Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Engineering

Moving towards an era of hybrid modelling: advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses

Apostolos Tsopanoglou and Ioscani Jiménez del Val

Monoclonal antibodies (mAbs) are the highest-selling class of biopharmaceuticals due to their capabilities in treating severe illnesses. Thus, strategies to optimise their manufacture have been at the centre of recent academic and industrial research. Mechanistic and statistical modelling approaches are considered useful tools in that quest, not only because they can be employed for online process monitoring but also due to their abilities to offer useful insight into how the underlying micro and macro-scale phenomena of upstream bioprocesses impact the yield and quality of biopharmaceuticals. This manuscript provides an overview of the mechanistic and statistical models of upstream mAb bioprocesses that have been published over the past five years and discusses their advantages and drawbacks. We conclude with an outline of synergistic, hybrid modelling strategies, which are emerging as key tools in the era of Biopharma 4.0.

Address
School of Chemical & Bioprocess Engineering, University College Dublin, D04 V1W8, Ireland

Corresponding author:
Jiménez del Val, Ioscani (ioscani.jimenezdelval@ucd.ie)

Current Opinion in Chemical Engineering 2021, **32**:100691

This review comes from a themed issue on **Biotechnology and bioprocess engineering: mechanistic and data-driven modelling of bioprocesses**

Edited by **Cleo Kontoravdi** and **Colin Clarke**

For a complete overview see the Issue and the Editorial

Available online 5th June 2021

https://doi.org/10.1016/j.coche.2021.100691

Introduction

The highest selling therapeutic glycoprotein products are monoclonal antibodies (mAbs) which, in 2018, accounted for US$103.4 billion in annual sales (~55% of the total revenue of all biopharmaceuticals) [1]. The demand for this class of biotherapeutics has been constantly increasing due to their efficacy in treating severe illnesses including cancer, autoimmune disorders, and infectious diseases [2]. Thus, mAb revenues are projected to exceed US$140 billion by 2024 [3]. The market for alternative immunotherapy formats, for example, antibody fragments, bispecific mAbs, and antibody conjugates is also expected to grow at an accelerated pace, with 16 such products undergoing late-phase clinical as of November 2020 [4]. Although the modelling strategies outlined in this review are relevant to all therapeutic glycoprotein formats, we focus on mAbs, given their commercial prevalence and recent advances in modelling their manufacturing bioprocesses.

Mammalian cell lines are the preferred expression platform for manufacturing biopharmaceuticals owing to their ability to produce complex proteins with human-like post-translational modifications [5]. The most popular expression system are Chinese Hamster Ovary (CHO) cells because they have been deemed as safe hosts for biopharmaceutical manufacturing and, hence, have high approval rates from regulatory agencies such as FDA and EMA [2].

Notably, cell culture conditions, such as temperature, pH, and media composition play a critical role in defining product quality during manufacturing processes [6,7]. Linking mathematical representations of mammalian systems with experimental work offers quantitative information on how Critical Process Parameters (CPPs) affect process outputs, such as product titre and glycosylation profiles and, hence, reduce process development time and costs [8]. Such strategies are employed within the Quality by Design (QbD) framework and must be supported by Process Analytical Technology (PAT) tools to ensure optimal safety and therapeutic efficacy of the final product [9••].

One of the challenges in employing model-based methods is the assurance of model accuracy, precision, and predictive capabilities. Mechanistic models can be used as digital twins for optimal design of experiments or for process monitoring [10]. To do so, it is imperative to initially generate process-specific experimental data, parametrise the constructed models and, with independent data, successfully validate their predictive capabilities.

In this review, we present a synopsis of the applications and challenges of mechanistic and statistical models in the context of upstream cell culture bioprocesses with particular focus on cell culture dynamics and product critical quality attributes (CQAs). Next, we provide an overview of hybrid modelling and its value in the era of Biopharma 4.0. We conclude with arguments in favour of hybrid modelling approaches, which integrate statistical

and mechanistic models to exploit their individual advantages while simultaneously minimising their respective limitations.

Mechanistic models of upstream cell culture: classification and challenges

Models based on physical, chemical, and biological principles offer a robust option in process engineering. Kinetic modelling is the preferred mechanistic approach to capture the underlying phenomena during cell culture, with expressions consisting of dynamic material balances (Eq. (1)) for viable/non-viable cell densities and nutrient/metabolite concentrations within the bioreactor.

$$\frac{d(V\cdot C_i)}{dt} = Q_{in}\cdot C_{i,in} - Q_{out}\cdot C_i + q_{i,prod}\cdot X_v\cdot V - q_{i,cons}\cdot X_v\cdot V \quad (1)$$

In Eq. (1), V is bioreactor liquid volume, $Q_{i,feed}$ and $Q_{i,out}$ are the inlet and outlet volumetric flowrates, $C_{i,in}$ is the concentration of nutrient i in the feed, C_i the concentration of nutrient/metabolite i within the bioreactor, $q_{i,prod}$ and $q_{i,cons}$ are the cell-specific production and consumption rates of nutrient/metabolite i, and X_v is viable cell density.

$$q_i = \alpha_i\left(\frac{C_i}{K_{m,i} + C_i}\right) \quad (2)$$

In mechanistic models, Monod-type equations (Eq. (2)) are most commonly used to define the cell-specific rates of biomass growth (μ_g) as well as the cell-specific uptake and secretion rates of each nutrient/metabolite i (q_i) [8]. Two kinetic parameters are present in Eq. (2): α_i is the maximum cell-specific rate and $K_{m,i}$ is the Monod saturation constant.

Derivation of such models is a complex process which requires expertise as Monod-expressions are commonly extended to include metabolite inhibition terms or other modifications [11,12•,13] to ensure that the behaviour of cultured mammalian cells is captured. Such modifications often result in overparametrised models. Another important challenge associated with this modelling framework is the definition of the stoichiometric relationships between nutrients, metabolites, cells, and product.

Depending on the level of mechanistic insight included in the kinetic model (and consequently contingent on the available mechanistic understanding for the mammalian production host to be modelled), these can be classified as belonging to one of the following: (*i*) unstructured-unsegregated, (*ii*) structured-unsegregated, (*iii*) unstructured-segregated, and (*iv*) structured-segregated [14].

Unstructured-unsegregated models utilise a black box approach where all biological reactions are lumped and defined as functions of abiotic variables (e.g. nutrient and metabolite concentrations) occurring within a single population of homogenous cells [15,16]. These models can be used to depict the dynamics of cell density, viability, nutrient/metabolite concentrations, and product titre [17,18]. In unstructured-unsegregated models, numerous assumptions and simplifications are required to compensate for the absence of mechanistic descriptions of intracellular phenomena.

When details about the intracellular environment of a homogenous cell population are described, the model is referred to as structured-unsegregated. Kotidis *et al.* [12•] proposed a mechanistic modelling approach linking a module describing CHO cell growth, extracellular nutrient and metabolite concentrations, and mAb titre with a second module for intracellular nucleotide sugar donor metabolism and a third module describing the enzymatic *N*-glycosylation process within the Golgi apparatus.

If the model captures the inherent cell population heterogeneity within the bioreactor, then it belongs to the unstructured-segregated category. These models typically perform population balances based on the dynamics of cell cycle phases or markers of cell health (e.g. apoptosis) [19,20]. However, the resulting models are heavily parametrised and obtaining high-confidence numerical values for the parameters requires extensive and bespoke experimental data which may pose challenges to model deployment in industry.

Lastly, structured-segregated models are the most complex because they consider both intracellular compartments and multiple cell populations to describe cell culture dynamics. Although these phenomena are key contributors to defining cell culture dynamics, lack of universality, and the ensuing requirement of extensive experimental effort for parameterisation make these models difficult to deploy in the industrial setting. A comprehensive review of mechanistic models for upstream pharmaceutical bioprocessing has been recently published elsewhere [14].

Statistical modelling

In this review, we refer to statistical models as those where non-mechanistic parametric equations are used to compute process outputs (response variables) from inputs (predictor variables). When applied to pharmaceutical bioprocesses, statistical modelling is embedded within a multivariate data analysis (MVDA) workflow that identifies the strongest correlations between critical process parameters (CPPs) and critical quality attributes (CQAs) without focussing on the mechanistic causalities governing the relationships [21•,21•,23•].

Table 1

Pros and Cons of upstream bioprocess modelling strategies and examples

Pros	Cons	Examples
Mechanistic		
• Based on mechanistic knowledge (biology, chemistry, physics) and yield enhanced process understanding. • Can be used for model-based design of optimally informative experiments. • Can be deployed for bioprocess control and optimisation.	• Extensive experimental effort for model validation. • Overparametrised models lack robustness and universality. • Difficult to automate/formalise model assembly. • Difficult to deploy within industry due to required expertise.	• Kinetic model for culture dynamics, metabolism, and glycosylation to optimise mAb quality [12•]. • Population balances, cell cycle dynamics, and markers of cell health to maximise mAb yield [20]. • Kinetic modelling for non linear model predictive control of glucose within a bioreactor [40].
Statistical		
• Automatic model assembly. • Reduced computational overhead. • Can be used for real-time bioprocess monitoring and fault detection.	• Predictive capabilities limited to validation space. • Limited bioprocess control and optimisation capabilities. • Requires easily accessible, representative, and reliable training datasets. • Data may require pre-treatment.	• MVDA links intracellular and extracellular metabolite concentrations with product glycosylation to accelerate process development [21–23]. • MVDA to create PLS model that predicts mAb titre and glycosylation based on amino acid, pH, and DO inputs [30]. • PLS-based soft sensor to control glucose within a bioreactor [31].
Hybrid		
• Constrains model outputs to physically feasible values via mechanistic relationships. • Automatic model assembly. • Model-based design of optimally informative experiments. • Can be deployed for bioprocess control and optimisation.	• Requires easily accessible, representative, and reliable training datasets. • Resource-consuming training process required to achieve adequate predictive capabillities. • Data may require pre-treatment.	• Hybrid mechanistic + ANN model to predict product titre based on nutrient/metabolite inputs [9••]. • Hybrid mechanistic + ANN model that predicts the effect of manganese feeding on mAb glycosylation [36••].

Statistical regression analysis has found industrial relevance in translating raw PAT signals into CPP and CQA measurements to deliver *in situ* bioprocess monitoring tools. Such strategies have translated raw signals of advanced PAT (e.g. capacitance [24], Raman [25], and near-infrared (NIR) spectra [26], into data for viable cell density [24], nutrient concentrations [25], and mAb quality, including glycosylation [26] and aggregation [27].

Most relevant to this manuscript is the use of multivariate data analysis (MVDA) to create statistical models that directly relate bioprocess CPPs with CQAs. Broadly, the MVDA workflow to create statistical models includes Principal Component Analysis (PCA) to identify correlations within the dataset and to quantify how the variables (latent or otherwise) contribute to dataset variance. The latent structures and correlations obtained through PCA are then used to construct the statistical models through Partial Least Squares (PLS) regression which may be complemented with variable importance in projection (VIP) or genetic algorithm analysis, to identify the subset of input variables that maximises model predictive capabilities [21•,22,23•].

Alternatively, model construction and parameterisation can be performed with ML algorithms (e.g. Random Forrest and Artificial Neural Networks — ANN) [28] that can offer improved model robustness over PLS regression because they are able to account for non-linear interactions between model inputs and outputs [29]. Irrespective of the strategy used for assembly, the final step in model development is cross-validation against independent datasets to assess robustnes and predictive capabilities.

MVDA workflows have been employed to create PLS models linking the effect of process inputs (culture temperature, media supplementation, and feeding times) on mAb CQAs (titre, fragmentation, aggregation, charge variants, and glycoforms) [21•] to support upstream bioprocess development across multiple scales [22,23•]. Green and Glassey [30] built PLS models to compute product titre and glycosylation using glucose and amino acid data as inputs, and Goldrick *et al.* [31] developed a PLS-based soft sensor to control glucose concentration based on online dissolved oxygen measurements. With further development, such strategies can be leveraged for real-time bioprocess optimisation [32].

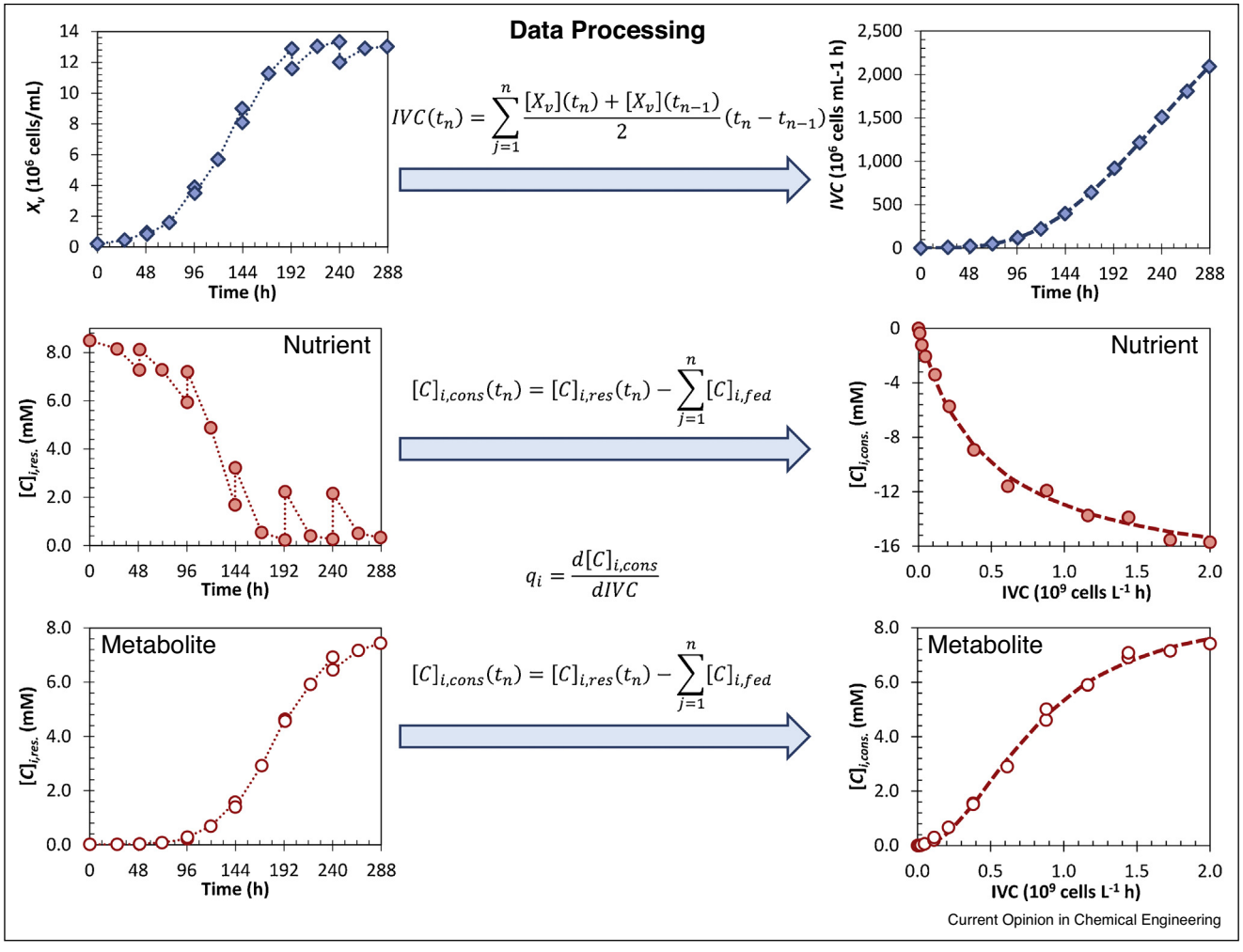
The key advantages of purely statistical models (Table 1) are that they can be assembled automatically and their relative simplicity incurs low computational expense. Although these advantages have accelerated industrial uptake of such workflows, purely models are, due to their

nature, unable to provide insight into causal effects relating bioprocess inputs and outputs.

In addition, MVDA predictions become less robust when modelling raw data (e.g. with discontinuities or missing datapoints) and, therefore, may require data pre-processing. This is particularly true for bolus fed batch processes (the most commonly used culture strategy in industry), where sharp increases in nutrient concentrations are observed after feeds.

Figure 1 outlines a variant of bolus fed-batch data pre-processing where a shift from the concentration space (i.e. cells, nutrients/metabolites, mAb) to the cell-specific rate space (i.e. cell-specific growth/death, uptake, and secretion rates) increases understanding of the intracellular processes governing the system [33], eliminates discontinuities from the data, and, thereby, enables faster and more robust model training. The cell-specific rates can then be included alongside concentration data for MVDA to yield more robust input/output regression models [9••]. Multivariate regression can also be used to overcome disparities associated with PAT/handling errors or fill in missing information while integrating datasets across different biomanufacturing runs [34]. Covering such data gaps is key in enhancing model robustness.

Figure 1



Data processing for statistical modelling.
Data for viable cell density ($\boldsymbol{X_v}$) is used to compute the integral of viable cells (***IVC***) up to culture time $\boldsymbol{n}$ ($\boldsymbol{t_n}$) the trapezoidal rule. The experimentally determined nutrient and metabolite residual concentrations ($[\boldsymbol{C}]_{\boldsymbol{i,res}}$) are processed to yield cumulative consumed concentrations ($[\boldsymbol{C}]_{\boldsymbol{i,cons}}$). Plotting $[\boldsymbol{C}]_{\boldsymbol{i,cons}}$ against ***IVC*** yields smooth datasets whose slopes are the cell-specific uptake/secretion rates. The dashed lines on the right-hand side graphs represent regressions which can be used to model the cell-specific consumption/production rates.

Hybrid modelling

A synergistic aproach of statistical and mechanistic models offers various advantages by exploiting the robust compartments of each strategy [35]. In hybrid modelling, the biological system is still described by a mechanistic framework (material balances) (Eq. (1)), but the the cell-specific rates (q_i) are defined as statistical expressions with input variables Y and parameters $\alpha_i, \ldots, \varpi_i$ (Eq. (3)).

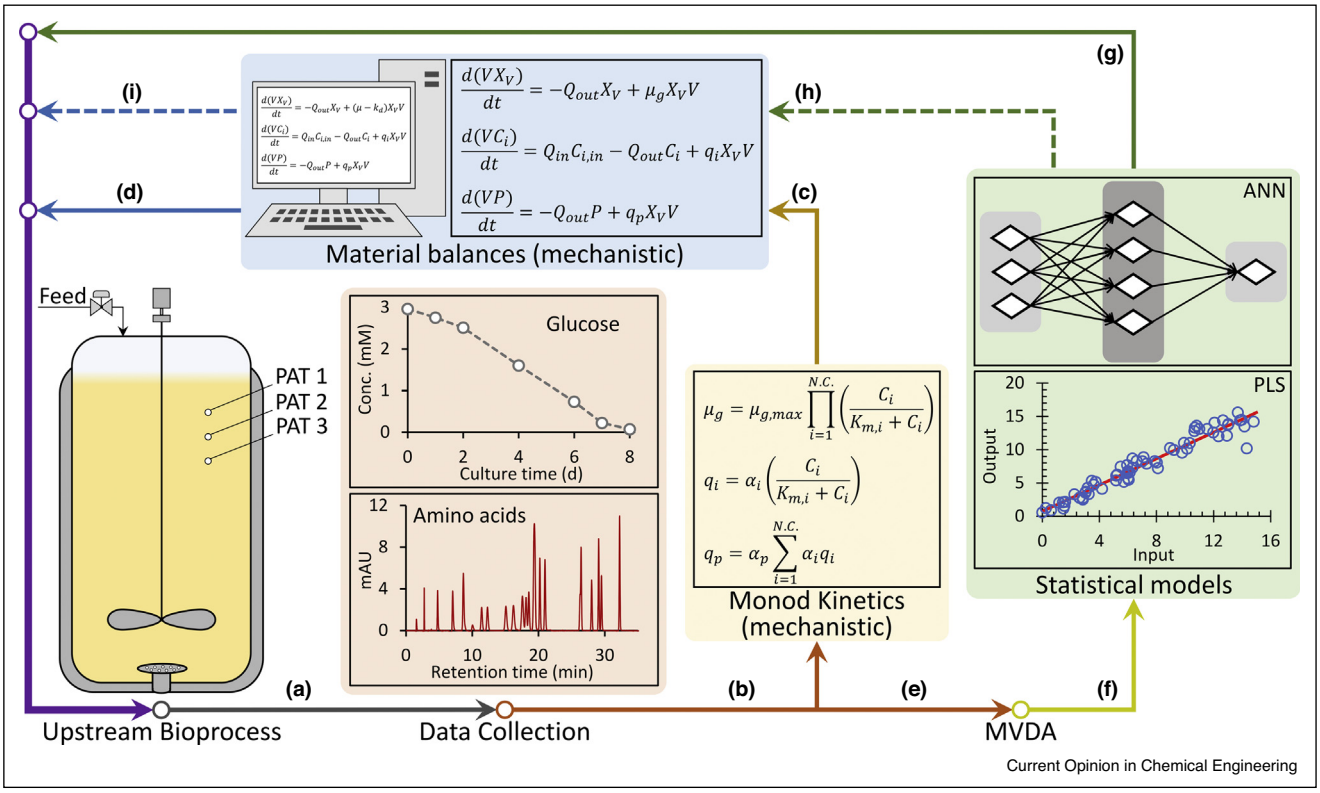
$$\frac{d(V{\cdot}C_i)}{dt} = Q_{in}{\cdot}C_{i,in} - Q_{out}{\cdot}C_i + q_{i,prod}{\cdot}X_v{\cdot}V - q_{i,cons}{\cdot}X_v{\cdot}V \quad (1)$$

$$q_i = \alpha_i + \beta_i{\cdot}Y + \gamma_i{\cdot}Y^2 + \ldots + \varpi_i{\cdot}Y^n \quad (3)$$

Two key advantages arise from this hybrid framework: (1) the solution space for the model is constrained by the material balances and (2) generation of the material balance and statistical cell-specific rate expressions can be automated. A constrained solution space will enhance model predictive capabilities, whereas automated model generation reduces the expertise required for industrial deployment. The structure and parameters of Eq. (3) can be obtained through the workflow outlined in the statistical modelling section, above. A general framework for assembling mechanistic, statistical, and hybrid models for upstream pharmaceutical bioprocesses is summarised in Figure 2.

Narayanan *et al.* [9••] developed a serial hybrid model, where an artificial neural network (ANN) was used for the estimation of the unknown-specific uptake/secretion rates of a bolus fed-batch CHO cell cultures. Once the ANN was trained by integration of experimental data from 81 independent fed-batch culture runs, the hybrid model produced accurate predictions for key nutrient/metabolite concentrations and product titer. It also

Figure 2



Mechanistic, statistical, and hybrid modelling workflows.
Offline and at/on-line PAT data is collected from the bioreactor **(a)** and used to define Monod-based (mechanistic) equations **(b)** which are then embedded within component material balances **(c)** to estimate unknown kinetic parameters ($\mu_{g,max}$, α_i, and $K_{m,i}$) through dynamic optimisation with the experimental data. The parameterised mechanistic model can then be used for bioreactor control and optimisation **(d)**. The collected bioreactor data can be subjected to MVDA **(e)**, the results of which are used to assemble statistical models through PLS regression or machine learning (e.g. ANN) **(f)**. The resulting statistical relationships can be used to directly model bioreactor behaviour **(g)** or can be fed into a material balance framework **(h)** to yield a hybrid model which can be used for bioreactor control and optimisation **(i)**. Notation and symbols are the same as outlined for Eq. (1).

exhibited good interpolation and extrapolation capabilities when compared to PLS regression models.

Kotidis and Kontoravdi [36••] constructed a kinetic/ANN hybrid modelling platform to estimate mAb glycosylation profiles by using intracellular NSD concentrations as inputs for an ANN describing the Golgi glycosylation process. This hybrid model was also used to describe the impact of different manganese feeding strategies on the product's final glycoform distribution. Overall, the resulting hybrid model is less parametrised and, hence, offers robust information about product N-linked glycosylation with substantially reduced experimental effort for model validation.

Hybrid modelling can potentially be used to accelerate process development when combined with design of experiments (DoE) [10,37,38••] where increased process knowledge reduces the number of required experiments and can, therefore, contribute to curbing product/process development costs. An alternative approach would be to deploy mechanistic models to generate training datasets for statistical or ML counterparts. This strategy would not provide enhanced process understanding but has potential applications in bioreactor soft sensing [37].

Hybrid models also have potential in designing optimally informative experiments which can be used to further enhance model performance. Along a similar vein, entity embedding vectors have been combined with Gaussian Process regression on data from sufficiently similar bioprocesses to curb the experimental burden required for model training [39]. Table 1 presents a summary of the pros, cons, and examples of mechanistic, statistical, and hybrid modelling strategies.

Hybrid modelling as a key enabler of Biopharma 4.0

Biopharma 4.0 — the next step in biopharmaceutical manufacturing digitalisation — aims to integrate cyberphysical systems, the industrial internet of things, and smart factories to yield automated end-to-end manufacturing bioprocesses that optimise biopharmaceutical production efficiency by minimising human intervention and the uncertainty which arises therefrom [41].

Purely statistical modelling approaches have greatly contributed towards Biopharma 4.0 objectives by delivering robust data analytics tools as well as enabling real-time monitoring with advanced PAT [24–27] and soft sensors [31,42]. However, their impact on the broader integration of systems within Biopharma 4.0 has been constrained by their lack of causal/mechanistic links between process inputs/outputs. Purely mechanistic models overcome the limitations of purely statistical models, but their complexity, experimental burden for parameterisation, and computational expense has limited their uptake by the biopharmaceutical sector.

Hybrid models overcome the limitations of purely statistical and mechanistic models to, therefore, progress the realisation of Biopharma 4.0. Coupling material balances with statistical models provides direct links between data and physical bioprocess systems to create cyberphysical platforms which facilitate integration of unit operations. Overall, hybrid models have the potential of realising the ultimate aims of Biopharma 4.0 by enabling model-based control [43] and optimisation to enhance the yield [9••,44–46] and quality [36••] of biopharmaceutical products.

Conclusions

In the context of the QbD initiative and Biopharma 4.0, mechanistic and statistical models have gained substantial traction across the biopharmaceutical sector in recent years. Mechanistic models have been successful in optimising upstream pharmaceutical bioprocesses [12•,19,20]. Although they are assembled based on the known chemical, physical, and biological phenomena governing the bioprocess, their widespread deployment has been limited due to the expertise required to construct the models, computational expense, and difficulties in obtaining accurate values for their kinetic parameters. Similarly, multivariate data analysis and regression has been extensively deployed to draw qualitative and quantitative correlations between process inputs and KPIs [21•,22,23•]. When coupled with advanced PAT tools, MVDA has been deployed to quantify CQAs in real time, thus offering new dimensions of process understanding [24–27]. MVDA is also playing a pivotal role in the systematic development of soft sensors [30,42]. These are vital steps towards the digital transformation of bioprocesses in what is fully encapsulated in the Biopharma 4.0 paradigm.

Hybrid modelling, which layers statistical regression onto a material balance framework, delivers the advantages of both mechanistic and statistical techniques while curbing their individual limitations. Hybrid modelling approaches have used cell culture data to build variance-based structural equations for the specific consumption/production rates of key nutrients/metabolites which thereafter can replace the semi-empirical Monod expressions commonly used in kinetic models [9••].

Within these strategies, the material balances constrain the solution space for the statistical modules, considerably reduce the number of kinetic parameters (and ensuing experimental effort required for parameterisation), and lead to enhanced model performance. A further advantage of hybrid models is that their assembly can be readily automated: the statistical model layer can be generated automatically, and material balances are universal. In the era of Biopharma 4.0, smart manufacturing

informed by hybrid models will lead to safer, more efficacious, personalised, and cost-effective biopharmaceuticals that are widely accessible.

Acknowledgement

The authors gratefully acknowledge funding from Science Foundation Ireland through the Solid-State Pharmaceutical Cluster — SFI's Research Centre for Pharmaceuticals (12/RC/2275_P2 SSPC).

Declaration of Competing Interest

The authors report no declarations of interest.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Walsh G: **Biopharmaceutical benchmarks 2018**. *Nat Biotechnol* 2018, **36**:1136-1145.

2. Lalonde M-E, Durocher Y: **Therapeutic glycoprotein production in mammalian cells**. *J Biotechnol* 2017, **251**:128-140.

3. Grilo AL, Mantalaris A: **The increasingly human and profitable monoclonal antibody market**. *Trends Biotechnol* 2019, **37**:9-16.

4. Kaplon H, Reichert JM: **Antibodies to watch in 2021**. *mAbs* 2021, **13**:1860476.

5. Dhara VG *et al.*: **Recombinant antibody production in CHO and NS0 cells: differences and similarities**. *BioDrugs* 2018, **32**:571-584.

6. Goey CH, Bell D, Kontoravdi C: **Mild hypothermic culture conditions affect residual host cell protein composition post-protein A chromatography**. *mAbs* 2018, **10**:476-487.

7. Villiger TK *et al.*: **High-throughput profiling of nucleotides and nucleotide sugars to evaluate their impact on antibody N-glycosylation**. *J Biotechnol* 2016, **229**:3-12.

8. Kyriakopoulos S *et al.*: **Kinetic modeling of mammalian cell culture bioprocessing: the quest to advance biomanufacturing**. *Biotechnol J* 2018, **13**:1700229.

9. •• Narayanan H *et al.*: **A new generation of predictive models: the added value of hybrid models for manufacturing processes of therapeutic proteins**. *Biotechnol Bioeng* 2019, **116**:2540-2549
A hybrid model utilising a single hidden layer ANN for the estimation of specific consumption/production rates which are fed into simple mass balance equations for the prediction of process variables. Once the ANN was trained and validated with 81 indipendant bolus fed-batch data, the hybrid model was able to quantitatively predict changes to cell density, nutrient/metabolite concentrations and product titre.

10. Bayer B, Striedner G, Duerkop M: **Hybrid modeling and intensified DoE: an approach to accelerate upstream process characterization**. *Biotechnol J* 2020, **15**:2000121.

11. Kornecki M, Strube J: **Accelerating biologics manufacturing by upstream process modelling**. *Processes* 2019, **7**.

12. • Kotidis P *et al.*: **Model-based optimization of antibody galactosylation in CHO cell culture**. *Biotechnol Bioeng* 2019, **116**:1612-1626
A mechanistic model consisting of three discrete compartments; one for cell culture dynamics, another for nucleotide sugar metabolism, and a final one for N-linked glycosylation. This modelling framework was employed to identify optimal galactose and uridine feeding regimes that resulted in maximal mAb galactosylation while eliminating negative impacts on product titre.

13. Jimenez del Val I, Fan Y, Weilguny D: **Dynamics of immature mAb glycoform secretion during CHO cell culture: an integrated modelling framework**. *Biotechnol J* 2016, **11**:610-623.

14. Moser A *et al.*: **Mechanistic mathematical models as a basis for digital twins**. *Adv Biochem Eng Biotechnol* 2021, **176**:133-180.

15. Lopez-Meza J *et al.*: **Using simple models to describe the kinetics of growth, glucose consumption, and monoclonal antibody formation in naive and infliximab producer CHO cells**. *Cytotechnology* 2016, **68**:1287-1300.

16. Kiparissides A, Pistikopoulos EN, Mantalaris A: **On the model-based optimization of secreting mammalian cell (GS-NS0) cultures**. *Biotechnol Bioeng* 2015, **112**:536-548.

17. Quiroga-Campano AL, Panoskaltsis N, Mantalaris A: **Energy-based culture medium design for biomanufacturing optimization: a case study in monoclonal antibody production by GS-NS0 cells**. *Metab Eng* 2018, **47**:21-30.

18. Maria G: **Model-based optimization of a fed-batch bioreactor for mAb production using a hybridoma cell culture**. *Molecules* 2020, **25**.

19. György R *et al.*: **Capturing mesenchymal stem cell heterogeneity during osteogenic differentiation: an experimental–modeling approach**. *Ind Eng Chem Res* 2019, **58**:13900-13909.

20. Grilo AL, Mantalaris A: **A predictive mathematical model of cell cycle, metabolism, and apoptosis of monoclonal antibody-producing GS–NS0 cells**. *Biotechnol J* 2019, **14**:1800573.

21. • Sokolov M *et al.*: **Enhanced process understanding and multivariate prediction of the relationship between cell culture process and monoclonal antibody quality**. *Biotechnol Prog* 2017, **33**:1368-1380
In this work, the authors-coupled PLS regression with a genetic algorithm (GA), to create a model which succesfully predicts the effect of media supplement addition and input process pertrubations to the final glycan profile of a monoclonal antibody produced in an Ambr-15® cell culture system.

22. Sokolov M *et al.*: **Sequential multivariate cell culture modeling at multiple scales supports systematic shaping of a monoclonal antibody toward a quality target**. *Biotechnol J* 2018, **13**:1700461.

23. • Zürcher P *et al.*: **Cell culture process metabolomics together with multivariate data analysis tools opens new routes for bioprocess development and glycosylation prediction**. *Biotechnol Prog* 2020, **36**:e3012
This statistical modelling strategy employs MVDA on a bioprocess dataset to identify correlations between CPPs and final product glycans. Two different approaches have been reported: firstly with PLS regression applied to data corresponding to each glycan group separately and secondly the use of a single PLS model for all glycan profiles. Another key feature of this work is the integration of metabolomics data within the framework, which substantially increased the model's predictive capabilities.

24. Metze S *et al.*: **Multivariate data analysis of capacitance frequency scanning for online monitoring of viable cell concentrations in small-scale bioreactors**. *Anal Bioanal Chem* 2020, **412**:2089-2102.

25. Bhatia H *et al.*: **In-line monitoring of amino acids in mammalian cell cultures using raman spectroscopy and multivariate chemometrics models**. *Eng Life Sci* 2018, **18**:55-61.

26. Zavala-Ortiz DA *et al.*: **Support vector and locally weighted regressions to monitor monoclonal antibody glycosylation during CHO cell culture processes, an enhanced alternative to partial least squares regression**. *Biochem Eng J* 2020, **154**:107457.

27. Ohadi K, Legge RL, Budman HM: **Intrinsic fluorescence-based at situ soft sensor for monitoring monoclonal antibody aggregation**. *Biotechnol Prog* 2015, **31**:1423-1432.

28. Antonakoudis A *et al.*: **The era of big data: genome-scale modelling meets machine learning**. *Comput Struct Biotechnol J* 2020, **18**:3287-3300.

29. Narayanan H *et al.*: **Bioprocessing in the digital age: the role of process models**. *Biotechnol J* 2020, **15**:1900172.

30. Green A, Glassey J: **Multivariate analysis of the effect of operating conditions on hybridoma cell metabolism and glycosylation of produced antibody**. *J Chem Technol Biotechnol* 2015, **90**:303-313.

31. Goldrick S *et al.*: **On-line control of glucose concentration in high-yielding mammalian cell cultures enabled through oxygen transfer rate measurements**. *Biotechnol J* 2018, **13**: e1700607.

32. Kontoravdi C, Jimenez del Val I: **Computational tools for predicting and controlling the glycosylation of biopharmaceuticals**. *Curr Opin Chem Eng* 2018, **22**:89-97.

33. Richelle A *et al.*: **Analysis of transformed upstream bioprocess data provides insights into biological system variation**. *Biotechnol J* 2020, **15**:2000113.

34. Mante J *et al.*: **A heuristic approach to handling missing data in biologics manufacturing databases**. *Bioprocess Biosyst Eng* 2019, **42**:657-663.

35. Bayer B *et al.*: **Comparison of modeling methods for DoE-based holistic upstream process characterization**. *Biotechnol J* 2020, **15**:1900551.

36. •• Kotidis P, Kontoravdi C: **Harnessing the potential of artificial neural networks for predicting protein glycosylation**. *Metab Eng Commun* 2020, **10**:e00131
An ANN/kinetic modeling framework predicting the dynamics of protein glycosylation under different manganese supplementation feeding and/or glycoengineering strategies. Once the neural network was trained and validated with five independent training/validation datasets, it was integrated with the model of reference [11] to succesfully overcome the limitations of the kinetic module.

37. Solle D *et al.*: **Between the poles of data-driven and mechanistic modeling for process operation**. *Chemie Ingenieur Tech* 2017, **89**:542-561.

38. •• Möller J *et al.*: **Model-assisted design of experiments as a concept for knowledge-based bioprocess development**. *Bioprocess Biosyst Eng* 2019, **42**:867-882
With this novel approach, the authors designed a model-assisted DOE methodology which utilized information collected by a mathematical framework, hence decreasing the extent of required experiments. The authors validated their strategy by experimentally reproducing the optimal scenarios obtained with the model. In addition, they applied their mDOE to optimise culture media and bolus feeding strategies.

39. Hutter C *et al.*: *Knowledge Transfer Across Cell Lines Using Hybrid Gaussian Process Models with Entity Embedding Vectors*. 2020.

40. Craven S, Whelan J, Glennon B: **Glucose concentration control of a fed-batch mammalian cell bioprocess using a nonlinear model predictive controller**. *J Process Control* 2014, **24**:344-357.

41. Steinwandter V, Borchert D, Herwig C: **Data science tools and applications on the way to Pharma 4.0**. *Drug Discov Today* 2019, **24**:1795-1805.

42. Randek J, Mandenius C-F: **On-line soft sensing in upstream bioprocessing**. *Crit Rev Biotechnol* 2018, **38**:106-121.

43. Papathanasiou MM, Kontoravdi C: **Engineering challenges in therapeutic protein product and process design**. *Curr Opin Chem Eng* 2020, **27**:81-88.

44. Papathanasiou MM *et al.*: **Advanced model-based control strategies for the intensification of upstream and downstream processing in mAb production**. *Biotechnol Prog* 2017, **33**:966-988.

45. Kappatou CD *et al.*: In *Sequential and Simultaneous Optimization Strategies for Increased Production of Monoclonal Antibodies, in Computer Aided Chemical Engineering.* Edited by Kiss AA. Elsevier; 2019:1021-1026.

46. Kroll P *et al.*: **Model-based methods in the biopharmaceutical process lifecycle**. *Pharm Res* 2017, **34**:2596-2613.