frontiers | Frontiers in Chemical Engineering

TYPE Review
PUBLISHED 15 May 2023
DOI 10.3389/fceng.2023.1157889

OPEN ACCESS

EDITED BY
Yinlun Huang,
Wayne State University, United States

REVIEWED BY
Alexandros Kiparissides,
Aristotle University of Thessaloniki, Greece
Ioscani Jiménez Del Val,
University College Dublin, Ireland

*CORRESPONDENCE
Alessandro Butté,
a.butte@datahow.ch

RECEIVED 03 February 2023
ACCEPTED 20 April 2023
PUBLISHED 15 May 2023

CITATION
Narayanan H, von Stosch M, Feidl F, Sokolov M, Morbidelli M and Butté A (2023), Hybrid modeling for biopharmaceutical processes: advantages, opportunities, and implementation.
Front. Chem. Eng. 5:1157889.
doi: 10.3389/fceng.2023.1157889

Hybrid modeling for biopharmaceutical processes: advantages, opportunities, and implementation

Harini Narayanan[1], Moritz von Stosch[2], Fabian Feidl[2], Michael Sokolov[2], Massimo Morbidelli[2] and Alessandro Butté[2]*

[1]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, United States, [2]DataHow AG, Zurich, Switzerland

Process models are mathematical formulations (essentially a set of equations) that try to represent the real system/process in a digital or virtual form. These are derived either based on fundamental physical laws often combined with empirical assumptions or learned based on data. The former has been existing for several decades in chemical and process engineering while the latter has recently received a lot of attention with the emergence of several artificial intelligence/machine learning techniques. Hybrid modeling is an emerging modeling paradigm that explores the synergy between existing these two paradigms, taking advantage of the existing process knowledge (or engineering know-how) and information disseminated by the collected data. Such an approach is especially suitable for systems and industries where data generation is significantly resource intensive while at the same time fundamentally not completely deciphered such as the processes involved in the biopharmaceutical pipeline. This technology could, in fact, be the enabler to meeting the demands and goals of several initiatives such as Quality by design, Process Analytical tools, and Pharma 4.0. In addition, it can aid in different process applications throughout process development and Chemistry, Manufacturing, and Control (CMC) to make it more strategic and efficient. This article focuses on providing a step-by-step guide to the different considerations to be made to develop a reliable and applicable hybrid model. In addition, the article aims at highlighting the need for such tools in the biopharmaceutical industry and summarizes the works that advocate its implications. Subsequently, the key qualities of hybrid modeling that make it a key enabler in the biopharmaceutical industry are elaborated with reference to the literature demonstrating such qualities.

KEYWORDS

hybrid modeling, artificial intelligence, machine learning, biopharmaceutical, process development

1 Introduction

Biopharmaceuticals are a growing class of therapeutic solutions attaining a global market size of USD 389.6 billion in 2021 and are projected to attain a market size of USD 856.9 billion by 2030. The number of biologics approved by the Food and Drug Association (FDA) and European Medicine Agencies (EMA) is increasing consistently with a surge of several different therapeutic modalities beyond the prevalent monoclonal

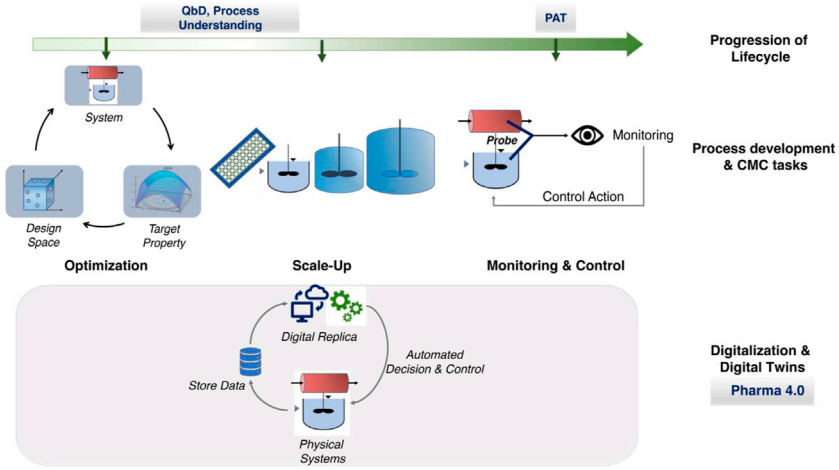

FIGURE 1
Schematic representation of the key process development and CMC tasks throughout the biopharmaceutical lifecycle with an indication of the key initiatives and goals of biopharma and the regulatory agencies.

antibodies (mAbs), recombinant proteins, and antibody-drug conjugates (ADCs) (Mullard, 2023). Some such modalities include mRNA-based vaccines (Sahin et al., 2014; Bhat et al., 2021; Qin et al., 2022), cell and gene therapies, extracellular vesicles (EVs) (Murphy et al., 2019; Klyachko et al., 2020; Bertolino et al., 2022), and living materials (Gilbert and Ellis, 2019; Rodrigo-Navarro et al., 2021), with the latter two still under development and have not reached the approval phase, yet (Walsh and Walsh, 2022).

Despite the increasing market sizes and approvals (Walsh, 2018; Walsh and Walsh, 2022), the development of such biologics is typically a slow and resource-intensive process taking up to 10–12 years and an investment of at least 2 billion USD (Narayanan et al., 2021a; Cardillo et al., 2021). Among these estimated times and costs, process development accounts for a significant portion taking up to 4 years (30% of the time to market) and an average investment of ~100 million USD (Cardillo et al., 2021) making it considerably resource-intensive, having important consequences in drug manufacturing. Additionally, the challenges for process development increase manifold when considering biosimilars where the complexity of matching drug quality is high in addition to the high time pressure due to competition.

In this regard, the key tasks associated with biopharmaceutical process development and Chemistry Manufacturing and Control (CMC) are: 1) Design and Optimization: To identify the suitable process parameters to produce a product with the desired quantity and quality attributes, 2) Scale-Up: To reproduce observations/attributes achieved in lab scale operations and also at commercial scale, 3) Monitoring and Control: Maintain the desired state/trajectory of the process. The different tasks in process development and CMC with the progression of the biopharmaceutical lifecycle are schematically represented in Figure 1 with the corresponding indication of the relevant initiatives.

In a highly multi-dimensional design space, where the interactions between the governing variables are complex, trial and error-based or restricted operations enforce sub-optimality and risk of not meeting quality specifications. In addition, it also increases the time and cost of development if such activities originate from an unfavorable region originating from an inefficient transfer of past learning. Currently, such transfers are performed as heuristics or conclusions determined based on observations drawn from particular cases (Narayanan and Love, 2022). For instance, working in a narrow range of pH or temperature based on previous product lines or using specific basal and feed media for all the products produced by a given host. Though the Quality by Design paradigm has come some way to reuse knowledge, mostly in the form of technical risk assessments, knowledge transfer is limited and despite the wide spread of platform processes, to some degree, the process pipeline needs to be developed from scratch, at least in parts, for every new molecule.

The increasing number of modalities and varieties within the same modality requires that the same development steps be undertaken multiple times. A standardized workflow for carrying out these steps is required to accelerate the pipeline while reducing the resources. What we envision through a standardized protocol is a guideline to the series of steps to be undertaken for each new product with an unbiased, generalized transfer of learnings from one case to another.

As a result, we require that all the data and learning acquired under a certain product are 1) strategically collected through efficient designs covering the design space, and 2) formalized via relevant mathematical modeling to capture the overall patterns in the data (Narayanan and Love, 2022). In addition to the goals of process development, such formalization of information and knowledge are inevitable in the digital era where all the process industries including biopharma are interested in digitalizing their operation and production, moving towards digital twins and smart factories (Steinwandter et al., 2019; Narayanan et al., 2020b; Chen et al., 2020; Gargalo et al., 2020; Sokolov, 2020; Zobel-Roos et al., 2020; Narayanan et al., 2021d).

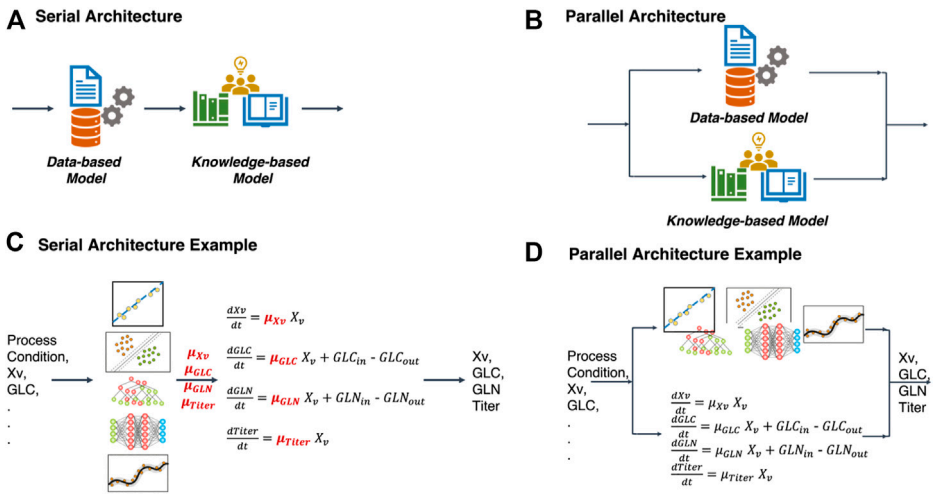

FIGURE 2
Schematic representation of different architectures of hybrid models, **(A)** serial and **(B)** parallel architecture, illustrated with an example **(C)**- Serial architecture, **(D)**- Parallel architecture.

To this aim, artificial intelligence/machine learning (AI/ML) are suitable techniques to learn patterns in data collected from high-dimensional design spaces with complex non-linear interactions among different factors (Narayanan et al., 2021a). However, using AI/ML solely requires good quality data to be collected in a considerable quantity to be able to develop useful and applicable models for the intended purposes of biopharmaceutical application. In contrast, biopharma is a data-limited industry specifically in terms of actively generated data, where each experiment and analytics is resource-intensive, thus, limiting the number of experiments and the corresponding data that can be generated (von Stosch et al., 2021). In addition, though the biopharmaceutical industry has a lot of reserves of the so-called "historical data," were not collected for the purpose of training an AI/ML algorithm. Subsequently, they harbor a lot of bias in terms of the design spaces explored and the information recorded as a result affecting the overall quality of the pre-existing data in the industry. These challenges have been described in depth in the following commentary article (Narayanan and Love, 2022).

However, data is not the only source of information available in biopharma. This is especially true for process applications where a basic understanding of the unit operations and some level of abstraction regarding the physicochemical phenomena at a unit operation level are available. However, this knowledge is not complete or sufficient to build a solely knowledge-based model.

Hybrid modeling has recently found increasing interest as it attains a balance between the purely data-based modeling offered by AI/ML and purely knowledge-based modeling (mostly less representative given the existing gaps in understanding the system). As a result, it presents remarkable qualities, detailed and summarized in Section 3, that make it suitable for all the process goals of the biopharmaceutical industry as highlighted in several works. For instance, (von Stosch et al., 2014; von Stosch et al., 2016), presents hybrid modeling as a solution for QbD and PAT, (Schubert et al., 1994a; Galvanauskas et al., 2004; Teixeira et al., 2006), for process optimization, (Bayer et al., 2021), for scale-up, (Schubert et al., 1994a; Chen et al., 2000; Galvanauskas et al., 2004; Teixeira et al., 2007; Ferreira et al., 2014; Narayanan et al., 2020a; Narayanan et al., 2020b), for process monitoring and control, (Sokolov, 2020; Narayanan et al., 2021d), for digitalization, and (Sokolov et al., 2021) for digital twins.

This article focuses on summarizing the different types of hybrid models that can be developed and subsequently highlighting the enabling qualities of such models that make them attractive for different process applications, following which a detailed step-by-step guideline of conceptual considerations to be made while developing these are provided.

2 Types of hybrid models

Though not so widely emphasized then, the hybrid modeling concept was already introduced in 1992 in some seminal works such as (Psichogios and Ungar, 1992; Schubert et al., 1994a; Thompson and Kramer, 1994). Traditionally, hybrid modeling has been categorized to have two architectures: serial and parallel architectures. As summarized in Figure 2A, the serial architecture essentially involves modeling fragments of the knowledge-based models using data-based models. In other words, the knowledge-based models rely on inputs from data-based modeling. On the other hand, in a parallel architecture (Figure 2B), both the knowledge-based and data-driven models are set up simultaneously and aggregation of the prediction from both is considered (Thompson and Kramer, 1994).

The knowledge-based models are mathematical representations of the engineering know-how and process knowledge typically in the form of system differential-algebraic equations. The data-based models, on the contrary, are statistical or machine learning models capable of learning from the patterns in the data instead of relying on prior knowledge. Figures 2C, D further illustrate this

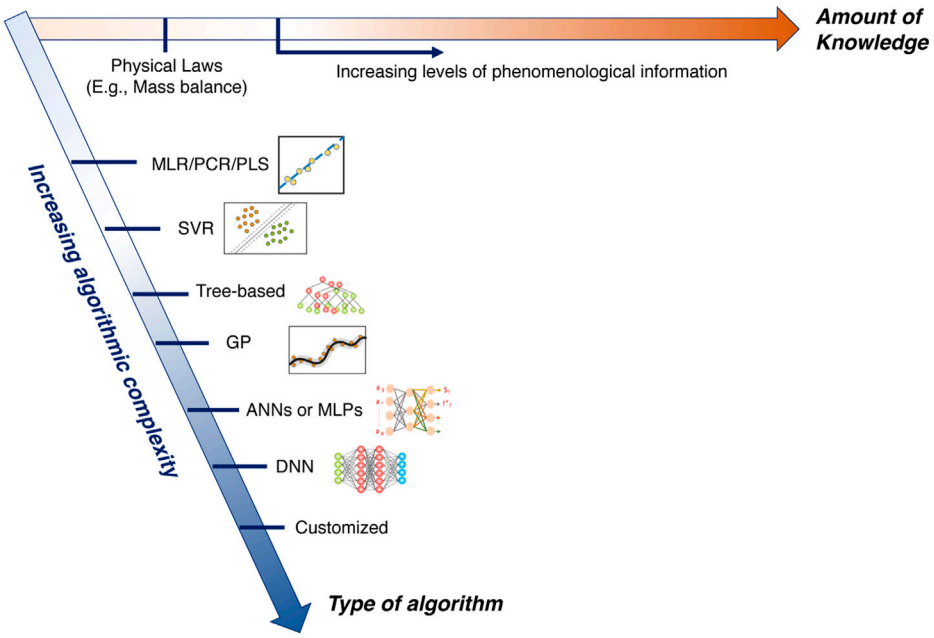

FIGURE 3
Schematic representation of the selection along two key axes to develop hybrid models: (i) type of data-based algorithm and (ii) amount of knowledge incorporation.

with a simple example of the case of a bioreactor both of which are extremely relevant processes in the biopharma industry.

Since its inception, several works have demonstrated the successful implementation of hybrid models for different types of bioprocesses with a large fraction of it based on the serial architecture (von Stosch et al., 2016; Narayanan et al., 2019; Bayer et al., 2021; Kotidis et al., 2021; Narayanan et al., 2021c; Narayanan et al., 2021b; Cruz-Bournazou et al., 2022; Narayanan et al., 2022b; Narayanan et al., 2022a; Nold et al., 2023). The serial architecture is in particular advantageous owing to the fact that it constrains and segregates the information learned by the data-based model. Subsequently, the models learned are more robust even with fewer data. The parallel architecture on the other hand trains the data-based model independently of the available knowledge, thus retaining all the downsides of a pure data-based model. It is here worth noting that, the current article will focus only on serial architecture.

Furthermore, the plethora of work on such serial architecture-based hybrid modeling can be further categorized mainly along two axes, as schematically illustrated in Figure 3. The first axis, more commonly exploited in the literature, is the type of machine learning/statistical algorithm used in data-based modeling. Several options presented in the literature include Artificial Neural networks or multi-layer perceptron (ANNs or MLPs) (Psichogios and Ungar, 1992; Schubert et al., 1994b; Schubert et al., 1994b; Feyo de Azevedo et al., 1997; Chen et al., 2000; Teixeira et al., 2006; Von Stosch et al., 2012a; Narayanan et al., 2019; Bayer et al., 2021; Narayanan et al., 2021b; Narayanan et al., 2021c; Narayanan et al., 2022b), Principal Component Regression (PCR) (Okamura et al., 2022), Partial Least Square regression (PLSR) (Von Stosch et al., 2011; Von Stosch et al., 2012b; Carvalho et al., 2022), Tree-based models (Hutter et al., 2017), Gaussian processes (GPs) (Hutter et al., 2021; Vega-Ramon et al., 2021; Cruz-Bournazou et al., 2022), and Deep Neural Networks (DNNs) (Pinto et al., 2022). Subsequently, efforts have also been devoted to using techniques such as symbolic regression and customized neural networks as data-based modeling approaches to have enhanced interpretability compared to the traditional ML approaches (Narayanan et al., 2022a; Doyle et al., 2023). The second axis is the extent to which prior knowledge is available and incorporated into the hybrid model. Conventionally, this was fixed prior to the development of the hybrid modeling to a pre-determined extent of knowledge support. In our recent works, we highlighted this and introduced the concept of the "degree-of-hybridization" (Narayanan et al., 2021b; Narayanan et al., 2022b) whereby the determinantal effects of introducing too little or too much knowledge were emphasized. Subsequently, we illustrated its implications for two systems of relevance for biomanufacturing, namely, the cell culture (Narayanan et al., 2022b) and the chromatographic process (Narayanan et al., 2021b). To summarize, for a given modeling task at hand, a family of hybrid models can be developed across these two dimensions: 1) type of data-based algorithm and 2) extent of knowledge incorporation.

Hybrid models are at the intersection of purely knowledge-based models and purely data-based models. As a result, it alleviates many of the shortcomings faced by these modeling paradigms individually. The next section describes the qualities of the hybrid models that make it attractive for different process applications and summarizes the literature evidences for the same.

3 Enabling qualities of hybrid models

Hybrid modeling finds a trade-off between constraining the relationships between the governing variables while allowing for the flexibility to learn potentially unknown interactions between them (Narayanan et al., 2021b; Narayanan et al., 2022b). This is especially useful when applied to a real system with many influencing factors that present complex interactions which are only partially understood. Some of the key qualities of hybrid models that have been rationalized and also demonstrated quantitatively are summarized below.

Since the ML model isn't solving the entire problem, it doesn't need as much data, it only needs to learn what the physics model can't explain well.

3.1 Better accuracy

A purely knowledge-based model suffers the risk of constraining too much based on biased assumptions while a purely data-driven model suffers from the excessive flexibility to learn any (even arbitrary) relationships based on data. The latter results in the requirement of a large amount of data to learn a generalized and representative model of the considered system. Given the ability of hybrid models to draw a trade-off between constraining and flexibility, in a limited data set-up such as biopharma, hybrid models in general shows better accuracy compared to either knowledge-based or data-based models.

Several hybrid modeling literature in the area of biotechnology (fermentation and cell culture) have compared their performance to ANN (Schubert et al., 1994b; Feyo de Azevedo et al., 1997) or recurrent neural networks (Feyo de Azevedo et al., 1997; Ferid et al., 2020) with the hybrid models consistently outperforming the other two. Subsequently, in many of the recent works including ours, the accuracy of hybrid models has been quantitatively shown to outperform the current industrial benchmark, PLSR1 (Von Stosch et al., 2012b; Narayanan et al., 2019; Narayanan et al., 2021d; Narayanan et al., 2022b). Furthermore, in our hybrid modeling work for bio chromatographic application, the accuracy of hybrid models was around 40% better than the typically used mechanistic model (Lumped kinetic model) for the protein-A capture step (Narayanan et al., 2021b; Narayanan et al., 2021c; Narayanan et al., 2021d). Such better prediction accuracy by the hybrid models correspondingly results in superior performance in relevant process applications. This has been quantitatively demonstrated, for instance, for process monitoring and control by Narayanan et al. (2020a), and von Stosch et al. (2012a) comparing against PLSR1 and ANN, respectively.

3.2 Lower data requirement

An additional advantage of the hybrid model is its ability to efficiently model systems where data generation capacity is limited. In theory, from a parameter estimation standpoint, knowledge-based models are probably the ones that require the least amount of data if specific and independent analytics exist for the evaluation of the different physically relevant parameters. In contrast, to estimate parameters using macroscopic data, a significantly large number of experiments might still be required to decorrelate the different parameters and estimate their value with reasonable certainty. In addition, a lot of experiments (and maybe more specialized experiments and analytics) must be performed *a priori* to generate and gather the required, detailed, and ideally unbiased knowledge to develop such a model in the first place. For most biopharmaceutical systems such unbiased generalized knowledge for different hosts and therapeutic modalities has not yet been deciphered. On the other hand, the flexibility offered by data-driven models requires a lot of data to establish a reliable model. Subsequently, hybrid models have shown superior predictive and extrapolation (c.f. Section 4.3) capabilities with much fewer data compared to their data-driven counterparts (Narayanan et al., 2022b). In our recent work, data requirements for hybrid models are compared against the PLSR models as well as the ANN with the former requiring much fewer data compared to either of the data-based alternatives (Narayanan et al., 2022b). This is rationalized by the fact that the knowledge-based part of the hybrid model constrains or segregates the problem to be learned by the ML (or data-based) model thus compensating for the reduced data.

In addition, it is worth noting here that the requirement of data to train a reliable model can be further minimized by using an appropriate design of experiments. In the industrial setting, it is standard practice to either perform experiments in a very narrow sub-region or use a classical design of experiments strategy (most often factorial designs or response surface models) (Narayanan and Love, 2022). However, such designs are extremely inefficient in capturing the non-linear interrelationships between variables (since they are based on perturbing corners and centers of the design space) thus resulting in poor representations for the model to learn (Narayanan and Love, 2022). In this direction, in our recent work, we have quantitatively demonstrated the reduction in the amount of data required and the difference in model performance between using a fractional factorial design as opposed to a uniform space-filling design such as Latin hypercube sampling [c.f. SI of (Narayanan et al., 2022b)].

3.3 Extrapolation and process optimization

In addition to improved predictive accuracy and reduced data burden, hybrid models also outperform their alternatives in extrapolation. Extrapolation capability is key for process optimization and the ability of the models necessitated to robustly predict beyond their training regimes. It is here noted that extrapolation is implied in terms of "operation" such as a change of feeding strategies, change of mode of operation, or expanding the boundaries of process parameters. The strong knowledge support in such models prevents the models from learning and subsequently predicting unrealistic behaviors when used for prediction outside their training region which has been a classical skepticism around the purely data-driven model. In theory, the knowledge-based models should be the most efficient in extrapolation as they are based on physical principles. However, this is true only if the knowledge-based models are unbiased and do not rely on empirical assumptions based on a case-specific basis (which unfortunately is often the case in fields such as bioprocesses).

The superior performance of hybrid models for extrapolation has been quantitatively demonstrated in the literature (Van Can et al., 1998; Narayanan et al., 2019; Narayanan et al., 2021b;

Narayanan et al., 2021c; Narayanan et al., 2021d; Narayanan et al., 2022a; Narayanan et al., 2022b). Subsequently, these capabilities have resulted in the successful application of hybrid models for process applications as highlighted, for instance, by Bayer et al. (2021) for scale transferability and Teixeira et al., for process optimization (Schubert et al., 1994a; Galvanauskas et al., 2004; Teixeira et al., 2006; Ferreira et al., 2014).

3.4 Physically relevant behavior and interpretability

Finally, hybrid models channel the information learned by the data-driven models to lumped physical parameters, resulting in physically relevant behaviors being learned and predicted by the model. Subsequently, the qualitative trends can be interpreted from these models. For instance, it has been previously shown how the imposition of mass balance in a bioreactor model results in the hybrid model learning biologically consistent glucose consumption patterns, as opposed to a purely statistical model which fails to learn such behavior [we refer the readers to (Narayanan et al., 2019) and (Narayanan et al., 2021d) for figures and in-depth discussions]. In addition, (Narayanan et al., 2022b), have also demonstrated the ability to extract other types of lumped kinetic information such as the glucose consumption rates, maintenance rates, etc., as well from such hybrid modeling approaches. Similarly, hybrid models for bio-chromatographic columns have been shown to learn relevant adsorption equilibrium representations (i.e., isotherms) when trained only on the breakthrough data (Narayanan et al., 2021b; Narayanan et al., 2021c). The ability to extract such information from a hybrid model makes it extremely attractive for process understanding applications and subsequently to attain the goals of QbD initiatives.

4 Steps to set up hybrid models

Subsequently, the following section will lay out the key step-by-step considerations for the development of hybrid models.

4.1 Define the modeling task

Prior to model development, the definition of the modeling task is key, irrespective of the type of model that is, being developed (mechanistic, data-driven, or hybrid models) (Bonvin et al., 2016). This is, in turn, determined by the associated application and process goals. Once the process goal and the corresponding modeling task are identified, the next step is to determine the inputs and outputs of the model.

This is typically determined by the ability to quantify different relevant variables. For instance, in most biopharmaceutical firms, in the cell culture (or bioreactor), macroscopic variables such as the concentration of extracellular metabolites, products, and cells are quantified in contrast to intracellular quantities such as the ones obtained through omics techniques. This constrains the inputs and outputs that can be used for the modeling tasks. In addition, if historical data is being used, the nature of the available dataset (missing information, frequency, *etc.*) determines the capacity of modeling and also if a reasonably reliable model can be set up, to begin with (Narayanan and Love, 2022).

Furthermore, which variables can be included as input in a model and how they are to be introduced may be for different applications. For instance, process optimization requires a modeling setup where the entire evolution of a process variable can be predicted based on just the initial condition. On the other hand, for process monitoring online or real-time information and variables can be used in the models.

Once the modeling task is defined and suitable inputs and outputs are decided, the model development process can commence. It is here noted that after the input and outputs have been identified there might be a requirement for certain data pre-processing such as missing data considerations, and frequency alignment of data points collected. Typically, for purely data-driven modeling certain data transformations are also performed to attain certain distribution or to achieve a certain type of non-linearity (typically in the linear model). In the case of hybrid modeling, the existing knowledge fragments often dictate or compartmentalize the functional form of the primary input, which might undergo transformations as required within the knowledge-based/mechanistic framework (Tsopanoglou and Jiménez del Val, 2021).

4.2 Identifying knowledge fragments and their mathematical representation

The first step in the development of a hybrid model involves the identification of the engineering know-how and the process knowledge available for the system under study. Subsequently, these knowledge fragments must be organized based on increasing levels of certainty. For instance, knowledge based on physical laws (e.g., mass balance, energy balance, thermodynamics) holds higher validity as compared to empirical assumptions (e.g., Monod kinetics for cell metabolism representation, Langmuir isotherm to represent adsorption) which may be valid for ideal, simple systems but may not hold try for complex systems.

Once the knowledge fragments are identified and sorted based on their certainty, their respective mathematical representations/formulation must be determined. The top level of such framework typically can be represented through a system of ordinary differential equations (e.g., to represent homogeneous macro-kinetic representation of cell cultures), partial differential equations (e.g., to represent chromatographic process), algebraic equations (e.g., to represent a system in steady state) or a mixture of these forms summarized as differential-algebraic equations. Within these overarching frameworks, other knowledge fragments assume various functional forms. For instance, linear representation to account for the proportional dependence of cell growth rate (nutrient consumption rate, and product production rate) on the viable cell density. Other manifestations of the knowledge could be in form of constraints, lower and upper bounds of different input or target variables, or even channeling the inputs received by data-based algorithm learning the missing information (discussed below in Section 3.2).

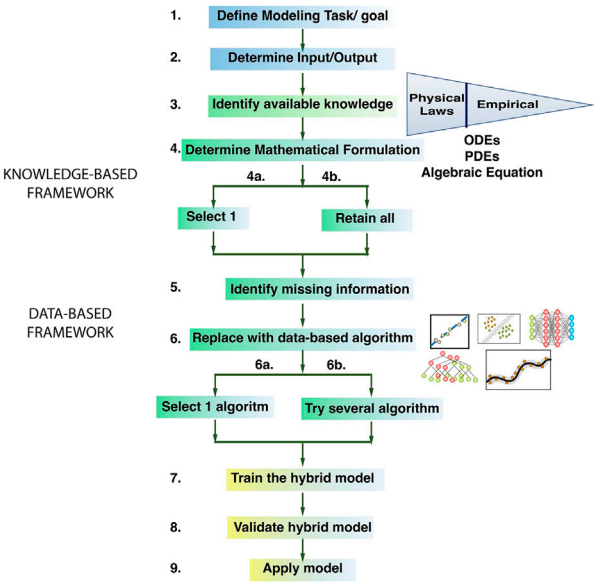

FIGURE 4
Flowchart representation of the step-by-step conceptual guide to developing a hybrid model.

It is here noted that the extent of knowledge to be incorporated, can be pre-determined (as popularly done in literature) or a family of hybrid models with different fractions of knowledge can be developed. In the later stage, once the models are trained, a choice can be made based on the performance of the different models on a validation set. An alternative option is to ensemble the prediction from all viable models (ignoring biased models that would result in poor performance already during the training phase) via stacking (Bishop, 2006; Hastie et al., 2017).

In the case that a pre-determined extent of knowledge is incorporated it must be ensured to avoid over-emphasizing the empirical assumptions-based knowledge fragments since they risk producing an extremely biased model. Though the flexibility of the data-based ML model might compensate for the biased or wrong knowledge enforced, it could lead to wrong interpretations, wrong extrapolations, and possibly require more data to correct for the misleading knowledge backbone imposed in the hybrid model. Thus, it is advised to introduce knowledge, that is representative of the physical correctness of the model. For instance, species-specific mass balances to account for the feed inlets and outlets are absolutely certain and can be incorporated into the knowledge-based model. Such similar knowledge fragments that aid in the learning of physically consistent behaviors can be incorporated into the knowledge-based layer of the hybrid model (Figure 4).

4.3 Identifying missing information and replacing it with data-based models

Subsequently, the fragments missing a mathematical representation based on unbiased prior knowledge or physico-chemical understanding are identified. These fragments are then modeled using statistical or machine-learning algorithms which offer a plethora of options from the simplest in the form of multi-linear regression to a latent space-based algorithm such as PLSR to complex non-linear models ranging from tree-based algorithms (XGBoost, Random Forest, etc.) to kernel-based methods (SVMs, GPs) to neural network-based algorithms (MLP, DNN) (Bishop, 2006; Hastie et al., 2017).

The choice of the algorithm can be made based on a pre-determined rationale or a family of hybrid models can be developed considering the different options of the machine learning algorithm. Currently, the literature is based on the former where the machine learning algorithm is fixed *a priori*, for instance, based on its ability to flexibly learn the underlying unknown function. The latter is still a viable option where a single model can be selected at a later step (i.e., using a validation test to evaluate the performance of hybrid models developed using different algorithms.) or all the models can be retained to perform ensemble modeling via stacking (Bishop, 2006; Hastie et al., 2017).

4.4 Training hybrid model(s)

Once the knowledge backbone and the algorithm to replace the missing information are determined, the hybrid modeling backbone is in place. The next is to train these models to optimize them for the considered system, from which relevant data is generated. There can be two approaches to training such hybrid models. The first involves substituting the parametric form of the machine learning algorithm into the mathematical formulation of the knowledge backbone and solving the system of differential-algebraic equations using appropriate numerical techniques. The other approach is to reorganize the terms in the hybrid model formulation such that the entities learned by a machine learning model can be segregated or collected as a function of all the inputs to a machine learning model. Subsequently, the hybrid models can be trained like any typical machine learning model (Bishop, 2006; Hastie et al., 2017). It is here worth noting that when the choice of the machine learning algorithm is non-parametric (e.g., tree-based algorithms or Gaussian processes), the latter approach is the only possible option since such algorithms lack a parametric representation.

4.5 Validate hybrid model(s)

The choice of the ML algorithm enforces the need to select certain hyperparameters dependent on the chosen algorithm. For instance, the number of layers, the number of hidden nodes in each layer, and the transfer function are some of the hyperparameters associated with an ANN; Gaussian processes have the kernel type and mean function type as the key hyperparameters, and so on and so forth for other ML algorithms. This requires the use of a validation set (or an internal test set) to select the relevant hyperparameter for the given study case.

In addition, the performance of the model on this validation set(s) can also be used to determine the most suitable ML algorithm and the key elements of knowledge fragments, in case a series of choices was retained in Sections 3.2, 3.3.

Summarizing, the final hybrid with the optimal corresponding architecture is chosen in this step which can subsequently be used for intended applications.

4.6 Apply model

Once the final model is calibrated or developed it is ready to be used for intended applications. Though this seems straightforward from a scientific perspective, a number of practical hurdles might make the application of hybrid models still challenging (Canzani and Timmer, 2021). However, it can be expected that with the ever-faster advancement of technologies, these hurdles become significantly smaller over time.

5 Conclusion

The biopharmaceutical industry is currently faced with the need to, on the one hand, accelerate process development and CMC activities for the new therapeutic modalities while on the other hand thriving to increase the production of an existing molecule to meet global demands. Additionally, it can benefit from the reduction in the market price of these therapeutics making them accessible to a larger population. This can be realized by enhancing the efficiency of process development and CMC, subsequently, reducing the development costs and thus the marketed price.

The growing capacities of mathematical algorithms, especially derived from AI and ML, have opened up many possibilities that can be utilized to achieve several of the aforementioned targets of the industry. A key bottleneck in this regard, however, is the ability of the biopharma industry to generate a large amount of data to develop such tools with reliability. The hybrid modeling paradigm is emerging as a pragmatic solution to enable all ML benefits while using limited data, especially for process development and CMC activities.

In this article, the key qualities of hybrid modeling that make it a suitable approach have been highlighted and summarized. The major aspects include lower data requirements, better predictive accuracy, extrapolation capabilities, and comparative ease of interpretability and physical relevance. Subsequently, a step-by-step guideline to develop a hybrid model has been detailed to provide a series of conceptual checklists to be considered to build a reliable and robust model for an intended purpose.

Despite the advancements on the algorithmic end to take into account knowledge to reduce data requirements, the quality of data collected is still to be ensured. It is required that the paradigm of experimental design and data generation is modified to justify the transition for process modeling. The past/current experimental design strategies rely heavily on one-factor-at-a-time approaches or performing statistical DoE with a small subset of design variables which makes it easier for human interpretation. However, in order for the process models to represent the physical system reliably, the effect of multivariate interactions must be studied and data must be collected. In addition, the existing paradigm of decoupling experimental design/data collection and modeling results in increased amounts of data points (especially by probing the unfavorable part of the design spaces). As a result, alternative approaches such as active learning coupled with iterative experimental design would be impactful and thus could be the future direction of research in this area. A more thorough sketch of these challenges and plausible solutions for the deployment of process models in the biopharmaceutical industry has been highlighted in the recent commentary article (Narayanan and Love, 2022).

Author contributions

HN: Prepare the original draft of the manuscript; MvS, FF, and MS: Review of the manuscript; AB and MM: Review of the manuscript and supervision. All authors contributed to the article and approved the submitted version.

Conflict of interest

Authors MvS, FF, MS, MM, and AB were employed by DataHow AG.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bayer, B., Duerkop, M., Striedner, G., and Sissolak, B. (2021). Model transferability and reduced experimental burden in cell culture process development facilitated by hybrid modeling and intensified design of experiments. *Front. Bioeng. Biotechnol.* 9, 740215–740312. doi:10.3389/fbioe.2021.740215

Bertolino, G. M., Maumus, M., Jorgensen, C., and Noël, D. (2022). Recent advances in extracellular vesicle-based therapies using induced pluripotent Stem cell-derived mesenchymal stromal cells. *Biomedicines* 10, 2281. doi:10.3390/biomedicines10092281

Bhat, B., Karve, S., and Anderson, D. G. (2021). mRNA therapeutics: Beyond vaccine applications. *Trends Mol. Med.* 27, 923–924. doi:10.1016/j.molmed.2021.05.004

Bishop, C. M. (2006). *Pattern recognition and machine learning*. 1st ed. Berlin, Germany: Springer.

Bonvin, D., Georgakis, C., Pantelides, C. C., Barolo, M., Grover, M. A., Rodrigues, D., et al. (2016). Linking models and experiments. *Industrial Eng. Chem. Res.* 55, 6891–6903. doi:10.1021/acs.iecr.5b04801

Canzani, E., and Timmer, S. W. (2021). *Beyond building predictive models: TwinOps in biomanufacturing*. Germany: Springer.

Cardillo, A. G., Castellanos, M. M., Desailly, B., Dessoy, S., Mariti, M., Portela, R. M. C., et al. (2021). Towards *in silico* process modeling for vaccines. *Trends Biotechnol.* 39, 1120–1130. doi:10.1016/j.tibtech.2021.02.004

Carvalho, M., Riesberg, J., and Budman, H. (2022). Hybrid model to predict the effect of complex media changes in mammalian cell cultures. *Biochem. Eng. J.* 186, 108560. doi:10.1016/j.bej.2022.108560

Chen, L., Bernard, O., Bastin, G., and Angelov, P. (2000). Hybrid modelling of biotechnological processes using neural networks. *Control Eng. Pract.* 8, 821–827. doi:10.1016/s0967-0661(00)00036-8

Chen, Y., Yang, O., Sampat, C., Bhalode, P., Ramachandran, R., and Ierapetritou, M. (2020). Digital twins in pharmaceutical and biopharmaceutical manufacturing: A literature Review. *Processes* 8, 1088. doi:10.3390/pr8091088

Cruz-Bournazou, M. N., Narayanan, H., Fagnani, A., and Butté, A. (2022). Hybrid Gaussian Process Models for continuous time series in bolus fed-batch cultures. *IFAC-PapersOnLine* 55, 204–209. doi:10.1016/j.ifacol.2022.07.445

Doyle, K., Tsopanoglou, A., Fejér, A., Glennon, B., and del Val, I. J. (2023). Automated assembly of hybrid dynamic models for CHO cell culture processes. *Biochem. Eng. J.* 191, 108763. doi:10.1016/j.bej.2022.108763

Farid, S. S., Baron, M., Stamatis, C., Nie, W., and Coffman, J. (2020). Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D. *mAbs* 12, 1754999. doi:10.1080/19420862.2020.1754999

Ferreira, a. R., Dias, J. M. L., Von Stosch, M., Clemente, J., Cunha, a. E., and Oliveira, R. (2014). Fast development of Pichia pastoris GS115 Mut+ cultures employing batch-to-batch control and hybrid semi-parametric modeling. *Bioprocess Biosyst. Eng.* 37, 629–639. doi:10.1007/s00449-013-1029-9

Feyo de Azevedo, S., Dahm, B., and Oliveira, F. R. (1997). Hybrid modelling of biochemical processes: A comparison with the conventional approach. *Comput. Chem. Eng.* 21, S751–S756. doi:10.1016/S0098-1354(97)87593-X

Galvanauskas, V., Simutis, R., and Lübbert, A. (2004). Hybrid process models for process optimisation, monitoring and control. *Bioprocess Biosyst. Eng.* 26, 393–400. doi:10.1007/s00449-004-0385-x

Gargalo, C. L., Caño, S., Heras, D., Jones, M. N., Udugama, I., Mansouri, S. S., et al. (2020). Towards the development of digital twins for the bio-manufacturing industry. *Adv. Biochem. Eng. Biotechnol.* 176, 1–34. doi:10.1007/10_2020_142

Gilbert, C., and Ellis, T. (2019). Biological engineered living materials: Growing functional materials with genetically programmable properties. *ACS Synth. Biol.* 8, 1–15. doi:10.1021/acssynbio.8b00423

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The elements of statistical learning the elements of statistical learning*. Berlin, Germany: Springer.

Hutter, C., von Stosch, M., Cruz Bournazou, M. N., and Butté, A. (2021). Knowledge transfer across cell lines using hybrid Gaussian process models with entity embedding vectors. *Biotechnol. Bioeng.* 118, 4389–4401. doi:10.1002/bit.27907

Hutter, S., Villiger, T. K., Brühlmann, D., Stettler, M., Broly, H., Soos, M., et al. (2017). Glycosylation flux analysis reveals dynamic changes of intracellular glycosylation flux distribution in Chinese hamster ovary fed-batch cultures. *Metab. Eng.* 43, 9–20. doi:10.1016/j.ymben.2017.07.005

Klyachko, N. L., Arzt, C. J., Li, S. M., Gololobova, O. A., and Batrakova, E. V. (2020). Extracellular vesicle-based therapeutics: Preclinical and clinical investigations. *Pharmaceutics* 12, 1171. doi:10.3390/pharmaceutics12121171

Kotidis, P., Pappas, I., Avraamidou, S., Pistikopoulos, E. N., Kontoravdi, C., and Papathanasiou, M. M. (2021). DigiGlyc: A hybrid tool for reactive scheduling in cell culture systems. *Comput. Chem. Eng.* 154, 107460. doi:10.1016/j.compchemeng.2021.107460

Mullard, A. (2023). 2022 FDA approvals. *Nat. Rev. Drug Discov.* 22, 83–88. doi:10.1038/d41573-023-00001-3

Murphy, D. E., de Jong, O. G., Brouwer, M., Wood, M. J., Lavieu, G., Schiffelers, R. M., et al. (2019). Extracellular vesicle-based therapeutics: Natural versus engineered targeting and trafficking. *Exp. Mol. Med.* 51, 32. doi:10.1038/s12276-019-0223-5

Narayanan, H., Behle, L., Luna, M. F., Sokolov, M., Guillén-Gosálbez, G., Morbidelli, M., et al. (2020a). Hybrid-EKF: Hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture. *Biotechnol. Bioeng.* 117, 2703–2714. doi:10.1002/bit.27437

Narayanan, H., Cruz Bournazou, M. N., Guillén Gosálbez, G., and Butté, A. (2022a). Functional-Hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. *Chem. Eng. J.* 430, 133032. doi:10.1016/j.cej.2021.133032

Narayanan, H., Dingfelder, F., Butté, A., Lorenzen, N., Sokolov, M., and Arosio, P. (2021a). Machine learning for biologics: Opportunities for protein engineering, developability, and formulation. *Trends Pharmacol. Sci.* 42, 151–165. doi:10.1016/j.tips.2020.12.004

Narayanan, H., and Love, J. C. (2022). Process modeling in the CMC of vaccines: Are we doing it right? *Vaccine Insights* 1, 299–314. doi:10.18609/vac.2022.042

Narayanan, H., Luna, M. F., von Stosch, M., Cruz Bournazou, M. N., Polotti, G., Morbidelli, M., et al. (2020b). Bioprocessing in the digital Age: The Role of process models. *Biotechnol. J.* 15, 19001722–e1900210. doi:10.1002/biot.201900172

Narayanan, H., Luna, M., Sokolov, M., Arosio, P., Butté, A., and Morbidelli, M. (2021b). Hybrid models based on machine learning and an increasing degree of process knowledge: Application to capture chromatographic step. *Industrial Eng. Chem. Res.* 60, 10466–10478. doi:10.1021/acs.iecr.1c01317

Narayanan, H., Luna, M., Sokolov, M., Butté, A., and Morbidelli, M. (2022b). Hybrid models based on machine learning and an increasing degree of process knowledge: Application to cell culture processes. *Industrial Eng. Chem. Res.* 61, 8658–8672. doi:10.1021/acs.iecr.1c04507

Narayanan, H., Seidler, T., Luna, M. F., Sokolov, M., Morbidelli, M., and Butté, A. (2021c). Hybrid Models for the simulation and prediction of chromatographic processes for protein capture. *J. Chromatogr. A* 1650, 462248. doi:10.1016/j.chroma.2021.462248

Narayanan, H., Sokolov, M., Morbidelli, M., and Butté, A. (2019). A new generation of predictive models: The added value of hybrid models for manufacturing processes of therapeutic proteins. *Biotechnol. Bioeng.* 116, 2540–2549. doi:10.1002/bit.27097

Narayanan, H., Sponchioni, M., and Morbidelli, M. (2021d). Integration and digitalization in the manufacturing of therapeutic proteins. *Chem. Eng. Sci.* 248, 117159. doi:10.1016/j.ces.2021.117159

Nold, V., Junghans, L., Bayer, B., Bisgen, L., Duerkop, M., Drerup, R., et al. (2023). Boost dynamic protocols for producing mammalian biopharmaceuticals with intensified DoE—A practical guide to analyses with OLS and hybrid modeling. *Front. Chem. Eng.* 4, 1044245. doi:10.3389/fceng.2022.1044245

Okamura, K., Badr, S., Murakami, S., and Sugiyama, H. (2022). Hybrid modeling of CHO cell cultivation in monoclonal antibody production with an impurity generation Module. *Ind. Eng. Chem. Res.* 61, 14898–14909. doi:10.1021/acs.iecr.2c00736

Pinto, J., Mestre, M., Costa, R. S., Striedner, G., and Oliveira, R. (2022). A general deep hybrid model for bioreactor systems: Combining first principles equations with deep neural networks. *Syst. Biol.* 2022, 495118. doi:10.1101/2022.06.07.495118

Psichogios, D. C., and Ungar, L. H. (1992). A hybrid neural network-first principles approach to process modeling. *AIChE J.* 38, 1499–1511. doi:10.1002/aic.690381003

Qin, S., Tang, X., Chen, Y., Chen, K., Fan, N., Xiao, W., et al. (2022). mRNA-based therapeutics: powerful and versatile tools to combat diseases. *Sig Transduct. Target Ther.* 7, 166. doi:10.1038/s41392-022-01007-w

Rodrigo-Navarro, A., Sankaran, S., Dalby, M. J., del Campo, A., and Salmeron-Sanchez, M. (2021). Engineered living biomaterials. *Nat. Rev. Mater* 6, 1175–1190. doi:10.1038/s41578-021-00350-8

Sahin, U., Karikó, K., and Türeci, Ö. (2014). mRNA-based therapeutics — developing a new class of drugs. *Nat. Rev. Drug Discov.* 13, 759–780. doi:10.1038/nrd4278

Schubert, J., Simutis, R., Dors, M., Havlik, I., and Lübbert, A. (1994a). Bioprocess optimization and control: Application of hybrid modelling. *J. Biotechnol.* 35, 51–68. doi:10.1016/0168-1656(94)90189-9

Schubert, J., Simutis, R., Dors, M., Havlík, I., and Lübbert, A. (1994b). Hybrid modelling of yeast production processes – combination of a priori knowledge on different levels of sophistication. *Chem. Eng. Technol.* 17, 10–20. doi:10.1002/ceat.270170103

Sokolov, M. (2020). Decision making and risk management in biopharmaceutical engineering - Opportunities in the Age of covid-19 and digitalization. *Industrial Eng. Chem. Res.* 59, 17587–17592. doi:10.1021/acs.iecr.0c02994

Sokolov, M., von Stosch, M., Narayanan, H., Feidl, F., and Butté, A. (2021). Hybrid modeling — A key enabler towards realizing digital twins in biopharma? *Curr. Opin. Chem. Eng.* 34, 100715–100717. doi:10.1016/j.coche.2021.100715

Steinwandter, V., Borchert, D., and Herwig, C. (2019). Data science tools and applications on the way to Pharma 4.0. *Drug Discov. Today* 24, 1795–1805. doi:10.1016/j.drudis.2019.06.005

Teixeira, A. P., Alves, C., Alves, P. M., Carrondo, M. J., and Oliveira, R. (2007). Hybrid elementary flux analysis/nonparametric modeling: Application for bioprocess control. *BMC Bioinforma.* 8, 30. doi:10.1186/1471-2105-8-30

Teixeira, A. P., Clemente, J. J., Cunha, A. E., Carrondo, M. J. T., and Oliveira, R. (2006). Bioprocess iterative batch-to-batch optimization based on hybrid parametric/nonparametric models. *Biotechnol. Prog.* 22, 247–258. doi:10.1021/bp0502328

Thompson, M. L., and Kramer, M. A. (1994). Modeling chemical processes using prior knowledge and neural networks. *AIChE J.* 40, 1328–1340. doi:10.1002/aic.690400806

Tsopanoglou, A., and Jiménez del Val, I. (2021). Moving towards an era of hybrid modelling: Advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses. *Curr. Opin. Chem. Eng.* 32, 100691. doi:10.1016/j.coche.2021.100691

Van Can, H. J. L., Te Braake, H. A. B., Dubbclman, S., Hellinga, C., Luyben, K. C. A. M., and Heijnen, J. J. (1998). Understanding and applying the extrapolation properties of serial gray-box models. *AIChE J.* 44, 1071–1089. doi:10.1002/aic.690440507

Vega-Ramon, F., Zhu, X., Savage, T. R., Petsagkourakis, P., Jing, K., and Zhang, D. (2021). Kinetic and hybrid modeling for yeast astaxanthin production under uncertainty. *Biotech Bioeng.* 118, 4854–4866. doi:10.1002/bit.27950

von Stosch, M., Davy, S., Francois, K., Galvanauskas, V., Hamelink, J. M., Luebbert, A., et al. (2014). Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry. *Biotechnol. J.* 9, 719–726. doi:10.1002/biot.201300385

von Stosch, M., Hamelink, J.-M., and Oliveira, R. (2016). Hybrid modeling as a QbD/PAT tool in process development: An industrial *E. coli* case study. *Bioprocess Biosyst. Eng.* 39, 773–784. doi:10.1007/s00449-016-1557-1

Von Stosch, M., Oliveira, R., Peres, J., and Feyo De Azevedo, S. (2012a). A general hybrid semi-parametric process control framework. *J. Process Control* 22, 1171–1181. doi:10.1016/j.jprocont.2012.05.004

Von Stosch, M., Oliveira, R., Peres, J., and Feyo De Azevedo, S. (2011). A novel identification method for hybrid (N)PLS dynamical systems with application to bioprocesses. *Expert Syst. Appl.* 38, 10862–10874. doi:10.1016/j.eswa.2011.02.117

Von Stosch, M., Oliveria, R., Peres, J., and De Azevedo, S. F. (2012b). Hybrid modeling framework for process analytical technology: Application to Bordetella pertussis cultures. *Biotechnol. Prog.* 28, 284–291. doi:10.1002/btpr.706

von Stosch, M., Portela, R. M., and Varsakelis, C. (2021). A roadmap to AI-driven *in silico* process development: Bioprocessing 4.0 in practice. *Curr. Opin. Chem. Eng.* 33, 100692. doi:10.1016/j.coche.2021.100692

Walsh, G. (2018). Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.* 36, 1136–1145. doi:10.1038/nbt.4305

Walsh, G., and Walsh, E. (2022). Biopharmaceutical benchmarks 2022. *Nat. Biotechnol.* 40, 1722–1760. doi:10.1038/s41587-022-01582-x

Zobel-Roos, S., Schmidt, A., Uhlenbrock, L., Ditz, R., Köster, D., and Strube, J. (2020). Digital twins in biomanufacturing. *Adv. Biochem. Eng. Biotechnol.* 176, 181–262. doi:10.1007/10_2020_146