Chapter 3

Predicting functional metabolic control mechanisms using Graph Convolutional Neural Networks

Martin Lempp[1]*, Niklas Farke[1]*, Hannes Link[1]

This chapter is written in manuscript style and is work in progress. My contribution to this work includedthe design of the study, coding and optimizing the neural networks and co-writing the manuscript.

Abstract

The systematic identification of functional metabolic control mechanisms relies mostly on the integration of data but barely considers the connectivity of the metabolic network. However, recent geometric deep learning approaches show promising performances in the prediction of links in network or graph structures. By using an *in silico* network of *Escherichia coli*'s central metabolism including metabolic and regulatory connections, we first generated data suitable for the prediction of regulatory links between metabolites and enzymes. By merging data with a Graph Convolutional Neural Network along the structure of the metabolic network we can correctly classify links in the graph as either 'regulatory' or 'non-regulatory'. Here we show, how such a neural network can be trained and also be used to fill knowledge gaps to identify functional metabolic control mechanisms.

* These authors contributed equally.

1 Max Planck Institute for Terrestrial Microbiology, D-35043 Marburg, Germany

Introduction

Although the stoichiometry and biochemistry of the metabolic network is fairly well described, it remains elusive how such systems maintain homeostasis. Besides of regulation of enzyme expression (transcriptional control) cells use metabolic control mechanisms that influence reaction rates (allosteric control) to maintain stable metabolic flux[1–4]. The systematic identification of allosteric metabolic regulations requires the confirmation of physical metabolite-enzyme interactions as well as information about the functionality of the interactions. State of the art to identify such control mechanisms in metabolic networks is the correlation of flux and metabolite changes across different conditions[5,6]. Another *in vitro* approach focuses on measuring enzyme-metabolite interactions directly[7]. Even though such high throughput methods have high potential to identify multiple interactions at once, they give only limited information about functional relevance of an interaction. All these methods have in common that regulatory link prediction is up to now based on data only and the metabolic network structure is usually not considered implicitly.

One way to use structural information of the network is to formulate the prediction of functional metabolite-enzyme interactions as a graph convolution-based link prediction task. As opposed to purely data-driven approaches, Graph Convolutional Neural Networks (GCNs) allow data integration while considering the network structure[8,9]. For each node in a network graph, GCNs merge the adjacent nodes' information and thereby capture the structural relationships of the network and how nodes are interconnected. By integrating network structure and data, GCNs have achieved state-of-the-art results in link prediction problems (for example friend recommendation in social media)[10]. In these tasks, the challenge is to predict the missing or future links of a network based on current knowledge. This strategy might be suitable for the identification of missing regulatory links in metabolic networks.

Here, we provide a first proof of principle based on *in silico* data how GCNs can be used to infer regulatory links in metabolic networks. We show that neural networks can be trained by using information about fluxes and metabolite changes in order to predict regulatory interactions in metabolic networks. Only the use of structural information of the network and functionality gives highest prediction accuracy and moreover, we show how such convolutional approaches can be used to fill knowledge gaps and to infer complete layers of information.

Results

Metabolic model generates data for link prediction

To use geometric deep learning for link prediction tasks we have to obtain data suitable to identify these links. To generate dream case data for such a proof of principle, we generate *in silico* data using a mechanistic model of *Escherichia coli*'s central metabolism[11] (**Fig. 1a**). The model contains the stochiometric information of 444 reactions and 314 metabolites in *E. coli* central carbon metabolism as well as 172 allosteric interactions between metabolites and enzymes. The majority of effector metabolites regulate only up to 2 reactions. The majority of regulated reactions have just up to two effectors. Only five regulations are unique, which means that there is only a single interaction between reaction and effector (**Fig.1b**). As we want to identify also such unique interactions with a link prediction approach, we added randomly 60 regulations with exclusive regulator-target pairs. By that, we can ensure that also interactions are identified, where neither effector nor target was part of a training set.

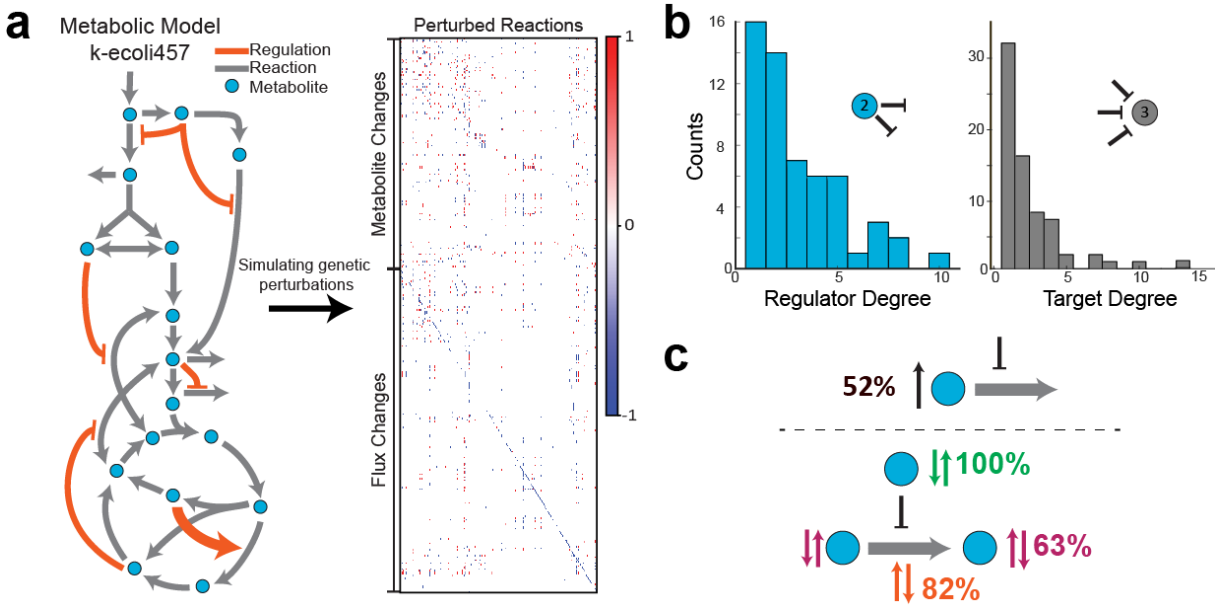

Figure 1. Metabolic flux analysis generates suitable data. a) By perturbing single reactions in a genome scale metabolic model, we generated discrete data of flux and metabolite changes in the network (1 - up, 0 – no change, -1 – down). The model consists not only of metabolic interactions between reactions (grey) and metabolites (blue), but also of regulatory interactions (orange). **b)** Node Degree Distributions of all regulators and targets in the model of the metabolism before adding unique regulations. **c)** Substrate accumulations of perturbed reactions (upper layer) and propagation of discrete changes of regulatory levels (green) along the regulations to the reaction (orange) and substrates or products (purple) level (lower layer).

To generate changes in reaction flux and metabolite levels, we perturbed all 444 reactions by lowering the flux through this reaction. Such perturbations simulate a genetic perturbation, e.g. the

knockdown of enzymes by CRISPR interference[3]. We simulated each perturbation 10 000 times with different kinetic parameters. We only considered a discrete change in flux or metabolite concentrations if we recorded this change in 66% of the parameter sets (1 – up, 0 – no change, -1 – down, see Methods section). We tested whether a changed metabolite or flux level affects connected fluxes and metabolites, either via regulatory or metabolic interactions. In 52% of the perturbations, the substrate of the perturbed reaction increased (**Fig. 1c**). In addition, we recorded in 82% of the cases where the effector level changed that also the flux in the regulated reaction changed. This effect propagates again in 62% and the substrates or product metabolite level of the regulated reaction changed. Based solely on the data, we would be able at best to recover 82% of the included regulations.

Graph convolutional networks identify regulatory links

We transformed the model network into a graph structure, with metabolites and reactions denoting nodes with the 444 features form the perturbations. Edges connecting the node pairs denoted relationships between reactions and metabolites and can either be regulatory or metabolic. The graph consisted of 1302 edges, from which 232 are regulatory and the remaining metabolic. To improve the power of the classification we added in a first step random edges between metabolites and reactions that are neither a metabolic edge nor a regulatory edge. During the convolution in the next step, the feature information of each node gets merged with the feature information from adjacent nodes to calculate an updated node feature (**Fig. 2a**). We only merge nodes in the immediate surroundings but in principle, the convolution could include several steps. After the convolution, merged features contained information about their neighbors. To generate a score for each edge, we toke each connected node pair and generated the respective edge score by merging the node features. Based on the calculated edge scores, edges are either classified as ‘regulatory edge’ or ‘non-regulatory edge’, where the second class contains artificial edges are as well as metabolic edges.

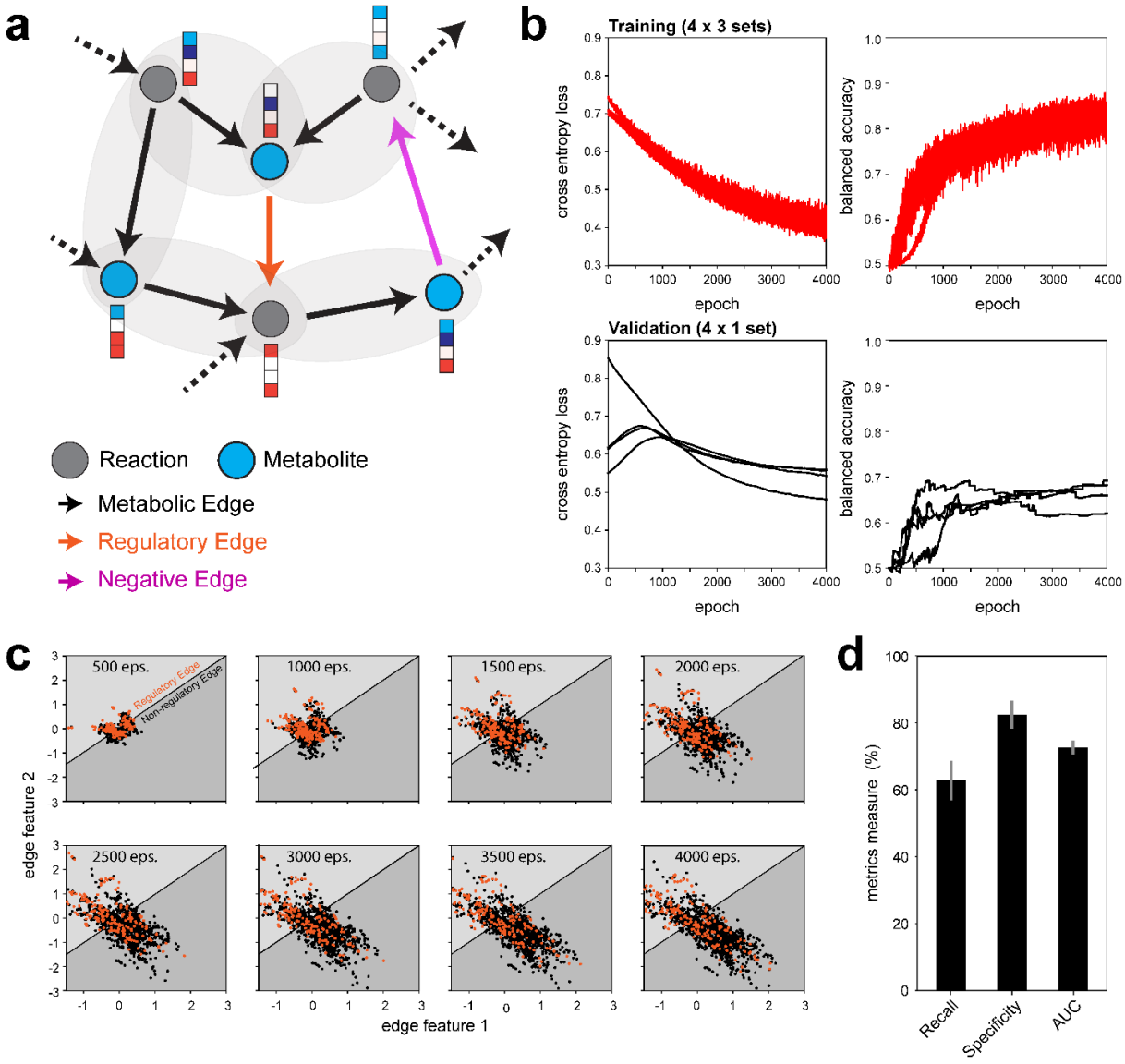

Figure 2. Graph convolution can learn to predict regulatory links on discrete flux and metabolite changes. a) Schematic how feature passaging follows the network structure. Features of each node get propagated along metabolic edges. After the convolution, merged features contain information about their neighbors. **b)** Cross entropy loss (left) decreases while accuracy (right) increases for all training (red) and validation (black) sets over 4 000 epochs. **c)** With increasing learning epochs (eps.), edges in a validation set get either classified as a regulatory edge (light grey area) or non-regulatory edge (dark grey). **d)** Classification metrics of the test set after 10 randomly initialized cross-validations. Grey bars indicate the standard deviation of the 10 cross-validation sets.

To train, validate and test the approach and the convolutional network, we split the edges into 5 even-sized batches. While 4 batches are always used for training and validation, 1 set is exclusively used to test the model. During each cross-validation, we always used 3 of the 4 training-validation batches for training and 1 batch for the cross-validation of the generated model. Each of the models was then used on the testing set to classify the edges. To confirm first whether our network was able to learn

on the discrete data, we checked the accuracy and loss for the training and validation data of the cross-validation sets (**Fig 2b**). The loss for the training sets decreased while the balanced accuracy increases. However, if we compared the accuracies there are slight differences between the validation and training set, indicating some overfitting. Checking the decision boundaries for the classification in the validation sets across the training epochs, we saw that the majority of negatives are correctly classified (**Fig. 2c**). This showed that our modeling approach would be able to learn regulatory links from simple metabolomics generated, discrete data. To access the final performance of the model we tried to predict the classes in the testing set. The respective class was predicted with each of the generated models from each cross-validation set and the results were averaged afterward. This ensemble approach should ensure for the evaluation of the models with the testing set highest precision in prediction. In this final testing set, our model predicted interactions with an average accuracy of 73%. We recovered real regulations with our model with a recall of 63%. With a specificity of 82%, non-regulatory edges we correctly identified and rejected (**Fig. 2d**). In addition, the model was also able to recover unique regulations with exclusive reaction-effector pairs.

Combining network structure and data recovers interactions best

To check whether the use of network information indeed increased the performance of the prediction, we compared the results of the graph convolution to a deep learning model where we substituted the convolutional layer with a fully connected linear layer (**Table 1**). As expected, the merging of data along the network structure gave a slight benefit in the classification of the test edges. As metric for our classification we use the F1 score, a common metric used if classes are imbalanced, so have unequal sizes. The F1 score is highest for the convolutional network. A model with a linear layer has a lower recall and specificity if we use the same training, validation and test sets like for a graph convolution approach. To assess how much the data influences the classification we also compared the predictions to a classification without the data information. In such a scenario, edges should just be classified as regulations, if the partners of this edge were classified before as a member of a regulator-effector pair. Indeed, without the data we are not able to identify the unique edges that were added to the network in the beginning.

Table 1. Recall, Specificity, balanced accuracy, F1 score and whether the model is able to recover unique effector-reaction pairs for a neural network with graph convolution, without graph convolution and a graph convolution without data features.

	Graph convolution with data	No Graph convolution	Graph convolution without data
Recall	62.83% ± 5.93%	56.30% ± 3.55%	61.30% ± 3.64%
Specificity	82.47% ± 4.23%	81.77% ± 3.15%	79.36% ± 2.49%
AUC	72.65% ± 2.02%	69.04% ± 1.69%	70.33% ± 1.61%
F1 score	0.42 ± 0.03	0.38 ± 0.03	0.38% ± 0.03
Uniques	YES	YES	NO

Graph Convolution recovers interactions despite information gaps

At this point, it seems that there is only a slight benefit of using structural information in the classification. However, we used up to this point *in silico* data, even though we only used discrete information about reaction flux or metabolite change for every reaction and metabolite of the network. This scenario will not apply to real experimental data. In such a case, we will most likely have only partial information about the network and knowledge gaps about fluxes and concentrations. In such a case, the aggregation of information along the edges of the metabolic network to classify edges as regulatory should highlight the performance and power of the convolutional approach. To test this hypothesis, we randomly removed 60% information about flux or metabolite changes from the data. Again, we test the performance of the convolutional network against a simple linear network. While the convolutional network is still able to discriminate between regulatory and non-regulatory edges with an F1 score of 0.31, a linear network has a lower score in the prediction if an edge is regulatory (**Table 2**). Especially the classification of non-regulatory edges decreases. This difference gets even more striking if we only use metabolic data and exclude information about fluxes. The convolutional network shows only little impairment by the missing flux data. The network without convolution has an even lower accuracy than with the random knowledge gaps (**Table 2**).

Table 2. Recall, Specificity, balanced accuracy, F1 score and whether the model is able to recover unique effector-reaction pairs for a neural network with graph convolution and without graph convolution in cases of knowledge gaps or missing flux information.

	60% knowledge gaps		only metabolites levels	
	Graph convolution with data	No Graph convolution	Graph convolution with data	No Graph convolution
Recall	37.61% ± 8.04%	70.22% ± 8.28%	58.26% ± 5.31%	57.17% ± 10.49%
Specificity	85.93% ± 8.51%	56.92% ± 7.95%	80.58% ± 4.36%	69.51% ± 8.97%
AUC	61.77% ± 2.88%	63.57% ± 1.40%	69.42% ± 1.66%	63.34% ± 2.02%
F1 score	0.31 ± 0.04	0.28 ± 0.016	0.38 ± 0.023	0.29 ± 0.028
Uniques	YES	YES	YES	YES

Discussion

Summarizing this proof of principle, deep learning approaches can identify regulatory metabolite-enzyme interactions. The model convolutes data along the stoichiometry of the metabolic network and can learn regulatory interactions on a combination of metabolite and flux data. In solely data-driven approaches the high number of false positives is a massive problem. Especially the high specificity in finding true negatives is an advantage of this approach[5,12]. At this point are the accuracies of the presented method not comparable with other classification problems. However, improving data quality could have a huge impact on the results. Here we belong only discrete values generated by an *in silico* model which is assumed to be a "dream case scenario". The dream case would be that we know the complete structure of the network and can perform the convolution without possible missing links in the structure[11]. Perhaps, real experimental and continuous data could generate higher accuracies. Even adding information about the similarity or correlation between potential effector-target pairs or about the metabolic subsystem could improve the performance[12].

Besides, we show the advantage of using such convolutional approaches on networks in case of feature gaps or completely missing information about one layer (e.g. fluxes of reactions). Many disciplines in systems biology rely on network structures where possible information gaps appear. These can be knowledge gaps for certain nodes in the used network or even complete missing layers like a network of transcription factors. This regulatory layer relies on inputs from a metabolic network and regulates the gene expression in a cell[12,13]. Even having only information about the protein levels and metabolites but missing the layer of regulation, we could use a combined regulatory and metabolic network. A convolutional network would learn to infer the intermediate regulatory layer so that it explains the final data the best.

Material & Methods

Generation of discrete *in silico* data

To generate data suitable for link prediction, we used the reaction stoichiometry and flux information from the model of Khodayari *et al.*[11]. At first, we adjusted the reaction stoichiometry to ensure that all fluxes are positive. Then, we removed all highly connected metabolites and isolated reactions. We then constructed a bipartite graph and added regulatory interactions[14]. We then removed the substrate and product regulations. We added 60 novel regulations. We then simulate flux and concentration control coefficients (FCC and CCC) using the reaction stoichiometry, reaction fluxes and 10 000 randomly sampled elasticities for reactions and regulations. The elasticities are sampled from a log-uniformly distribution in the range 0.001 and 1 for reactions (1 means that the enzyme operates in the linear regime, 0 means that the enzyme is saturated) and in the range (-)1 and (-)4 for regulatory interactions. The control coefficients tell us about the sensitivity of each reaction towards a local perturbation for a given set of elasticities. We then discretize the control coefficients. If 66% of the coefficients are bigger than an arbitrary cutoff of $5x10^{-4}$, a 1 is assigned for upregulation. If 66% of the coefficients fall below -$5x10^{-4}$, then a -1 is assigned.

Graph convolution for regulatory link prediction

To perform the graph convolution with the data of the metabolic flux analysis, we used the Pytorch geometric implementation of GCNs[9]. We first randomly sampled negative edges in the network. After the addition of the negative samples, we randomly split the network into 5 sets, where 1 set is used as the final testing set while the other 4 are used for the training and cross-validation of the model. To generate confidence values for our model, we performed 10 randomly initiated cross-validations. The network consists of one graph convolutional layer and one linear layer with two outputs for the final classification of regulatory or non-regulatory edges. The model was trained over 4 000 epochs with a drop out 0.5. The discrete features and the network of the metabolic flux analysis were used as input. To generate knowledge gaps, we set 60% of the node features to zero. To use only metabolite changes as input, we set all flux changes to zero. To use only structural information, we one-hot-encoded each node of the network instead of using the generated features. To use a neural network without a convolutional layer we substituted the graph convolutional layer by a fully connected linear layer.

Code and Data availability

Matlab code to perform Metabolic Flux Analysis and Python code to perform the Link prediction can be accessed from the GitHub repository https://github.com/mlempp/Metabolic_Control_Link_Prediction.

References

1. Chubukov, V., Gerosa, L., Kochanowski, K. & Sauer, U. Coordination of microbial metabolism. *Nature Reviews Microbiology* **12**, 327–340 (2014).

2. Gerosa, L. & Sauer, U. Regulation and control of metabolic fluxes in microbes. *Current Opinion in Biotechnology* **22**, 566–575 (2011).

3. Sander, T. *et al.* Allosteric Feedback Inhibition Enables Robust Amino Acid Biosynthesis in E. coli by Enforcing Enzyme Overabundance. *Cell Systems* **8**, 66-75.e8 (2019).

4. Lindsley, J. E. & Rutter, J. Whence cometh the allosterome? *Proceedings Of The National Academy Of Sciences* **103**, 10533–10535 (2006).

5. Hackett, S. R. *et al.* Systems-level analysis of mechanisms regulating yeast metabolic flux. *Science* **354**, aaf2786 (2016).

6. Link, H., Kochanowski, K. & Sauer, U. Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nature Biotechnology* **31**, 357–361 (2013).

7. Piazza, I. *et al.* A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* **172**, 358-372.e23 (2018).

8. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *arXiv:1607.00653* (2016).

9. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907* (2016).

10. Zhang, M. & Chen, Y. Link Prediction Based on Graph Neural Networks. *arXiv:1802.09691* (2018).

11. Khodayari, A. & Maranas, C. D. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nature Communications* **7**, 1–12 (2016).

12. Lempp, M. *et al.* Systematic identification of metabolites controlling gene expression in E. coli. *Nature Communications* **10**, 1–9 (2019).

13. Ortmayr, K., Dubuis, S. & Zampieri, M. Metabolic profiling of cancer cells reveals genome-wide crosstalk between transcriptional regulators and metabolism. *Nature Communications* **10**, 1841 (2019).

14. Reznik, E. *et al.* Genome-Scale Architecture of Small Molecule Regulatory Networks and the Fundamental Trade-Off between Regulation and Enzymatic Activity. *Cell Reports* **20**, 2666–2677 (2017).