Bilevel Optimisation for Targeted Metabolic Network Reduction

Mariana Monteiro * **Fengqi You** ** **Cleo Kontoravdi** *

* *Sargent Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom*
** *Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York, 14853, USA*

Abstract: Metabolic network models are powerful tools for understanding cellular functions and guiding biotechnological applications. Yet, the complexity of these models poses challenges in accurately predicting intracellular flux distributions, considering limited measurement availability. To address this, we propose a novel bilevel optimisation framework to metabolic network reduction using Bayesian optimisation. Our method assigns continuous probability values to reactions and iteratively refines a reduced model that balances network simplification with predictive performance. The upper-level Bayesian optimisation process selects reaction removal probabilities, while the lower level evaluates model feasibility and performance through flux sampling. A Gaussian Process surrogate is trained to approximate the impact of reaction removals on model accuracy, guiding the optimisation toward a minimal yet representative network. We applied our methodology to a Chinese Hamster Ovary (CHO) cell metabolic model using multiple datasets, demonstrating its ability to adapt to different datasets and suggest targeted measurements. By unifying lumping and sensitivity analysis concepts in a data-driven framework, our approach systematically simplifies metabolic models, increasing their applicability in both development and manufacturing processes.

Keywords: Systems biology, synthetic biology, metabolic flux modeling; Biopharmaceutical processes

1. INTRODUCTION

Metabolic network models are a community resource that supports understanding metabolic function, identifying cell engineering targets and guiding the optimisation of biotechnological applications (Fouladiha et al., 2019; Schinn et al., 2021; Kol et al., 2020). Metabolic models describe the reactions that occur inside the cell in either a lumped or detailed form and may include other pathways such as protein synthesis, folding and secretion. The number of reactions in metabolic models can range from a few dozen to a few thousand, with the largest models that are based on detailed gene-protein-reaction networks being genome-scale metabolic models (Strain et al., 2023a). Industrial and academic researchers do not routinely measure intracellular fluxes, as that would require intricate experimental setups. Instead, extracellular metabolite concentrations are typically measured instead and then used to determine the exchange flux rates. This means that the number of unknown fluxes typically far exceeds the number of measurements, leading to an underdetermined system of equations that is solved via optimisation. The most common methodology for this is flux balance analysis (FBA) (Orth et al., 2010), where a cellular objective is assumed to generate a set of steady-state flux predictions.

There are several sources of uncertainty in this activity, one of which is the model network itself, the choice of which greatly influences modeling outputs (Bernstein et al., 2021). Furthermore, overly complex metabolic models may lead to solutions with high uncertainty, while simplified models may overlook insights from the data. There have been several efforts towards reducing reaction network size, either in biochemical networks described parametrically (Danø et al., 2006; Snowden et al., 2017) or via the introduction of constraints (Erdrich et al., 2015; Ataman and Hatzimanikatis, 2017; Ataman et al., 2017). Regarding the former, literature divides model reduction techniques into three types: lumping, sensitivity analysis and time-scale based Okino and Mavrovouniotis (1998); Danø et al. (2006); Snowden et al. (2017); Radulescu et al. (2012). Lumping involves aggregating kinetic expressions to lower the dimensions of the system by decreasing the number of species to include in the model. Sensitivity analysis includes perturbing the biochemical system and eliminating sections of the network that contribute less to the overall system. Time-scale-based methods identify fast and slow reactions, which simplifies the dynamical system via the application of the quasi-steady-state assumption.

There is no superior method for reducing network and the choice of method should be determined both by the complexity and the objective of the reduction (Danø et al., 2006; Snowden et al., 2017). For example, some methods require the assumption of a fully parameterizable initial model, which in genome-scale models is currently impossible. Hence, in constraint-based models, researchers

have employed other types of techniques such as pruning based on reactions with lower fluxes (Erdrich et al., 2015), identifying and lumping the minimal size subnetworks that produce each key component of the biomass equation (Ataman and Hatzimanikatis, 2017), deriving a reduced stoichiometric matrix via graph search that makes exclusions based on pre-determined rules (Ataman et al., 2017) and automating reduction methods for faster reduced model creation (van Rosmalen et al., 2021). The limitations of these approaches include requiring a user-defined initial model and having an imposed length of subsystem connections, leading users to decide between more rigid or more flexible model reduction techniques (Singh and Lercher, 2019). While most studies focus on structural reduction or sensitivity analysis in isolation, there lacks an approach that could unify these strategies while still allowing for flexibility to adapt to new data.

A potential approach to address the limitations of reduction techniques is via bilevel optimisation. Bilevel optimisation, largely studied in Process Systems Engineering, specifically for hierarchical supply chain planning and scheduling problems. These problems are typically expensive to solve directly as they become intractable in real-life examples Chu et al. (2015). Hence, there has been a lot of literature proposing alternatives to bypass this hurdle, such as, developing bilevel decomposition algorithms, heuristics or replacing the lower level problem by a cheaper model, surrogate to the original one Erdirik-Dogan and Grossmann (2008); You et al. (2011); Chu et al. (2015).

This study proposes a framework for iteratively reducing the size of metabolic networks given available data. The proposed framework is applied to a Chinese Hamster Ovary (CHO) cell metabolic model and tested with experimental datasets from carbon labeling experiments. Our approach is inspired by both lumping and sensitivity analysis methods, as it suggests which reactions could be removed (or lumped) via exploration and exploitation of different reaction removal combinations. Furthermore, it allows for adaptation to new data, bridging a gap between structural and functional reduction techniques.

The contributions of this work are as follows: (1) we introduce a bilevel approach for the iterative reduction of metabolic network models, adaptable to data; (2) we aggregate structural reduction with sensitivity analysis and make use of that for the decision which reaction to remove; (3) by leveraging a surrogate model for the lower-level, our framework is amenable to any initial network model size; and (4) we demonstrate the applicability of our framework for different CHO cell culture datasets. Our proposed framework can be used at the process development and manufacturing levels, when deciding which model to use given measurement availability, but also at the metabolic model curation stage, as it can consider different gene annotations in the search space.

2. METHODOLOGY

2.1 Framework Overview

Our approach, shown in the following figure 1, is bilevel. The two levels comprise the following:

(1) Upper-Level Problem: Iterative reduction of the metabolic network via the assignment of a continuous probability value (between 0 and 1) to each unmeasured reaction. The exploration/exploitation is conducted through Bayesian Optimisation.
(2) Lower-Level Problem: Checking the feasibility of the metabolic network model followed by conducting flux sampling in order to evaluate how well the new model performs. The performance is then used as feedback to refine the upper-level decisions.

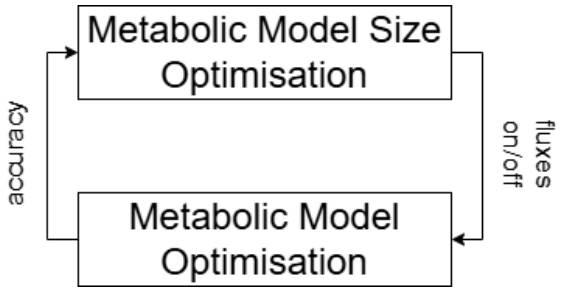


Fig. 1. Bilevel Approach. The upper level iteratively reduces the metabolic network size by assigning reaction removal probabilities to each reaction. The lower level evaluates the feasibility and performance of the reduced model from the upper level. The stopping criteria was the number of iterations (250).

As shown in figure 1, solving the metabolic model requires optimisation. This means that this bilevel approach contains an optimisation in both layers. Solving the bilevel optimisation problem directly was considered intractable. As such, the metabolic network optimisation is replaced by a Gaussian process, which can be sampled, rather than optimised. The proposed framework to solve the bilevel problem is shown in the following figure 2. It is composed of a training and a validation phase. In the training phase, we define a train a Gaussian Process (GP) surrogate that predicts the impact of reaction removals on the model performance. This process includes performing kernel selection, kernel hyperparameter optimisation and model training. We use Bayesian Optimisation to iteratively refine the best set of reactions to be removed from the model, using the GP. This process includes allowing the optimisation to suggest an set of reaction removal probabilities, checking for feasibility of the model with FBA, flux sampling the new model, calculating the error between predicted and experimental fluxes and iteratively updating the model with the lowest error. Each of these major steps will be explained in detail in the following subsections.

2.2 Outer Loop: Training Phase

The training process consists of iterating through each of the multiple datasets (6 in total, 80% used for training), and for each identifying the set of reaction probabilities with the lowest loss function. The final reaction probabilities set will be the lowest of all of the datasets.

2.3 Inner Loop: Bayesian Optimisation

For the inner loop, a custom Bayesian Optimisation process was setup. We define a Gaussian Process (GP) as a

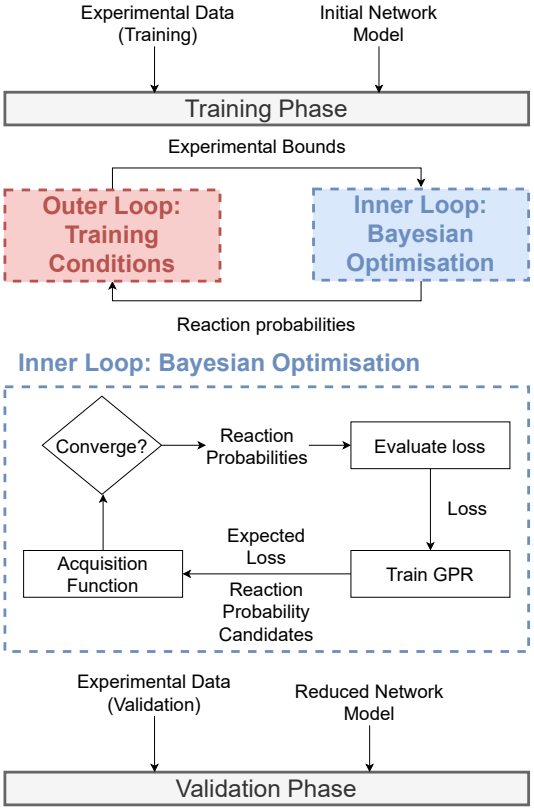


Fig. 2. General framework

surrogate model that relates the probability of reaction removal with metabolic model performance. A custom acquisition function is defined as:

$$a(x) = \mu(x) - 0.01\sigma(x) + 0.1(\frac{1}{R}\sum_{i=1}^{R} x_i) \quad (1)$$

where $\mu(x)$ is the predicted mean loss for reaction probabilities, $\sigma(x)$ is the predicted standard deviation, R the number of reactions and the last term is a regularization term to prevent removal of a lot of reactions. This acquisition function aims to balance exploration and exploiting, while encouraging a sensible network sized model. The optimisation of the acquisition function uses scipy minimize routine (Virtanen et al., 2020), specifically the L-BFGS-B optimiser.

The GP surrogate is trained by fitting a Gaussian Process Regressor to the training data. For efficiency reasons, only the first dataset is used to for model selection as well as train the kernel's hyperparameters. Each kernel is evaluated using k-fold cross validation, by computing the negative log-marginal likelihood. The kernel with the lowest score is selected and used for the subsequent iterations.

Loss Function

The loss function is composed of four main steps:

(1) Reaction Removal based on Probabilities: Reactions are removed if their probabilities are below a user-defined threshold.
(2) Feasibility Check: A standard FBA Orth et al. (2010) is run to check that a feasible solution exists. If not, the the irreducible infeasible subsystem of the problem is calculated to define the constraints that are infeasible (Gurobi Optimization, LLC, 2024). Those constraints are relaxed, and this iteration is repeated until the problem has a solution or a maximum of 50 iterations is achieved.
(3) Flux Sampling of 5000 samples is conducted (Saunders et al., 2019).
(4) Calculation of Total Loss (explained below).

The total loss is computed as follows:

$$\begin{aligned}\text{Total Loss} = \alpha \times \text{WMAE} + \beta \times \text{Flux Variability} \\ + \gamma \times \text{Sparsity Penalty}\end{aligned} \quad (2)$$

The Weight Mean Absolute Error (WMAE) is defined as the difference between the average sampled fluxes (from flux sampling) and experimental fluxes (the mean values of the datasets), normalized by the standard deviation (assumed to be half the gap between lower and upper bounds of the experimental values). N stands for the number of measured reactions.

$$\text{WMAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|v_i^{\text{sample}} - v_i^{\text{exp}}\right|}{\sigma_i^{\text{exp}}} \quad (3)$$

Flux Variability represents the standard deviation across samples:

$$\text{Flux Variability} = \frac{1}{N}\sum_{i=1}^{N}\sigma\left(v_i^{\text{samples}}\right) \quad (4)$$

The Sparsity Penalty acts like a regularization term that discourages removing too many reactions, as it scales with the sum of $1 - \textit{reaction probability}$, represented by p_i.

$$\text{Sparsity Penalty} = \frac{1}{R}\sum_{i=1}^{R}(1 - p_i) \quad (5)$$

α, β and γ represent the importance of each of different parts that compose the loss, are were given the values of 1.0, 0.1, 0.01.

2.4 Validation Phase

The validation phase consisted in sampling the model, applying the reaction removals obtained in the training phase and resampling the model, with the experimental bounds of the validation datasets.

2.5 Experimental Data and Pre-Processing

The applicability of the framework is demonstrated using published data from carbon labeling Chinese hamster ovary (CHO) cell culture experiments presented in various

studies and summarised in Strain et al. (2023b). The data comprises uptake and secretion rates for CHO cells grown under various conditions, as well as intracellular fluxes estimated from ^{13}C labelling. Two cases were considered, with two different thresholds and different training and validation datasets. Table 1 summarizes the setup. The

Threshold	0.4	0.6
Training	SVGS	SV
	SVM1	SVM1
	SVM2	SVM2
	SVM4	SVM3
Validation	SV	SVGS
	SVM3	SVM4

Table 1. Training and Validation Cases for Different Thresholds

base model used for reduction is CHOmpact (Jiménez del Val et al., 2023). It contains 144 reactions and 156 metabolites. This model was chosen for its relevance to CHO cell metabolism, making it an ideal starting point for applying the iterative reduction methodology. The framework is modelled in Python 3.10. We used an Intel Core i7 CPU with 6 Cores and 12 Logical Processors, and Microsoft Windows Pro as our operating system.

3. RESULTS & DISCUSSION

3.1 Hyperparameter Tuning of Gaussian Process Model

The two Gaussian Process (GP) models have a similar structure but differ in their kernel hyperparameters. Both models use a kernel of the form:

$$k(x,x') = \sigma_1^2 \cdot \text{Matern}(\ell_1,\nu) + \sigma_2^2 \cdot \text{RationalQuadratic}(\alpha,\ell_2) + \sigma_n^2 \cdot \text{WhiteKernel}. \quad (6)$$

where:

$$\text{Matern}(\ell,\nu) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}\|x-x'\|}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\|x-x'\|}{\ell}\right),$$

$$\text{RationalQuadratic}(\alpha,\ell) = \left(1+\frac{\|x-x'\|^2}{2\alpha\ell^2}\right)^{-\alpha},$$

$$\text{WhiteKernel} = \sigma_n^2\delta(x,x').$$

where the Matern and Rational Quadratic kernels capture different smoothness properties of the function, and the White Kernel accounts for noise.

3.2 Reduced Model

Figure 3 shows the reaction probabilities for both thresholds. The probability of each reaction is computed in the inner loop of the methodology presented in 2 based on whether its inclusion improves model accuracy when compared to experimentally determined flux values. A threshold of 0.6 results in a model with 64 reactions, whereas a threshold of 0.4 results in a model with 84 reactions. This is to be expected, as a lower threshold leads to more reactions being retained in the model.

Kernel Parameters	**Threshold 0.6**	**Threshold 0.4**
Matern		
Length Scale (ℓ)	1.54×10^{-5}	0.138
Smoothness (ν)	1.5	1.5
Rational Quadratic		
Alpha (α)	0.0342	1.0×10^{5}
Length Scale (ℓ)	1.0×10^{-5}	3.52
White	1.0×10^{7}	1.0×10^{7}

Table 2. Comparison of Gaussian Process (GP) Models

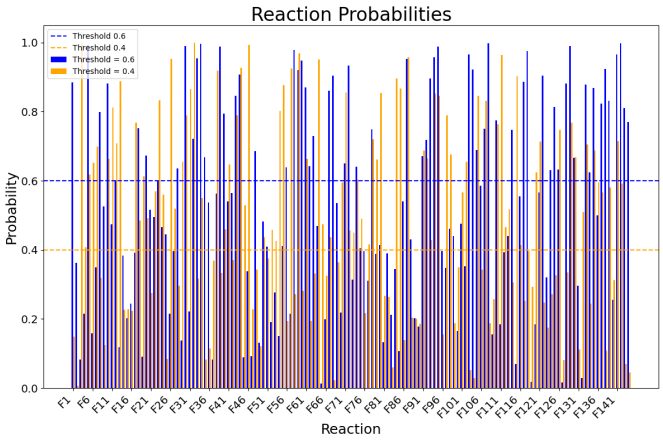


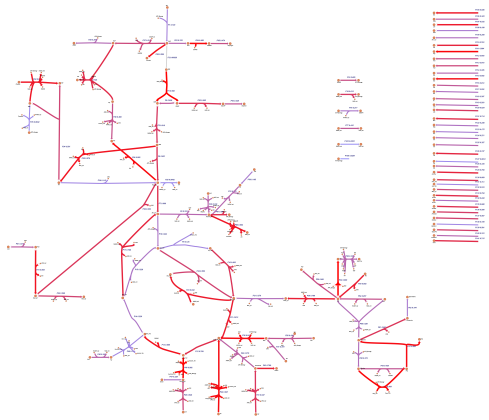
Fig. 3. Reaction Probabilities for threshold of 0.6 and 0.4.

Figure 4 depicts the metabolic network with the reaction probabilities for the two different threshold cases. Darker and thicker lines have a higher reaction probability while grey lines exhibit lower reaction probabilities. The higher threshold model appears to retain stronger flux across major pathways, with fewer low-flux pathways being present, whereas the lower threshold model allows for a broader range of fluxes, given that it eliminated less reactions.

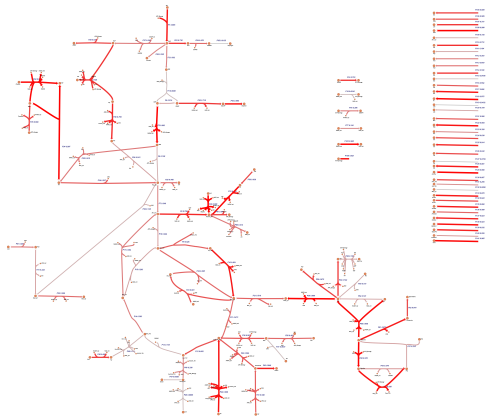
Some reactions are present in both reduced models and in the list of measured reactions, including important exchange reactions for glucose and lactate, and biomass synthesis. When moving to a higher threshold, there are reactions that are initially excluded, such as certain glycolytic fluxes (F3, F4 and F6). Instead and counterintuitively, carbon flux is channelled through the pentose phosphate pathway. The opposite also occurs; the product flux (F143) has a very high probability of retention in the higher threshold (0.81) and a very low probability in the lower threshold (0.07). This might be due to the fluxes being more sparse in the lower threshold, which may cause the flux leading to the product precursors being insufficient to justify keeping that particular flux (figure 4). We therefore proceed to enforce the retention of measured reactions *a posteriori*.

3.3 Performance

The original model was reduced according to the probability values; however, the measured reactions were always included regardless of their probability value. This rule was imposed *a posteriori* to give the optimiser maximum flexibility. After reduction, both the initial and the reduced models were sampled. Similarly to what was described in



(a) Threshold = 0.4



(b) Threshold = 0.6

Fig. 4. Reduced network model

section 2.3, after applying the bounds, a feasibility check was conducted and bounds were selectively relaxed based on its results.

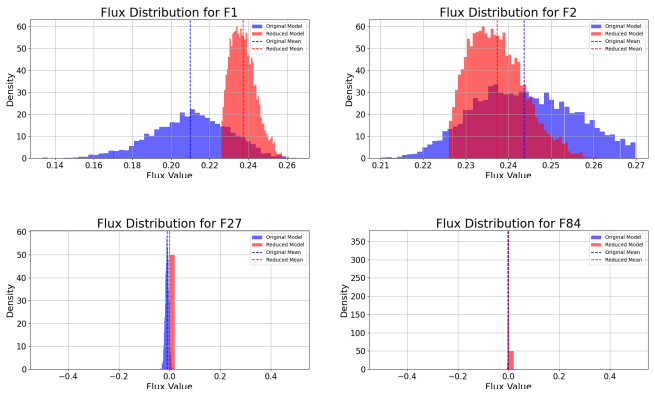


Fig. 5. Flux Distributions before and after model reduction on validation dataset SVM3 (threshold = 0.4).

Figure 5 depicts the flux distributions before and after model reduction, for fluxes 1,2 27 and 84. In the original model, the product of reaction 1 can flow to reaction 2, 27, 84 or 63. Flux 63 was removed from the model due to its probability being below the threshold (0.33). Of the remaining ones, only flux 84 is unmeasured. It is observable that fluxes 27 and 84 retained their mean and shape after reduction. The distribution for flux 2 became narrower but the mean was practically unaffected. Flux 1, however, exhibited a significantly narrower distribution despite the mean only increasing half of a tenth-unit. In this example, the model reduction narrowed the sampling space. This could be either attributed to the feasibility relaxation or the fact that reduced models have fewer sinks, and as such the density per flux needs to be higher.

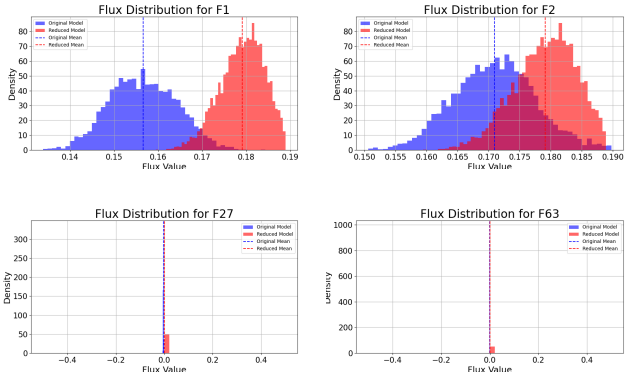


Fig. 6. Flux Distributions before and after model reduction on validation dataset SVM4 (threshold = 0.6).

Figure 6 provides an analysis for the same set of reactions, but for another validation set and for threshold = 0.6. In this case, flux 84 was removed, having a probability of 0.35. Similarly to the previous case, the mean of the samples is relatively unaffected before and after sampling. However, the spread of the reduced model is narrower and the density higher.

3.4 Implications for Experimental Design

The results show that this tool can help reduce the size of metabolic models, while retaining key metabolic pathways, in a systematic and automated way. The trade-off between model size and reaction retention is context-dependent, and as such different thresholds should be studied. The proposed methodology could also be used as a tool to prioritize future experimental measurements.

4. CONCLUSION

This study presented a framework for metabolic network reduction based on flux sampling and Bayesian optimisation. The framework presents an alternative to existing methods that consider both lumping and sensitivity analysis techniques, and capitalizes on existing fluxomic data from carbon labeling experiments.

The workflow includes a tuning step of the surrogate model used for the optimisation, which increases the computational burden of the framework. Future work will focus on reducing this computational effort by exploring other hyperparameter tuning techniques, such as automated or adapted methods. Additionally, we aim to explore alternative metrics for the quantification of model performance

and incorporate a cross-validation step for gene essentiality to make this workflow amenable to genome-scale models.

ACKNOWLEDGEMENTS

MM thanks the UK Biotechnology and Biological Sciences Research Council (BBSRC) and GSK for her studentship. The research was supported by BBSRC grant BB/I017011/1 and Imperial Global Fellows Fund. MM thanks James Morrissey for kindly offering the metabolic network map.

REFERENCES

Ataman, M., Gardiol, D., Fengos, G., and Hatzimanikatis, V. (2017). redgem: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. *PLoS Computational Biology*, 13.

Ataman, M. and Hatzimanikatis, V. (2017). lumpgem: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. *PLoS Computational Biology*, 13.

Bernstein, D.B., Sulheim, S., Almaas, E., and Segrè, D. (2021). Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biology*, 22.

Chu, Y., You, F., Wassick, J.M., and Agarwal, A. (2015). Integrated planning and scheduling under production uncertainties: Bi-level model formulation and hybrid solution method. *Computers & Chemical Engineering*, 72, 255–272.

Danø, S., Madsen, M.F., Schmidt, H., and Cedersund, G. (2006). Reduction of a biochemical model with preservation of its basic dynamic properties. *The FEBS Journal*, 273.

Erdirik-Dogan, M. and Grossmann, I.E. (2008). Simultaneous planning and scheduling of single-stage multi-product continuous plants with parallel lines. *Computers & Chemical Engineering*, 32(11), 2664–2683. doi: https://doi.org/10.1016/j.compchemeng.2007.07.010. Enterprise-Wide Optimization.

Erdrich, P., Steuer, R., and Klamt, S. (2015). An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. *BMC Systems Biology*, 9.

Fouladiha, H., Marashi, S.A., Torkashvand, F., Mahboudi, F., Lewis, N.E., and Vaziri, B. (2019). A metabolic network-based approach for developing feeding strategies for cho cells to increase monoclonal antibody production. *Bioprocess and Biosystems Engineering*, 43, 1381–1389.

Gurobi Optimization, LLC (2024). Gurobi Optimizer Reference Manual. URL `https://www.gurobi.com`.

Jiménez del Val, I., Kyriakopoulos, S., Albrecht, S., Stockmann, H., Rudd, P.M., Polizzi, K.M., and Kontoravdi, C. (2023). Chompact: A reduced metabolic model of chinese hamster ovary cells with enhanced interpretability. *Biotechnology and Bioengineering*, 120(9), 2479–2493. doi:https://doi.org/10.1002/bit.28459.

Kol, S., Ley, D., Wulff, T., Decker, M., Arnsdorf, J., Schoffelen, S., Hansen, A.H., Jensen, T.L., Gutierrez, J.M., Chiang, A.W., et al. (2020). Multiplex secretome engineering enhances recombinant protein production and purity. *Nature communications*, 11(1), 1908.

Okino, M.S. and Mavrovouniotis, M.L. (1998). Simplification of mathematical models of chemical reaction systems. *Chemical Reviews*, 98(2), 391–408. doi: 10.1021/cr950223l. PMID: 11848905.

Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? *Nature biotechnology*, 28(3), 245–248.

Radulescu, O., Gorban, A.N., Zinovyev, A.Y., and Noel, V. (2012). Reduction of dynamical biochemical reactions networks in computational biology. *Frontiers in Genetics*, 3.

Saunders, H., Dyson, B., Vass, L., Johnson, G., and Schwartz, J.M. (2019). Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *npj Systems Biology and Applications*, 5. doi: 10.1038/s41540-019-0109-0.

Schinn, S., Morrison, C., Wei, W., Zhang, L., and Lewis, N. (2021). A genome-scale metabolic network model and machine learning predict amino acid concentrations in chinese hamster ovary cell cultures. *Biotechnology and Bioengineering*, 118. doi:10.1002/bit.27714.

Singh, D. and Lercher, M.J. (2019). Network reduction methods for genome-scale metabolic models. *Cellular and Molecular Life Sciences*, 77, 481 – 488.

Snowden, T.J., Graaf, P.H., and Tindall, M.J. (2017). A combined model reduction algorithm for controlled biochemical systems. *BMC Systems Biology*, 11.

Strain, B., Morrissey, J., Antonakoudis, A., and Kontoravdi, C. (2023a). How reliable are chinese hamster ovary (cho) cell genome-scale metabolic models? *Biotechnology and Bioengineering*, 120(9), 2460–2478. doi:https://doi.org/10.1002/bit.28366.

Strain, B., Morrissey, J., Antonakoudis, A., and Kontoravdi, C. (2023b). How reliable are chinese hamster ovary (cho) cell genome-scale metabolic models? *Biotechnology and Bioengineering*, 120(9), 2460–2478. doi:https://doi.org/10.1002/bit.28366.

van Rosmalen, R., Smith, R., Martins dos Santos, V., Fleck, C., and Suarez-Diez, M. (2021). Model reduction of genome-scale metabolic models as a basis for targeted kinetic models. *Metabolic Engineering*, 64, 74–84. doi: https://doi.org/10.1016/j.ymben.2021.01.008.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272. doi:10.1038/s41592-019-0686-2.

You, F., Grossmann, I.E., and Wassick, J.M. (2011). Multisite capacity, production, and distribution planning with reactor modifications: Milp model, bilevel decomposition algorithm versus lagrangean decomposition scheme. *Industrial & Engineering Chemistry Research*, 50(9), 4831–4849.