A review and perspective on hybrid modeling methodologies

Schweidtmann, Artur M.; Zhang, Dongda; von Stosch, Moritz

DOI
10.1016/j.dche.2023.100136

Publication date
2024

Document Version
Final published version

Published in
Digital Chemical Engineering

Citation (APA)
Schweidtmann, A. M., Zhang, D., & von Stosch, M. (2024). A review and perspective on hybrid modeling methodologies. *Digital Chemical Engineering*, *10*, Article 100136. https://doi.org/10.1016/j.dche.2023.100136

Digital Chemical Engineering 10 (2024) 100136

Contents lists available at ScienceDirect

Digital Chemical Engineering

journal homepage: www.elsevier.com/locate/dche

Review article

A review and perspective on hybrid modeling methodologies

Artur M. Schweidtmann, Dongda Zhang, Moritz von Stosch *

Department of Chemical Engineering, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands
Department of Chemical Engineering, the University of Manchester, Engineering Building, Manchester M13 9PL, United Kingdom
DataHow AG, Hagenholzstrasse 111, Zurich, Switzerland

ARTICLE INFO

Keywords:
Hybrid modeling
Hybrid semi-parametric modeling
Grey-box
Neural networks
Parameter identification

ABSTRACT

The term hybrid modeling refers to the combination of parametric models (typically derived from knowledge about the system) and nonparametric models (typically deduced from data). Despite more than 20 years of research, over 150 scientific publications (Agharafeie et al., 2023), and some recent industrial applications on this topic, the capabilities of hybrid models often seem underrated, misunderstood, and disregarded by other disciplines as "simply combining some models" or maybe it has gone unnoticed at all. In fact, hybrid modeling could become an enabling technology in various areas of research and industry, such as systems and synthetic biology, personalized medicine, material design, or the process industries. Thus, a systematic investigation of the hybrid model properties is warranted to scoop the full potential of machine learning, reduce experimental effort, and increase the domain in which models can predict reliably.

1. Introduction

Machine-learning has obtained a lot of attention in recent years performing tasks unthinkable before (Wang et al., 2020). Much hope also relies on machine learning in natural and life science-related fields, dreaming of the description of highly complex systems solely by using data, regressing some inputs (e.g., features, factors, predictors, regressors) to some outputs (e.g., response, target) (Montáns et al., 2019; Jumper et al., 2021). Similarly, seeking to unravel the mechanisms of the system by mechanistic modeling has in the past given rise to fundamental modeling research (Sun et al., 2019; Gernaey et al., 2010; Horstemeyer, 2010). At first sight, machine learning approaches seem to compete with the more traditional fundamental modeling. However, fundamental modeling can be combined with machine-learning approaches as highlighted in literature (Antoniewicz, 2015; Baker et al., 2018; Bikmukhametov and Jäschke, 2020; Hamilton et al., 2017; Zhang et al., 2019, 2020).

Indeed, the idea of combining mechanistic modeling with data-driven models has been around from the 1990th (Psichogios and Ungar, 1992; Su et al., 1993; Kramer et al., 1992; Johansen and Foss, 1992; Thompson and Kramer, 1994). Since then a significant amount of research has been published using the terms "*hybrid modeling*" in the more process engineering-related research fields and "*grey-box modeling*" in the control and automation field. Though grey-box modeling is understood to include a wider range of models than hybrid modeling, e.g., a system of equations that is derived from first principles and complemented by empirically derived equations or structuring the machine-learning model based on process knowledge (Alhajeri et al., 2022; Wu et al., 2020) qualify as grey-box but not as hybrid models. Hybrid modeling is understood as the combination of models that are different in their traits, i.e., one part of the model structure is derived from knowledge (hence each parameter has a physical meaning and is normally identifiable, this type of model is named 'parametric' and it is typically represented by white boxes) whereas the other part of the structure is derived from data (hence parameters do not have any physical meaning and are normally not identifiable, and this type of model is named 'nonparametric' and it is typically represented by black boxes). As such, and in order to reduce the ambiguity in that the term hybrid modeling could be understood, the term hybrid semi-parametric modeling has been suggested (Thompson and Kramer, 1994; von Stosch et al., 2014b). In what follows, we use the term hybrid modeling as a short version of hybrid semi-parametric modeling.

The current hybrid modeling research and applications have evolved from the area of artificial neural networks, starting with Psichogios and Ungar (Psichogios and Ungar, 1992). They showed that the integration of fundamental knowledge into neural networks can (1) improve the model's extrapolation performance (the model faithfully predicts the system behavior beyond prior tested conditions), (2) reduce its data requirements, and (3) increase process understanding (e.g., model interpretability). The origin of these key properties can easily be comprehended considering the examples discussed in Box 1. These key properties were demonstrated and further extended in the literature (van Can et al., 1999; Van Can et al., 1998; Schuppert, 2000;

* Corresponding author at: DataHow AG, Hagenholzstrasse 111, Zurich, Switzerland.
E-mail address: m.vonstosch@datahow.ch (M. von Stosch).

https://doi.org/10.1016/j.dche.2023.100136
Received 25 October 2023; Received in revised form 12 December 2023; Accepted 13 December 2023
Available online 17 December 2023

von Stosch et al., 2014b; Fiedler and Schuppert, 2008; Kahrs and Marquardt, 2007). Coming from the mechanistic modeling perspective, machine-learning models can significantly extend their applicability for cases where parameters of the mechanistic model exhibit great confidence intervals, as they could be observed to vary in function of experimental conditions not captured by the mechanistic model or evolve over time, see e.g. Shah et al. (2022), Vega-Ramon et al. (2021).

Several recent reviews have discussed the application of hybrid models in the process industries, e.g., for separation processes (Schäfer et al., 2020; McBride et al., 2020), chemical, petroleum and energy systems (Sharma and Liu, 2022; Zendehboudi et al., 2018; Bradley et al., 2022), biochemical processes (Galvanauskas et al., 2018; Agharafeie et al., 2023; Mahanty, 2023), water systems (Schneider et al., 2022). Moreover, the importance of hybrid modeling is emphasized in multiple recent reviews and perspective articles on machine learning in chemical engineering (Lee et al., 2018; Venkatasubramanian, 2019; Schweidtmann et al., 2021b; Daoutidis et al., 2023). The challenges for the application of hybrid models to biopharmaceutical processes have been highlighted in Tsopanoglou and Jiménez del Val (2021). Also, the role of hybrid modeling for smart manufacturing (Yang et al., 2020), industry 4.0 (Sansana et al., 2021) and digital twins (Sokolov et al., 2021) has been explored. Beyond the process industries, systems and synthetic biology seems a very promising field for applications (Hamilton et al., 2017; Portela et al., 2018; von Stosch et al., 2014a; Pinto et al., 2023; Lee et al., 2020) as well as the field of personalized medicine (Clifton et al., 2017) and pharmacokinetics/pharmacodynamics modeling (Antontsev et al., 2021).

Following the above achievements, we argue that modeling complex, high-dimensional, and/or computationally expensive systems could greatly benefit from the combination and integration of more fundamental modeling and machine-learning techniques because of three main reasons:

1. The curse of dimensionality. The number of factors is often large in practical applications, and the data demand of nonparametric models scales super-linearly (exponentially) with the feature dimensionality. Thus, we either need to focus our studies on subsystems, hence limiting the number of factors and dimension, or generate humongous amount of informative data (not just big data). Given that the cost to generate data in may disciplines is quite high (e.g. a 250 mL mammalian cultivation run costs several hundreds of dollars), the application of pure nonparametric models for the entire system seem cost-prohibitive.
2. The trade-off between performance and effort. Even if we had enough data, we may still not be able to decode the hidden physical knowledge (and this could be very time-consuming). We always face the situation with some amount of data and some partial understanding of the process, however, we only have limited time to build a model that is as accurate as possible for process operation.
3. The quest for computational efficiency. Using hybrid surrogate models for computationally expensive process simulation is advantageous as they can speed up calculating and address tasks that are otherwise not feasible (e.g., real-time or fast decision-making). Though it may be argued that pure black-box based reduced models may achieve the same, the extrapolation capabilities of hybrid models render them invaluable for design, optimization, or control in many practical applications.

Recent reviews on hybrid modeling methodologies have provided new views on the structuring/typology of hybrid models (Sharma and Liu, 2022; Bradley et al., 2022; Rajulapati et al., 2022) and also revisited their framing within the evolving area of machine-learning models and in particular physics informed neural networks. In this review and perspective article, we suggest a more refined definition of hybrid models and their structures. We revisit the methodologies for hybrid model construction and training in light of potential applications, highlight the existing challenges and propose potential research directions to address them in connecting with other research fields.

We also discuss the perspectives of ongoing and future research in this field. The distinctive contribution of this review paper lies in its emphasis on novel solutions aimed at enhancing current practices in hybrid model construction, thereby charting a course for the broader utilization of hybrid models across a wider spectrum of applications (see Fig. 1). .

Box 1: Showcasing the properties of serial hybrid modeling structures

Imagine a system can be manipulated by three factors (inputs), x_1, x_2 and x_3, and the response of the system, which can be measured, is y. If the system is to be modeled by a data-driven model, e.g. a neural network, the model can be posed as:

$$y = f(x_1, x_2, x_3, \mathbf{w}) \tag{1}$$

which implies that x_1, x_2 and x_3, need to be modulated such that the function $f(\cdot)$ (described by the data-driven model) can be inferred and the parameters $\mathbf{w}$ identified (Fig. 1a). Thus, a three-dimensional space needs to be explored to be capable of drawing any conclusions regarding potential interactions between x_1, x_2 and x_3 and/or nonlinearity of the system (within the studied ranges). Considering the system exhibits only main effects and interactions, $2^3 = 8$ experiments are required to decipher the impact of the factors on the system response. Suppose that it is known that the impact of x_3 on y can be described with $x_3/(x_3+p)$ (with p some parameter), such that the system can be modeled by:

$$y = x_3/(x_3 + p) \cdot f(x_1, x_2, w) \tag{2}$$

The space that needs to be explored in this case comprises only two dimensions (Fig. 1b), namely that of x_1 and x_2. This implies that the number of experiments can be reduced, four experiments are required for the example. In addition, the model will predict reliably for any value of x_3, i.e., the model can extrapolate in x_3 beyond tested values. Regardless of the values in x_3 (except if $x_3 = 0$) the model $f(x_1, x_2, w)$ can be inferred given sufficient variation in x_1 and x_2 (which can exhibit a particular advantage in the case that x_1 or x_2 cannot be controlled). The model $f(x_1, x_2, w)$ will typically be simpler than $f(x_1, x_2, x_3, w)$ which also simplifies the modeling exercise. However, not all of these properties are unique to this specific nonparametric–parametric serial hybrid model sequence. Consider the model shown in Fig. 1c. This structure allows for a reduction in the number of experiments required for the characterization of the system because of the introduced structure (only the impact of g rather than that of x_1 and x_2 needs to be investigated). However, the extrapolation capabilities are different to the model shown in Fig. 1b. This model can be expected to predict well as long as g stays within the prior investigated ranges and hence x_1 and x_2 should only be varied such that g does not exceed this range. In the contrary, the extrapolation limits for x_3 shown in Fig. 1b (and discussed in the example before), are of physical nature in that the model prediction performance will deteriorate when the described mechanism is no longer governing the system behavior.

2. Fundamentals of hybrid modeling

In this section, we describe the hybrid modeling structures (Section 2.1, the main advantages of hybrid modeling (Section 2.2), and the training of hybrid models (Section 2.3).

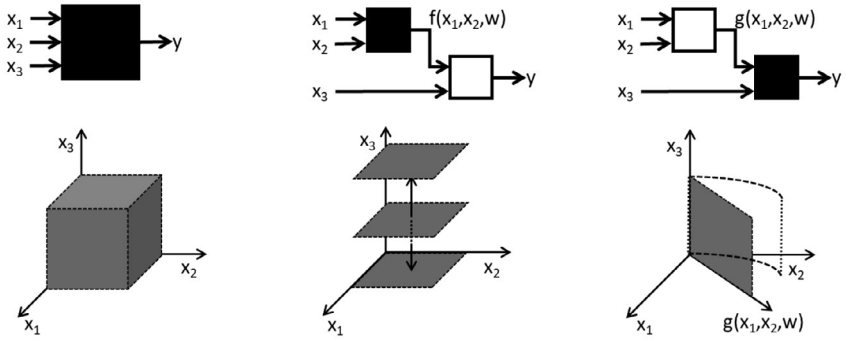

Fig. 1. Considering a system with 3 factors (x_1 to x_3) and response y which only exhibits main effects and interactions, it can be seen that the integration of knowledge has the potential to reduce the experiments. It is considered that $f(x_1, x_2, w)$ is a function that is given by a machine-learning model, whereas $g(x_1, x_2, w)$ is a function that is derived from mechanistic knowledge.

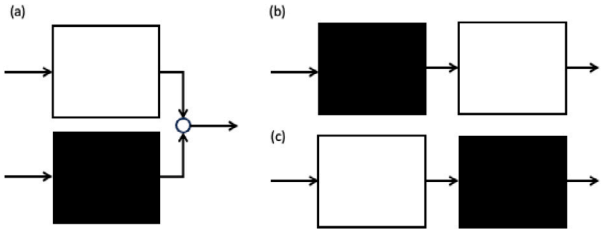

Fig. 2. Basic hybrid modeling structures based on Psichogios and Ungar (1992) where (a) shows a parallel model structure and (b,c) show serial model structures. Note that the black boxes represent machine learning models and the white boxes represent mechanistic models.

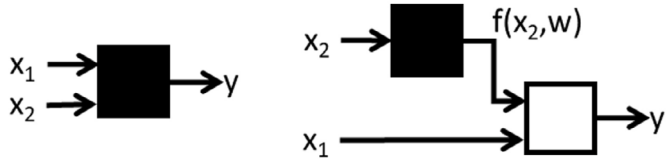

Fig. 3. An example, where the differences in inputs of the nonparametric model might results in scenarios where for certain extrapolative variations in the inputs the predictions of the nonparametric model (left) might be significantly less accurate than those of the hybrid semi-parametric model (right).

2.1. Hybrid modeling structures

There are three basic configurations of parametric and nonparametric model that determine whether a hybrid model is understood to have a parallel or serial structure. These basic configurations are shown in Fig. 2, where the black-box typically represents the nonparametric, data-driven component, and the white-box represents the parametric, knowledge derived component.

These basic configurations have a high relevance for the mathematical foundations of hybrid models. Also, the model's advantages and limitations can be more easily understood as a result of the differentiation. Hence, we provide a more restrictive definition of the basic structures in the following.

In practice, hybrid models are often more complex intertwined structures where a clear classification as parallel or serial is not possible (Bradley et al., 2022). For these more complex hybrid models that consist of several white and black boxes, the properties are not so straightforward to assess and the mathematical savvy reader is referred to Fiedler and Schuppert for a mathematically more rigorous analysis of such tree structured hybrid models (Fiedler and Schuppert, 2008), which constitute a particular class of serial hybrid models. Moreover, one can distinguish hybrid models using concepts from network science where model structures can be classified into acyclic graphs and cyclic graphs. One example is that of dynamic hybrid models where the predicted quantity is used as an input for the next timestep, which are discussed in more detail in Section 3.1. Hence, methods that have been developed for graphs could perhaps readily be applied to hybrid models, though we are not aware of any such application.

2.1.1. Parallel structures

A hybrid model is defined here to have a parallel structure if (1) the parametric model can independently of the nonparametric model describe the system's behavior; and (2) the nonparametric model "only" improves the prediction quality of the parametric model, aiming to obtain a good agreement with the real system. While this is a "mechanism correction" approach as referred to in Bradley et al. (2022), Zhang et al. (2020), not all mechanism correction approaches are automatically hybrid due to the point on independence.

The extrapolation performance of this parallel structure can generally be assumed to be limited as the nonparametric model is typically not expected to describe the behavior of the system outside the training ranges. However, scenarios exist in which the parametric model is extrapolating, whereas the nonparametric model operates within the prior training ranges using complimentary measurements (e.g. measured spectra, Fig. 3) or describing one part of the system that does not extrapolate effectively reducing the input dimensionality of the black box model part (Quaghebeur et al., 2022).

In the simplest parallel case the nonparametric model's correction is added to the parametric model's prediction, but more sophisticated structures are available where the weighting of each prediction is made in accordance with the expected prediction accuracy of the model (Dors et al., 1995; Peres et al., 2001, 2008). In this regard, the Kalman, Sigma Point, Particle or other alike filters could also be understood as a hybrid model where the parametric component describes the process dynamics and the nonparametric component consists of a soft-sensor model, which are then combined by a weighting component (Simutis and Lübbert, 2017; Cabaneros Lopez et al., 2021). Hence, a lot of learnings can be drawn from developments in this field and the interested reader is referred to Narayanan et al. (2020), Cabaneros Lopez et al. (2021). In general though, the added value of the nonparametric model is an increase in performance within the "domain" of the training data (i.e., the data used for the development (training) of the nonparametric model) (Kahrs and Marquardt, 2007; Schweidtmann et al., 2021c). One way to improve the extrapolation potential and reduce data requirements of parallel hybrid models is to reduce the input dimensionality of the black box by selecting only a few inputs (e.g., based on mechanistic knowledge) (Fiedler and Schuppert, 2008).

2.1.2. Serial structures

The most interesting hybrid modeling properties, namely a reduction in the data requirements, improvement in extrapolation performance, and improved systems understanding, can be obtained with

serial structures, the basic ones are discussed in more detail in Box 1 and shown in Fig. 2b and c. Feature engineering can be regarded as a serial hybrid model, as it resembles the serial structure shown in Fig. 2c. The good performance that can be improved by feature engineering demonstrates the capabilities of this structure. For instance, Richelle et al. (2020) used material balance equations and simple biochemical considerations to transfer the focus of the variation analysis from the process to the underlying biological system for two upstream biopharmaceutical processes. Tantamount to feature engineering, the output can be engineering, as represented by Fig. 2c, giving rise to similar performance. For instance, de Azevedo et al. (Rodrigues de Azevedo et al., 2017; Azevedo et al., 2019) used an established empirical equation to transform variations in controlled drug release profiles into variations into two distinct parameters that characterize the release profile, i.e., the total amount of released drug and the kinetic of drug release. The link function in generalized linear regression can perhaps be understood as conceptually similar. However, rather than using an equation that stems from fundamental knowledge, the link function many times is chosen to resemble the observed behavior, and the parameter of the function are subsequently used as output of linear regression (Note that there exist several requirements for the link function, such as monotonicity, differentiability, etc.) (Hilbe, 2011; Lindsey, 1999).

2.2. Why hybrid models perform better?

The benefits of hybrid models arise from the incorporation of fundamental knowledge (Rogers et al., 2023b). Generally, one can differentiate two types of parametric knowledge, according to von Stosch et al. (von Stosch et al., 2014b). One type, referred to as "structuring knowledge", considers the structure of the interactions of the different variables (which typically is at least to some degree time-invariant) or systems components. A structuring knowledge example, is the stoichiometric reaction matrix that describes the interconnection of reactions and compounds. The other type of knowledge, referred to as "forming knowledge" by von Stosch et al. (von Stosch et al., 2014b), describes the functional form in that two or more variables are related. Kinetic rate functions are an example for forming knowledge.

As should become apparent from the considerations described in Box 1, the integration of structural or forming knowledge, can potentially (1) reduce the number of experiments/data points required to characterize the model application domain; and (2) increase the extrapolation properties along certain domains. However, an error in the structural or forming knowledge (i.e., the behavior of the system is not appropriately described by the mathematical equations) will bias the model, constraining its performance (Rogers et al., 2023b). Thus, it is of paramount importance to assess to which extend can the structuring knowledge be trusted and the assumptions that the fundamental knowledge is based upon. The evaluation of the fundamental knowledge base prior to model creation and experiment execution can help to assess the advantages of this approach a priori (Rogers et al., 2023b). The creation of hybrid modeling structures by trial and error, i.e., without consideration why the model should perform better, will increase performance or decrease the number of experiments at best by chance.

It is noteworthy that, while other advanced machine learning techniques exist that also integrate domain knowledge and process data, hybrid models are recognized for presenting distinct advantages beyond those offered by these methods. For instance, physics-informed neural network (PINN) has been extensively studied recently for different applications (Raissi et al., 2019; Rogers et al., 2023a; Sansana et al., 2021; Wu et al., 2020; Zheng and Wu, 2023). However, PINN and hybrid models are proposed for diverse applications. PINN primarily serves as a surrogate model, substituting a computationally-intensive, first-principles-derived physical model—exemplified by the Navier–Stokes equations. While PINN has the capability to integrate pertinent information from both process data and a physical model, it is imperative to note that the existence of a rigorous mechanistic model is a prerequisite for PINN construction. This stands in stark contrast to the hybrid model, which operates independently of the need for a highly accurate physical model. The hybrid model is principally employed to simulate complex systems where only partial understanding is available. It employs a (simple) mechanistic model to quantify partial process understanding and utilizes process data (constructing a data-driven model) to address gaps in knowledge. Simultaneously, it maintains a flexible model structure for computational efficiency. Notably, the hybrid model distinguishes itself from PINN by not functioning as a surrogate model but as an enhancement of the underlying mechanistic model. As such, hybrid models demonstrate broader applicability across various domains. However, this does not imply that hybrid models preclude the utilization of other machine learning methodologies, such as PINN. Indeed, the collaborative integration of PINN can contribute to the advancement of hybrid model development, as elucidated in Section 3.4.

In addition, hybrid models offer distinct advantages over other machine learning based physical model construction strategies such as symbolic regression (Forster et al., 2023; Narayanan et al., 2022) and sparse regression (Massonis et al., 2023; Brunton et al., 2016). Firstly, unlike sparse regression techniques e.g. sparse identification of nonlinear dynamics (SINDy), hybrid models eliminate the necessity for a predefined library of potential mechanistic expressions. Secondly, the inclusion of a mechanistic model structure within hybrid models enhances interpretability, a feature not consistently guaranteed by symbolic regression. Symbolic regression often generates mathematical expressions solely based on statistical considerations, lacking inherent physical meaning. A parallel concern arises in sparse regression, where the process of selecting expressions from the predefined library may lack the incorporation of valuable physical insights. Moreover, when confronted with the simulation of highly nonlinear systems, such as bioprocesses or chemical reaction networks, both sparse regression and symbolic regression may prove inefficient in capturing the underlying process complexity. These methods often rely on simplistic expressions to describe process behaviors. Contrastingly, hybrid models adeptly address this limitation by combining a data-driven model with a mechanistic model, making them well-suited for the simulation of intricate physical systems commonly seen in chemical engineering applications.

2.3. Hybrid model training

The fitting of the model to data by adapting the model's parameter values is referred to as training or learning in the machine-learning field, in statistics called parameter estimation, whereas for fundamental models one typically speaks of parameter identification. Due to the fixed model structure, identifiability of the parameters given the data and model structure is to be considered when fitting fundamental, parametric models (Iliadis, 2019; Karlsson et al., 2012; Villaverde, 2019; Massonis et al., 2023). Parameters in a machine-learning model are known to be nonidentifiable given the symmetric structure of many types of machine learning models and the large number of parameters giving the machine learning models the flexibility to approximate nonlinear functions (Hornik, 1993). However, this flexibility in structure of machine-learning models gives rise to the problem of overfitting and several methods have been proposed to alleviate this issue, such as drop-out, regularization, early-stopping, pruning, etc. Hinton et al. (2006, 2012), Wang and Raj (2017). One could also look at this difference in training fundamental and machine-learning models in terms of bias and variance. Whereas with fundamental models the focus is on fitting the model to reduce the error from bias (introduced by the rigid structure), machine-learning approaches fit the model to reduce the error from variance (as the model structure can be adapted counteracting the bias). Notably, black box machine learning approaches like

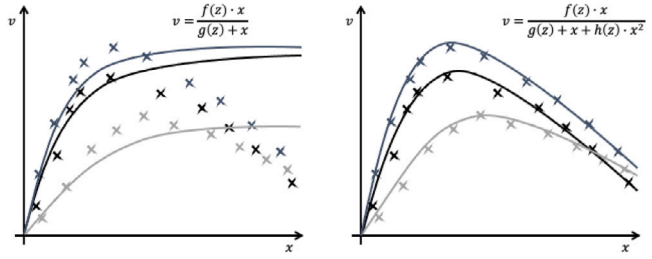

Fig. 4. Values of v over x for experiments with 3 different levels of $z = [z_1, z_2, \ldots, z_n]$ as well as the fitted function, indicated in the top right corner of each plot.

neural networks are proven to have universal approximation ability. Therefore, it is possible that hybrid models may not perform as well as machine learning models if the mechanistic knowledge is inaccurate, i.e. the mechanistic knowledge provides an inductive bias (Psichogios and Ungar, 1992). This potential impact of inductive bias is further highlighted in Sections 3.4 and 3.5.

When fitting hybrid models, both problems, i.e., identifiability and overfitting, need to be addressed. The inductive bias, stemming from the knowledge back-bone, could hinder the successful development of a model (if the introduced structure does not match the underlying system), while at the same time it could also facilitate model development (if the introduced structure matches the behavior of the system and therefore constrains the parameter space). This is visualized for an example in Fig. 4.

The prevailing approaches for parameter identification in hybrid models give very little attention to structural identifiability and inductive bias. A general recommendation for hybrid model development is to only integrate those parts of the knowledge for which one is certain that they are "correct", creating a baseline model. Subsequently, the knowledge part can be increased while evaluating the performance of the extended hybrid against the baseline hybrid model to check whether the integration of knowledge has led to a deterioration or increase in performance (Rogers et al., 2023b). While this manual procedure can help with the inductive bias issue, it can only to some extent and implicitly address the issue of structural identifiability. Hence, research in this direction would be important to address these shortcomings better.

Hybrid models can have parameters in both, the parametric and nonparametric models. The practical approach to identifying them considers first fitting and then fixing the parameters of the parametric part (either through knowledge or, if the structure allows it, through data) and subsequently training the nonparametric (black box) model. This can potentially be followed by a re-adaptation of the parametric model parameters, which provides good model performance as shown by Yang et al. (2011), one of the few approaches proposed to identify the parameters in both parts. It is evident that this sequential approach might converge to scenarios where the model parameter and structure are not globally optimal. The incremental identification approach, proposed by Kahrs and Marquardt (2008), can also be used to identify parameters in both parts. This approach somewhat works its way backwards from the outputs through the knowledge to the parameters, decomposing the problem and therefore making it easier to solve. However, for this, all the outputs must have been measured. Also, along with many serial structures, it might not be possible to obtain a closed form (the equations of the parametric model cannot be inverted) or the estimation is numerically ill-conditioned, e.g., when the system is dynamic. When this direct parameter identification is not an option, the indirect approaches can be used, also referred to as sensitivities approach (Psichogios and Ungar, 1992; Oliveira, 2004). However, this approach has its limitations, in particular, if the system is dynamic, since the system of ordinary differential equations or even partial differential equations need to be numerically integrated which is a sequential approach taking considerable time as it cannot be parallelized. In addition, parameter identification typically is a nonconvex optimization wherefore multiple restarts from random parameters are required to obtain a good approximator. However, recent work suggests that the number of restarts can be limited when using a stochastic gradient descent algorithm (Pinto et al., 2022), which would decrease the computational load at least to some degree, though the multiple restarts could of course be carried out in parallel. Nevertheless, it is not clear how to use the sensitivities approach with a number of other machine-learning models, such as Gaussian process models.

3. Current challenges and research perspectives

In this section, we discuss a number of current challenges and future research perspectives of hybrid models.

3.1. Dynamic hybrid models

Depending on the mechanistic model expression, a hybrid model can be categorized as static or dynamic (Glassey and von Stosch, 2018). For instance, imagine a first-order chemical reaction with catalyst deactivation. One way to quantify the mechanistic knowledge of this reaction is to express the reaction concentration as an algebraic equation: $c_A = c_{A0} \cdot e^{-k \cdot t}$. As catalyst deactivation is a complex process dependent on the operating conditions (e.g., temperature T), reaction mixture composition (e.g., reactant concentration c_A), and time duration (t), a machine learning model can be used to estimate the time-varying reaction rate constant $k = f(T, c_A, t, w)$ (Bui et al., 2022) with w the weights/parameters of the machine-learning model. Thus, the hybrid model can be expressed as $c_A = c_{A0} \cdot e^{-f(T, c_A, t, w) \cdot t}$. Alternatively, one can also formulate the mechanistic knowledge using a differential equation: $\frac{dc_A}{dt} = k \cdot c_A$ with the same machine learning model employed for reaction rate constant estimation. In this way, the hybrid model can be expressed as: $\frac{dc_A}{dt} = f(T, c_A, t, w) \cdot c_A$. When comparing the two approaches, it is easy to see that the first approach does not involve the time derivative of the state variable, meaning that the hybrid model is directly simulating the state over time. Instead, the second approach aims to calculate the derivative of the state (i.e., change of the state at each time), hence numerical integration is needed in order to estimate reactant concentration. Therefore, the first approach is considered a static hybrid model (i.e., no time derivative involved), and the second approach is called a dynamic model (i.e., direct involvement of derivative).

Intuitively, one would prefer the use of a static model whenever possible as the resulting parameter estimation problem could be less mathematically challenging compared to a dynamic model (c.f. Section 2.3). This is particularly true if the states are not measured frequently or if the measurements are of high noise, as errors can be amplified when propagating from the state space to its derivative space (Bayer et al., 2020a). In other words, one may build a dynamic hybrid model that gives an adequate fitting performance for state measurements but completely miscalculates the derivative of the state. However, regardless of the type of the hybrid model, due to the embedding of a machine learning model compartment, hybrid model parameter estimation is always a challenge. This is because hybrid models can be highly nonlinear and nonconvex, and they contain a large number of parameters that are non-identifiable. As a result, when simultaneously estimating all the parameters in a hybrid model, the resulting parameter estimation problem is often ill-defined (Kahrs and Marquardt, 2008), thus requiring substantial manual tuning during model construction. While the incremental identification strategy by Kahrs and Marquardt (Kahrs and Marquardt, 2008) has been proposed to resolve this issue it is to date only applied to static systems.

In 2021, a more efficient two-step parameter estimation strategy has been developed by Vega-Ramon et al. (2021). Compared to the simultaneous parameter estimation approach, this strategy decouples

the original problem into two steps. In the first step, it converts a time-varying parameter p into a vector $\boldsymbol{p} = [p_1, p_2, \dots, p_{tf}]^T$, i.e., assigning an independent parameter p_i at each time step i. In this way, it explicitly decouples the machine learning model from the mechanistic model. Then, parameter estimation can be conducted via a well-established dynamic model parameter estimation approach (del Rio-Chanona et al., 2015; Cruz-Bournazou et al., 2022). To prevent overfitting, a regularization term can be added in the objective function to penalize changes in parameter values between adjacent time steps. Once completed, in the second step, new data points $(\boldsymbol{x_i}, p_i)$ can be directly generated to train a machine learning to approximate the relation between process variables $\boldsymbol{x}$ and parameter p at each time step. Moreover, data augmentation can be applied to generate a large amount of synthetic data to facilitate machine learning model construction. The advantages of this strategy are two-fold. Firstly, dynamic model parameter estimation results in less complex mathematical model which is easier to optimize and more likely to obtain a high-quality solution, and the regularization term can effectively prevent overfitting of the time-varying parameters. Secondly, data augmentation can be effectively applied to stabilize the training of the machine learning model, thus increasing the accuracy and reliability of the hybrid model and reducing the time for real data generation. In addition, this strategy has been successfully applied to dynamic hybrid model construction and is verified to be of high efficiency (Vega-Ramon et al., 2021; Rogers et al., 2023b; Cruz-Bournazou et al., 2022). However, this strategy requires all time-varying outputs to be measured and it might underestimate the dynamic properties of the system. The rich behavior of dynamic systems (such as oscillations, etc.) can be attributed to the dependence of the system on its state (which could be represented by a cyclic graph). This dependence is explicitly accounted for by the indirect learning approach, namely the sensitivity equations. Using the example, the sensitivity equations read:

$$\frac{\mathrm{d}}{\mathrm{d}t} \cdot \frac{dc_A}{dw} = \frac{d(f(T, c_A, t) \cdot c_A)}{dc_A} \cdot \frac{dc_A}{dw} + \frac{d(f(T, c_A, t) \cdot c_A)}{dw} \tag{3}$$

where the term $\frac{d(f(T,c_A,t)\cdot c_A)}{dc_A}$ captures the propagation. This term is not explicitly taken into account by Vega-Ramon et al. (2021). The semi-direct learning approach proposed by Pinto et al. (2022) considers this dependence and a combination of the two approaches would be envisioned to be even more effective.

Although not being tested yet, one can expect that this decoupling strategy can be effectively adopted to construct static hybrid models.

It is worth noticing that although static hybrid models can potentially alleviate numerical issues arising from parameter estimation, in practice dynamic hybrid models are more widely used within the field of chemical engineering. This is because for complex systems, the majority of physical knowledge (i.e., mechanistic models) is expressed as ordinary differential equations (e.g., the Langmuir–Hinshelwood model for catalytic reaction kinetics and the Droop model for fermentation process) or partial differential equations (e.g., Navier–Stokes equations for fluid dynamics), and they usually do not have closed-form solutions in the state space. Moreover, these differential equations directly describe the interdependence of different state variables (i.e., human interpretable knowledge), which are more reliable when extrapolated for process predictive modeling. As a result, it is more practical to build a hybrid model to simulate the derivative space rather than the state space.

Though most dynamic hybrid modeling applications have focused on first order dynamic systems, more complex high order dynamics can relatively easily be accounted for using a series of time-lagged inputs in the machine-learning model (von Stosch et al., 2010; Sitapure and Sang-Il Kwon, 2023). Considering a bioprocess, where the underlying biological system might exhibit a "memory effect", one could aim at modeling the system using a second order dynamic system, i.e.

$$\frac{dc}{dt} = r \tag{4}$$

$$\frac{dr}{dt} = g(c, w) \tag{5}$$

where c is a vector of concentrations, r is a vector of reaction rates and $g(c, w)$ is a nonparametric model with parameters w. Alternatively, the system could be modeled by considering n time-lagged inputs ($c(t-\tau_n)$), i.e.

$$\frac{dc}{dt} = g(c, c(t-\tau), \dots, c(t-\tau_n), w) \tag{6}$$

with τ the delay (Mowbray et al., 2023). The same rational as for the integration of mechanistic knowledge likely can also be used to define whether the dynamics are explicitly formulated as part of the mechanistic backbone or integrated into the parametric part. This might provide an additional direction of research for the field of hybrid modeling, though this is a topic that has been widely studied in process control (Seborg et al., 2016).

3.2. Automatic structure discrimination

Automatic structure discrimination is an underdeveloped research area for hybrid modeling, whereas significant achievements have been made in discriminating data-driven models and mechanistic models. For data-driven models, a range of methods have been developed to reduce model complexity and minimize risks in overfitting. For example, for artificial neural networks, several approaches have been proposed to either make the neural network structure leaner (e.g., regularization, drop-out learning, pruning) (Hinton et al., 2006, 2012) or to discriminate between the performance of different structures (AIC, BIC, adjusted R^2). Moreover, different hyperparameter selection frameworks (e.g., Bayesian optimization) have also been developed to systematically identify the optimal neural network structure (Yu and Zhu, 2020). Meanwhile, other machine learning models such as Gaussian processes are intrinsically immune to structure discrimination (although a kernel function still needs to be pre-defined) given their unique characteristic (Rasmussen and Williams, 2006). Through the development of the two-step parameter estimation strategy (Vega-Ramon et al., 2021), these data-driven model structure discrimination methods can be effectively used to identify a suitable machine learning model structure for hybrid model construction.

For mechanistic models, similarly, extensive research has been conducted within this topic. Mechanistic model structure identification can be either addressed by using statistical criteria (e.g., AIC, BIC, Hannah Quinn Criterion, Bridge Criterion) based on different assumptions (Ward, 2008), or through the use of advanced optimization algorithms (e.g., mixed-integer programming, sparse regression) (Willis and von Stosch, 2016; Brunton et al., 2016), or a combination of the two.

This work is extended to a specific class of hybrid models by Willis and von Stosch (2017), who propose a method based on mixed integer linear programming (MILP) that allows to simultaneously identify the parameters of a polynomial/rational model and discriminate its structure. The work by Narayanan et al. (2022) can be seen as a further extension. These methods are conceptually similar to those of symbolic and sparse regression, which either screen a model library consisting of different mechanistic knowledge-derived expressions to select the most appropriate model (Daume et al., 2020; Herold and King, 2014; Kroll et al., 2017; Sahinidis, 2016; Willis and von Stosch, 2016; Wilson and Sahinidis, 2017; Žegklitz and Pošík, 2021; Chakraborty et al., 2021), or iteratively "distill" the underlying equations using genetic programming (Hinchliffe and Willis, 2003; McKay et al., 1997; Schmidt and Lipson, 2009; Searson, 2015; Willis et al., 1997).

However, as the inference of the determined model structure is critically dependent on the data quality (the amount of information captured in the data), one could argue that the derived models based on the above approach are indeed *nonparametric* (i.e., data-driven models rather than mechanistic models) as the model structure is inferred from data without consideration of the underlying mechanism. This is particularly true if the model library consists of randomly formulated terms to improve diversity. Claims that "laws of nature" (Schmidt and Lipson, 2009) or "mechanistic models" (Daume et al., 2020; Kroll et al.,

2017) are derived could be potentially problematic without rigorous justification, as the approaches rather seem a model-fitting exercise, similar to statistical models. Philosophically, one would expect the formulation of mechanistic models to follow the hypothesis-driven scientific principle giving rise to reproducible findings, whereas the derivation of the model structure from an experiment point of view could give rise to purely circumstantial observation fitting equations. As a result, once these statistic models are identified, it is important to investigate which hypotheses and conditions can be derived from these identified models (e.g. the Langmuir–Hinshelwood model is derived based on quasi-steady state of reaction intermediates, Navier–Stokes equations are derived based on the conservation of momentum).

Despite the aforementioned achievements, effective strategies for hybrid model structure discrimination which take into account both the mechanistic model compartment and the data-driven model compartment have been barely proposed. For instance, one fundamental assumption for the majority of the mechanistic model structure discrimination methods is that model parameters are time-independent, which is not the case for hybrid models (i.e., the presence of time-varying parameters). As a result, they are not capable (or at least suitable) for hybrid model structure discrimination. Similarly, although the data-driven model structure discrimination methods can be applied to build the data-driven model compartment for a hybrid model, they are not applicable to help identify the structure of the mechanistic model compartment. At this moment, simultaneous parameter estimation and model structure discrimination can be only applied to some specific types of hybrid models. For example, the MILP-based technique developed by Willis and von Stosch (2017) can be used to address this challenge if the hybrid model can be reformulated to be linear in the parameters. Here, recent works on optimization formulations of machine learning models could be used (Fischetti and Jo, 2018; Schweidtmann and Mitsos, 2019; Tsay et al., 2021; Grimstad and Andersson, 2019; Schweidtmann et al., 2021a). Zhang et al. (2020), Forster et al. (2023) proposed an MINLP-based technique if the hybrid model is a discrepancy model (i.e., a machine learning model is appended to a mechanistic model to rectify model-process mismatch). In 2022, the first hybrid model structure discrimination methodology that could be applicable to general hybrid models is proposed through the use of reinforcement learning (RL) (Mowbray et al., 2023). This strategy also requires a library of possible mechanistic model candidates (each of which must have strong physical meaning) and then employs RL to estimate which candidates should be selected as well as if their parameters are time-independent or time-varying. Through a number of in-silico tests, this strategy is proven of high potential. Moreover, this strategy is also applicable to modeling history-dependent systems (i.e., current dynamics dependent on historical conditions). However, this framework is still at its infant stage, thus substantial improvements are required for future real-world applications. Finally, another potential strategy that has never been explored before is physics-informed neural network. A more detailed investigation about this approach for hybrid model structure discrimination is discussed in Section 3.4.

3.3. Incremental learning

The methods presented hitherto are batch learning methods, i.e., the model is fitted for a given data set. However, on many practical applications new experiments/runs are executed and new data become available extending the original data set. Thus, strategies for incremental learning on new data are of high relevance. On approach is the re-training using either all data, a combination of previous and new data, or just the new data. While in principle, the batch learning methods (i.e., batch-incremental learning (Read et al., 2012) - sometimes also iterative learning, which essentially means training a model de novo on all data) could be used for this purpose, for an increasing quantity of data, re-learning will become computationally expensive. Alternatively, the model could be trained only on the new data, i.e., instance-incremental learning (Read et al., 2012), which however might give rise to inferior overall model performance. For instance, the training might be influenced by a particularity of the new data (e.g., sample bias, outlier behavior) or by a drift of the behavior of the underlying system, concept drift (Read et al., 2012). In general, this can be framed as the stability/plasticity dilemma (Chefrour, 2019), where stability is the capacity to not forget the already learned data and plasticity is the capacity to assimilate new data. Instance-incremental learning methods need to exhibit a good stability/plasticity compromise and according to Chefrour (2019) should exhibit four criteria:

1. “it should be able to learn additional information from new data (plasticity);
2. it should not require access to the original data, used to train the existing classifier;
3. it should preserve previously acquire knowledge it should not suffer from significant loss of originally learned knowledge (stability);
4. it should be able to accommodate new classes that may be introduced with new data; ”

While incremental learning is an active field of research in the machine-learning community (Chefrour, 2019; Read et al., 2012; Wang et al., 2020), research on this topic in the field of hybrid modeling is limited.

3.4. Hybrid models for adaptive and evolving systems

So far, we have discussed hybrid models with fixed structures. However, hybrid models can also adapt and evolve (Chefrour, 2019). In adaptive models, the structure is fixed and only the parameters are adapted. In evolving models, the structure and parameters are learned from the new data. This ties back to the parameter estimation and structure identification challenges addressed before. There is no clear borderline to distinguish whether a system should be categorized as adaptive or evolving, as they are interchangeable in many cases. For example, mammalian cells used for recombinant protein synthesis often go through two distinct phases including an initial cell growth stage and a later product synthesis stage. Metabolic activities within cells are changed significantly between the two stages. One can either regard this system as adaptive by designing a lumped macro-kinetic model (e.g., the Monod model) within which the model structure can simulate both phases but parameter values have to be changed, or consider the system as evolving by directly constructing different models for each phase. Similarly, there can be multiple reaction mechanisms for an organic reaction (e.g., substitution vs. elimination such as S_N1:E1, S_N2:E2). The dominating reaction mechanism depends on the operating conditions and can be changed throughout the reaction process. For a gas–solid catalytic reaction operated at a high temperature, due to continuous changes in catalyst configuration and reactivity, the underlying reaction mechanism is also evolving. These systems can also be modeled either as an adaptive system or an evolving system.

Both mechanistic models and machine learning models have been extensively studied to simulate the two types of systems. For machine learning models, given their data-driven nature, it is straightforward to update their parameters and even structures based on new data if needed. For mechanistic models, parameter re-estimation has been the primary approach if they are applied to adaptive systems. Although in principle the same concept can be extended to evolving systems through mixed-integer programming or sparse regression (i.e., automatically determining suitable model structures using new data), in practice, this is not widely used as only updating model parameters can already result in good performance for process prediction, optimization, and control. The use of NLP is also more effective than MINLP if the process model is highly nonlinear and fast decision-making is required.

As a result, reducing an evolving system to an adaptive system can often simplify the model construction challenge and meanwhile remain good accuracy.

Nevertheless, research on the systematic use of hybrid models for adaptive and evolving systems has been absent and critical questions are yet to be answered. For example, could it be that fixing some part of the hybrid model structure by introducing knowledge increases stability and robustness for the learning of the remaining structure and parameters, or does the introduced inductive bias, and therefore reduced plasticity, hinder the constructed hybrid model from being accurate and reliable? In addition, given the small size of data collected from an ongoing process, how to determine which part of the model structure and parameters should be updated meanwhile minimizing risks in over-parameterization and overfitting (Rogers et al., 2022)?

To answer these questions, recent achievements in process systems engineering and machine learning may offer extra insight. On the one hand, discrepancy hybrid models have been widely used in process control. A discrepancy hybrid model $\frac{dy}{dt} = f(\cdot) + g(\cdot)$ consists of a mechanistic model $f(\cdot)$ and a machine learning model $g(\cdot)$ to rectify mismatch between $f(\cdot)$ and the real process. During process control, the mechanistic model is fixed whilst the machine learning model (e.g., Gaussian processes) is updated using real-time data to guarantee hybrid model accuracy. Such an update becomes more time-efficient when the differential equation is converted to a difference equation for control purposes. Borrowing this concept, for adaptive or evolving systems, one can also build a discrepancy model consisting of a comprehensive but fixed mechanistic model and a simple but adaptive machine learning model. Such models have the potential to effectively solve tasks related to process control and online optimization. Transfer learning can somewhat be framed in the same manner, just that the data-driven model would estimate parameters of the mechanistic model, e.g., the reaction rate. Hence, the data-driven model would evolve when applying the hybrid model, e.g., to a bioprocess with the same species but a different product.

Transfer learning has been successfully adopted to develop data-driven models for new system prediction (Rogers et al., 2022; Xiao and Wu, 2023). Two strategies have been developed for transfer model construction based on neural network topology selection and parameter regularization, both only requiring minimum process data. By integrating transfer learning within hybrid model construction, it is possible to efficiently update model parameters during an ongoing process. Alternatively, one could also use, e.g., an embedding approach to bridge between systems that exhibit generally the same behavior but show some sub-system-specific behavior (Hutter et al., 2021) or, e.g., a meta-learner (Weiss et al., 2016). The incorporated knowledge also in this case might facilitate or hinder the learning exercise.

Finally, recent advances in physics-informed neural network (PINN) (Raissi et al., 2019) may also help with hybrid model construction. PINN can be used to solve two classes of problems: data-driven solution (i.e., forward problem) and data-driven discovery of differential equations (i.e., inverse problem). For a forward problem, both process data and process model are known. The PINN is trained to approximate process states across the spatial and temporal dimension. In other words, the inputs of a PINN are time (t) and space ($\boldsymbol{x}$), whereas the output of a PINN is state ($\boldsymbol{s}$) (i.e., $\boldsymbol{s} = P(\boldsymbol{x}, t)$). The forward problem of PINN can also be viewed as surrogate modeling an area that is widely investigated in process systems and design engineering (Misener and Biegler, 2023; Alizadeh et al., 2020; Viana et al., 2021; Sansana et al., 2021). However, its inverse problem can help identify hybrid model structure. For an inverse problem, only process data is available. A process model structure (normally a comprehensive model structure incorporating different possible mechanisms) is available but parameters within each term are unknown. The PINN is trained to identify the correct model structure and its associated parameters. Thus, PINN can be used as an alternative technique for hybrid model automatic structure discrimination, and by feeding new process data, PINN may help identify the evolved model structure throughout the process time course. This is similar to the concept of sparse identification of nonlinear dynamics (SINDy). However, so far the inverse problem of PINN is only applied to systems containing time-independent parameters. The traditional PINN structure ($\boldsymbol{s} = P(\boldsymbol{x}, t)$) is not capable of simulating systems with time-varying parameters. To resolve this challenge, one could use the approach of Vega-Ramon et al. (2021) which converts a time-varying parameter p into a vector, as described before. Similarly to the RL based strategy discussed in Section 3.2, the PINN based approach is at its infant stage and it remains unknown if it can be an efficient method. The only preliminary investigation conducted so far (Rogers et al., 2023a) suggests that the RL based approach has better performance than the PINN based approach for hybrid model structure discrimination. However, with the support of transfer learning, PINN could be more data-efficient than RL when being applied for hybrid model structure and parameter update.

3.5. Uncertainty quantification of hybrid models

Uncertainty estimation has been considered as one of the most pressing challenges for hybrid model development. Given the commonly observed batch-to-batch variations within the formulated chemicals industry and pharmaceutical industry, predicting process uncertainty is of great importance to guarantee product quality and minimize batch failure. As uncertainty includes both epistemic uncertainty (i.e., due to a lack of information about a particular situation) and aleatory uncertainty (i.e., inherent to the nature of a random chance determined situation), predicting process uncertainty is a difficult task for most real-world engineering applications. From a model construction point of view, model uncertainty can be divided into parametric uncertainty and structural uncertainty. Depending on the nature of the model, different techniques have been developed to approximate model uncertainty.

For data-driven models, as their parameters are non-identifiable, it is not common to divide their uncertainty into parametric and structural uncertainty. Instead, only the overall model uncertainty is considered. One of the most widely used techniques is bootstrapping through which a machine learning model (e.g., artificial neural network) is trained using different sub-sets of data to generate a sample of prediction results. This sample is then used to approximate the mean and standard deviation of true model prediction. Bootstrapping is easy to implement and has been adopted for hybrid models (Pinto et al., 2019; Bayer et al., 2020b; Polak et al., 2022), but this method does not well estimate uncertainty in many cases (Mostofian and Zuckerman, 2019).

Therefore, recently heteroscedastic noise neural networks (HNNs) have been proposed to simulate model uncertainty by reformulating the network structure and the training loss function (Kay et al., 2022). HNNs automatically output the mean of the prediction and the variance of the residuals by using the negative log-likelihood (NLL) loss function. Although more rigorous, HNNs only include aleatoric uncertainty due to its frequentist characteristic. To account for both aleatoric and epistemic uncertainty, probabilistic machine learning models such as Gaussian processes and Bayesian neural networks have been applied to different studies and have shown great potential. Though Gaussian process models have also been integrated into hybrid models (Vega-Ramon et al., 2021; Hutter et al., 2021; Cruz-Bournazou et al., 2022), these applications have not investigated uncertainty. In particular, it would be interesting to understand how parameter identification might impact on the uncertainty representation, as the parametric part induces an inductive bias.

For mechanistic models, as their model structure is formulated based on process knowledge, both structural uncertainty and parametric uncertainty have been investigated extensively. For model structural uncertainty, this is mainly addressed through model structure discrimination and model-based design of experiments. Although these

frameworks have been well established in the literature, it is critical to highlight that given the adaptive and evolving nature of many chemical and biochemical processes, the dominating process mechanism can shift over time and therefore the identified model structure may only hold true within a certain range of operating conditions. Automatically predicting mechanistic model structural uncertainty is an open challenge. Nevertheless, as discussed in Section 3.4, in practice when applying a mechanistic model for process modeling and optimization, it is more common to only update model parameters while fixing the model structure. As a result, in most cases the model uncertainty is approximated using parametric uncertainty. For parametric uncertainty, classic statistic methods (e.g., using Hessian matrix to approximate the Fisher information matrix) have been well established and widely used in many studies and this concept has also been extended to hybrid models (Kahrs and Marquardt, 2007). Again, caution must be taken that this parametric uncertainty is only a linear approximation of the true parametric uncertainty at the proxy of the optimal solution to the model parameters.

As one can expect, due to the difficulty in hybrid model parameter estimation and structure discrimination, accurately quantifying hybrid model uncertainty has been rarely explored. The traditional simultaneous parameter estimation approach for hybrid model construction is incapable of estimating parametric uncertainty given that the data-driven model compartment is simultaneously constructed along with parameter estimation for the mechanistic model compartment (i.e., parametric uncertainty is influenced by the data-driven model and cannot be estimated accurately). It is also inefficient to estimate structural uncertainty as this would be extremely time consuming to compare various combinations of mechanistic model candidates with data-driven model candidates. Nonetheless, the recently proposed two-step approach can greatly ease the procedure for hybrid model uncertainty estimation. As described above in Section 3.1, the two-step approach can effectively decouple hybrid model construction into mechanistic model construction and data-driven model construction. Therefore, within the first step, mechanistic model structural uncertainty can be initially conducted by assuming all parameters are time-independent. In this way, the best performed mechanistic model structure can be identified. Then, parameter estimation can be conducted to determine which parameters should be switched from time-independent to time-varying in order to reduce uncertainty and improve model accuracy. Once determined, parameter estimation can be carried out again to calculate the means and variances of both time-varying parameters (i.e., the mean and variance of the parameter at each time step) and time-independent parameters. This information is finally passed to the second stage so that a machine learning model can be constructed accurately. Within this stage, variances of time-varying parameters can be used to generate a large number of synthetic data and probabilistic machine learning models can be applied to for data-driven model construction. Through this approach, the overall model uncertainty can be estimated efficiently. A more detailed illustration can be found in the recent study (Rogers et al., 2023b).

Nonetheless, the two-step approach also has its drawback. The greatest advantage of hybrid models is the ability to integrate domain knowledge with process data. However, there are several key questions that fundamentally influence hybrid model uncertainty but have not been answered in the literature. For instance, how to systematically balance the amount of process knowledge (i.e., inductive bias) and the amount of process data? To which extent should we trust and prioritize domain knowledge over measured data? Moreover, how to determine the level of our confidence on hypothesized prior knowledge (particularly if the system is only partially understood) against experimental measurements? As the two-step approach separates mechanistic model construction and data-driven model construction, it cannot automatically take these questions into account (i.e. the mechanistic model constructed in Step 1 will dictate performance of the data-driven model in Step 2, meaning that this two-step approach is naturally biased towards domain knowledge).

In the recent study (Rogers et al., 2023b), the researchers attempted to answer the above questions by investigating a fermentation process. Three hybrid models were constructed using the two-step approach to embed different amount of first-principle knowledge, namely a black hybrid model (i.e., only using minimum process knowledge), a grey hybrid model (i.e., only using process knowledge with high confidence), and a white hybrid model (i.e., using all possible process knowledge based on literature). Once constructed, the models were used to predict dynamics of the fermentation process operated in a larger bioreactor. They observed that the black hybrid model fails to provide an acceptable model fitting result and exhibits significantly larger uncertainty compared to the other two; whilst compared to the grey hybrid model, the white hybrid model has lower prediction accuracy but is more confident (i.e., false confidence). This suggests that without carefully balancing the selection of domain knowledge vs. process data, the uncertainty estimated by a hybrid model cannot be fully trusted (e.g., either over-conservative or overconfident). As a result, developing an efficient and automatic uncertainty quantification framework remains a top priority for hybrid model studies.

3.6. The validity domain of hybrid models

Rather than assessing the uncertainty of the model outputs in form of prediction or confidence intervals (as obtained from the approaches described hitherto), the domain of inputs for which hybrid models (or data-driven models) produce "valid" results can be sought, i.e., the validity domain. This domain is spanned by the inputs of the nonparametric model as well as the variation in the input data on which the hybrid model (and implicitly the nonparametric model) was trained. While this seems to imply that the nonparametric model is the only component carrying uncertainty, uncertainty stemming from the parametric part also seems to be implicitly accounted for. This is because, if the model describes the process accurately for the data the model was trained on (i.e., the model is valid), then likely the uncertainty in this domain is low.

A simple approach to characterize the validity domain, would be the use of ranges for each input (i.e., constructing a hyperrectangle in the input space). However, this disregards the multivariate nature of processes and in many cases will overestimate the validity of the model. While using a convex hull, build on around the training data (Kahrs and Marquardt, 2007; Bae et al., 2020), would overcome this shortcoming for scenarios where data are scattered across the input domain the convex hull might also overestimate validity because the input data domain might be nonconvex or there might be "holes" in the input data domain. Still, it might be attractive to use the convex hull approach when the data is homogeneously distributed, as, e.g., the validity criteria can be formulated in form of a linear constraint, which renders numerical optimization efficient.

In case of sparse data and potential holes in the input data domain, a topological data analysis such as persistent homology can be performed to identify such holes (Schweidtmann et al., 2021c). Moreover, two alternative approaches exist to model scattered data domains in hybrid models: k-means clustering (Teixeira et al., 2006; Ferreira et al., 2014; von Stosch et al., 2016; Bangi and Kwon, 2023) and one-class classification (e.g., using support vector machines) (Schweidtmann et al., 2021c). For both approaches, a threshold needs to be set that determines what is within and what without the validity domain. As both approaches allow for a more gradual understanding of how distant a novel data point is in relation to past data and as "small" excursions into a space for which the model was not trained could be allowed for an overall, integral threshold level can be set. As suggested by Teixeira et al. (2006) a risk-measure can be defined and integrated into the optimization as a nonlinear constraint.

4. Conclusion

The concept of hybrid modeling—integrating both parametric (knowledge-based) and nonparametric (data-derived) models—holds significant promise for the advancement of various scientific and industrial domains. Despite relevant scientific advances over two decades and its contribution to over 150 scientific publications, there remains a considerable gap in its comprehensive understanding and application. For instance, how to best extend an existing mechanistic model with a machine-learning model to account for changes in process parameters or material attributes, i.e., how to avoid combining the shortcomings of both models but rather combine their strengths? This fundamental development of hybrid modeling is not just a matter of academic nuance; it potentially undermines the methodological advancements that such an approach can bring to critical fields, including systems biology, personalized medicine, material design, and process industries. It is paramount to delve methodically into the inherent properties and capabilities of hybrid models. Such an exploration will not only harness the capabilities of machine learning but also optimize experimental protocols and enhance the reliability of predictive domains.

Future research and application of hybrid modeling research should focus on dispelling prevailing misconceptions and emphasizing its pivotal role in fostering scientific innovation and precision. As promising future research challenges in the field of hybrid modeling, we identify the (1) identification and use of dynamic hybrid models, (2) the automatic structure discrimination, (3) incremental learning, (4) adaptive and evolving systems, (5) uncertainty quantification, and (6) modeling the validity domain.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Moritz von Stosch reports a relationship with DataHow AG that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agharafeie, Roshanak, Oliveira, Rui, Rodrigues, João, Ramos, Correia, Mendes, Jorge M., 2023. Application of hybrid neural models to bioprocesses: A systematic literature review. Authorea Preprints.

Alhajeri, Mohammed S., Luo, Junwei, Wu, Zhe, Albalawi, Fahad, Christofides, Panagiotis D., 2022. Process structure-based recurrent neural network modeling for predictive control: A comparative study. Chem. Eng. Res. Des. 179, 77–89.

Alizadeh, Reza, Allen, Janet K., Mistree, Farrokh, 2020. Managing computational complexity using surrogate models: a critical review. Res. Eng. Des. 31 (3), 275–298.

Antoniewicz, Maciek R., 2015. Methods and advances in metabolic flux analysis: a mini-review. J. Ind. Microbiol. Biotechnol. 42 (3), 317–325.

Antontsev, Victor, Jagarapu, Aditya, Bundey, Yogesh, Hou, Hypatia, Khotimchenko, Maksim, Walsh, Jason, Varshney, Jyotika, 2021. A hybrid modeling approach for assessing mechanistic models of small molecule partitioning in vivo using a machine learning-integrated modeling platform. Sci. Rep. 11 (1), 11143.

Azevedo, Cristiana Rodrigues, Díaz, Victor Grisales, Prado-Rubio, Oscar Andrés, Willis, Mark J., Préat, Véronique, Oliveira, Rui, Stosch, Moritz, 2019. Hybrid semiparametric modeling: A modular process systems engineering approach for the integration of available knowledge sources. In: Systems Engineering in the Fourth Industrial Revolution. Wiley, pp. 345–373.

Bae, Jaehan, Lee, Hye Ji, Jeong, Dong Hwi, Lee, Jong Min, 2020. Construction of a valid domain for a hybrid model and its application to dynamic optimization with controlled exploration, 59 (37). pp. 16380–16395.

Baker, Ruth E., Peña, Jose-Maria, Jayamohan, Jayaratnam, Jérusalem, Antoine, 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol. Lett. 14 (5), 20170660.

Bangi, Mohammed Saad Faizan, Kwon, Joseph Sang-Il, 2023. Deep hybrid model-based predictive control with guarantees on domain of applicability. AIChE J. 69 (5), e18012.

Bayer, B., Sissolak, B., Duerkop, M., von Stosch, M., Striedner, G., 2020a. The shortcomings of accurate rate estimations in cultivation processes and a solution for precise and robust process modeling. Bioprocess Biosyst. Eng. 43 (2), 169–178.

Bayer, Benjamin, von Stosch, Moritz, Striedner, Gerald, Duerkop, Mark, 2020b. Comparison of modeling methods for DoE-based holistic upstream process characterization. Biotechnol. J. 15 (5), 1900551.

Bikmukhametov, Timur, Jäschke, Johannes, 2020. Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. Comput. Chem. Eng. 138, 106834.

Bradley, William, Kim, Jinhyeun, Kilwein, Zachary, Blakely, Logan, Eydenberg, Michael, Jalvin, Jordan, Laird, Carl, Boukouvala, Fani, 2022. Perspectives on the integration between first-principles and data-driven modeling. Comput. Chem. Eng. 166, 107898.

Brunton, Steven L., Proctor, Joshua L., Kutz, J. Nathan, 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. 113 (15), 3932–3937.

Bui, Linh, Joswiak, Mark, Castillo, Ivan, Phillips, Ailene, Yang, Jin, Hickman, Daniel, 2022. A hybrid modeling approach for catalyst monitoring and lifetime prediction. ACS Eng. Au 2 (1), 17–26.

Cabaneros Lopez, Pau, Udugama, Isuru A., Thomsen, Sune T., Roslander, Christian, Junicke, Helena, Iglesias, Miguel M., Gernaey, Krist V., 2021. Transforming data to information: A parallel hybrid model for real-time state estimation in lignocellulosic ethanol fermentation. Biotechnol. Bioeng. 118 (2), 579–591.

Chakraborty, Arijit, Sivaram, Abhishek, Venkatasubramanian, Venkat, 2021. AI-DARWIN: A first principles-based model discovery engine using machine learning. Comput. Chem. Eng. 154, 107470.

Chefrour, Aida, 2019. Incremental supervised learning: algorithms and applications in pattern recognition. Evol. Intell. 12 (2), 97–112.

Clifton, Sara M., Kang, Chaeryon, Li, Jingyi Jessica, Long, Qi, Shah, Nirmish, Abrams, Daniel M., 2017. Hybrid statistical and mechanistic mathematical model guides mobile health intervention for chronic pain. J. Comput. Biol. 24 (7), 675–688, PMID: 28581814.

Cruz-Bournazou, M. Nicolas, Narayanan, Harini, Fagnani, Alessandro, Butte, Alessandro, 2022. Hybrid Gaussian process models for continuous time series in bolus fed-batch cultures. IFAC-PapersOnLine 55 (7), 204–209, 13th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems DYCOPS 2022.

Daoutidis, Prodromos, Lee, Jay H., Rangarajan, Srinivas, Chiang, Leo, Gopaluni, Bhushan, Schweidtmann, Artur M., Harjunkoski, Iiro, Mercangöz, Mehmet, Mesbah, Ali, Boukouvala, Fani, et al., 2023. Machine learning in process systems engineering: Challenges and opportunities. Comput. Chem. Eng. 108523.

Daume, Sven, Kofler, Sandro, Kager, Julian, Kroll, Paul, Herwig, Christoph, 2020. Generic workflow for the setup of mechanistic process models. pp. 189–211.

del Rio-Chanona, Ehecatl Antonio, Dechatiwongse, Pongsathorn, Zhang, Dongda, Maitland, Geoffrey C., Hellgardt, Klaus, Arellano-Garcia, Harvey, Vassiliadis, Vassilios S., 2015. Optimal operation strategy for biohydrogen production. Ind. Eng. Chem. Res. 54 (24), 6334–6343.

Dors, M., Simutis, R., Lübbert, A., 1995. Advanced supervision of mammalian cell cultures using hybrid process models. IFAC Proc. Vol. 28 (3), 72–77.

Ferreira, A.R., Dias, J.M.L., von Stosch, M., Clemente, J., Cunha, A.E., Oliveira, Rui, 2014. Fast development of pichia pastoris GS115 Mut+ cultures employing batch-to-batch control and hybrid semi-parametric modeling. Bioprocess Biosyst. Eng. 37 (4), 629–639.

Fiedler, Bernold, Schuppert, Andreas, 2008. Local Identification of Scalar Hybrid Models with Tree Structure, Vol. 73, No. 3. Oxford Academic, pp. 449–476.

Fischetti, Matteo, Jo, Jason, 2018. Deep neural networks and mixed integer linear optimization. Constraints 23 (3), 296–309.

Forster, Tim, Vázquez, Daniel, Cruz-Bournazou, Mariano Nicolas, Butté, Alessandro, Guillén-Gosálbez, Gonzalo, 2023. Modeling of bioprocesses via MINLP-based symbolic regression of S-system formalisms. Comput. Chem. Eng. 170, 108108.

Galvanauskas, Vytautas, Simutis, Rimvydas, Lübbert, Andreas, 2018. Hybrid modeling of biochemical processes. Hybrid Model. Process Ind. 89–127.

Gernaey, Krist V., Lantz, Anna Eliasson, Tufvesson, Pär, Woodley, John M., Sin, Gürkan, 2010. Application of mechanistic models to fermentation and biocatalysis for next-generation processes. Trends Biotechnol. 28 (7), 346–354.

Glassey, Jarka, von Stosch, Moritz (Eds.), 2018. Hybrid Model. Process Ind.. CRC Press.

Grimstad, Bjarne, Andersson, Henrik, 2019. Relu networks as surrogate models in mixed-integer linear programs. Comput. Chem. Eng. 131, 106580.

Hamilton, Franz, Lloyd, Alun L., Flores, Kevin B., 2017. Hybrid modeling and prediction of dynamical systems. PLoS Comput. Biol. 13 (7), e1005655.

Herold, Sebastian, King, Rudibert, 2014. Automatic identification of structured process models based on biological phenomena detected in (fed-)batch experiments. Bioprocess Biosyst. Eng. 37 (7), 1289–1304.

Hilbe, Joseph M., 2011. Generalized linear models. In: Lovric, Miodrag (Ed.), International Encyclopedia of Statistical Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 591–596.

Hinchliffe, Mark P., Willis, Mark J., 2003. Dynamic systems modelling using genetic programming. Comput. Chem. Eng. 27 (12), 1841–1854.

Hinton, Geoffrey E., Osindero, Simon, Teh, Yee Whye, 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554.

Hinton, Geoffrey E., Srivastava, Nitish, Krizhevsky, Alex, Sutskever, Ilya, Salakhutdinov, Ruslan R., 2012. Improving neural networks by preventing co-adaptation of feature detectors.

Hornik, Kurt, 1993. Some new results on neural network approximation. Neural Netw. 6 (8), 1069–1072.
Horstemeyer, Mark F., 2010. Multiscale modeling: a review. In: Practical Aspects of Computational Chemistry: Methods, Concepts and Applications. Springer, pp. 87–135.
Hutter, Clemens, von Stosch, Moritz, Cruz Bournazou, Mariano N., Butte, Alessandro, 2021. Knowledge transfer across cell lines using hybrid Gaussian process models with entity embedding vectors. Biotechnol. Bioeng. 118 (11), 4389–4401.
Iliadis, Athanassios, 2019. Structural identifiability and sensitivity. J. Pharmacokinet. Pharmacodyn. 46 (2), 127–135.
Johansen, Tor A., Foss, Bjarne A., 1992. Representing and learning unmodeled dynamics with neural network memories. In: 1992 American Control Conference. pp. 3037–3043.
Jumper, John, Evans, Richard, Pritzel, Alexander, Green, Tim, Figurnov, Michael, Ronneberger, Olaf, Tunyasuvunakool, Kathryn, Bates, Russ, Žídek, Augustin, Potapenko, Anna, et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589.
Kahrs, O., Marquardt, W., 2007. The validity domain of hybrid models and its application in process optimization. Chem. Eng. Process. Process Intensif. 46 (11), 1054–1066.
Kahrs, O., Marquardt, W., 2008. Incremental identification of hybrid process models. Comput. Chem. Eng. 32 (4–5), 694–705.
Karlsson, Johan, Anguelova, Milena, Jirstrand, Mats, 2012. An efficient method for structural identifiability analysis of large dynamic systems. In: IFAC Proceedings Volumes (IFAC-PapersOnline), Vol. 16, No. PART 1. Elsevier, pp. 941–946.
Kay, Sam, Kay, Harry, Mowbray, Max, Lane, Amanda, Mendoza, Cesar, Martin, Philip, Zhang, Dongda, 2022. Integrating autoencoder and heteroscedastic noise neural networks for the batch process soft-sensor design. Ind. Eng. Chem. Res. 61 (36), 13559–13569.
Kramer, Mark A., Thompson, Michael L., Bhagat, Phiroz M., 1992. Embedding theoretical models in neural networks. In: 1992 American Control Conference. pp. 475–479.
Kroll, Paul, Hofer, Alexandra, Stelzer, Ines V., Herwig, Christoph, 2017. Workflow to set up substantial target-oriented mechanistic process models in bioprocess engineering. Process Biochem. 62, 24–36.
Lee, Dongheon, Jayaraman, Arul, Kwon, Joseph S., 2020. Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling. PLoS Comput. Biol. 16 (12), 1–31.
Lee, Jay H., Shin, Joohyun, Realff, Matthew J., 2018. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. Comput. Chem. Eng. 114, 111–121.
Lindsey, J.K., 1999. A review of some extensions to generalized linear models. Stat. Med. 18 (17–18), 2223–2236.
Mahanty, Biswanath, 2023. Hybrid modeling in bioprocess dynamics: Structural variabilities, implementation strategies, and practical challenges. Biotechnol. Bioeng. 120 (8), 2072–2091.
Massonis, Gemma, Villaverde, Alejandro F., Banga, Julio R., 2023. Distilling identifiable and interpretable dynamic models from biological data. PLoS Comput. Biol. 19 (10), 1–27.
McBride, Kevin, Sanchez Medina, Edgar Ivan, Sundmacher, Kai, 2020. Hybrid semiparametric modeling in separation processes: A review. Chem. Ing. Tech. 92 (7), 842–855.
McKay, Ben, Willis, Mark, Barton, Geoffrey, 1997. Steady-state modelling of chemical process systems using genetic programming. Comput. Chem. Eng. 21 (9), 981–996.
Misener, Ruth, Biegler, Lorenz, 2023. Formulating data-driven surrogate models for process optimization. Comput. Chem. Eng. 179, 108411.
Montáns, Francisco J., Chinesta, Francisco, Gómez-Bombarelli, Rafael, Kutz, J. Nathan, 2019. Data-driven modeling and learning in science and engineering. C. R. Méc. 347 (11), 845–855.
Mostofian, Barmak, Zuckerman, Daniel M., 2019. Statistical uncertainty analysis for small-sample, high log-variance data: Cautions for bootstrapping and Bayesian bootstrapping. J. Chem. Theory Comput. 15 (6), 3499–3509, PMID: 31002504.
Mowbray, Max R., Wu, Chufan, Rogers, Alexander W., Rio-Chanona, Ehecatl A. Del, Zhang, Dongda, 2023. A reinforcement learning-based hybrid modeling framework for bioprocess kinetics identification. Biotechnol. Bioeng. 120 (1), 154–168.
Narayanan, Harini, Behle, Lars, Luna, Martin F., Sokolov, Michael, Guillén-Gosálbez, Gonzalo, Morbidelli, Massimo, Butté, Alessandro, 2020. Hybrid-EKF: Hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture. Biotechnol. Bioeng. 117 (9), 2703–2714.
Narayanan, Harini, Cruz Bournazou, Mariano Nicolas, Guillén Gosálbez, Gonzalo, Butté, Alessandro, 2022. Functional-hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. Chem. Eng. J. 430, 133032.
Oliveira, R., 2004. Combining first principles modelling and artificial neural networks: a general framework. Comput. Chem. Eng. 28 (5), 755–766.
Peres, J., Oliveira, R., de Azevedo, S. Feyo, 2008. Bioprocess hybrid parametric/nonparametric modelling based on the concept of mixture of experts. Biochem. Eng. J. 39 (1), 190–206.
Peres, J., Oliveira, R., Feyo De Azevedo, S., 2001. Knowledge based modular networks for process modelling and control. In: Comput. Chem. Eng.. 25, (4–6), Pergamon, pp. 783–791.
Pinto, Jose, Mestre, Mykaella, Ramos, J., Costa, Rafael S., Striedner, Gerald, Oliveira, Rui, 2022. A general deep hybrid model for bioreactor systems: Combining first principles with deep neural networks. Comput. Chem. Eng. 165, 107952.
Pinto, Jose, Ramos, Joao R.C., Costa, Rafael S., Oliveira, Rui, 2023. A general hybrid modeling framework for systems biology applications: Combining mechanistic knowledge with deep neural networks under the SBML standard. AI 4 (1), 303–318.
Pinto, José, Rodrigues de Azevedo, Cristiana, Oliveira, Rui, von Stosch, Moritz, de Azevedo, Cristiana Rodrigues C.R., Oliveira, Rui, von Stosch, Moritz, 2019. A bootstrap aggregated hybrid semi-parametric modeling framework for bioprocess development. Bioprocess Biosyst. Eng. 42 (11), 1853–1865.
Polak, Jakub, v. Stosch, Moritz, Sokolov, Michael, Piccioni, Lorenzo, Streit, Alexander, Schenkel, Berthold, Guelat, Bertrand, 2022. Hybrid modeling supported development of an industrial small-molecule flow chemistry process. Comput. Chem. Eng. 108127.
Portela, Rui M.C., von Stosch, Moritz, Oliveira, Rui, 2018. Hybrid semiparametric systems for quantitative sequence-activity modeling of synthetic biological parts. Synth. Biol. ysy010.
Psichogios, Dimitris C., Ungar, Lyle H., 1992. A hybrid neural network-first principles approach to process modeling. AIChE J. 38 (10), 1499–1511.
Quaghebeur, Ward, Torfs, Elena, De Baets, Bernard, Nopens, Ingmar, 2022. Hybrid differential equations: Integrating mechanistic and data-driven techniques for modelling of water systems. Water Res. 213, 118166.
Raissi, Maziar, Perdikaris, Paris, Karniadakis, George E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707.
Rajulapati, Lokesh, Chinta, Sivadurgaprasad, Shyamala, Bala, Rengaswamy, Raghunathan, 2022. Integration of machine learning and first principles models. AIChE J. 68 (6), e17715.
Rasmussen, Carl Edward, Williams, Christopher K.I., 2006. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, Cambridge, p. 266.
Read, Jesse, Bifet, Albert, Pfahringer, Bernhard, Holmes, Geoff, 2012. Batch-Incremental Versus Instance-Incremental Learning in Dynamic and Evolving Data. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7619 LNCS, Springer, Berlin, Heidelberg, pp. 313–323.
Richelle, Anne, Lee, Boung Wook, Portela, Rui M.C., Raley, Jonathan, Stosch, Moritz, 2020. Analysis of transformed upstream bioprocess data provides insights into biological system variation. Biotechnol. J. 2000113.
Rodrigues de Azevedo, Cristiana, von Stosch, Moritz, Costa, M.S. Mariana S., Ramos, A.M., Cardoso, M.M. Margarida, Danhier, Fabienne, Préat, Véronique, Oliveira, Rui, 2017. Modeling of the burst release from PLGA micro- and nanoparticles as function of physicochemical parameters and formulation characteristics. Int. J. Pharm. 532 (1), 229–240.
Rogers, Alexander William, Cardenas, Ilya Orson Sandoval, Del Rio-Chanona, Ehecatl Antonio, Zhang, Dongda, 2023a. Investigating physics-informed neural networks for bioprocess hybrid model construction. In: Kokossis, Antonios C., Georgiadis, Michael C., Pistikopoulos, Efstratios (Eds.), 33rd European Symposium on Computer Aided Process Engineering. In: Computer Aided Chemical Engineering, vol. 52, Elsevier, pp. 83–88.
Rogers, Alexander W., Song, Ziqi, Ramon, Fernando Vega, Jing, Keju, Zhang, Dongda, 2023b. Investigating 'greyness' of hybrid model for bioprocess predictive modelling. Biochem. Eng. J. 190, 108761.
Rogers, Alexander W., Vega-Ramon, Fernando, Yan, Jiangtao, del Río-Chanona, Ehecatl A., Jing, Keju, Zhang, Dongda, 2022. A transfer learning approach for predictive modeling of bioprocesses using small data. Biotechnol. Bioeng. 119 (2), 411–422.
Sahinidis, Nick, 2016. The ALAMO approach to machine learning. p. 2410.
Sansana, Joel, Joswiak, Mark N., Castillo, Ivan, Wang, Zhenyu, Rendall, Ricardo, Chiang, Leo H., Reis, Marco S., 2021. Recent trends on hybrid modeling for industry 4.0. Comput. Chem. Eng. 151, 107365.
Schäfer, Pascal, Caspari, Adrian, Schweidtmann, Artur M., Vaupel, Yannic, Mhamdi, Adel, Mitsos, Alexander, 2020. The potential of hybrid mechanistic/data-driven approaches for reduced dynamic modeling: Application to distillation columns. Chem. Ing. Tech. 92 (12), 1910–1920.
Schmidt, Michael, Lipson, Hod, 2009. Distilling free-form natural laws from experimental data. Science 324 (5923), 81–85.
Schneider, Mariane Yvonne, Quaghebeur, Ward, Borzooei, Sina, Froemelt, Andreas, Li, Feiyi, Saagi, Ramesh, Wade, Matthew J., Zhu, Jun-Jie, Torfs, Elena, 2022. Hybrid modelling of water resource recovery facilities: status and opportunities. Water Sci. Technol. 85 (9), 2503–2524.
Schuppert, Andreas A., 2000. Extrapolability of structured hybrid models: a key to optimization of complex processes. In: Equadiff 99. pp. 1135–1151.
Schweidtmann, Artur M., Bongartz, Dominik, Grothe, Daniel, Kerkenhoff, Tim, Lin, Xiaopeng, Najman, Jaromił, Mitsos, Alexander, 2021a. Deterministic global optimization with Gaussian processes embedded. Math. Program. Comput. 13 (3), 553–581.
Schweidtmann, Artur M., Esche, Erik, Fischer, Asja, Kloft, Marius, Repke, Jens-Uwe, Sager, Sebastian, Mitsos, Alexander, 2021b. Machine learning in chemical engineering: A perspective. Chem. Ing. Tech. 93 (12), 2029–2039.

Schweidtmann, Artur M., Mitsos, Alexander, 2019. Deterministic global optimization with artificial neural networks embedded. J. Optim. Theory Appl. 180 (3), 925–948.
Schweidtmann, Artur M., Weber, Jana M., Wende, Christian, Netze, Linus, Mitsos, Alexander, 2021c. Obey validity limits of data-driven models through topological data analysis and one-class classification. Opt. Eng. 23 (2), 855–876.
Searson, Dominic P., 2015. GPTIPS 2: An open-source software platform for symbolic data mining. In: Handbook of Genetic Programming Applications. Springer International Publishing, Cham, pp. 551–573.
Seborg, Dale E., Edgar, Thomas F., Mellichamp, Duncan A., Doyle, Francis J. (Eds.), 2016. Process Dynamics and Control, fourth ed. Wiley.
Shah, Parth, Sheriff, M. Ziyan, Bangi, Mohammed Saad Faizan, Kravaris, Costas, Kwon, Joseph Sang-Il, Botre, Chiranjivi, Hirota, Junichi, 2022. Deep neural network-based hybrid modeling and experimental validation for an industry-scale fermentation process: Identification of time-varying dependencies among parameters. Chem. Eng. J. 441, 135643.
Sharma, Niket, Liu, Y.A., 2022. A hybrid science-guided machine learning approach for modeling chemical processes: A review. AIChE J. 68 (5), e17609.
Simutis, Rimvydas, Lübbert, Andreas, 2017. Hybrid approach to state estimation for bioprocess control. Bioengineering 4 (4), 21.
Sitapure, Niranjan, Sang-Il Kwon, Joseph, 2023. Introducing hybrid modeling with time-series-transformers: A comparative study of series and parallel approach in batch crystallization. Ind. Eng. Chem. Res.
Sokolov, Michael, von Stosch, Moritz, Narayanan, Harini, Feidl, Fabian, Butté, Alessandro, 2021. Hybrid modeling a key enabler towards realizing digital twins in biopharma? Curr. Opin. Chem. Eng. 34, 100715.
Su, Hong-Te, Bhat, N., Minderman, P.A., McAvoy, T.J., 1993. Integrating neural networks with first principles models for dynamic modeling. In: Balchen, J.G. (Ed.), Dynamics and Control of Chemical Reactors, Distillation Columns and Batch Processes. In: IFAC Symposia Series, Pergamon, Oxford, pp. 327–332.
Sun, Sheng, Ouyang, Runhai, Zhang, Bochao, Zhang, Tong-Yi, 2019. Data-driven discovery of formulas by symbolic regression. MRS Bull. 44 (7), 559–564.
Teixeira, A.P. Ana P., Clemente, J.J. João J., Cunha, A.E. António E., Carrondo, M.J.T. Manuel J.T., Oliveira, Rui, 2006. Bioprocess iterative batch-to-batch optimization based on hybrid parametric/nonparametric models. Biotechnol. Prog. 22 (1), 247–258.
Thompson, Michael L., Kramer, Mark A., 1994. Modeling chemical processes using prior knowledge and neural networks. AIChE J. 40 (8), 1328–1340.
Tsay, Calvin, Kronqvist, Jan, Thebelt, Alexander, Misener, Ruth, 2021. Partition-based formulations for mixed-integer optimization of trained ReLU neural networks. Adv. Neural Inf. Process. Syst. 34, 3068–3080.
Tsopanoglou, Apostolos, Jiménez del Val, Ioscani, 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.
van Can, H.J.L., te Braake, H.A.B., Bijman, A., Hellinga, C., Luyben, K.Ch.A.M., Heijnen, J.J., 1999. An efficient model development strategy for bioprocesses based on neural networks in macroscopic balances: Part II. Biotechnol. Bioeng. 62 (6), 666–680.
Van Can, Henricus J.L., Te Braake, Hubert A.B., Dubbelman, Sander, Hellinga, Chris, Luyben, Karel Ch.A.M., Heijnen, Joseph J., 1998. Understanding and applying the extrapolation properties of serial gray-box models. AIChE J. 44 (5), 1071–1089.
Vega-Ramon, Fernando, Zhu, Xianfeng, Savage, Thomas R., Petsagkourakis, Panagiotis, Jing, Keju, Zhang, Dongda, 2021. Kinetic and hybrid modeling for yeast astaxanthin production under uncertainty. Biotechnol. Bioeng. 118 (12), 4854–4866.
Venkatasubramanian, Venkat, 2019. The promise of artificial intelligence in chemical engineering: Is it here, finally? AIChE J. 65 (2), 466–478.
Viana, Felipe A.C., Gogu, Christian, Goel, Tushar, 2021. Surrogate modeling: tricks that endured the test of time and some recent developments. Struct. Multidiscip. Optim. 64 (5), 2881–2908.
Villaverde, Alejandro F., 2019. Observability and structural identifiability of nonlinear biological systems. In: Complexity, Vol. 2019. Hindawi Limited.
von Stosch, Moritz, Carinhas, Nuno, Oliveira, Rui, 2014a. Hybrid modeling for systems biology: Theory and practice. Model. Simul. Sci. Eng. Technol. 65, 367–388.
von Stosch, Moritz, Hamelink, Jan-Martijn, Oliveira, Rui, 2016. Toward intensifying design of experiments in upstream bioprocess development: An industrial Escherichia coli feasibility study. Biotechnol. Prog. 32 (5), 1343–1352.
von Stosch, Moritz, Oliveira, Rui, Peres, Joana, Feyo de Azevedo, Sebastião, 2014b. Hybrid semi-parametric modeling in process systems engineering: Past, present and future, 60. pp. 86–101.
von Stosch, Moritz, Peres, Joana, de Azevedo, Sebastião Feyo, Oliveira, Rui, 2010. Modelling biochemical networks with intrinsic time delays: a hybrid semi-parametric approach. BMC Syst. Biol. 4, 131.
Wang, Haohan, Raj, Bhiksha, 2017. On the origin of deep learning.
Wang, Xizhao, Zhao, Yanxia, Pourpanah, Farhad, 2020. Recent advances in deep learning. Int. J. Mach. Learn. Cybern. 11 (4), 747–750.
Ward, Eric J., 2008. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. Ecol. Model. 211 (1–2), 1–10.
Weiss, Karl, Khoshgoftaar, Taghi M., Wang, DingDing, 2016. A survey of transfer learning. J. Big Data 3 (1), 9.
Willis, Mark, Hiden, Hugo, Hinchliffe, Mark, McKay, Ben, Barton, Geoffrey W., 1997. Systems modelling using genetic programming. Comput. Chem. Eng. 21, S1161–S1166.
Willis, Mark J., von Stosch, Moritz, 2016. Inference of chemical reaction networks using mixed integer linear programming. Comput. Chem. Eng. 90, 31–43.
Willis, Mark J., von Stosch, Moritz, 2017. Simultaneous parameter identification and discrimination of the nonparametric structure of hybrid semi-parametric models. Comput. Chem. Eng. 104, 366–376.
Wilson, Zachary T., Sahinidis, Nikolaos V., 2017. The ALAMO approach to machine learning. Comput. Chem. Eng. 106, 785–795.
Wu, Zhe, Rincon, David, Christofides, Panagiotis D., 2020. Process structure-based recurrent neural network modeling for model predictive control of nonlinear processes. J. Process Control 89, 74–84.
Xiao, Ming, Wu, Zhe, 2023. Modeling and control of a chemical process network using physics-informed transfer learning. Ind. Eng. Chem. Res. 62 (42), 17216–17227.
Yang, Aidong, Martin, Elaine, Morris, Julian, 2011. Identification of semi-parametric hybrid process models, 35 (1). pp. 63–70.
Yang, Shu, Navarathna, Pranesh, Ghosh, Sambit, Bequette, B. Wayne, 2020. Hybrid modeling in the era of smart manufacturing. Comput. Chem. Eng. 140, 106874.
Yu, Tong, Zhu, Hong, 2020. Hyper-parameter optimization: A review of algorithms and applications.
Žegklitz, Jan, Pošík, Petr, 2021. Benchmarking state-of-the-art symbolic regression algorithms. Genet. Program. Evol. Mach. 22 (1), 5–33.
Zendehboudi, Sohrab, Rezaei, Nima, Lohi, Ali, 2018. Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review. Appl. Energy 228, 2539–2566.
Zhang, Dongda, Del Rio-Chanona, Ehecatl Antonio, Petsagkourakis, Panagiotis, Wagner, Jonathan, 2019. Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization. Biotechnol. Bioeng. 116 (11), 2919–2930.
Zhang, Dongda, Savage, Thomas R., Cho, Bovinille A., 2020. Combining model structure identification and hybrid modelling for photo-production process predictive simulation and optimisation. Biotechnol. Bioeng. 117 (11), 3356–3367.
Zheng, Yingzhe, Wu, Zhe, 2023. Physics-informed online machine learning and predictive control of nonlinear processes with parameter uncertainty. Ind. Eng. Chem. Res. 62 (6), 2804–2818.