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

Norah Alghamdi1+, Wennan Chang1,2+, Pengtao Dang1,2, Xiaoyu Lu1, Changlin Wan1,2, Silpa Gampala3, Zhi Huang1,2, Jiashi Wang1, Qin Ma4, Yong Zang1, 5, Melissa Fishel3\*, Sha Cao1, 5\*, Chi Zhang1,2\*

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

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

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

\*To whom correspondence should be addressed. +1 317-278-9625; Email: [czhang87@iu.edu](mailto:czhang87@iu.edu). Correspondence is also addressed to Melissa Fishel, Email: mfishel@iu.edu, and Sha Cao, Email: [shacao@iu.edu](mailto: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 22 super module classes of 169 modules. For the metabolism part, all reactions were collected from Kyoto Encyclopedia of Genes and Genomes database (KEGG). 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 is 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 |
| 22 | Steroid hormone synthesis | 3 | 177 |

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

A metabolic module is defined by a number of connected metabolic reactions. Denoted as a module contains reactions . Denote the flux of a reaction as and the flux of a module as .

**Definition 1. Independency of reaction flux.** We call two reactions and have independent fluxes if a perturbation in (or will not affect the solution space of (or ) under flux balance condition, denoted as . Similarly, we can define the independence between a reaction and a metabolic module , by if , 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 and are conditionally independent given the flux of , here is a set of reactions, if the and have independent fluxes when the fluxes of the reactions in are fixed, denoted as . 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 generates the inputs of and generates the inputs of , i.e., . Here 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 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 .

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, and , , if , then and if , then , here 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 and , , we have if , and if . □

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 . 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. 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. 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 .

*Scalability and Identifiability.*

*Scalability analysis.* The most time-consuming of scFEA is the training of neural networks. In the training step, we update the parameter of neural network to minimize the loss function. Specifically, the training process consists of a forward pass process and a back-propagation update process for each epoch. The forward pass process is formed by a matrix multiplication operation, namely, input multiplied by the weights and plus the bias. Then an activation function determines whether each neural node is activated or non-activated, where the activation function has time complexity. In total, the time complexity is for three layers of forwarding and back-propagation update, where is the number of nodes in the input layer, is the number of nodes of the hidden layers, is the number of nodes of the output layer, is the cell number, is number of iterations. In this study, we set , , , , is cell number for each dataset. Noting that to implement scFEA, parallel GPU computing is encouraged since sub-networks are independent. We tested the running time of scFEA on a personal laptop of Intel i7-7600 CPU and 16GB RAM. The scFEA analysis of the complete network took 14 and 23 minutes on 4486 cells of the GSE72056 data and 5902 cells of the GSE103322 datasets, respectively.

*Identifiability.* The number of parameters in the complete model is around , where is the number genes in module ( and for the complete map) and is the number of layers of ( as the empirical setting). The number of constraints is the total number of metabolites, ( for the complete map). Hence the number of total constraints divided by the number of parameters is , which is and when or . For a scRNA-seq data with for the data generated by constructing a library for each individual cell or for drop-seq data, selecting the 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 22 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 . 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 , where 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. Based on the synthetic data, we had a predicted metabolic flux matrix which is 1000 single cells by 169 modules. To validate the predicted flux, we calculated the Pearson’s correlation of true flux matrix and predicted flux in row-wise (sample-wise) and column-wise (module-wise). Two groups of correlation coefficients are shown in Figure 5D as violin plot. For sample-wise correlation, each point represents Pearson’s correlation of cell , calculated by . All correlation coefficients under same null hypothesis, which is two variables have no linear correlation. Thus, we calculate Person’s correlation value with probability of 0.05 to reject the null hypothesis, using Student’s *t*-distribution. Then we calculate the ratio of points which greater than the correlation under null hypothesis.

Other validation experiments based on our pancreatic cancer dataset. To generate perturbed data, we set parameter to control the ratio of perturbed single cells. In each setting, we randomly selected 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.