METABOLICALLY CONSTRAINED NEURAL NETWORKS FOR BIOPROCESS OPTIMIZATION

Remy Kusters
Gourmey (SUPRÊME)
91058 Évry-Courcouronnes, France
remy@gourmey.com

In this work, we present a neural network-based model that predicts growth kinetics by incorporating *multiple* biochemical constraints into the loss function of the neural network, in particular material (carbon and nitrogen) balance and metabolic flux balance. These constraints effectively reduce the data requirements, narrow the solution spaces to biologically feasible solutions, and enhance interpretability. Additionally, it enables the inference of critical, yet often difficult-to-measure, bioprocess parameters such as gas consumption rates and biomass composition. We demonstrate the application of this approach through a case study on E. coli cell culture. Finally, we discuss potential extensions of this approach, emphasizing how the application of multiple biological constraints can serve as a foundation for a multilevel framework in AI-driven virtual cell models, enabling interpretable and biologically grounded predictions Bunne et al. (2024).

Data generation: To demonstrate this approach we generate synthetic data by performing Flux Balance Analysis (FBA) simulations of E. coli cultures using the CORBApy package Ebrahim et al. (2013). O_2, glucose, glutamine, glutamate, and lactate are used as substrates to generate biomass, along with CO_2 and ammonia as by-products ($\mathrm{e_coli_core}$ dataset: 72 metabolites, 95 reactions). A dataset is generated which optimizes the instantaneous biomass production rate, q_{bm}, as a function of metabolite and gas consumption rates, q_i, under varying constrained media compositions (Fig. 1a). To reflect biological variability, 5% of white noise is added to (q_{bm}, q_i).

Predicting the biomass composition and CO_2 production: We model the instantaneous biomass production rate, q_{bm}, as a function of specific metabolite and gas consumption rates, q_i, using a simple feed-forward neural network: $\bar{q}_{bm} = f(q_i)$. To train the model under material balance constraints (carbon and nitrogen), we define a constrained loss function Raissi et al. (2019),

$$\mathcal{L} = \mathrm{mse}\left(\bar{\mathrm{q}}_{\mathrm{bm}}, \mathrm{q}_{\mathrm{bm}}\right) + \lambda_1 \mathcal{L}_{\mathrm{c}}^2 + \lambda_2 \mathcal{L}_{\mathrm{n}}^2. \tag{1}$$

Besides the mse, the loss function contains a carbon $\mathcal{L}_c = \sum a_i q_i - q_{co2} - \alpha_{bm} q_{bm}$ and nitrogen $\mathcal{L}_n = \sum b_i q_i - q_{nh4^+} - \alpha_{bm} q_{bm}$ material balance constraint. Here, a_i and b_i are the respective carbon and nitrogen numbers, and α_{bm}, β_{bm} the (unknown) biomass composition's carbon and nitrogen numbers. λ_1 and λ_2 are model parameters that regulate the magnitude of the constraint and are set to 10^{-4} throughout the rest of the paper. When the biomass composition (α_{bm}, β_{bm}) and the CO_2 production (q_{co2}) are know, training the model to predict the biomass production rates, q_{bm}, using Eq. 1, constrains the solution space for $\bar{q}_{bm}$ to solutions permitted under the carbon and nitrogen balance.

When the relative carbon, α_{bm}, and nitrogen, β_{bm}, content are unknown, the model can *learn* these parameters, alongside the neural network parameters to predict $\bar{q}_{bm}$, by defining them as trainable parameters Raissi et al. (2019); Lagergren et al. (2020). As demonstration, this task has been performed on the aforementioned e-coli data-set and, with as few as 100 training samples, the values for $(\alpha_{bm}, \beta_{bm})$ were recovered with an error inferior to 2% [1]. These values provide valuable insights in the biomass composition, an important process and quality control parameter.

In many experimental settings, the CO2 production rate, q_{co2}, is often challenging to measure directly due to factors like pH buffers. By including this as hidden variable $\bar{q}_{co2}(q_i, q_{bm})$ directly in the architecture (Fig. 1b) it can be learned alongside the prediction of $\bar{q}_{bm}$ Tartakovsky et al. (2018). Using the previously utilized dataset, but now with q_{co2} treated as unknown, allows for its prediction while ensuring compliance with the enforced carbon and nitrogen constraints. In Fig. 1(d) we

[1] Throughout this paper work, we use a two layer FFNN for $f(q_i)$ (64 neurons), trained with SGD, with default hyper-parameters.

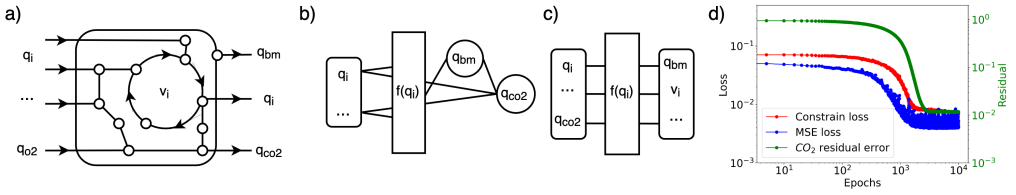

Figure 1: a) Metabolic flux model and two network architectures covered: b) with hidden variable prediction and c) including the learned internal fluxes. d) MSE and Constrain loss, as well as the relative residual error in the hidden CO_2 production rate (compared to ground truth data).

display the mse and constrain loss (Eq. 1) during training as well as the relative error in the the CO_2 production rate $|q_{co2} - \bar{q}_{co2}|/q_{co2}$. This shows that both the biomass production as well as the CO_2 production rate can be learned simultaneously.

Learning Internal Fluxes: Beyond material balance constraints, more complex constraints, such as the stoichiometric relationships governing metabolic flux balance can be imposed to i) uncover internal metabolic fluxes and ii) further constrain the biomass production rate $\bar{q}_{bm}$. For each internal metabolic node, incoming fluxes equal outgoing fluxes (See Fig 1a and c), represented by $S \cdot \bar{v} = 0$, where S is the stoichiometric matrix and $\bar{v}$ the *learned* internal flux vectors. Adding this supplemental constraint to the loss function,

$$\mathcal{L} = \mathrm{mse}\left(\bar{\mathrm{q}}_{\mathrm{bm}}, \mathrm{q}_{\mathrm{bm}}\right) + \lambda_1 \mathcal{L}_{\mathrm{c}}^2 + \lambda_2 \mathcal{L}_{\mathrm{n}}^2 + \lambda_3 |\mathrm{S} \cdot \mathrm{v}|^2, \tag{2}$$

enforces internal consistency of the metabolic fluxes Faure et al. (2023). Using the data-set of the previous paragraph and including specific consumption rates of all relevant amino acids, q_i, as input, we successfully infer the internal fluxes, $\bar{v}_i$, coherent with the stoichiometric constraint imposed, $S \cdot v < 10^{-3}$. Uncovering these internal fluxes $\bar{v}$ permits to identify metabolic limitations and bottlenecks, informing further process and media optimization.

Discussion: The proposed framework incorporates material and flux balance constraints into neural networks and serves a dual purpose: (i) Narrowing the solution space to biologically plausible regions and (ii) facilitating the identification of hidden variables and parameters. Key design choices, such as neural network architecture and loss function formulation, play an important role in improving generalization beyond the training distribution Xu et al. (2020). Further refinements—such as specialized network architectures tailored for metabolic dependencies, or introducing uncertainty quantification Yang et al. (2021) could enhance model extrapolation and predictive accuracy further.

Future extensions will explicitly incorporate the temporal dynamics of the cell culture to capture metabolic changes in cellular growth under varying nutrient availability, leveraging the full potential of Physics-Informed Neural Networks (PINNs) Raissi et al. (2019). Additionally, Integrating further levels of domain-specific knowledge, such as gene-regulatory interactions Fortelny & Bock (2020) offers further potential to enhance interpretability and biological relevance. These advancements would make the framework a more comprehensive and scalable solution for optimizing industrial bioprocesses through metabolic constraints.

Meaningfulness Statement

Life, at its core, is governed by a complex interplay of biochemical reactions, resource allocation, and regulatory mechanisms to sustain essential functions such as metabolism and proliferation. To capture these dynamics with predictive machine learning, we must constrain these models by incorporate "known" fundamental biological principles.

References

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. Cobrapy: constraints-based reconstruction and analysis for python. *BMC systems biology*, 7:1–6, 2013.

Léon Faure, Bastien Mollet, Wolfram Liebermeister, and Jean-Loup Faulon. A neural-mechanistic hybrid approach improving the predictive power of genome-scale metabolic models. *Nature Communications*, 14(1):4669, 2023.

Nikolaus Fortelny and Christoph Bock. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome biology*, 21:1–36, 2020.

John H Lagergren, John T Nardini, Ruth E Baker, Matthew J Simpson, and Kevin B Flores. Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. *PLoS computational biology*, 16(12):e1008462, 2020.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Alexandre M Tartakovsky, Carlos Ortiz Marrero, Paris Perdikaris, Guzel D Tartakovsky, and David Barajas-Solano. Learning parameters and constitutive relationships with physics informed deep neural networks. *arXiv preprint arXiv:1808.03398*, 2018.

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*, 2020.

Liu Yang, Xuhui Meng, and George Em Karniadakis. B-pinns: Bayesian physics-informed neural networks for forward and inverse pde problems with noisy data. *Journal of Computational Physics*, 425:109913, 2021.