Dynamic Multiscale Hybrid Modelling of a CHO cell system for Recombinant Protein Production

Oliver Pennington*. Sebastián Espinel Ríos Mauro Torres Sebastian***

Alan Dickson* Dongda Zhang*

**University of Manchester, Manchester, Oxford Road, M1 3AL UK (e-mail: oliver.pennington-3@postgrad.manchester.ac.uk
mauro.torressebastian@manchester.ac.uk
alan.dickson@manchester.ac.uk
dongda.zhang@manchester.ac.uk).*

***Princeton University, New Jersey, NJ 08544, USA (e-mail: s.espinelrios@princeton.edu)*

Abstract: Multiscale hybrid modelling of biosystems utilises advantageous aspects of several modelling approaches, from the physical interpretations of kinetic modelling to the power of a data-driven Artificial Neural Network (ANN). This study implements multiscale modelling to gain insight into the production of Trastuzumab (Herceptin) from Chinese Hamster Ovary (CHO) cells under challenging dynamics. A reduced metabolic network is subject to enzyme constraints with a Dynamic Metabolic Flux Analysis (ecDMFA) approach and integrated within a macro-scale hybrid kinetic model. The model can simulate fed-batch processes with optimized feed control, as well as providing insight into the control gained by alteration to the cell culture media. On the intracellular level, the influence from extracellular perturbations can be observed, in addition to giving an estimated production rate of unmeasured by-products. Overall, this model can be used as a reliable digital twin to estimate the underlying fed-batch process dynamics for future model predictive control and process optimisation.

Keywords: Machine Learning Assisted Modelling, Optimal Control, Experiment Design

1. INTRODUCTION

1.1 Motivation

Bioprocesses link masses of global research interests, including the production of renewable fuels, plastics, and many other high value bioproducts. From an economic standpoint, the UK bioeconomy alone is worth roughly £220 billion (as of 2018), and is set to double by 2030, providing a lucrative research platform (Harrington, 2018). Such growth requires overcoming many challenges, including deficient metabolic and secretory phenotypes for protein production, low yields in reactor scale-up, by-product accumulation (leading to heightened separation costs and product loss), and finally significant batch-to-batch variation, generating quality-control challenges. These are currently the main hurdles to conquer in the field of metabolic engineering and understanding both the metabolic pathways and macro-scale system, as well as how they are interlinked, is essential to overcoming such challenges. Reaching these goals will lead to the development of industrially desired microbial strains for large scale fermentation.

With the fourth industrial revolution becoming ever prevalent, the transition to digitalisation is underway. The utilisation of digital twins has enormous potential for design of experiments, process control, and process optimisation. The concept of machine learning is continuously being employed in new ways to uncover safer, more economical, sustainable, and efficient chemical processing approaches. With the current growth of interest within the fields of both artificial intelligence and bioprocess engineering, there has never been a better time to combine and harness the advantages of each. Mammalian cell lines alone account for around 70% of therapeutic recombinant protein production (O'Flaherty et al., 2020), despite being inherently complex systems; utilising machine learning can play a key role in accurately simulating, controlling, and optimising these complex cell lines, where mechanistic models can struggle to find a balance between oversimplification and over parameterisation.

1.2 Aims

This study aims to unite the macro-scale and micro-scale aspects of a batch cell culture, while incorporating the advantages of an ANN to overcome dynamics that are challenging for physically derived bio-kinetic models to capture. The design of a stable model is key for plausible simulations of the macro-scale and micro-scale systems, upon which further insight can be made regarding relationships between control parameters and the system, process optimisation, and process control. Another challenge to overcome is the infinitely possible solution space in flux-based modelling – it is essential to add constraints and an objective function that greatly narrows the solution space into a region of plausible simulations.

This modelling methodology aspires to be simple in application, with minimal data requirement and computational expense, while maximising simulation accuracy, process insight, and extrapolation potential for process control and optimisation. This will make the modelling methodology described in this work best applicable to complex bioprocesses, such as mammalian cell lines, and new bioprocesses, where knowledge is limited but rapid scale-up is desired. With 20-30 new mammalian made products gaining FDA approval each year (O'Flaherty et al., 2020), there is certainly demand for modelling processes based on limited knowledge.

1.3 Case study

The study at hand looks at the growth of glucose-fed CHO cells across a period of 10 days, with measurements of twenty-nine medium component concentrations (including biomass, glucose, glutamine, lactate, ammonia, Trastuzumab and other amino acids) taken every 24 hours, provided by a previous study (Torres et al., 2019). The modelling focuses on the first 6 days, which are of particular interest since substrate depletion, and therefore cell death, occurs beyond this period.

In this work, the term *hybrid modelling* refers to the combination of a mechanistic (or white-box) model, with a data-drive (or black-box) model to make a combined hybrid (or grey-box) model for the macro-scale simulation.

2. METHODOLOGY

2.1 Macro-scale kinetic modelling

Due to the lack of observable substrate inhibition and cell death, a basic Monod-inspired model was applied to the dynamic extracellular medium. The extracellular medium concentrations of 6 key components were chosen to be simulated: biomass (X), glucose (G), glutamine (Gln), Trastuzumab (P), lactate (Lac) and ammonia (Amm). The remaining components are simulated with a data-driven model. The system of 6 Ordinary Differential Equations (ODEs) is described as in

$$\frac{dX}{dt} = X\,\mu_{\max_G}\,\frac{G}{K_G + G} \tag{1}$$

$$\frac{dG}{dt} = -\,X\,v_{\max_G}\,\frac{G}{K_G + G} \tag{2}$$

$$\frac{dGln}{dt} = -\,X\,v_{\max_{Gln}}\,\frac{Gln}{K_{Gln} + Gln} \tag{3}$$

$$\frac{dP}{dt} = X\,Y_{PX} \tag{4}$$

$$\frac{dLac}{dt} = X\,Y'_{LacG}\,\frac{G}{K_G + G} \tag{5}$$

$$\frac{dAmm}{dt} = X\,Y_{AmmX} \tag{6}$$

where $\mu_{\max_G}$ refers to the maximum specific growth rate of biomass, with vmax being the maximum substrate uptake flux, K_i being the affinity constant for a given substrate i, and Y_{ij} being the yield coefficient between products i and j. It should be noted that an alteration is made to the production rate of lactate due to the strong link between glucose and lactate via the central carbon metabolism pathway within the metabolic network. It is of utmost importance that all parameters remain positive in value to retain physical interpretability, as in

$$\boldsymbol{\beta} \geq \mathbf{0} \tag{7}$$

where $\boldsymbol{\beta}$ is the vector of parameters. For essential amino acids, model structures were trialled based on concepts of constant requirement for growth, constant cell consumption rate, and treating amino acids like a secondary substrate (similar to the glutamine model in Equation 3). Parameters were found using a stochastic optimisation algorithm, Particle Swarm Optimisation (PSO), that minimised a mean squared error objective function, Z, as in

$$\min Z \quad \text{st.}$$

$$Z = \frac{1}{n}\sum_{t,meas}\left(\left(\frac{X_t - X_{meas_t}}{\sigma_X}\right)^2 + \sum_i\left(\frac{C_{i_t} - C_{i_{meas_t}}}{\sigma_i}\right)^2\right) \tag{8}$$

where n is he number of datapoints. The objective function minimises the difference between the simulated extracellular concentration profiles and the measured averages at each timepoint; each term is weighted by the experimental standard deviation of component i; σ_i.

2.2 Hybrid modelling

Challenging growth dynamics in the initial 3 days lead to the requirement of time-varying parameters. However, parameters allowed to vary over time are not defined as functions of time, but as functions of state variables, such as substrate and biomass concentrations, uptake rates, and concentrations of inhibitors such as ammonia. Such functions are defined by an Artificial Neural Network (ANN), and it is the combination of the ANN with the physical model described in Equations 1-6 that forms the hybrid macro-scale model.

The first parameter allowed to vary is the maximum permitted glucose flux; $v_{\max_G}$. This parameter was deemed dynamic due to the excessive initial glucose consumption rates alongside low cell counts. Physically, it can be inferred that initial uptake fluxes may be high due to the initial challenge that the cell faces of adapting to the medium, hence not all substrate consumption drives cell growth. In addition, the presence of less cells means the cell culture can distribute the glucose amongst the cells more readily. It should be noted that deviations in $v_{\max_G}$were penalised to avoid overfitting of the experimental data.

The second time varying parameter was the lactate yield coefficient; Y'_{LacG}. A similar observation was made with extremely high initial lactate production fluxes; often referred to as the Warburg effect (O'Brien et al., 2020). The two aforementioned parameters, $v_{\max_G}$and Y'_{LacG}, are the two outputs of the ANN, which can then be used to simulate the model. This is shown in *ANN representation for the simulation of time-varying parameters based on system components and dynamics* (Fig. 1).

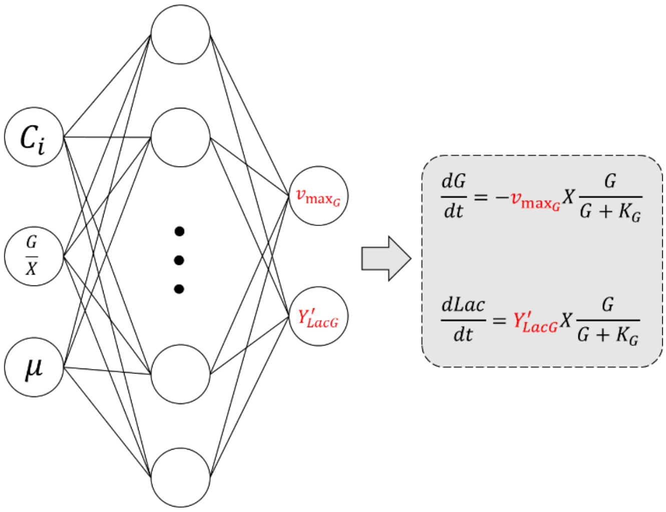
C_i
$\frac{G}{X}$
μ
$v_{\max_G}$
Y'_{LacG}
$\frac{dG}{dt} = -v_{\max_G} X \frac{G}{G + K_G}$
$\frac{dLac}{dt} = Y'_{LacG} X \frac{G}{G + K_G}$

Figure 1 - ANN representation for the simulation of time-varying parameters based on system components and dynamics

For the remaining extracellular components, the aim is to shift towards the use of another ANN to predict requirements of essential amino acids where a basic kinetic model does not hold true. Amino acid consumption rates will be the ANN outputs, with inputs being factors such as inhibitors, cell count and cell growth rate.

In order to determine an appropriate ANN structure, the performance of several ANN structures should be compared. This is done by accounting for the simulated error, as well as the total number of parameters, k. An Akaike Information Criterion with correction for small sample sizes ($AICc$) is employed to quantify the performance of each ANN structure, as in

$$AICc = n \ln Z + 2\,k + \frac{2\,k^2 + 2\,k}{n - k - 1} \tag{9}$$

where correction for a small sample size n was chosen due to the shortage of data, which is likely to be a recurring theme in novel bioprocesses – a target of this modelling methodology. The network with the lowest AICc score was chosen as the best compromise between accuracy and minimal overfitting.

2.3 ecDMFA

A metabolic network developed in a recent CHO cell study regarding the reduction of a genome-scale model (Jiménez del Val et al., 2023) was obtained. Reduced metabolic networks have the benefit of greatly reduced computational expense for dynamic systems, while still being able to capture key intracellular behaviour. The reduced metabolic network was employed to develop a constraint seen in traditional Metabolic Flux Analysis (MFA). Dynamic Metabolic Flux Analysis (DMFA) is effectively a series of consecutive MFA problems solved simultaneously with different extracellular inputs and outputs. The reduced network contains 144 reactions linked by 101 metabolites and forms the mass balance defined by a stoichiometric matrix ($\mathbf{S}$) and the vector of fluxes ($\mathbf{v}_t$), as in

$$\mathbf{S}\,\mathbf{v}_t = \mathbf{0} \tag{10}$$

where the time subscript (t) refers to the constraint being valid at every time-step the constraint is applied to, despite changes in the values of intracellular flux. The biomass specific growth rate is utilised within this constraint to account for intracellular metabolite dilution, however the product of specific growth rate and intracellular metabolite concentrations (rate of dilution) is extremely low (in comparison to simulated fluxes), making it non-essential to include if no estimation of intracellular metabolite concentration can be made.

The network is further constrained by mass balances in accordance with the macro-scale hybrid kinetic model, where the vector of measured transport ($trans$) fluxes (a subset of $\mathbf{v}_t$) is defined by changes in the vector of measured extracellular component concentrations ($\mathbf{C}$), as in

$$\mathbf{v}_{trans_t} = -\frac{1}{X}\frac{d\mathbf{C}}{dt_t} \tag{11}$$

where biomass concentration (X) accounts for the change from macro-scale to micro-scale dynamics. The biomass specific growth rate (μ) is effectively a transport flux for the biomass component and is accounted for in the stoichiometric matrix in Equation 10.

As aforementioned, flux-based modelling approaches have large numbers of degrees of freedom; there can be infinite solutions. To help narrow down the solution space to a more realistic region, further constraints can be employed. In this study, enzyme-constraints are utilised to direct flux through biologically favourable pathways. Each reaction j has an associated flux and enzyme, with a corresponding molar mass (M_{r_j}) and turnover rate (k_{cat_j}). These parameters can be used alongside a reaction flux to estimate the mass of each enzyme required for the simulate flux and the total enzyme mass is constrained to a known maximum (M_E), as in

$$\sum_j \frac{M_{r_j} v_{j_t}}{k_{cat\,j}} \leq M_E \tag{12}$$

where parameters for Equation 12 were found from an existing enzyme-constrained modelling study on CHO cells (Yeo et al., 2020). It is paramount to account for the fact that a reduced metabolic network is being used so less reactions are accounted for. Therefore, M_E must be adjusted accordingly. The subscript time (t) once again indicates that the constraint in Equation 12 is applied to every MFA time-step. Note that the total allowed enzymatic mass per cell is time-invariant. For this constraint to retain credibility, all fluxes must be non-negative; reversible reactions j (with net flux v_{rev_j}) are split into separate non-negative forward and backward fluxes (v_{F_j} and v_{R_j}, respectively), as in

$$v_{rev_j} = v_{F_j} - v_{R_j} \tag{13}$$

$$v_{F_j}, v_{R_j} \geq 0 \tag{14}$$

Another way in which to simulate the most realistic flux profiles possible is to employ an objective function that imitates observable cell phenotypes. One example is efficient operation, where a cell aims to minimise the magnitude of all reaction rates within it to minimise wasted energy sources, such as ATP. Another is the cell stability – drastically inconsistent fluxes are unlikely to be correct so substantial changes in the same reactions flux should be penalised over each iteration time gap Δt. The flux reducing objective and flux inconsistency penalty are therefore combined into one quadratic (and therefore convex) objective function, as in

$$\min \sum_{t=t_0}^{T} \mathbf{v}_t + \lambda \sum_{t=t_0+\Delta t}^{T} (\mathbf{v}_t - \mathbf{v}_{t-\Delta t})^T (\mathbf{v}_t - \mathbf{v}_{t-\Delta t}) \qquad (15)$$

where λ is the penalty weight – a tuneable parameter that was determined through an iterative approach that ensures minimal impact on the main objective term.

2.3 Methodology summary

The overall modelling methodology is conducted in order to maximise simulation accuracy and plausibility, while minimising data requirement and computational expense. The overall methodology is broken down as follows in *Summary of multi-scale hybrid modelling approach* (Fig. 2).

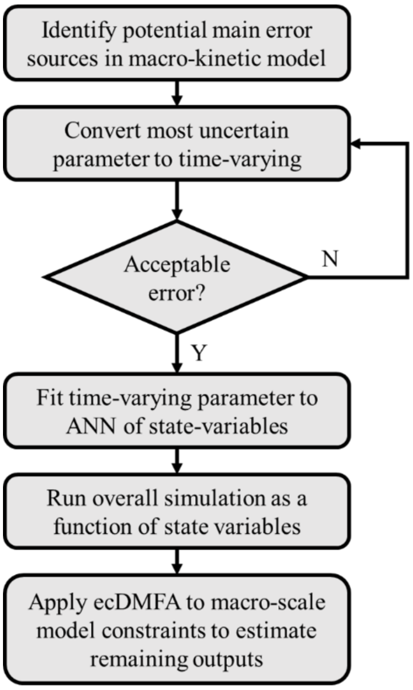
Identify potential main error sources in macro-kinetic model
Convert most uncertain parameter to time-varying
Acceptable error?
N
Y
Fit time-varying parameter to ANN of state-variables
Run overall simulation as a function of state variables
Apply ecDMFA to macro-scale model constraints to estimate remaining outputs

Figure 4 - Summary of multi-scale hybrid modelling approach

3. RESULTS AND DISCUSSION

3.1 Macro-scale hybrid modelling

The purely mechanistic macro-scale kinetic model was first fit with Equations 1-6 with no time-varying parameters. *Sample plots of substrate (glucose) and biomass concentration profiles for original Macro-Kinetic Model (MKM)* (Fig. 3) shows a sample of the simulation with substrate (glucose) and biomass concentration profiles. As shown, significant improvements can be made to the 26.8% mean error present, and so the next step is to allow $v_{\max_G}$ to behave as a dynamic parameter to account for excessive initial glucose uptake rates.

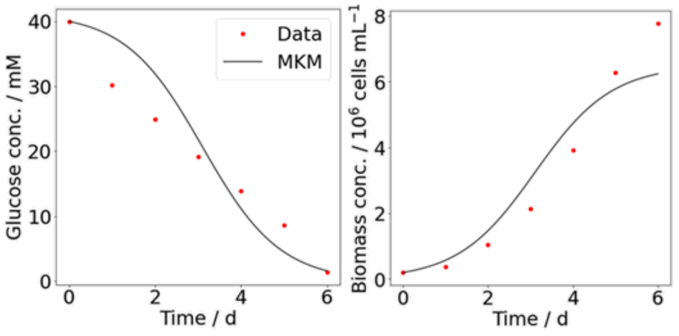

Figure 2 - Sample plots of substrate (glucose) and biomass concentration profiles for original Macro-Kinetic Model (MKM)

The introduction of a time-varying parameter, $v_{\max_G}$, to the glucose consumption rate drastically improved the experimental data fits of both glucose and biomass without overfitting, as seen in *Sample plots of substrate (glucose) and biomass concentration profiles for original Macro-Kinetic Model (MKM)* (Fig. 3), and *Sample plots of substrate (glucose) and biomass concentration profiles for Macro-Kinetic Model (MKM) with time varying $v_{\max_G}$, which has a profile shown (bottom left)* (Fig. 4). The mean error of the model was reduced to 6.9% from 26.8% through the introduction of only two time-varying parameters ($v_{\max_G}$ and Y'_{LacG}).

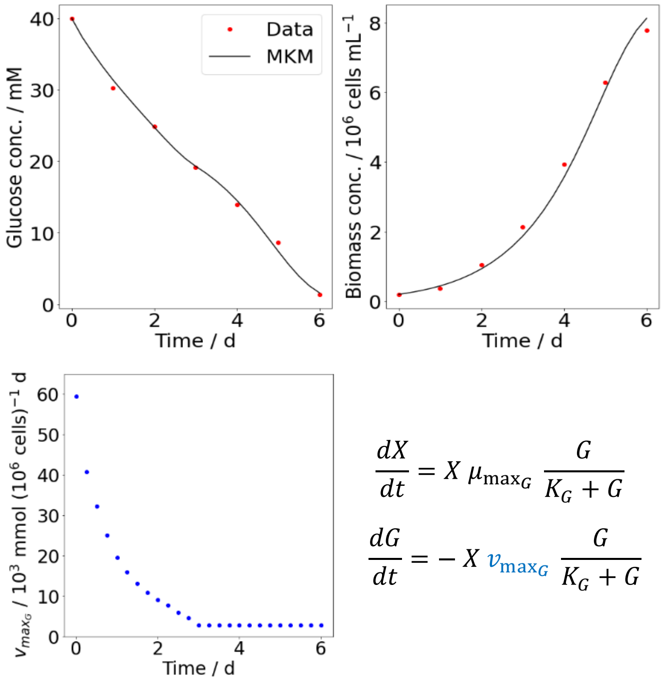

Figure 3 - Sample plots of substrate (glucose) and biomass concentration profiles for Macro-Kinetic Model (MKM) with time varying $v_{\max_G}$, which has a profile shown (bottom left)

The parameter shows a decreasing trend that stabilise in the latter stages of the batch process. This aligns with the idea of the cell culture beginning to stabilise, further supporting the theory that glucose is not initially directly driving cell growth like it is in the latter stages of the batch process, meaning it is either being wasted or utilised in an alternative manner that the intracellular model could uncover.

Time-varying parameters were not allowed to vary beyond a time of 3 days since the culture should be in a more stable growth phase. This further supported our preliminary studies which showed excellent fits with constant parameter values when simulating days 3 to 6.

Parameter convergence occurred within 500 iterations using a stochastic Particle Swarm Optimisation (PSO) algorithm, which was chosen for its capability of handling highly non-convex optimisation programming problems due to having the ability to escape local minima, unlike gradient-descent approaches. The problem was run 5 times, each with 1000 particles, to confirm the same optima was being found, increasing the confidence in it being a global optimal solution.

3.2 ecDMFA

The ecDMFA was successfully applied to the macro-scale hybrid model concentration profiles with samples being taken 4 times per day ($\Delta t = 0.25$ d). Such sampling was chosen as a trade-off between computational expense (which increases with more frequent sampling) and avoiding overly linearised flux profiles (which becomes a source of significant error with less frequent sampling).

Enzyme constraints were not violated, helping guide flux down plausible reaction pathways. Example fits to Equation 11, from components described in Equations 1-3 (glucose, biomass, and glutamine), are shown below in *Transport fluxes for glucose consumption, biomass specific growth and glutamine consumption as defined by the Macro-Kinetic Model (MKM) and applied to the ecDMFA samples* (Fig. 5).

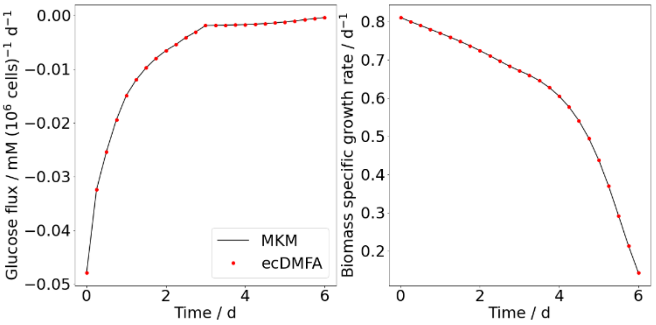

Figure 6 - Transport fluxes for glucose consumption, biomass specific growth and glutamine consumption as defined by the Macro-Kinetic Model (MKM) and applied to the ecDMFA samples

Convergence occurred using an Interior Point Optimisation (IPOPT) algorithm. Gradient-based optimisation was chosen since Equations 11-15 describe a convex optimisation problem, allowing the global optimum to be found upon convergence with or without the enzyme-constraints.

The flux results for the metabolic network can be assessed to further validate ecDMFA simulation results throughout the experimental timeframe, which in turn helps validate the feasibility of the macro-scale simulation while giving insight into the dynamic intracellular reaction network. An example of a flux distribution sample is given in *Sample flux distribution through the central carbon metabolism from day 3 of the simulation, with the shaded region representing the mitochondria* (Fig. 6). In the stable cell growth phase, the flux distribution stemming from glucose uptake is shown to be channelled into turning the TCA cycle, with less carbon being wasted in the form of lactate production. This observation aligns with current cellular understanding that lactate inhibits its own production, so central carbon flux will be directed away from its production beyond the initial high-production phase.

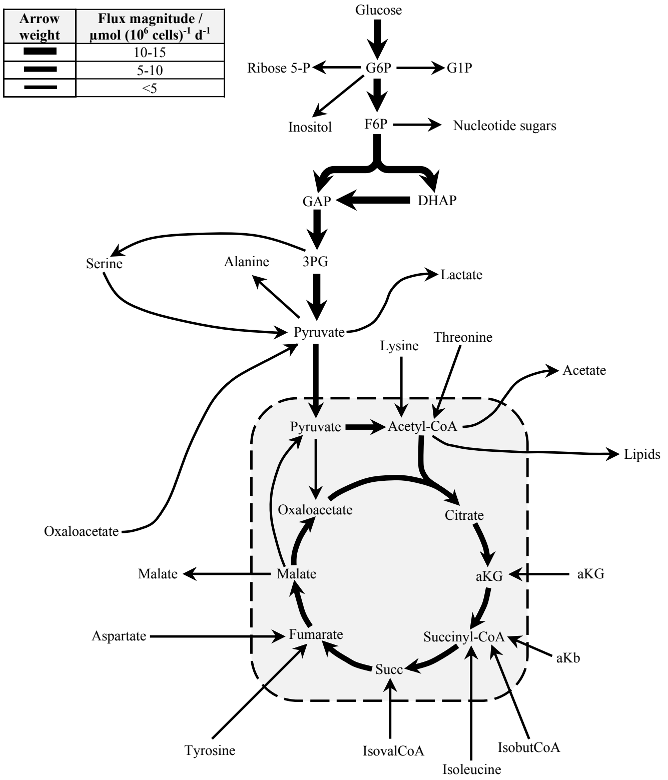

Figure 5 - Sample flux distribution through the central carbon metabolism from day 3 of the simulation, with the shaded region representing the mitochondria.

Individual fluxes can also be analysed and compared to further validate the ecDMFA results. An example of this is shown, again from the central carbon metabolism, in *Flux profiles for reactions 1 and 2 - the conversion of glucose into Fructose 6-Phosphate (F6P)* (Fig. 7).

The enzyme-constraint was met but was not limiting, as shown in *Simulated cellular enzyme mass requirement, with the "Max Quantity" referring to the parameter* M_E (Fig. 8). This is intuitive since not all enzymes are activated at once, meaning there should be a noticeable buffer for enzyme activation. Fluxes, and therefore enzyme requirements, decrease over time due to reduced extracellular substrate (glucose and glutamine) concentrations in combination with an increased cell concentration. This decreases the specific uptake rates per cell, leading to lower fluxes throughout the metabolic network.

Individual enzyme requirements can also be scrutinised, with noticeable contributions coming from the central carbon

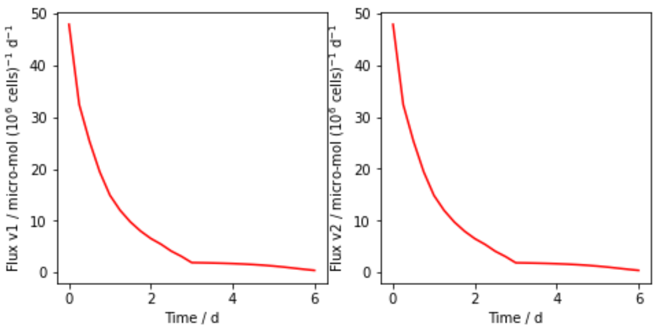

Figure 7 - Flux profiles for reactions 1 and 2 - the conversion of glucose into Fructose 6-Phosphate (F6P)

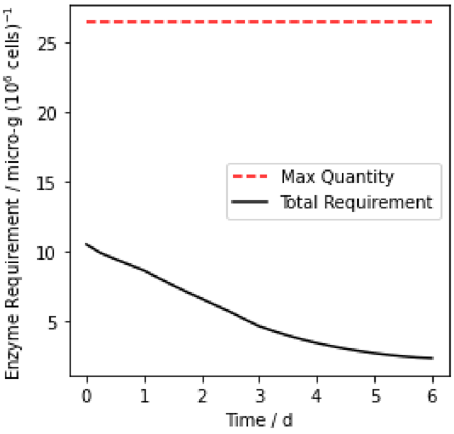

Figure 8 - Simulated cellular enzyme mass requirement, with the "Max Quantity" referring to the parameter M_E

metabolism where a large proportion of flux is directed. Not all large fluxes require significant quantities of active enzymes however, since some enzymes have particularly high turnover rates (k_{cat_j}), such as the conversion from Glucose 6-Phosphate (G6P) to Fructose 6-Phosphate (F6P).

The ecDMFA methodology has proven to be capable of capturing a dynamic system, while giving physically plausible estimates for unmeasured extracellular metabolite demands and production rates. This can potentially be used to indicate when essential amino acids have been depleted for longer fed-batch processes.

4. CONCLUSIONS

In this study a multi-scale hybrid model has been successfully constructed and scrutinised to gain insight into a complex CHO cell batch process for the production of Trastuzumab. The novelty of the work lies in the combination of a macro-scale hybrid kinetic model with a micro-scale ecDMFA modelling methodology. The introduction of time-varying parameters allowed complex macro-scale dynamics to be captured while sticking as closely as possible to the original kinetic model structure, thus maximising the capacity for extrapolation needed for process optimisation and control. Using an ANN has allowed the macro-scale model to remain a function of observable components of the extracellular medium only, thus removing the need for profiles written as functions of time that have been seen in previous work (Pennington et al., 2023). Utilizing an underlying macro-kinetic model also greatly reduces the amount of data required to accurately simulate the extracellular experimental data.

The use of the ecDMFA methodology described encourages plausible flux simulations for identifying trends between extracellular and intracellular dynamics, while remaining a convex programming problem to minimise computational cost. A combined model allows the simulation and optimisation of batch and fed-batch processes while utilising a micro-scale model to uncover cell functionality under feasible operation. A key benefit of this work is the ability to bypass a trial-and-error approach to several model structures for a poorly understood system, while still incorporating some fundamental knowledge to improve the ability to extrapolate. Future work looks to incorporate more detailed genomic data for constraints and validation.

Therefore, successes of this model paves the way to developing an effective digital twin for the modelling and prediction of the underlying process, and can help identify an optimal control strategy for maximising Trastuzumab production for future fed-batch operations. The ability of this model to incorporate machine learning with minimal data gives it huge potential for control and optimisation applications to novel bioprocesses, where both mechanistic understanding and experimental data is limited.

REFERENCES

Harrington, R. (2018). *Growing the bioeconomy: a national strategy to 2030*, HM Government.

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*, volume 120, issue 9, pages 2479-2493.

O'Brien, C.M., Mulukutla, B.C., Mashek, D.G., and Hu, W.S. (2020). Regulation of Metabolic Homeostasis in Cell Culture Bioprocesses. *Trends in Biotechnology*, volume 38, issue 10, pages 113-1127.

O'Flaherty, R., Bergin, A., Flampouri, E., Mota, L.M., Obaidi, I., Quigley, A., Xie, Y., and Butler, M. (2020). Mammalian cell culture for production of recombinant proteins: A review of the critical steps in their biomanufacturing. *Biotechnology Advances*, volume 43.

Pennington, O., and Zhang, D. (2023). Comparing different modelling approaches for metabolic network dynamic simulation under uncertainty. *Computer Aided Chemical Engineering*, volume 52, pages 2589-2594.

Torres, M., Julio, B., Rigual, Y., Latorre, Y., Vergara, M., Dickson, A.J., and Altamirano, C. (2018). Metabolic flux analysis during galactose and lactate co-consumption reveals enhanced energy metabolism in continuous CHO cell cultures. *Chemical Engineering Science*, volume 205, pages 201-211.

Yeo, H.C., Hong, J., Lakshmanan, M., and Lee, D.Y. (2020). Enzyme capacity-based genome scale modelling of CHO cells. *Metabolic Engineering*, volume 60, pages 138-147.