NAR Genomics and Bioinformatics, 2025, **7**, lqaf003
https://doi.org/10.1093/nargab/lqaf003
Advance access publication date: 31 January 2025
Computational Methods for Transcriptome Analysis

OXFORD

GEMCAT—a new algorithm for gene expression-based prediction of metabolic alterations

Suraj Sharma[1,2,†], Roland Sauter[3,†], Madlen Hotze[4], Aaron Marcellus Paul Prowatke[4], Marc Niere[1], Tobias Kipura[4], Anna-Sophia Egger[4], Kathrin Thedieck[4,5,6,7,8], Marcel Kwiatkowski[4], Mathias Ziegler[1], Ines Heiland[2,3,*]

[1]Department of Biomedicine, University of Bergen, 5020 Bergen, Norway
[2]Neuro-SysMed, Department of Neurology, Haukeland University Hospital, 5021 Bergen, Norway
[3]Department of Arctic and Marine Biology, UiT The Arctic University of Norway, 9019 Tromsø, Norway
[4]Institute of Biochemistry and Center for Molecular Biosciences Innsbruck, University of Innsbruck, A-6020 Innsbruck, Austria
[5]Department Metabolism, Senescence and Autophagy, Research Center One Health Ruhr, University Alliance Ruhr & University Hospital Essen, University Duisburg–Essen, 45147 Essen, Germany
[6]German Cancer Consortium (DKTK), partner site Essen, a partnership between German Cancer Research Center (DKFZ) and University Hospital Essen, 69120 Heidelberg and University Hospital Essen, 45147 Essen, Germany
[7]Laboratory of Pediatrics, Section Systems Medicine of Metabolism and Signaling, University of Groningen, University Medical Center Groningen, 9713 GZ Groningen, The Netherlands
[8]Freiburger Materialforschungszentrum, Stefan-Meier-Straße 21, 79104 Freiburg, Germany

[*]To whom correspondence should be addressed. Email: Ines.Heiland@uit.no
[†]The first two authors should be regarded as Joint First Authors.

Abstract

The interpretation of multi-omics datasets obtained from high-throughput approaches is important to understand disease-related physiological changes and to predict biomarkers in body fluids. We present a new metabolite-centred genome-scale metabolic modelling algorithm, the Gene Expression-based Metabolite Centrality Analysis Tool (GEMCAT). GEMCAT enables integration of transcriptomics or proteomics data to predict changes in metabolite concentrations, which can be verified by targeted metabolomics. In addition, GEMCAT allows to trace measured and predicted metabolic changes back to the underlying alterations in gene expression or proteomics and thus enables functional interpretation and integration of multi-omics data. We demonstrate the predictive capacity of GEMCAT on three datasets and genome-scale metabolic networks from two different organisms: (i) we integrated transcriptomics and metabolomics data from an engineered human cell line with a functional deletion of the mitochondrial NAD transporter; (ii) we used a large multi-tissue multi-omics dataset from rats for transcriptome- and proteome-based prediction and verification of training-induced metabolic changes and achieved an average prediction accuracy of 70%; and (iii) we used proteomics measurements from patients with inflammatory bowel disease and verified the predicted changes using metabolomics data from the same patients. For this dataset, the prediction accuracy achieved by GEMCAT was 79%.

Introduction

The amount of transcriptome data available has increased exponentially over the past decade. Statistical analyses and functional classification approaches such as Gene Ontology term analyses [1, 2] are useful to generate hypotheses, but their functional interpretation is often challenging. Thus, new approaches are required that allow a better integration of expression data to predict physiologically relevant and measurable changes such as metabolic alterations. As linked transcriptomics, proteomics, and metabolomics data have become increasingly available, another challenge is to effectively combine these multi-omics datasets to gain new insights into biological pathways and their regulation.

The most promising strategies for multi-omics data integration to date are based on the molecular mechanism connecting the different layers. These approaches are rooted in the understanding of gene functions and network topology [3–5]. Among these approaches, genome-scale metabolic modelling is one of the most common approaches, with flux balance analysis (FBA) being the pre-eminent formalism. FBA is a constraint-based optimization approach and has been used extensively for biotechnological applications [5–9]. There are different approaches to integrate gene expression data to predict metabolic flux alterations. However, FBA always requires the definition of an optimization target, such as growth or ATP production, which are not always clear and easy to define, especially in mammalian systems. Although attempts have been made to interpret results from FBA to infer metabolite concentration changes [10], the primary output of conventional FBA approaches is predictions of metabolic fluxes [9, 11, 12]. Experimental verification of metabolic fluxes in complex biological systems such as mammalian cells and whole organisms is difficult and cost-intensive, whereas direct prediction

Received: March 5, 2024. **Revised:** December 18, 2024. **Editorial Decision:** January 6, 2025. **Accepted:** January 11, 2025

of metabolite concentration changes can be verified using targeted metabolomics and provide the basis for the prediction of metabolic biomarkers. This has been done successfully using ordinary differential equations resembling the kinetics of individual biochemical reactions [13–16]. However, due to the scarcity of kinetic data and the time required to construct these models, they are limited to a small, targeted set of reactions (up to 100) that are relevant for a specific research question [17–20]. In 2005, Patil and Nielsen presented a network-based approach that enables the prediction of the likelihood of changes in metabolite concentrations at genome scale [21] without the necessity to define optimization targets and pathway constraints. This approach, however, cannot predict the direction of change, which would be important for the interpretation in terms of physiological consequences.

In this study, we present a novel metabolic modelling approach that enables genome-scale integration and analysis of both transcriptomics and proteomics data. In our algorithm, the genome-scale metabolic network is represented as a directed graph, with metabolites as nodes and reactions as edges. Prediction of changes in metabolite concentrations is derived through the comparison of enzyme abundances between two conditions. For instance, the metabolite concentration will remain unchanged if the abundance of enzymes producing or consuming a metabolite is unchanged or all change with the same magnitude. However, if the measured abundance of enzymes that produce a metabolite increases while the abundance of the ones consuming it decreases or remains constant, the concentration of this metabolite is predicted to increase. Importantly, we not only take into account local changes in reactions directly connected to the metabolite, but also consider changes both upstream and downstream in the network. The weights of the edges are scaled based on the changes in the inferred (transcriptomics) or measured (proteomics) enzyme abundances. Through the development of an algorithm that combines gene expression or proteomics data integration with the calculation of the PageRank (PR) centrality [22] of nodes, we can predict changes in the centrality (ranking) of the nodes within a given network.

Among the various centrality measures available, PR stands out as it assesses the quality of connections within a network, making it highly effective for identifying influential nodes. Furthermore, it evaluates the global importance of nodes, providing a comprehensive network-wide perspective. Designed to handle large networks efficiently, PR is scalable. The iterative refinement process ensures stable and accurate rankings. Given these strengths, PR is particularly well suited for our approach.

We consider the changes in the ranking equal to qualitative alterations in metabolite concentrations. An overview of our approach is shown in Fig. 1. We named the approach GEMCAT (Gene Expression-based Metabolite Centrality Analysis Tool) and demonstrate its efficacy using three test cases: (i) the analysis of a transcriptome dataset from an engineered human cell line with a functional deletion of the mitochondrial NAD transporter *SLC25A51* [23]; (ii) the analyses of tissue-specific transcriptome and proteome datasets from a longitudinal study of training-induced metabolic changes in rats [24]; and (iii) the integration of a proteomics dataset from patients with inflammatory bowel disease (IBD) [25–27]. The predictions were compared to the corresponding metabolomics data.

Furthermore, to identify the expression changes that lead to certain metabolic alterations, we developed a method to calculate centrality control coefficients that links expression changes to metabolic changes and enables sensitivity analysis in genome-scale metabolic networks. This method allows us to integrate different types of omics data and to perform a sensitivity analysis that can reveal the causes of metabolic alterations.

Materials and methods

Genome-scale metabolic model

For the analysis of the cell line data and the data from IBD patients, we used Recon3D as a genome-scale human metabolic reconstruction that represents a comprehensive human metabolic network model, accounting for 3288 open reading frames that encode 3695 enzymes and 13 543 reactions on 8399 metabolites localized across seven subcellular locations [28]. For the analysis of the data from rats, we used RatGEM as a genome-scale metabolic model of rat (*Rattus norvegicus*). It represents 2804 enzymes and 12 995 reactions involving 8458 metabolites across nine subcellular compartments [29].

A graph theoretical representation of a genome-scale metabolic network

A graph is composed of nodes linked by edges. A graph is called a directed graph if the direction of the edges linking the nodes is defined. A genome-scale metabolic network can be represented as a directed graph $G(M, R, \Phi)$, where M is a finite set of metabolites and R is a finite set of interactions, which are ordered pairs of distinct interactions ($R \subseteq \{(x, y) \mid (x, y) \in M^2;\ x \neq y\}$) contained in the metabolic network. A metabolic reaction is then represented by several edges, which is less intuitive, but makes the graph easier to process as the nodes represent the same kind of entities [30]. This approach allows the integration of information about the metabolites' transition on edges. Several ways of representing a metabolic network as a graph exist [30, 31]. A stoichiometric matrix $S := (s_{i,j})$, where rows represent metabolites and columns denote reactions, can be used to derive a metabolic graph. $s_{i,j}$ takes negative integers for metabolites that are substrates and positive integers for products of a reaction [19]. The elements are zero otherwise. S can be transformed to yield an adjacency matrix $A{:=}(a_{i,j})$, where $a_{i,j}$ becomes 1 when a reaction exists between metabolites $m_i \in M$ and $m_j \in M$ $(i \neq j)$. The entries are zero otherwise. The edges are characterized by the weighted adjacency matrix $\Phi := (\phi_{i,j})$, where $\phi_{i,j}$ is the sum of weights of the edges between metabolites m_i and m_j. We derived $\Delta\Phi$ from the differential abundance of enzymes by comparing two different strains or conditions.

Integration of the differential gene expression or proteomics data

The integration of differential measured (proteomics) or inferred (transcriptomics) protein abundance starts with the mapping of the corresponding genes or proteins onto the metabolic network. This is done by processing the gene–protein reaction (GPR) relations described by Fang *et al.* [32], according to which the following scenarios are possible: (i) if only one protein is associated with a metabolic reaction, the abundance of this protein is assigned to the reaction; (ii) if several proteins are jointly required for a reaction to take place, the geometric mean of the abundance is assigned to the

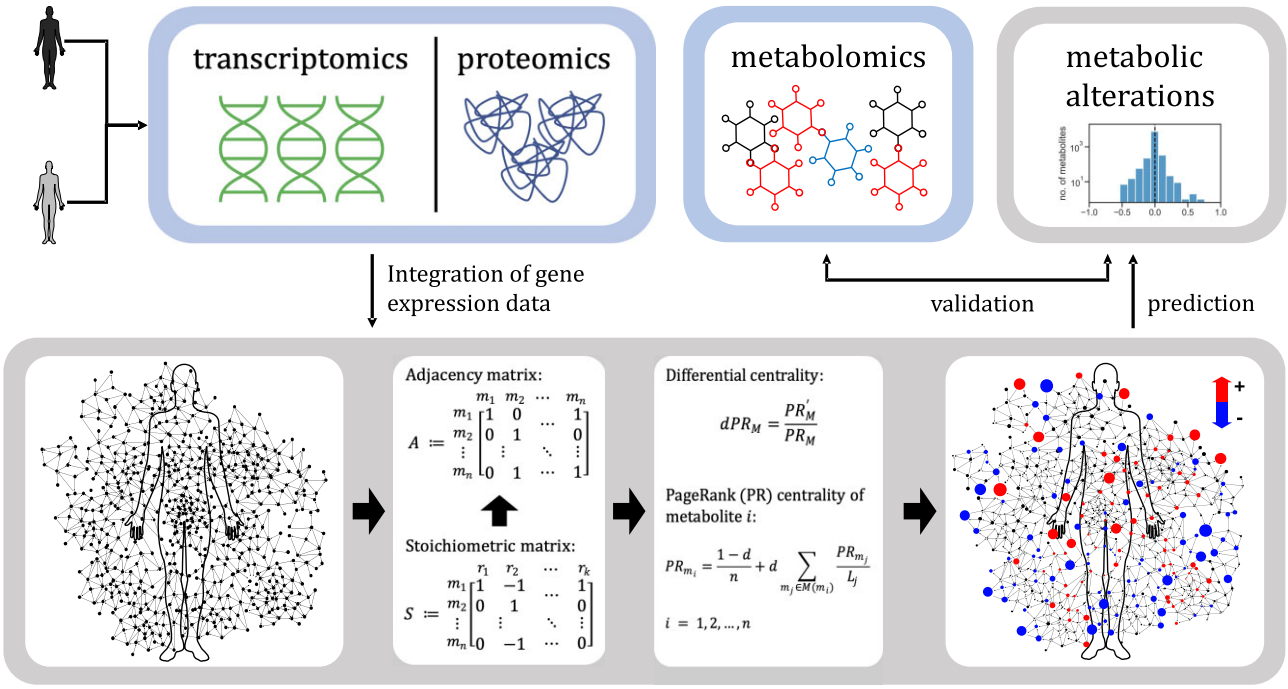

Figure 1. An overview of GEMCAT, our PR-assisted method to integrate transcriptomics or proteomics data to predict metabolic alterations in genome-scale metabolic models (human metabolic model, HMM). A stoichiometric matrix *S* and an adjacency matrix *A* can be derived from the reactions in a metabolic network. Thus, the metabolic network is represented as a directed graph composed of nodes (metabolites) linked by edges (enzymatic reactions). Upon integration of the gene expression data into the graph, PR centrality of every metabolite in the HMM is calculated. The differential PR centrality of metabolites is used to predict changes in their concentrations. The predicted metabolic alterations can be validated using the experimentally measured changes in the metabolomics data.

reaction; (iii) if any one of several proteins is sufficient for a reaction to occur, the arithmetic mean of the abundance is assigned to the reaction; and (iv) any GPR containing combinations of both scenarios (ii) and (iii) is parsed according to their logical relation. The processing of GPRs results in an abundance vector, where each element denotes the net abundance of a protein catalysing a particular reaction.

Calculation of the weighted adjacency from a stoichiometric matrix

Let E be the protein abundance vector calculated by processing the corresponding GPRs and S be the stoichiometric matrix of the given metabolic network with rows and columns representing metabolites and reactions, respectively. All reversible reactions contained in S were split into their unidirectional component reactions, and the corresponding entry in E was duplicated accordingly. A weighted stoichiometric matrix is created as

$$S' = E \cdot S. \tag{1}$$

We split this matrix into its components for products and substrates, respectively, as

$$S^{+}_{i,j} = \begin{cases} S'_{i,j}, & \text{if } x > 0, \\ 0, & \text{otherwise} \end{cases} \tag{2a}$$

and

$$S^{-}_{i,j} = \begin{cases} \left|S'_{i,j}\right|, & \text{if } x < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2b}$$

These matrices are used to calculate the weighted adjacency matrix as

$$\Phi = S^{-} \cdot S^{+T}, \tag{3}$$

which is a square matrix used to represent a finite graph, which in our case is the human metabolic network from the Recon3D model. The elements of the matrix indicate the weight shared by the adjacent pairs of metabolites in the metabolic network.

Calculation of the differential centrality to predict metabolic alterations

The PR centrality of metabolites M can be calculated as

$$\mathrm{PR} = \begin{bmatrix} \mathrm{PR}_{m_1} \\ \mathrm{PR}_{m_2} \\ \vdots \\ \mathrm{PR}_{m_n} \end{bmatrix}. \tag{4}$$

The PR centrality provides a ranking that identifies the importance of a node in a network. In its original form, the algorithm calculates a probability distribution to represent the likelihood of a person randomly clicking on links and arriving at a particular page on the internet [22, 33]. This probability is expressed as a numeric value between 0 and 1. In the context of a metabolic network, the PR value for any metabolite m_i is dependent on the PR values for each metabolite m_j $(i \neq j)$ contained in the set $M(m_i)$ containing all metabolites linking to metabolite m_i divided by the total number of links from metabolite m_j. Further, metabolites with no outbound edges, i.e. metabolites that are sinks, are assumed to link out to all other metabolites in the network. Their PR scores are there-

fore divided evenly among all other metabolites. There is a residual probability d, also known as a damping factor, that is added to each node as a likelihood of a transition to any other node in the network [22]. The PR value of a metabolite m_i can be calculated as

$$\mathrm{PR}_{m_i} = \frac{1-d}{n} + d \sum_{m_j \in M(m_i)} \frac{\mathrm{PR}_{m_j}}{L_{m_j}}, \quad (5)$$

where L_{m_j} is the number of outbound links on m_j and n is the total number of metabolites in the metabolic network. The PR values correspond to the elements of the dominant right eigenvector of the weighted adjacency matrix, which has been normalized to ensure that each column adds up to 1. The eigenvector is given in Equation (4). The PR centralities are iteratively calculated for each metabolite until the convergence is achieved [34]. A differential PR centrality (dPR) can be calculated by comparing the relative change in the PR centralities of M for the perturbed and baseline conditions. We used dPR to predict changes in the abundance of M (Equation 6):

$$\mathrm{dPR} = \frac{\mathrm{PR}_{\mathrm{perturbed}}}{\mathrm{PR}_{\mathrm{baseline}}} \approx \frac{[M_{\mathrm{perturbed}}]}{[M_{\mathrm{baseline}}]}. \quad (6)$$

Calculation of centrality control coefficients

To study the effect of perturbation in the network, we devised a metric that we have referred to as centrality control coefficient. It relates the fractional change in the PR centrality of metabolite M_m to the fractional change in the abundance of enzyme E_e as

$$C^{\mathrm{PR}_m}_{E_e} = \left(\frac{E_e}{\mathrm{PR}_m} \frac{\Delta \mathrm{PR}_m}{\Delta E_e} \right)_{\Delta E_e \to 0}. \quad (7)$$

These coefficients are characteristic for a given network and are not dependent on the expression data. Please see Equations (1)–(3) describing how E relates to the PR centrality of metabolites M.

Gene expression and metabolomics measurements

Cells (293-*SLC25A51*-ko [35] and parental HEK293) were cultivated in Dulbecco's modified Eagle medium, high glucose (Merck/Sigma, D5671) supplemented with 10% (v/v) fetal bovine serum (FBS), 2 mM L-glutamine, 1 mM sodium pyruvate, and penicillin/streptomycin at 37°C in humidified atmosphere with 5% CO_2. Cells were harvested, washed with phosphate-buffered saline, and counted. For RNA sequencing (RNA-seq) analyses, 5×10^6 cells in three technical replicates per cell line were frozen in liquid nitrogen and shipped on dry ice to Novogene Co., Ltd (Cambridge, UK) for processing. RNA sequences were generated using the Illumina NovaSeq platforms. Fragments per kilobase of transcript per million mapped reads were used directly to infer enzyme abundance changes. For the metabolomics measurements, five times 2×10^6 cells in five technical replicates from each parental HEK293 and 293-*SLC25A51*-ko cells were subjected to chloroform–methanol-based simultaneous proteo-metabolomics liquid–liquid extraction [36]. Glycolytic metabolites, metabolites of tricarboxylic acid (TCA) cycle, and nucleotides (ATP, ADP, AMP) were analysed by ion chromatography–single ion monitoring mass spectrometry (IC–SIM-MS) [36] using a quadrupole orbitrap (Exploris 480) and an ICS-6000 IC system (both Thermo Fisher Scientific). Free amino acids and tryptophan metabolites were analysed by multiple reaction monitoring (MRM) using a triple quadrupole mass spectrometer (TQ-XS, Waters) coupled to a UPLC system (ACQUITY Premier, Waters) as described previously [37]. IC–SIM-MS data were processed using TraceFinder 5.0 (version 5.0.889.0, Thermo Scientific) and the LC–MRM-MS data were processed using MS Quan (Waters Connect, Waters) [36, 37].

Gene expression and metabolomics data from previous studies

The tissue-specific transcriptomics, proteomics, and metabolomics from a longitudinal study of training-induced metabolic changes in rats were taken from [24]. The dataset consists of transcriptome and metabolome measurements from 18 tissues from sedentary rats and rats trained for 1, 2, 4, or 8 weeks. The matching proteome was only available for seven tissues. The differential proteomics and metabolomics from mucosa biopsies of IBD patients and healthy controls were obtained from [25–27].

Chemicals

HPLC-grade acetonitrile, methanol, formic acid, Micro BCA Protein Assay Kit, Gibco Qualified FBS, and ammonium bicarbonate were obtained from Thermo Fisher Scientific (Dreieich, Germany). [U-^{13}C]-labelled yeast extract of *Pichia pastoris* (2×10^9 cells) was purchased from ISOtopic Solutions (Vienna, Austria), reconstituted in 2 ml HPLC-H_2O, aliquoted, and stored at −80°C. [U-^{13}C]-labelled lactate [20% (w/w) dissolved in H_2O] and [U-^{13}C–^{15}N]-labelled canonical amino acids (dissolved in 0.1 M HCl) were purchased from Eurisotop (Saarbruecken, Germany).

Calculation of error metrics

We have calculated the following metrics to evaluate the error in GEMCAT's predictions:

- Accuracy: It is defined as the ratio of correct predictions made by GEMCAT relative to the total number of predictions as follows:

$$\mathrm{accuracy} = \frac{\text{no. of correct predictions}}{\text{total no. of predictions}}, \quad (8)$$

where the term 'correct predictions' refers to predictions that have the same direction of change as the measurements.

- Spearman's rank: It measures the strength and direction of the monotonic relationship between two ranked datasets. It was calculated using Equation (9):

$$\rho = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)}, \quad (9)$$

where d_i is the difference between the ranks of corresponding observations and n is the number of observations. The coefficient ρ ranges from −1 to 1, where 1 indicates perfect positive correlation, −1 indicates perfect negative correlation, and 0 suggests no correlation.

- Symmetric mean absolute percentage error (SMAPE) is a normalized error metric that quantifies the accuracy of

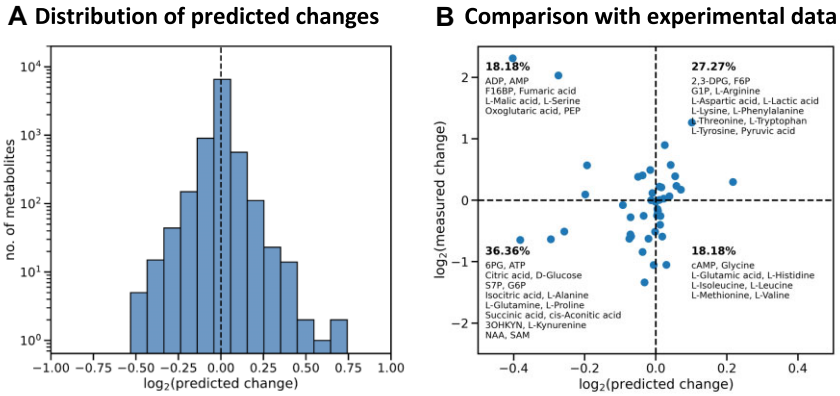

Figure 2. Comparison of predicted and measured changes in the abundance of metabolites in *SLC25A51*-deficient (293-*SLC25A51*-ko) cells relative to parental HEK293 cells. (**A**) A histogram showing the distribution of changes predicted in the abundance of 8399 metabolites in the HMM (Recon3D model) using RNA-seq data from three replicates. (**B**) A scatter plot showing the distribution of the mean of predicted metabolic changes calculated based on the integration of RNA-seq data from three replicates in comparison to the mean of the experimentally measured changes in metabolite concentrations in 293-*SLC25A51*-ko relative to parental HEK293 cells from five replicates (for details see the 'Materials and methods' section). The dashed lines are used to divide the plot into four quadrants. The upper right and lower left quadrants show metabolites, whose concentration changes are predicted correctly. In each quadrant, the percentage and names of metabolites corresponding to it are shown. Metabolite abbreviations: AMP, adenosine monophosphate; ATP, adenosine triphosphate; cAMP, cyclic AMP; 2,3-DPG, 2,3-diphosphoglyceric acid; 6PG, 6-phosphogluconic acid; S7P, D-sedoheptulose 7-phosphate; F16BP, fructose 1,6-bisphosphate; F6P, fructose 6-phosphate; G1P, glucose 1-phosphate; G6P, glucose 6-phosphate; PEP, phosphoenolpyruvic acid; SAM, *S*-adenosylmethionine; NAA, *N*-acetyl-L-aspartic acid; 3OHKYN, hydroxykynurenine.

predictions in a way that accounts for both the magnitude of the errors and the relative scale of the measured and predicted values. It was calculated using Equation (10):

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| - |\hat{y}_i|)/2} \times 100, \qquad (10)$$

where y_i is the measured value, $\hat{y}_i$ is the predicted value, and n is the number of observations. SMAPE provides a percentage error, with values ranging from 0 to 100%.

- Mean squared logarithmic error (MSLE) is used to evaluate the accuracy by penalizing under-predictions more than over-predictions and is useful when the target variable (residual) spans multiple orders of magnitude. It was calculated using Equation (11):

$$\text{MSLE} = \frac{1}{n}\sum_{i=1}^{n}\left(\log(1+y_i) - \log\left(1+\hat{y}_i\right)\right)^2, \qquad (11)$$

where y_i is the measured value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

- Normalized root mean square error (NRMSE) is a variation that normalizes the root mean square error by the range of the measured values, thus allowing a comparison across data with different scales. It was calculated using Equation (12):

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}}, \qquad (12)$$

where y_i is the measured value, $\hat{y}_i$ is the predicted value, n is the number of observations, and $y_{\max}$ and $y_{\min}$ represent the maximum and minimum values of the measurements.

Results

Metabolic alterations in an engineered HEK293 cell line

We used RNA-seq data from an *SLC25A51* CRISPR–Cas9 knockout cell line that lacks a functional mitochondrial NAD transporter [35]. Differential gene expression was calculated by comparing the RNA-seq from three different replicates of *SLC25A51* knockout (293-*SLC25A51*-ko) and parental HEK293 cells, respectively. We used Recon3D [28] as a genome-scale HMM. Two thousand six hundred fifteen transcripts could be mapped to the Recon3D model. As the model contains 3695 enzymes that catalyse 7675 out of 13 543 reactions in total, we assumed that the remaining reactions did not change and thus set the corresponding values to 1. Truncated transcripts of *SLC25A51* can still be detected in the 293-*SLC25A51*-ko cell line, which is common in CRISPR–Cas9 deletions. However, the massive decrease of mitochondrial NAD has been shown experimentally, thereby functionally validating the knockout [35]. We thus set the transport reaction of NAD across the mitochondrial membrane to 0. We used our newly developed algorithm GEMCAT to calculate the differential PR centrality (dPR) of metabolites M ($M = m_1, m_2, \ldots, m_n$; $n = 8399$) in the HMM and used dPR as an estimate for metabolic alteration that is caused by changes in transcript abundance in 293-*SLC25A51*-ko compared to parental HEK293 cells. The predicted $\log_2$ fold changes vary between -0.75 and 0.75 (see Fig. 2A). As Recon3D is a compartmentalized model, but we only have whole-cell metabolomics measurements, we calculated the arithmetic mean of predicted metabolite changes across all subcellular compartments (compartment-specific predictions are shown in Supplementary Fig. S1). We compared the mean values with experimentally measured concentration changes in 44 metabolites. Twenty-eight out of 44, thus ~64% of the metabolites, were predicted correctly (see Fig. 2B). This includes several intermediates of the TCA cycle that have previously been reported to decrease in *SLC25A51*-deficient

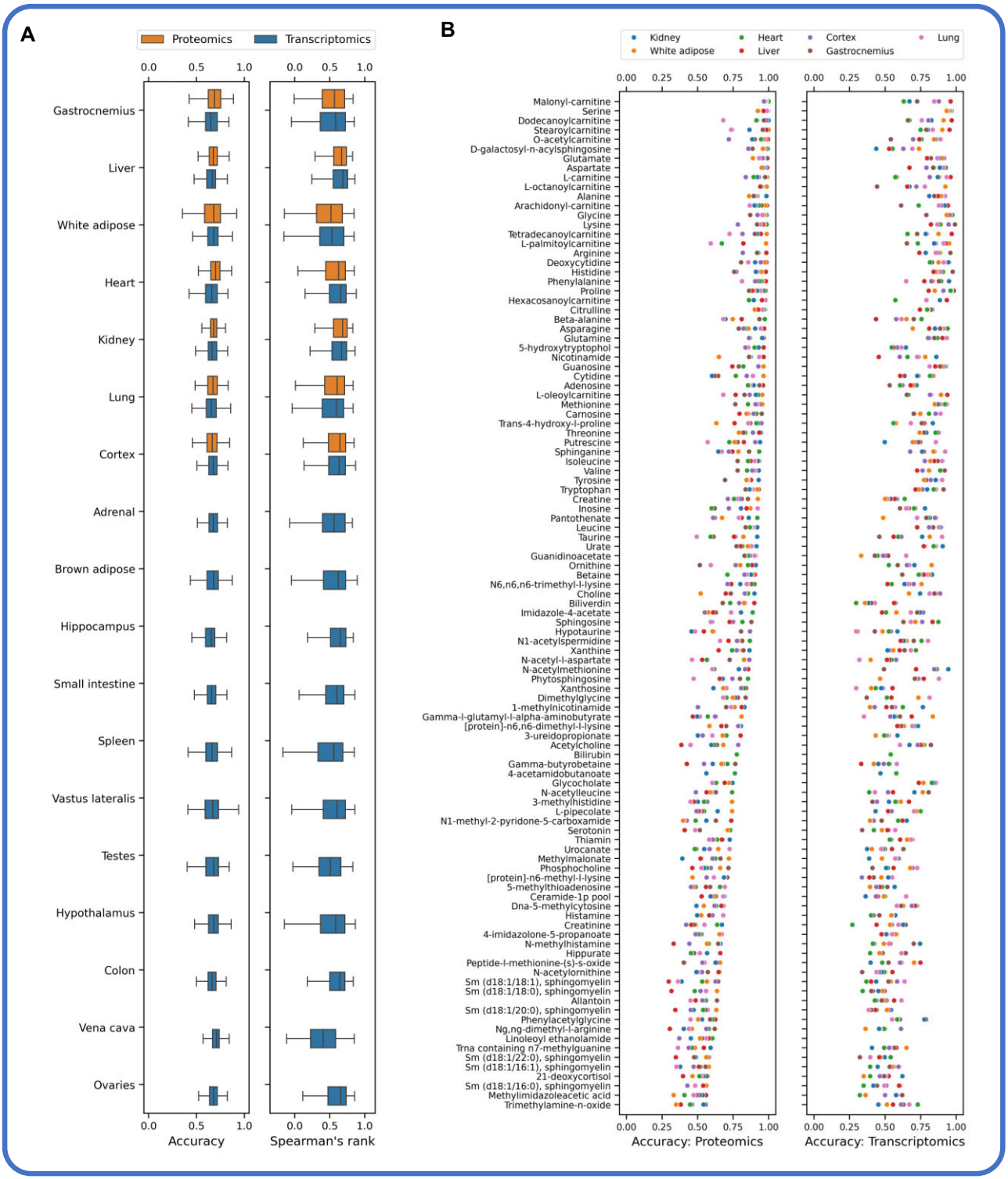

Figure 3. Efficacy of GEMCAT in predicting training-induced metabolic alterations in rats. (**A**) A box plot representation of GEMCAT's prediction accuracy distribution for >100 metabolites, compared to experimentally measured changes across various tissues. Accuracy is the ratio of correctly predicted metabolites to the total number of predictions. Prediction efficiency is further evaluated using Spearman's rank coefficient, which ranges from −1 to +1, with 0 indicating no correlation. A coefficient of −1 or +1 implies an exact monotonic relationship. Each box represents the interquartile range, with the line inside the box indicating the median. The whiskers extend to show the distribution. (**B**) A dot plot distribution of mean accuracy per metabolite for the seven tissues that had metabolomics, proteomics, and transcriptomics data available. Detailed results for all metabolites are provided in the extended data sheets available at https://doi.org/10.6084/m9.figshare.28170524.

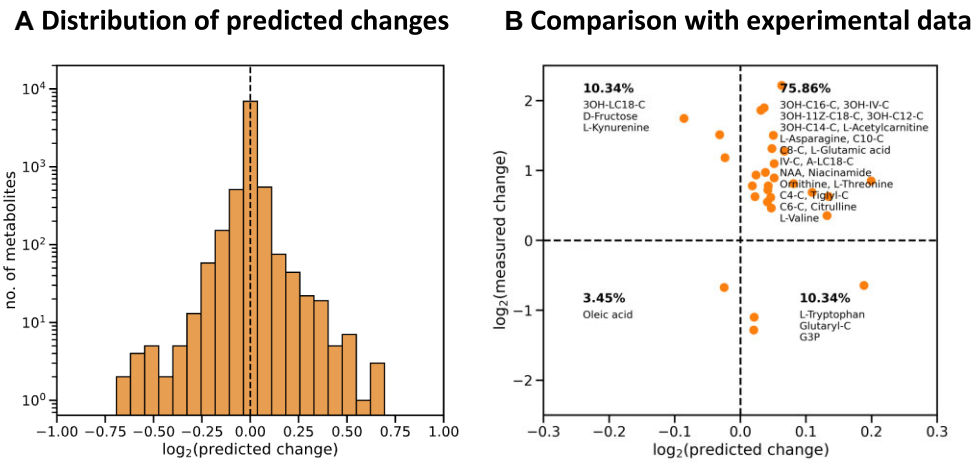

Figure 4. Comparison of predicted and measured changes in the abundance of metabolites in UC patients relative to healthy controls. (**A**) The histogram shows the distribution of predicted changes in all 8399 metabolites contained in the HMM. (**B**) A scatter plot showing the comparison of the predicted changes against the experimentally measured metabolic changes. The dashed lines separate the quadrants between correctly (upper right and lower left quadrants) and incorrectly predicted metabolites. All metabolites measured to be changed significantly ($p < 0.05$ [26]) and having a dPR > ||0.01|| are shown here. The complete results for 137 metabolites are provided in Supplementary Fig. S5 and data at https://doi.org/10.6084/m9.figshare.28170524. Metabolites are shown in orange and corresponding names are indicated in the respective quadrants. The percentage of metabolites in each quadrant is also shown. Metabolite abbreviations: 3OH-C16-C, 3-hydroxyhexadecanoylcarnitine; 3OH-IV-C, 3-hydroxyisovalerylcarnitine; 3OH-11*Z*-C18-C: 3-hydroxy-11*Z*-octadecenoylcarnitine; 3OH-C12-C, 3-hydroxydodecanoylcarnitine; 3OH-C14-C, 3-hydroxytetradecanoylcarnitine; C10-C, decanoylcarnitine; C8-C, L-octanoylcarnitine; 3OH-LC18, (3*S*)-3-hydroxylinoleoyl-CoA; IV-C, isovalerylcarnitine; NAA, *N*-acetyl-L-aspartic acid; A-LC18-C, alpha-linolenylcarnitine; C4-C, butyrylcarnitine; Tiglyl-C, tiglylcarnitine; Glutaryl-C, glutarylcarnitine; C6-C, hexanoylcarnitine; G3P, glycerol 3-phosphate.

HAP1 cells [38]. Further details about variation between replicates are provided in Supplementary Figs S1 and S2 and extended data sheets are available at https://doi.org/10.6084/m9.figshare.28170524.

Predicting metabolic alterations in rats

We used a published transcriptome and proteome dataset comparing the tissue-specific training-induced changes in rats to that of sedentary (control) rats [24]. Both transcriptome and proteome datasets comprise measurements at five time points, i.e. 0, 1, 2, 4, and 8 weeks, from different tissue samples. While the longitudinal measurements of the transcriptome from 18 different tissues were available, proteomics data were only available from 7 tissues. We used a comprehensive genome-scale metabolic model of rat (RatGEM [29]) covering 8458 metabolites to integrate the gene expression data. We could map the abundance of 2617 transcripts and 1759 proteins to the annotated enzymes in the RatGEM. The values for the enzymes not covered by the transcriptome were set to 1 and thus assumed to be unchanged between the two compared samples. Upon expression data integration, we performed the calculation of the dPR scores for all samples from each tissue against all samples from the same tissue independent of the time point and treatment, thus resulting in over 10 000 comparisons. Since the RatGEM is also a compartmentalized model and we have metabolome measurement from the whole cell, we took the arithmetic mean across all intracellular compartments. We compared the mean to the experimentally measured concentrations in >100 metabolites. We summarize our predictions in terms of accuracy, i.e. the ratio between the correctly predicted metabolites and the total predictions (Fig. 3). Additionally, we evaluated the quality of our predictions by calculating the Spearman's rank coefficient. A distribution of the prediction accuracy and Spearman's rank coefficient is shown in Fig. 3A. On average, we predicted 70% of the metabolites correctly (accuracy ≈ 0.7). It is worth noting that the prediction based on proteomics is slightly better (average accuracy $= 0.73$) than prediction based on transcriptomics data (average accuracy $= 0.67$; for accuracy per tissue, see Fig. 3A). An analysis of individual metabolite predictions is given in Fig. 3B and Supplementary Fig. S3. Some metabolites are consistently predicted much more accurately than others. A more detailed analysis evaluating the quality of prediction is shown in Supplementary Fig. S4. For a comprehensive list of accuracies for each metabolite, see extended data sheets provided at https://doi.org/10.6084/m9.figshare.28170524.

Predicting metabolic alterations in IBD patients

To predict metabolic alterations in IBD patients with ulcerative colitis (UC), we used a published mucosa proteome dataset comparing protein abundance in patients with severe UC to that of healthy adults [27]. The dataset comprises only significantly changed protein abundances. These were mapped to 158 enzymes in the Recon3D model. The values for the enzymes not covered by the proteomics were set to 1 and thus assumed to be unchanged between patients and controls. Upon integration of the differential proteomics data, we calculated the dPR scores of all metabolites in our HMM to predict the changes in their concentrations. The distribution of predicted changes in the metabolite concentrations was centred around 0 (see Fig. 4A), with most of the metabolites being unchanged as expected as very few enzymes in the network were significantly changed. To validate our predictions, we compared them with the metabolite measurements from the same set of UC patients and healthy adults [25, 26]. We correctly predict changes in ~79% of the metabolites measured to change significantly (i.e. $p < 0.05$ [26], Fig. 4B). See Supplementary Fig. S5 for the comparison of all measured metabolites and extended data sheets available at

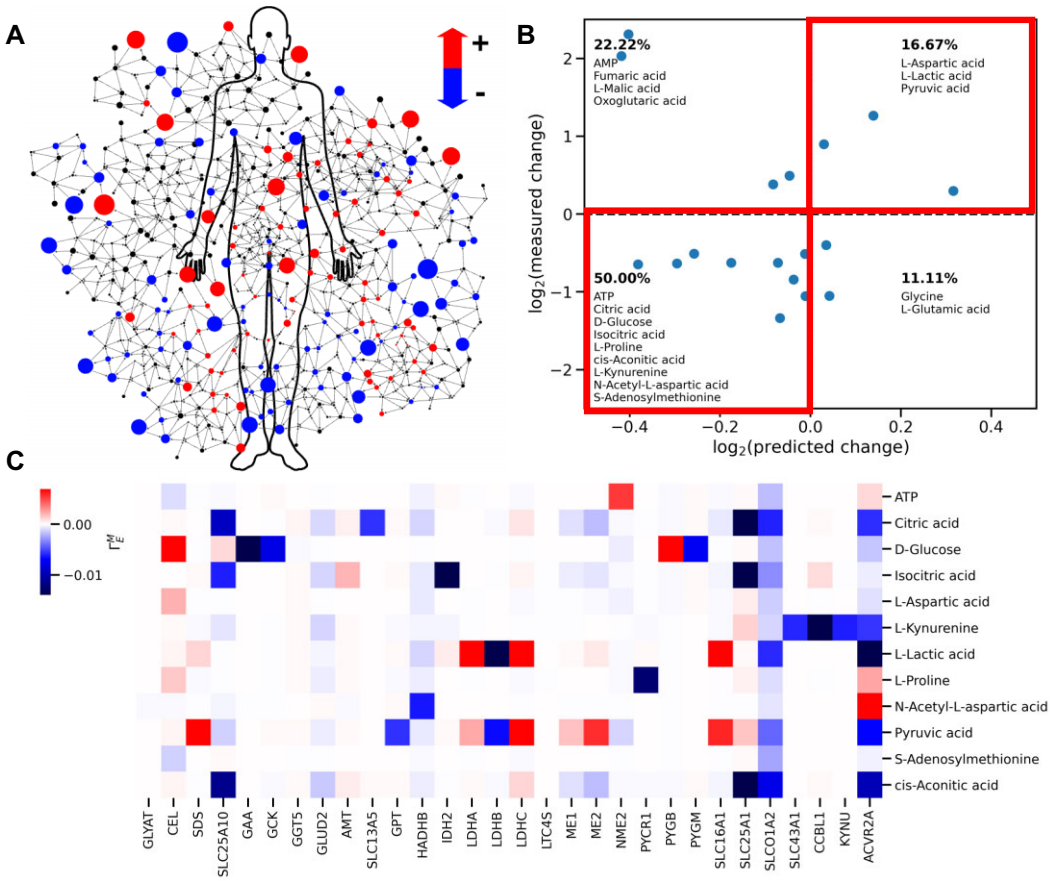

Figure 5. Calculation of response coefficients (Γ_E^M) to trace metabolic alterations back to the underlying changes in gene expression. (**A**) Changes in metabolomics data mapped onto the human metabolic network. (**B**) The scatter plot is based on the comparison shown in Fig. 2B but limited to metabolites that show consistent predicted changes for three replicates of transcriptomics data and significant changes in the measurements ($p < 0.05$, two-tailed Student's *t*-test). Each dot represents the mean value of a metabolite. The bold lines highlight the upper right and bottom left quadrants, where the direction of the predicted changes agrees with the experimentally measured changes. In each quadrant, the percentage and names of metabolites corresponding to it are shown. (**C**) A heatmap showing the response coefficients ($\Gamma_E^M > \|0.005\|$) of correctly predicted metabolites.

https://doi.org/10.6084/m9.figshare.28170524 for values and standard deviation.

Multi-omics integration allows tracing of metabolic changes to changes in enzyme abundance

As the accessibility of metabolomics measurements increases, it becomes important to integrate them with other omics datasets such as transcriptomics and proteomics. We therefore extended GEMCAT to enable tracing of metabolic changes back to the changes in gene expression or proteomics. To this end, we first analysed the impact of abundance changes in each enzyme on the predicted change of all metabolites. This approach is similar to metabolic control analysis and is an important component of mathematical analysis of complex systems [39, 40]. It describes the system behaviour in terms of the properties of its variables. To systematically analyse the effect of perturbation of the metabolic network, we estimated the centrality control coefficient ($C_{E_e}^{\mathrm{PR}_m}$, Equation 7). If this coefficient is positive, the PR centrality of the metabolite m increases as the weight of the reaction increases, which in turn is increased due to the increase in abundance of enzyme E_e. These coefficients are characteristic for a given network and are not dependent on the expression data. Furthermore, these coefficients can be used to estimate the change in centrality of metabolites M for a given differential in enzyme abundance [$\Delta E = (E' - E)/E$, where E' and E denote enzyme abundances for compared and reference conditions, respectively] and refer to this as response coefficient ($\Gamma_E^{\mathrm{PR}_M} = C_E^{\mathrm{PR}_M} \cdot \Delta E$). The response coefficients $\Gamma_E^{\mathrm{PR}_M}$ can be used to trace metabolic alterations back to the underlying changes in the gene expression or proteomics, as shown in Fig. 5. (The full set of centrality control coefficients C_E^M and response coefficients Γ_E^M is provided at https://doi.org/10.6084/m9.figshare.28170524.) Alternatively, the approach can be used to analyse discrepancies between predictions and experiments and thus identify potential points of post-translational regulation or potential incomplete pathway information.

Discussion

In this study, we present GEMCAT, a new algorithm to predict metabolic alterations in a large set of metabolites using genome-scale metabolic networks. Unlike FBA, GEMCAT does not require an objective function, and only uses the intrinsic properties of the metabolic network. Furthermore, it can predict qualitative changes in metabolite concentra-

tions directly. Upon integration of enzyme abundances into a graph-based representation of the metabolic network, we calculated the change in PR centrality of metabolites contained in the graph. This change in PR centrality, referred to as differential PR (dPR) centrality in this paper, was used to predict metabolic alterations originating from changes in enzyme abundances. Using different genome-scale metabolic models (Recon3D and RatGEM as human and rat genome-scale metabolic reconstructions, respectively), GEMCAT can predict changes in >8000 metabolites for different subcellular compartments. As metabolomics measurements are limited to a much lower number of metabolites and are difficult, if not impossible, to perform for all subcellular compartments, we can only verify our predictions on a limited set of metabolites. Since most metabolomics studies comprise whole-cell analyses, we combined the predicted subcellular changes into whole-cell changes, and here used a simplified assumption, i.e. the calculation of arithmetic means across all subcellular compartments. To accurately determine the contribution of each subcellular compartment to changes at the whole-cell level, compartment-specific concentrations of each metabolite would be required. Such data are not available to date. Compared to transcriptomics, proteomics data are considered a better proxy for enzyme activities, and we indeed achieved slightly better prediction accuracy for the integration of rat proteomics data compared to the integration of transcriptomics data from the same rats. However, protein coverage is often still much lower in proteomics compared to transcriptomics datasets. Despite these limitations, GEMCAT achieves a remarkably high prediction accuracy for the proteomics-based predictions for rat and patient datasets even though they only covered 1759 and 158 enzymes of the RatGEM and Recon3D network, respectively. Given that untargeted metabolomics measurements are still rather expensive and incomplete, GEMCAT provides a new approach to derive hypotheses about metabolic alterations that can be validated by far more accurate targeted metabolite measurements. In this way, GEMCAT can also assist in the identification of disease-specific metabolic biomarkers or signatures.

In the analysis of the rat dataset, we see that GEMCAT achieves a higher prediction accuracy for a subset of metabolites. The origin of this bias can be manifold: Besides potential differences in metabolite measurements' accuracy, there could be errors in the network reconstruction or a bias resulting from the integration algorithm towards certain parts of the network. To understand the reasons for incorrectly predicted metabolic changes, one can trace back the predicted changes to the corresponding expression changes and network components using our network-based control analysis approach. Consequently, our approach can be applied to improve metabolic network reconstruction by tracing incorrectly predicted metabolites to identify errors or gaps in the network reconstructions.

We envisage that GEMCAT can be extended in several ways to improve prediction quality. It should, for example, be noted that it is not yet possible to incorporate changes in media composition or diet that do of course impact nutrient availability and thus metabolite changes. We furthermore believe that prediction accuracy can be further improved through integration of kinetic parameters and other relevant information such as post-translational modifications and their functional effects by adjusting the weights of the edges.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

The project has been funded by the following funding agencies: Norwegian Research Council (Project No. 288164, 325172) to M.Z. and I.H.; University of Innsbruck (Project No. 316826) to M.K.; the Tyrolian Research Fund (Project No. 18903) to M.K.; European Union's Horizon 2020 research and innovation programme (MESI-STRAT Grant Agreement No. 754688) to K.T., I.H., and M.Z.; European Partnership for the Assessment of Risks from Chemicals PARC (Grant Agreement No. 101057014) to K.T.; and European Research Council (ERC AdG BEYOND STRESS, Grant Agreement No. 101054429) to K.T. The computations were partially performed on resources provided by UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway (Project No. NN9795K). Funding to pay the Open Access publication charges for this article was provided by UiT The Arctic University of Norway.

Data availability

A computational framework to allow GEMCAT calculations is developed using Python programming. The framework is available as a Python package and can be downloaded from Python Package Index (PyPI). GEMCAT is also hosted on GitHub (https://github.com/MolecularBioinformatics/GEMCAT) and has been provided via figshare (https://doi.org/10.6084/m9.figshare.25020101.v1). The full set of results, the example files as well as the differential expression dataset, and the proteomics and metabolomics data along with the Python scripts used to produce images presented in this paper can be downloaded from figshare (https://doi.org/10.6084/m9.figshare.28170524). The RNA-seq data are available at GEO GSE255209.

References

1. Ashburner M, Ball CA, Blake JA *et al.* Gene Ontology: tool for the unification of biology. *Eur J Biochem* 2000;**25**:89–95.
2. Gene Ontology Consortium, Aleksander SA, Balhoff J *et al.* The Gene Ontology Knowledgebase in 2023. *Genetics* 2023;**224**:iyad031.
3. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods. *Nat Rev Microbiol* 2012;**10**:291–305.
4. Blazier A, Papin J. Integration of expression data in genome-scale metabolic network reconstructions. *Front Physiol* 2012;**3**:299.
5. Nielsen J. Systems biology of metabolism. *Annu Rev Biochem* 2017;**86**:245–75.
6. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol* 2008;**4**:e1000082.

7. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods. *Nat Rev Microbiol* 2012;**10**:291–305.
8. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol* 2010;**28**:245–48.
9. Palsson BO. *Systems Biology: Constraint-Based Reconstruction and Analysis*. Cambridge: Cambridge University Press, 2015.
10. Reznik E, Mehta P, Segrè D. Flux imbalance analysis and the sensitivity of cellular growth to changes in metabolite pools. *PLoS Comput Biol* 2013;**9**:1003195.
11. O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. *Cell* 2015;**161**:971–87.
12. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 2009;**5**:320–20.
13. Dewi DL, Mohapatra SR, Blanco Cabañes S *et al.* Suppression of indoleamine-2,3-dioxygenase 1 expression by promoter hypermethylation in ER-positive breast cancer. *OncoImmunology* 2017;**6**:e1274477.
14. Odendaal C, Jager EA, Martines A-CMF *et al.* Personalised modelling of clinical heterogeneity between medium-chain acyl-CoA dehydrogenase patients. *BMC Biol* 2023;**21**:184.
15. Mohapatra SR, Sadik A, Tykocinski L-O *et al.* Hypoxia inducible factor 1α inhibits the expression of immunosuppressive tryptophan-2,3-dioxygenase in glioblastoma. *Front Immunol* 2019;**10**:2762.
16. Shlomi T, Cabili MN, Ruppin E. Predicting metabolic biomarkers of human inborn errors of metabolism. *Mol Syst Biol* 2009;**5**:263.
17. Stavrum AK, Heiland I, Schuster S *et al.* Model of tryptophan metabolism, readily scalable using tissue-specific gene expression data. *J Biol Chem* 2013;**288**:34555–66.
18. Mazat J-P, Devin A, Ransac S. Modelling mitochondrial ROS production by the respiratory chain. *Cell Mol Life Sci* 2020;**77**:455–65.
19. Heinrich R, Schuster S. *The Regulation of Cellular Systems*. Boston, MA: Springer US, 1996.
20. Voit EO. The best models of metabolism. *Wiley Interdiscip Rev Syst Biol Med* 2017;**9**:e1391.
21. Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Natl Acad Sci USA* 2005;**102**:2685–89.
22. Page L, Brin S. The anatomy of a large-scale hypertextual web search engine. *Comput Netw ISDN Syst* 1998;**30**:107–17.
23. Luongo TS, Eller JM, Lu M-J *et al.* SLC25A51 is a mammalian mitochondrial NAD^+ transporter. *Nature* 2020;**588**:174–79.
24. Amar D, Gay NR, Jimenez-Morales D *et al.* The mitochondrial multi-omic response to exercise training across rat tissues. *Cell Metab* 2024;**36**:1411–29.
25. Diab J, Hansen T, Goll R *et al.* Lipidomics in ulcerative colitis reveal alteration in mucosal lipid composition associated with the disease State. *Inflamm Bowel Dis* 2019;**25**:1780–87.
26. Diab J, Hansen T, Goll R *et al.* Mucosal metabolomic profiling and pathway analysis reveal the metabolic signature of ulcerative colitis. *Metabolites* 2019;**9**:291.
27. Schniers A, Goll R, Pasing Y *et al.* Ulcerative colitis: functional analysis of the in-depth proteome. *Clin Proteomics* 2019;**16**:4.
28. Brunk E, Sahoo S, Zielinski DC *et al.* Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat Biotechnol* 2018;**36**:272–81.
29. Wang H, Robinson JL, Kocabas P *et al.* Genome-scale metabolic network reconstruction of model animals as a platform for translational research. *Proc Natl Acad Sci USA* 2021;**118**:e2102344118.
30. Lacroix V, Cottret L, Thébault P *et al.* An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans Comput Biol Bioinform* 2008;**5**:594–617.
31. Palsson BO. Metabolic systems biology. *FEBS Lett* 2009;**583**:3900–4.
32. Fang X, Wallqvist A, Reifman J. Modeling phenotypic metabolic adaptations of *Mycobacterium tuberculosis* H37Rv under hypoxia. *PLoS Comput Biol* 2012;**8**:e1002688.
33. Page L, Brin S, Motwani R.The PageRank Citation Ranking: Bringing Order to the Web. Technical report. Stanford InfoLab, 1999.
34. Arasu A, Novak J, Tomkins A *et al.* PageRankcomputation and the structure of the web: experiments and algorithms. 2002. *Proceedings of the eleventh international World Wide Web conference*, pp. 107-117.
35. Goyal S, Paspureddi A, Lu MJ *et al.* Dynamics of SLC25A51 reveal preference for oxidized NAD^+ and substrate led transport. *EMBO Rep* 2023;**24**:e56596.
36. van Pijkeren A, Egger A-S, Hotze M *et al.* Proteome coverage after simultaneous proteo-metabolome liquid–liquid extraction. *J Proteome Res* 2023;**22**:951–66.
37. Kipura T, Hotze M, Hofer A *et al.* Automated liquid handling extraction and rapid quantification of underivatized amino acids and tryptophan metabolites from human serum and plasma using dual-column U(H)PLC–MRM-MS and its application to prostate cancer study. *Metabolites* 2024;**14**:370.
38. Girardi E, Agrimi G, Goldmann U *et al.* Epistasis-driven identification of SLC25A51 as a regulator of human mitochondrial NAD import. *Nat Commun* 2020;**11**:6145.
39. Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur J Biochem* 1974;**42**:89–95.
40. Kacser H, Burns JA. The control of flux. *Symp Soc Exp Biol* 1973;**27**:65–104.

Received: March 5, 2024. **Revised:** December 18, 2024. **Editorial Decision:** January 6, 2025. **Accepted:** January 11, 2025