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

Norah Alghamdi[1+], Wennan Chang[1,2+], 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]*, Chi Zhang[1,2]*

[1]Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, [3]Department of Pediatrics, [5]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA.

[2]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN 46202, USA

[4]Department of Biomedical Informatics, Ohio State University, Columbus, OH 43210, USA

*To whom correspondence should be addressed. +1 317-278-9625; Email: czhang87@iu.edu. Correspondence is also addressed to Melissa Fishel, Email: mfishel@iu.edu, and Sha Cao, Email: shacao@iu.edu.

[+]These authors have an equal contribution to this work.

SUPPLEMENTARY METHODS

Collection and reorganization of human metabolic map

We reorganized the human metabolic network into different reaction types including metabolism, transporter, and biosynthesis. The reorganized network includes 21 super module classes of 175 modules. For the metabolism part, all reactions were collected from Kyoto Encyclopedia of Genes and Genomes database (KEGG) (61). The first super module includes 121 Glucose and TCA cycle reactions. The glycolysis pathway has major out-branches including polysaccharides synthesis, pentose phosphate, serine metabolism, lactate production and acetyl-coA downstream metabolism, hence were split into seven modules. Most of the TCA cycle intermediate substrates are with branches, so the TCA cycle was split into six modules. This super module is regarded as the central metabolism pathway. The main role of this super module are for energy (ATP) production and fueling other metabolic and biosynthesis pathways with acetyl-coA. The second super module is serine metabolism, which contains 220 reactions. This pathway plays a crucial role in controlling the balance and demand of amino acid types [1]. The Pentose Phosphate pathway (PPP) forms the third super module, contains 44 reactions involved in the biosynthesis of PRPP, a precursor for nucleic acids biosynthesis [2]. The fourth super module is biosynthesis and metabolism of fatty acids, which connects the main metabolic map only via the acetyl-coA. The fatty acids biosynthesis and metabolism pathways have a series of parallel reactions chains for different types of fatty acids. This super module contains two modules of fatty acid synthesis and metabolism, totaling 148 reactions [3]. We collected all amino acid metabolic pathways from KEGG database and rebuild super modules based on the network topology. In total, we generated six super modules of amino acids metabolism, namely Aspartate, Beta-Alanine, Glutamate, and Leucine/Valine/Isoleucine metabolism pathways and Urea Cycle. The aspartate metabolism pathway has 16 enzymes catalyzing 37 reactions, B-alanine metabolism pathway includes 21 enzymes carrying 130 reactions, glutamate metabolism pathway is with 10 enzymes and 21 reactions, and 16 enzymes for urea cycle, respectively. Each of the three essential metabolite leucine, isoleucine, and valine, has a

separate pathway. Two additional metabolic super modules are Propionyl-CoA metabolism for exchange of multiple coenzyme A types and spermidine metabolism related to the glutathione and S-adenosyl-L-methionine (SAM) metabolisms.

Transporters enable the movement of molecules between two side of cell membranes. We collect human transporter genes and annotations from Transporter Classification Database, by using the symbol and description in this database [4, 5]. We collected 116 transporter genes of 35 metabolites presented in the metabolic and biosynthesis modules.

An essential part of metabolic map is biosynthesis pathways. KEGG database and literature [6-11] are the main information sources used for building biosynthesis modules. We collected 69 biosynthesis modules forming 10 super modules, 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.

We also conducted the same approach to reconstruct the mouse metabolic map and enable the capability of mouse data analysis to scFEA. Detailed statistics of the mouse metabolic map, super modules, the number of modules and genes are given in the table below.

Table 1 in Supplementary Methods. Statistics of mouse modules and genes

SM ID	Super Module class	#Modules	#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

Reduction and reconstruction of the metabolic map into a factor graph.

A metabolic module is defined by a number of connected metabolic reactions. Denoted $M^i = \{R_1^i, \dots, R_{k_i}^i\}$ as a module i contains k_i reactions $R_1^i, \dots, R_{k_i}^i$. Denote the flux of a reaction R

as $Flux_R$ and the flux of a module M^i as $Flux_{M^i} = \{Flux_{R_1^i}, \dots, Flux_{R_{k_i}^i}\}$.

Definition 1. Independency of reaction flux. We call two reactions R_1 and R_2 have independent fluxes if a perturbation in R_1 (or R_2) will not affect the solution space of R_2 (or R_1) under flux balance condition, denoted as $R_1 \perp R_2$. Similarly, we can define the independence between a reaction R_1 and a metabolic module $M^i = \{R_1^i, \dots, R_{k_i}^i\}$, by $R_1 \perp M^i$ if $R_1 \perp R_1^i, \dots, R_1 \perp R_{k_i}^i$, and similarly for two modules.

Intuitively, for any connected two reactions (or modules) with a potential metabolic flow exchange, the flux of the two reactions (or modules) cannot be independent.

Definition 2. Conditional independency of reaction flux. We call two reactions R_1 and R_2 are conditionally independent given the flux of C, here C is a set of reactions, if the R_1 and R_2 have independent fluxes when the fluxes of the reactions in C are fixed, denoted as $R_1 \perp R_2|C$. Similarly, we can define the conditional independency between one reaction and one module, or between two modules.

One straightforward example of conditional independency is a linear reaction chain, in which R_1 generates the inputs of R_2 and R_2 generates the inputs of R_3, i.e., $R_1 \to R_2 \to R_3$. Here $R_1 \perp R_3|R_2$ under flux balance condition.

The goal of our network reduction is to reduce the network complexity for a more efficient learning. By merging multiple connected reactions into a module, we utilize the module to represent the merged reactions. Intuitively, the first condition needs to be satisfied in the network reduction is that a merged module should have a unique and meaningful flux that could represent the fluxes of the reactions in the module, i.e. (i) the flux of the module outputs needs to have a unique solution when the flux of inputs is fixed.

Definition 3. Flux of a merged module. If a merged module M^i satisfies the flux of the module outputs needs to have a unique solution when the flux of inputs is fixed, we define the module flux as a vector of the flux of out its outputs, denoted as $\widetilde{Flux}_{M^i}$.

In addition to the necessary condition of a unique and meaningful solution, the reduction a series of reactions into a module should not affect the uncertainty of other reactions, i.e., (ii) for any two reactions, R_p and R_q, $R_p, R_q \notin M^i$, if $R_p \perp R_q|M^i$, then $R_p \perp R_q|\widetilde{Flux}_{M^i}$ and if $R_p \not\perp R_q|M^i$, then $R_p \not\perp R_q|\widetilde{Flux}_{M^i}$, here $\not\perp$ indicates not independent.

Lemma 1. A merged module satisfies (i) and (ii) if:

(1) None of the merged intermediate metabolites has more than one out-flux reactions that correspond to more than one module outputs.

(2) None of the merged intermediate metabolites has an in-flux or out-flux other than merged reactions or the module input and output.

Proof of condition (i): If none of the merged intermediate metabolites has more than one out-flux reactions that correspond to more than one module outputs and none of the merged intermediate metabolites has an in-flux or out-flux other than merged reactions or the module input and output, i.e., each intermediate metabolite does not result into different branches of outs, hence the flux of the module outputs needs to have a unique solution when the flux of inputs is fixed. On the other hand, when an intermediate metabolite has multiple out-flux reactions, if these fluxes result into more than one module outputs, we can also identify such

an intermediate metabolite C that is closest to the module output, and all the intermediate metabolites between this metabolite to the module outputs are either (a) has more than one out-flux reactions that correspond to one module outputs or (b) only has one out-flux, hence the solution of the out-flux reactions of C is not unique given fixed module inputs. If the merged intermediate metabolite has an in-flux or out-flux other than merged reactions or the module input and output, under flux balance condition, the module output is unfixed due to this in-flux or out-flux is unfixed, hence the outflux of the module is unique. □

Proof of condition (ii): If the condition (i) holds and none of the merged intermediate metabolites has an in-flux or out-flux other than merged reactions or the module input and output, i.e., the module outputs are fixed and the intermediate metabolites of the module does have any biochemical mess exchange with other reactions other than through the module inputs or outputs, for any R_p and R_q, $R_p, R_q \notin M^i$, we have $R_p \perp R_q | \widetilde{Flux}_{M^i}$ if $R_p \perp R_q | M^i$, and $R_p \overline{\perp} R_q | \widetilde{Flux}_{M^i}$ if $R_p \overline{\perp} R_q | M^i$. □

Noted, based on Lemma 1, if the two conditions hold, i.e., (1) None of the merged intermediate metabolites has more than one out-flux reactions that correspond to more than one module outputs, and (2) None of the merged intermediate metabolites has an in-flux or out-flux other than merged reactions or the module input and output, the module outputs have a unique solution under fixed inputs and changes of the reactions inside the module are independent to reactions outside the module conditional to a fixed flux rate of the module, i.e., solving the flux of each individual reaction in a merged module is equivalent to solve the flux of the module.

Model Implementation

The deep neural network is implemented based on pytorch version 1.6.0. Structure of neural network is costumed in a *Flux* Class object. For each metabolic module, a three layers neural network was created. The number of hidden nodes is eight. The number of output node is one. Since gene number in metabolic modules are different, we adopt a dynamic way to create input nodes. In *Flux* Class definition, the number of input nodes is fixed at the total gene number for all metabolic modules. However, we set the input value as zero if current gene does not exist in the current metabolic module. In addition, we do not allow the *bias* parameter for the input layer. In this way, only existed genes are connected to the hidden layer and actual input nodes of sub-networks are different. Input gene expression value of sub-network is normalized by logarithm if the input value is larger than 30. An activation function calculates a weighted sum of its input, add a bias and then decides whether it should be active or not. A Hyperbolic Tangent activation function, named *Tanhshrink*, is used here. The element-wise of *Tanhshrik* function is defined as $Tanhshrink(x) = x - \tanh(x)$. To build all sub-networks in a large deep neural network, we use *torch.nn.ModuleList* to store parallel sub-networks. The second part of the large parallel neural network is the constrain function for estimated flux value. The flux balance of each metabolite can be formed as a linear equation. In other words, inflow value is supposed to equal to outflow value for each metabolite. In total, the number of linear equations is equal to the number of metabolites. The calculation of linear equations is a child function of *Flux* object to ensure balance status is updated in every step of optimization. The stoichiometric matrix, which stored the corresponding relationship between metabolite and modules, is used

to update linear equations. We use a stochastic optimization method to update the parameters for all sub-networks. To avoid the trivial solution, we add penalty term in the objective function $\lambda \sum_{j=1}^{N}\left(\sum_{m=1}^{M} Flux_{m,j} - TA_j\right)^2$. TA_j is the summation of gene expression value of all metabolic genes. each supermodule modules. This penalty term also makes sure the estimated flux value proportion to the gene expression value scale for each single cell. λ is the hyperparameter to balance the importance of two terms in the objective function. Learning rate in optimization is another hyperparameter. Small learning rate will cause slow converge while large learning rate will cause too oscillatory to converge.

Neural network of the flux of each metabolic module

The metabolic network is a complex biological topological structure. To mimic the inclusiveness and flexibility of metabolic network in single cell resolution level, we model it by deep neural network which is powerful to describe nonlinear relationship and capture the latent information in large scale data with unknown noise. Although the metabolic network has high connectivity, each metabolic module is independent and only regulate the specific functionality in individual cells. In our model, each metabolic module is implemented as an independent sub-network. The sub-network is a deep neural network consist of input layer, hidden layers, and output layer. The input of sub-network is SC gene expression value and the output is the estimated flux. These sub-networks have high connectivity of output nodes by paralleling them as a large-scale deep neural network. If there are common genes in several modules, these sub-networks have connection in hidden layer via common input nodes. In each metabolic module, the node number of input layer is matched with gene number in module and thus a dynamic deep network construction method is proposed (see detailed implementation). The total node number of output layer is equal to the module number M.

Scalability and Identifiability.

Scalability analysis. The mainly time-consuming comes from the training of deep neural network. Maturing of a neural network consists of forward pass process and back-propagation update process. Forward pass process can be formed as matrix multiplication, where input multiply the weights on the link and plus the bias. Then activation function has $O(1)$ time complexity. The time complexity is $O(e * N * (i * h + h * m))$ for three layers forwarding and back-propagation update, where i is input layer node number, h is hidden layer node number, m is output layer node number, N is the cell number, e is number of iterations. In our work, $i = \sum_{m=1}^{M} i_m = 726$, $h = M \times 8$, $m = 175$, $e = 100$, N is cell number for each dataset. Paralleling GPUs matrix operation is encouraged since sub-networks are independent.

Identifiability. The number of parameters in the complete model is around $\sum_{m=1}^{M} i_m + 4^x M$, where i_m is the number genes in module m ($M = 175$ and $\sum_{m=1}^{M} i_m = 726$ for the complete map) and x is the number of layers of f_{nn}^{m} ($x = 2\ or\ 3$ as the empirical setting). The number of constraints is the total number of metabolites, K ($K = 84$ for the complete map). Hence the number of total constraints divided by the number of parameters is $\frac{KN}{\sum_{m=1}^{M} i_m + 4^x M}$, which is $0.0238 * N$ and $0.007 * N$ when $x = 2$ or 3. For a scRNA-seq data with $N \sim 10^2$ for the data generated by constructing a library for each individual cell or

$N \sim 10^3$ for drop-seq data, selecting $x = 2\ or\ 3$ the $\frac{\#constraints}{\#parameters}$ is much larger can 1, hence guarantee the identifiability and mathematical correctness of the formulation.

Data simulation, perturbation, cross-validation, and drop-out experiment details.

To validate scFEA predicted metabolic fluxes, we simulated pseudo scRNA-seq data where the true cell-wise flux is known. The difficulty of simulation is to mimic the non-linear relationship between genes and metabolic modules. We took two-step recurrent way to solve the challenge. Firstly, we separate 1000 SCs and total 21 super modules in 10 groups. The genes thus been divided as 10 groups as well. For each group, the expression value forms an independent normal distribution $N(10,2)$. Then, feed generated scRNA-seq data into scFEA and get the predicted flux as the basis for second round simulation. The current predicted fluxes are ground truth. In second step, the expression value in each module and each single cell forms an independent normal distribution $N(\mu, \sigma^2)$, where μ is predicted flux in last step. Then, new pseudo dataset can also be applied to scFEA and correlate the predicted flux with the ground truth in sample/module-wise. Pearson correlation was used in analysis.

Other validation experiments based on our pancreatic cancer dataset. To generate perturbed data, we set parameter α to control the ratio of perturbed single cells. In each setting, we randomly selected α single cells and shuffled the genes. Five repetitions executed in each setting. To validate the robustness of scFEA, we executed both cross-validation and drop-out experiment. In cross-validation experiments, we separated total single cells into 5 or 10 groups and 80% (4/5) or 90% (9/10) single cells were used in each experiment. In drop-out experiments, we randomly sampled metabolic module related genes. After fixing the iterative addition drop-out rate, we sample single cells and make the expression as zero. The lower expression value has a high drop-out probability.

Public scRNA-seq data processing and analysis

We collected six datasets from public domain. Basic QC for SC using the Seurat (version 3) default parameter to filter out cells with high expressions of MT-coding genes. The cell type label and sample information provided in the original work were directly utilized.

GSE132581: This dataset is collected on mouse perivascular adipose tissue. The original work indicated the two distinct subpopulations existed in PVAT-derived mesenchymal stem cells.

GSE72056: This dataset is collected on human melanoma tissues. The original paper provided cell classification and annotations including B cells, cancer-associated fibroblast (CAF) cells, endothelial cells, macrophage cells, malignant cells, NK cells, T cells, and unknown cells.

GSE103322: This dataset is collected on head and neck cancer tissues. The original paper provided cell classification and annotations including B cells, dendritic cells, endothelial cells, fibroblast cells, macrophage cells, malignant cells, mast cells, myocyte cells, and T cells. Notably, as indicated by the original work, malignant cells have high intertumoral heterogeneity.

CCLE data: This dataset was downloaded from Broad Institute CCLE data portal (https://portals.broadinstitute.org/ccle). In total, pan-cancer cell lines (n = 1076) were included in this paper.

Spatial breast cancer data: This dataset was downloaded from 10x spatial official website. Block A section 1 was used in this paper. (https://support.10xgenomics.com/spatial-gene-

expression/datasets)

ROSMAP data: This dataset is generated from the Religious Orders Study (ROS) or the Rush Memory and Aging Project (MAP), mainly focus on the Alzheimer's disease research. The dataset was download from RADC Research Resource Sharing Hub (https://www.radc.rush.edu/)

SUPPLEMENTARY REFERENCES

1. Mattaini, K.R., M.R. Sullivan, and M.G. Vander Heiden, *The importance of serine metabolism in cancer.* The Journal of cell biology, 2016. **214**(3): p. 249-257.
2. Jin, L. and Y. Zhou, *Crucial role of the pentose phosphate pathway in malignant tumors (Review).* Oncol Lett, 2019. **17**(5): p. 4213-4221.
3. Mikalayeva, V., et al., *Fatty Acid Synthesis and Degradation Interplay to Regulate the Oxidative Stress in Cancer Cells.* International journal of molecular sciences, 2019. **20**(6): p. 1348.
4. Bhutia, Y.D., et al., *SLC transporters as a novel class of tumour suppressors: identity, function and molecular mechanisms.* The Biochemical journal, 2016. **473**(9): p. 1113-1124.
5. Lin, L., et al., *SLC transporters as therapeutic targets: emerging opportunities.* Nature reviews. Drug discovery, 2015. **14**(8): p. 543-560.
6. DeAngelis, P.L., J. Liu, and R.J. Linhardt, *Chemoenzymatic synthesis of glycosaminoglycans: Re-creating, re-modeling and re-designing nature's longest or most complex carbohydrate chains.* Glycobiology, 2013. **23**(7): p. 764-777.
7. Gao, C. and K.J. Edgar, *Efficient Synthesis of Glycosaminoglycan Analogs.* Biomacromolecules, 2019. **20**(2): p. 608-617.
8. Krasnova, L. and C.-H. Wong, *Understanding the Chemistry and Biology of Glycosylation with Glycan Synthesis.* 2016. **85**(1): p. 599-630.
9. Lv, X., et al., *Synthesis of Sialic Acids, Their Derivatives, and Analogs by Using a Whole-Cell Catalyst.* Chemistry (Weinheim an der Bergstrasse, Germany), 2017. **23**(60): p. 15143-15149.
10. Moffatt, B.A. and H. Ashihara, *Purine and pyrimidine nucleotide synthesis and metabolism.* The arabidopsis book, 2002. **1**: p. e0018-e0018.
11. Zulueta, M.M., et al., *Synthesis of glycosaminoglycans*. 2016. p. 235-261.