



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Flux-dependent graphs for metabolic networks

Citation for published version:

Beguirisse-Díaz, M, Bosque, G, Oyarzún, D, Picó, J & Barahona, M 2018, 'Flux-dependent graphs for metabolic networks' *npj Systems Biology and Applications*, vol. 4, no. 1, 32. DOI: 10.1038/s41540-018-0067-y

Digital Object Identifier (DOI):

[10.1038/s41540-018-0067-y](https://doi.org/10.1038/s41540-018-0067-y)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

npj Systems Biology and Applications

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



ARTICLE OPEN

Flux-dependent graphs for metabolic networks

Mariano Beguerisse-Díaz^{1,2}, Gabriel Bosque³, Diego Oyarzún¹, Jesús Picó³ and Mauricio Barahona¹

Cells adapt their metabolic fluxes in response to changes in the environment. We present a framework for the systematic construction of flux-based graphs derived from organism-wide metabolic networks. Our graphs encode the directionality of metabolic flows via edges that represent the flow of metabolites from source to target reactions. The methodology can be applied in the absence of a specific biological context by modelling fluxes probabilistically, or can be tailored to different environmental conditions by incorporating flux distributions computed through constraint-based approaches such as Flux Balance Analysis. We illustrate our approach on the central carbon metabolism of *Escherichia coli* and on a metabolic model of human hepatocytes. The flux-dependent graphs under various environmental conditions and genetic perturbations exhibit systemic changes in their topological and community structure, which capture the re-routing of metabolic flows and the varying importance of specific reactions and pathways. By integrating constraint-based models and tools from network science, our framework allows the study of context-specific metabolic responses at a system level beyond standard pathway descriptions.

npj Systems Biology and Applications (2018)4:32; doi:10.1038/s41540-018-0067-y

INTRODUCTION

Metabolic reactions enable cellular function by converting nutrients into energy, and by assembling macromolecules that sustain the cellular machinery.¹ Cellular metabolism is usually thought of as a collection of pathways comprising enzymatic reactions associated with broad functional categories. Yet metabolic reactions are highly interconnected: enzymes convert multiple reactants into products with other metabolites acting as co-factors; enzymes can catalyse several reactions, and some reactions are catalysed by multiple enzymes, and so on. This enmeshed web of reactions is thus naturally amenable to network analysis, an approach that has been successfully applied to different aspects of cellular and molecular biology, e.g., protein-protein interactions,² transcriptional regulation,³ or protein structure.^{4,5}

Tools from graph theory⁶ have previously been applied to the analysis of structural properties of metabolic networks, including their degree distribution,^{7–10} the presence of metabolic roles,¹¹ and their community structure.^{12–15} A central challenge, however, is that there are multiple ways to construct a network from a metabolic model.¹⁶ For example, one can create a graph with metabolites as nodes and edges representing the reactions that transform one metabolite into another,^{7,8,17,18} a graph with reactions as nodes and edges corresponding to the metabolites shared among them,^{19–21} or even a bipartite graph with both reactions and metabolites as nodes.²² Importantly, the conclusions of graph-theoretical analyses are highly dependent on the chosen graph construction.²³

A key feature of metabolic reactions is the directionality of flows: metabolic networks contain both irreversible and reversible reactions, and reversible reactions can change their direction depending on cellular and environmental contexts.¹ Existing graph constructions have been useful for developing an intuitive understanding of metabolic complexity. Many of these

constructions, however, lead to graphs that do not include directional information that is central to metabolic function.^{8,16} In addition, current graph constructions are usually derived from the whole set of metabolic reactions in an organism, and thus correspond to a generic metabolic ‘blueprint’ of the cell. Yet cells switch specific pathways ‘on’ and ‘off’ to sustain their energetic budget in different environments.²⁴ Hence, such blueprint graphs might not capture the specific metabolic connectivity in a given environment, thus limiting their ability to provide biological insights in different growth conditions.

In this paper, we present a flow-based approach to construct metabolic graphs that encapsulate the directional flow of metabolites produced or consumed through enzymatic reactions. The proposed graphs can be tailored to incorporate flux distributions under different environmental conditions. To introduce our approach, we proceed in two steps. We first define the *Normalised Flow Graph* (NFG), a weighted, directed graph with reactions as nodes, edges that represent supplier-consumer relationships between reactions, and weights given by the probability that a metabolite chosen at random from all reactions is produced/consumed by the source/target reaction. This graph can be used to carry out graph-theoretical analyses of organism-wide metabolic organisation independent of cellular context or environmental conditions. We then show that this formalism can be adapted seamlessly to construct the *Mass Flow Graph* (MFG), a directed, environment-dependent, graph with weights computed from Flux Balance Analysis (FBA),²⁵ the most widespread method to study genome-scale metabolic networks.

Our formulation addresses several drawbacks of current constructions of metabolic graphs. Firstly, in our flow graphs, an edge indicates that metabolites are produced by the source reaction and consumed by the target reaction, thus accounting for metabolic directionality and the natural flow of chemical mass from reactants to products. Secondly, the Normalised Flow Graph

¹Department of Mathematics, Imperial College London, London SW7 2AZ, UK; ²Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK and ³Institut Universitari d'Automàtica i Informàtica Industrial, Universitat Politècnica de València, Camí de Vera s/n, 46022 Valencia, Spain

Correspondence: Mariano Beguerisse-Díaz (beguerisse@maths.ox.ac.uk) or Mauricio Barahona (m.barahona@imperial.ac.uk)

Received: 27 January 2017 Revised: 28 June 2018 Accepted: 3 July 2018

Published online: 14 August 2018

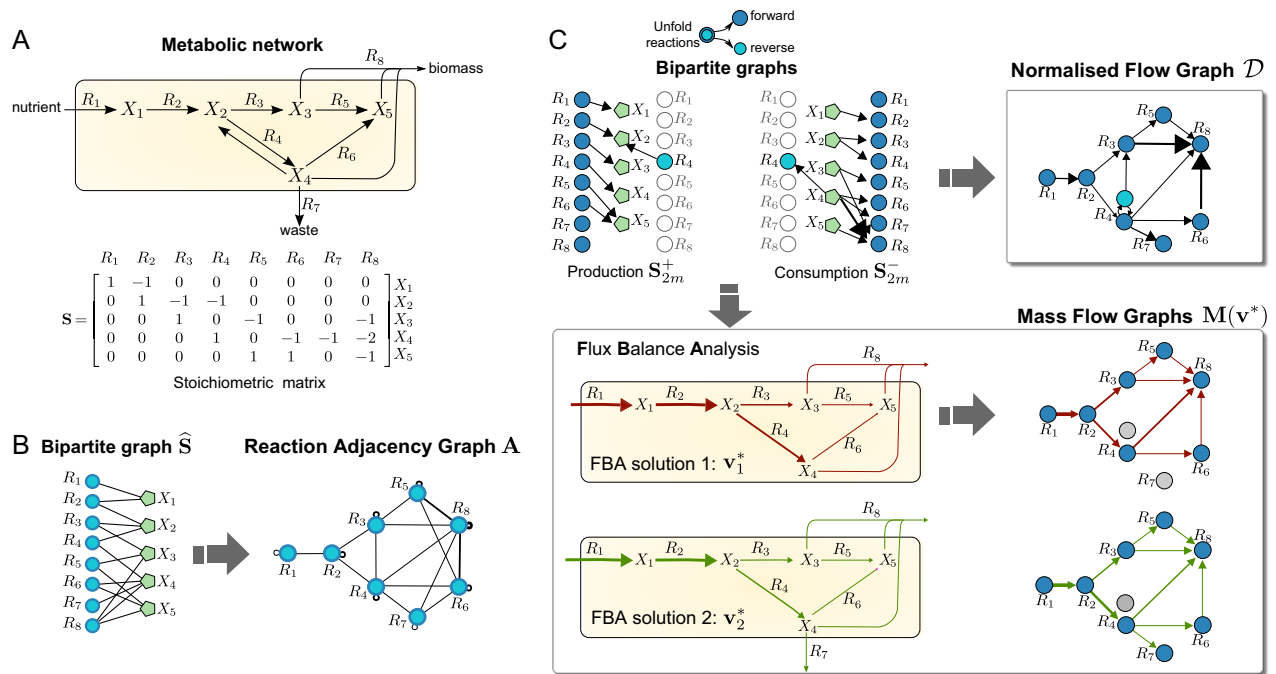


Fig. 1 Graphs from metabolic networks. **a** Toy metabolic network describing nutrient uptake, biosynthesis of metabolic intermediates, secretion of waste products, and biomass production.³² The biomass reaction is $R_8: X_3 + 2X_4 + X_5$, **b** Bipartite graph associated with the boolean stoichiometric matrix \hat{S} , and the Reaction Adjacency Graph (RAG)¹⁶ with adjacency matrix $A = \hat{S}^T \hat{S}$. The undirected edges of **A** indicate the number of shared metabolites among reactions. **c** The Normalised Flow Graph (NFG) \mathcal{D} and two Mass Flow Graphs (MFG) $M(v^*)$ constructed from the consumption and production stoichiometric matrices (5). Note that the reversible reaction R_4 is unfolded into two nodes. The NFG in Eq. (8) is a directed graph with weights representing the probability that a metabolite chosen uniformly at random from the stoichiometric matrix is produced by source reaction is consumed by the target reaction. The MFGs in Eq. (12) are constructed from two different Flux Balance Analysis solutions (v_1^* and v_2^*) obtained by optimising a biomass objective function under different flux constraints representing different environmental or cellular contexts (see Sec. SI 2 in the Supplementary Information for details). The weighted edges of the MFGs represent mass flow from source to target reactions in units of metabolic flux. The computed FBA solutions translate into different connectivity in the resulting MFGs

discounts naturally the over-representation of pool metabolites (e.g., adenosine triphosphate (ATP), nicotinamide adenine dinucleotide (NADH), protons, water, and other co-factors) that appear in many reactions and tend to obfuscate the graph connectivity. Our construction avoids the removal of pool metabolites from the network, which can change the graph structure drastically.^{26–30} Finally, the Mass Flow Graph incorporates additional biological information reflecting the effect of the environmental context into the graph construction. In particular, since the weights in the MFG correspond directly to fluxes (in units of mass per time), different biological scenarios can be analysed using balanced fluxes (e.g., from different FBA solutions) under different carbon sources and other environmental perturbations.^{16,25,31,32}

After introducing the mathematical framework, we showcase our approach with two examples. Firstly, in the absence of environmental context, our analysis of the NFG of the core model of *Escherichia coli* metabolism³³ reveals the importance of including directionality and appropriate edge weights in the graph to understand the modular organisation of metabolic sub-systems. We then use FBA solutions computed for several relevant growth conditions for *E. coli*, and show that the structure of the MFG changes dramatically in each case (e.g., connectivity, ranking of reactions, community structure), thus capturing the environment-dependent nature of metabolism. Secondly, we study a model of human hepatocyte metabolism evaluated under different conditions for the wild-type and in a mutation found in primary hyperoxaluria type 1, a rare metabolic disorder,³⁴ and show how the changes in network structure of the MFGs reveal new information that is complementary to the flux analysis predicted by FBA.

RESULTS

Definitions and background

Consider a metabolic network composed of n metabolites X_i ($i = 1, \dots, n$) that participate in m reactions

$$R_j: \sum_{i=1}^n \alpha_{ij} X_i \rightleftharpoons \sum_{i=1}^n \beta_{ij} X_i, \quad j = 1, 2, \dots, m, \quad (1)$$

where α_{ij} and β_{ij} are the stoichiometric coefficients of species i in reaction j . Let us denote the concentration of metabolite X_i at time t as $x_i(t)$. We then define the n -dimensional vector of metabolite concentrations: $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))^T$. Each reaction takes place with rate $v_j(\mathbf{x}, t)$, measured in units of concentration per time.³⁵ We compile the reaction rates in the m -dimensional vector: $\mathbf{v}(t) = (v_1(t), \dots, v_m(t))^T$.

The mass balance of the system can then be represented compactly by the system of ordinary differential equations

$$\dot{\mathbf{x}} = S\mathbf{v}, \quad (2)$$

where the $n \times m$ matrix S is the stoichiometric matrix with entries $S_{ij} = \beta_{ij} - \alpha_{ij}$, i.e., the net number of X_i molecules produced (positive S_{ij}) or consumed (negative S_{ij}) by the j -th reaction. Figure 1a shows a toy example of a metabolic network including nutrient uptake, biosynthesis of metabolic intermediates, secretion of waste products, and biomass production.³²

There are several ways to construct a graph for a given metabolic network with stoichiometric matrix S . A common approach¹⁶ is to define a *unipartite graph* with reactions as nodes

and $m \times m$ adjacency matrix

$$\mathbf{A} = \widehat{\mathbf{S}}^T \widehat{\mathbf{S}}, \quad (3)$$

where $\widehat{\mathbf{S}}$ is the boolean version of \mathbf{S} (i.e., $\widehat{S}_{ij} = 1$ when $S_{ij} \neq 0$ and $\widehat{S}_{ij} = 0$ otherwise). This is the *Reaction Adjacency Graph* (RAG), in which two reactions (nodes) are connected if they share metabolites, either as reactants or products. Self-loops represent the total number of metabolites that participate in a reaction (Fig. 1b).

Though widely studied,^{8,16} the RAG has known limitations and overlooks key aspects of the connectivity of metabolic networks. The RAG is blind to the directionality of flows, as it does not incorporate the irreversibility of reactions (by construction \mathbf{A} is a symmetric matrix). Furthermore, the structure of \mathbf{A} is dominated by the large number of edges introduced by pool metabolites that appear in many reactions, such as water, ions or enzymatic cofactors. Computational schemes have been introduced to mitigate the bias caused by pool metabolites,²⁷ but these do not follow from biophysical considerations and need manual calibration. Finally, the construction of the graph \mathbf{A} from is not easily extended to incorporate the effect of environmental changes.

Metabolic graphs that incorporate flux directionality and biological context

To address the limitations of the reaction adjacency graph \mathbf{A} , we propose a graph formulation that follows from a flux-based perspective. To construct our graph, we unfold each reaction into two separate directions (forward and reverse) and redefine the links between reaction nodes to reflect producer-consumer relationships. Specifically, two reactions are connected if one produces a metabolite that is consumed by the other. As shown below, this definition leads to graphs that naturally account for the reversibility of reactions, and allows for the seamless integration of biological contexts modelled through FBA. Inspired by matrix formulations of chemical reaction network kinetics,³⁶ we rewrite the reaction rate vector \mathbf{v} as:

$$\mathbf{v} := \mathbf{v}^+ - \mathbf{v}^- = \mathbf{v}^+ - \text{diag}(\mathbf{r})\mathbf{v}^-,$$

where \mathbf{v}^+ and \mathbf{v}^- are non-negative vectors containing the forward and backward reaction rates, respectively. Here the $m \times m$ matrix $\text{diag}(\mathbf{r})$ contains \mathbf{r} in its main diagonal, and \mathbf{r} is the m -dimensional reversibility vector with components $r_j = 1$ if reaction R_j is reversible and $r_j = 0$ if it is irreversible. With these definitions, we can rewrite the metabolic model in Eq. (2) as:

$$\dot{\mathbf{x}} = \mathbf{S}\mathbf{v} = \underbrace{[\mathbf{S} \quad -\mathbf{S}]}_{\mathbf{S}_{2m}} \begin{bmatrix} \mathbf{I}_m & 0 \\ 0 & \text{diag}(\mathbf{r}) \end{bmatrix} \begin{bmatrix} \mathbf{v}^+ \\ \mathbf{v}^- \end{bmatrix} := \mathbf{S}_{2m} \mathbf{v}_{2m}, \quad (4)$$

where $\mathbf{v}_{2m} := [\mathbf{v}^+ \mathbf{v}^-]^T$ is the unfolded $2m$ -dimensional vector of reaction rates, \mathbf{I}_m is the $m \times m$ identity matrix, and we have defined \mathbf{S}_{2m} , the unfolded version of the stoichiometric matrix of the $2m$ forward and reverse reactions.

Normalised Flow Graph: a directional blueprint of metabolism

The unfolding into forward and backward fluxes leads us to the definition of production and consumption stoichiometric matrices:

$$\begin{aligned} \text{Production :} & \quad \mathbf{S}_{2m}^+ = \frac{1}{2}(\text{abs}(\mathbf{S}_{2m}) + \mathbf{S}_{2m}) \\ \text{Consumption :} & \quad \mathbf{S}_{2m}^- = \frac{1}{2}(\text{abs}(\mathbf{S}_{2m}) - \mathbf{S}_{2m}), \end{aligned} \quad (5)$$

where $\text{abs}(\mathbf{S}_{2m})$ is the matrix of absolute values of the corresponding entries of \mathbf{S}_{2m} . Note that each entry of the matrix \mathbf{S}_{2m}^+ , denoted s_{ij}^+ , gives the number of molecules of metabolite X_i produced by reaction R_j . Conversely, the entries of \mathbf{S}_{2m}^- , denoted

s_{ij}^- , correspond to the number of molecules of metabolite X_i consumed by reaction R_j .

Within our directional flux framework, it is natural to consider a purely probabilistic description of producer-consumer relationships between reactions, as follows. Suppose we are given a stoichiometric matrix \mathbf{S} without any additional biological information, such as metabolite concentrations, reaction fluxes, or kinetic rates. In the absence of such information, the probability that a metabolite molecule X_k chosen uniformly at random from \mathbf{S} is produced by reaction R_i and consumed by reaction R_j is:

$$P(\text{one molecule of } X_k \text{ is produced by } R_i \text{ and consumed by } R_j) = \frac{s_{ki}^+}{w_k^+} \frac{s_{kj}^-}{w_k^-}, \quad (6)$$

where $w_k^+ = \sum_{h=1}^{2m} s_{kh}^+$ and $w_k^- = \sum_{h=1}^{2m} s_{kh}^-$ are the total number of molecules of X_k produced and consumed by all reactions that have been accounted for in \mathbf{S}_{2m} . Unlike models that rely on stochastic chemical kinetics,³⁷ the probabilities in Eq. (6) do not contain information on kinetic rate constants, which are typically not available for genome-scale metabolic models.³⁸ In our formulation, the relevant probabilities contain only the stoichiometric information included in the matrix \mathbf{S}_{2m} and should not be confused with the reaction propensity functions in Gillespie-type stochastic simulations of biochemical systems.

We thus define the weight of the edge between reaction nodes R_i and R_j as the probability that any metabolite chosen at random is produced by R_i and consumed by R_j . Summing over all metabolites and normalizing, we obtain the edge weights of the adjacency matrix of the NFG:

$$\mathcal{D}_{ij} = \frac{1}{n} \sum_{k=1}^n \frac{s_{ki}^+}{w_k^+} \frac{s_{kj}^-}{w_k^-}, \quad (7)$$

in which $\sum_{i,j} \mathcal{D}_{ij} = 1$ (i.e., the probability that any metabolite is consumed/produced by any reaction is 1). Rewritten compactly in matrix form, we obtain the

$$\text{Normalised Flow Graph (NFG) :} \quad \mathcal{D} = \frac{1}{n} \left(\mathbf{W}_+^\dagger \mathbf{S}_{2m}^+ \right)^T \left(\mathbf{W}_-^\dagger \mathbf{S}_{2m}^- \right), \quad (8)$$

where $\mathbf{W}_+^\dagger = \text{diag}(\mathbf{S}_{2m}^+ \mathbf{1}_{2m})^\dagger$, $\mathbf{W}_-^\dagger = \text{diag}(\mathbf{S}_{2m}^- \mathbf{1}_{2m})^\dagger$, $\mathbf{1}_{2m}$ is a vector of ones, and \dagger denotes the Moore-Penrose pseudoinverse. In Fig. 1c we illustrate the creation of the NFG for a toy network. The NFG is a weighted, directed graph which encodes a blueprint of the whole metabolic model, and provides a natural scaling of the contribution of pool metabolites to flux transfer. We remark that the NFG is distinct from directed analogues of the RAG constructed from boolean production and consumption stoichiometric matrices, as shown in Sec. SI 1.

We now extend the construction of the NFG to accommodate specific environmental contexts or growth conditions.

Mass flow graphs: incorporating information of the biological context

Cells adjust their metabolic fluxes to respond to the availability of nutrients and environmental requirements. Flux Balance Analysis (FBA) is a widely used method to predict environment-specific flux distributions. FBA computes a vector of metabolic fluxes \mathbf{v}^* that maximise a cellular objective (e.g., biomass, growth or ATP production). The FBA solution is obtained assuming steady state conditions ($\dot{\mathbf{x}} = 0$ in Eq. (2)) subject to constraints that describe the availability of nutrients and other extracellular compounds.¹⁶ The core elements of FBA are briefly summarised in the Appendix A.1.

To incorporate the biological information afforded by FBA solutions into the structure of a metabolic graph, we again define the graph edges in terms of production and consumption fluxes.

Similarly to Eq. (4), we unfold the FBA solution vector \mathbf{v}^* into forward and backward components: positive entries in the FBA solution correspond to forward fluxes, negative entries in the FBA solution correspond to backward fluxes. From the unfolded fluxes

$$\mathbf{v}_{2m}^* = \begin{bmatrix} \mathbf{v}^{*+} \\ \mathbf{v}^{*-} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \text{abs}(\mathbf{v}^*) + \mathbf{v}^* \\ \text{abs}(\mathbf{v}^*) - \mathbf{v}^* \end{bmatrix},$$

we compute the $2m \times 1$ vector of production and consumption fluxes as

$$\mathbf{j}(\mathbf{v}^*) = \mathbf{S}_{2m}^+ \mathbf{v}_{2m}^* = \mathbf{S}_{2m}^- \mathbf{v}_{2m}^*. \quad (9)$$

The k -th entry of $\mathbf{j}(\mathbf{v}^*)$ is the flux at which metabolite X_k is produced and consumed; the equality of the production and consumption fluxes follows from the steady state condition, $\dot{\mathbf{x}} = 0$ (i.e., the fluxes are balanced).

To construct the flow graph, we define the weight of the edge between reactions R_i and R_j as the total mass flow of metabolites produced by R_i that are consumed by R_j . Assuming that the amount of metabolite produced by one reaction is distributed among the reactions that consume it in proportion to their flux (and respecting the stoichiometry), the flux of metabolite X_k from reaction R_i to R_j is given by

$$\text{Flow of } X_k \text{ from } R_i \text{ to } R_j = (\text{flow of } X_k \text{ produced by } R_i) \times \left(\frac{\text{flow of } X_k \text{ consumed by } R_j}{\text{total consumption flow of } X_k} \right). \quad (10)$$

For example, if the total flux of metabolite X_k is 10 mmol/gDW/h, with reaction R_i producing X_k at a rate 1.5 mmol/gDW/h and reaction R_j consuming X_k at a rate 3.0 mmol/gDW/h, then the flow of X_k from R_i to R_j is 0.45 mmol/gDW/h.

Summing (10) over all metabolites, we obtain the edge weight relating reactions R_i and R_j :

$$M_{ij}(\mathbf{v}^*) = \sum_{k=1}^n s_{ki}^+ v_{2m_i}^* \times \left(\frac{s_{kj}^- v_{2m_j}^*}{\sum_{j=1}^{2m} s_{kj}^- v_{2m_j}^*} \right). \quad (11)$$

In matrix form, these edge weights are collected into the adjacency matrix of the

$$\text{Mass Flow Graph (MFG): } \mathbf{M}(\mathbf{v}^*) = (\mathbf{S}_{2m}^+ \mathbf{V}^*)^T \mathbf{J}_v (\mathbf{S}_{2m}^- \mathbf{V}^*), \quad (12)$$

where $\mathbf{V}^* = \text{diag}(\mathbf{v}_{2m}^*)$, $\mathbf{J}_v = \text{diag}(\mathbf{j}(\mathbf{v}^*))$ and \dagger denotes the matrix pseudoinverse. The MFG is a directed, weighted graph with edge weights in units of mmol/gDW/h. Self-loops describe the metabolic flux of autocatalytic reactions, i.e., those in which products are also reactants.

The MFG provides a versatile framework to create environment-specific metabolic graphs from FBA solutions. In Fig. 1c, we illustrate the creation of MFGs for a toy network under different biological scenarios. In each case, an FBA solution is computed under a fixed uptake flux with the remaining fluxes constrained to account for differences in the biological environment: in scenario 1, the fluxes are constrained to be strictly positive and no larger than the nutrient uptake flux, while in scenario 2 we impose a positive lower bound on reaction R_7 . Note how the MFG for scenario 2 displays an extra edge between reactions R_4 and R_7 , as well as distinct edge weights to scenario 1 (see Sec. SI 2 for details). These differences illustrate how changes in the FBA solutions translate into different graph connectivities and edge weights.

Graphs for *Escherichia coli* core metabolism

To illustrate our framework, we construct and analyse the flow graphs of the well-studied core metabolic model of *E. coli*.³³ This model (Fig. 2a) contains 72 metabolites and 95 reactions, grouped into 11 pathways, which describe the main biochemical routes in central carbon metabolism.^{39–41} We provide a Supplemental

Spreadsheet with full details of the reactions and metabolites in this model, as well as all the results presented below.

The Normalised Flow Graph: the impact of directionality

To examine the effect of flux directionality on the metabolic graphs, we compare the Reaction Adjacency Graph (**A**) and our proposed Normalised Flow Graph (**D**) for the same metabolic model in Fig. 2. The **A** graph has 95 nodes and 1,158 undirected edges, whereas the **D** graph has 154 nodes and 1,604 directed and weighted edges. The increase in node count is due to the unfolding of reversible reactions into forward and backward reaction nodes. Unlike the **A** graph, where edges represent shared metabolites between two reactions, the directed edges of the **D** graph represent the flow of metabolites from a source to a target reaction. A salient feature of both graphs is their high connectivity, which is not apparent from the traditional pathway representation in Fig. 2a.

The effect of directionality becomes apparent when comparing the importance of reaction nodes in both graphs (Fig. 2b–d), as measured with the PageRank score for node centrality.^{42,43} The overall node hierarchy is maintained across both graphs: exchange reactions tend to have low PageRank centrality scores, core metabolic reactions have high scores, and the biomass reaction has the highest scores in both graphs. Yet we also observe substantial changes in several reactions. For example, the reactions for ATP maintenance (ATPM, irreversible), phosphoenolpyruvate synthase (PPS, irreversible) and ABC-mediated transport of L-glutamine (GLNabc, irreversible) drop from being among the top 10% most important reactions in the **A** graph to the bottom percentiles in the **D** graph. Conversely, reactions such as aconitase A (ACONTa, irreversible), transaldolase (TALA, reversible) and succinyl-CoA synthetase (SUCCoAS, reversible), and formate transport via diffusion (FORTi, irreversible) gain substantial importance in the **D** graph. For instance, FORTi is the sole consumer of formate, which is produced by pyruvate formate lyase (PFL), a reaction that is highly connected to the rest of the network. Importantly, in most of the reversible reactions, such as ATP synthase (ATPS4r), there is a wide gap between the PageRank of the forward and backward reactions, suggesting a marked asymmetry in the importance of metabolic flows.

Community detection is frequently used in the analysis of complex graphs: nodes are clustered into tightly related communities that reveal the coarse-grained structure of the graph, potentially at different levels of resolution.^{44–46} The community structure of metabolic graphs has been the subject of multiple analyses.^{12,14,44} However, most community detection methods are applicable to undirected graphs only, and thus fail to capture the directionality of the metabolic graphs we propose here. To account for graph directionality, we use the Markov Stability community detection framework,^{46–48} which uses diffusion on graphs to detect groups of nodes where flows are retained persistently across time scales. Markov Stability is ideally suited to find multi-resolution community structure⁴⁵ and can deal with both directed and undirected graphs^{46,49} (see “Methods” section). In the case of metabolic graphs, Markov Stability can reveal groups of reactions that are closely interlinked by the flow of metabolites that they produce and consume.

Figure 3 shows the difference between the community structure of the undirected RAG and the directed NFG of the core metabolism of *E. coli*. For the **A** graph, Markov Stability reveals a partition into seven communities (Fig. 3b, see also Supplemental Sec. SI 3), which are largely dictated by the many edges created by shared pool metabolites. For example, community C 1(**A**) is mainly composed of reactions that consume or produce ATP and water. Yet, the biomass reaction (the largest consumer of ATP) is not a member of C 1(**A**) because, in the standard **A** graph construction, any connection involving ATP has

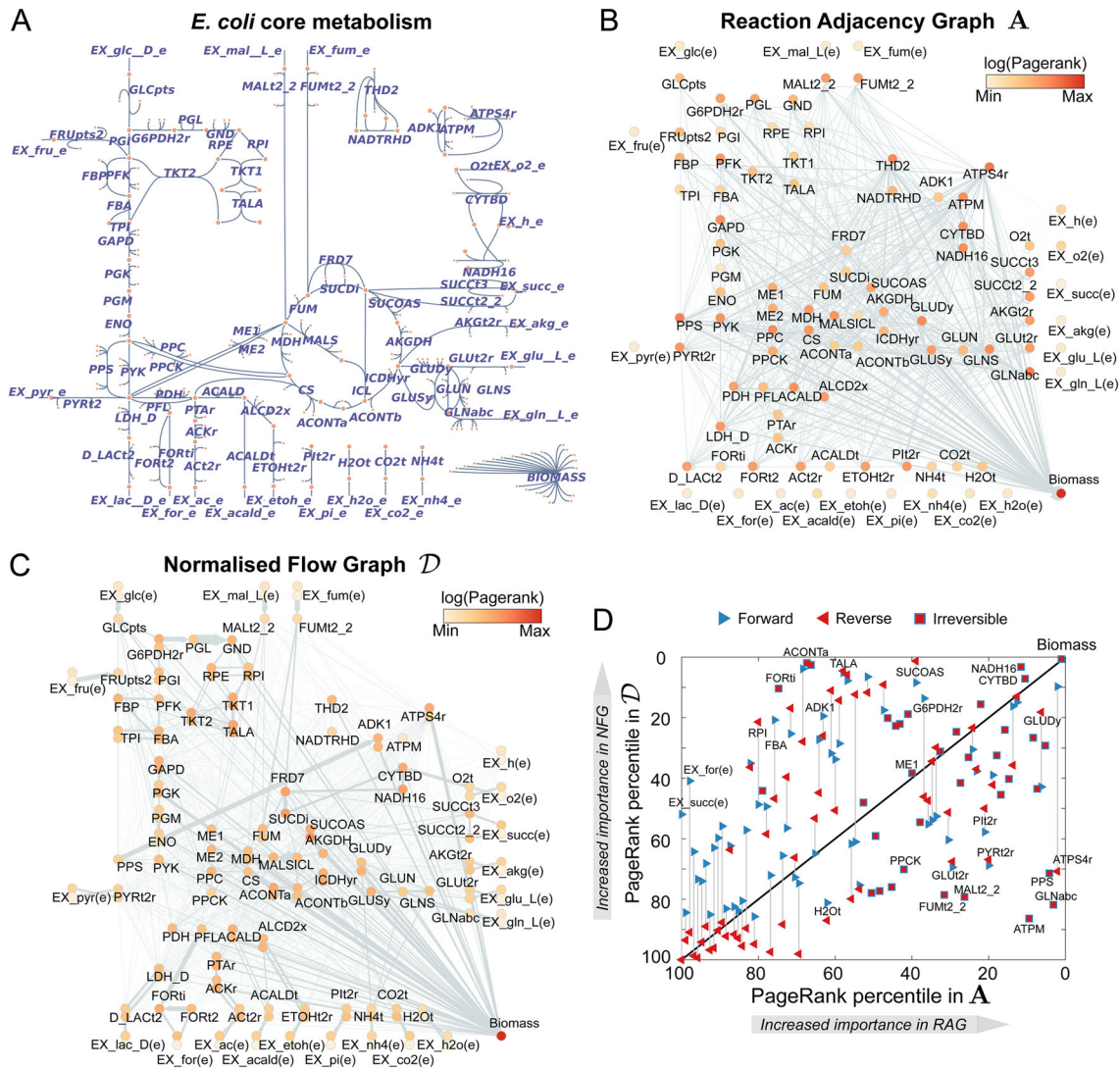


Fig. 2 Graphs for the core metabolism of *Escherichia coli*. **a** Map of the *E. coli* core metabolic model created with the online tool Escher.^{33,54} **b** The standard Reaction Adjacency Graph **A**, as given by Eq. (3). The nodes represent reactions; two reactions are linked by an undirected edge if they share reactants or products. The nodes are coloured according to their PageRank score, a measure of their centrality (or importance) in the graph. **c** The directed Normalised Flow Graph **D**, as computed from Eq. (8). The reversible reactions are unfolded into two overlapping nodes (one for the forward reaction, one for the backward). The directed links indicate flow of metabolites produced by the source node and consumed by the target node. The nodes are coloured according to their PageRank score. **d** Comparison of PageRank percentiles of reactions in **A** and **D**. Reversible reactions are represented by two triangles connected by a line; both share the same PageRank in **A**, but each has its own PageRank in **D**. Reactions that appear above (below) the diagonal have increased (decreased) PageRank in **D** as compared to **A**.

equal weight. Other communities in **A** are also determined by pool metabolites, e.g. C 2(**A**) is dominated by H^+ , and C 3(**A**) is dominated by NAD^+ and $NADP^+$, as illustrated by word clouds of the relative frequency of metabolites in the reactions within each community. The community structure in **A** thus reflects the limitations of the RAG construction due to the absence of biological context and the large number of uninformative links introduced by pool metabolites.

For the **D** graph, we found a robust partition into five communities (Fig. 3c, Supplementary Sec. SI 3), which comprise reactions related consistently through biochemical pathways. Community C1(**D**) contains the reactions in the pentose phosphate pathway together with the first steps of glycolysis involving D-fructose, D-glucose, or D-ribulose. Community C2(**D**) contains the main reactions that produce ATP from substrate level as well as oxidative phosphorylation and the biomass reaction. Community C3(**D**) includes the core of the citric acid cycle, anaplerotic reactions related to malate syntheses, as well as the

intake of cofactors such as CO_2 . Community C4(**D**) contains reactions that are secondary sources of carbon (such as malate and succinate), as well as oxidative phosphorylation reactions. Finally, community C5(**D**) contains reactions that are part of the pyruvate metabolism subsystem, as well as transport reactions for the most common secondary carbon metabolites such as lactate, formate, acetaldehyde and ethanol. Altogether, the communities of the **D** graph reflect metabolite flows associated with specific cellular functions, as a consequence of including flux directionality in the graph construction. As seen in Fig. 3c, the communities are no longer exclusively determined by pool metabolites (e.g., water is no longer dominant and protons are spread among all communities). For a more detailed explanation and comparison of the communities found in the **A** and **D** graphs, see Supplementary Section SI 3. Full information about PageRank scores and communities is provided in the Supplementary Spreadsheet.

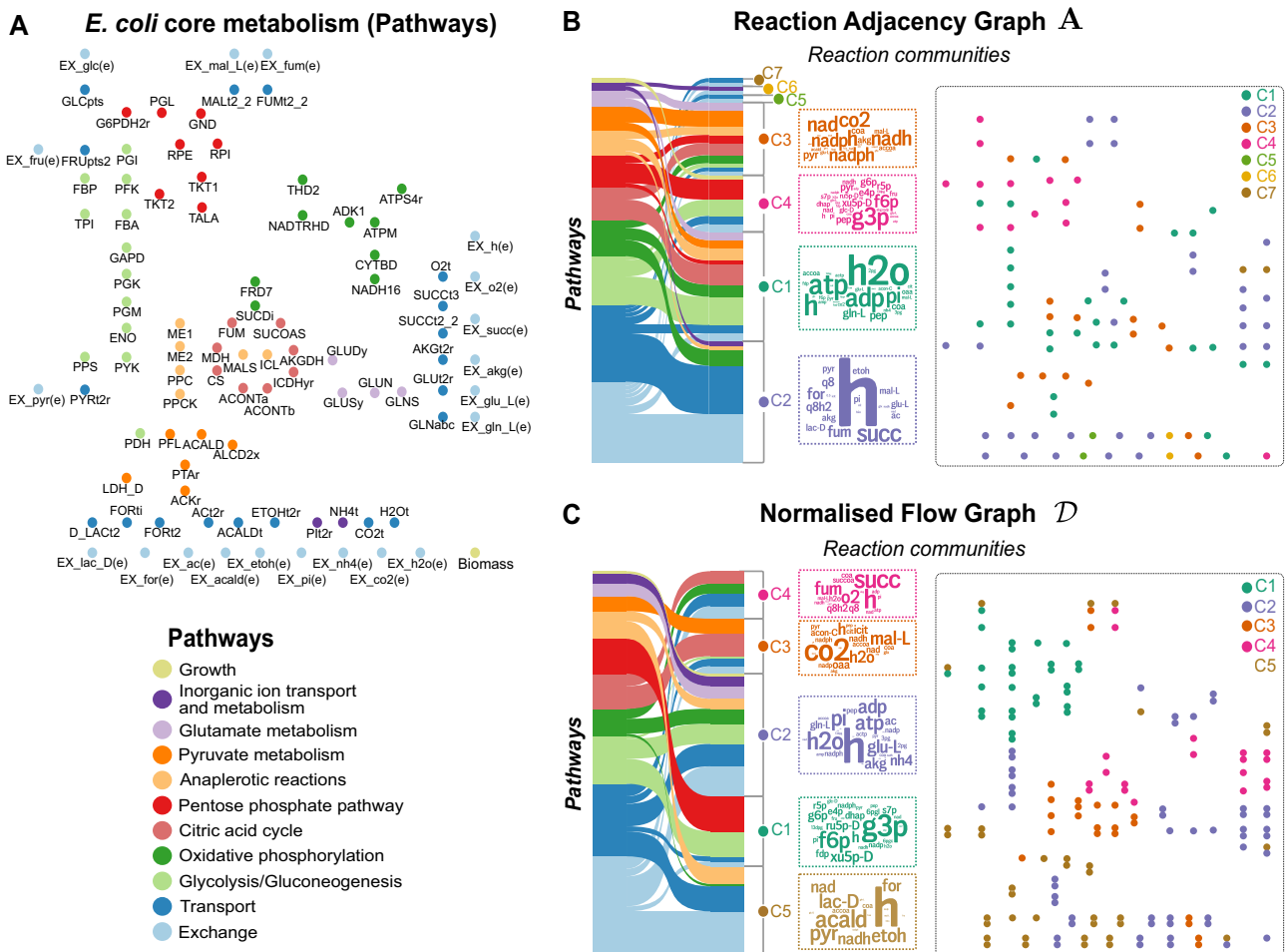


Fig. 3 Directionality and community structure of graphs for *Escherichia coli* metabolism. **a** Reactions of the core model of *E. coli* metabolism grouped into eleven biochemical pathways, **b, c** Graphs **A** and **D** from Fig. 2b–c partitioned into communities computed with the Markov Stability method; for clarity, the graph edges are not shown. The Sankey diagrams^{68,69} show the correspondence between biochemical pathways and the communities found in each graph. The word clouds contain the metabolites that participate in the reactions of each community, and the word size is proportional to the number of reactions in which each metabolite participates

Mass Flow Graphs: the impact of growth conditions and biological context

To incorporate the impact of environmental context, we construct the Mass Flow Graphs in Eq. (12) using FBA solutions of the core model of *E. coli* metabolism in four relevant growth conditions: aerobic growth in rich media with glucose; aerobic growth in rich media with ethanol, anaerobic growth in glucose; and aerobic growth in glucose but phosphate- and ammonium-limited. The results in Fig. 4 show how changes in metabolite fluxes under different biological contexts have a direct effect in the MFG. Note that, in all cases, the MFGs have fewer nodes than the blueprint graph **D** since the FBA solutions contain numerous reactions with zero flux.

Next we summarise how the changes in the community structure of the MFGs for the four conditions reflect the distinct relationships of functional pathways in response to growth requirements.

Aerobic growth in D-glucose (M_{glc}). We found a robust partition into three communities with an intuitive biological interpretation (Fig. 4a and Supplementary Fig. S12A). C1(M_{glc}) is the carbon-processing community, comprising reactions that process carbon from D-glucose to pyruvate including most of the glycolysis and pentose phosphate pathways, together with related transport and exchange reactions. C2(M_{glc}) harbours the bulk of reactions

related to oxidative phosphorylation and the production of energy in the cell, including the electron transport chain of NADH dehydrogenase, cytochrome oxidase, and ATP synthase, as well as transport reactions for phosphate and oxygen intake and proton balance. C2(M_{glc}) also includes the growth reaction, consistent with ATP being the main substrate for both the ATP maintenance (ATPM) requirement and the biomass reaction in this growth condition. Finally, C3(M_{glc}) contains reactions related to the citric acid cycle (TCA) and the production of NADH and NADPH (i.e., the cell's reductive power), together with carbon intake routes strongly linked to the TCA cycle, such as those starting from phosphoenolpyruvic acid (PEP).

Aerobic growth in ethanol (M_{etoh}). The robust partition into three communities that we found for this scenario resembles the structure of M_{glc} with subtle, yet important, differences (Fig. 4b and Supplementary Fig. S12B). Most salient are the differences in the carbon-processing community C1(M_{etoh}), which reflects the switch from D-glucose to ethanol as a carbon source. C1(M_{etoh}) contains gluconeogenic reactions (instead of glycolytic), due to the reversal of flux induced by the change of carbon source, as well as anaplerotic reactions and reactions related to glutamate metabolism. In particular, the reactions in this community are related to the production of precursors such as PEP, pyruvate, 3-phospho-D-glycerate (3PG), glyceraldehyde-3-phosphate (G3P),

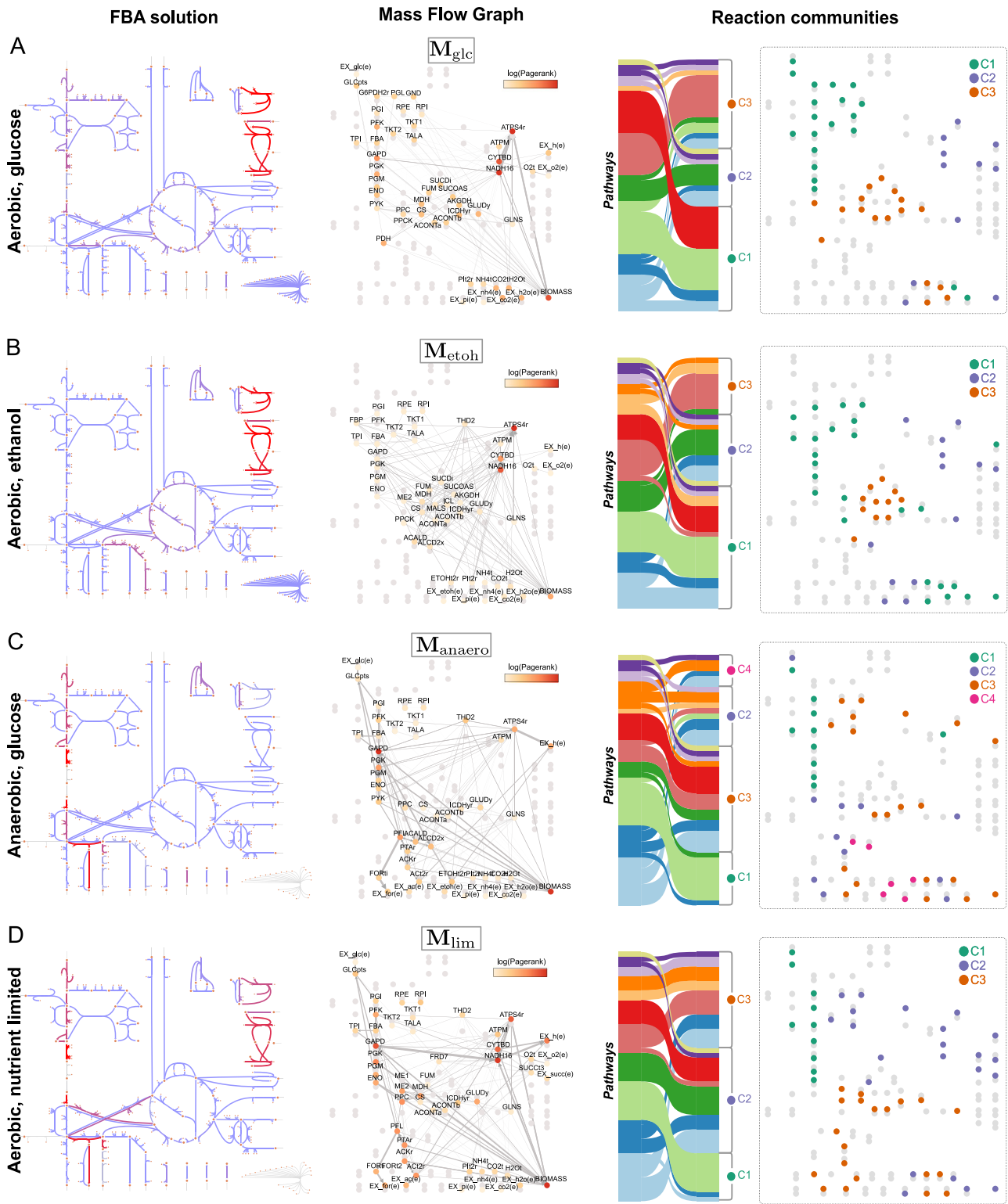


Fig. 4 Mass Flow Graphs for *Escherichia coli* in different growth conditions. The MFGs are computed from Eq. (12) and the FBA solutions in four different environments: **a** Aerobic growth in D -glucose, **b** aerobic growth in ethanol, **c** anaerobic growth in D -glucose, and **d** aerobic growth in D -glucose but with limited ammonium and phosphate. Each subfigure shows: (left) flux map obtained with Escher, ⁵⁴ where the increased red colour of the arrows indicates increased flux; (centre) Mass Flow Graph with nodes coloured according to their PageRank (zero flux reactions are in grey; thickness of connections proportional to fluxes); (right) community structure computed with the Markov Stability method together with Sankey diagrams showing the correspondence between biochemical pathways and MFG communities

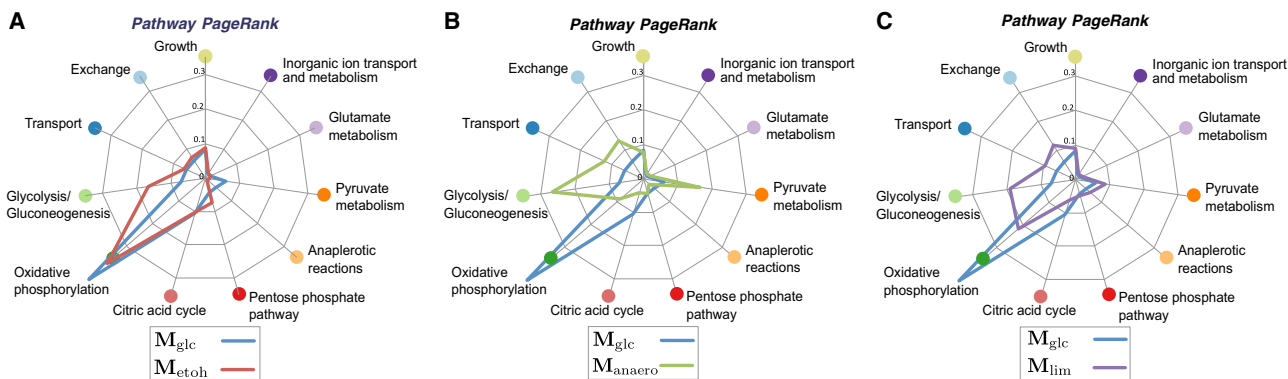


Fig. 5 Pathway centrality (PageRank) computed from the MFG of different growth conditions. The cumulative pathway PageRank reflects the relative importance of metabolic pathways in each MFG. Changes in pathway centrality indicate the overall rearrangement of fluxes within the pathways in response to environmental shifts: **a** from aerobic glucose-rich to aerobic ethanol-rich; **b** from aerobic glucose-rich to anaerobic glucose-rich; **c** from aerobic glucose-rich to a similar medium with limited phosphate and ammonium. Variations in cumulative PageRank highlight changes across most cellular pathways

D-fructose-6-phosphate (F6P), and D-glucose-6-phosphate, all of which are substrates for growth. Consequently, the biomass reaction is also grouped within C1(M_{etoH}) due to the increased metabolic flux of precursors relative to ATP production in this biological scenario. The other two reaction communities (energy-generation C2(M_{etoH}) and citric acid cycle C3(M_{etoH})) display less prominent differences relative to the M_{glc} graph, with additional pyruvate metabolism and anaplerotic reactions as well as subtle ascriptions of reactions involved in NADH/NADPH balance and the source for acetyl-CoA.

Anaerobic growth in D-glucose (M_{anaero}). The profound impact of the absence of oxygen on the metabolic balance of the cell is reflected in drastic changes in the MFG (Fig. 4c and Supplementary Fig. SI 2C). Both the connectivity and reaction communities in the MFG are starkly different from the aerobic scenarios, with a much diminished presence of oxidative phosphorylation and the absence of the first two steps of the electron transport chain (CYTBD and NADH16). We found that M_{anaero} has a robust partition into four communities. C1(M_{anaero}) still contains carbon processing (glucose intake and glycolysis), yet now decoupled from the pentose phosphate pathway. C3(M_{anaero}) includes the pentose phosphate pathway grouped with the citric acid cycle (incomplete) and the biomass reaction, as well as the growth precursors including alpha-D-ribose-5-phosphate (r5p), D-erythrose-4-phosphate (e4p), 2-oxalacetate and NADPH. The other two communities are specific to the anaerobic context: C2(M_{anaero}) contains the conversion of PEP into formate (more than half of the carbon secreted by the cell becomes formate⁵⁰); C4(M_{anaero}) includes NADH production and consumption via reactions linked to glyceraldehyde-3-phosphate dehydrogenase (GAPD).

Aerobic growth in D-glucose but limited phosphate and ammonium (M_{lim}). Under growth-limiting conditions, we found a robust partition into three communities (Fig. 4d and Supplementary Fig. SI 2D). The community structure reflects *overflow metabolism*,⁵¹ which occurs when the cell takes in more carbon than it can process. As a consequence, the excess carbon is secreted from the cell, leading to a decrease in growth and a partial shutdown of the citric acid cycle. This is reflected in the reduced weight of the TCA pathway and its grouping with the secretion routes of acetate and formate within C3(M_{lim}). Hence, C3(M_{lim}) comprises reactions that are not strongly coupled in favourable growth conditions, yet are linked together by metabolic responses to limited ammonium and phosphate. Furthermore, the carbon-processing community C1 (M_{lim}) contains the glycolytic pathway, yet detached from the pentose phosphate pathway (as in M_{anaero}), highlighting its role in

precursor formation. The bioenergetic machinery, contained in community C2(M_{lim}), includes the pentose phosphate pathway, with a smaller role for the electron transport chain (21.8% of the total ATP as compared to 66.5% in M_{glc}).

In addition to the effect on community structure, Fig. 4 also shows the changes induced by the environment on the MFG connectivity and relative importance of reactions, as measured by their PageRank score. To provide a global snapshot of the effect of growth conditions on cellular metabolism, Fig. 5 shows the cumulative PageRank of each pathway for each of the MFGs. The cumulative PageRank quantifies the relative importance of pathways, and how their importance changes upon environmental shifts.

In aerobic growth, a shift from glucose to ethanol ($M_{\text{glc}} \rightarrow M_{\text{etoH}}$) as carbon source decreases the importance of pyruvate metabolism and oxidative phosphorylation, while increasing the importance of the pentose phosphate pathway. A shift from aerobic to anaerobic growth in glucose ($M_{\text{glc}} \rightarrow M_{\text{anaero}}$) sees a large reduction in the importance of oxidative phosphorylation and the citric acid cycle, coupled with a large increase in the importance of gluconeogenesis, pyruvate metabolism, and transport and exchange reactions. The effect of growth-limiting conditions in aerobic growth under glucose ($M_{\text{glc}} \rightarrow M_{\text{lim}}$) is reflected on the increased importance of pyruvate metabolism and a reduction in the importance of oxidative phosphorylation, citric acid cycle, and the pentose phosphate pathway. The importance of transport and exchange reactions is also increased under limiting conditions. Such qualitative relations between growth conditions and the importance of specific pathways highlights the utility of the MFGs to characterise systemic metabolic changes in response to environmental conditions.

A more detailed discussion of the changes in pathways and reactions can be found in Section SI 4 and Fig. SI 2 in the Supplementary Information, with full details of all the results in the Supplemental Spreadsheet.

Multiscale organisation of mass flow graphs

The definition of the MFGs as *directed graphs* opens up the application of network-theoretic tools for detecting modules of reaction nodes and the hierarchical relationships among them.

In contrast with methods for undirected graphs, the Markov Stability framework^{47,52} can be used to detect multi-resolution community structure in directed graphs (Sec. 4.2), thus allowing the exploration of the multiscale organisation of metabolic reaction networks. The modules so detected reflect subsets of reactions where metabolic fluxes tend to be contained.

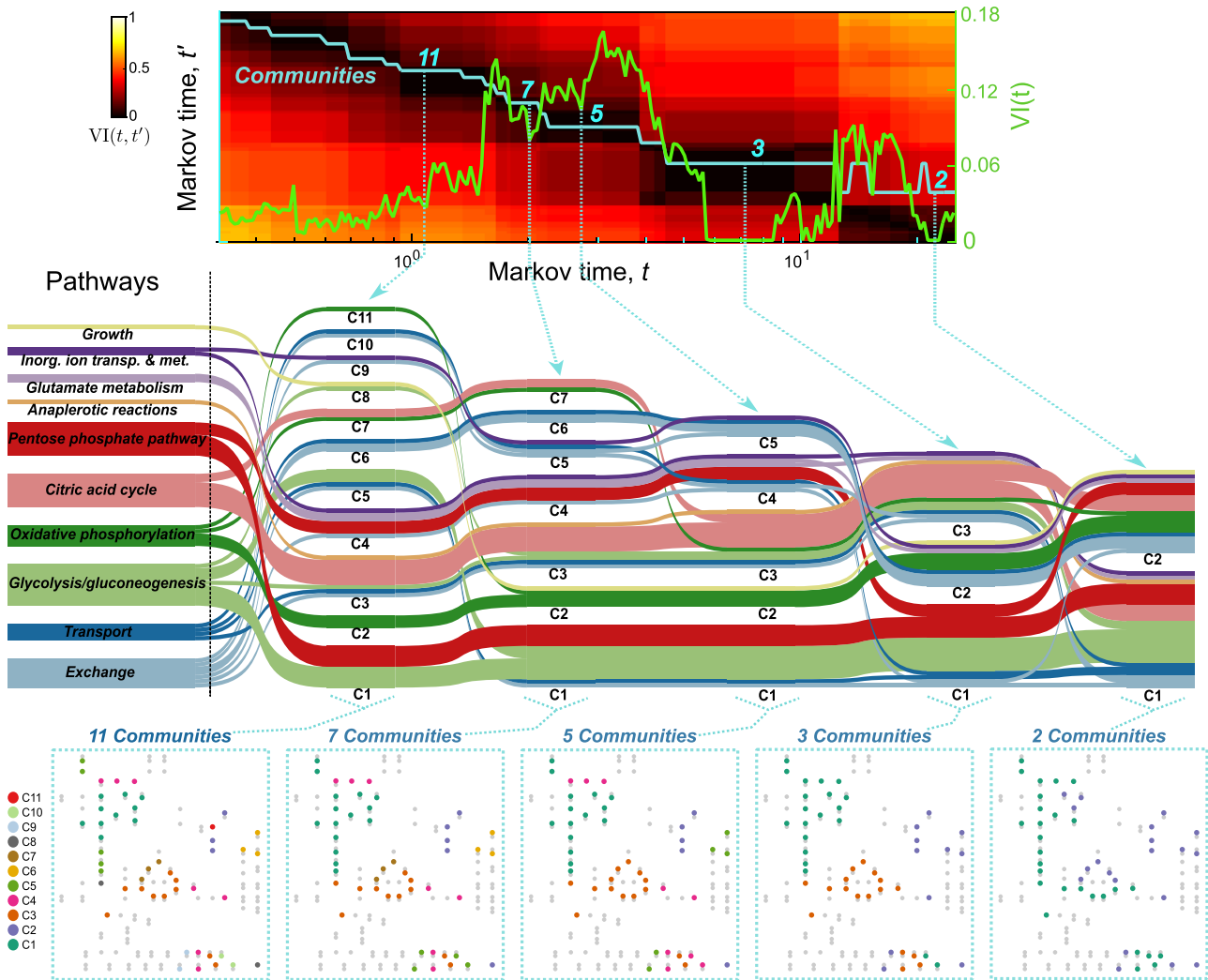


Fig. 6 Community structure of flow graphs across different scales. We applied the Markov Stability method to partition the mass flow graph for *E. coli* aerobic growth in glucose (\mathbf{M}_{glc}) across levels of resolution. The top panel shows the number of communities of the optimal partition (blue line) and two measures of its robustness ($VI(t)$ (green line) and $VI(t, t')$ (colour map)) as a function of the Markov time t (see text and “Methods” section). The five Markov times selected correspond to robust partitions of the graph into 11, 7, 5, 3, and 2 communities, as signalled by extended low values of $VI(t, t')$ and low values (or pronounced dips) of $VI(t)$. The Sankey diagram (middle panel) visualises the multiscale organisation of the communities of the MFG across Markov times, and the relationship of the communities with the biochemical pathways. The bottom panel shows the five partitions at the selected Markov times. The partition into 3 communities corresponds to that in Fig. 4A

Figure 6 illustrates this multiscale analysis on \mathbf{M}_{glc} , the MFG of *E. coli* under aerobic growth in glucose. By varying the Markov time t , a parameter in the Markov Stability method, we scanned the community structures at different resolutions. Our results show that, from finer to coarser resolution, the MFG can be partitioned into 11, 7, 5, 3, and 2 communities of high persistence across Markov time (extended plateaux over t , as shown by low values of $VI(t, t')$) and high robustness under optimisation (as shown by dips in $VI(t)$). For further details, see Section 4.2 and Refs.^{45–47,52}

The Sankey diagram in Fig. 6 visualises the pathway composition of the graph partitions and their relationships across different resolutions. As we decrease the resolution (longer Markov times), the reactions in different pathways assemble and split into different groupings, reflecting both specific relationships and general organisation principles associated with this growth condition. A general observation is that glycolysis is grouped together with oxidative phosphorylation across most scales, underlining the fact that those two pathways function as cohesive metabolic sub-units in aerobic conditions. In contrast, the

exchange and transport pathways appear spread among multiple partitions across all resolutions. This is expected, as exchange/transport are enabling functional pathways, in which reactions do not interact amongst themselves but rather feed substrates to other pathways.

Other reaction groupings reflect more specific relationships. For example, the citric acid cycle (always linked to anaplerotic reactions) appears as a cohesive unit across most scales, and only splits in two in the final grouping, reflecting the global role of the TCA cycle in linking to both glycolysis and oxidative phosphorylation. The pentose phosphate pathway, on the other hand, is split into two groups (one linked to glutamate metabolism and another one linked to glycolysis) across early scales, only merging into the same community towards the final groupings. This suggests a more interconnected flux relationship of the different steps of the pentose phosphate pathway with the rest of metabolism. In Supplementary Figure SI2, we present the multiscale analyses of the reaction communities for the other three growth scenarios (\mathbf{M}_{etoh} , $\mathbf{M}_{\text{anaer}}$, \mathbf{M}_{lim}).

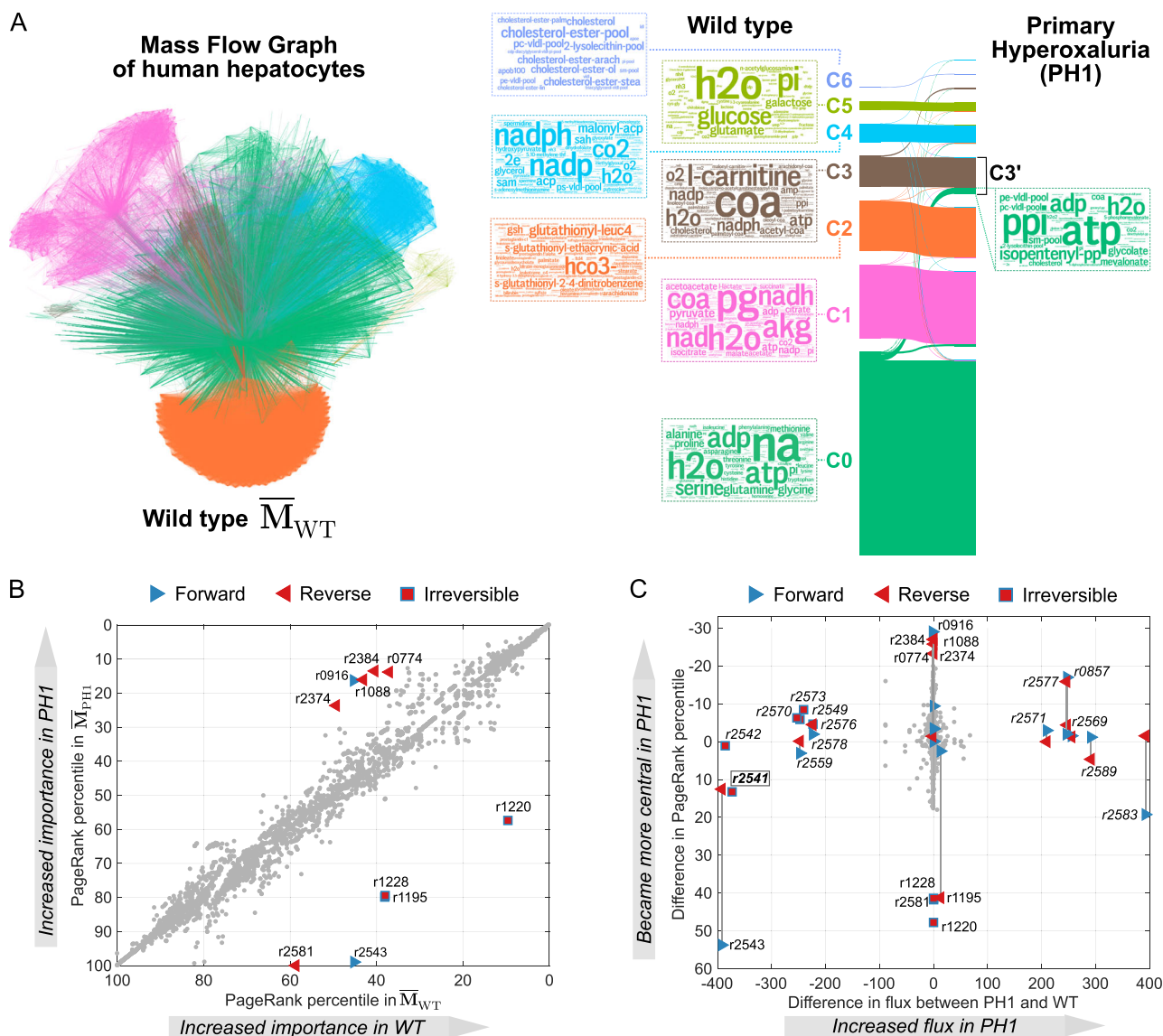


Fig. 7 MFG analysis of a model of human hepatocyte metabolism and the genetic condition PH1. **a** Average MFG of wild-type hepatocytes cells over 442 metabolic objectives. The reaction nodes are coloured according to communities in a 7-way partition obtained with Markov Stability. The Sankey diagram shows the consistency between the communities in the wild-type MFG and the communities independently found in the MFG of the mutated PH1 cells. Word clouds of the most frequent metabolites in the reactions of the WT communities reveal functional groupings (see text). Under the PH1 mutation, the only large change relates to metabolites that join C3' from community C0 in WT. **b** Comparison of the PageRank percentiles in the WT and PH1 MFGs, with reactions whose rank changes by more than 20 percentiles labelled. **c** Difference in FBA flux between WT and PH1 vs difference in PageRank percentile between WT and PH1. Reactions whose flux difference is greater than 100 mmol/gDW/h (italics) or whose change in PageRank percentile is greater than 20 are labelled. The differences in centrality (PageRank) provide complementary information, revealing additional important reactions affected by the PH1 mutation that knocks out reaction r2541

Using MFGs to analyse hepatocyte metabolism in wild type and PH1 mutant human cells

To showcase the applicability of our framework to larger metabolic models, we analyse a model of human hepatocyte (liver) metabolism with 777 metabolites and 2589 reactions,³⁴ which extends the widely used HepatoNet1 model⁵³ with an additional 50 reactions and 8 metabolites. This extended model was used in Ref.³⁴ to compare wild-type cells (WT) and cells affected by the rare disease Primary Hyperoxaluria Type 1 (PH1), which lack alanine:glyoxylate aminotransferase (AGT) due to a genetic mutation. AGT is an enzyme found in peroxisomes and its mutation decreases the breakdown of glyoxylate, with subsequent accumulation of calcium oxalate that leads to liver damage.

Following,³⁴ we first obtain 442 FBA solutions for different sets of metabolic objectives for both the wild-type (WT) model and the PH1 model lacking AGT (reaction r2541). We then generate the corresponding 442 MFGs for each WT and PH1, and obtain the averages over each ensemble: \bar{M}_{WT} and \bar{M}_{PH1} . Of the 2589 reactions in the model, 2448 forward and 1362 reverse reactions are present in at least one of the FBA solutions. Hence the average MFGs have 3810 nodes each (see Supplementary Spreadsheet for full details about the reactions).

Figure 7a shows the MFG for the wild-type (\bar{M}_{WT}) coloured according to a robust partition into 7 communities obtained with Markov Stability. The seven communities are broadly linked to amino acid metabolism (C0), energy metabolism (C1 and C5), glutathione metabolism (C2), fatty acid and bile acid metabolism

(C3 and C4) and cholesterol metabolism and lipoprotein particle assembly (C6). As expected, the network community structure of the MFG is largely preserved under the AGT mutation: the Sankey diagram in Fig. 7b shows a remarkable match between the partitions of $\bar{\mathbf{M}}_{\text{WT}}$ and $\bar{\mathbf{M}}_{\text{PH1}}$ found independently with Markov Stability. Despite this similarity, our method also identified subtle but important differences between the healthy and diseased networks. In particular, C3' in $\bar{\mathbf{M}}_{\text{PH1}}$ receives 60 reactions, almost all taking place in the peroxisome and linked to mevalonate and isopentenyl pathways, as well as highly central transfer reactions of PP_i, O₂ and H₂O₂ between the peroxisome and the cytosol (r1152, r0857, r2577).

Overall, the centrality (PageRank) of most reactions in the MFG is relatively unaffected by the PH1 mutation, as shown by the good correlation between the PageRank percentiles in $\bar{\mathbf{M}}_{\text{WT}}$ and $\bar{\mathbf{M}}_{\text{PH1}}$ in Fig. 7c. Yet, there are notable exceptions, and the reactions that exhibit the largest change in PageRank centrality (labelled in Fig. 7b) provide biological insights into the disease state. Specifically, the four reactions (r0916, r1088, r2384, r2374) that undergo the largest increase in centrality from $\bar{\mathbf{M}}_{\text{WT}}$ to $\bar{\mathbf{M}}_{\text{PH1}}$ are related to the transfer of citrate out of the cytosol in exchange for oxalate and PEP; whereas those with the largest decrease of PageRank from $\bar{\mathbf{M}}_{\text{WT}}$ to $\bar{\mathbf{M}}_{\text{PH1}}$ are related to VLDL-pool reactions (r1228, r1195, r1220) and to transfers of hydroxypyruvate and alanine from peroxisome to cytosol (r2581, r2543).

It is worth remarking that although oxalate and citrate reactions are directly linked to metabolic changes associated with the PH1 diseased state, none of them exhibits large changes in their flux predicted by FBA, yet they show large changes in PageRank centrality.

These observations underscore how the information provided by our network analysis provides complementary information to the results from FBA. As shown in Fig. 7d, there is a group of reactions (labelled with italics in the Figure) that exhibit large gains or decreases in their flux under the PH1 mutation, yet they only undergo relatively small changes in their PageRank scores. Closer inspection reveals that most of these reactions are close to the AGT reaction (r2541, highlighted in the Figure) in the pathway and involve the conversion of glycolate, pyruvate, glycine, alanine and serine. Hence the changes in flux follow from the *local rearrangement* of flows as a consequence of the deletion of reaction r2541. On the other hand, the citrate and oxalate reactions (r0916, r1088, r2384, r2374) with large changes in their centrality yet undergo small changes in flux, thus reflecting *global changes* in the flux structure of the network. Importantly, the transport reactions of O₂, H₂O₂, serine and hydroxypyruvate between cytosol and peroxisome (r0857, r2577, r2583, r2543) all undergo large changes both in centrality and flux, highlighting the importance of peroxisome transfer reactions in PH1. We provide a full spreadsheet with these analyses as Supplementary Material for the interested reader.

DISCUSSION

Metabolism is commonly understood in terms of functional pathways interconnected into metabolic networks,⁵⁴ i.e., metabolites linked by arrows representing enzymatic reactions between them as in Fig. 2a. However, such standard representations are not amenable to rigorous graph-theoretic analysis. Fundamentally different graphs can be constructed from such metabolic reactions depending on the chosen representation of species/interactions as nodes/edges, e.g., reactions as nodes; metabolites as nodes; or both reaction and metabolites as nodes.¹⁶ Each of those graphs can be directed or undirected and with weighted links computed according to different rules. The choices and subtleties in graph construction are crucial both to capture the relevant metabolic information and to interpret their topological properties.^{10,17}

Here we have presented a flux-based strategy to build graphs for metabolic networks. Our graphs have reactions as nodes and directed weighted edges representing the flux of metabolites produced by a source reaction and consumed by a target reaction. This principle is applied to build both 'blueprint' graphs (NFG), which summarise probabilistically the fluxes of the whole metabolism of an organism, as well as context-specific graphs (MFGs), which reflect specific environmental conditions. The blueprint Normalised Flow Graph has edge weights equal to the probability that source/target reactions produce/consume a molecule of a metabolite chosen at random from the stoichiometric matrix in the absence of any other information, and can thus be used when this matrix is the only information available. The NFG construction naturally tames the over-representation of pool metabolites without the need to remove them from the graph arbitrarily, as often done in the literature.^{26,28–30} Context-specific Mass Flow Graphs (MFGs) can incorporate the effect of the environment, e.g., with edge weights corresponding to the total flux of metabolites between reactions as computed by Flux Balance Analysis (FBA). FBA solutions for different environments can then be used to build different metabolic graphs in different growth conditions.

The proposed graph constructions provide complementary tools for studying the organisation of metabolism and can be embedded into any FBA-based modelling pipeline. Specifically, the NFG relies on the availability of a well-curated stoichiometric matrix, which is produced with metabolic reconstruction techniques that typically precede the application of FBA. The MFG, on the other hand, explicitly uses the FBA solutions in its construction. Both methods provide a systematic framework to convert genome-scale metabolic models into a directed graph on which analysis tools from network theory can be applied.

To exemplify our approach, we built and analysed NFG and MFGs for the core metabolism of *E. coli*. Through the analysis of topological properties and community structure of these graphs, we highlighted the importance of weighted directionality in metabolic graph construction, and revealed the flow-mediated relationships between functional pathways under different environments. In particular, the MFGs capture specific metabolic adaptations such as the glycolytic-gluconeogenic switch, overflow metabolism, and the effects of anoxia. The proposed graph construction can be readily applied to large genome-scale metabolic networks.^{12,19,21,22,39}

To illustrate the scalability of our analyses to larger metabolic models, we studied a genome-scale model of a large metabolic model of human hepatocytes with around 3000 reactions in which we compared the wild-type and a mutated state associated with the disease PH1 under more than 400 metabolic conditions.³⁴ Our network analysis of the MFGs revealed a consistent organisation of the reaction graph, which is highly preserved under the mutation. Our analysis also identified notable changes in the network centrality score and community structure of certain reactions, which is linked to key biological processes in PH1. Importantly, network measures computed from the MFGs reveal complementary information to that provided by the sole analysis of perturbed FBA fluxes.

Our flow graphs provide a systematic connection between network theory and constraint-based methods widely employed in metabolic modelling,^{21,22,25,32} thus opening avenues towards environment-dependent, graph-based analyses of cell metabolism. An area of interest for future research is the use of MFGs to study how network measures of flow graphs can help characterise metabolic conditions that maximise the efficacy of drug treatments or disease-related distortions, e.g., cancer-related metabolic signatures.^{55–58} In particular, MFGs can quantify metabolic robustness via graph statistics upon removal of reaction nodes.²²

The proposed framework for graph construction for metabolic networks can be extended in different directions. The core idea is

the distinction between production and consumption fluxes (5), and how to encode the relationship between produced and consumed mass flows in the weighted links of a graph. This general principle can also be used to build other potentially useful graphs. For example, two other graphs that describe relationships between reactions are:

$$\text{Competition graph : } \mathcal{D}_c = \frac{1}{n} \mathbf{S}_{2m}^{-T} (\mathbf{W}_-^i)^2 \mathbf{S}_{2m}^- \quad (13)$$

$$\text{Synergy graph : } \mathcal{D}_s = \frac{1}{n} \mathbf{S}_{2m}^{+T} (\mathbf{W}_+^i)^2 \mathbf{S}_{2m}^+ \quad (14)$$

The competition and synergy graphs are undirected and their edge weights represent the probability that two reactions consume (\mathcal{D}_c) or produce (\mathcal{D}_s) metabolites chosen uniformly at random. Similarly to the MFG construction, we can also create the corresponding FBA versions of *competition and synergy mass flow graphs*, which follow directly from (12–14). These additional graphs could help reveal further relationships between metabolic reactions in the cell, and will be the subject of future studies.

The framework can also be extended to include dynamic adaptations of metabolic activity in different ways: by using dynamic extensions of FBA,^{59,60} by incorporating static⁶¹ or time-varying⁶² enzyme concentrations; or by considering kinetic models (with kinetic constants when available) to generate probabilistic reaction fluxes in the sense of stochastic chemical reaction networks.^{37,63} Of particular interest to metabolic modelling, we envision that MFGs could provide a route to evaluate the robustness of FBA solutions^{25,64} by exploiting the non-uniqueness of the MFG from each FBA solution in the space of graphs. Such results could enhance the interface between network science and metabolic analysis, allowing for the systematic exploration of the system-level organisation of metabolism in response to environmental constraints and disease states.

METHODS

Flux balance analysis

Flux Balance Analysis (FBA)^{25,32} is a widely-adopted approach to analyse metabolism and cellular growth. FBA calculates the reaction fluxes that optimise growth in specific biological contexts. The main hypothesis behind FBA is that cells adapt their metabolism to maximise growth in different biological conditions. The conditions are encoded as constraints on the fluxes of certain reactions; for example, constraints reactions that import nutrients and other necessary compounds from the exterior.

The mathematical formulation of the FBA is described in the following constrained optimisation problem:

$$\begin{aligned} &\text{maximise : } \mathbf{c}^T \mathbf{v} \\ &\text{subject to } \begin{cases} \mathbf{S} \mathbf{v} = 0 \\ \mathbf{v}_{lb} \leq \mathbf{v} \leq \mathbf{v}_{ub}, \end{cases} \end{aligned} \quad (15)$$

where \mathbf{S} is the stoichiometry matrix of the model, \mathbf{v} the vector of fluxes, \mathbf{c} is an indicator vector (i.e., $c(i) = 1$ when i is the biomass reaction and zero everywhere else) so that $\mathbf{c}^T \mathbf{v}$ is the flux of the biomass reaction. The constraint $\mathbf{S} \mathbf{v} = 0$ enforces mass-conservation at stationarity, and \mathbf{v}_{lb} and \mathbf{v}_{ub} are the lower and upper bounds of each reaction's flux. Through these vectors, one can encode a variety of different scenarios.³³ The biomass reaction represents the most widely-used flux that is optimised, although there are others that can be used as well.^{31,65}

In our simulations, we set the individual carbon intake rate to 18.5 mmol/gDW/h for every source available in each scenario. We allowed oxygen intake to reach the maximum needed in to consume all the carbon except in the anaerobic condition scenario, in which the upper bound for oxygen intake was 0 mmol/gDW/h. In the scenario with limited phosphate and ammonium intake, the levels of NH_4 and phosphate intake were fixed at 4.5 mmol/gDW/h and 3.04 mmol/gDW/h respectively (a reduction of 50% compared to a glucose-fed aerobic scenario with no restrictions).

Markov Stability community detection framework

We extract the communities in each network using the Markov Stability community detection framework.^{47,48} This framework uses diffusion processes on the network to find groups of nodes (i.e., communities) that retain flows for longer than one would expect on a comparable random network; in addition, Markov Stability incorporates directed flows seamlessly into the analysis.^{46,49}

The diffusion process we use is a continuous-time Markov process on the network. From the adjacency matrix \mathbf{G} of the graph (in our case, the RAG, NFG or MFG), we construct a rate matrix for the process: $\mathbf{M} = \mathbf{K}_{out}^{-1} \mathbf{G}$, where \mathbf{K}_{out} is the diagonal matrix of out-strengths, $k_{out,i} = \sum_j g_{i,j}$. When a node has no outgoing edges then we simply let $k_{out,i} = 1$. In general, a directed network will not be strongly-connected and thus a Markov process on \mathbf{M} will not have a unique steady state. To ensure the uniqueness of the steady state we must add a teleportation component to the dynamics by which a random walker visiting a node can follow an outgoing edge with probability λ or jump (teleport) uniformly to any other node in the network with probability $1 - \lambda$.⁴² The rate matrix of a Markov process with teleportation is:

$$\mathbf{B} = \lambda \mathbf{M} + \frac{1}{N} [(1 - \lambda) \mathbf{I}_N + \lambda \text{diag}(\mathbf{a})] \mathbf{1} \mathbf{1}^T, \quad (16)$$

where the $N \times 1$ vector \mathbf{a} is an indicator for dangling nodes: if node i has no outgoing edges then $a_i = 1$, and $a_i = 0$ otherwise. Here we use $\lambda = 0.85$. The Markov process is described by the ODE:

$$\dot{\mathbf{x}} = -\mathbf{L}^T \mathbf{x}, \quad (17)$$

where $\mathbf{L} = \mathbf{I}_N - \mathbf{B}$. The solution of (17) is $\mathbf{x}(t) = e^{-t\mathbf{L}^T} \mathbf{x}(0)$ and its stationary state (i.e., $\dot{\mathbf{x}} = 0$) is $\mathbf{x} = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is the leading left eigenvector of \mathbf{B} .

A hard partition of the graph into C communities can be encoded into the $N \times C$ matrix \mathbf{H} , where $h_{ic} = 1$ if node i belongs to community c and zero otherwise. The $C \times C$ clustered autocovariance matrix of (17) is

$$\mathbf{R}(t, \mathbf{H}) = \mathbf{H}^T \left(\mathbf{I} e^{-t\mathbf{L}^T} - \boldsymbol{\pi} \boldsymbol{\pi}^T \right) \mathbf{H}, \quad (18)$$

and the entry (c,s) of $\mathbf{R}(t, \mathbf{H})$ measures how likely it is that a random walker that started the process in community c finds itself in community s after time t when at stationarity. The diagonal elements of $\mathbf{R}(t, \mathbf{H})$ thus record how good the communities in \mathbf{H} are at retaining flows. The Markov stability of the partition is then defined as

$$r(t, \mathbf{H}) = \text{trace } \mathbf{R}(t, \mathbf{H}). \quad (19)$$

The optimised communities are obtained by maximising the cost function (19) over the space of all partitions for every time t to obtain an optimised partition $\hat{\mathcal{P}}(t)$. This optimisation is NP-hard; hence, with no guarantees of optimality. Here we use the Louvain greedy optimisation heuristic,⁶⁶ which is known to give high quality solutions $\hat{\mathcal{P}}(t)$ in an efficient manner. The value of the Markov time t , i.e. the duration of the Markov process, can be understood as a resolution parameter for the partition into communities.^{45,47} In the limit $t \rightarrow 0$, Markov stability will assign each node to its own community; as t grows, we obtain larger communities because the random walkers have more time to explore the network.⁴⁸ We scan through a range of values of t to explore the multiscale community structure of the network. The code for Markov Stability can be found at http://www.imperial.ac.uk/mpbara/Partition_Stability/.

To identify the important partitions across time, we use two criteria of robustness.⁴⁵ Firstly, we optimise (19) 100 times for each value of t and we assess the consistency of the solutions found. A relevant partition should be a robust outcome of the optimisation, i.e., the ensemble of optimised solutions should be similar as measured with the normalised variation of information:⁶⁷

$$VI(\mathcal{P}, \mathcal{P}') = \frac{2\Omega(\mathcal{P}, \mathcal{P}') - \Omega(\mathcal{P}) - \Omega(\mathcal{P}')}{\log(n)}, \quad (20)$$

where $\Omega(\mathcal{P}) = -\sum_c p(c) \log p(c)$ is a Shannon entropy and $p(c)$ is the relative frequency of finding a node in community \mathcal{P} in partition \mathcal{P} . We then compute the average variation of information of the ensemble of solutions from the $\ell = 100$ Louvain optimisations $\mathcal{P}_i(t)$ at each Markov time t :

$$V(t) = \frac{1}{\ell(\ell - 1)} \sum_{i \neq j} VI(\mathcal{P}_i(t), \mathcal{P}_j(t)). \quad (21)$$

If all Louvain runs return similar partitions, then $V(t)$ is small, indicating robustness of the partition to the optimisation. Hence we select partitions with low values (or dips) of $V(t)$. Secondly, relevant partitions should also

be optimal across Markov time, as indicated by a low values of the cross-time variation of information:

$$VI(t, t') = VI(\hat{P}(t), \hat{P}(t')). \quad (22)$$

Therefore, we also search for partitions with extended low value plateaux of $VI(t, t')$.^{45,46,52}

Data statement

No new data were generated during the course of this research. The results of the analysis are available in the Supplementary Spreadsheet.

Code availability statement

All computations performed on MATLAB. Code for Markov Stability available at http://www.imperial.ac.uk/mpbara/Partition_Stability/

ACKNOWLEDGEMENTS

M.B.D. acknowledges support from the James S. McDonnell Foundation Postdoctoral Program in Complexity Science/Complex Systems Fellowship Award (#220020349-CS/PD Fellow), and the Oxford-Emirates Data Science Lab. G.B. acknowledges the support from the Spanish Ministry of Economy FPI Program (BES-2012-053772). D.O. acknowledges support from an Imperial College Research Fellowship and from the Human Frontier Science Program through a Young Investigator Grant (RGY0076-2015). J.P. acknowledges the support from MINECO/AEI/FEDER, UE through the SynBioFactory grant (DPI2014-55276-CS-1). M.B. acknowledges funding from the EPSRC through grants EP/I017267/1 and EP/N014529/1.

AUTHOR CONTRIBUTIONS

All authors contributed to the design of the analysis. M.B.D. and G.B.C. performed the analyses. All authors contributed to the manuscript and approved its final version.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Systems Biology and Applications* website (<https://doi.org/10.1038/s41540-018-0067-y>).

Competing interests: The authors declare no competing interests.

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

REFERENCES

1. Berg, J., Tymoczko, J. & Stryer, L. *Biochemistry*, 5th edn, New York City, NY, USA: W. H. Freeman (2002).
2. Thomas, A., Cannings, R., Monk, N. & Cannings, C. On the structure of protein-protein interaction networks. *Biochem. Soc. Trans.* **31**, 1491–1496 (2003).
3. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–61, <https://doi.org/10.1038/nrg2102> (2007).
4. Amor, B., Yaliraki, S. N., Woscholski, R. & Barahona, M. Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection. *Mol. Biosyst.* **10**, 2247–58 (2014).
5. Amor, B. R., Schaub, M. T., Yaliraki, S. N. & Barahona, M. Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nat. Commun.* **7**, 12477 (2016).
6. Newman, M. *Networks: An Introduction*. (Oxford University Press, Inc., New York, NY, USA, 2010).
7. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–4, <https://doi.org/10.1038/35036627> (2000).
8. Wagner, A., & Fell, D. A. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* **268**, 1803–1810 (2001).
9. Gleiss, P. M., Stadler, P. F., Wagner, A. & Fell, D. A. Relevant cycles in chemical reaction networks. *Adv. Complex Syst.* **04**, 207–226 (2001).
10. Arita, M. The metabolic world of *Escherichia coli* is not small. *Proc. Natl. Acad. Sci. USA* **101**, 1543–7 (2004).
11. Guimerà, R. & Nunes Amaral, L. A. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
12. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–5 (2002).

13. Takemoto, K. Does habitat variability really promote metabolic network modularity? *PLoS One* **8**, e61348 (2013).
14. Zhou, W. & Nakhleh, L. Convergent evolution of modularity in metabolic networks through different community structures. *BMC Evol. Biol.* **12**, 181 (2012).
15. Cooper, K. & Barahona, M. Role-based similarity in directed networks. *arXiv:1012.2726*, <http://arxiv.org/abs/1012.2726> (2010).
16. Palsson, B. O. *Systems Biology: Properties of Reconstructed Networks*. (Cambridge University Press, New York, NY, USA, 2006).
17. Ouzounis, C. A. & Karp, P. Global Properties of the Metabolic Map of *Escherichia coli*. *Genome Res.* **10**, 568–576 (2000).
18. Ma, H.-W. & Zeng, A.-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinforma. (Oxf., Engl.)* **19**, 1423–30 (2003).
19. Ma, H.-W., Zhao, X.-M., Yuan, Y.-J. & Zeng, A.-P. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinforma. (Oxf., Engl.)* **20**, 1870–6 (2004).
20. Vitkup, D., Kharchenko, P. & Wagner, A. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* **7**, R39 (2006).
21. Samal, A. et al. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinforma.* **7**, 118 (2006).
22. Smart, A. G., Amaral, L. A. N. & Ottino, J. M. Cascading failure and robustness in metabolic networks. *Proc. Natl. Acad. Sci. USA* **105**, 13223–8 (2008).
23. Winterbach, W., Mieghem, P. V., Reinders, M., Wang, H. & de Ridder, D. Topology of molecular interaction networks. *BMC Syst. Biol.* **7**, 90 (2013).
24. Sauer, U. et al. Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. *J. Bacteriol.* **181**, 6679–88 (1999).
25. Orth, J. D., Thiele, I. & Palsson, B. What is flux balance analysis. *Nat. Biotechnol.* **28**, 245–248 (2010).
26. Ma, H. & Zeng, A.-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277 (2003).
27. Croes, D., Couche, F., Wodak, S. J. & van Helden, J. Inferring meaningful pathways in weighted metabolic networks. *J. Mol. Biol.* **356**, 222–236 (2006).
28. da Silva, M. R., Sun, J., Ma, H., He, F. & Zeng, A.-P. In: Björn H. Junker, Falk Schreiber (ed) *Metabolic Networks Analysis of Biological Networks*. 233–253. John Wiley & Sons, Inc.: Hoboken, NJ, USA (2007).
29. Kreimer, A., Borenstein, E., Gophna, U. & Ruppin, E. The evolution of modularity in bacterial metabolic networks. *Proc. Natl. Acad. Sci. USA* **105**, 6976–6981 (2008).
30. Samal, A. & Martin, O. C. Randomizing genome-scale metabolic networks. *PLoS ONE* **6**, e22295 (2011).
31. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**, 119 (2007).
32. Rabinowitz, J. D. & Vastag, L. Teaching the design principles of metabolism. *Nat. Chem. Biol.* **8**, 497–501 (2012).
33. Orth, J., Fleming, R. & Palsson, B. Reconstruction and use of microbial metabolic networks: the core *Escherichia coli* metabolic model as an educational guide. *EcoSal Plus*, <https://doi.org/10.1128/ecosalplus.10.2.1> (2010).
34. Pagliarini, R. et al. In Silico modeling of liver metabolism in a human disease reveals a key enzyme for histidine and histamine homeostasis. *Cell Rep.* **15**, 2292–2300 (2016).
35. Heinrich, R., & Schuster, S. *The Regulation of Cellular Systems*. (Springer: US, 2012).
36. Chellaboina, V., Bhat, S. P., Haddad, W. M. & Bernstein, D. S. Modeling and analysis of mass-action kinetics. *IEEE Control Syst.* **29**, 60–78 (2009).
37. Gillespie, D. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
38. Srinivasan, S., Cluett, W. R. & Mahadevan, R. Constructing kinetic models of metabolism at genome-scales: A review. *Biotechnology Journal* **10**, (1345–1359 (2015)).
39. Folch-Fortuny, A. et al. MCR-ALS on metabolic networks: Obtaining more meaningful pathways. *Chemom. Intell. Lab. Syst.* **142**, 293–303 (2015).
40. Schuster, S., Fell, D. A. & Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18**, 326–332 (2000).
41. Schilling, C. H., Letscher, D. & Palsson, B. O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248 (2000).
42. Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web* Report No. 1999–66 Palo Alto, CA, USA: Stanford InfoLab, (1999). <http://ilpubs.stanford.edu:8090/422/>.
43. Gleich, D. F. Pagerank beyond the web. *SIAM Rev.* **57**, 321–363, <https://doi.org/10.1137/140976649> (2015).
44. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).

45. Schaub, M. T., Delvenne, J.-C., Yaliraki, S. N. & Barahona, M. Markov dynamics as a zooming lens for multiscale community detection: non clique-like communities and the field-of-view limit. *PLoS ONE* **7**, e32210 (2012).
46. Lambiotte, R., Delvenne, J. & Barahona, M. Random walks, markov processes and the multiscale modular organization of complex networks. *Netw. Sci. Eng., IEEE Trans. on* **1**, 76–90 (2014).
47. Delvenne, J.-C., Yaliraki, S. & Barahona, M. Stability of graph communities across time scales. *Proc. Nat. Acad. Sci. USA* **107**, 12755–12760 (2010).
48. Delvenne, J.-C., Schaub, M. T., Yaliraki, S. N., & Barahona, M. in: *Dynamics On and of Complex Networks* (eds. Mukherjee A, Choudhury M, Peruani F., Ganguly N., & Mitra B.) 221–242 (Springer: New York, 2013).
49. Beguerisse-Díaz, M., Garduño Hernández, G., Vangelov, B., Yaliraki, S. N. & Barahona, M. Interest communities and flow roles in directed networks: the Twitter network of the UK riots. *J. R. Soc. Interface* **11** (2014). <http://rsif.royalsocietypublishing.org/content/11/101/20140940>.
50. Sawers, R. Formate and its role in hydrogen production in escherichia coli. *Biochem. Soc. Trans.* **33**, 42–46 (2005).
51. Vemuri, G. N., Eiteman, M. A., McEwen, J. E., Olsson, L. & Nielsen, J. Increasing nadh oxidation reduces overflow metabolism in saccharomyces cerevisiae. *Proc. Natl Acad. Sci. USA* **104**, 2402–2407 (2007).
52. Bacik, K. A., Schaub, M. T., Beguerisse-Díaz, M., Billeh, Y. N. & Barahona, M. Flow-based network analysis of the caenorhabditis elegans connectome. *PLoS Comput. Biol.* **12**, 1–27 (2016).
53. Gille, C. et al. Hepatonet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol. Syst. Biol.* **6**, <http://msb.embopress.org/content/6/1/411> (2010).
54. King, Z. A. et al. Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS Comput. Biol.* **11**, 1–13 (2015).
55. Cserehely, P., Ágoston, V. & Pongor, S. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.* **26**, 178–182 (2005).
56. Chang, R. L., Xie, L., Xie, L., Bourne, P. E. & Palsson, B. Ø. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput. Biol.* **6**, e1000938 (2010).
57. Folger, O. et al. Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* **7**, <http://msb.embopress.org/content/7/1/501.full.pdf> (2011).
58. Vaitheesvaran, B. et al. The warburg effect: a balance of flux analysis. *Metabonomics* **11**, 787–796, <https://doi.org/10.1007/s11306-014-0760-9> (2015).
59. Waldherr, S., Oyarzún, D. A. & Bockmayr, A. Dynamic optimization of metabolic networks coupled with gene expression. *J. Theor. Biol.* **365**, 469–485 (2015).
60. Rügen, M., Bockmayr, A. & Steuer, R. Elucidating temporal resource allocation and diurnal dynamics in phototrophic metabolism using conditional FBA. *Sci. Rep.* **5**, 15247 (2015).
61. Colijn, C. et al. Interpreting expression data with metabolic flux models: Predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput. Biol.* **5**, e1000489 (2009).
62. Oyarzún, D. A. Optimal control of metabolic networks with saturable enzyme kinetics. *IET Syst. Biol.* **5**, 110–9 (2011).
63. Oyarzún, D. A., Lugagne, J.-B. & Stan, G.-B. Noise propagation in synthetic gene circuits for metabolic control. *ACS Synth. Biol.* **4**, 116–125 (2015).
64. Gudmundsson, S. & Thiele, I. Computationally efficient flux variability analysis. *BMC Bioinforma.* **11**, 489 (2010).
65. Feist, A. M. & Palsson, B. O. The biomass objective function. *Curr. Opin. Microbiol.* **13**, 344–349 (2010).
66. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008 (2008).
67. Meila, M. Comparing clusterings: an information based distance. *J. Multivar. Anal.* **98**, 873–895 (2007).
68. Sankey, H. The thermal efficiency of steam-engines. *Minutes Proc. Inst. Civil. Eng.* **125**, 182–242 (1896).
69. Rosvall, M. & Bergstrom, C. T. Mapping change in large networks. *PLoS ONE* **5**, e8694, <https://doi.org/10.1371/journal.pone.0008694> (2010).



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

© The Author(s) 2018