12th IFAC International Symposium on
Advanced Control of Chemical Processes
July 14-17, 2024. Toronto, Canada

Graph Neural Network Representation of State Space Models of Metabolic Pathways

Mohammad Aghaee * **Stephane Krau** ** **Melih Tamer** ***
Hector Budman ****

* *Department of Chemical Engineering, University of Waterloo, Waterloo, N2L3G1, Canada (e-mail: maghaeef@uwaterloo.ca).*
** *Manufacturing Technologies, Sanofi, Toronto, M2R3T4, Canada (e-mail: stephane.krau@sanofi.com)*
*** *Manufacturing Technologies, Sanofi, Toronto, M2R3T4, Canada (e-mail: melih.tamer@sanofi.com)*
**** *Department of Chemical Engineering, University of Waterloo, Waterloo, N2L3G1, Canada (e-mail: hbudman@uwaterloo.ca)*

Abstract: A novel Metabolic Graph Neural Network (MGNN) model is proposed for simulating the dynamic behavior of metabolites involved in oxidative stress metabolic pathways in a bacterial cell culture. The developed MGNN model is trained and validated with in-silico data generated from the mechanistic model. By using the a priori known metabolic network, the proposed MGNN model effectively reduces the overfitting issue as compared to a fully connected network that does not uses the metabolic network knowledge. The MGNN exhibits a superior fit for both training and testing datasets. The proposed MGNN is highly interpretable since it efficiently computes the relevance of each metabolite on any other metabolite by applying gradient computation and back-propagation operations to the neural network. The proposed model is also shown to be useful for fault detection.

Keywords: Graph Neural Network, Metabolic Pathways, State Space Model, Bordetella pertussis, Oxidative Stress, Fault Detection and Diagnosis

1. INTRODUCTION

With the development of improved computer technology and the availability of large amounts of biological data, accurate modeling of biochemical processes has become increasingly important to leveraging these data to predict biological systems' behavior.

The traditional kinetic modelling (KM) approach that has been widely used for modelling dynamic bio-processes is based on the formulation of mass balances for key metabolites. For example, Lei et al. (2001) proposed a continuous mechanistic model for pyruvate metabolism in *saccharomyces cerevisiae* fermentation. The model presents the dynamic of eight key metabolites of pyruvate metabolism in *saccharomyces cerevisiae* with 12 reactions which all were modelled based on Michaelis–Menten kinetics with respect to a given substrate and with a first order dependency on the active biomass pool. These kinetic models rely on explicit functional relationships connecting the rate of change of metabolites and the enzyme kinetics involved in each reactions. Michaelis–Menten kinetics is the most commonly used kinetic rate expression although the true mechanistic kinetic rate law for each specific reaction is unknown a priori for most enzymes. Because each such kinetic expression involves a number of parameters, overall KMs require a large number of parameters that must be calibrated via nonlinear optimization based on generally noisy and scarce data. The problem becomes more pronounced in large metabolic networks due to the large number of metabolites and reactions involved. Generally, the structure of these models is only loosely based on prior knowledge about the interactions between metabolites.

Dynamic metabolic flux modelling, especially the dynamic flux balance analysis (DFBA) is an alternative approach that uses detailed knowledge of the metabolic network of reactions. Mahadevan et al. (2002) described two different formulations of DFBA, which incorporate rate of change of flux constraints. The resulting DFBA models were able to describe the dynamics of diauxic growth of *Escherichia coli* on glucose and acetate as well as the dynamics of key metabolites' concentrations. The DFBA model of Mahadevan et al. (2002) consists of four ordinary differential equations for extracellular glucose, acetate, oxygen and biomass. DFBA are generally formulated by a constrained optimization problem where a biological objective, e.g. growth rate is maximize subject to stoichiometric and other constraints. Although this approach has the potential to reduce the number of model parameters as compared to the traditional kinetic modelling approach, stoichiometric constraints alone are not sufficient to fully describe dynamic behaviour as they ignore enzyme kinetics limitations. To address this problem, additional kinetic rate constraints are added to the constrained optimization problem. Since these additional constraints involve tuning

⋆ This research was supported by Mitacs through Mitacs-Accelerate program

parameters, a crucial challenge in DFBA is to identify a minimal number of kinetic constraints to fit data while avoiding over-fitting. This search often requires the solution of mixed integer optimization problems that are difficult to solve. Also, DFBA and KM are less suitable for monitoring due to model error unless the latter will be accounted for as in e.g. Du et al. (2015).

As an alternative to the aforementioned KM and DFBA approaches, data-driven models such as machine learning models can be used when large datasets and computing power. Costello and Martin (2018) developed a machine learning algorithm to predict pathway dynamics in an automated fashion. For example, a Python code called TPOT (Tree-based Pipeline Optimization Tool) which relies on a Genetic optimization algorithm is used to automatically select the most suitable model among different candidates and to choose among different preprocessing algorithms that are included in the scikit-learn Python package. This method was tested on Limonene pathways and compared to the Michaelis-Menten kinetic model with 10 ordinary differential equations. However, the results exhibited considerable offset with respect to experimental data. The lack of fit was explained by the occurrence of highly nonlinear behaviour combined with scarce data for training.

Deep learning (DL) models can be used to effectively reconstruct highly nonlinear trajectories but there is always the possibility of data overfitting due to the high number of parameters in DL models. In case of over-fitting, this can be reduced by either increasing the amount of data or by pruning the neural network model to decrease the number of parameters. Data is generally limited and pruning of the network requires the use of empirical thresholds that are generally difficult to determine in the presence of noise. In this work we propose an alternative DL modelling approach where a priori knowledge about the metabolic network stoichiometry is used to constraint the structure of the neural network thus eliminating the need for empirical pruning methods.

This paper introduces Metabolic Graph Neural Networks (MGNN) as a novel approach for dynamic prediction of metabolites' concentrations and for detection and diagnosis of faults. In this special neural networks, the neurons contain information about metabolites and the network architecture is constrained based on a priori biological information about the network of metabolic reactions. The use of such constrained architecture is expected to lead to several benefits: i- overfitting of data will be avoided as compared to a fully interconnected network that does not use a priori information about the network of reactions, ii- by capitalizing on currently available powerful deep learning algorithms, this MGNN model is easier to tune as compared to other stocihiometry based approaches such as DFBA that require the solution of a challenging NLP, iii- the nonlinear form of the kinetic constraints in DFBA, e.g. Monod, Hill and other, must be decided a priori whereas MGNN does not require such choice , iv- the resulting MGNN can provide explainability since the neurons are associated to metabolites and thus the interconnection signals are associated to fluxes. The proposed methodology is tested on the oxidative stress metabolic network of B. Pertussis which is a bacteria used for the manufacturing of the whooping cough vaccine. The model is compared with a dynamic mechanistic model developed by Vitelli et al. (2023) for the same biochemical system.

2. PROPOSED METHODOLOGY

2.1 Graph Neural Networks

Processes involved a variety of phenomena that can be understood in terms of the relationships between the elements. A set of elements, and the connections between them, can naturally be expressed as a graph. In recent years, neural networks that operate on graph data (called graph neural networks, or GNNs) have been developed for a wide range of applications, such as antibacterial discovery Stokes et al. (2020) or for quantum chemistry Gilmer et al. (2017).

In the graph based modelling approach, the nodes of a graph represent objects or concepts, and the edges represent their relationships or interactions. For example, in the graph description of metabolic pathways, the nodes in the graph represent the metabolites and the connections between the nodes represent the reactions.

Typically, biochemical systems exhibit highly nonlinear behaviour due to the nonlinear kinetics regulating individual reactions among metabolites. On the other hand the connections between nodes in neural networks are also regulated by nonlinear activation functions but these can only represent simple nonlinear behaviour such as sigmoidal or semi-linear behaviour (Relu functions). Hence, single connections between two neurons can not capture the full nonlinear dependencies of the biochemical process. Instead, we propose in this work the use of one or more layers of neurons to connect between two neurons that are associated to individual metabolites. In addition to nonlinearity, the operation of bioprocesses in batch or perfusion modes, requires consideration of dynamic behaviour. To address the expected nonlinear dynamic behaviour a graph neural network model is proposed in this study as shown in Figure 1. This particular neural network is used for predicting state x_i at current time $k+1$ which is fed by those states which are connected to state x_i according to the graph edges at time k.

2.2 Oxidative Stress Metabolic Pathways

Oxidative stress metabolism in *Bordetella pertussis* fermentations involves a complex network of reactions occurring among various key molecules. At the heart of this process lies NADPH (Nicotinamide Adenine Dinucleotide Phosphate), a critical cofactor that plays a central role in countering the damaging effects of reactive oxygen species (ROS). When B. pertussis encounters ROS in the host environment, NADPH becomes crucial for maintaining redox balance and protecting the bacterium from oxidative harm. The motivation for focusing on the mechanism of oxidative stress is that it was found in previous studies by our group to be highly correlated with low productivity of antigens required in the formulation of the whooping cough vaccine.

NADPH serves as an essential reactant (co-factor) in various enzymatic reactions. It is involved in the reduction

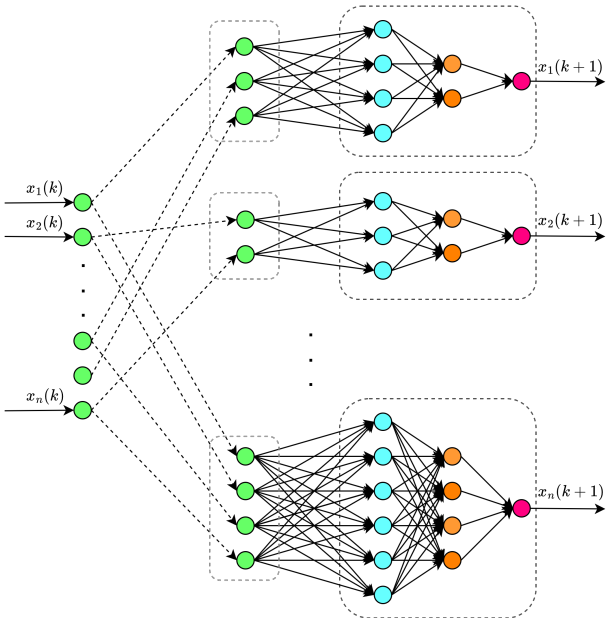

Fig. 1. A schematic of graph neural networks (GNNs).

of NADP+ (the oxidized form of NADPH) into its reduced form, NADPH, through specific enzymatic reactions, Vitelli et al. (2023). In this context, NADPH helps activate antioxidative enzymes like superoxide dismutase and catalase, which are crucial for neutralizing ROS, i.e. superoxide radicals and hydrogen peroxide into less harmful molecules, thus preventing cellular damage.

In response to oxidative stress, *Bordetella pertussis* shifts its metabolism to emphasize reactions related to generation and quenching of ROS. NADPH is a co-factor for anti-oxidative reactions catalyzed by superoxide dismutase and catalase. By diverting energy away from anabolic processes, e.g. cell growth, the bacterium uses NADPH to quench ROS.

This metabolic adaptation helps *Bordetella pertussis* strike a balance between growing and staying alive in the challenging host environment. By using NADPH and its antioxidative enzymes, the bacterium can thrive and successfully establish itself within the respiratory system despite the presence of harmful ROS.

2.3 Graph Neural Network for Oxidative Stress Metabolic Pathways

In this study, we have developed a MGNN model to represent the dynamic behavior of oxidative stress metabolic pathways. Figure 2 illustrates the pathways in the metabolic network of oxidative stress in *Bordetella pertussis* bacteria. Metabolites S_{ext}, S, A, B, C and A_{ext} represent extracellular glutamate, intracellular glutamate, NADPH, ROS, $NADP^+$ and extracellular NADPH respectively. A metabolic neural network is developed based on the presented metabolic network in order to predict the concentrations of metabolites over time. Also, a fully connected neural network is also developed for comparison with the metabolic neural network model in terms of performance and number of parameters. The training data have been synthetically generated using the mechanistic model developed by Vitelli et al. (2023) which was fully

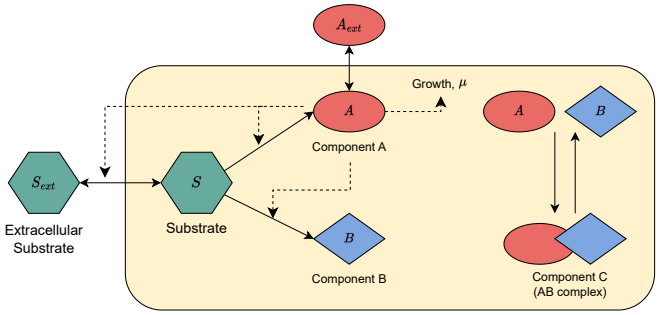

Fig. 2. An illustration of the metabolic pathways involved in oxidative stress.

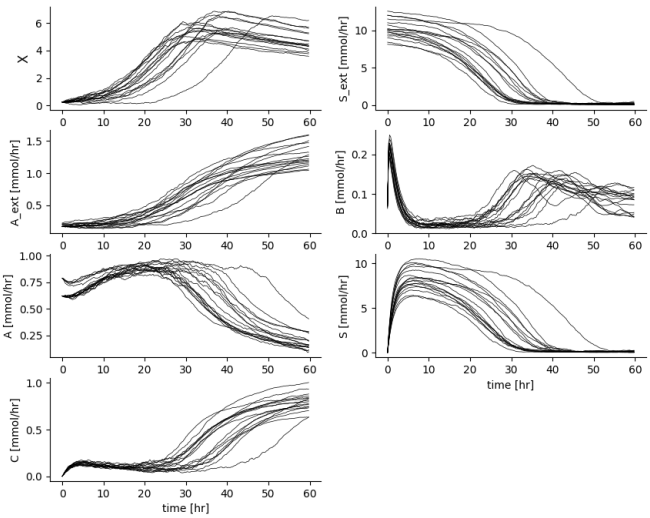

Fig. 3. Synthetic data produced by the mechanistic model developed by Vitelli et al. (2023).

trained and validated with experimental data. In practice, the proposed MGNN could be also directly trained with data. In this study, in silico data for 20 batches are generated by considering different initial conditions corresponding to perturbations around the experimental initial conditions that were used in the actual experiments . The data was split into 3 data sets to be used for training, validation and testing respectively. The training set is used to train the weights of the network for a particular set of hyper-parameters, the validation set is used to find a best set of hyperparameters and the testing set is used to test the accuracy with data that is not used for neither training or validation. Hence, the overall calibration of the model involves successive iterations between training and validation steps from which a best combination of weights and hyperparameters is found. In total, fifteen batches were used in the training process, two batches in the validation process, and three batches in the testing process. The total fermentation time of each batch is 60 hours (120 samples per batch) and includes the 6 metabolites presented above plus biomass. Figure 3 shows the in silico dynamic profiles of 6 metabolites and biomass concentrations obtained with the mechanistic model in Vitelli et al. (2023) during the *Bordetella pertussis* fermentation.

3. RESULTS AND DISCUSSION

Two different deep learning models were trained to predict metabolites' concentrations over time: i- the MGNN model, and ii- the associated fully connected neural network model. The hyperparameters considered for these models are: the number of layers, the number of units within each layer, and the learning rate. Using the validation data set, these parameters are tuned to minimize the loss function. The hyperparameter search is implemented using Python's Keras-tuner. To perform this search, a grid of hyperparameters is defined, for the number of hidden layers for correctly capturing nonlinear dynamics of metabolites' behavior = [1, 2, 3], number of neurons for each layer = [1, 2, 4, 8], and learning rate = [0.1, 0.01, 0.001, 0.0001]. Figure 4 illustrates the architecture of the metabolic graph neural network model that was developed for predicting the concentrations of different metabolites during oxidative stress in *Bordetella pertussis* fermentation. Figures 5 and 6 compare the dynamics of metabolite concentrations between the fully connected neural network and the mechanistic model in a training and in a testing batch respectively. As depicted in these figures, the fully connected neural network demonstrates a good fit for the training batch but a notably poor fit for the testing batch, indicating an overfitting issue in the fully connected neural network model. As previously mentioned, overfitting is caused by the high number of parameters in the fully connected neural network. However, Figure 7 exhibits a very good fit for the testing batch, addressing the overfitting issue due to the model's significantly lower number of parameters. A comparison between the MGNN model and the corresponding fully connected neural network model in terms of the mean square error is shown in Table 1.

Physical explainability (interpretability) of the network is a desired property of deep learning models. Also, the proposed MGNN model has the ability to quantify the effect of input variables on the outputs by back-propagation operations through the network. To interpret the MGNN we consider the gradients of the time derivative of each metabolite's concentration with respect to each metabolite's concentration that can be directly obtained from the network from the difference between the metabolite's concentrations at two consecutive times k and k+1. The absolute values of the gradients are averaged over the fermentation duration to quantify the average impact of each metabolite on any other metabolite. Also, if the signs of the gradients are considered, the actual positive or negative correlation among metabolites can be inferred.

Figure 8 illustrates the absolute value of the impact of each metabolite on each other metabolite in the metabolic network. The figure accurately demonstrates that metabolites not connected to a specific metabolite in the metabolic graph exhibit no contributions in the change of that specific metabolite. Figure 9 also illustrate the positive or negative impact of various metabolites on the change of a particular metabolite.

For example, the first row of Figures 8 and 9, shows a positive impact from metabolites X (biomass), A (NADPH), and S (glutamate) on the change in biomass concentration which aligns with the mechanistic ODE equation developed for the dynamics of this metabolite:

$$\begin{cases} \dfrac{dX}{dt} = (\mu - D)X \\ \mu = \alpha F_A A \end{cases} \tag{1}$$

For A_{ext} (NADPH extracellular), Figure 9 displays a positive effect from X and A, in accordance with the ODE equation below:

$$\frac{dA_{ext}}{dt} = (K_m + D)AX \tag{2}$$

Eq. 2, indicates that the change dA_{ext} has a direct positive correlation with variables A and X. Correspondingly, according to the metabolic graph, A_{ext} either converts into intracellular A or accumulates outside the cell, aligning with the negative influence of A_{ext} on its own alteration, as determined by the MGNN model.

Similarly, Figure 8 and 9 shows positive influences of metabolites A, B, C, and X, as well and negative influence of S on the change of metabolite A (NADPH). These observations can be verified with the mechanistic ODE equation describing the concentration dynamics of this metabolite:

$$\frac{dA}{dt} = F_A A - k_p AB + k_m C - \mu A - K_m A \tag{3}$$

A (NADPH), S (glutamate), C ($NADP^+$), and X (biomass) contribute to the increase in B, while the consumption of metabolite B in producing C and degradation of B result in its decrease.

$$\frac{dB}{dt} = F_B A - k_p AB + k_m C - \mu B - d_B B \tag{4}$$

Regarding metabolite C ($NADP^+$), Figures 8 and 9 depict a positive influence from metabolites A, B, C, and S, and a negative influence from X. This observation can be again verified with the mechanistic model concerning the dynamics of this metabolite.

$$\begin{cases} \dfrac{dC}{dt} = k_p AB - k_m C - \mu C \\ F_A = \dfrac{(\frac{v}{1+S/K_{si}})S}{KX + S}(\dfrac{S}{K_t + S})^{1.5} \\ \mu = \alpha F_A A \end{cases} \tag{5}$$

Regarding metabolite S_{ext} (extracellular glutamate), only the intracellular glutamate (S) has the potential to positively impact the increase in S_{ext} concentration, as it can only be transported to the exterior of the cell. The concentrations of metabolites S_{ext}, X, and A at the previous time lead to a decrease in the concentration of S_{ext}. This can also be corroborated from the metabolic graph (Figure 2) and from the mechanistic model (Eq. 6):

$$\frac{dS_{ext}}{dt} = -\frac{\mu X}{Y_{X/S}} \tag{6}$$

Furthermore, the mechanistic model ODE (Eq. 7) confirms that the metabolites X, S, and A have positive impact and metabolite S at the previous time has a negative impact on the change of S.

$$\frac{dS}{dt} = -F_A A - F_B A + \alpha A(S_{ext} - S) - \mu S \tag{7}$$

An interesting application of the MGNN model is for fault detection. To assess this capability, a change (fault) was intentionally introduced into the media (S_{ext}) during the

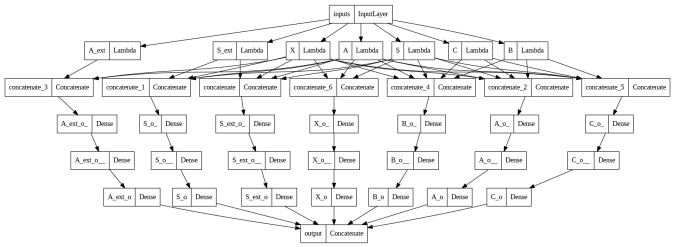

Fig. 4. Architecture of the metabolic graph neural network for the oxidative stress metabolic network.

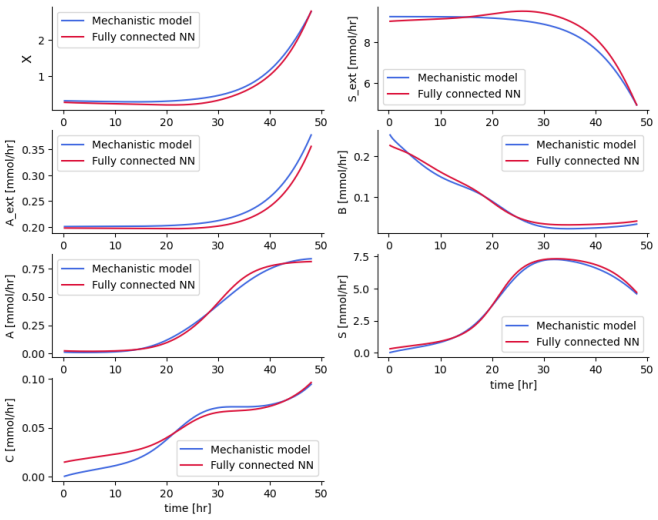

Fig. 5. Comparison of the predicted concentrations of metabolites over time in a training batch using the developed fully connected neural network with the mechanistic model.

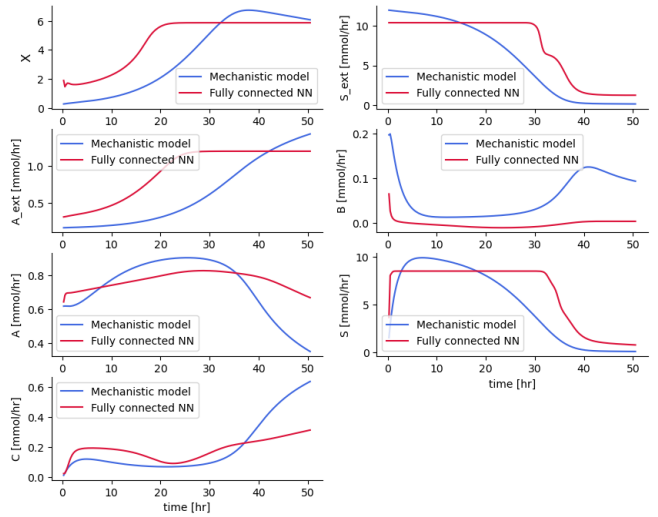

Fig. 6. Comparison of the predicted concentrations of metabolites over time in a testing batch using the developed fully connected neural network with the mechanistic model.

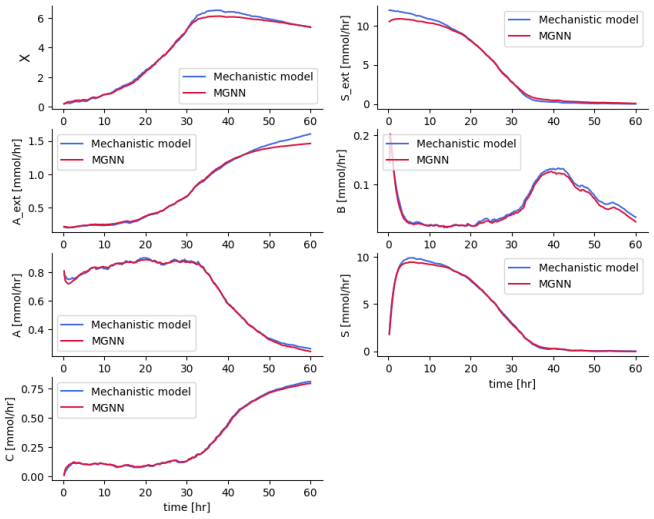

Fig. 7. Comparison of the predicted concentrations of metabolites over time in a testing batch using the developed Metabolic Graph Neural Network (MGNN) with the mechanistic model.

Table 1. Comparison of the Metabolic Neural Network model (MGNN) and the fully connected neural network (FCNN).

	MGNN	FCNN
MSE in the training dataset	3.6157e-04	8.2261e-05
MSE in the testing dataset	4.1534e-04	1.379e-01
Number of parameters	575	2247
Overfitting problem	No	Yes

Fig. 8. The relevance of metabolites at the previous time to the change of metabolites at the current time using the MGNN model.

time interval $t = [20, 26]$ by increasing the concentration of this metabolite by 10. Based on an online fluorescence probe currently available in our lab, we assumed that the concentrations of biomass (X) and extracellular NADPH (A_{ext}) can be measured online. The MGNN model, initialized with the same conditions as normal batches, was simulated while forcing the predicted values of biomass (X) and extracellular NADPH (A_{ext}) to be equal to the

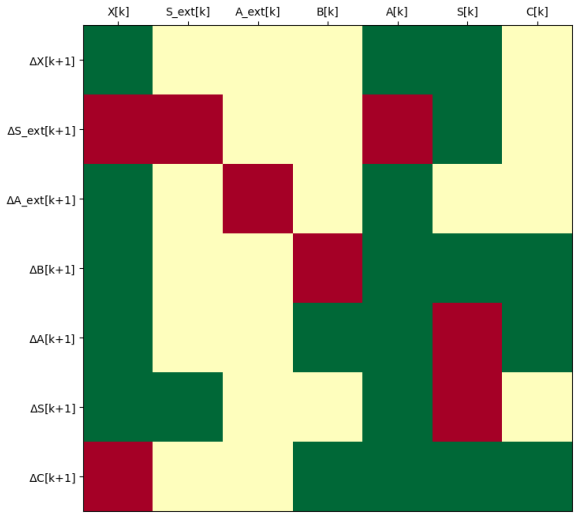

Fig. 9. The positive or negative influence of each input on the output direction; Green (Positive), Red (Negative), Yellow (Neutral)

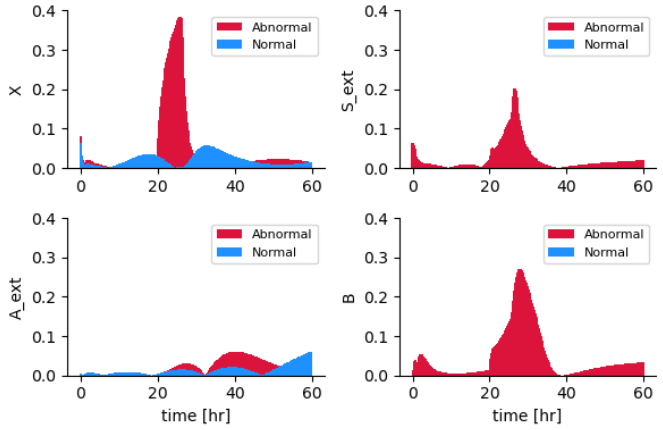

Fig. 10. The contribution of different metabolites to detect the imposed fault in the media (S_{ext}) in $t = [20, 26]$ compared to their contributions in normal media.

measured values. Using the contribution plots' algorithms developed in Aghaee et al. (2023), the difference between the response of the MGNN model for the faulty batch and the average response for normal batches was back-propagated through the MGNN model to determine the contribution of different metabolites, as illustrated in Figure 10. These plots clearly show the detection of the change in nutrient concentration and its impact on the other metabolites.

4. CONCLUSION

This study proposes a novel Metabolic Graph Neural Network (MGNN) as a modelling tool for simulating the dynamic behavior of metabolites within a metabolic network. By incorporating prior knowledge of metabolic graph pathways, a neural network model with a significantly lower number of parameters is obtained thus leading to less overfitting of data as compared to a fully connected NN that does not use prior metabolism knowledge. The developed MGNN model results in significantly better accuracy as compared to the fully connected in terms of the mean square errors.

Furthermore, the MGNN model is highly explainable since it comprises multiple sub-neural network models, each representing the dynamics of a specific metabolite. In contrast, the fully connected neural network model cannot be physically interpreted.

Moreover, the methodology can be used to generate a relevance matrix that reveals the influence of different metabolites on the change of a specific metabolite. The relevance matrix correctly predicts (positive or negative) the correlations among metabolites as verified by the mechanistic model that was used to generate the in silico data and by the metabolic graph associated with oxidative stress phenomena. The derivation of MGNN is found to be significantly faster than the development of a mechanistic model since the MGNN. While the mechanistic model requires a priori decisions regarding the kinetic expressions relating different metabolites, the MGNN does not require such a priori assumptions regarding the forms of the nonlinear kinetic terms relating the metabolites thus simplifying the development of the model. The proposed model is also shown to be useful for fault detection applications.

ACKNOWLEDGEMENTS

This work is the result of the research project supported by MITACS Grant IT16479 through MITACS-Accelerate Program and Sanofi, Toronto.

REFERENCES

Aghaee, M., Krau, S., Tamer, M., and Budman, H. (2023). Unsupervised fault detection of pharmaceutical processes using long short-term memory autoencoders. *Industrial & Engineering Chemistry Research*.

Costello, Z. and Martin, H.G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ systems biology and applications*, 4(1), 1–14.

Du, Y., Duever, T.A., and Budman, H. (2015). Fault detection and diagnosis with parametric uncertainty using generalized polynomial chaos. *Computers & Chemical Engineering*, 76, 63–75.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., and Dahl, G.E. (2017). Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.

Lei, F., Rotbøll, M., and Jørgensen, S.B. (2001). A biochemically structured model for saccharomyces cerevisiae. *Journal of biotechnology*, 88(3), 205–221.

Mahadevan, R., Edwards, J.S., and Doyle, F.J. (2002). Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical journal*, 83(3), 1331–1340.

Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4), 688–702.

Vitelli, M., Tamer, I.M., Pritzker, M., and Budman, H. (2023). Modeling the effect of oxidative stress on bordetella pertussis fermentations. *Biotechnology Progress*, e3335.