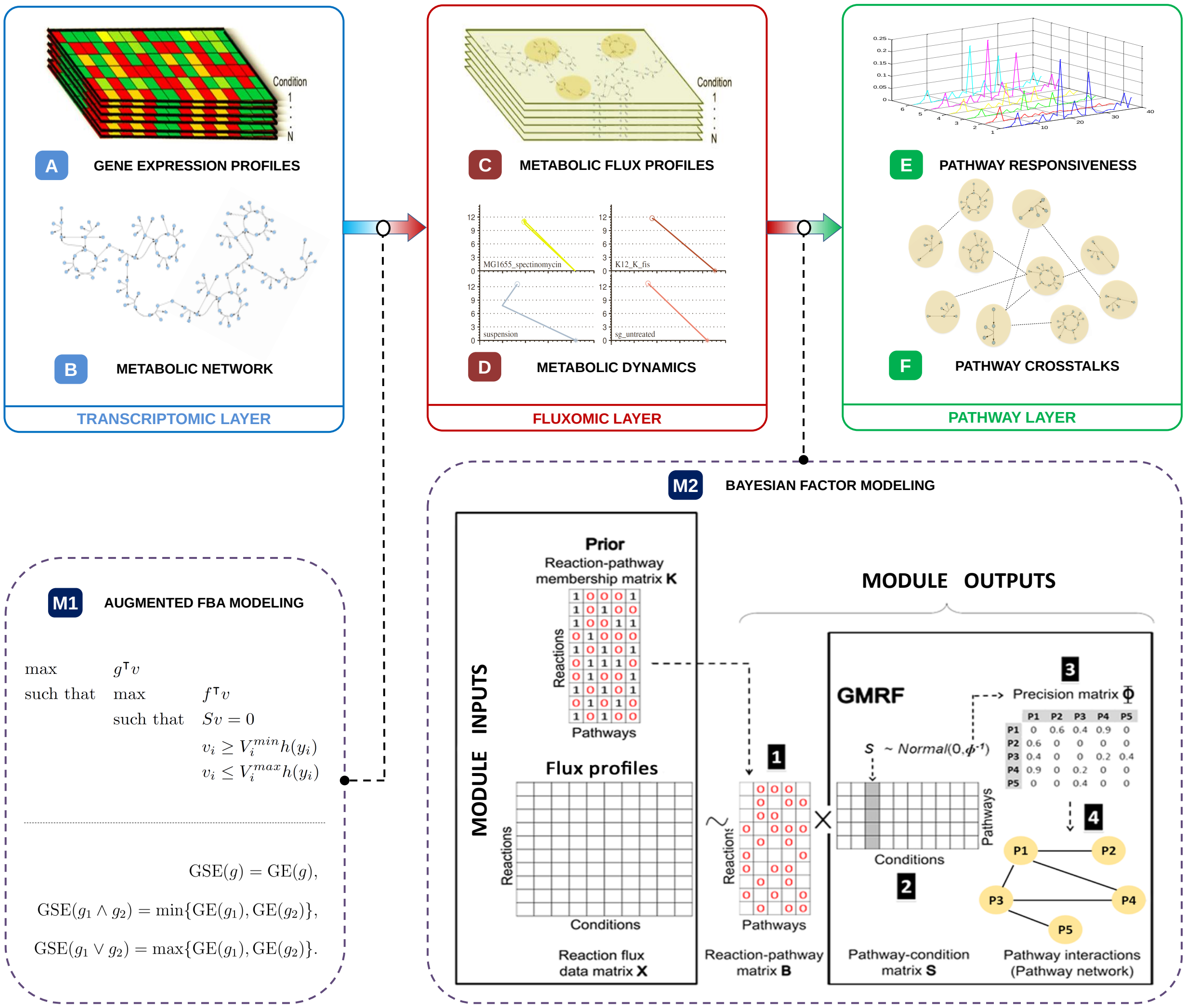

Figure 1: Our modeling framework combines Bayesian machine learning and metabolic flux modeling to analyze metabolic pathways from gene expression data. First, we propose an augmented FBA method (M1) to map gene expression data (A) on the metabolic network (B). The method includes a bilevel maximization framework in the case study of biomass-acetate objective space, therefore producing optimal flux profiles in different environmental conditions (C). Using a set of conditions with time-series gene expression profiles, we use our framework to elucidate the metabolism dynamics, involving rearrangements in the objective space during growth (D). Finally, we perform Bayesian factor modeling (M2) on the reaction flux distribution, by taking pre-defined reaction-pathway memberships as prior knowledge. This enables us to infer the pathway responsiveness to each environmental condition (E) and the cross-correlations between pathways (F), elucidating the underlying mechanisms of bacterial response to dynamically changing environments.

of as Kirchoff's laws applied to any node representing a metabolite in the network. The flux through a metabolite must be constant, namely the input flux equals the output flux. If one allows the concentrations of metabolites to increase linearly, the conditions become $dX/dt \geq 0$, which is equivalent to $Sv \geq 0$. This approach represents Von Neumann's optimal growth scenario, and may be useful when one has to ensure that some metabolites are available also outside of the chemical reaction network modeled (e.g. when modeling only a compartment of a larger organism (*14*), or when applying external optimization algorithms to maximize or minimize the metabolite concentration (*15*)).

To analyze the effects caused by the change of external media and conditions in which an *E. coli* strain was grown, we map each gene expression array (microarray profiles) to the acetate-biomass space of objective functions. We take these two fluxes as objectives because of the common assumption that microorganisms tend to optimize their metabolic network in order to maximize the growth rate, and possibly produce additional chemicals, in order to cope with multiple, sometimes conflicting, objectives to optimize simultaneously (*16*). As well as being an important target for biotechnology, with multiple industrial applications (*17*), acetate is central to many pathways in both aerobic and anaerobic *E. coli*. Being an intermediate metabolite, it is representative of processes not directly related to growth, and therefore it is highly indicative of metabolic flexibility for possible reorganizations that need to be performed during adaptations to environmental changes. When acetate is present at high levels, it inhibits cell growth and recombinant protein productivity (*18*).

The gene regulation process in bacteria is used to respond to the variations taking place in the metabolism or in the external environment. Here we take into account 466 *E. coli* Affymetrix Antisense2 microarray expression profiles collected in various media and conditions (*19*), such as low or high glucose, aerobic or anaerobic environment, pH changes, antibiotics, and heat shock. For the dynamical analysis of growth, we will consider 41 growth conditions, each of which has been sampled at four time steps (164 microarray profiles in total).

Each expression profile is mapped to the *E. coli* model (*20*), which we have augmented with a map from gene expression (GE) to constraints for metabolic fluxes. First, in order map a gene expression profile to a *gene set expression* (GSE) profile, we use recursively the following rules valid for the three basic cases of gene set:

$$\begin{aligned}
\text{Single gene:} \qquad & \mathrm{GSE}(g) = \mathrm{GE}(g), \\
\text{Enzymatic complex:} \qquad & \mathrm{GSE}(g_1 \wedge g_2) = \min\{\mathrm{GE}(g_1), \mathrm{GE}(g_2)\}, \\
\text{Isozymes:} \qquad & \mathrm{GSE}(g_1 \vee g_2) = \max\{\mathrm{GE}(g_1), \mathrm{GE}(g_2)\}.
\end{aligned} \tag{1}$$

Then, we solve the two-level maximization problem

$$\begin{aligned}
&\max && g^{\intercal} v \\
&\text{such that} && \max \ \left\{ f^{\intercal} v \,\middle|\, Sv = 0, v_i \geq V_i^{min} h(y_i), v_i \leq V_i^{max} h(y_i) \right\}
\end{aligned} \tag{2}$$

where V_i^{min} and $V_i^{max}$are the default lower- and upper-bound for each flux v_i, f and g are n-dimensional Boolean arrays that select the fluxes to be maximized (in this paper, only one flux is selected for each level of the maximization problem). The gene set expression $y_i = \mathrm{GSE}_i$ represents the "expression" of the ith reaction of the model. The map from each gene set to the associated flux upper- and lower-bound is defined as

$$h(x) = \frac{\gamma(1 + |log(x)|)^{\mathrm{sgn}(x-1)}}{\sigma_i^2} \tag{3}$$

(and $h(x) = \gamma/\sigma_i^2$ if $x = 1$), where $\mathrm{sgn}(x-1) = (x-1)/\,|x-1|$; γ is the weight for the variance σ^2, which is the variance of the gene set, computed from the variance of its genes using the same rules (1) used for the gene set expression (Figure 1-M1).

The importance of a gene - and therefore the ability to change the reaction flux of the corresponding reaction in the FBA model - is inversely proportional to the variance of that gene across all the experimental conditions. The idea underlying this assumption is that those genes whose expression is only slightly varied across conditions must be regarded as key genes for the organism (*21*). We adopt γ as a multiplicative factor for the inverse of the variance of each gene, representing the weight attributed to the variance as an indicator of the importance of a gene. Here, γ was chosen as the value that causes the smallest loss of information (quantified using a normalized root mean square error, NRMSE) when using the Bayesian factor model. However, if further experimental data is available (e.g. protein abundance, translation rate, codon usage or post-translational modifications), we expect this parameter to be varied individually for each reaction.

The reasons for choosing this mathematical structure are as follows. First, a recent model suggests a protein synthesis rate growing fast with increasing mRNA abundance, but decreasing the growth speed for high values of mRNA abundance (*22*). Second, there is empirical evidence that logarithmic maps are useful to map biological processes (*23*). Third, the approximation of this behavior with a logarithmic function simplifies the task of avoiding that unrealistically high values of measured gene expression levels are translated into overly weak constraints. The correlation between gene expression and metabolic phenotype is still a matter of debate, but recent evidence suggests that protein abundance is mainly determined by the transcript level (*24*, *25*). Therefore, we use the logarithmic map only to set constraints, while we solve the bilevel linear program (2) to find the final flux distribution under each condition. The solution of the bilevel problem is a pair representing the maximum natural objective (biomass) allowed by the constraints, and the maximum second objective (acetate) possible in the computed biomass-maximizing flux distribution. Each experimental condition is associated with a gene expression array, and therefore with a bilevel problem.

3 Pathway-based Bayesian factor modeling

Factor modeling is an unsupervised learning technique in a family of latent variable models, assuming that high-dimensional data are generated from the hidden lower-dimensional factors that are shared across data samples. Specifically, matrix factorization assumes that an observed $R \times C$ data matrix can be explained by two low-rank matrices with the dimension of $R \times P$ and $P \times C$ where $P << R, C$. Moreover, under Gaussian Markov Random Field (GMRF) properties, assuming a Gaussian distribution on the underlying factors in the matrix factorization process can capture the dependency between each element in P.

Here we regard pathways as the latent factors underlying bacterial flux responses. We assume that the observed flux rates arise from specific combinations of pathways that are activated by a stimulus. In particular, the values of a flux rate for each reaction depends on the degrees of pathway activation and on the association strength between that reaction and the activated pathways. Having obtained the $R \times C$ flux rate matrix from the bilevel FBA where R and C are the number of reactions and conditions respectively, we use the Bayesian matrix factorization modeling with GMRF to perform pathway analysis. Specifically, we decompose the flux rate data matrix into two low-rank matrices whose dimensions are $R \times P$ and $P \times C$, as well as a $P \times P$ correlation matrix, where P is the number of latent pathways. Note that, in general, we take a pathway into account if it is associated with at least one reaction in R. Figure 1-M2 demonstrates the factor model graphically and mathematically. On the basis of the rationale of matrix factorization, the $R \times P$ matrix represents the association strength between reaction fluxes and pathways, while the $P \times C$ matrix denotes the association strength between pathways and conditions, suggesting degrees of pathway responsiveness. In addition, the correlations between metabolic pathways are denoted by the $P \times P$ matrix.

Formally, let R, C, and P be the number of reactions, conditions, and pathways respectively. The flux data matrix $\mathbf{X} \in \mathbb{R}^{R \times C}$ is decomposed into two matrices: $\mathbf{X} \sim \mathbf{BS}$. The first matrix $\mathbf{B} \in \mathbb{R}^{R \times P}$ denotes the membership strength of reactions in each pathway. The second matrix $\mathbf{S} \in \mathbb{R}^{P \times C}$ corresponds to the degree of pathway responsiveness specific to each condition. We transform the pre-defined reaction-pathway memberships of the *E. coli* model into the binary matrix $\mathbf{K} \in \{0, 1\}^{R \times P}$ in order to guide the clustering of reactions into pathways in $\mathbf{B}$. On the basis of our assumption of pathway cross-correlations, we model pathway dependencies by assuming a Gaussian distribution on $\mathbf{S}$ with a zero mean and a precision (inverse covariance) matrix $\boldsymbol{\Phi} \in \mathbb{R}^{P \times P}$ from which the correlations between pathways are computed.

Based on the Gaussian Markov Random Field (GMRF) framework, the off-diagonal elements of the precision (inverse covariance) matrix can be interpreted as the partial correlations between any two random variables as follows (*26*). The zero pattern in the precision matrix encodes the independence relations of two variables generated by a Gaussian distribution conditioned on the other random variables. Therefore, the precision matrix can be used to form a pathway cross-correlation network, where nodes represent individual pathways and neighboring nodes represent highly correlated pathways. More precisely, since we assume that for any given condition c, the pathway responsiveness vector $s_c \sim \text{Normal}(0, \boldsymbol{\Phi}^{-1})$ is randomly drawn from a P-dimensional multivariate Gaussian with a zero mean μ and a precision matrix $\boldsymbol{\Phi}$, we can form an undirected graph $\mathcal{G}(\Phi) = (V, E)$ with vertices V corresponding to the random variables (pathways) $V_1, V_2, V_3, \ldots, V_P$ and edges E satisfying $(V_i, V_j) \in E$ if and only if $\phi_{ij} \neq 0$. If $\mathcal{N}(i) = j : (i, j) \in E$ denotes the set of neighboring nodes of s in the graph $\mathcal{G}$, the independence correlation of $V_i \perp V_u | V_{\mathcal{N}(i)}$ holds for any node $u \notin \mathcal{N}(i)$ that is not a neighbor of V_i. Moreover, the correlation strength between V_i and V_j is calculated as

$$Corr(V_i, V_j | V_{\setminus ij}) = \left| \frac{\phi_{ij}}{\sqrt{\phi_{ii}\phi_{jj}}} \right|, \tag{4}$$

where the subscript "$\setminus ij$" indicates all nodes except V_i and V_j. Consequently, the correlation strength will range from zero to one, indicating the weakest correlation (conditional independence) and strongest correlation respectively. The remarkable characteristic of the GMRF is that the conditional independence can be interpreted directly from the precision matrix. More importantly, it can encode any arbitrary structure of a graph.

Having observed the data matrix $\mathbf{X}$, the goal is to make inference on $\mathbf{S}, \mathbf{B}$, and $\mathbf{\Phi}$. Our interest for biological interpretation is only on $\mathbf{S}$ and $\mathbf{\Phi}$, which indicate the degree of pathway responsiveness to each condition and the cross-correlations between pathways, respectively. At each time step, we use the inferred cross-correlations as the underlying mechanism of the *E. coli* system to predict the temporal progression of pathway activation.

In order to infer all unknown variables, we apply a Bayesian approach starting from the construction of a full probabilistic model, in which all relevant entities (i.e. observed data, latent variables, and nuisance variables) are treated as random variables having uncertainties described by a probability distribution. To avoid the optimization of the parameters, we apply the Bayesian hierarchical modeling, in which each parameter is given a prior distribution with a set of fixed hyperparameters.

The model is mainly based on Gaussian distributions containing two parameters, a mean and a precision (an inverse variance). According to the matrix factorization method, $\mathbf{X}$ is modeled with a mean $\mathbf{BS}$ and a precision τ_ϵ (Equations (5) and (6)).

$$\mathbf{x}_{rc} = \sum_{p=1}^{P} b_{rp}s_{pc} + \epsilon = \mathbf{b}_r\mathbf{s}_c + \epsilon \; ; \quad \epsilon = \text{random noise, } \epsilon \in \mathbb{R}, \tag{5}$$

$$\epsilon \sim \text{Normal}(0, \tau_\epsilon^{-1}) \tag{6}$$

where $\mathbf{b}_r$ is the rth row of $\mathbf{B}$, and $\mathbf{s}_c$ is the cth column of $\mathbf{S}$. Equations (7) and (8) illustrate that the latent variables $\mathbf{B}$ and $\mathbf{S}$ are modeled with a zero mean because of the sparsity constraints. While elements in $\mathbf{B}$ are assumed to be independently distributed with a precision τ_B, elements in $\mathbf{S}$ are presumably correlated within each column with a precision matrix $\mathbf{\Phi}$, expressing the hypothesis of pathway cross-correlations.

$$b_{rp} = \begin{cases} 0, & \text{if } k_{rp} = 0 \\ \text{Normal}(b_{rp}|0, \tau_B^{-1}), & \text{if } k_{rp} = 1 \end{cases} \tag{7}$$

$$\mathbf{s}_c \sim \text{GMRF}(\mathbf{0}, \mathbf{\Phi}^{-1}) \tag{8}$$

All parameters (τ_ϵ, τ_B, and $\mathbf{\Phi}$) are given conjugate priors with their own fixed hyperparameters ($\alpha_\epsilon, \beta_\epsilon, \alpha_B, \beta_B, \nu, \mathbf{\Psi}$), which are a Gamma distribution for τ_ϵ and τ_B and a Wishart distribution for $\mathbf{\Phi}$ (Equation (9),(10) and (11)).

$$\tau_\epsilon \sim \text{Gamma}(\alpha_\epsilon, \beta_\epsilon) \tag{9}$$

$$\tau_B \sim \text{Gamma}(\alpha_B, \beta_B) \tag{10}$$

$$\mathbf{\Phi} \sim \text{Wishart}(\nu, \mathbf{\Psi}) \tag{11}$$

The set of Equation (5) - (11) entirely specifies each model entity with its probability distribution (*5*). The use of conjugate priors makes the inference computable analytically.

Upon the model specification, inference of unknown quantities can be made through the computation of the posterior distribution, which is composed by the probability conditioned on the observed data, known as likelihood, and its prior. Due to the conjugacy, we apply a Gibbs sampling algorithm, where each unknown variable is alternately sampled until the convergence is satisfactory. In order to reduce auto-correlation between samples, every ten iterations, we collect the samples of the matrix $\mathbf{S}$ for the comparison between conditions over time, and those of $\mathbf{\Phi}$ for the analysis of pathway cross-correlations.

4 Results and discussion

We first applied FBA modeling to map 466 gene expression data profiles, plus 164 time-series data profiles, into 2583 reaction flux rates of the *E. coli* metabolic network, subject to the maximization of acetate and biomass production. Next, we performed Bayesian factor modeling on those reaction fluxes by taking pre-defined reaction-pathway memberships as prior knowledge to infer the responsiveness degree of the 37 pathways of the *E. coli* model.

4.1 Sensitivity and mean square error analysis

Based on a normalized root mean square error (NRMSE) as shown in Equation (12), we performed a convergence analysis indicating how well the model fits the data. We compute the error as

$$\text{NRMSE} = \frac{\sqrt{\sum_{r,c} (x_{rc} - \mathbf{b}_r \mathbf{s}_c)^2 / (R \times C)}}{x_{max} - x_{min}}, \tag{12}$$

where x_{max} and x_{min} are the highest and lowest values in $\mathbf{X}$.

More specifically, the NRMSE indicates how much the estimated flux rates inferred from the model assumption according to Equation (5) are deviated from the original flux rate data matrix. In other words, it intuitively represents the information loss we expect if we use this model in place of the actual flux rates. The smaller the error is, the better the model fits the data. Figure 2 shows that the error decreases to approximately 0.853% at the stationary state.

We also performed a sensitivity analysis on γ in order to show the robustness of our approach, by testing the perturbation of the flux rates induced by various perturbations of γ. Specifically, small changes of γ (1%) caused a very small average perturbation of 0.0079 mmolh^{-1}gDW^{-1} of the flux rates, 0.024 mmolh^{-1}gDW^{-1} for acetate and 0.0009 h^{-1} for biomass. A stronger (10%) perturbation of γ yields an average perturbation of 0.0081

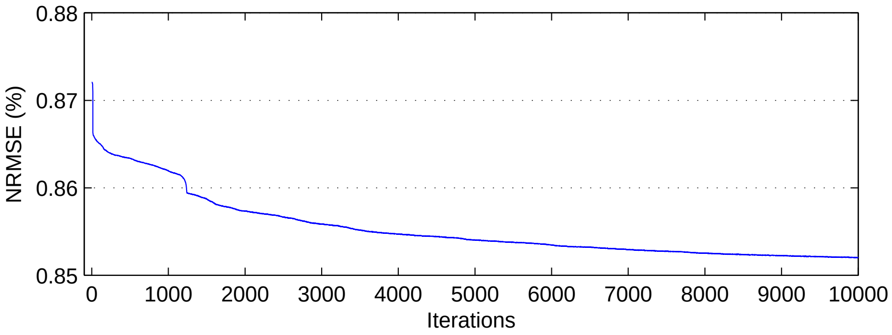

Figure 2: Normalized root mean square error (NRMSE) across the iterations.

mmolh^{-1}gDW^{-1} of the flux rates, 0.2200 mmolh^{-1}gDW^{-1} for acetate and 0.0085 h^{-1} for biomass. Finally, a strong perturbation (one order of magnitude) resulted in an average change of 0.2385 mmolh^{-1}gDW^{-1} in the flux rates, 2.0696 mmolh^{-1}gDW^{-1} for acetate, and 0.044 h^{-1} for biomass. We would like to remark that γ can be assigned individually for each reaction where information on the translation rate, codon usage or post-translational modifications is available.

4.2 Bacterial flux responses and pathway cross-correlations

With the objective to maximize both biomass and acetate production as shown in Figure 3, the *E. coli* strains grown in conditions with 10 mmolh^{-1}gDW^{-1} of glucose uptake rate produce more biomass and acetate than the strains in lower glucose. Under aerobic conditions, the maximum biomass is 2.31 h^{-1} with 39.52 mmol h^{-1} gDW^{-1} of acetate, while the maximum acetate reached is 48.20 mmol h^{-1} gDW^{-1} with a biomass of 2.16 h^{-1}. Under anaerobic conditions and low glucose, the maximum biomass is 1.04 h^{-1} (with 4.36 mmol h^{-1} gDW^{-1} of acetate production). Under anaerobic conditions and high glucose, the *E. coli* is only able to produce 1.36 h^{-1} of biomass, with a maximum of 19.50 mmol h^{-1} gDW^{-1} of acetate. The full table with the experimental conditions and the values of acetate and biomass are in Supplementary Information. We are interested in investigating the underlying mechanisms in different oxygen conditions and glucose uptake rates (high/low glucose environments with a threshold of 10 mmolh^{-1}gDW^{-1}).

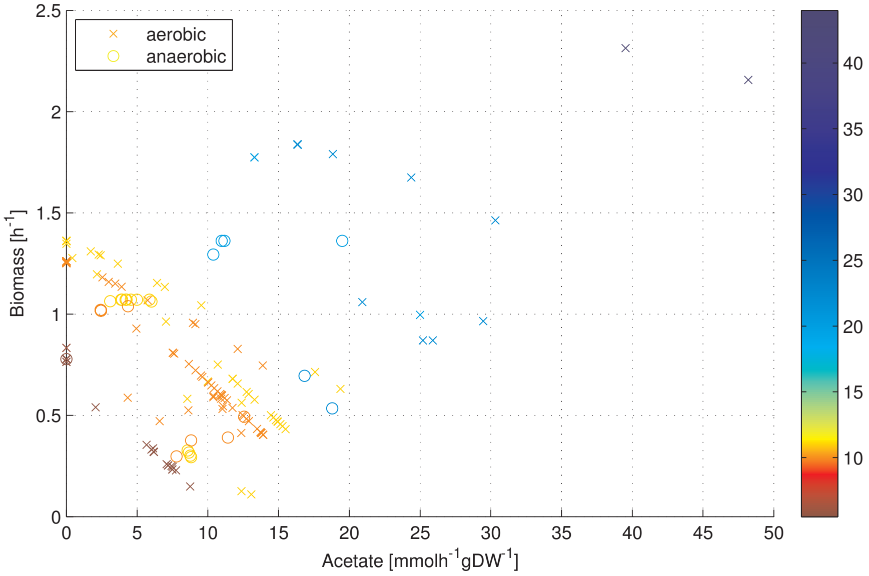

Figure 3: Through augmented FBA, all the 466 gene expression profiles are mapped to the acetate-biomass objective space. Each point contains 2583 reaction flux rates, while only acetate and biomass are shown, representing the objectives of the bilevel linear program. The color bar shows the glucose uptake rate [mmolh^{-1}gDW^{-1}]. Interestingly, the *E. coli* grown in some conditions with 10 mmolh^{-1}gDW^{-1} of glucose uptake rate is able to produce more biomass and acetate than the strains grown on higher glucose. The pathway-based Bayesian analysis is performed on the flux rates in the four conditions based on the two criteria of oxygen and glucose.

The Bayesian factor modeling with GMRF allows us to extract the behaviors of *E. coli* in response to a variety of experimental conditions and cross-correlations between pathways. Since we assume that the pathway cross-correlations represent a static, intrinsic property of *E. coli*, the model computes the cross-correlations as a global factor which is shared by individual conditions and over time. Figure 4 illustrates the sparse network of inferred cross-correlations between metabolic pathways. It is remarkable that nucleotide pathway (PID:5) acts as a central hub of the cross-correlation network, and is involved with multiple pathways. The strongest correlation is the cross-correlation between the alanine and aspartate metabolism pathway (PID:25) and the pivot pathway of another small community pertaining to the valine, leucine, and isoleucine metabolism (PID:17). A recent review of amino acids and their functions shows that alanine is the primary amino acid gluconeogenesis, and valine directly synthesizes glutamine and alanine (*27*).

Interestingly, there is a modest link between these two clusters PID:5 and PID:17 with correlation ≈ 0.15 (top 3% in the correlation matrix) by a transcription factor called leucine-responsive regulatory protein (Lrp) (*28*). This link plays an important role depending on the availability of oxygen over time, which we will discuss in the next section. Table 1 shows the average responsiveness degree of the most responsive pathways to different oxygen and glucose conditions. While the nucleotide pathway is important under anaerobic conditions on low glucose, the valine, leucine, and isoleucine metabolism pathway and the alanine and aspartate metabolism pathway both exhibit a key role under aerobic conditions on high glucose, highlighting a pathway cross-correlation between them (Figure 4).

Table 1: Average responsiveness of the most responsive pathways across aerobic and anaerobic conditions of high and low glucose. PID:5 is important under anaerobic conditions on low glucose, while PID:17 and PID:25 both exhibit a key role under aerobic conditions on high glucose, highlighting a pathway cross-correlation between them.

		High glucose		Low glucose	
PID	Pathway name	Aerobic	Anaerobic	Aerobic	Anaerobic
5	Nucleotide salvage	0.0865	0.0965	0.1366	**0.1714**
17	Valine, leucine, and isoleucine metabolism	**0.2219**	0.2147	0.1974	0.1590
25	Alanine and aspartate metabolism	**0.1544**	0.1487	0.1285	0.1076

4.3 Flux rate progression and temporal pathway activation

Not only do the *E. coli* strains respond to the environment differently from condition to condition within the same time frame, but their temporal behaviors also progress over time in a different way. In order to perform a dynamical analysis of pathway activation, we consider the gene expression profiles of 41 time-course experiments after the exposure to stimuli or in stress conditions.

Figure 5 shows the acetate secretion/assimilation and the growth rate as a result of the metabolic regulation system of *Escherichia coli* in response to the changing environment. We also compare the Euclidean distance “covered” by each of the 41 conditions from the start to the end of the observations (four time steps). The fluctuations in acetate production depend on the balance between pyruvate fermentation to acetate and Krebs cycle. Interestingly, the

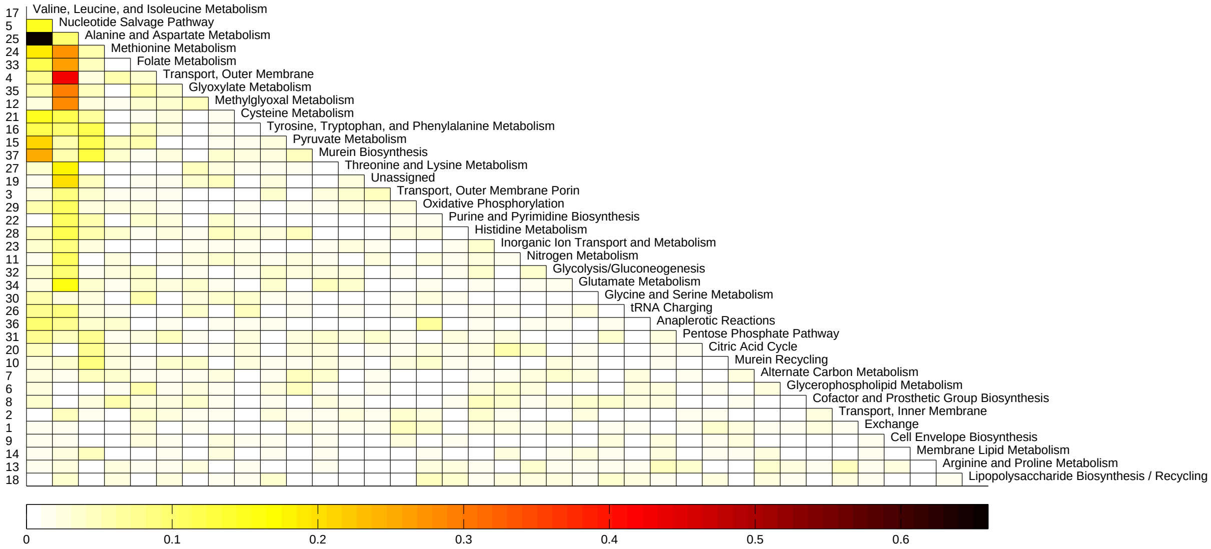

Figure 4: Pathway correlation matrix derived from $\mathbf{\Phi}$, an output of the Bayesian factor modeling. On the y-axis we report the pathway identifiers (PID) that correspond to their pathway names labeled on the diagonal. The color bar shows the correlation between pathways computed with the Bayesian factor modeling. These values suggest the strength of pathway cross-correlations underlying the bacterial response through flux rates in both steady-state and time-course experiments.

most remarkable fluctuations are those caused by the presence of ampicillin, norfloxacin and spectinomycin, antibiotics used to treat a number of bacterial infections. This result confirms that in specific conditions, some antibiotics can have a high impact on the metabolism and physiology of *E. coli*, while others (e.g. kanamycin) have no effect (for instance, due to the development of antimicrobial resistance). These metabolic transitions confirm well-known experimental results on the acetate switch during growth under different conditions, and on the inverse relation between growth rate and acetate secretion (*29*, *30*). The global response of the biomass to a changing environment can be explained also as a separated response of two subpopulations reacting at different speeds to the environmental change (*31*).

Pathway activation shows an apparent progress over time in the aerobic condition shown in Figure 6a, while the anaerobic activation is more static (Figure 6b). As expected, the nucleotide salvage pathway (PID:5) highly responds to the environment where oxygen was present, while valine, leucine, and isoleucine metabolism pathway (PID:17) and alanine and

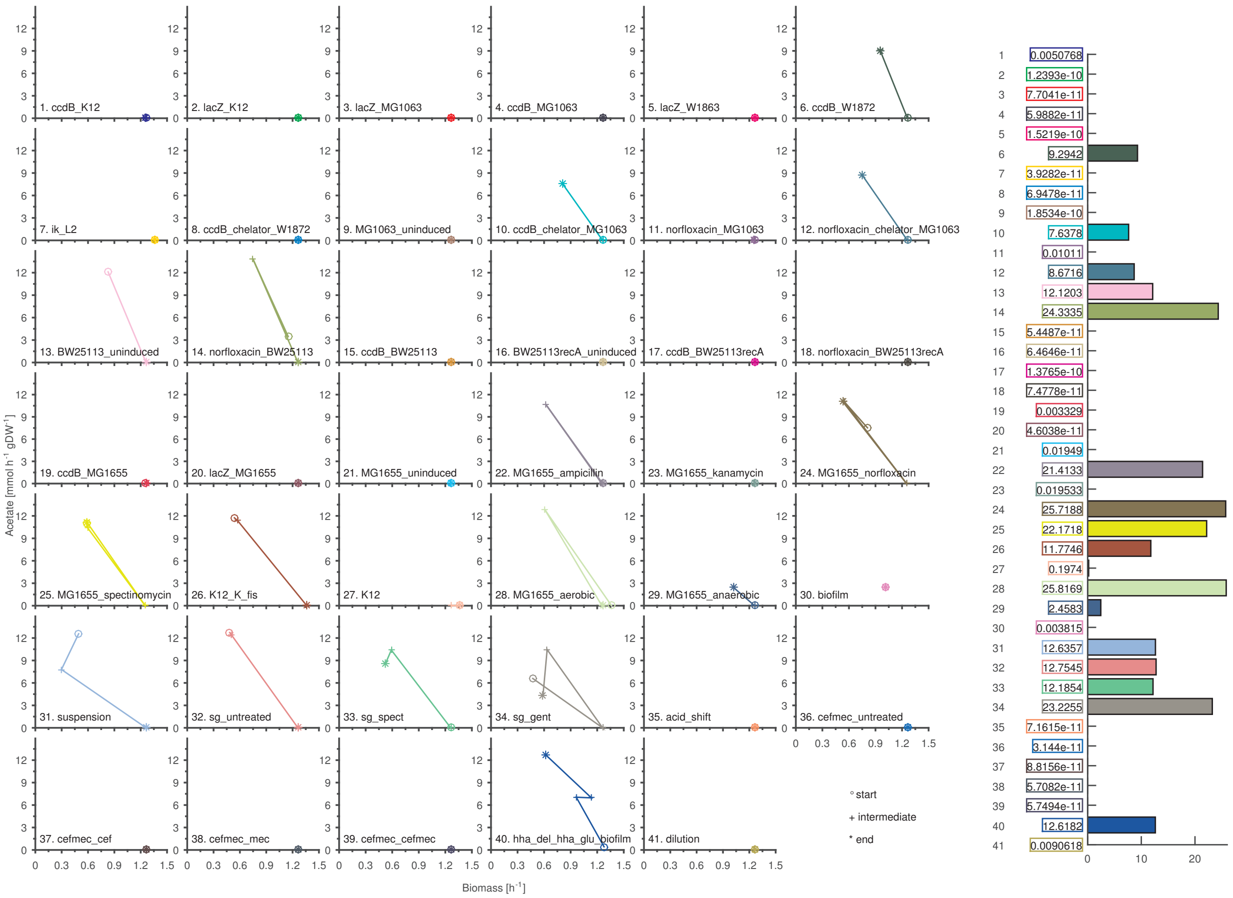

Figure 5: Acetate production/assimilation and biomass produced by *Escherichia coli* in 41 different environmental conditions. We used a set of 164 expression arrays consisting of 41 conditions with 4 measurements each. Growth conditions include pH changes, antibiotics, genetic perturbations, heat shock, different growth media, carbon source, oxygen and glucose concentrations (see Supplementary Information for further details). Some conditions show quick metabolic rearrangement during growth, whereas under other conditions the bacterium metabolism remains stable. In the right panel, we show the Euclidean distance covered by each of the 41 conditions in the biomass-acetate space during the four time steps. The conditions whose bars are not shown have a negligible (less than 10^{-9}) distance covered in the acetate-biomass space during the four time steps.

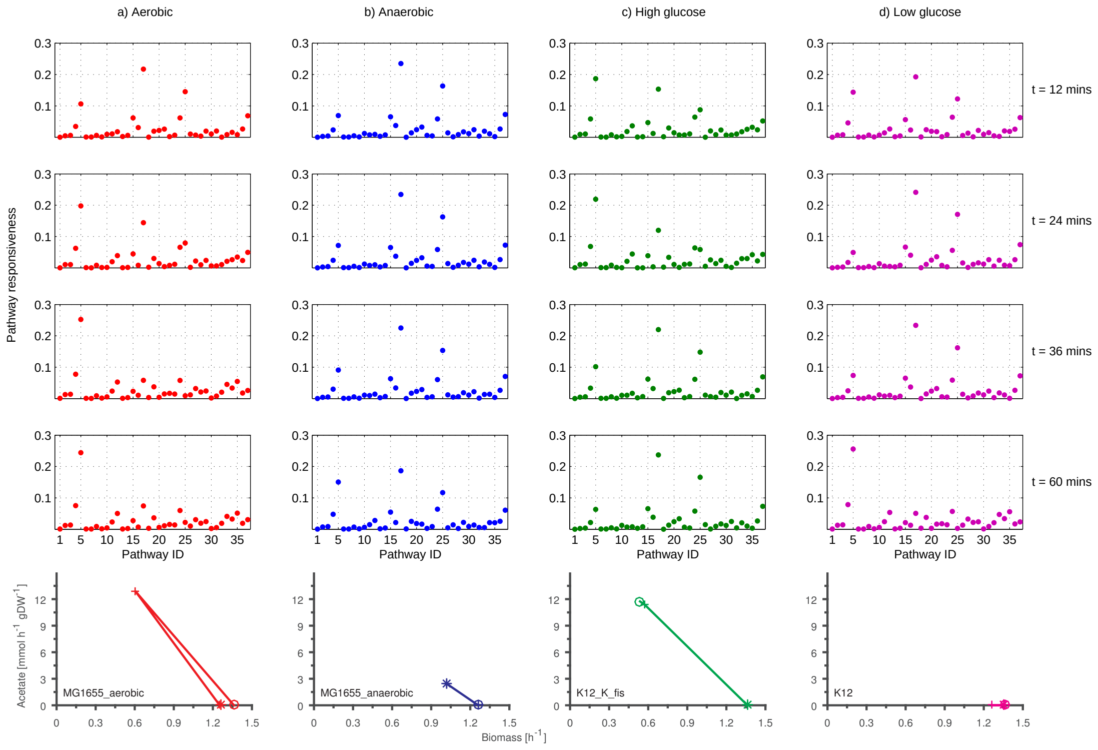

Figure 6: Pathway responsiveness in four conditions: (a) aerobic, (b) anaerobic, (c) high glucose, and (d) low glucose at four time steps. We performed the Bayesian factor modeling on the time-course flux rate responses sharing the same underlying cross-correlations. As a result, pathway responsiveness indicates how much a pathway is likely to be responsive to a given condition at a given time. The pathway responsiveness plots provide us the comprehensive interpretation of the *E. coli* progression of phenotypic behavior in response to each condition, as shown in the plots at the bottom. Specifically, conditions with large variations in the biomass and acetate production are also likely to cause large variation in the pathway responsiveness.

aspartate metabolism (PID:25) are less active. In contrast, all these pathways are active in response to the low-oxygen environments, and maintain their activity over time. These results support the inferred cross-correlation between two communities of PID:5 and PID:17 linked by the transcription factor called leucine-responsive regulatory protein (Lrp) (*28*), which is designated as a global transcription factor in *E. coli* (*32*). A recent experiment confirmed that the activity of Lrp decreases as aerobiosis increases (*33*). This may explain the mechanism of the inactive PID:17 after the exposure to oxygen.

In addition, we observed that the cysteine metabolism pathway (PID:21) was activated in the anaerobic condition and lasted for three time steps before it was deactivated in the last time point. The case of cysteine has long been studied in bacteria: although cysteine residues of FNR, another global transcription factor in *E. coli* (*32*), are alkylated with iodoacetate in permeabilized aerobic or anaerobic conditions, the process takes 50 minutes in anaerobic bacteria and 6 minutes in aerobic bacteria (*34*). Thus, we could not observe that the cysteine metabolism was active at any time steps after six minutes in aerobic bacteria, but we could observe the activation of cysteine pathway at the first three time frames within the first 50 minutes under anaerobic condition.

Different levels of glucose uptake also change pathway activities over time as shown in Figures 6c and 6d. At the early stage of glucose prevalence, the methylglyoxal metabolism pathway (PID:12) was activated, unlike when the bacterium was starved of glucose. Methylglyoxal (MG) is usually synthesized under a condition with low phosphate and high dihydroxyacetone phosphate (DHAP), an environment that occurs most frequently under high-glucose settings (*35*). A recent experiment confirmed that the increase of glucose uptake rates result in the temporary excretion of methylglyoxal synthase (MgsA) (*36*, *37*). As MG accumulation will lead to cell death, *E. coli* requires a mechanism for MG degradation (*37*). The detoxification of methylglyoxal includes the conversion MG to S-lactoyl glutathione and then to D-lactate by glyoxalase enzymes I and II (*37*). This procedure has also been presented to be the predominant MG detoxification system in *E. coli* (*37*, *38*). This exposition also endorses our discovery about the modest correlation (≈ 0.05) between the methylglyoxal metabolism pathway (PID:12) and the glyoxalase metabolism pathway (PID:35), which appears at the top 10% of the inferred sparse cross-correlation network.

5 Conclusion

As a result of many recent research efforts to elucidate the relation between genotype and phenotype, we currently have models for a better understanding of the individual components, but arguably a less clear picture of the interactions between the biological components that result in a given phenotype (*10*). We still have, moreover, limited knowledge about how to use these models to predict a phenotypic response to a changing environmental condition, due to the lack of comprehensive data across different conditions and accurate training processes performed on the models (*39*). Probabilistic, sometimes rare interactions between molecules and metabolic activation depend on external conditions and may change over time. These interactions and active components give rise to important effects in the bacterial physiology, such as non-linear activation of toggle switches or master regulators.

When a particular bacterial phenotype has to be achieved, some pathways are required to respond more than others. Our idea is to analyze a compendium of experimental conditions to investigate the combinations of pathway activation that will allow the bacterium to mimic the behavior of the desired phenotype. In order to map the environmental changes to the *E. coli*, we started from the most recent genome-scale metabolic reconstruction (*20*) and we used bilevel flux-balance-analysis to modify the constraints on the metabolic fluxes according to the gene expression profile associated with each condition. Each environmental condition is converted into a flux profile and mapped to a single point in a bidimensional objective space, therefore translating dynamic genetic activities into dynamic reactions fluxes. (In this paper we focused on the acetate-biomass space, but the methodology is readily applicable to any multidimensional space by extending bilevel FBA to many-level FBA.)

At this step, thousands of reaction flux rates for each experimental condition would need a lot of expertise and manual work for their interpretation. We therefore summarized the reaction fluxes by developing a Bayesian factor model able to identify the *pathway responsiveness*, representing the responsive degree of each pathway under each environmental condition. Our method is also able to achieve a systematic prediction of *E. coli* metabolic pathway responses to time-varying signals. The Bayesian factor model simultaneously elucidates all the pathway-pathway interactions (pathway cross-correlations), which are also the underlying process behind the pathway activation at each time step. While the activation of some metabolic pathways is kept at the same level, other pathways fluctuate as part of the global response to the fluctuating environment.

While the cross-correlations are considered an intrinsic property of the *E. coli* metabolic network, and thus being computed globally, the degrees of responsiveness depend on the particular environmental conditions at each time step. This characterizes the temporal progression of pathway activation, throughout the time series, in response to given stimuli. Instead of considering a single condition a time, performing the analysis across all conditions provides insights into pathway connectivity and pathway islands that rarely co-activate with others. Pathway activation profiles allows different conditions in which the *E. coli* responds similarly to be clustered. This is also advantageous for discovering alternative antibiotic treatments by replacing stimuli with different chemicals (*40*). Many interesting applications can be implemented from the particular pathway cross-correlations predicted from a set of environmental experiments. For instance, they facilitate the prediction of bacterial behaviors in specific situations involving reciprocal action or influence between different organisms, e.g. the interaction between bacteria and plants in mycorrhiza, or sepsis and health conditions (e.g. gut microbiota). Our results can also help to shed light on why different conditions can show the same response in a given multi-objective output space. Finally, assessing the pathway correlations with our framework can indicate an *operating distance* between pathways, thus enhancing the current knowledge of the metabolic network and providing foundations of methodological value for analyzing multi-omic data.

Acknowledgement

This work is supported by the EU Mission T2D project. NP acknowledges the Royal Thai Government Scholarship.

Supporting Information Available

Supplementary Table 1. Flux rates obtained as output of the augmented FBA for the 466 experimental conditions.

Supplementary Table 2. Pathway cross-correlations inferred by the Bayesian factor model.

Supplementary Table 3. Pathway responsiveness to the 466 experimental conditions and the 41 conditions at four time steps, resulted from the Bayesian factor modeling. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

1. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., and Lander, E. S. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America 102*, 15545–15550.

2. Hood, L. (2013) Systems biology and p4 medicine: past, present, and future. *Rambam Maimonides medical journal 4*, e0012.

3. Wang, T., Gu, J., Yuan, J., Tao, R., Li, Y., and Li, S. (2013) Inferring pathway crosstalk networks using gene set co-expression signatures. *Mol. BioSyst. 9*, 1822–1828.

4. Li, Y., Agarwal, P., and Rajagopalan, D. (2008) A global pathway crosstalk network. *Bioinformatics 24*, 1442–1447.

5. Pratanwanich, N., and Lio', P. (2014) *Molecular biosystems DOI:10.1039/c4mb00014e.*

6. Pan, X.-H. (2012) Pathway Crosstalk Analysis Based on Protein-protein Network Analysis in Ovarian Cancer. *Asian Pacific Journal of Cancer Prevention 13*, 3905–3909.

7. Kim, H. U., Kim, T. Y., and Lee, S. Y. (2011) Framework for network modularization and Bayesian network analysis to investigate the perturbed metabolic network. *BMC systems biology 5*, S14.

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

9. Lee, D., Smallbone, K., Dunn, W., Murabito, E., Winder, C., Kell, D., Mendes, P., and Swainston, N. (2012) Improving metabolic flux predictions using absolute gene expression data. *BMC Systems Biology 6*, 73.

10. Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014) Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics 15*, 107–120.

11. von Kamp, A., and Klamt, S. (2014) Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS computational biology 10*, e1003378.

12. Chindelevitch, L., Trigg, J., Regev, A., and Berger, B. (2014) An exact arithmetic toolbox for a consistent and reproducible structural analysis of metabolic network models. *Nature communications 5*.

13. Edwards, J. S., Covert, M., and Palsson, B. (2002) Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology 4*, 133–140.

14. Angione, C., Carapezza, G., Costanza, J., Lió, P., and Nicosia, G. (2013) Pareto optimality in organelle energy metabolism analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 10*, 1032–1044.

15. Costanza, J., Carapezza, G., Angione, C., Lió, P., and Nicosia, G. (2012) Robust design of microbial strains. *Bioinformatics 28*, 3097–3104.

16. Angione, C., Costanza, J., Carapezza, G., Lió, P., and Nicosia, G. (2013) A design automation framework for computational bioenergetics in biological networks. *Molecular BioSystems 9*, 2554–2564.

17. Potera, C. (2005) Making succinate more successful. *Environmental health perspectives* A833–A835.

18. Dittrich, C. R., Bennett, G. N., and San, K.-Y. (2005) Characterization of the Acetate-Producing Pathways in Escherichia coli. *Biotechnology progress 21*, 1062–1067.

19. Faith, J. J., Driscoll, M. E., Fusaro, V. A., Cosgrove, E. J., Hayete, B., Juhn, F. S., Schneider, S. J., and Gardner, T. S. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic acids research 36*, D866–D870.

20. Orth, J., Conrad, T., Na, J., Lerman, J., Nam, H., Feist, A., and Palsson, B. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism2011. *Molecular systems biology 7*, 535.

21. Mar, J. C., Matigian, N. A., Mackay-Sim, A., Mellick, G. D., Sue, C. M., Silburn, P. A., McGrath, J. J., Quackenbush, J., and Wells, C. A. (2011) *PLoS genetics 7*, e1002207.

22. Firczuk, H., Kannambath, S., Pahle, J., Claydon, A., Beynon, R., Duncan, J., Westerhoff, H., Mendes, P., and McCarthy, J. E. (2013) An in vivo control map for the eukaryotic mRNA translation machinery. *Molecular systems biology 9*, 635.

23. Paltanea, M., Tabirca, S., Scheiber, E., and Tangney, M. Logarithmic Growth in Biological Processes. Computer Modelling and Simulation (UKSim), 2010 12th International Conference on. 2010; pp 116–121.

24. Guimaraes, J. C., Rocha, M., and Arkin, A. P. (2014) Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic acids research 42*, 4791–4799.

25. Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M., and Drummond, D. A. (2014) Accounting for experimental noise reveals that mRNA levels, amplified by post-

transcriptional processes, largely determine steady-state protein levels in yeast. *bioRxiv* 009472.

26. Rue, H., and Held, L. *Gaussian Markov random fields: theory and applications*; CRC Press, 2005.

27. Wu, G. (2009) Amino acids: metabolism, functions, and nutrition. *Amino acids 37*, 1–17.

28. Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology 5*, e8.

29. Kleman, G. L., and Strohl, W. R. (1994) Acetate metabolism by Escherichia coli in high-cell-density fermentation. *Applied and environmental microbiology 60*, 3952–3958.

30. Wolfe, A. J. (2005) The acetate switch. *Microbiology and Molecular Biology Reviews 69*, 12–50.

31. Kotte, O., Volkmer, B., Radzikowski, J. L., and Heinemann, M. (2014) Phenotypic bistability in Escherichia coli's central carbon metabolism. *Molecular systems biology 10*.

32. Martınez-Antonio, A., and Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Current opinion in microbiology 6*, 482–489.

33. Rolfe, M. D., Ocone, A., Stapleton, M. R., Hall, S., Trotter, E. W., Poole, R. K., Sanguinetti, G., and Green, J. (2012) Systems analysis of transcription factor activities in environments with stable and dynamic oxygen concentrations. *Open biology 2*, 120091.

34. Trageser, M., and Unden, G. (1989) Role of cysteine residues and of metal ions in the

regulatory functioning of FNR, the transcriptional regulator of anaerobic respiration in Escherichia coli. *Molecular microbiology 3*, 593–599.

35. Hopper, D., and Cooper, R. (1972) The purification and properties of¡ italic¿ Escherichia coli¡/italic¿ methylglyoxal synthase. *Biochem. J 128*, 321–329.

36. Pepper, E. D., Farrell, M. J., Nord, G., and Finkel, S. E. (2010) Antiglycation effects of carnosine and other compounds on the long-term survival of Escherichia coli. *Applied and environmental microbiology 76*, 7925–7930.

37. Ferguson, G. P. (1999) Protective mechanisms against toxic electrophiles in¡ i¿ Escherichia coli¡/i¿. *Trends in microbiology 7*, 242–247.

38. MacLean, M., Ness, L., Ferguson, G., and Booth, I. (1998) The role of glyoxalase I in the detoxification of methylglyoxal and in the activation of the KefB K+ efflux system in Escherichia coli. *Molecular microbiology 27*, 563–571.

39. Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A., and Tagkopoulos, I. (2014) An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. *Molecular systems biology 10*, 735.

40. Kohanski, M. A., Dwyer, D. J., and Collins, J. J. (2010) How antibiotics kill bacteria: from targets to networks. *Nature Reviews Microbiology 8*, 423–435.