A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data

Norah Alghamdi,[1,6] Wennan Chang,[1,2,6] Pengtao Dang,[1,2] Xiaoyu Lu,[1] Changlin Wan,[1,2] Silpa Gampala,[3] Zhi Huang,[1,2] Jiashi Wang,[1] Qin Ma,[4] Yong Zang,[1,5] Melissa Fishel,[3] Sha Cao,[1,5] and Chi Zhang[1,2]

[1] Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA; [2] Department of Electrical and Computer Engineering, Purdue University, Indianapolis, Indiana 46202, USA; [3] Department of Pediatrics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA; [4] Department of Biomedical Informatics, Ohio State University, Columbus, Ohio 43210, USA; [5] Department of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

The metabolic heterogeneity and metabolic interplay between cells are known as significant contributors to disease treatment resistance. However, with the lack of a mature high-throughput single-cell metabolomics technology, we are yet to establish systematic understanding of the intra-tissue metabolic heterogeneity and cooperative mechanisms. To mitigate this knowledge gap, we developed a novel computational method, namely, single-cell flux estimation analysis (scFEA), to infer the cell-wise fluxome from single-cell RNA-sequencing (scRNA-seq) data. scFEA is empowered by a systematically reconstructed human metabolic map as a factor graph, a novel probabilistic model to leverage the flux balance constraints on scRNA-seq data, and a novel graph neural network–based optimization solver. The intricate information cascade from transcriptome to metabolome was captured using multilayer neural networks to capitulate the nonlinear dependency between enzymatic gene expressions and reaction rates. We experimentally validated scFEA by generating an scRNA-seq data set with matched metabolomics data on cells of perturbed oxygen and genetic conditions. Application of scFEA on this data set showed the consistency between predicted flux and the observed variation of metabolite abundance in the matched metabolomics data. We also applied scFEA on five publicly available scRNA-seq and spatial transcriptomics data sets and identified context- and cell group–specific metabolic variations. The cell-wise fluxome predicted by scFEA empowers a series of downstream analyses including identification of metabolic modules or cell groups that share common metabolic variations, sensitivity evaluation of enzymes with regards to their impact on the whole metabolic flux, and inference of cell–tissue and cell–cell metabolic communications.

[Supplemental material is available for this article.]

Metabolic dysregulation is a hallmark of many disease types including cancer, diabetes, cardiovascular disease, and Alzheimer's disease (Mattson and Chan 2001; Rask et al. 2001; Matsuzawa 2006; Dunn et al. 2014; Hirschey et al. 2015; Kochanek et al. 2019; Sun et al. 2020). In cancer, the diseased cells are well understood to rewire their metabolism and energy production to support rapid proliferation, sustain viability, and promote acquired drug resistance (Thompson et al. 2005; DeBerardinis et al. 2008; Hanahan and Weinberg 2011; Ward and Thompson 2012). Here, the diseased cells often react differently to the microenvironmental stress. Such heterogeneity often results in an increased repertoire of possible cellular responses to compromise the efficacy of drug therapies, leading to the enhanced survival of the entire diseased cell population (Bishop et al. 2007; Lidstrom and Konopka 2010). The metabolome is an excellent indicator of phenotypic heterogeneity owing to its high dynamics and plasticity (Zenobi 2013). Current high-throughput metabolic profiling has been largely applied to bulk cell or tissue samples from which we could only observe an averaged metabolic signal over a large number of cells, whereas single-cell metabolomics is still in its infancy because of its relatively low throughput and low sensitivity (Zenobi 2013; Fessenden 2016; Emara et al. 2017; Zampieri et al. 2017; Ali et al. 2019, 2020; Duncan et al. 2019). Overall, our understanding of metabolic dysregulation of human disease has been immensely limited by our technology to study the metabolic landscape at the single-cell level and in the context of their tissue microenvironment (Jaenisch and Bird 2003; Feinberg 2007; Heintzman et al. 2007; Harris et al. 2010; The ENCODE Project Consortium 2012; Roadmap Epigenomics Consortium et al. 2015; Robertson-Tessi et al. 2015; Kim and DeBerardinis 2019).

Single-cell RNA-seq (scRNA-seq) data has been widely used to characterize cell type–specific transcriptional states and its underlying phenotypic switches in a complex tissue (Vasdekis and Stephanopoulos 2015; Damiani et al. 2019; Evers et al. 2019; Honkoop et al. 2019; Saurty-Seerunghen et al. 2019; Xiao et al. 2019, 2020; Rohlenova et al. 2020; Zhang et al. 2020; Levine et al. 2021). Realizing the strong connections between transcriptomic and metabolomic profiles (Hirayama et al. 2009; Lee et al. 2012; Mehrmohamadi et al. 2014; Damiani et al. 2019; Xiao

[6] **These authors contributed equally to this work.**
Corresponding authors: czhang87@iu.edu, mfishel@iu.edu, shacao@iu.edu

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.271205.120. Freely available online through the *Genome Research* Open Access option.

et al. 2019b; Wagner et al. 2021), scRNA-seq data has found its application in portraying metabolic variations. Using scRNA-seq data, the existing research studied metabolic changes of predefined cell groups relying on differential expression and enrichment analysis of key metabolic enzymes and pathways (Vasdekis and Stephanopoulos 2015; Evers et al. 2019; Honkoop et al. 2019; Saurty-Seerunghen et al. 2019; Xiao et al. 2019, 2020; Rohlenova et al. 2020; Levine et al. 2021). However, for this type of analysis, the node/edge structures in a metabolic pathway graph or the mass balance constraints of metabolic network is not considered. Studies coupling single-cell transcriptomics data and the flux balance analysis (FBA) at steady-state framework have only recently emerged (Damiani et al. 2019; Zhang et al. 2020). The FBA describes the potential flux over the topological structure of a metabolic network, with a set of equations governing the mass balance at steady state. The advantage of incorporating FBA into the model is twofold: considering the chemical stoichiometry in FBA could lead to more accurate estimation of the metabolite abundance; and flux estimation for each individual metabolite can be solved, leading to high-resolution characterization of the metabolic profiling. Damiani et al. (2019) developed scFBA that uses the cell group–specific gene expression status derived from scRNA-seq data to regularize the network topology for FBA. Wagner et al. (2021) proposed a method, namely, Compass, that maximizes the coherence between scRNA-seq expression profile and predicted flux in solution space of FBA. However, as stated in the original works (Wagner et al. 2021), the stringent flux balance and steady-state assumption in scFBA and Compass may not be rational for certain disease types with constantly severe “imbalance” of many metabolites, namely, cancer. Another limitation of the FBA-based methods is that the single cells’ gene expression is not used directly to model metabolic flux. Both scFBA and Compass used single-cell gene expression as certain constraints to guide the search in the solution space of flux balance condition. In addition, both models are intended for modeling the fluxes for cells of predefined groups instead of at a single-cell resolution, and they are restricted to a small portion of the whole metabolic map. Therefore, it remains an urgent task to design advanced computational tools for reliable estimation of cell-wise metabolic flux and states by translating single-cell transcriptomes to single-cell fluxomes. Such a tool is vital to unravel the principles of how the disease microenvironment may affect the metabolic phenotypes for the heterogeneous cell types (Damiani et al. 2019; Evers et al. 2019).

Computational challenges to estimate cell-wise metabolic flux arise from the following aspects: (1) multiple key factors determine cells’ metabolic states, including exogeneous nutrient availability, leading to the discrepancy of cell type–specific markers and metabolic phenotypes and states; (2) the whole metabolic network is of high complexity, hence, a proper computational reduction and reconstruction of the network is needed to reach a balance between resolution of metabolic state characterization and computational feasibility; (3) the intricate nonlinear dependency between transcriptomic expressions and metabolic reaction rates calls for a more sophisticated model to fully capitulate the relationships; and (4) alternative enzymes with similar functions may result in common metabolic phenotypes, however, exactly which enzymes share such common effect to the metabolic flux change remains largely unknown.

In this study, we developed a novel computational method, namely, single-cell flux estimation analysis (scFEA), to estimate the relative rate of metabolic flux at single-cell resolution from scRNA-seq data. Specifically, scFEA can effectively solve the aforementioned challenges with the following computational innovations: (1) an optimization function derived based on a probabilistic model to consider the flux balance constraints among a large number of single cells with varied metabolic fluxomes, (2) a metabolic map reduction approach based on network topology and gene expression status, (3) a multilayer neural network model to capture the nonlinear dependency of metabolic flux on the enzymatic gene expressions, and (4) a novel graph neural network architecture and solution to maximize the overall flux balance of intermediate substrates throughout all cells. The central hypotheses of scFEA are (1) the flux variations of a metabolic module can be modeled as a nonlinear function of the transcriptomic-level changes of the catalyzing enzymes; and (2) the total flux imbalance of all intermediate substrates should be minimized throughout all single cells. The cell-wise fluxome estimated by scFEA enables a series of downstream analyses, including identification of cell- or tissue-level metabolic stress, sensitivity evaluation of individual enzymes to the whole metabolic network, and inference of cell–tissue and cell–cell metabolic exchanges. To validate scFEA, we generated an scRNA-seq data set with matched tissue-level metabolomic profiles under different biochemical perturbations. Applications of scFEA on synthetic data sets, the newly generated data set with matched scRNA-seq and metabolomic profiles, and six other independent real-world data sets, validated the prediction accuracy, robustness, and biological interpretability of scFEA.

Results

Systems biology considerations, hypotheses, and analysis pipeline of scFEA

The reaction rate of a simple enzyme catalyzed metabolic reaction follows the Michaelis–Menten kinetic model: $V = K_{\text{cat}}\ ([E][S]/K_m + [S])$, which is a nonlinear function of enzyme concentration $[E]$, substrate concentration $[S]$, and kinetic parameters K_{cat} and K_m. On one hand, the reaction rate is approximately a linear function of the enzyme concentration when the substrate concentration is much larger than K_m, that is, when $[S]/(K_m + [S])$ is ~1; on the other hand, the enzyme concentrations could often serve as a surrogate for the substrate concentration considering the regulatory effect of substrate availability on the enzyme transcription. Overall, we consider the reaction rate to be a (non)linear function of the enzyme concentration. Obviously, the flux of a reaction chain is mostly determined by the rate-limiting steps, which depend on the flux distribution, substrate concentration, and kinetic parameters. Hence, the rate-limiting steps are often context specific and unknown because of the dynamics of the physiological and biochemical conditions of the cells. Based on these considerations, we developed scFEA to estimate cell-wise metabolic flux from scRNA-seq data. scFEA consists of three major computational components: (1) network reduction and reconstruction, (2) estimation of cell-wise metabolic flux, and (3) downstream analyses including estimation of metabolic stress, perturbation of metabolic genes, and clustering of cells with different metabolic states (Fig. 1). The required input of scFEA is an scRNA-seq data set, whereas optional inputs, including cell group labels or the subset of metabolic reactions of interest, can be specified for additional analysis.

To reduce the complexity of the metabolic map, we reconstructed it into a factor graph composed by connected metabolic modules as variables and intermediate metabolites as factors (Fig. 1A). Specifically, connected reactions are merged into one module,

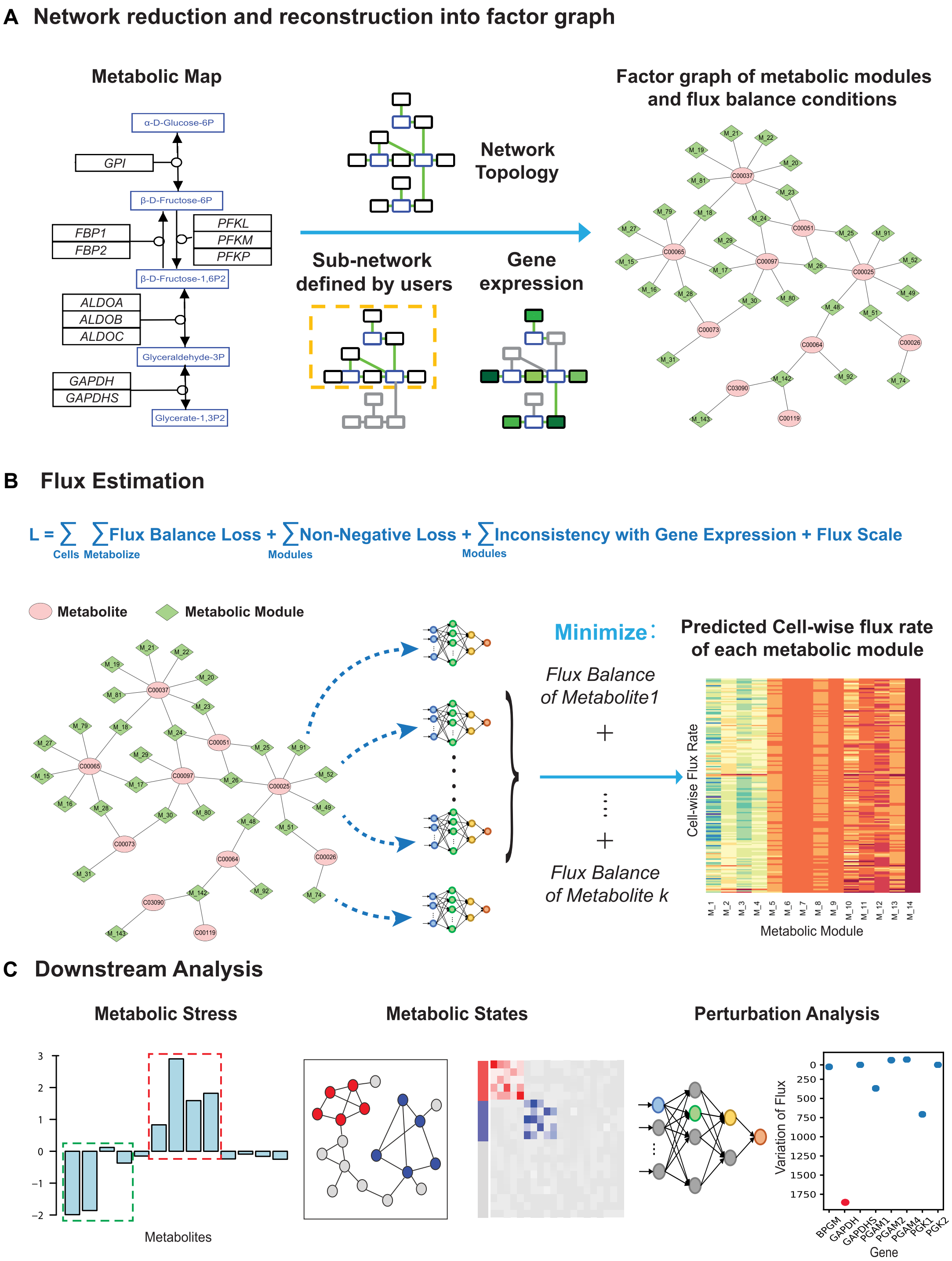

Figure 1. The computational framework of scFEA. (*A*) Metabolic reduction and reconstruction. A metabolic map was reduced and reconstructed into a factor graph based on network topology, significantly non-zero gene expressions, and users' input. (*B*) A novel graph neural network architecture–based prediction of the cell-wise fluxome. A loss function (L) composed by loss terms of flux balance, non-negative flux, coherence between predicted flux and gene expression, and constraint of flux scale, were used to estimate cell-wise metabolic flux from scRNA-seq data. See detailed models and formulations in Results and Methods. (*C*) Downstream analysis of scFEA is provided, including inference of metabolic stress, cell and module clusters of distinct metabolic states, and the genes of the top impact to the whole metabolic flux.

if changes in the reaction rates within the module do not affect the rates of the rest of the reactions, given a fixed flux rate of the module. In other words, the estimated flux of a module stays the same, with or without merging the reactions, under the flux balance condition. This approach increases the robustness of flux estimation and reduces the computational complexity.

The central computational component of scFEA is a novel graph neural network architecture, which models the cell-wise metabolic flux of each metabolic module using gene expression levels of the catalyzing enzymes (Fig. 1B). We hypothesize that the metabolic flux throughout all the single cells in a tissue sample should minimize the overall imbalance of the influx/outflux of intermediate substrates. The rationality of this assumption is that cells within the same tissue exchange metabolites with each other, hence the total flux balance constraint on all the single cells from one tissue sample is more robust than in individual cells. In scFEA, we use the gene expression variations to reflect the protein level change of enzymes and transporters. Note that this assumption is supported by many existing studies that reveal the high explainability of the transcriptome for the proteome (Schnell 2014; Roadmap Epigenomics Consortium et al. 2015; Liu et al. 2016). We assume the flux variations of a module generally impact its neighboring modules, which can be reflected by aggregating the expression variations of the genes in its neighborhood over the metabolic network. The nonlinear dependency between gene expression and metabolic flux is modeled as a fully connected neural network of 2–4 layers, which could be considered as a nonlinear approximation of the Michaelis–Menten model. To solve the neural network parameters, scFEA minimizes a loss function that mimics the overall flux imbalance of all modules in all cells, with further non-negativity and other prior assumptions on the module fluxome. The large number of single cells in scRNA-seq data grants sufficient statistical power to detect the flux variations and avoids the overfitting of the neural network training (for details, see Methods). It is noteworthy that the parameters of the neural network could serve as sensitivity measures of the metabolic flux balance to the variations of the genes. In other words, genes with higher impact are likely to be associated with rate-limiting reactions under the particular context.

The estimated cell-wise metabolic flux enables the prediction of (1) the metabolites or pathways with high imbalance in certain cell groups, (2) groups of metabolic modules or cells with varied metabolic states, and (3) the metabolic genes whose perturbation highly impacts the overall metabolic flux (Fig. 1C). In this study, we mainly focus on solving cell-wise metabolic flux and states and method validations in human cells. A capability for mouse data analysis is also provided in the software package of "scFEA."

Metabolic map reduction and reconstruction

The whole metabolic network in human and mouse have been well studied. However, although databases including the Kyoto Encyclopedia of Genes and Genomes (KEGG) provide well categorized metabolic pathways and the comprehensive set of metabolic genes (Kanehisa and Goto 2000), the network topological structure needs to be further optimized for fluxome estimation because of the following reasons: (1) the flux balance constraints depend on the optimization goal or computational assumption, such as the balance of carbon, redox, or pH; (2) the network complexity needs to be reduced to enable computational feasibility; and (3) a manual correction and annotation of the directions of reactions and transporters is needed. In addition, cells of different types or physiological states naturally have varied metabolic states. In scFEA, we first manually curated and annotated the metabolic map of human and mouse retrieved from the KEGG database. The global metabolic map is further reduced and reconstructed into a factor graph based on its topological property. scFEA also allows the selection of a connected subnetwork in the global metabolic network for flux estimation.

Collection of human and mouse metabolic map

The metabolic map consists of pathways and reactions that fall under four major types, namely, import, metabolism, biosynthesis, and export. To ensure a comprehensive coverage of the global metabolic map, we collected reactions of metabolism and biosynthesis as well as transporters for import and export from different data sources. Specifically, metabolic reactions were directly retrieved from the KEGG database (Kanehisa and Goto 2000); the transporters and annotations of import and export reactions were accessed from the transporter classification database (Saier et al. 2006); biosynthesis reactions were collected from the biosynthesis pathways encoded in KEGG and curated by using additional literature (for details, see Supplemental Methods). The final metabolic map covers the metabolism, transport, and biosynthesis of carbohydrate, amino acids, fatty acids and lipids, glycan, and nucleic acids in human and mouse, including 862 genes of 390 enzymes, 1880 reactions, 1219 metabolites, and 116 transporter genes of 35 metabolites in human. Complete lists of genes and reactions of the collected human metabolic map is given in Supplemental Table S1.

Reconstruction of the metabolic map into a factor graph

The metabolic reaction map naturally forms a directed factor graph when considering each reaction as a variable and each metabolite as a factor. A directed factor graph was first reconstructed by the stoichiometric matrix of all reactions in the global metabolic map. In the factor graph, variable, factor, and directed edge are reactions, metabolites, and whether or not a reaction involves a metabolite as the substrate or product, respectively. In this study, we use a flux balance assumption of carbon-based metabolites. Therefore, 273 compounds that do not affect the flux balance of carbon-based molecules were excluded from the stoichiometric matrix, such as H_2O, ATP, NADH, or other cofactors (for a complete list, see Supplemental Table S1). We further reduced the complexity of the factor graph based on its topological structure. In this step, connected reactions were merged into a module if (1) none of the merged intermediate metabolites had more than one influx or outflux reaction that corresponded to more than one module's inputs or outputs; and (2) none of the merged intermediate metabolites had an influx or outflux other than the merged reactions or the module input and output. We have proved that under these two conditions and the flux balance condition, changes of the reactions inside the module will not affect the reactions outside the module conditional on a fixed flux rate of the module. In other words, solving the flux of each individual reaction in a merged module is equivalent to solving the flux of the module (for details, see Supplemental Methods). The merged reactions will form a variable node containing multiple reactions in the factor graph, while the factor nodes are still individual metabolites. In addition, we identified certain classes of metabolites, including different types of fatty acids, pyrimidines, purines, and steroid hormones, that form highly connected web-like metabolic pathways. Instead of solving the flux for each individual metabolite, we consider the metabolites of the same class as one factor. The network reduction approach enables a more robust flux estimation of reaction modules instead of individual reactions and a more efficient computation over the simplified network topological structure.

We reconstructed the human metabolic map into a factor graph consisting of 169 modules of 22 supermodule classes, 862 genes, and 128 metabolites, of which 66 are intermediate substrates. Here, each supermodule is a manually curated group of modules with similar functions (Table 1). More details on the

Table 1. Supermetabolic module information

Supermodule ID	Supermodule class	Number of modules	Number of genes
1	Glycolysis + TCA cycle	14	83
2	Serine metabolism	18	114
3	Pentose phosphate	1	28
4	Fatty acids metabolism/ synthesis	2	81
5	Aspartate metabolism	5	35
6	Beta-alanine metabolism	5	48
7	Propionyl-CoA metabolism	2	25
8	Glutamate metabolism	5	13
9	Leucine + valine + isoleucine	8	99
10	Urea cycle	8	30
11	Spermine metabolism	2	7
12	Transporters	35	80
13	Hyaluronic acid synthesis	5	26
14	Glycogen synthesis	1	4
15	Glycosaminoglycan synthesis	1	14
16	N-linked glycan synthesis	12	88
17	O-linked glycan synthesis	4	17
18	Sialic acid synthesis	3	12
19	Glycan synthesis	1	5
20	Purine synthesis	17	67
21	Pyrimidine synthesis	17	49
22	Steroid hormone synthesis	3	177

factor graphs can be found in Supplemental Table S1 and Supplemental Figure S1. Figure 2A illustrates the functional group and complete topological structure of the collected metabolic modules and supermodules in human. Figure 2B illustrates several examples of how network motifs in the input metabolic network are merged into one metabolic module.

For a given scRNA-seq data and a user defined task, scFEA further refines the task-specific metabolic factor graph by (1) limiting the analysis to user-selected metabolic networks, and (2) excluding the modules without significantly expressed genes. For (2), scFEA will first determine for all the genes whether they have a significant non-zero expression state, using our in-house left-truncated mixture Gaussian model (Methods; Wan et al. 2019). By default, scFEA considers a module as blocked if it becomes disconnected with other modules after removing the reactions whose associated genes do not have a non-zero expression state. The blocked modules will be excluded from further analysis. On account of the common dropout events in scRNA-seq data, scFEA allows keeping a module as long as at least one of the genes involved in this module has significantly non-zero expressions.

The topological structure of metabolic modules including input, output, and intermediate metabolites and genes associated with each module, serve as the input to our graph neural network model.

Mathematical formulation of metabolic flux estimation in individual cells

For a clear model setup, we first formulate the metabolic network as a directed factor graph. Here, each metabolic module is represented as a variable, and each compound is represented as a factor node carrying a loss function that evaluates the level of flux imbalance among modules, and the direction represents whether a metabolite is the input or output of a metabolic module (Supplemental Fig. S1). We denote $FG(C^{1\times K}, RM^{1\times M}, E = \{E_{C\rightarrow R}, E_{R\rightarrow C}\})$ as the factor graph, where $C^{1\times K} = \{C_k, k = 1, \ldots, K\}$ is the set of K compounds; $RM^{1\times M} = \{R_m, m = 1, \ldots, M\}$ is the set of M metabolic modules; and $E_{R\rightarrow C}$ and $E_{C\rightarrow R}$ represent direct edges from module to compound and from compound to module, respectively. For the kth compound C_k, we define the set of reactions producing and consuming C_k as $F_{\text{in}}^{C_k} = \{R_m | (R_m \rightarrow C_k) \in E_{R\rightarrow C}\}$ and $F_{\text{out}}^{C_k} = \{R_m | (C_k \rightarrow R_m) \in E_{C\rightarrow R}\}$, which is derived from the stoichiometric matrix of the whole metabolic map. For an scRNA-seq data set with N cells, we denote $\text{Flux}_{m,j}$ as the flux of the mth module in the cell j, $j = 1 \ldots N$, and let $F_j = \{\text{Flux}_{1,j}, \ldots, \text{Flux}_{M,j}\}$. Our computational hypothesis is that the total flux imbalance of the intermediate metabolites throughout all the collected cells should be minimized, based on which we developed the likelihood function of the flux of all modules throughout all cells as

$$\phi(C, F) = \prod_{j=1}^{N}\prod_{k=1}^{K} \phi(C_{k,j}|F_j)\varphi(F_j),$$

where

$$\phi(C_{k,j}|F_j) = \phi(C_{k,j}|F_{\text{in}}^{C_k}, F_{\text{out}}^{C_k}) \propto e^{-\frac{\lambda\left(\sum_{m\in F_{in}^{C_k}} \text{Flux}_{m,j} - \sum_{m'\in F_{\text{out}}^{C_k}} \text{Flux}_{m',j}\right)^2}{2}}$$

and $\phi(F_j)$ represents the prior distribution of the fluxome in cell j, and λ is a tuning hyperparameter. scFEA models the flux of reach reaction, $\text{Flux}_{m,j}$, as a nonlinear function of the expression changes of the genes associated with the module. Denote $\boldsymbol{G^m} = \{G_1^m, \ldots, G_{i_m}^m\}$ as the genes associated with the reactions in R_m, and $\boldsymbol{G_j^m} = \{G_{i_1,j}^m, \ldots, G_{i_m,j}^m\}$ as their expressions in sample j, where i_m stands for the number of genes in R_m. We model $\text{Flux}_{m,j} = f_{nn}^m(\boldsymbol{G_j^m}|\ \boldsymbol{\theta_m})$ as a multilayer fully connected neural network with the input $\boldsymbol{G_j^m}$, where $\boldsymbol{\theta_m}$ denotes the parameters of the neural network (Fig. 3). It is noteworthy that the cell group and tissue context–specific distribution of the flux $\phi(F_j)$ and the reaction parameters $\boldsymbol{\theta_m}$ are always unknown. Apparently, without further constraints, $\text{Flux}_{m,j} \equiv 0$ is a trivial solution. To provide a robust and rational solution, we introduced two additional constraints to $\text{Flux}_{m,j}$, namely, (1) the predicted flux, $\text{Flux}_{m,j}$, should be nonnegative; and (2) within a supermodule (Fig. 2A), the total predicted flux should be correlated with gene expression variation. The second assumption assumes that the metabolic flux variation within large metabolic modules should be coherent to their gene expression change, which is supported by recent studies (Damiani et al. 2019; Wagner et al. 2021). This assumption effectively avoids the trivial solution. Hence, instead of directly maximizing $\phi(C, F)$, we solve the $\boldsymbol{\theta_m}$ and cell-wise flux $\text{Flux}_{m,j}$ by minimizing the following loss function L:

$$L = \sum_{j=1}^{N}\sum_{k=1}^{K}\left(\sum_{m\in F_{in}^{C_k}} \text{Flux}_{m,j} - \sum_{m'\in F_{\text{out}}^{C_k}} \text{Flux}_{m',j}\right)^2 + \alpha\sum_{j=1}^{N}\sum_{m=1}^{M}(|\text{Flux}_{m,j}| - \text{Flux}_{m,j})$$
$$+ \beta\sum_{j=1}^{N}[1 - |\text{cor}(\text{Flux}_{:,j}^{SM}, \text{GE}_{:,j}^{SM})|] + \gamma\sum_{j=1}^{N}\left(\sum_{m=1}^{M}|\text{Flux}_{m,j}| - TA_j\right)^2,$$

where α, β, and γ are hyperparameters; cor represents Pearson correlation coefficient; Flux^{SM} and GE^{SM} are two NSM × N matrices, here NSM is number of supermodules; $\text{Flux}_{m,j}^{SM}$ represents the sum of the flux of the modules in the supermodule m; $\text{GE}_{m,j}^{SM}$ represents the sum of expression of the genes in the supermodule m, in cell j; and TA_j is a surrogate for total metabolic activity level of cell j,

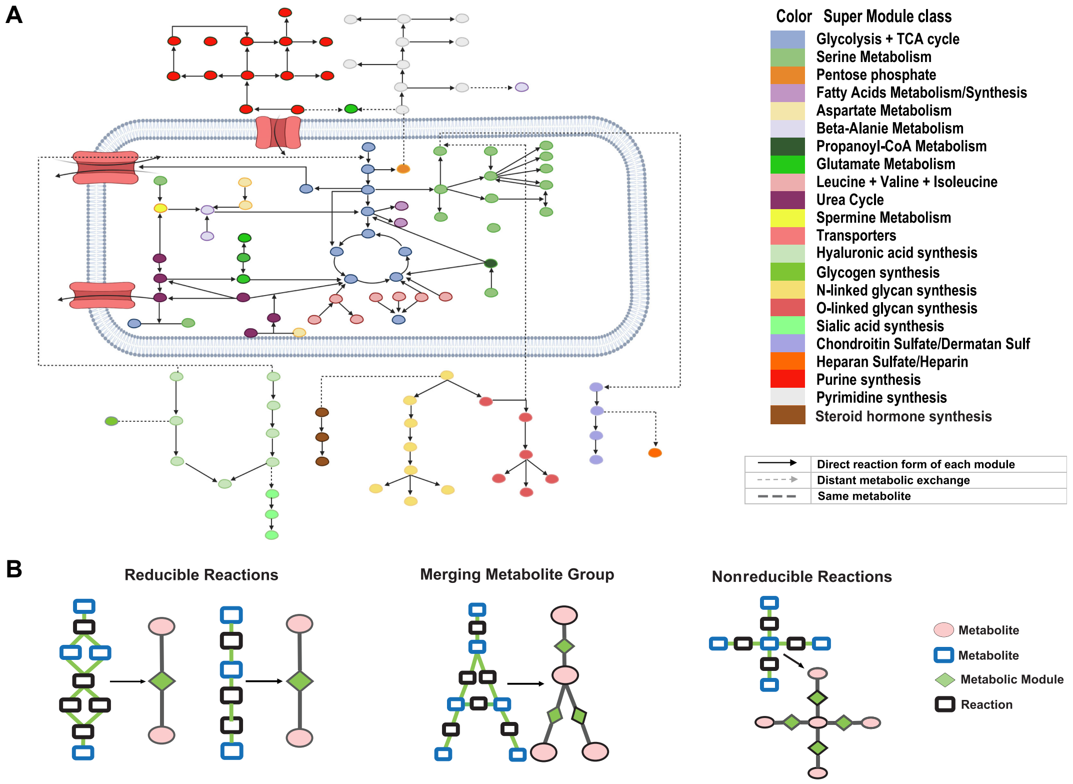

Figure 2. Reduced and reconstructed human metabolic map. (*A*) Collected human metabolic modules and supermodule classes. (*B*) Examples of how the network motifs in the metabolic map are simplified into metabolic modules, where the reactions and metabolites are represented by black and blue rectangles, and modules and metabolites are colored by green and pink. Chainlike reactions can be directly simplified; a complicated module connected by multiple branches can be shrunk into one point linked with the multiple in/out branches; and complicated intersections cannot be simplified.

which is assigned as the total expression of metabolic genes in cell *j*. The first, second, third, and fourth terms of *L* are related to constraints on flux balance, non-negative flux, the coherence between predicted flux and total gene expression level of each supermodule, and the relative scale of flux, respectively. Here Pearson's correlation, which is scale-free, is used to model the coherence between gene expression and predicted flux, as genes may have varied intrinsic expression range. Our empirical and robustness analyses suggested that $\alpha=1, \beta=0, \gamma=1$, and $\alpha=1, \beta=1, \gamma=1$ result in a good leverage of the flux balance loss and other constraints for Smart-seq2 and 10x Genomics data, respectively (for details, see robustness analysis and Supplemental Methods).

The preceding formulation defines a new graph neural network architecture for flux estimation over a factor graph: on one hand, each variable is defined as a neural network of biologically meaningful attributes, that is, the genes participating in each metabolic module; on the other hand, the information aggregation between adjacent variables is constrained by the balance of the influx and outflux of each intermediate metabolite. The number of intermediate constraints (*K*) and large number of cells (*N*) of scRNA-seq data ensures the identifiability of $\boldsymbol{\theta_m}$ for the multilayer f_{nn}^m at a certain complexity level. Detailed analysis of the computational feasibility, scalability, tuning of hyperparameters, and options of additional loss terms are provided in Methods and Supplemental Methods.

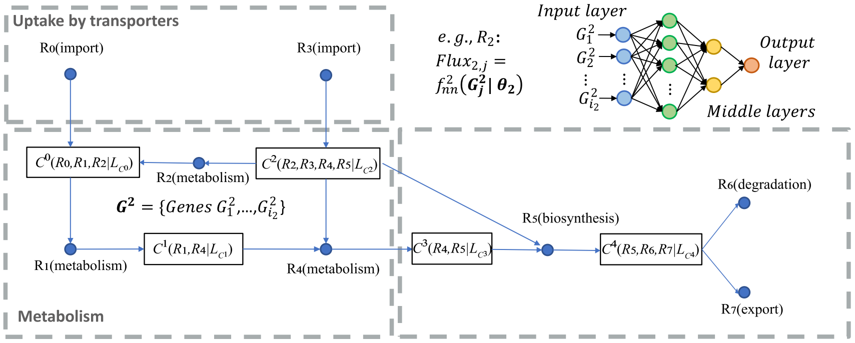

Figure 3. A toy model of the factor graph of metabolic modules, flux balance conditions, and the flux model for the module R_2 (*top right*). In the factor graph, each *C* (metabolites) corresponds to one flux balance condition and serves as a factor, and each *R* (can be a reaction or a module) is a variable. For example, $C^0(R_0, R_1, R_2|L_{c0})$ simply represents that the metabolite C^0 is determined by the flux balance loss of R_0, R_1, R_2, where L_{c0} is the flux balance term of C^0. Import and export/degradation reactions are considered as having no input or output substrates.

The challenges to minimize the loss function *L* include the following: (1) the balance of one intermediate substrate is influenced by multiple modules, hence updating the module flux one at a time may not be computationally efficient;

and (2) the updating strategy for a large group of fluxes cannot be theoretically derived. The two challenges prohibit a direct usage of back propagation or gradient descending methods. We developed an effective optimization strategy for L by adopting the idea of information transfer in belief propagation, which has been commonly used in analyzing cyclic networks such as Markov random field (Lan et al. 2006). Specifically, L is minimized by iteratively minimizing the flux imbalance of C_k and the weighted sum of the flux imbalance of the Hop-2 neighbors of C_k in the factor graph, as the L_k^* defined below:

$$L_k^* = \sum_{j=1}^{N} \left(\sum_{m \in F_{in}^{C_k}} \mathrm{Flux}_{m,j} - \sum_{m' \in F_{\mathrm{out}}^{C_k}} \mathrm{Flux}_{m',j} \right)^2 + \sum_{k'} W_{k'} \sum_{j=1}^{N} \left(\sum_{m \in F_{in}^{C_{k'}}} \mathrm{Flux}_{m,j} - \sum_{m' \in F_{\mathrm{out}}^{C_{k'}}} \mathrm{Flux}_{m',j} \right)^2,$$

where $C_{k'}$ are the Hop-2 neighbors of C_k, and $W_{k'}$ is proportional to the current total imbalance of all the Hop-2 neighbors of $C_{k'}$ except for C_k itself (for more details, see Methods). Here the Hop-2 neighbors of a compound (or module) on the factor graph is defined as all other compounds (or modules) having a connection with the modules (or compounds) who connect to the compound (or module). Such a regional perturbation strategy over the whole graph can effectively leverage the search of global minimum and computational feasibility.

The output of scFEA includes f_{nn}^m, $\boldsymbol{\theta_m}$ for each module and predicted cell-wise metabolic flux $\mathrm{Flux}_{m,j}$. The predicted flux $\mathrm{Flux}_{m,j}$ is a relative measure of unfixed scale. However, $\mathrm{Flux}_{m,j}$ is comparable among cells ($\mathrm{Flux}_{m,}$) or metabolic modules ($\mathrm{Flux}_{,j}$).

Method validation on an scRNA-seq data with perturbed metabolic conditions and matched metabolomics data

To validate the cell-wise flux estimated by scFEA, we generated an scRNA-seq data set consisting of 162 patient-derived pancreatic cancer cells (Pa03c cell) under two crossed experimental conditions: *APEX1* knockdown (*APEX1* KD) or control, and under hypoxia or normoxia conditions (for detailed experimental procedure and data processing, see Methods). Metabolomics profiling of 14 metabolites were collected on bulk wild-type Pa03c cells and *APEX1* inhibition cells under the normoxia conditions, each with three replicates (Supplemental Table S2). The 14 metabolites include glucose, glucose-1 phosphate, glucose-6 phosphate, pyruvate, and lactate in the glycolysis pathway, citrate, 2-oxoglutarate, succinate, fumarate, malate in the TCA cycle, and amino acids glutamate, glutamine, serine, and ornithine. We used the Smart-seq2-fluidigm protocol for single-cell RNA sequencing. It allows for saturated gene detection of each single cell, to enable a more accurate modeling of metabolic flux. *APEX1* is a multifunctional protein that interacts with multiple transcriptional factors (TFs) to regulate cellular responses to hypoxia and oxidative stress (Kelley et al. 2012). Our previous studies identified significant roles of *APEX1* in the regulation of Pa03c cells' response to metabolic environment changes (Shah et al. 2017; Wan et al. 2019).

To the best of our knowledge, scFEA is the first computational tool to estimate metabolic flux at single-cell level. Without baseline methods for comparisons, we validate scFEA by examining the consistency between the metabolic flux variation predicted by scFEA and experimental observations. We identified 126 up-regulated and 443 down-regulated genes in *APEX1* KD versus control under the normoxia condition, and 260 up-regulated and 1496 down-regulated genes under hypoxia condition. Pathway enrichment analysis showed that the TCA cycle (normoxia: $P = 0.003$, hypoxia: $P = 1.12 \times 10^{-7}$) and oxidative phosphorylation (normoxia: $P = 3.17 \times 10^{-4}$, hypoxia: $P = 1.77 \times 10^{-8}$) pathways are significantly enriched by down-regulated genes, under both normoxia and hypoxia conditions. This suggests that the knockdown of *APEX1* may lead to inhibited cellular aerobic respiration. In addition, genes regulated by hypoxia-inducible factor 1-alpha (*HIF1A*), including glycolysis and TCA cycle genes, were observed to be up-regulated in hypoxia conditions compared with normoxia conditions in the control Pa03c cells. This is consistent with the common knowledge of hypoxia response. Of the 14 metabolites, we have seen an increase of abundance in glucose, glucose-1 phosphate, glucose-6 phosphate, and lactate, and decrease in 2-oxoglutarate, succinate, fumarate, and malate in *APEX1*-KD versus control cells under the normoxia condition. In summary, analysis of the single-cell gene expression and bulk cell metabolomic data revealed that knockdown of *APEX1* affects the cells' glucose metabolism and inhibits the cells' TCA cycle pathway under both normoxia and hypoxia condition. Figure 4A illustrates the variation of genes and metabolites involved in glycolysis, pentose phosphorylation, TCA cycle, glutaminolysis, and aspartate metabolism pathways in *APEX1*-KD versus control under the normoxia condition. We conducted a qRT-PCR experiment to confirm the down-regulated genes in glycolysis, TCA cycle, and oxidative phosphorylation pathways (Supplemental Fig. S2). A complete list of differentially expressed genes and pathway enrichment results are provided in Supplemental Table S3.

Consistency between the scFEA-predicted flux variation and the metabolomics data

We applied scFEA to the aforementioned scRNA-seq data of the four conditions, with hyperparameters $\alpha = 1$, $\beta = 0$, and $\gamma = 1$. We first focus on the normoxia conditions in which matched single-cell expression and metabolomics data are available. scFEA-predicted decreased metabolic flux for the modules in glycolysis and TCA cycle in *APEX1*-KD versus control, that is, glucose → D-Glucose 1-phosphate (G1P) → alpha-D-Glucose 6-phosphate (G6P) → glyceraldhyde-3P (G3P) → 3-Phospho-D-glyceroyl phosphate (3PD) → pyruvate → acetyl-CoA → citrate → 2-Oxoglutarate (2OG) → succinate-CoA →succinate → fumarate → malate → oxaloacetate (OAA) and pyruvate → lactate. Particularly, the reactions toward the downstream from this reaction chain has even lower flux in *APEX1*-KD versus control (Fig. 4B). We then examined the Pearson's correlation between the averaged predicted flux change with the observed metabolomic change of intermediate metabolites in glycolysis and TCA cycle pathways. In *APEX1*-KD versus control cells under normoxia condition, we observed a Pearson correlation coefficient (PCC) of 0.86 ($P = 0.006$) (Fig. 4B), suggesting the high consistency between predicted flux variation with the observed metabolic changes. Using metabolomics data, we observed an increase of production for glucose, G1P, G3P, and lactate, and a decrease of production for 2OG, succinate, fumarate, and malate in *APEX1*-KD versus control (Fig. 4C). By the Michaelis–Menten model, the substrates of largely varied concentration determine the reaction rate in a nonlinear manner (close to linear when the reaction is less saturated). Hence, variations in the concentration of the metabolites with one dominating outflux could partially reflect the changes of the outflux rate. We also correlated the metabolomic change with the averaged expression change of the

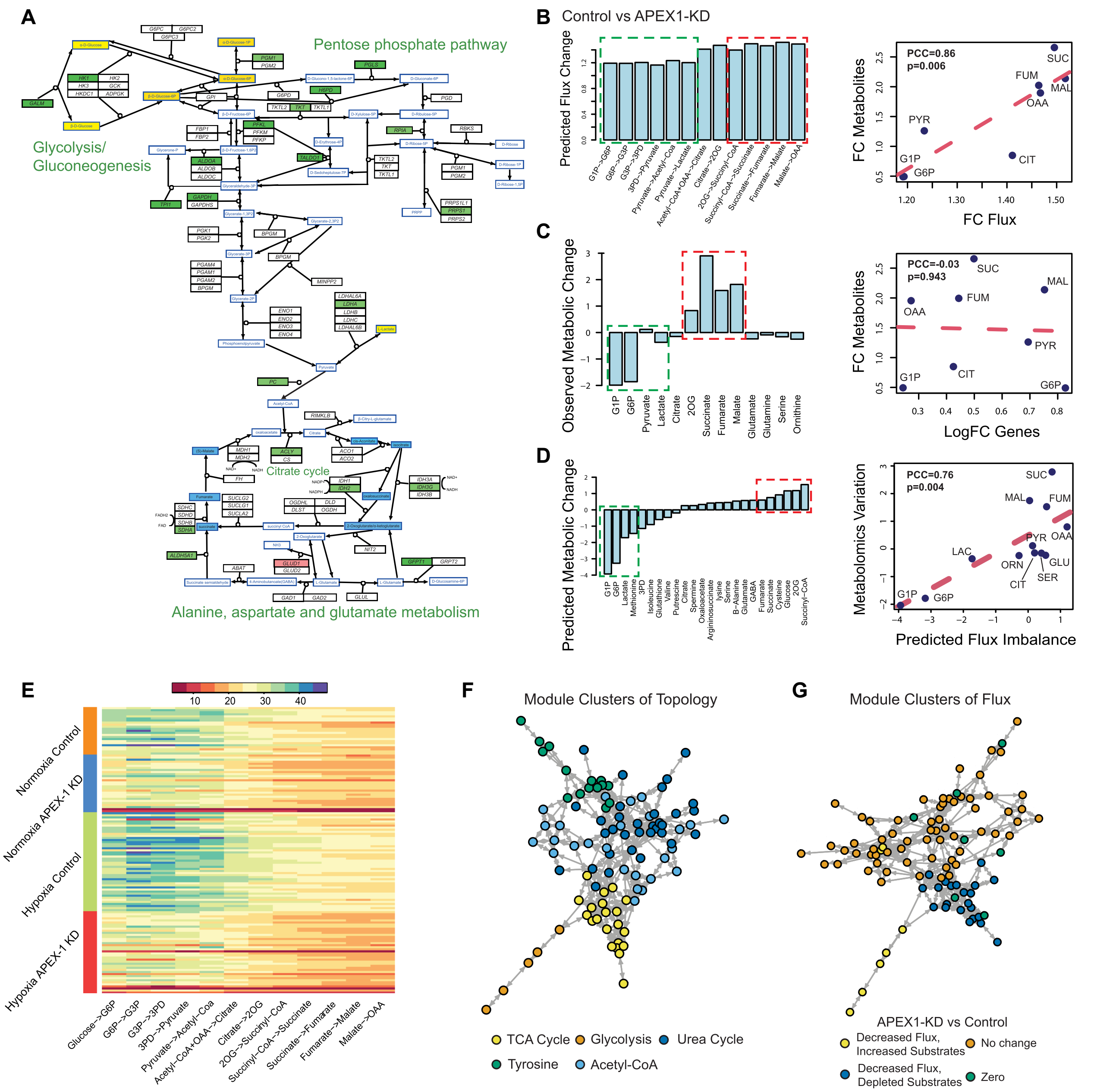

Figure 4. Application of scFEA on matched scRNA-seq and metabolomics data of Pa03C cells. (*A*) Gene expression and metabolomic variations of the glycolysis, pentose phosphate, TCA cycle, glutamine, and aspartate metabolic pathways in *APEX1*-KD versus control under normoxia condition. Genes/metabolites are shown in rectangular boxes with black/blue borders, up-regulated/down-regulated genes are colored in red/green, increased and decreased metabolites are colored in yellow/blue, respectively. The darker color suggests a higher variation. (*B*) Predicted flux fold change (*left*, *x*-axis: metabolic module, *y*-axis: predicted flux change) in control versus *APEX1*-KD, and correlation between fold change of predicted flux and observed metabolite change (*right*, *x*-axis: fold change of predicted flux, *y*-axis: fold change of observed metabolite abundance, each data point is one metabolite). (PYR) pyruvate; (CIT) citrate; (FUM) fumarate; (SUC) succinate; (MAL) malate. (*C*) Observed metabolomic change (*left*, *x*-axis: metabolites, *y*-axis: abundance difference observed in the metabolomics data) in control versus *APEX1*-KD, and correlation between log fold change of gene expressions involved in each reaction and observed metabolomics change (right, *x*-axis: log fold change of the averaged expression of the genes involved in each reaction, *y*-axis: fold change of observed metabolites abundance observed in the metabolomics data, each data point is one metabolite). (*D*) Predicted metabolic stress (*left*, *x*-axis: metabolites, *y*-axis: predicted abundance difference) in control versus *APEX1*-KD and correlation between predicted metabolic stress and observed difference in metabolite abundance (*right*, *x*-axis: top scFEA-predicted imbalance of the influx/outflux of intermediate metabolites, *y*-axis: difference of observed metabolomic abundance, in control versus *APEX1*-KD, each data point is one metabolite: (LAC) lactate; (SER) serine; (GLU) glutamine; (ORN) ornithine. In *B–D*, all comparisons were made by comparing control versus *APEX1*-KD under normoxia. The fold change of metabolomic abundance is used in calculating the correlation in *B–C* and difference of metabolomic abundance is used in *D*. The green and red dashed blocks represents the accumulated (green) and depleted (red) metabolites in Control versus *APEX1*-KD. (*E*) Profile of the predicted fluxome of 13 glycolytic and TCA cycle modules. Here, each column represents the flux between two metabolites (shown on the *x*-axis) for all the cells of the four experimental conditions (shown on the *y*-axis). For two neighboring fluxes, the product of the reaction on the *left* is the substrate of the reaction on the *right*, and in a perfectly balanced flux condition, the two neighboring fluxes should be equal. (*F*) Clusters of metabolic modules inferred by using the network connectivity structure only. (*G*) Clusters of metabolic modules inferred by using the network topological structure (weight of 0.3) combined with the predicted fluxome (weight of 0.7).

enzymes catalyzing the reactions. However, no significant correlation was observed (PCC = −0.03, P = 0.943) (Fig. 4C), suggesting that single-cell gene expression alone does not produce a good estimate of single-cell metabolomic landscape. In addition, single sample gene set enrichment analysis (ssGSEA) has been used to model cell-wise pathway activity in scRNA-seq data (Chen et al. 2020). Again, the correlation between the metabolomic changes and the differences in averaged ssGSEA score in *APEX*1-KD versus control cells is not significantly large (PCC = 0.42, P = 0.299) (Supplemental Fig. S3). Here, we showed that scFEA-predicted metabolic flux is much more consistent to the true metabolomics changes, as it leveraged the nonlinear relationships between gene expression and enzymatic reaction rate and the flux balance constraints of the metabolites.

High consistency of the predicted metabolic stress with experimentally observed metabolomic changes

scFEA allows us to investigate the cell-wise metabolic stress, which was defined as the imbalance of the influxes/outfluxes of each intermediate metabolite in each cell. Figure 4D shows that the G1P, G6P, and lactate were accumulated while 2OG, succinate, succinyl-CoA, and fumarate were depleted in *APEX1*-KD versus control. A PCC of 0.75 (P = 0.004) was observed between the predicted metabolic stress and the true metabolic change on 12 metabolites with both measured metabolomic profile and predicted metabolic stress. This shows the high accuracy of the predicted and observed metabolic stress level. Details on the predicted and observed metabolic imbalance were provided in Supplemental Table S3. Figure 4E shows the predicted cell-wise fluxome of the glycolysis and TCA cycle modules for cells of the four conditions. We observed, in general, higher flux of the glycolytic modules than the TCA cycle modules, with the largest average flux gap seen on pyruvate → acetyl-CoA and acetyl-CoA → citrate. In addition, the flux of the downstream reactions (citrate → 2OG → succinyl-CoA → succinate) of the TCA cycle is lower than the upstream reactions (succinate → fumarate → malate → OAA). A possible explanation for the leaky metabolic flux is that some of the intermediate substrates flow to other branches, primarily for biosynthesis of amino acids. Among the four conditions, we identified that the hypoxia control group has the highest flux rate of glycolysis and TCA cycle modules. Clearly, the inhibition of *APEX1* significantly decreased the metabolism rate of glucose. Seeing the accumulations of glycolytic substrates and depletions of TCA cycle substrates, we speculate that the knockdown of *APEX1* may directly impact the downstream part of glycolysis, the whole TCA cycle and further oxidative phosphorylation, leading to accumulation of G1P and G6P as a result of the blockage. Up-regulation of glucose transporters was also observed in *APEX1* KD versus control, further suggesting the accumulation of glycolytic substrates.

Perturbation analysis to detect key flux determining genes

We also conducted a perturbation analysis to detect the key genes with high impact on each metabolic module (for details, see Methods). The following genes were identified to have the highest impact on metabolic flux: *HK1* and *HK2* (Glucose→G6P, EC: 2.7.1.1); *ALDOA*, *PFKL*, and *GPI* (G6P→G3P, EC: 5.3.1.9); *GAPDH* and *PGK1* (G3P→3PD, EC: 1.2.1.12, 2.7.2.3); *ENO1*, *PGAM1*, and *PKM* (3PD→Pyruvate, EC: 5.4.2.11, 4.2.1.11); *PDHA2* (Pyruvate→acetyl-CoA, EC: 1.2.4.1); *LDHA* (Pyruvate→Lactate, EC: 1.1.1.27); *ACLY* (acetyl-CoA + OAA→Citrate, EC: 2.3.3.8); *IDH2* (Citrate→2OG, EC: 1.1.1.42); *DLD* and *OGDH* (2OG→Succinyl-CoA, EC: 1.2.4.2); *SUCLG1* (Succinyl-CoA→Succinate, EC: 6.2.1.4); *SDHA* (Succinate→Fumarate, EC: 1.3.5.1); *FH* (Fumarate→Malate, EC: 4.2.1.2); and *MDH1* (Malate→OAA, EC: 1.1.1.37). Detailed results of the perturbation analysis were illustrated in Supplemental Figure S4. A qRT-PCR experiment was conducted to confirm the down-regulation of the aforementioned key metabolic genes, including *HK1*, *PFKL*, *ACLY*, *SDHA*, and *IDH2* (Supplemental Fig. S2). We also compared the predicted high impact enzymes in the modules containing multiple enzymes (seven in total) with the rate-limiting enzymes reported in the rate-limiting enzymes database (RLEdb) (Zhao et al. 2009). We observed that six of the seven predicted high impact enzymes, namely, 2.7.1.1, 1.2.1.12, 2.7.2.3, 5.4.2.11, 1.2.4.1, and 1.2.4.2, have been reported in RLEdb, suggesting a significant enrichment (P = 0.0005 by Fisher's exact test) of our predictions to RLEdb. We further conducted a module level perturbation analysis by increasing or decreasing the expression of genes in a certain module (Methods). Consistent to our experimental observations, a decrease of expression on genes of the downstream part of glycolysis pathway in the control cells will lower the flux of the TCA cycle, causing the accumulation of glycolytic intermediate substrates and depletion of TCA cycle metabolites.

Detecting groups of metabolic modules with similar variations and cells with distinct metabolic states

We also applied scFEA to a larger metabolic map, with the 11 metabolic supermodules and transporters. Figure 4F illustrated five distinct groups of metabolic modules derived using a spectral clustering method purely based on their network topology (see Methods), namely, (1) glycolysis, (2) TCA cycle and glutamine metabolism–related modules, (3) tyrosine and serine metabolism, (4) urea cycle–related modules, and (5) acetyl-CoA-related metabolisms such as fatty acids and propanoyl-CoA metabolisms. To examine the high-level structure based on the flow of flux, we conducted a clustering analysis of the metabolic modules by considering both the network connectivity and flux similarity. The distance between two modules R_i and R_j is defined as $\alpha d(R_i, R_j) + (1-\alpha)d^F(R_i, R_j)$, where $d(R_i, R_j)$ is the normalized spectral distance based on the metabolic network connectivity, and $d^F(R_i, R_j)$ is the normalized similarity based on the estimated flux of all the normoxia cells (Methods). Here, $\alpha = 0.3$ is used in the analysis. Figure 4G shows the metabolic module clusters by integrating topological structure and flux similarity. Four distinct clusters were identified, including (1) glycolysis and fatty acids metabolism of decreased flux and accumulated substrates in *APEX1*-KD versus control, (2) TCA cycle and pyruvate metabolism with decreased flux and depleted substrates, (3) metabolism of amino acids and other metabolites with unchanged flux and metabolites, and (4) a few other modules of 0 flux rates, respectively. This observation further validated the rationality of the scFEA-predicted fluxome.

We also conducted cell clustering based on the estimated single-cell flux (Methods). It is no surprise that the cell clusters coincide with experimental conditions, forming five groups of cells of high, intermediate, and low metabolic rates, high lactate production, and low TCA cycle rate (Supplemental Fig. S5).

Method validation and robustness analysis on synthetic and independent real-world data sets

Method validation on independent real-world data

We also validated scFEA on an independent scRNA-seq data of perivascular adipose tissue derived mesenchymal stem cells (PV-

ADSC) (NCBI Gene Expression Omnibus [GEO; https://www.ncbi.nlm.nih.gov/geo/] accession number GSE132581) (Gu et al. 2019) by using hyperparameters $\alpha = 1$, $\beta = 0$, and $\gamma = 1$. To the best of our knowledge, this data set, in addition to our newly generated data set, are the only two scRNA-seq data with matched tissue-level targeted metabolomics profiling available in the public domain. We first reconducted cell clustering analysis and identified two distinct PV-ADSC cell clusters corresponding to different levels of differentiation, as reported in the original work (Fig. 5A; Gu et al. 2019). Here, the clusters were visualized using UMAP (McInnes et al. 2018). Owing to the small sample size (85 cells), scFEA was applied to estimate only the fluxome of glycolysis and TCA cycle pathways. We observed an increased flux of glycolytic reactions ($P < 1.56 \times 10^{-6}$), lactate production ($P = 0.002$), and the reactions from *cis*-aconitate to oxaloacetate in the TCA cycle ($P < 0.02$) in the more differentiated (MD) versus the high stemness (HS) PV-ADSC cells. The reactions from acetyl-CoA to citrate were not significantly changed ($P = 0.887$) (Fig. 5B). This is consistent with the observations made on the metabolomics data in the original work; that is, the glycolytic intermediate metabolites, lactate production, and metabolites in the later part of TCA cycle were elevated in the MD cells, but citrate was not significantly changed. We also analyzed the metabolic modules of two amino acid supermodules with metabolomics profile reported in the original study, namely, valine and isoleucine metabolism and glutamate and glutathione metabolism (Supplemental Fig. S6). Elevated valine and isoleucine metabolic flux in MD versus HS cells has been predicted by scFEA, which is consistent to the original report. scFEA also predicted an increased flux of the modules from glutathione → glutamate → glutamine → TCA cycle; this could explain the increased flux rate of TCA cycle but less increase in citrate production. The original study only reported a depletion of glutathione and glutamate; our metabolic stress analysis also predicted more decreased glutathione and glutamate in MD versus HS cells. Our analysis suggested that the elevated glutamate and glutathione metabolism is to fuel the substrate source for TCA cycle in MD cells, which depleted the concentration of glutathione and glutamate.

Method validation on randomly shuffled gene expression profile

In scFEA, we assume that the flux distribution in each single cell should be constrained by the flux balance condition, whereas the reaction rate of each module could be modeled as a nonlinear function of the gene expressions involved in this module. These two assumptions suggested that the distribution of the gene expressions involved in the metabolic modules was constrained by a set of equations governed by the metabolic flux distribution and the flux balance condition. One existing evidence that directly supports our assumptions is that the expression of closely related metabolic genes tend to be co-up-regulated or co-down-regulated

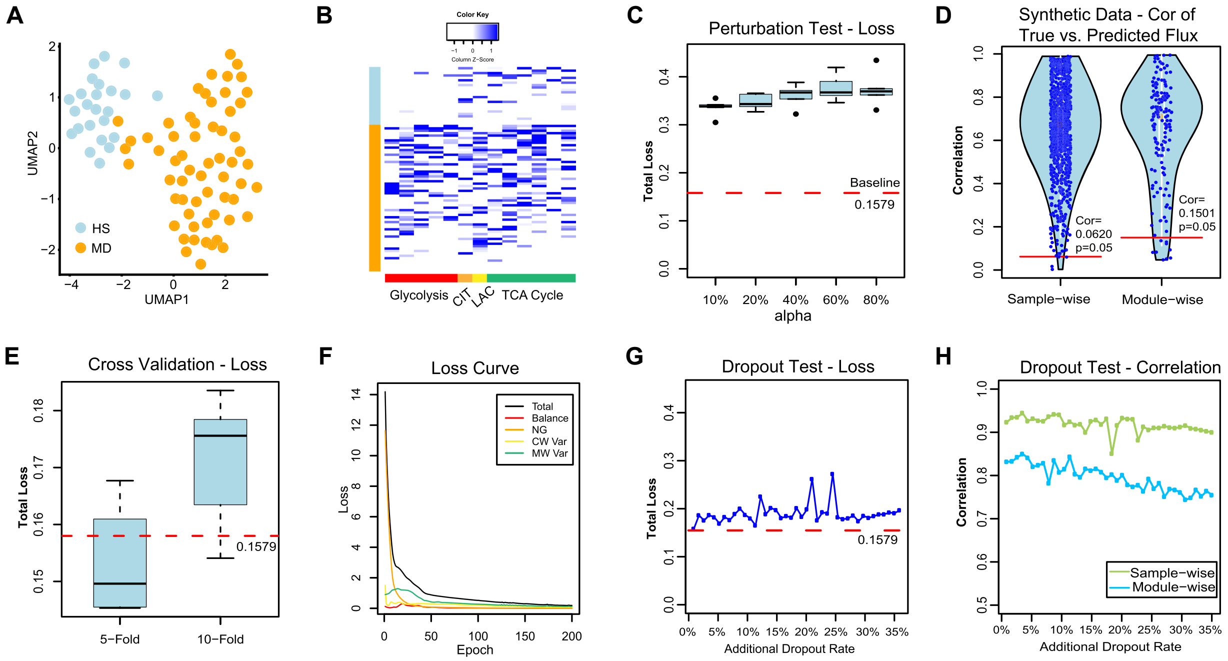

Figure 5. Method validations on real-world and synthetic data sets. (*A*) UMAP-based clustering visualization of the GSE132581 PV-ADSC data, in which HS and MD stand for PV-ADSC of HS and more differentiation, respectively. (*B*) Distribution of predicted cell-wise flux of glycolytic and TCA cycle modules. Each row is one cell, and row side color bar represents HS and MD PV-ADSC by blue and orange, respectively. Each column is one module. The *left* five columns (red) are glycolytic modules from glucose to acetyl-CoA, the CIT column (orange) is the reaction from acetyl-CoA to Citrate, the LAC column (yellow) is the reaction from pyruvate to lactate, and the *right* six columns (green) are TCA cycle modules from citrate to oxaloacetic acid. (*C*) The total loss (*y*-axis) for cases in which different proportion (*x*-axis) of cell samples have randomly shuffled gene expressions of the pancreatic cancer cell line data. The baseline loss 0.1579 was computed using the original expression profile of all 166 cells. (*D*) The sample-wise and module-wise correlation (*y*-axis) between the true and predicted module flux in synthetic data-based method validation with multiple repetitions, in which Cor = 0.5775 ($P = 0.05$) and 0.5778 ($P = 0.05$) correspond to the sample-wise and module-wise correlation, respectively. (*E*) Total loss (*y*-axis) computed under 5-fold/10-fold cross-validation (*x*-axis) versus baseline loss. (*F*) Convergency of the total loss and four loss terms during the training of neural networks on the pancreatic cancer cell line data. (*G*) Total loss (*y*-axis) computed from the robustness test by adding 0%–35% artificial dropouts to the original data (50.22% zero rate) versus baseline loss. (*H*) Sample-wise and module-wise correlation (*y*-axis) of the module flux predicted from the data with 0%–35% additional artificial dropouts with the module flux predicted from the original data.

(van der Knaap and Verrijzer 2016; Li et al. 2018). To further validate our assumption, we randomly shuffled the expression profile of each gene in a certain proportion (10%, 20%, 40%, 60%, and 80%) of cells in our pancreatic cancer cell line data and applied scFEA to each shuffled data (for details, see Supplemental Methods). We observed that the minimized total loss is positively associated with the level of perturbations (Fig. 5C) and the original scRNA-seq data achieved the smallest total loss, which partially supports our underlying assumption.

Method validation on synthetic data

We simulated matched metabolic flux and gene expression data on 1000 single cells. For the 1000 cells, we first randomly generated a different flux distribution of 169 connected modules from the solution space satisfying flux balance condition of these modules. The expression profile of the genes involved in each module was reversely simulated by assuming that its flux follows a fixed nonlinear function of the gene expressions. A detailed data simulation approach was provided in Supplemental Methods. We applied scFEA on the simulated single-cell gene expression profile and compared the fluxome predicted by scFEA and the known fluxome. We observed that scFEA-predicted fluxes are highly consistent to the true flux distribution, on both directions of the cells and metabolic modules (Fig. 5D). Specifically, >99.6% single cells achieved at least 0.0620 ($P = 0.05$) sample-wise correlation and >84.79% modules achieved at least 0.1501 ($P = 0.05$) module-wise correlation. Our analysis showed that under the assumption of scFEA, that is, if the flux balance constraint and nonlinear dependency between gene expression and metabolic hold, the formulation and solution strategy of scFEA could accurately estimate the cell-wise fluxome from single-cell gene expression data.

Robustness analysis based on perturbed sample inputs, cross-validation, and analysis of hyperparameters

We also tested the robustness of scFEA by 5-fold/10-fold cross validations on the pancreatic cancer cell line data. Compared with the baseline total loss achieved by using all cells, the total loss of the testing data does not change significantly when using different training cells to train the model (Fig. 5E). In training the neural networks, scFEA used Adam as the optimizer (Kingma and Ba 2015), which can adaptively adjust the learning rate. To choose the most suitable hyperparameters of the four terms in the loss function, we conducted experiments by changing the relative scale of any two terms and fixing the remaining two on the pancreatic cancer cell line data. We changed the relative ratio of two hyperparameters from 10 to 1000. Our experiments suggested a similar optimal solution can always be achieved under our hyperparameter perturbation range (Supplemental Fig. S7). Figure 5F showcases the convergence of the four loss terms and total loss in the model fitting of the pancreatic cancer cell line data. In addition, the applications on six real-world data (see further results) and simulated data suggested that the default hyperparameters always generate results of good convergence of the total loss and high biological implications. The default hyperparameters of the current version and details in hyperparameter running codes were provided via GitHub (https://github.com/changwn/scFEA).

Robustness analysis with respect to a different level of dropout

To further examine the method's robustness, we simulated different levels of additional dropout events to our pancreatic cancer cell line data. Our data was collected by using the Smart-seq2-fluidigm protocol, whose original ratio of zero expressions of the metabolic genes is 50.22%. We simulated additional dropout rate ranging from 4.34% to 34.78%, to reach a typical dropout level of a droplet-based scRNA-seq data (~85%), and applied scFEA on the tampered data (for details, see Supplemental Methods). We observed that the total loss slightly increases from 0.1649 to 0.2722 when the zero ratio increased from 50% to 85% (Fig. 5G). The module-wise and cell-wise correlation between the flux estimated from the original data and the tampered data are consistently higher than 0.7437 and 0.8505 (Fig. 5H), suggesting the high robustness of scFEA with respect to different levels of dropout events.

Application of scFEA on scRNA-seq data of tumor and brain microenvironment revealed distinct metabolic stress, exchange, and varied metabolic states in different types of cells

In this section, we primarily focused on validating the computational concept and applicability of scFEA on five real-world data sets, including two scRNA-seq data of cancer microenvironment, one single nuclei RNA-seq data of brain tissue, and one spatial transcriptomics data of breast cancer tissue. The data information is detailed in Supplemental Methods. All 169 metabolic modules across the whole metabolomic network were used in the analysis. Owing to the lack of matched metabolomics information, we focused on demonstrating the capability of scFEA in inferring metabolic flux, metabolic stress, and subgroups of cells and metabolic modules having distinct variations on these data sets.

Application on scRNA-seq data of cancer microenvironment

We applied scFEA on two publicly available scRNA-seq data sets collected from the microenvironment of melanoma (GEO GSE72056) and head and neck cancer (GEO GSE103322) by using hyperparameters $\alpha = 1$, $\beta = 0$, and $\gamma = 1$. In both data sets, we generated UMP-based cell and cell group visualization by using predicted fluxomes of the 169 modules (Fig. 6A–D). We identified that the metabolic flux distributions are quite homogeneous within cancer cells while being distinct from immune and stromal cells in both data sets (Fig. 6A,C). Distinct cell clusters of immune and stromal cells corresponding to varied metabolic fluxomes were also identified (Fig. 6B,D). A possible explanation is that cancer cells having a reprogrammed metabolism are more robust to the biochemical variations than immune and stromal cells in the tumor microenvironment.

We observed that the malignant cells have the highest metabolic rates in most metabolic reactions compared to other cell types in both melanoma and head and neck cancer, especially for the glucose and amino acids metabolic modules (Fig. 6E,F). On average, the TCA cycle and lactate production account for 43.4% and 52.5% of the total glycolysis flux in head and neck cancer, and 65.3% and 46.1% of the total glycolysis flux (with additional carbon flow from other metabolites such as glutaminolysis) in melanoma, respectively. In the nonmalignant cells, the ratio of lactate production is much lower. Our observation clearly suggested the existence of Warburg effect and metabolic shift in cancer cells, which is consistent to our previous findings of high lactate production in melanoma (Xu et al. 2012).

We identified that the malignant cells have the highest metabolic stress, which is defined as the total imbalance of intermediate substrates, followed by fibroblast and endothelial cells and then immune cells. Similar to the pancreatic cancer cell line data, we identified that both cancer and stromal cells in both

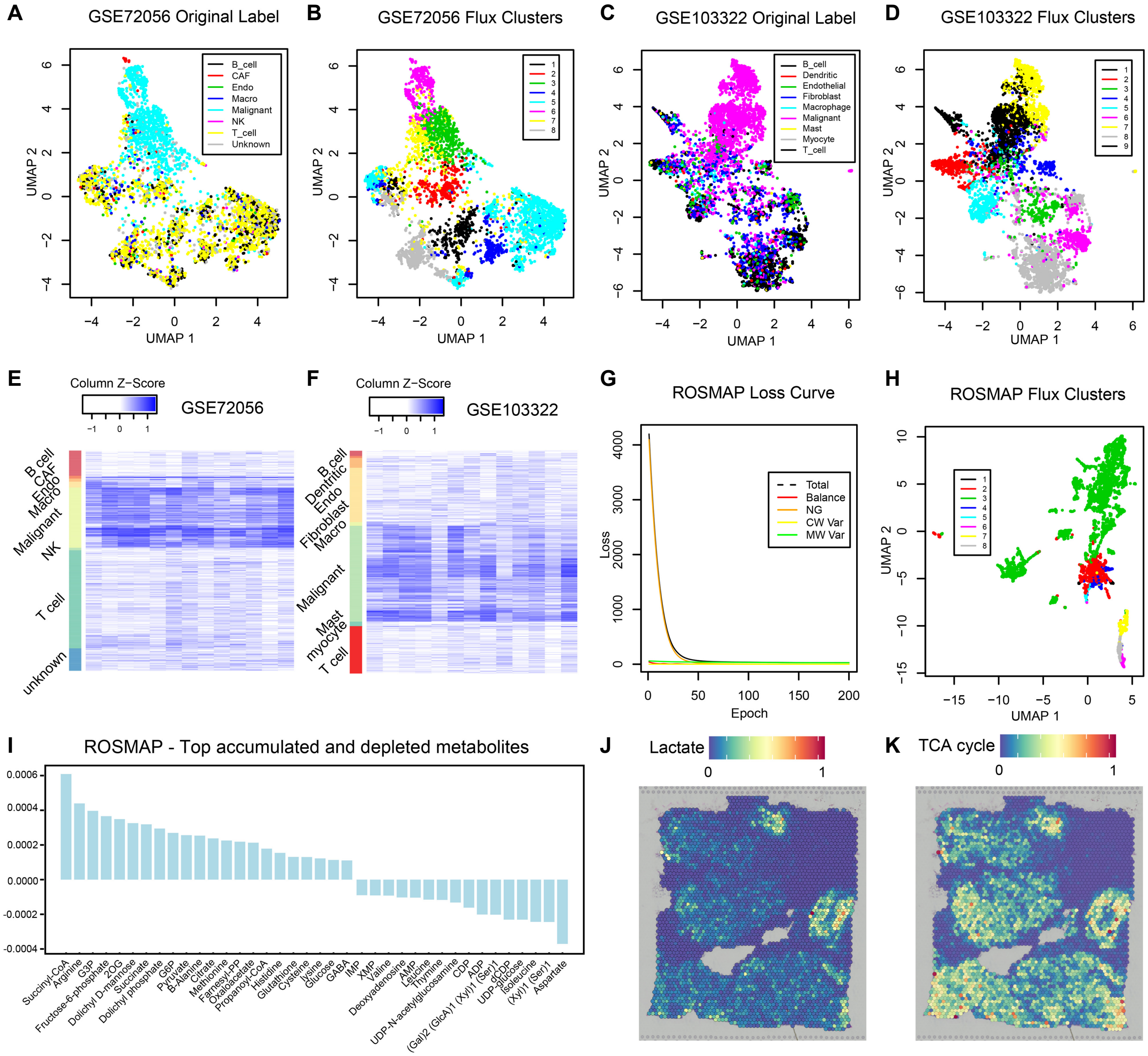

Figure 6. Application on two tumor scRNA-seq data sets, ROSMAP, and one breast cancer spatial transcriptomics data set. (*A*) UMAP-based clustering visualization using predicted metabolic fluxes of the GSE72056 melanoma data; the cell label was provided in the original work (Tirosh et al. 2016). (*B*) UMAP-based clustering visualization using predicted metabolic fluxes of the GSE72056; *k*-means clustering was used for cell clustering. (*C*) UMAP-based clustering visualization using predicted metabolic fluxes of the GSE103322 head and neck cancer data; the cell label was provided in the original work (Puram et al. 2017). (*D*) UMAP-based clustering visualization using predicted metabolic fluxes of the GSE103322; *k*-means clustering was used for cell clustering. (*E*) Distribution of predicted cell-wise flux of glycolytic and TCA cycle modules of GSE72056 melanoma data. Each row is one cell, and row side color bar represents eight cell types. Each column is one module. The *left* five columns are glycolytic modules from glucose to acetyl-CoA, the sixth column is the reaction from acetyl-CoA to Citrate, the seventh column is the reaction from pyruvate to lactate, and the *right*-most six columns (columns 8–13) are TCA cycle modules from citrate to oxaloacetic acid. (*F*) Distribution of predicted cell-wise flux of glycolytic and TCA cycle modules of GSE103322 head and neck cancer data. Each row is one cell, and row side color bar represents nine cell types, respectively. The columns are the same as *E*. (*G*) UMAP-based clustering visualization using predicted metabolic fluxes of the ROSMAP data; *k*-means clustering was used for cell clustering. (*H*) Convergency curve of the total loss and four loss terms during the training of neural networks on the ROSMAP data. (*I*) Top accumulated and depleted metabolites predicted in the AD neuron cells in the ROSMAP data. The *y*-axis is metabolism stress level (or level of accumulation and depletion), where a positive value represents accumulation and a negative value represents depletion. The *x*-axis are metabolites in a decreasing order of the accumulation level. (*J*) scFEA-predicted flux rate of lactate product on the spatial breast cancer data. The color of each point represents the spatial-wise predicted lactate product rate. The spatial plot is overlaid on the tissue slice image. (*K*) scFEA-predicted flux rate of TCA cycle (citrate to 2OG) on the spatial breast cancer data.

cancer types tend to have depleted glucose, G1P, and G6P. In addition, cancer cells tend to suffer from a high depletion level of acetyl-CoA. On the other hand, TCA cycle intermediates and amino acids tend to be accumulated in cancer cells. These observations are consistent with the findings derived from quantitative metabolomics data collected on solid cancer (Hirayama et al. 2009).

We noticed that the direction of imbalance for most intermediate metabolites seem to be the same throughout different cell types, although the imbalance level is much lower in stromal cells compared to cancer cells. A possible explanation is that these cells were collected in a small region of the same microenvironment, and the similar stresses, such as hypoxia and altered pH level, cause a similar impact on the metabolic landscape of cells of different types. The predicted cell type–specific fluxome and imbalance level of metabolites were given in Supplemental Table S4.

Application on droplet-based snRNA-seq data of Alzheimer's disease

We also applied scFEA on the ROSMAP single nuclei RNA-sequencing (snRNA-seq) data collected from cells in the central nervous systems of Alzheimer's disease (AD) patients and healthy donors (Mathys et al. 2019) by using hyperparameters $\alpha = 1$, $\beta = 1$, and $\gamma = 1$. Specifically, the ROSMAP snRNA-seq data was collected using the 10x Genomics Chromium droplet-based protocol. Comparing to the Smart-seq based scRNA-seq data, droplet-based data often have lower total expression signals and a higher dropout rate. scFEA has been successfully applied on this data set. Changes of the total loss over the running epochs suggested the total loss converge to a small value during the training of the scFEA model (Fig. 6G). Specifically, the flux imbalance loss forms the major loss term in the beginning of the training and quickly converges to a small value, suggesting a solution with good flux balance condition has been identified in this data set. Based on the scFEA-predicted flux, we identified that metabolic activity is higher in neuron cells than in other brain cell types. Cell clusters of different metabolic states were computed (Fig. 6H), in which a large cluster consisting of cells with more active metabolism has been identified (in green). We further studied the metabolic stress of this cell cluster, which is enriched by neuron cells from AD patients (Fig. 6I). We found that glucose, glycolytic and TCA cycle substrates, and glutathione are among the top accumulated metabolites. Suppressed glycolysis and dysfunctional TCA cycle that may lead to increased glucose and other intermediate metabolites, and elevated glutathione in response to reactive oxygen species, have been reported in AD (Atamna and Frey 2007; Mandal et al. 2019; Le Douce et al. 2020). On the other hand, molecules involved in DNA synthesis and valine/leucine/isoleucine metabolism are most depleted in the AD neuron cells, which are consistent to the recently reported observations of suppressed DNA synthesis and valine metabolism in AD (Yurov et al. 2011; Polis and Samson 2020). We predicted aspartate and metabolites involved in glycosaminoglycan synthesis are largely depleted in the AD neuron cells. Previous studies reported the association of these metabolites to AD (Doraiswamy 2003; Huynh et al. 2019), however their abundance change has been less studied. We anticipate that the cell-wise metabolic stress prediction enabled by scFEA could offer novel and systematic insight for biomarker prioritization.

Application on spatial transcriptomics data

As discussed above, distinct cell clusters of different metabolic states were identified in two cancer microenvironment data (GSE72056 and GSE103322) by using hyperparameters $\alpha = 1$, $\beta = 0$, and $\gamma = 1$. We speculate that the different metabolic states are caused by varied biochemical conditions, such as hypoxia or oxidative stress level, in the tumor microenvironment. To further validate this hypothesis and the method, we applied scFEA on a spatial transcriptomics data of human breast cancer collected from 10x Genomics visium protocol by using hyperparameters $\alpha = 1$, $\beta = 0$, and $\gamma = 1$. Clearly, cells that are spatially near each other should be exposed to similar biochemical stress conditions. We predicted spatial spot-specific metabolic flux by first applying scFEA on the spatial gene expression profile and then conducting associations of the predicted flux with spatial positions. scFEA identified two distinct spatial regions of high lactate production flux (Fig. 6J) and six spatial regions of high TCA cycle flux (Fig. 6K). Ratio of pyruvate → lactate flux and pyruvate → TCA cycle flux were computed, and the two high lactate production regions were predicted as of high hypoxia level, which were further validated by the high expression level of HIF1A-regulated genes in cells of these regions.

Discussion

Despite the ample knowledge we have gained on metabolic dysregulation for many disease types, there are still major gaps in our understanding of the integrated behavior and metabolic heterogeneity of cells in the context of tissue microenvironment. Essentially, the metabolic behavior can vary dramatically from cell to cell driven by the need to cope with various dynamic metabolic stress. A large amount of scRNA-seq has proven its potential in delivering information on a cell functioning state and its underlying phenotypic switches. Designing advanced computational tools to harness the power of large-scale scRNA-seq data for reliable prediction of cell-wise metabolic flux and states is crucial to bridge the technological gap of single-cell metabolomics. Knowledge derived therefrom will substantially benefit our understanding of the metabolic regulation intrinsic to diseased cells and on microenvironmental factors imposed upon the diseased cells and shed light on potential therapeutic vulnerabilities.

scFEA is developed to predict metabolic flux at single-cell resolution from scRNA-seq data, to construct a compendium of metabolic states for different cell types and tissue contexts, and understand their relevance to various disease phenotypes. To experimentally validate scFEA, we generated an scRNA-seq data of a patient-derived pancreatic cancer cells under four conditions of perturbed oxygen level and metabolic regulators and matched tissue-level metabolomics data and qRT-PCR profiles of key metabolic genes. We validated that the variations of metabolic flux predicted by scFEA are highly consistent with the observed metabolomic changes under different conditions. We also applied scFEA on in-drop or droplet-based scRNA-seq data and spatial transcriptomics data. Our analysis suggested that scFEA could robustly predict cell and cell group–specific metabolic shift for the data generated from different protocols. The fluxome estimated by scFEA enables a series of downstream analysis including identification of cell or tissue-level metabolic stress, sensitivity evaluation of enzymes to the whole metabolic flux, and inference of cell–tissue and cell–cell metabolic exchanges.

The scFEA model has the following advantages: (1) the model characterizes true biological flux by leveraging the metabolic network structure, and it is generally applicable as it requires only the input of scRNA-seq data; (2) the number of constraints, that is, the number of flux balance conditions multiplied by the single-cell number, is larger than the number of parameters, avoiding possible overfitting; and (3) the neural network–based flux estimation can well handle the nonlinear dependency between enzymatic gene expression and reaction rates. In the network reduction and reconstruction of scFEA, connected reactions were merged to form one metabolic module. The neural network model allows for a nonlinear dependency between gene expression and module flux. Theoretically, the flux rate could be determined by an "OR" operation of the high expression of any gene involved in the module. scFEA uses neighboring genes on the metabolic map to infer the metabolic flux of connected metabolic reactions, which increases robustness to dropout events and prediction accuracy. Our analysis suggested that scFEA is capable of identifying the interactive effect of multiple rate-limiting enzymes in one module.

scFEA seeks for a constrained optimization of flux balance, and each flux was modeled as a nonlinear function of the relevant enzymatic gene expression levels. The flux of each module is constrained by three additional loss terms, namely, (1) non-negativity, (2) consistency between predicted metabolic flux and gene expression level, and (3) the cell-wise total metabolic activity, TA_j. Although our current setting has been validated using matched scRNA-seq and metabolomics data, applications to publicly available cancer data suggested a similar trend metabolic "imbalance" for both cancer and stromal cells. Our analysis suggested that setting A_{mj} for each supermodule m in cell j may increase the flexibility of cell-specific metabolic imbalance, but at the price of possible overfitting. A more sensitive approach is to train a specific model for each predefined cell group. The biological rationale is that the neural network parameters contain the information of "kinetic parameters" that link gene expression with metabolic rate, which may differ among distant cell types, or cells under different conditions. Hence it is rational to assume cell type–specific parameters.

In this study, we did not provide a theoretical proof of the correctness of the scFEA model. Future efforts on generating high quality validating data or more comprehensive systems biological derivations could improve the understanding of the dependency between gene expression and metabolic state in individual cells. Compared with the existing FBA-based solutions, which tend to ignore kinetics and assume stringent flux balance condition, our new model treating flux balance as a loss function and seeking for modeling the nonlinear dependency between transcriptome and fluxome is more flexible, robust, and scientifically reasonable. Unlike other FBA-based approaches, scFEA does not require a prior knowledge of the imports and exports of the whole system. The flux distribution, including the influxes/outfluxes of the system, is estimated by minimizing the loss terms through a large number of cells. We consider such a consideration is more suitable for cell-wise flux estimation because the influxes and outfluxes are always cell and context specific and unknown. Although the flux balance in the scFEA model is robust to the stoichiometric coefficients, the predicted fluxome is represented by relative reaction rates scaled by total metabolic activity.

The neural network–based optimization framework of scFEA could enable a potential integration of metabolomics data, kinetic parameters, spatial information, or other prior knowledge of metabolic imbalance, to better characterize cell and tissue-level metabolic shifts of the target system. One future direction is to use metabolomics data, kinetic parameters, or other prior knowledge to better design the first layer of the neural network in modeling the flux of each module. Spatial information can be used to preselect a group of cells for training a spatially dependent model. Another future direction is to implement the current flux estimation analysis in spatial transcriptomics to infer (1) metabolic shifts specific to spatial patterns and (2) metabolic exchange between adjacent cells. This application to spatial transcriptomics data will be particularly interesting for cancer studies, to reveal how the metabolism and macromolecule biosynthesis in stromal cells such as cancer-associated fibroblasts, affect the adjacent cancer cells.

Overall, scFEA can efficiently delineate the sophisticated metabolic flux and imbalance specific to certain cell groups. We anticipate that it has the potential to decipher metabolic heterogeneity, tease out the metabolomic susceptibility to certain drugs, and ultimately warrant novel mechanistic and therapeutic insights of a diseased biological system at an unprecedented resolution.

Methods

Collection of the human metabolic map

We consider the human metabolic network as composed of different reaction types including metabolism, transport (including uptake and export), and biosynthesis. As detailed in Results, the reconstructed network consists of 22 supermodule classes of 169 modules. All reactions related to metabolism were collected from the Kyoto Encyclopedia of Genes and Genomes database (KEGG) (Kanehisa and Goto 2000). In total, 11 metabolism-related supermodules were manually summarized and are made up of glycolysis, the TCA cycle, pentose phosphate, fatty acids metabolism and synthesis; metabolism of amino acids, namely, serine, aspartate, beta-alanine, glutamate, leucine/valine/isoleucine; and the urea cycle, propionyl-CoA, and spermidine metabolism (Cao et al. 2017). The 11 metabolism supermodules contain 1388 reactions, 317 enzymes, which corresponds to 563 genes.

Transporters enable the trafficking of molecules in and out of cell membranes. We collected the human transporter proteins, their corresponding genes, and metabolite substrates from the Transporter Classification Database (Lin et al. 2015; Bhutia et al. 2016). In total, 80 transporter genes and 35 related metabolites were collected.

An essential part of the metabolic map are the biosynthesis pathways. The KEGG database and literature (Moffatt and Ashihara 2002; DeAngelis et al. 2013; Zhang et al. 2015a,b; Krasnova and Wong 2016; Zulueta et al. 2016; Lv et al. 2017; Sun et al. 2018, 2020; Gao and Edgar 2019) are the main information sources used for building biosynthesis modules. We collected 69 biosynthesis modules forming 10 supermodules, namely, biosynthesis of hyaluronic acid, glycogen, glycosaminoglycan, N-linked glycan, O-linked glycan, sialic acid, glycan, purine, pyrimidine, and steroid hormones. Overall, the biosynthesis modules include 459 genes of 269 enzymes catalyzing 869 reactions.

More details of the collection of the human metabolic map and the statistics of the mouse metabolic map are provided in Supplemental Methods.

Selecting genes of significant expression

We applied our in-house method, Left Truncated Mixture Gaussian model (LTMG), to determine the expression status of each gene in each single cell. LTMG considers the multimodality of the expression profile of each gene throughout all the single cells, by assuming that the gene's expression follows a mixture of suppressed state and activated states as represented by the following likelihood function (Wan et al. 2019):

$$\prod_{j=1}^{N}\left(\sum_{i=1}^{S} a_i p_i(x_j|u_i,\ \sigma_i) + a_{S+1} p_{S+1}(x_j|u_{S+1},\ \sigma_{S+1})\right),$$

where x_j, $j=1\ldots N$ is the expression profile of gene x in N cells; the index $1\ldots S$ are the S active expression states and $S+1$ is the suppressed expression state; a_i is the proportion of each state with $a_1+\cdots+a_{S+1}=1$; $a_{1\ldots S}>0$ and $a_{S+1}\geq 0$; and p_i, u_i, and σ_i are the pdf, mean, and standard deviation of each expression state. Specifically, LTMG considers the distribution of each mixing component, p_i, as a left-truncated Gaussian distribution to account for the noise of dropout events. In this work, LTMG was used to fit to each gene's expression and a gene was determined to have significant expression if $\sum_{i=1}^{S} a_i \geq 0.1$; that is, the gene has active expression states in at least 10% cells.

Prefiltering of active modules based on gene expression

Each metabolic module contains an input, an output, and a number of enzymes catalyzing the reactions. A reaction is considered as disconnected if none of the genes encoding its catalyzing enzymes is significantly expressed. A metabolic module is considered as blocked if there is no connected path from the input to the output. Considering the common dropout events in scRNA-seq data, especially for the drop-seq data, we adopted a conservative approach to pretrim the metabolic modules: essentially, a module will be removed from further analysis if none of the genes involved in all reactions of this module has significantly active expressions.

scFEA model setup and a belief propagation–based solution of the flux model

Model setup

We developed a novel optimization strategy to minimize L similar to the idea of belief propagation (Yedidia et al. 2001). Specifically, the flux balance of each metabolite C_k, $L_K \triangleq \sum_{j=1}^{N} \left(\sum_{m \in F_{in}^{C_k}} \text{Flux}_{m,j} - \sum_{m' \in F_{\text{out}}^{C_k}} \text{Flux}_{m',j} \right)^2$, will be iteratively optimized, by taking into account all the Hop-2 neighbors in the factor graph (metabolites), denoted as $Ne(C_k)$, and Hop-4 neighbors (metabolites), that is, $Ne^2(C_k) := \{C_{k'} | C_{k'} \in Ne(Ne(C_k)) \backslash C_k \}$. Specifically, for a more efficient optimization, we adopt the idea of belief propagation by minimizing a reweighted flux imbalance: $L_K^* \triangleq L_K + \sum_{C_{k'} \in Ne^2(C_k)} W_{k'} L_{k'}$ at each iteration, where $W_{k'}$ is a weight value in (0, 1] representing the reliability of the current flux balance of $C_{k'}$. We set $W_{k'} = \exp\left(-\sum_{C_{k''} \in Ne(Ne(C_{k'})) \backslash \{C_{k'}, C_k\}} L_{k''} / |Ne^2(C_{k'}) \backslash \{C_{k'}, C_k\}| \right)$ as an exponential function of the negative averaged imbalance level of 2-hop neighbors (metabolite) of $C_{k'}$ excluding C_k, with higher $W_{k'}$ denoting lower imbalance level of the metabolites. The intuition is that the more reliable the current flux is estimated for the modules involving $C_{k'}$, which is reflected by the averaged imbalance level of its 2-hop neighbors, a higher weight $W_{k'}$ should be given to $C_{k'}$. Therefore, that when minimizing L_K, a disruption of the flux balance of $C_{k'}$ of higher weight will be more heavily penalized and less desirable.

Neural network model setup

For each module, a neural network is used to represent the nonlinear dependency between gene expressions and reaction rates. Each neural network has a_1 hidden layers, each with a_2 hidden nodes, and one output node. In this study, we took $a_1 = 3$ and $a_2 = 8$. A Hyperbolic Tangent activation function, $\text{Tanhshrink}(x) = x - \tanh(x)$, is used. The number of nodes and the number of hidden layers determines the complexity of network structure, which impacts the convergence time of optimization. A too simple structure may not fully capture the nonlinear relationship, but a too complex structure may make it difficult to train all parameters and reach convergence. Our organized metabolic modules have an average gene number of eight, which determines the input nodes of scFEA. Because scFEA has 169 parallel subnetworks for each metabolic module, we decide that three hidden layers can leverage the level of nonlinearity and overfitting and ensure a feasible computational cost (for details, see Supplemental Methods).

Clustering analysis of cells with distinct metabolic states

scFEA adopts an attributed graph clustering approach to identify the groups of cells and metabolic modules forming a distinct metabolic state. Two clustering approaches were provided to the results of scFEA for different purposes, namely, clustering of (1) metabolic modules based on the metabolic map and the predicted flux, and (2) cells sharing a common state on the overall metabolic map based on the predicted flux.

Clustering of metabolic modules

Denote the adjacency matrix of the context-specific metabolic map as $A^{M \times M}$ and predicted metabolic flux as $\text{Flux}^{M \times N}$, where $\text{Flux}_{m,j}$ represents the predicted flux rate of the module m in cell j, and a two-stage spectral clustering was applied to cluster the metabolic modules based on $A^{M \times M}$ and $\text{Flux}^{M \times N}$. Specifically, denote $A^{F,\, M \times M}$ as the Euclidean distance of the M modules calculated using $\text{Flux}^{M \times N}$, and $D^{M \times M}$ and $D^{F,\, M \times M}$ as two diagonal matrices, in which $D_{ii} = \sum_{j=1}^{M} A_{ij}$ and $D_{ii}^{F} = \sum_{j=1}^{M} A_{ij}^{F}$. The normalized graph Laplacian matrices for the network topology and attributes similarity were defined as $L = I - D^{-1/2} A D^{-1/2}$ and $L^F = I - D^{F-1/2} A^F D^{F-1/2}$. The normalized graph Laplacian matrices scale the topological similarity and attributes similarly into the same scale. Denote $d(R_i, R_j)$ and $d^F(R_i, R_j)$ as the Euclidean distance between the metabolic modules R_i and R_j calculated using the smallest P_1 eigenvectors of L and the smallest P_2 eigenvectors of L^F, respectively, the modules were clusters by the k-means method with the following distance measure:

$$\alpha d(R_i, R_j) + (1 - a) d^F(R_i, R_j),$$

where α, P_1, and P_2, and the number of clusters are hyperparameters. Our empirical analysis suggested a default setting as $\alpha = 0.3$, which assigns a higher weight to the similarity of the predict flux; $P_1 = \max\{3, \text{floor}(\# SM/2)\}$, where $\# SM$ is the number of supermodules in the current metabolic map; and $P_2 = \max\{3, \text{floor}(\# M/17)\}$, where $\# M$ is the number of nonzero modules in the current metabolic map. The number of clusters should be predetermined by users and depends on the number of cells, cell types, and metabolic modules.

Analysis of cell group–specific metabolic stress and metabolic exchanges among cell groups

The cell-wise metabolic flux estimated by scFEA enables the analysis of metabolic stress. For a predefined cell group such as cells of the same type, the total imbalance of each compound will be computed and ranked. A one-way t-test was applied to test if the imbalance is significantly different from 0. The metabolic exchange among different cell groups from one tissue sample was identified as the metabolites with different signs of metabolic imbalance in different cell groups, such as accumulation and depletion, or exporting or importing. Tissue-level metabolic stress is computed as the total imbalance throughout multiple cells.

Perturbation analysis

In scFEA, to evaluate the impact of the change in gene expression on the whole metabolic map, a perturbation analysis is conducted that includes three components: (1) the direct impact of each gene G_i^m to the flux module m can be directly computed by its derivative df_{nn}^m / dG_i^m for all the modules containing G_i^m; (2) the impact of the flux change of one module A on a target module B could be estimated as the variations of flux in B calculated under different values of flux in A, while keeping the other parameters/input fixed, that is, a Monte Carlo–based method; (3) the impact of each gene's expression to the flux of distant modules can be evaluated by integrating (1) and (2) using a chain rule, that is, by first computing the flux change of the modules containing the gene and then evaluating the change of other modules.

Patient-derived cell line models of pancreatic cancer

Pa03C cells were obtained from Dr. Anirban Maitra's laboratory at The Johns Hopkins University (Jones et al. 2008). All cells were maintained at 37°C in 5% CO_2 and grown in DMEM (Invitrogen) with 10% serum (Hyclone). Cell line identity was confirmed by DNA fingerprint analysis (IDEXX BioResearch) for species and baseline short tandem repeat analysis testing in February 2017. All cell lines were 100% human, and a nine-marker short tandem repeat analysis is on file. They were also confirmed to be mycoplasma free.

scRNA-seq experiment

Cells were transfected with either Scrambled (SCR) (5′-CCAUGAGGUCAGCAUGGUCUG-3′, 5′-GACCAUGCUGACCUCAUGGAA-3′) or siAPEX1 (5′-GUCUGGUACGACUGGAGUACC-3′, 5′-UACUCCAGUCGUACCAGACCU-3′ siRNA). Briefly, 1×10^5 cells are plated per well of a six-well plate and allowed to attach overnight. The next day, Lipofectamine RNAiMAX reagent (Invitrogen) was used to transfect in the *APEX1* and SCR siRNA at 20 nM following the manufacturer's indicated protocol. Opti-MEM, siRNA, and Lipofectamine was left on the cells for 16 h and then regular DMEM media with 10% serum was added.

Three days posttransfection, SCR/siAPEX1 cells were collected and loaded into 96-well microfluidic C1 Fluidigm array (Fluidigm). All chambers were visually assessed, and any chamber containing dead or multiple cells was excluded. The SMARTer system (Clontech) was used to generate cDNA from captured single cells. The dscDNA quantity and quality was assessed using an Agilent Bioanalyzer (Agilent Technologies) with the High Sensitivity DNA Chip. The Purdue Genomics Facility prepared libraries using a Nextera kit (Illumina). Unstrained 2 × 100 bp reads were sequenced using the HiSeq 2500 on rapid run mode in one lane.

scRNA-seq data processing and analysis

FastQC was applied to evaluate the quality of the single-cell RNA sequencing data. Counts were called for each cell sample by using STAR alignment pipeline (Dobin et al. 2013) against human GRCh38 reference genome. Cells with fewer 250 or more than 10,000 non-zero expressed genes were excluded from the analysis. Cells with >15% counts mapped to the mitochondrial genome were excluded as low-quality cells, resulting in 40 *APEX1* KD and 48 Control cells under hypoxia condition and 27 *APEX1* KD and 46 Control cells under normoxia condition for further analysis.

We used our in-house left truncated mixture Gaussian model to identify differentially expressed genes (Wan et al. 2019). Pathway enrichment analysis of the genes in the identified biclusters are computed using a hypergeometric test against the 1329 canonical pathways in the MSigDB database (Liberzon et al. 2011), with $P < 0.001$ as a significance cutoff.

Metabolomic profiling and data analysis

To address the function of the mitochondria, S-1 MitoPlates (Biolog) Mitochondrial Function Assays were performed following the manufacturer's protocol. The assay covers 14 metabolites in central metabolic pathways, namely, glucose, glucose-1 phosphate, glucose-6 phosphate, pyruvate, and lactate in the glycolysis pathway; citrate, 2-oxoglutarate, succinate, fumarate, malate in the TCA cycle; and amino acids glutamate, glutamine, serine, and ornithine. Specifically, assay mix (60 min at 37°C) was added to the plates to dissolve the substrates. We collected, counted, resuspended PDAC cells in provided buffer and plated them at 5 × 104 cells/well after treatment. Readings at 590 nm were taken every 5 min for 4 h at 37°C. Experiments were performed in triplicate with three biological replicates for the siAPEX1 and control PDAC cells under the normoxia condition. Raw data was analyzed using GraphPad Prism 8, and statistical significance was determined using the two-way ANOVA, and *P*-values < 0.05 were considered statistically significant.

qRT-PCR

qRT-PCR was used to measure the mRNA expression levels of the various genes identified from the scRNA-seq analysis. Following transfection, total RNA was extracted from cells using the Qiagen RNeasy Mini kit (Qiagen) according to the manufacturer's instructions. First-strand cDNA was obtained from RNA using random hexamers and MultiScribe reverse transcriptase (Applied Biosystems). Quantitative PCR was performed using SYBR Green Real Time PCR master mix (Applied Biosystems) in a CFX96 Real Time detection system (Bio-Rad). The relative quantitative mRNA level was determined using the comparative Ct method using ribosomal protein L6 (*RPL6*) as the reference gene. Experiments were performed in triplicate for each sample. Statistical analysis performed using the 2−ΔΔCT method and analysis of covariance (ANCOVA) models, as previously published (Fishel et al. 2015).

Data access

The raw and processed scRNA-seq data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE99305 (normoxia) and GSE173433 (hypoxia). The scFEA package, full set of process scRNA-seq data, metabolomic profile, and analysis codes used in this work are available at GitHub (https://github.com/changwn/scFEA) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the National Science Foundation (NSF) DBI IIBR 2047631, the National Institutes of Health (National Institute of General Medical Sciences), NSF IIS 1850360, National Institutes of Health (NIH) 5U54AG065181, NIH 1R01GM131399, Showalter Young Investigator Award, and Precision Health Initiative of Indiana University. C.Z. and S.C. thank Dr. Ying Xu from the University of Georgia, Dr. Yunlong Liu, Dr. Kun Huang, Dr. Ji Zhang, and Dr. Xiongbin Lu from Indiana University for their constructive suggestions and advice to this work.

References

Ahl PJ, Hopkins RA, Xiang WW, Au B, Kaliaperumal N, Fairhurst AM, Connolly JE. 2020. Met-Flow, a strategy for single-cell metabolic analysis highlights dynamic changes in immune subpopulations. *Commun Biol* **3:** 305. doi:10.1038/s42003-020-1027-9

Ali A, Abouleila Y, Shimizu Y, Hiyama E, Emara S, Mashaghi A, Hankemeier T. 2019. Single-cell metabolomics by mass spectrometry: advances, challenges, and future applications. *Trends Analyt Chem* **120:** 115436. doi:10.1016/j.trac.2019.02.033

Atamna H, Frey WH II. 2007. Mechanisms of mitochondrial dysfunction and energy deficiency in Alzheimer's disease. *Mitochondrion* **7:** 297–310. doi:10.1016/j.mito.2007.06.001

Bhutia YD, Babu E, Ramachandran S, Yang S, Thangaraju M, Ganapathy V. 2016. SLC transporters as a novel class of tumour suppressors: identity, function and molecular mechanisms. *Biochem J* **473:** 1113–1124. doi:10.1042/BJ20150751

Bishop AL, Rab FA, Sumner ER, Avery SV. 2007. Phenotypic heterogeneity can enhance rare-cell survival in 'stress-sensitive' yeast populations. *Mol Microbiol* **63:** 507–520. doi:10.1111/j.1365-2958.2006.05504.x

Cao S, Zhu X, Zhang C, Qian H, Schuttler HB, Gong J, Xu Y. 2017. Competition between DNA methylation, nucleotide synthesis, and antioxidation in cancer versus normal tissues. *Cancer Res* **77:** 4185–4195. doi:10.1158/0008-5472.CAN-17-0262

Chen YP, Yin JH, Li WF, Li HJ, Chen DP, Zhang CJ, Lv JW, Wang YQ, Li XM, Li JY, et al. 2020. Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma. *Cell Res* **30:** 1024–1042. doi:10.1038/s41422-020-0374-x

Damiani C, Maspero D, Di Filippo M, Colombo R, Pescini D, Graudenzi A, Westerhoff HV, Alberghina L, Vanoni M, Mauri G. 2019. Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *PLoS Comput Biol* **15:** e1006733. doi:10.1371/journal.pcbi.1006733

DeAngelis PL, Liu J, Linhardt RJ. 2013. Chemoenzymatic synthesis of glycosaminoglycans: re-creating, re-modeling and re-designing nature's longest or most complex carbohydrate chains. *Glycobiology* **23:** 764–777. doi:10.1093/glycob/cwt016

DeBerardinis RJ, Lum JJ, Hatzivassiliou G, Thompson CB. 2008. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab* **7:** 11–20. doi:10.1016/j.cmet.2007.10.002

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Doraiswamy PM. 2003. The role of the *N*-methyl-D-aspartate receptor in Alzheimer's disease: therapeutic potential. *Curr Neurol Neurosci Rep* **3:** 373–378. doi:10.1007/s11910-003-0019-8

Duncan KD, Fyrestam J, Lanekoff I. 2019. Advances in mass spectrometry based single-cell metabolomics. *Analyst* **144:** 782–793. doi:10.1039/C8AN01581C

Dunn L, Allen GF, Mamais A, Ling H, Li A, Duberley KE, Hargreaves IP, Pope S, Holton JL, Lees A, et al. 2014. Dysregulation of glucose metabolism is an early event in sporadic Parkinson's disease. *Neurobiol Aging* **35:** 1111–1115. doi:10.1016/j.neurobiolaging.2013.11.001

Emara S, Amer S, Ali A, Abouleila Y, Oga A, Masujima T. 2017. Single-cell metabolomics. In *Metabolomics: from fundamentals to clinical applications* (ed. Sussulini A), pp. 323–343. Springer, Cham, Switzerland.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

Evers TMJ, Hochane M, Tans SJ, Heeren RMA, Semrau S, Nemes P, Mashaghi A. 2019. Deciphering metabolic heterogeneity by single-cell analysis. *Anal Chem* **91:** 13314–13323. doi:10.1021/acs.analchem.9b02410

Feinberg AP. 2007. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447:** 433–440. doi:10.1038/nature05919

Fessenden M. 2016. Metabolomics: small molecules, single cells. *Nature* **540:** 153–155. doi:10.1038/540153a

Fishel ML, Wu X, Devlin CM, Logsdon DP, Jiang Y, Luo M, He Y, Yu Z, Tong Y, Lipking K. 2015. Apurinic/apyrimidinic endonuclease/redox factor-1 (APE1/Ref-1) redox function negatively regulates NRF2. *J Biol Chem* **290:** 3057–3068. doi:10.1074/jbc.M114.621995

Gao C, Edgar KJ. 2019. Efficient synthesis of glycosaminoglycan analogs. *Biomacromolecules* **20:** 608–617. doi:10.1021/acs.biomac.8b01150

Gu W, Nowak WN, Xie Y, Le Bras A, Hu Y, Deng J, Issa Bhaloo S, Lu Y, Yuan H, Fidanis E, et al. 2019. Single-cell RNA-sequencing and metabolomics analyses reveal the contribution of perivascular adipose tissue stem cells to vascular remodeling. *Arterioscler Thromb Vasc Biol* **39:** 2049–2066. doi:10.1161/ATVBAHA.119.312732

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144:** 646–674. doi:10.1016/j.cell.2011.02.013

Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong CB, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao YJ, et al. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* **28:** 1097–1105. doi:10.1038/nbt.1682

Heintzman ND, Stuart RK, Hon G, Fu YT, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu CX, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318. doi:10.1038/ng1966

Hirayama A, Kami K, Sugimoto M, Sugawara M, Toki N, Onozuka H, Kinoshita T, Saito N, Ochiai A, Tomita M, et al. 2009. Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. *Cancer Res* **69:** 4918–4925. doi:10.1158/0008-5472.CAN-08-4806

Hirschey MD, DeBerardinis RJ, Diehl AME, Drew JE, Frezza C, Green MF, Jones LW, Ko YH, Le A, Lea MA, et al. 2015. Dysregulated metabolism contributes to oncogenesis. In *Seminars in cancer biology* (ed. Bishayee A, Block K), Vol. 35, pp. S129–S150. Elsevier. doi:10.1016/j.semcancer.2015.10.002

Honkoop H, de Bakker DE, Aharonov A, Kruse F, Shakked A, Nguyen PD, de Heus C, Garric L, Muraro MJ, Shoffner A, et al. 2019. Single-cell analysis uncovers that metabolic reprogramming by ErbB2 signaling is essential for cardiomyocyte proliferation in the regenerating heart. *eLife* **8:** e50163. doi:10.7554/eLife.50163

Huynh MB, Ouidja MO, Chantepie S, Carpentier G, Maïza A, Zhang G, Vilares J, Raisman-Vozari R, Papy-Garcia D. 2019. Glycosaminoglycans from Alzheimer's disease hippocampus have altered capacities to bind and regulate growth factors activities and to bind tau. *PLoS One* **14:** e0209573. doi:10.1371/journal.pone.0209573

Jaenisch R, Bird A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33:** 245–254. doi:10.1038/ng1089

Jones S, Zhang X, Parsons DW, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno AJs. 2008. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321:** 1801–1806. doi:10.1126/science.1164368

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28:** 27–30. doi:10.1093/nar/28.1.27

Kelley MR, Georgiadis MM, Fishel ML. 2012. APE1/Ref-1 role in redox signaling: translational applications of targeting the redox function of the DNA repair/redox protein APE1/Ref-1. *Curr Mol Pharmacol* **5:** 36–53. doi:10.2174/1874467211205010036

Kim J, DeBerardinis RJ. 2019. Mechanisms and implications of metabolic heterogeneity in cancer. *Cell Metab* **30:** 434–446. doi:10.1016/j.cmet.2019.08.013

Kingma DP, Ba J. 2015. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, Conference Track Proceedings.

Kochanek KD, Murphy SL, Xu J, Arias E. 2019. Deaths: final data for 2017. *Natl Vital Stat Rep* **68:** 1–77.

Krasnova L, Wong CH. 2016. Understanding the chemistry and biology of glycosylation with glycan synthesis. *Annu Rev Biochem* **85:** 599–630. doi:10.1146/annurev-biochem-060614-034420

Lan X, Roth S, Huttenlocher D, Black MJ. 2006. Efficient belief propagation with learned higher-order Markov random fields. In *European conference on computer vision*, pp. 269–282. Springer, Berlin.

Le Douce J, Maugard M, Veran J, Matos M, Jégo P, Vigneron PA, Faivre E, Toussay X, Vandenberghe M, Balbastre Y, et al. 2020. Impairment of glycolysis-derived L-serine production in astrocytes contributes to cognitive deficits in Alzheimer's disease. *Cell Metab* **31:** 503–517.e8. doi:10.1016/j.cmet.2020.02.004

Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, Mendes P, Swainston N. 2012. Improving metabolic flux predictions using absolute gene expression data. *BMC Syst Biol* **6:** 73. doi:10.1186/1752-0509-6-73

Levine LS, Hiam-Galvez KJ, Marquez DM, Tenvooren I, Madden MZ, Contreras DC, Dahunsi DO, Irish JM, Oluwole OO, Rathmell JC, et al. 2021. Single-cell metabolic dynamics of early activated CD8 T cells during the primary immune response to infection. *Immunity* **54:** 829–844.e5. doi:10.1101/2020.01.21.911545

Li X, Egervari G, Wang Y, Berger SL, Lu Z. 2018. Regulation of chromatin and gene expression by metabolic enzymes and metabolites. *Nat Rev Mol Cell Biol* **19:** 563–578. doi:10.1038/s41580-018-0029-7

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JPJB. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27:** 1739–1740. doi:10.1093/bioinformatics/btr260

Lidstrom ME, Konopka MC. 2010. The role of physiological heterogeneity in microbial population behavior. *Nat Chem Biol* **6:** 705–712. doi:10.1038/nchembio.436

Lin L, Yee SW, Kim RB, Giacomini KM. 2015. SLC transporters as therapeutic targets: emerging opportunities. *Nat Rev Drug Discov* **14:** 543–560. doi:10.1038/nrd4626

Liu Y, Beyer A, Aebersold R. 2016. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165:** 535–550. doi:10.1016/j.cell.2016.03.014

Lv X, Cao H, Lin B, Wang W, Zhang W, Duan Q, Tao Y, Liu XW, Li X. 2017. Synthesis of sialic acids, their derivatives, and analogs by using a whole-cell catalyst. *Chemistry (Easton)* **23:** 15143–15149. doi:10.1002/chem.201703083

Mandal PK, Shukla D, Tripathi M, Ersland L. 2019. Cognitive improvement with glutathione supplement in Alzheimer's disease: a way forward. *J Alzheimers Dis* **68:** 531–535. doi:10.3233/JAD-181054

Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, Menon M, He L, Abdurrob F, Jiang X, et al. 2019. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570:** 332–337. doi:10.1038/s41586-019-1195-2

Matsuzawa Y. 2006. Therapy insight: adipocytokines in metabolic syndrome and related cardiovascular disease. *Nat Clin Pract Cardiovasc Med* **3:** 35–42. doi:10.1038/ncpcardio0380

Mattson MP, Chan SL. 2001. Dysregulation of cellular calcium homeostasis in Alzheimer's disease: bad genes and bad habits. *J Mol Neurosci* **17:** 205–224. doi:10.1385/JMN:17:2:205

McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* **3:** 861. doi:10.21105/joss.00861

Mehrmohamadi M, Liu X, Shestov AA, Locasale JW. 2014. Characterization of the usage of the serine metabolic network in human cancer. *Cell Rep* **9:** 1507–1519. doi:10.1016/j.celrep.2014.10.026

Moffatt BA, Ashihara H. 2002. Purine and pyrimidine nucleotide synthesis and metabolism. *Arabidopsis Book* **1:** e0018. doi:10.1199/tab.0018

Polis B, Samson AO. 2020. Role of the metabolism of branched-chain amino acids in the development of Alzheimer's disease and other metabolic disorders. *Neural Regen Res* **15:** 1460. doi:10.4103/1673-5374.274328

Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, et al. 2017. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171:** 1611–1624.e24. doi:10.1016/j.cell.2017.10.044

Rask E, Olsson T, Soderberg S, Andrew R, Livingstone DE, Johnson O, Walker BR. 2001. Tissue-specific dysregulation of cortisol metabolism in human obesity. *J Clin Endocrinol Metab* **86:** 1418–1421. doi:10.1210/jcem.86.3.7453

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518:** 317–330. doi:10.1038/nature14248

Robertson-Tessi M, Gillies RJ, Gatenby RA, Anderson AR. 2015. Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. *Cancer Res* **75:** 1567–1579. doi:10.1158/0008-5472.CAN-14-1428

Rohlenova K, Goveia J, García-Caballero M, Subramanian A, Kalucka J, Treps L, Falkenberg KD, de Rooij LP, Zheng Y, Lin L, et al. 2020. Single-cell RNA sequencing maps endothelial metabolic plasticity in pathological angiogenesis. *Cell Metab* **31:** 862–877.e14. doi:10.1016/j.cmet.2020.03.009

Saier MH Jr, Tran CV, Barabote RD. 2006. TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res* **34:** D181–D186. doi:10.1093/nar/gkj001

Saurty-Seerunghen MS, Bellenger L, El-Habr EA, Delaunay V, Garnier D, Chneiweiss H, Antoniewski C, Morvan-Dubois G, Junier MP. 2019. Capture at the single cell level of metabolic modules distinguishing aggressive and indolent glioblastoma cells. *Acta Neuropathol Commun* **7:** 155. doi:10.1186/s40478-019-0819-y

Schnell S. 2014. Validity of the Michaelis–Menten equation—steady-state or reactant stationary assumption: that is the question. *FEBS J* **281:** 464–472. doi:10.1111/febs.12564

Shah F, Goossens E, Atallah NM, Grimard M, Kelley MR, Fishel M. 2017. APE1/Ref-1 knockdown in pancreatic ductal adenocarcinoma—characterizing gene expression changes and identifying novel pathways using single-cell RNA sequencing. *Mol Oncol* **11:** 1711–1732. doi:10.1002/1878-0261.12138

Sun H, Zhang C, Cao S, Sheng T, Dong N, Xu Y. 2018. Fenton reactions drive nucleotide and ATP syntheses in cancer. *J Mol Cell Biol* **10:** 448–459. doi:10.1093/jmcb/mjy039

Sun H, Zhou Y, Skaro MF, Wu Y, Qu Z, Mao F, Zhao S, Xu Y. 2020. Metabolic reprogramming in cancer is induced to increase proton production. *Cancer Res* **80:** 1143–1155. doi:10.1158/0008-5472.CAN-19-3392

Thompson C, Bauer D, Lum J, Hatzivassiliou G, Zong WX, Zhao F, Ditsworth D, Buzzai M, Lindsten T. 2005. How do cancer cells acquire the fuel needed to support cell growth? *Cold Spring Harb Symp Quant Biol* **70:** 357–362. doi:10.1101/sqb.2005.70.011

Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352:** 189–196. doi:10.1126/science.aad0501

van der Knaap JA, Verrijzer CP. 2016. Undercover: gene control by metabolites and metabolic enzymes. *Genes Dev* **30:** 2345–2369. doi:10.1101/gad.289140.116

Vasdekis AE, Stephanopoulos G. 2015. Review of methods to probe single cell metabolism and bioenergetics. *Metab Eng* **27:** 115–135. doi:10.1016/j.ymben.2014.09.007

Wagner A, Wang C, Fessler J, DeTomaso D, Avila-Pacheco J, Kaminski J, Zaghouani S, Christian E, Thakore P, Schellhaass B, et al. 2021. Metabolic modeling of single Th17 cells reveals regulators of autoimmunity. *Cell* **184:** 4168–4185.e21. doi:10.1016/j.cell.2021.05.045

Wan C, Chang W, Zhang Y, Shah F, Lu X, Zang Y, Zhang A, Cao S, Fishel ML, Ma Q, et al. 2019. LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. *Nucleic Acids Res* **47:** e111. doi:10.1093/nar/gkz655

Ward PS, Thompson CB. 2012. Metabolic reprogramming: a cancer hallmark even Warburg did not anticipate. *Cancer Cell* **21:** 297–308. doi:10.1016/j.ccr.2012.02.014

Xiao Z, Dai Z, Locasale JW. 2019. Metabolic landscape of the tumor microenvironment at single cell resolution. *Nat Commun* **10:** 3763. doi:10.1038/s41467-019-11738-0

Xiao Z, Locasale JW, Dai Z. 2020. Metabolism in the tumor microenvironment: insights from single-cell analysis. *Oncoimmunology* **9:** 1726556. doi:10.1080/2162402X.2020.1726556

Xu K, Mao X, Mehta M, Cui J, Zhang C, Xu Y. 2012. A comparative study of gene-expression data of basal cell carcinoma and melanoma reveals new insights about the two cancers. *PLoS One* **7:** e30750. doi:10.1371/journal.pone.0030750

Yedidia JS, Freeman WT, Weiss Y. 2001. Generalized belief propagation. In *Advances in neural information processing systems 13 (NIPS 2000)* (ed. Leen T, et al.), pp. 689–695.

Yurov YB, Vorsanova SG, Iourov IY. 2011. The DNA replication stress hypothesis of Alzheimer's disease. *ScientificWorldJournal* **11:** 2602–2612. doi:10.1100/2011/625690

Zampieri M, Sekar K, Zamboni N, Sauer U. 2017. Frontiers of high-throughput metabolomics. *Curr Opin Chem Biol* **36:** 15–23. doi:10.1016/j.cbpa.2016.12.006

Zenobi R. 2013. Single-cell metabolomics: analytical and biological perspectives. *Science* **342:** 1243259. doi:10.1126/science.1243259

Zhang C, Cao S, Toole BP, Xu Y. 2015a. Cancer may be a pathway to cell survival under persistent hypoxia and elevated ROS: a model for solid-cancer initiation and early development. *Int J Cancer* **136:** 2001–2011. doi:10.1002/ijc.28975

Zhang C, Liu C, Cao S, Xu Y. 2015b. Elucidation of drivers of high-level production of lactates throughout a cancer development. *J Mol Cell Biol* **7:** 267–279. doi:10.1093/jmcb/mjv031

Zhang Y, Kim MS, Nguyen E, Taylor DM. 2020. Modeling metabolic variation with single-cell expression data. bioRxiv doi:10.1101/2020.01.28.923680

Zhao M, Chen X, Gao G, Tao L, Wei L. 2009. RLEdb: a database of rate-limiting enzymes and their regulation in human, rat, mouse, yeast and *E. coli*. *Cell Res* **19:** 793–795. doi:10.1038/cr.2009.61

Zulueta MM, Lin SY, Hu YP, Hung SC. 2016. Synthesis of glycosaminoglycans. In *Glycochemical synthesis: strategies and applications* (ed. Hung SC, Zulueta MM), Chapter 10, pp. 235–261. Wiley, Hoboken, NJ. doi:10.1002/9781119006435.ch10

Received September 3, 2020; accepted in revised form July 1, 2021.

A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data

Norah Alghamdi, Wennan Chang, Pengtao Dang, et al.

Genome Res. 2021 31: 1867-1884 originally published online July 22, 2021
Access the most recent version at doi:10.1101/gr.271205.120