Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Cell culture process dynamics and metabolic flux distributions using hybrid models

Rajiv Kailasanathan[a,b], Abhishek Sivaram[a], and Seyed Soheil Mansouri[a]*

[a] Department of Chemical and Biochemical Engineering, Technical University of Denmark
[b] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark
* Corresponding Author: seso@kt.dtu.dk.

ABSTRACT

Cell culture processes play a central role in the production of various therapeutic compounds. These processes are multiscale and highly complex, making them challenging to describe comprehensively using fully mechanistic models. In this study, we employ an integrated hybrid machine learning and first principles model to predict the viable cell density, product titer, and metabolite concentration profiles. We employ the concept of degree of hybridization, where we create a family of hybrid models each with increasing degree of process knowledge. Predictions from the feasible hybrid architecture were integrated with a genome scale metabolic model to evaluate the flux distribution of reactions related to the central carbon metabolism of the cell throughout the process duration. We demonstrate that the current approach not only reasonably predicts the bioprocess profile but also provides biologically relevant information that can uncover dynamics of intracellular metabolism which can open opportunities for new optimization strategies.

Keywords: Modelling and Simulations, Machine Learning, Hybrid Modelling, Metabolic flux distribution

INTRODUCTION

Mathematical modelling of mammalian cell based biomanufacturing process is now at the heart of academic and industrial research due to quality-by-design standards[1] that are recommended by regulatory agencies as well as its potential in optimizing the processes driven by global sustainability demands[2]. Mechanistic (or knowledge-driven, first principles) and statistical (or data-driven) are two ends of a spectrum of modelling strategies. Knowledge-driven models are preferred when fundamental knowledge of the underlying process is available. But, construction of fully mechanistic models for bioprocesses is extremely challenging, as the knowledge available to describe the biology of the process is limited or too complex for practical applications[3]. Hence, conventional mechanistic models attempt to achieve a trade-off between the process knowledge they attempt to describe using the experimental data available and the computational complexity of a robust model.

On the other hand, data-driven models are preferred when there is limited understanding of the underlying process. Statistical and machine learning techniques are employed to infer non-linear relationships between process variables. While these models are often simple to generate and use, they often cannot extrapolate, and a large quantum of training data is required to adequately capture process dynamics.

An attractive combination of mechanistic and data-driven techniques to integrate the advantages of both approaches is being increasingly investigated over the past decades. Several architectures of hybrid models have been built which are described in detail in [4]. In our study, we use a serial architecture in which we establish a mechanistic backbone that is derived from the principle of conservation of mass and use a data-driven model to estimate the unknown parts of the equation. We investigate the concept of 'degree of hybridization'[5], [6] to identify the optimal level of process knowledge that can be described by the hybrid model. To achieve this, we construct a family of hybrid models with varying levels of process knowledge and evaluate their predictive capability.

A connection between macro-scale process model and microscale metabolic behaviour can provide scientific

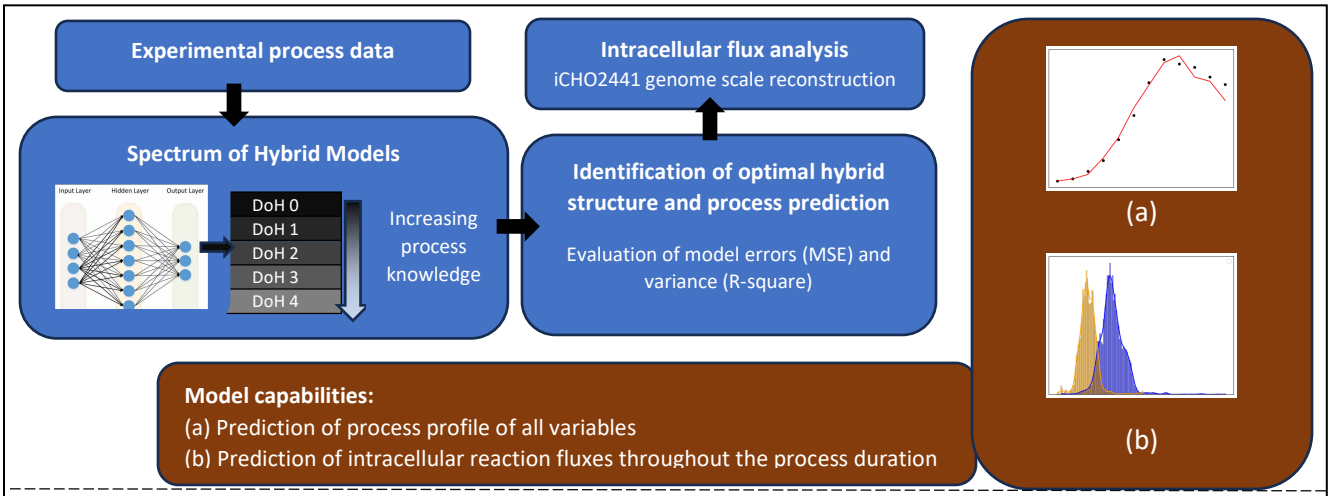

Figure 1: Hybrid modelling framework outline.

understanding of the biological process and uncover new strategies for optimization. Genome scale metabolic models are a mathematical representation of all known metabolic reactions for an organism and are a valuable tool to understand the interplay between various cellular pathways. Usually these models are expressed as an underdetermined system of equations and the flow of metabolites through the network is studied by optimizing the fluxes to achieve a 'cellular objective'[7], which results in an infinite solution space that satisfies the cellular objective as well as other enforced constraints that are determined through experimental and theoretical knowledge. Flux sampling is another powerful method that allows us to study the steady state distribution of fluxes throughout the feasible solution space without assuming a cellular objective[8] and as a result eliminates observer bias.

In this study we harness the power of both hybrid process models and genome scale metabolic models to simultaneously model the process dynamics of CHO cell cultures and describe the intracellular behaviour that drives the observed process. The overall modelling framework is outlined in Figure 1. First, we demonstrate the concept of degree of hybridization and use it to determine the optimal level of process knowledge that accurately describes the process dynamics. Subsequently, we use the predictions of the hybrid process model to determine the distributions of important reactions belonging to the TCA cycle and pentose phosphate pathway of CHO cell metabolism using a recent genome scale reconstruction to study the intracellular behaviour throughout the fed-batch process. We demonstrate that the optimal hybrid model makes predictions that are not only accurate, but also biologically relevant. This framework can be a useful tool in building process monitoring tools and developing optimization strategies.

METHODS

Family of hybrid models

The family of hybrid models are constructed as described by Narayanan[5], where they define the degree of hybridization (DoH) as a measure of the process knowledge that is described explicitly by the mechanistic part of the hybrid model. This strategy was adopted to determine the optimal degree of hybridization that best describes the process dynamics of the cell culture under study. A family of 5 models with increasing degree of hybridization were constructed. The purely data-driven model (DoH =0) has no process knowledge incorporated and is tasked with predicting the process variables directly. The succeeding model (DoH=1) attempts to learn the rate of accumulation of the process variables, and the mechanistic part of the model calculates the value of the process variable at the next time step by integrating the system of differential equations. Here, the data driven part, a neural network, must learn to close the mass balance. The next hybridized model (DoH=2) is tasked with learning only the rates of production or consumption of the process variables, because the mechanistic model describes the mass balance through inflow and outflow rates of the process variables.

Further knowledge is added in the next model (DoH = 3) by encoding the proportionality of the reaction rates to the viable cell density. This allows the data-driven model to learn the specific reaction rates of the various process variables. The final model of this study (DoH = 4) incorporates the phenomenon of cell death, which is also modelled as a reaction proportional to the viable cell density. In all models, we employ only shallow neural networks as the data-driven part of the hybrid model. It is also assumed that the reactor is fully mixed with a homogenous cell population. Table 1 describes the model equations of the family of hybrid models.

Datasets for modelling

Degree of Hybridization (DoH)	Process knowledge described	Mechanistic model	Data-driven model
0	-	-	$[Z, Y_{t-1}, F] \xrightarrow[ANN]{} Y_t$
1	Rate of accumulation	$\frac{d\boldsymbol{Y}}{dt} = \boldsymbol{\alpha}$	$[Z, Y_{t-1}, F] \xrightarrow[ANN]{} \boldsymbol{\alpha}$
2	Mass balance	$\frac{d\boldsymbol{Y}}{dt} = \boldsymbol{\beta} + \boldsymbol{F}_{in} - \boldsymbol{F}_{out}$	$[Z, Y_{t-1}, F] \xrightarrow[ANN]{} \boldsymbol{\beta}$
3	Cell specific reaction rate	$\frac{d\boldsymbol{Y}}{dt} = \boldsymbol{\gamma} * X_v + \boldsymbol{F}_{in} - \boldsymbol{F}_{out}$	$[Z, Y_{t-1}, F] \xrightarrow[ANN]{} \boldsymbol{\gamma}$
4	Cell specific reaction rates, growth and death rates	$\frac{dX_v}{dt} = \boldsymbol{\mu_g} * X_v - \boldsymbol{\mu_d} * X_v + \boldsymbol{F}_{in} - \boldsymbol{F}_{out}$ $\frac{d\boldsymbol{Y}}{dt} = \boldsymbol{\gamma} * X_v + \boldsymbol{F}_{in} - \boldsymbol{F}_{out}$	$[Z, Y_{t-1}, F] \xrightarrow[ANN]{} \boldsymbol{\mu_g}, \boldsymbol{\mu_d}, \boldsymbol{\gamma}$

Table 1: Description of family of hybrid models explored in this study. In all hybrid structures, the data-driven part and mechanistic part are sequentially connected and the data driven part is tasked with learning a different phenomenon for each model. $\boldsymbol{\alpha}$ describes the rate of accumulation, $\boldsymbol{\beta}$ describes rate of production/consumption, $\boldsymbol{\gamma}$ describes cell specific reaction/growth rates, $\boldsymbol{\mu_g}$ is the specific growth rate, and $\boldsymbol{\mu_d}$ is specific death rate. *Z*: process conditions (feed concentrations of glucose and glutamine), *Y*: Response variables measured daily, *F*: Feeding profile and feed rate.

The dataset consists of 33 fed-batch shake flask cultivations that were collected and provided publicly by Bayer et al[9]. In brief, the cultivations were carried out with defined process conditions designed using Design of Experiments (DoE). The 4 important process conditions that were varied in the DoE were glucose concentration in the feed, glutamine concentration in the feed, rate of bolus feed, and the feeding profile. The cultivations were sampled once per day, the response variables (Viable and dead cell density, concentrations of glucose, glutamine, ammonia, lactate, and product) were analytically determined.

Model training and evaluation

The family of hybrid models were constructed and trained in Python using the package PyTorch (2.5.1). The data driven part of the models had similar structure with only 1 hidden layer composed of 10 nodes activated by a *tanh* function. Of the 33 fed batch runs, 5 batches were set aside for testing and the rest were split into a ratio of 80:20 for training and validation after normalization.

The performance of the different models was compared using mean squared error for the test data:

$$MSEP = \frac{1}{N\_time \cdot N\_var} \sum_{i=1}^{N\_var} \sum_{t=1}^{N\_time} \left(C_{i,t}^{meas} - C_{i,t}^{pred} \right)^2$$

In addition, R-squared values were computed for each process variable in the test data for each model, to assess the ability of the model in capturing the variance in the dataset.

Metabolic flux distribution

The metabolic flux analysis part of this multiscale framework is developed using iCHO2441, a genome scale metabolic model of Chinese Hamster Ovary cell[10]. Transport fluxes which carry metabolites in and out of the cell are determined by the hybrid process model described previously. The cell-specific reaction rates of glucose, glutamine, ammonia, and lactate are incorporated in the metabolic model as the transport fluxes. In addition, cell specific growth rate was also used to constrain the 'biomass transport flux'. The assumption of a pseudo-steady state results in equation (2):

$$\boldsymbol{S}\boldsymbol{v_t} = 0, \tag{2}$$

$$\boldsymbol{v_t^{trans}} = \boldsymbol{r_i^{hybrid}}, \quad i \in \{Glc.Gln, Lac, Amm, Biomass\} \tag{3}$$

In equation (2), $\boldsymbol{S}$ corresponds to the stoichiometric matrix of the genome scale model, which is composed of the stoichiometric coefficient of every metabolite in every equation. $\boldsymbol{v_t}$ refers to the flux of every reaction at time **t**. This equation imposes the pseudo-steady state by ensuring that there is no intracellular accumulation of metabolites. The metabolic network is subject to the constraints imposed by equation (3) on the extracellular transport fluxes $\boldsymbol{v_t^{trans}}$, which correspond to the specific exchange reaction rates $\boldsymbol{r_i^{hybrid}}$ that are determined by the hybrid process model. Equation (2) and (3) will result

in infinite possible solutions for $\boldsymbol{v_t}$, as the number of variables is much higher than the number of equations in the model. Biologically, this means that the flow of metabolites can be re-routed in infinitely possible ways to satisfy the requirements of the cell. To study the solution space generated by equations (2) and (3), we perform uniform sampling of the solution space using an artificial centering hit-and-run sampling implemented in Python[11]. By generating 10000 samples of the metabolic network at each time step, we analysed the changes in the distribution of fluxes of reactions belonging to the central carbon metabolism (TCA cycle and Pentose Phosphate Pathway).

RESULTS AND DISCUSSION

Identification of optimal hybrid architecture

Figure 2 compares the predictive performance of the family of hybrid models in the test data set. The purely data driven model exhibits the lowest performance both in terms of mean squared error of the predictions and the coefficient of determination.

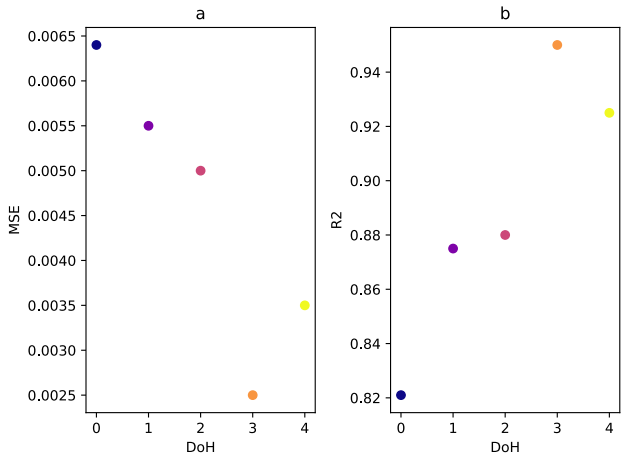

Figure 2: MSE (a) and coefficient of determination (b) of the family of hybrid models applied on the test data set. DoH=3 shows the best performance interms of least error (2.5×10^{-3}) and highest R2 (0.95).

Addition of the first layer of process knowledge (DoH = 1) brings the MSE down by about 14% to 5.5×10^{-3}. This trend continues in the next two steps, where the MSE lowers to 5.0×10^{-3} for the model with DoH = 2 and further down to a minimum of 2.5×10^{-3} for the model with DoH = 3. Interestingly, the DoH=4 model does not follow this trend. Addition of process knowledge about cell death increased the MSE error of DoH=4 model to 3.5×10^{-3}. A similar trend is seen in the coefficient of determination (Fig 2B), where the highest R squared value is observed for the model with DoH = 3. It is important to note that despite the addition of further process knowledge from DoH =3 to 4, the predictive performance of the hybrid model decreases. This observation underlines the trade-off between the generalizability of the hybrid model and the degree of freedom imposed by the hybrid architecture. Based on these observations, DoH=3 was selected to be the optimal level of hybridization.

Intracellular flux distributions

The selected optimal hybrid model provides biologically interpretable information in terms of cell specific uptake and secretions of Glucose, Glutamine, Lactate, and Ammonia. In addition, it is also able to describe the specific growth rate of the cells. The profile of the cell specific exchange fluxes predicted by the hybrid model for a test data set throughout its process profile is shown as illustrative example in Figure 4.

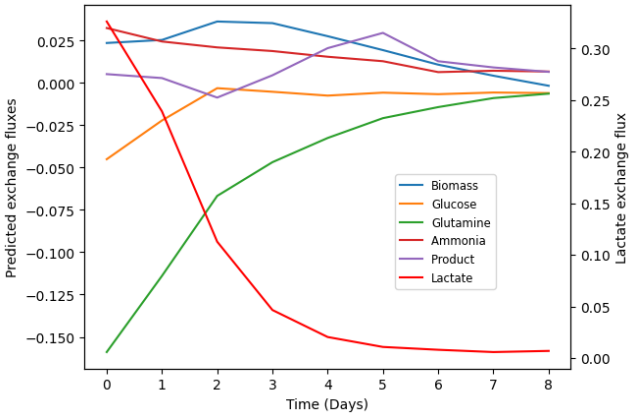

Figure 4: Cell specific uptake and secretion fluxes predicted by the hybrid process model. All fluxes except biomass are expressed in $mmol\,(10^6\,cells)^{-1}day^{-1}$. Biomass flux is expressed conventionally in day^{-1} .

The exchange fluxes predicted by the hybrid process model describe a rapid proliferation/growth phase between days 1-3, and a production phase between days 4-7. In the growth phase, the cells have a high uptake rate of glutamine and glucose. As expected, the secretion flux of lactate is also high, to account for the ATP generation required to meet the cells energy demand. In the production phase, as secretion flux of lactate is low, the carbon flux is directed towards the product.

To understand the intracellular metabolism that underlies the observed process, a genome scale metabolic reconstruction of CHO cell was employed. This reconstruction, iCHO2441[10], has been reported to have improved intracellular predictive performance. By using the predictions from the hybrid model as constraints, 10000 points were sampled from the solution space generated by assuming a pseudo-steady state at each time step. The sampled fluxes were used to visualize the difference in cellular metabolism between the growth phase and the

stationary/production phase.

TCA cycle

The distribution of fluxes of 2 important reactions belonging to TCA cycle is visualized in Figure 5. It is observed that the activity of reactions in the TCA cycle is higher in the production phase than in the growth phase.

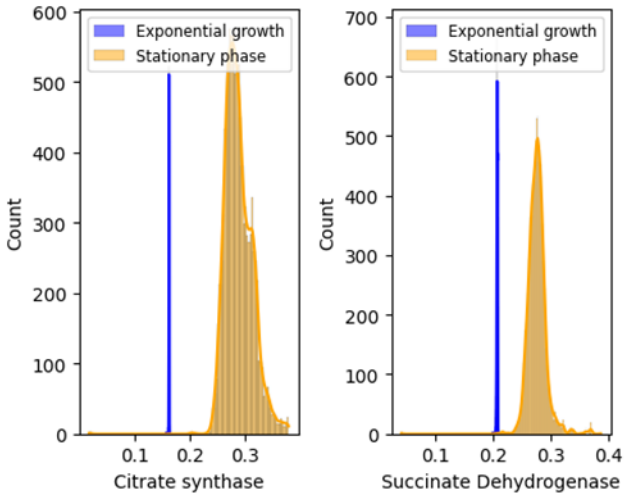

Figure 5: Distribution of fluxes of reaction in TCA cycle (in $mmol\ (10^6\ cells)^{-1} day^{-1}$).

This has been previously verified through experimental approaches through ^{13}C metabolic flux analysis[12]. In the growth phase, lactate production is preferred as it is faster but energetically less efficient route for ATP synthesis. But in the production phase, TCA cycle is favoured to generate ATP. A possible reason for this phenomenon is the inhibiting nature of the accumulated lactate[13].

Pentose Phosphate Pathway

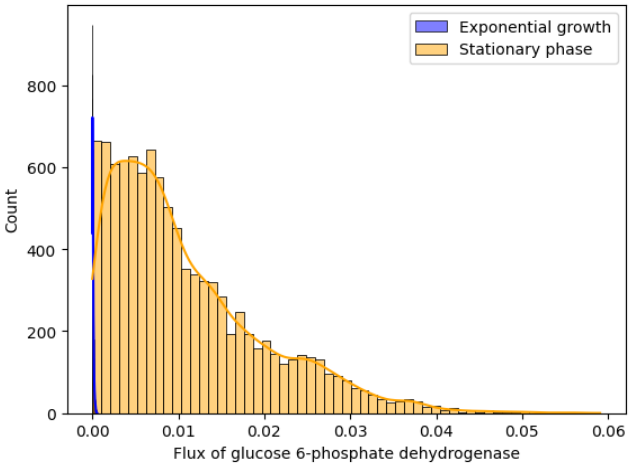

Figure 6: Distribution of flux of G6PDH reaction in the oxidative PP pathway in the exponential and stationary phase.

Oxidative pentose phosphate pathway (oxPPP) flux was significantly higher after the growth phase, while it was close to zero during the growth phase. The flux of glucose-6-phosphate dehydrogenase (G6PDH) increased during production phase even as the uptake fluxes of glucose and glutamine fell considerably when compared to the growth phase. The distribution of G6PDH is visualized in figure 6. This phenomenon of increased oxPPP activity during post growth phase has been reported previously[14]. It has been suggested that the increased flux towards PPP is a result of increased antibody production, which requires constant supply of NADH for protein folding[15].

CONCLUSION AND OUTLOOK

In this study, we demonstrate that for the cell culture process described by the dataset, hybrid architectures work much better than purely data-driven models and that the introduction of fundamental process knowledge greatly improves the predictive capability of the model. By constructing a family of hybrid models with increasing degree of hybridization, we observe that the predictive accuracy of the hybrid model increases as long as the process knowledge introduced does not introduce a bias in the model. There exists a trade-off between the model parameters and the fundamental process knowledge to be added, as a more complex model needs more data points for robust training. In this case, a model with moderate degree of hybridization (DoH=3) was found to be the optimal model by comparing the error and predictive accuracy across the family of models. This model can efficiently learn the cell specific uptake/secretion/growth rates of the cell culture process, which aids in both macro-scale process monitoring as well as understanding of biologically relevant phenomena. While the hybrid process model shows good accuracy within the dataset, it will be interesting to test the capabilities of the model in extrapolating with respect to the process operating conditions. This

A cellular scale understanding of the cell culture process is achieved by employing a genome scale metabolic reconstruction of CHO metabolism constrained by the predictions of the hybrid model. By performing flux sampling, we capture the distribution of the entire feasible solution space of the metabolic network within these predicted constraints. This allows us to study the flux distribution without the assumption of metabolic objectives, which cannot be experimentally determined at most time steps. The captured flux distributions match well with previously available experimental data and provides a micro-scale understanding of the bioprocess. This hybrid framework opens avenues for further process optimization strategies by targeting cellular level processes; for example, a glucose feeding profile that directs carbon

flux towards the TCA cycle instead of the energetically inefficient lactate synthesis.

Overall, the outlined hybrid framework has the ability to combine data-driven techniques with fundamental process knowledge to simultaneously predict macroscale process profile and biologically relevant process information that can provide microscopic information about cellular behaviour. This strategy is particularly relevant for advanced modelling and simulation of cell cultures for process monitoring and optimization.

ACKNOWLEDGEMENTS

This work was funded by the Novo Nordisk Foundation through grant number NNF20CC0035580.

REFERENCES

1. Q. Su *et al.*, "A perspective on Quality-by-Control (QbC) in pharmaceutical continuous manufacturing," *Comput. Chem. Eng.*, vol. 125, pp. 216–231, Jun. 2019, doi: 10.1016/j.compchemeng.2019.03.001.
2. K. Pandey, M. Pandey, V. Kumar, U. Aggarwal, and B. Singhal, "Bioprocessing 4.0 in biomanufacturing: paving the way for sustainable bioeconomy," *Syst. Microbiol. Biomanufacturing*, vol. 4, no. 2, pp. 407–424, Apr. 2024, doi: 10.1007/s43393-023-00206-y.
3. S. Craven, N. Shirsat, J. Whelan, and B. Glennon, "Process model comparison and transferability across bioreactor scales and modes of operation for a mammalian cell bioprocess," *Biotechnol. Prog.*, vol. 29, no. 1, pp. 186–196, 2013, doi: 10.1002/btpr.1664.
4. D. Solle *et al.*, "Between the Poles of Data-Driven and Mechanistic Modeling for Process Operation," *Chem. Ing. Tech.*, vol. 89, no. 5, pp. 542–561, 2017, doi: 10.1002/cite.201600175.
5. H. Narayanan, M. Luna, M. Sokolov, A. Butté, and M. Morbidelli, "Hybrid Models Based on Machine Learning and an Increasing Degree of Process Knowledge: Application to Cell Culture Processes," *Ind. Eng. Chem. Res.*, vol. 61, no. 25, pp. 8658–8672, Jun. 2022, doi: 10.1021/acs.iecr.1c04507.
6. "Hybrid Models Based on Machine Learning and an Increasing Degree of Process Knowledge: Application to Capture Chromatographic Step | Industrial & Engineering Chemistry Research." Accessed: Nov. 11, 2024. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.iecr.1c01317
7. J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?," *Nat. Biotechnol.*, vol. 28, no. 3, pp. 245–248, Mar. 2010, doi: 10.1038/nbt.1614.
8. [8] H. A. Herrmann, B. C. Dyson, L. Vass, G. N. Johnson, and J.-M. Schwartz, "Flux sampling is a powerful tool to study metabolism under changing environmental conditions," *Npj Syst. Biol. Appl.*, vol. 5, no. 1, pp. 1–8, Sep. 2019, doi: 10.1038/s41540-019-0109-0.
9. B. Bayer, M. Duerkop, R. Pörtner, and J. Möller, "Comparison of mechanistic and hybrid modeling approaches for characterization of a CHO cultivation process: Requirements, pitfalls and solution paths," *Biotechnol. J.*, vol. 18, no. 1, p. 2200381, 2023, doi: 10.1002/biot.202200381.
10. B. Strain, J. Morrissey, A. Antonakoudis, and C. Kontoravdi, "How reliable are Chinese hamster ovary (CHO) cell genome-scale metabolic models?," *Biotechnol. Bioeng.*, vol. 120, no. 9, pp. 2460–2478, Sep. 2023, doi: 10.1002/bit.28366.
11. W. Megchelenbrink, M. Huynen, and E. Marchiori, "optGpSampler: An Improved Tool for Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks," *PLOS ONE*, vol. 9, no. 2, p. e86587, Feb. 2014, doi: 10.1371/journal.pone.0086587.
12. J. Dean and P. Reddy, "Metabolic analysis of antibody producing CHO cells in fed-batch production," *Biotechnol. Bioeng.*, vol. 110, no. 6, pp. 1735–1747, 2013, doi: 10.1002/bit.24826.
13. O. Pennington, S. Espinel Ríos, M. T. Sebastian, A. Dickson, and D. Zhang, "A multiscale hybrid modelling methodology for cell cultures enabled by enzyme-constrained dynamic metabolic flux analysis under uncertainty," *Metab. Eng.*, vol. 86, pp. 274–287, Nov. 2024, doi: 10.1016/j.ymben.2024.10.013.
14. N. Templeton, J. Dean, P. Reddy, and J. D. Young, "Peak antibody production is associated with increased oxidative metabolism in an industrially relevant fed-batch CHO cell culture," *Biotechnol. Bioeng.*, vol. 110, no. 7, pp. 2013–2024, 2013, doi: 10.1002/bit.24858.
15. S. Chakravarthi, C. E. Jessop, and N. J. Bulleid, "The role of glutathione in disulphide bond formation and endoplasmic-reticulum-generated oxidative stress," *EMBO Rep.*, vol. 7, no. 3, pp. 271–275, Mar. 2006, doi: 10.1038/sj.embor.7400645.