A Hybrid Science-Guided Machine Learning Approach for Modeling Chemical Processes: A Review

Niket Sharma and Y. A. Liu*, AspenTech Center of Excellence in Process System Engineering, Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.  *Corresponding author: design@vt.edu; (540) 231-7800

Abstract:

This study presents a broad perspective of hybrid process modeling combining the scientific knowledge and data analytics in bioprocessing and chemical engineering with a science-guided machine learning (SGML) approach. We divide the approach into two major categories: ML complements science, and science complements ML. We review the literature relating to the hybrid SGML approach, and propose a systematic classification of hybrid SGML models. For applying ML to improve science-based models, we present expositions of direct serial and parallel hybrid modeling and their combinations, inverse modeling, reduced-order modeling, quantifying uncertainty in the process and even discovering governing equations of the process model. For applying scientific principles to improve ML models, we discuss the science-guided design, learning and refinement. For each sub-category, we identify its requirements, strengths and limitations, together with their published and potential applications. We also present several examples to illustrate different hybrid SGML methodologies for modeling chemical processes.

1 | INTRODUCTION

Modeling of many physiochemical systems requires detailed scientific knowledge of the system which is not always feasible for complex processes. We make some assumptions when modeling
the system with first principles that ultimately leads to some knowledge gaps in describing the original system. Even for the systems where the scientific knowledge is sufficient to model the system and there is limited data to estimate the multiple parameters of a first-principles model. We often apply data-based models to study the systems where scientific data are available since they are more accurate in prediction. However, data-based/machine learning models are *black-box models* which can over-fit the data and also produce scientifically inconsistent results. For better accuracy, ML models also require more data which is not always feasible for many problems. Therefore, it is important to integrate science-based knowledge and data-based knowledge for an accurate and scientifically consistent prediction, which we will refer to as *hybrid science-guided machine learning (SGML) approach*.

The most popular hybrid SGML approach that is being practiced in different fields of science is to combine a data-based ML model with a science-based first-principles model. However, there are more ways to combine scientific knowledge and data-based knowledge. In this work, we focus on both aspects of science complementing ML, and ML complementing science.

In our development of the hybrid SGML approach, we have benefited from two latest references. In their 2017 article, Karpatne et. al.\(^1\) suggest the theory-guided data science as a new paradigm for scientific discovery from data. They classify the theory-guided data science methods into different categories, such as theory-guided design of models, initialization, theory-guided refinement of data science outputs, hybrid models of theory of data science, and augmenting theory-based models using data science. In their 2020 article, Willard et. al.\(^2\) classify the integration of physics-based modeling with ML methodology according to the modeling objectives. The latter include, for example, improving the predictions beyond physical models,
downscaling the complexity of physics-based models, generating data, quantifying uncertainty, and discovering governing equations of the data-based model.

The objective of this paper is to present a review and exposition of scientific and engineering literature relating to the hybrid SGML approach, and propose a systematic classification of hybrid SGML models focusing on both science complementing ML models, and ML complementing science-based models. Section 2 gives a review of the broad applications of hybrid SGML approach in bioprocessing and chemical engineering. As the number of reported methodologies and applications continues to rise significantly, it is hard for a person unfamiliar with the subject to identify the appropriate approach for a specific application. This leads to our key focus in Sections 3 to 5, beginning with a systematic classification and exposition of hybrid SGML methodologies in Section 3. Section 4 explains different categories of applying ML to complement science-based models, discuss their requirements, strengths and limitations, suggest potential areas of applications, and present illustrative examples from chemical manufacturing. Section 5 focuses on different categories of applying scientific principles to complement ML models, together with their requirements, strengths and limitations, as well as their potential applications and illustrative examples. Section describes the challenges and opportunities for hybrid SGML approach for modeling chemical processes. Section 7 summarizes our conclusions.

This work differentiates itself from several recent reviews of hybrid modeling in bioprocessing and chemical engineering through the following contributions: (1) presentation of a broader hybrid SGML methodology of integrating science-guided and data-based models, and not just the direct combinations of first-principles and ML models; (2) classification of the hybrid model applications according to their methodology and objectives, instead of their areas of
applications; (3) identification of the themes and methodologies which have not been explored much in bioprocessing and chemical engineering applications, like the use of scientific knowledge to help improve the ML model architecture and learning process for more scientifically consistent solutions; and (4) illustrations of the use of these hybrid SGML methodologies applied to industrial polymer processes, such as inverse modeling and science-guided loss which have not been applied previously in such applications.

2 | APPLICATIONS OF HYBRID SGML APPROACH IN BIOPROCESSING AND CHEMICAL ENGINEERING

The integration of science-based models with data-based models has appeared in various fields like fluid mechanics, turbulence modeling, quantum physics, climate science, geology and biological sciences.

This study focuses on applications of hybrid SGML methodologies in bioprocessing and chemical engineering. Among the earliest applications is the direct hybrid modeling involving the integration of first-principles model with data-based neural networks. Psychogios and Unger combine a first-principles model based on prior process knowledge with a neural network, which serves as an estimator of unmeasured process parameters that are difficult to model from first principle. They apply the hybrid model to a fed-batch bioreactor, and the integrated model has better properties than the standard “black-box” neural network models. In particular, the integrated model is able to interpolate and extrapolate much more accurately, is easier to analyze and interpret, and requires significantly fewer training examples. Thompson and Kramer later
demonstrate how to integrate simple process model and first-principles equations to improve the neural network predictions of cell biomass and secondary metabolite in a fed-batch penicillin fermentation reactor when trained on sparse and noisy process data.

Agarwal\textsuperscript{12} develops a general qualitative framework for identifying the possible ways of combining neural networks with the prior knowledge and experience embedded in the available first-principles models, and discusses the direct hybrid modeling with series or parallel configuration to combine the outputs of the science-based model and the ML model. Asprion, et al.\textsuperscript{13} present the term, grey-box modeling, for optimization of chemical processes. They consider the case where a predictive model is missing for a process unit within a larger process flowsheet, and use measured operating data to set up hybrid models combining physical knowledge and process data. They report results of optimization using different gray-box models for process simulators applied to a cumene process. Actually, in a number of earlier studies, Bohlin and his coworkers have explored in details the concepts of gray-box identification for process control and optimization, and Bohlin has summarized the concepts, tools and applications of grey-box hybrid modeling in an excellent book.\textsuperscript{14}

Over the years, we have seen a growing number of applications of hybrid modeling in bioprocessing and chemical engineering as part of the advances in smart manufacturing \textsuperscript{15-17}.

In their 2021 paper, Sansana et al.\textsuperscript{16} discuss mechanistic modeling, data-based modeling, hybrid modeling structures, system identification methodologies, and applications. They classify their hybrid model into parallel, series, surrogate models (which are simpler mathematical representations of more complex models and similar to reduced-order models that we discuss
below), and alternate structures (which include gray-box modeling mentioned above). In the alternate structures, they refer to some applications of semi-mechanistic model structures where the best hybrid model is selected using optimization concepts. They also classify the hybrid models based on some of the chemical industry applications into analysis of model-plant mismatch\(^{17}\), model transfer, feasibility analysis and predictive maintenance, apart from the previous mentioned applications like process control, monitoring and optimization.

Von Stosch et. al.\(^{18}\) have used the term, *hybrid semi-parametric modeling*, in their 2014 review, and have summarized applications in bioprocessing for monitoring, control, optimization, scale-up and model reduction. They emphasize that the application of hybrid semi-parametric techniques does not automatically lead to better results, but that rational knowledge integration has potential to significantly improve model-based process design and operation.

Qin and Chiang\(^ {19} \) review the advances in statistical machine leaning and process data analytics that can provide efficient tools in developing future hybrid models. In a latest paper, Qin et. al.\(^ {20} \) propose a statistical learning procedure integrating with process knowledge to handle a challenging problem of developing a predictive model for process impurity levels from more than 40 process variables in an industrial distillation system. Both studies highlight the power of statistical machine leaning for developing future hybrid process models.

A survey of the literature has shown applications of hybrid modeling in bioprocesses\(^ {21-27} \), chemical and oil and gas process industries\(^ {28-32} \), and polymer processes\(^ {33,34} \) for more accurate and scientifically consistent predictions. This survey has also shown many topical focuses of applications in bioprocessing and chemical engineering, including process control\(^ {35-38} \), design of
In a recent study, Zhou et al.\textsuperscript{46} present a hybrid approach for integrating material and process design that holds much promise in process and product design. Cardillo et. al.\textsuperscript{47} demonstrate the importance of hybrid models in silico production of vaccines to accelerate the manufacturing process. Chopda et. al.\textsuperscript{23} apply integrated process analytical techniques, and modeling and control strategies to enable the continuous manufacturing of monoclonal antibodies. McBride et. al.\textsuperscript{48} classify the hybrid modeling applications in different separation processes in chemical industry, namely, distillation\textsuperscript{49-51}, crystallization\textsuperscript{52,53}, extraction\textsuperscript{54-56}, flotation\textsuperscript{57,58}, filtration\textsuperscript{59,60} and drying\textsuperscript{61,61}. Venkatasubramanian\textsuperscript{63} gives an excellent exposition of the current state of development and applications of artificial intelligence in chemical engineering. The author highlights the intellectual challenges and rewards for developing the conceptual frameworks for hybrid models, mechanism-based causal explanations, domain-specific knowledge discovery engines, and analytical theories of emergence, and presents examples from optimizing material design and process operations.

In an excellent edited volume, Glassey and Stosch\textsuperscript{64} discuss some of the key strengths of hybrid modeling in chemical processes, particularly in the prediction of scientifically consistent results beyond the experimentally tested process conditions, which is crucial for process development, scale-up, control and optimization. They also identify some challenges. For example, incorrect fundamental knowledge in a science-based model could impose bias on predictions, thus the underlying assumptions used in a model are important for analysis. Also, time and accuracy of parameter estimation is critical when deciding on a hybrid modeling strategy. Kahrs and
Marquardt\textsuperscript{65} discuss the approach of simplifying the complex hybrid models into sequence of simpler problems, such as data preprocessing, solving nonlinear equations, parameter estimation and building empirical models using ML.

Herwing and Portner in their latest book showcase the applications of hybrid modeling in digital twins for smart biomanufacturing\textsuperscript{124}.

A recent patent by Chan et al.\textsuperscript{66} presents Aspen Technology’s approach on asset optimization using integrated modeling, optimization and artificial intelligence. In a later white paper, Beck and Munoz\textsuperscript{67} describe Aspen Technology’s current focus on hybrid modeling, combining AI and domain expertise to optimize assts. In particular, based on their application experience in in chemical industries, Aspen Tech have classified hybrid models into three categories: AI-driven, first-principles driven and reduced-order models.\textsuperscript{67} They define an AI-driven hybrid model as an empirical model based on plant or experimental data and use first principles, constraints and domain knowledge to create a more accurate model. Examples of AI-driven models are inferential sensors or online equipment models. They define a first-principles driven hybrid model as an existing first-principles model augmented with data and AI to improve model’s accuracy and predictability, which has seen many applications in bioprocessing and chemical engineering. Lastly, they define a reduced-order model where we use ML to create an empirical data-based model based on data from numerous first-principles process simulation runs, augmented with constraints and domain expertise, in order to build a fit-for-purpose low-dimensional model that can run more quickly. With reduced-order models, we can extend the scale of modeling from units to the plant-wide models that can be deployed faster.
3 | A CLASSIFICATION AND EXPOSITION OF HYBRID SCIENCE-GUIDED MACHINE LEARNING MODELS

As we have seen thus far, the majority of work in hybrid model applications in bioprocessing and chemical engineering focuses on the direct combination of science-based and data-based models. In this article, we portray a broad perspective of the combination of scientific knowledge and data analysis in bioprocessing and chemical engineering as inspired by some of the applications in physics and other areas \(^1,^2\). We categorize these hybrid SGML applications in chemical process industry into two major categories, namely, ML compliments science and science compliments ML, together with their sub-categories based on the methodologies and objectives of hybrid modeling as illustrated in Figure 1. We also classify the applications in bioprocessing and chemical engineering according to our hybrid SGML approach. We present examples in several areas of SGML which have not been explored much thus far, and which have great potential for process improvement and optimization.
We can integrate a first-principles scientific model with a data-based model to improve the model accuracy and consistency. In the following, we introduce the sub-categories of direct hybrid
modeling, inverse modeling approach, reducing model complexity, quantifying uncertainty in the process, and discovering governing equations.

4.1 | Direct Hybrid Modeling

A direct hybrid model combines the output of a first-principles or science-based model with the output of a data-based ML model to improve the prediction accuracy of dependent variables. These combinations could occur in a series configuration, a parallel configuration, or a series-parallel configuration. The direct hybrid modeling strategy is the most widely used approach in hybrid modeling in bioprocessing and chemical engineering.

4.1.1 | Parallel Direct Hybrid Model

Figure 2 illustrates the concept of a parallel direct hybrid model. The science-based model may use the initial conditions and boundary conditions as inputs to make a prediction (Ym), while the ML model uses dynamic time-varying data to make the predictions (Yml). We then combine both outputs directly or with assigned weights (w1, w2) to achieve higher prediction accuracy. We can determine the weights by least squares optimization to minimize the total sum of squares of errors for the difference between the plant and the hybrid model.
Figure 2. Parallel direct hybrid model: Ym and Yml are model predictions, and w1 and w2 are weights.

Galvanauskas et al.\textsuperscript{68} combine directly the data-based neural networks for kinetics and viscosity predictions with the first-principles mass balance ordinary differential equations to optimize the production rate of an industrial penicillin process. Chang et al.\textsuperscript{33} showcase a parallel hybrid model for the dynamic simulation of a batch free-radical polymerization of methyl methacrylate. They combine an approximate rate function for the concentration of the immeasurable initiator concentration with a black-box time-dependent or recurrent neural network model\textsuperscript{9} of the dependent variables representing the mass and moment balance equations of the polymerization reactor. They use the resulting hybrid neural network and rate function (HNNRF)
model to optimize the batch polymerization system, identifying the optimal recipe or operating conditions of the batch polymerization system.

*Hybrid residual modeling or parallel direct hybrid residual model* is a class of the parallel direct hybrid model, where we use a first-principles or science-based process model to quantify the time-dependent prediction error or residual, $Y_{res}$, between plant data $Y(t)$ and science-based model prediction $Y_m$ as a function of process variables 41,69-71. Figure 3 illustrates the concept of the parallel direct hybrid residual model. The correction to the model output taking care of the prediction error or residual of the ML model in the hybrid residual configuration improves the model accuracy over the non-residual configuration of Figure 2.
Figure 3. Parallel direct hybrid residual model: Ym represents model outputs, Res are the time-dependent prediction errors or residues between plant data Y(t) and science-based model outputs Ym, and Ym + Yres are the corrected model outputs

We recommend that the use of hybrid models will generally perform better than standalone ML model for applications like process development. This follows because hybrid models are better at extrapolation, while standalone ML models can be adequate for prediction in a steady running plant.

Tian et al. develop a hybrid residual model for a batch polymerization reactor. First, they develop a simplified process model based on polymerization kinetics, and mass and energy balances to predict the monomer conversion, number-average molecular weight MWN, and weight-average molecular weight MWW. This first-principles process model cannot predict these product quality targets accurately because of its neglect of the gel effect at high monomer conversion and other factors. Next, the authors develop a parallel configuration of three data-based, time-dependent or recurrent neural networks trained by process data to predict the residuals of monomer conversion, MWN and MWW of the simplified first-principles process model. The predicted residuals are added to the predictions from the simplified process model to form the final hybrid model predictions. Because of focus in batch process control is on the end-of-batch product quality targets, the use of time-dependent or recurrent neural networks can usually offer good long-range predictions. Therefore, the resulting hybrid residual model performs well in many batch process control and optimization applications.
Simutis and Lubnert\textsuperscript{36} present another application of the direct hybrid modeling methodology to state estimation for bioprocess control. This work combines a first-principles state Kalman filter based on mass balances of biomass, substrate and product, and an ML-based observation model for quantifying relationship between less established variables and measurements. Recently, Ghosh et. al.\textsuperscript{72-73} apply the parallel hybrid modeling framework in process control, where they combine first-principles models with data-based model built by applying subspace identification for better prediction of batch polymer manufacturing and seed crystallization system. Hanachi et. al.\textsuperscript{74} showcase the application of direct hybrid modeling methodology for predictive maintenance. They combine a physics-based model with a data-based inferential model in an iterative parallel combination for predicting manufacturing tool wear.

### 4.1.2 | Series Direct Hybrid Model

Figure 4 illustrates the series direct hybrid model. The science-based process model serves to augment the data needs of the ML model, while the ML model can help in estimating the parameters of the science-based model. Babanezhad et al.\textsuperscript{75} consider the computational fluid dynamics (CFD) for two-phase flows in chemical reactors, and couple science-based CFD results to a ML model based on an adaptive network-based fuzzy inference system (ANFIS). Once the ML model captures the pattern of the CFD results, they use the hybrid model for process simulation and optimization. Some features calculated from a science-based CFD model can \textit{augment the data} as inputs to a ML model. Chan et. al.\textsuperscript{66} have discussed the advantages of data augmentation by combining simulation and plant data to generate a more accurate data-based analysis. In an application to crude distillation in petroleum refining, Mahalec and Sanchez\textsuperscript{51} use a science-based model to calculate the internal reflux to augment other plant data as inputs to a ML model,
in order to calculate the relationship to the product true boiling point curves for quality analysis.

The data augmentation in series hybrid models is more relevant when some feature measurements are missing in the original data, so we use a first-principles model to calculate those features and then augment those calculated data to the ML model to study the combined multivariate effects. The goal is more towards causal effect of the added science model features and less towards improving accuracy. If we find that some missing feature measurements cause a mismatch between a science-based model and the actual plant, data augmentation may improve the training performance of the hybrid model.

Krippel et. al.\textsuperscript{76} present the hybrid modeling of an ultrafiltration process where they calculate the flux using a ML model to act as an input to a science-based model. Similarly, Luo et al.\textsuperscript{29} develop a hybrid model for a fixed-bed reactor for ethylene oxidation, integrating first-principles reaction kinetics and reactor model with a ML catalyst deactivation model. The latter is developed with support vector regression from operating data, assuming the deactivation property decreasing monotonically with time. With the hybrid model, the prediction error is less than 5% for the prediction of an industrial reactor. The approach can predict the production more accurately and have more reliable extrapolation.

Figure 4 shows that a ML model can also help in estimating the parameters of the science-based model. Mantovanelli et al.\textsuperscript{77} develop a hybrid model for an industrial alcoholic fermentation process, combining first-principles mass and energy balance equations for a series of five fermenters with a data-based, functional link network\textsuperscript{75} to identify the kinetic parameters of the
fermentation reactors trained by plant data. The hybrid model includes the effect of temperature on the fermentation kinetics and show good nonlinear approximation capability. Sharma and Liu\textsuperscript{78} show how to use plant data to estimate kinetic parameters of first-principles models for industrial polyolefin processes. In a recent study Bangi and Kwong \textsuperscript{125} estimate process parameters in hydraulic fracturing process using deep neural network which are then input to a first-principles model. Finally, we note that as illustrated in Figure 4, we can interchangeably use a science-based model or a ML model first in the hybrid framework, depending on if we require to add more features to augment the data set or to estimate model parameters.

Figure 4. Series direct hybrid model

4.1.3 | Series-Parallel or Combined Direct Hybrid Model

Figure 5 shows a combined direct hybrid model, where we use the steady-state data from the plant to estimate the unknown parameters of a science-based process model and then uses the
hybrid residual modeling strategy of Figure 3 for prediction. This series-parallel combination or feedback system can improve model predictions depending on the application.

Figure 5. Combined Direct Hybrid Model: Ym are outputs, Yres are residuals, and Ym+ Tres are corrected outputs

Bhutani et. al.\textsuperscript{79} present a definitive study comparing first-principles, data-based and hybrid models applied to an industrial hydrocracking process. In particular, they couple a first-principles hydrocracking model based on pseudocomponents with data-based neural network models of different configurations of Figures 3 to 5 that quantify the variations in operating conditions, feed quality and catalyst deactivation. The neural network component of the hybrid model either provides updated model parameters in the first-principles process model connected in series, or correct predictions of the first-principles process models. The hybrid models are able to represent
the behavior of an industrial hydrocracking unit to provide accurate and consistent predictions in the presence of process variations and changing operating scenarios.

Song et. al.\textsuperscript{80} also apply the direct hybrid model configurations of Figure 3 to 5 to an industrial hydrocracking process and analyze the strengths and weaknesses of these configurations. They call a model a \textit{mechanism-dominated model} if the accuracy of its outputs is mainly dominated by the available theoretical knowledge used to develop the model; and they also call a model a \textit{data-dominated model} if the accuracy of its outputs is mainly dominated by the quality of the training data and the performance of the resulting data-based model. In particular, they give both the first-principles model and the series direct hybrid model of Figure 4 as examples of mechanism-dominated models, and cite the data-based model, parallel direct residual model of Figure 3, and the combined direct hybrid model of Figure 5 as examples of data-dominated models.

In their work, Song et al.\textsuperscript{80} combine a mechanism-dominated model with a data-dominated model as a hybrid direct model of Figure 2, with the weighting factors for the outputs of two individual models being determined in an adaptive fashion. For their application, Song et al. work with a mechanism-dominated model of an industrial hydrocracking process based on kinetic lumping\textsuperscript{79,80}, and with a data-dominated model based on a self-organizing map (SOM) followed by a convolutional neural network (CNN), with both being trained by simulated process data based on Aspen HYSYS\textsuperscript{80}. They evaluate the performance of the hybrid model for operational optimization of the hydrocracking producing different product scenarios. While this study includes new conceptual development, it needs much simplification of its relatively complex methodology to make it readily applicable by data scientists and practicing engineers.
In a recent study, Chen and Lerapetritou\textsuperscript{17} demonstrate how to use partial correlation analysis from multivariate statistics and mutual information analysis from information theory to identify and improve the plant-model mismatch in using a direct combined hybrid model for a pharmaceutical manufacturing process. As the authors state, implementing this plant-model mismatch strategy requires active excitation of variables online in order to capture the corresponding response data from the plant, which is often difficult to perform in manufacturing plants and in experimental settings, and could benefit from new development in computing and information technology.

Lima et al.\textsuperscript{81} propose a semi-mechanistic model building framework based on selective and localized model extensions. They use a symbolic reformulation of a set of first-principles model equations in order to derive hybrid mechanistic–empirical models. The symbolic reformation permits the addition of empirical elements selectively and locally to the model. They apply the approach to the identification of a non-ideal reactor and to the optimization of the Otto–Williams benchmark reactor.

This combined strategy is generally more useful for the case where the science-based model has unknown parameters. We could use ML to determine these unknown parameters and then apply a hybrid residual ML approach. By doing so, we could improve the model prediction accuracy as well.

\textbf{4.1.4 | An Application of Combined Direct Hybrid Modeling to Polymer Manufacturing}

We apply the combined direct modeling strategy to an industrial polyethylene process for the prediction of melt index. We build a first-principles steady-state model of a Mitsui slurry high-
density polyethylene (HDPE) process by following the methodology and kinetic parameters presented in Supplement 1b of Sharma and Liu\textsuperscript{78}. For this application, it is easier to first estimate the complex multisite Ziegler-Natta polymerization kinetic parameters using steady-state production targets, and then convert the resulting steady-state simulation model based on Aspen Plus to a dynamic simulation model using Aspen Plus Dynamics. The resulting dynamic simulation model has similar independent process variables, including the feed flow and compositions and the reactor operating conditions. For less complex applications, dynamic data could be used for parameter estimation.

Figure 6 compares the predictions of the first-principles dynamic simulation model (in red) with the plant data with grade transitions (in green). We see much deviation between the model predictions and the plant data. We compare the MI values from the model with the plant data and calculate the error residuals. The \textit{root-mean-squares-error (RMSE) values of the model residual is to 1.5} for the actual MI data with a standard deviation of 5.1.
To improve the accuracy of model predictions, we develop a regression model to predict the error residues as a function of independent process variables using a ML method called random forest algorithm\textsuperscript{82} with Python. This leads to a hybrid model that predicts the MI value as a sum of the dynamic simulation model prediction (first-principles-based) and the predicted error residual (data-based) corresponding to a give set of independent process variable values, as illustrated in Figure 5. Figure 6 shows that the hybrid model predictions (\textit{with a RMSE value of 0.21}) match the plant data much better than a first-principles dynamic simulation model alone. We note that a data-based model alone has also a similar accuracy, but it may give scientifically inconsistent results for predictions beyond process operating data which the model uses. Thus,
the hybrid model is not only accurate, but also gives scientifically consistent results beyond current operating range.

4.2 | Inverse Modeling

In inverse modeling, we use the output of a system to infer its corresponding input or independent variables; this is different from the forward modeling where we use the known independent variables to predict the output of the system. Figure 7 illustrates the inverse modeling framework. We see that in the traditional data-based approach, we use process variable data (X) and quality target data (Y) to train and test a ML model. Because the plant does not measure most quality targets continuously, we can apply a science-based process model, developed by first principles and validated by plant data, to predict and augment the quality target data (Y) for given process variable (X).

Figure 7. Inverse Modeling Framework
One of the earliest applications of inverse modeling for chemical process was by Savkovic-Stevanovic et. al.\textsuperscript{83} They use a neural network controller for product composition control of a distillation plant based on the process inverse dynamic model relating the product composition to the reflux flow rate. The results illustrate the feasibility of using neural network for learning nonlinear dynamic model of the distillation column from plant input-output data. Their results also demonstrate the importance to take the time-delay of the plant into account.

Pharmaceutical product design and development typically uses the design of experiments (DOE) and response surface modeling (RSM) for steady-state process modeling, while neglecting the process dynamics and time delays. Tomba et. al.\textsuperscript{84,85} demonstrate how to use the inverse modeling concept to generate process understanding with dynamic process models, quantifying the impact of temporal deviations and production dynamics. Specifically, they perform data-based, latent variable regression model inversion to find the best combination of raw materials and process variables to achieve the desired quality targets. The authors propose to combine design-of-experiments studies with hybrid modeling for process characterization.

Recently, Bayer al.\textsuperscript{86} apply the inverse modeling approach to Escherichia coli fed-batch cultivations, evaluating the impact of three critical process variables. They compare the performance of a hybrid model to a pure data-driven model and the widely adopted RSM of the process endpoints, and show the superior behavior of the hybrid model compared to the pure black-box approaches for process characterization. The inverse modeling methodology makes
the decision-making process in pharmaceutical product development faster, while minimizing the number of experiments and reducing the raw material consumption.

Raccuglia et al.\textsuperscript{87} train the ML learning model using reaction data to predict reaction outcomes for the crystallization of templated vanadium selenites. They demonstrate the use of ML to assist material discovery using data from previously unsuccessful or failed material synthesis experiments. The resulting ML model outperforms traditional human strategies, and successfully predicts conditions for new organically templated, inorganic product formation with a success rate of nearly 90%. Significantly, they show that inverting the machine-learning model reveals new hypotheses regarding the conditions for successful product formation.

There is a growing interest in the inverse approach to material design, in which the desired target properties are used as input to identify the atomic identity, composition and structure (ACS) that exhibit such properties. Liao et al.\textsuperscript{88} present a metaheuristic approach to material design that incorporates the inverse modeling framework.

Vankatasunramanian\textsuperscript{61} also mentions the importance of inverse problem being solved by the application of artificial intelligence in chemical engineering processes.

Note the inverse modeling approach may lead to non-unique solutions which can give a range of predictions of input parameters within the operating range. By adding additional constraints to the input parameters (such as their operating range), we may obtain a unique solution.

4.3 | An Application of Inverse Modeling to Polymer Manufacturing
We illustrate the application of an inverse modeling approach that integrates steady-state and dynamic simulation models of a Mitsui slurry HDPE process, developed from first principles and validated by plant data, with a data-based ML model. The goal is to predict the operating conditions for producing new polymer grades, given the desired product quality targets, such as melt index (MI), polymer density (Rho), polydispersity index (PDI) and polymer production rate (P). The details of the steady-state simulation model are available in Supplement 1b of reference 78.

We first estimate the polymerization kinetic parameters from plant production targets in a steady-state model using Aspen Polymers based on our reported methodology. This results in a validated Aspen Polymers steady-state simulation model. Next, we convert the steady-state model to a dynamic model using Aspen Plus Dynamics. We use the dynamic model to simulate the product quality data for different process operating conditions, which include the data characterizing the polymer grade transitions. Then, we use a Python-based, ensemble machine learning regression model to regress the simulated data, with the simulated product quality data as input, and the process operating conditions (flow rates of all input streams) as the output.

Given the desired quality targets for a new polymer grade, we apply the trained ML model to predict the operating conditions for the new polymer grade.

Figure 8 illustrates that the inverse modeling approach predicts the hydrogen feed flow rate with a high accuracy (low RMSE = 0.9) when compared to actual plant data for a standard deviation of 20). Thus, if we want to produce a new polymer grade given its quality targets, we can predict the operating conditions required to produce that polymer grade using the inverse modeling approach.
4.4 | Reduced-Order Models

Reduced-order models (ROMs) are simplified models that represent a complex process in a computationally inexpensive manner, but also maintain high degree of accuracy of prediction in simulating the process. In bioprocessing and chemical engineering, we can apply the ROM methodology to simulate complex processes and then use ML models to optimize the processes. See Figure 9. We can use ROMs to simulate different scenarios and sensitivities in order to generate process data, which in turn can be combined with ML models to build accurate soft sensors to predict quality variables. This approach helps to make sure that the ML model is trained on process data with multiple variations which is not possible in a steady plant run.
Hence, data-based sensors will be accurate for any future process optimization, scale up etc. and it is also easier to deploy such models online.

![Reduced-order process modeling framework](image)

**Figure 9. Reduced-order process modeling framework**

In one of the earliest applications of ROM, MacGregor et al.\(^9\) apply a PLS (projection to latent squares or partial least squares) ML model of a polyethylene using process data simulated from a process model to develop inferential prediction models for polymer properties. This application involves a high-pressure tabular reactor system producing low-density polyethylene, in which all the fundamental polymer properties are extremely difficult to measure and are usually unavailable, and some on-line measurements such as the temperature profile down the reactor and the solvent flow rate are available on a frequent basis. The dimensionality reduction aspects of PLS facilitates the development of a multivariate statistical control plot for monitoring the operating performance of the reactors.

Model reduction can be achieved through dimensional reduction methods like principal component analysis. Another approach is to apply the residual combination with ML model for a ROM model, or to build a ML-based surrogate model for the full-order model. Reduced-order models have been called *surrogate models* in the context of grey-box modeling techniques where
first-principles models are combined with data-based optimization techniques. Rogers and Lerapetritou\textsuperscript{91,92} propose the use of surrogate models as reduced-order models that approximate the feasibility function for a process in order to evaluate the flexibility and operability of a science-based process model, since it is difficult to directly evaluate the feasibility due to black-box constraints.

In a recent study, Abdullah et. al.\textsuperscript{93} showcase a data-based reduced-order modeling of non-linear processes that have time-scale multiplicity to identify the slow process state variables that can be used in a dynamic model. Agarwal et. al.\textsuperscript{94} use ROM for modeling pressure swing adsorption process where they use a low-dimensional approximation of a dynamic partial differential equation model, which is more computationally efficient. In another study, Kumar et. al.\textsuperscript{45} use a reduced-order steam methane reformer model to optimize furnace temperature distribution. In a recent study, Shafer et al.\textsuperscript{95} use a reduced-dimensional dynamic model for the optimal control of air separation unit. The model combines compartmentalization to reduce the number of differential equations with artificial neural networks to quantify the nonlinear input–output relations within compartments. This work reduces the size of the differential equation system by 90\%, while limiting the additional error in product purities to below 1 ppm compared to a full-order stage-by-stage model.

Kumari et. al.\textsuperscript{126} use data based reduced order methods for computational fluid dynamic modeling applied to a case study of super critical carbon dioxide rare event. They propose a k-nearest neighbor (kNN)-based parametric reduced-order model (PROM) for consequence estimation of rare events to enhance numerical robustness with respect to parameter change. Recently, many operator-theoretic modeling identification and model reduction approaches like
the Koopman operators have been applied to integrate first-principles knowledge into finding relationship among multiple process variables in chemical processes. Koopman operator offers great utility in data-driven analysis and control of nonlinear and high-dimensional systems. Narsingham and Kwon develop a new local Dynamic Mode Decomposition (DMD) method to better capture local dynamics which does temporal clustering of snapshot data using mixed integer nonlinear programming. The developed models are subsequently used to compute approximate solutions to the original high-dimensional system and to design a feedback control system of hydraulic fracturing processes for the computation of optimal pumping schedules.

Our focus on ROM is more towards using the science-based model to simulate process data that can be used by ML models to derive empirical correlations for process optimization. ROM are particularly useful in chemical processes for dynamic optimization of a complex large-scale process.

4.5 | An Application of Reduced-Order Modeling to Polymer Manufacturing

We illustrate the ROM methodology in a HYPOL polypropylene production process. The details of the steady-state simulation model are available in Supplement 1a of reference 78.

The Hypol process is complex with series of reactors, separators and recycle loops. The process has many operating variables, such as feed flow rates of propylene, hydrogen to each reactor, and temperature and pressure in each reactor. It is critical to quantify the effects of operating variables on the polymer quality targets, particularly melt index, in order to design or optimize
the process. To achieve this, we need multivariate process data which are not usually available in a steady running plant. Hence, we use the ROM methodology.

We model the HYPOL polypropylene production process following the methodology of Sharma and Liu\textsuperscript{78} and then run multiple steady-state simulations to generate multivariate data with varying operating variables and the corresponding melt index predictions. We use a random forest ML model\textsuperscript{89} to train the simulated data to predict the melt index as a function of the process variables and also understand the causality of important features affecting the polymer quality. The empirical ML model can serve as an approximate quality sensor. We can use it to predict the melt index at varying process variables. See Figure 10a.

Figure 10. (a) Melt Index prediction at varying process variables
The ML model also decides the relative importance of different operating variables in reducing the mean decrease in “node impurity”, which is a measure of how much each operating variable feature reduces the variance in the model. Figure 10b illustrates that the ROM calculates the important features like hydrogen flow rate (H24) and the temperature to the fourth reactor (R4T) as the most important variables affecting the melt index, which can then be used to find the optimum conditions to produce polymer of a specified melt index value and to improve the process design for a new process.

Figure 10. (b) Feature importance for melt index prediction: RxT and RxP refer to the temperature and pressure of reactor x; C3x and H2x represent the mass flow rates of propylene and hydrogen to reactor x; C24 is the mass flow rate of ethylene to reactor 4; and CAT is the catalysts mass flow rate.

4.6 | Hybrid SGML Modeling for Uncertainty Quantification
A science-based model can produce results with some uncertainties which can be quantified by ML-based techniques. The uncertainties in science-based models arise from uncertainty in model parameters, and boundary and initial conditions. In some cases, the model bias and assumptions can be a source of uncertainty as well. We can use the predictions from a calibrated model to quantify uncertainties. Data-based ML models like Gaussian process, neural networks etc. are used to help build a surrogate model that defines a relation between model inputs and outputs which can then be used to quantify the uncertainty.

Because of uncertainty in process inputs and process states in a chemical process model, the uncertainty propagates to the process outputs as well. The uncertainty in a science-based model due to any of the parameters or any of the prior knowledge can be used by a ML model to quantify uncertainty in a chemical process as shown in Figure 11.

Figure 11. Uncertainty quantification of melt Index prediction of a slurry HDPE process
This surrogate data-based ML modeling reduces the computational expense of Monte Carlo methods, which are traditionally used for uncertainty quantification (UQ)\textsuperscript{96}.

Because of uncertainty in process inputs and process states in a chemical process model, the uncertainty propagates to the process outputs as well. Duong et. al.\textsuperscript{97} uses UQ for process design and sensitivity analysis of complex chemical processes using the polynomial chaos theory. Fenila et. al.\textsuperscript{98} utilize UQ for electrochemical synthesis, where they calculate simulation uncertainties and global parameter sensitivities for the hybrid model. UQ has also been applied to understand complex reaction mechanisms. Proppe et. al.\textsuperscript{99} showcase kinetic simulations in discrete-time space considering the uncertainty in free energy and detecting regions of uncertainty in reaction networks. UQ techniques are popular in the field of catalysis and material science as they are used to quantify the uncertainty of models based on density functional theory\textsuperscript{100,101}. In another study, Boukouval and Lerapetritou\textsuperscript{102} demonstrate the feasibility analysis of a science-based process model over a multivariate factor space. They use a stochastic data-based model for feasibility evaluation, referred to as Kriging and develop an adaptive sampling strategy to minimize sampling cost while maintaining feasibility.

4.7 | An Application of SGML Modeling to Uncertainty Quantification in Polymer Manufacturing

We quantify the uncertainty of the chemical process model in predicting the melt index for the industrial HDPE process described in Section 2.1.4. This uncertainty in prediction may result from the estimated kinetic parameters of the process, which propagates to the quality output as well. We simulate the data using the chemical process model and calculate the prediction intervals
using a gradient boosting ML model\textsuperscript{89}. In this case, we use the concept of prediction intervals to determine the range of model prediction. We use the quantile regression loss with gradient boosting model to predict the prediction intervals\textsuperscript{103}. We define a lower and an upper quantile according to the desired prediction interval. Figure 12 illustrates the uncertainty in the prediction of melt index given by the range of the 90\% prediction interval which implies that there is 90\% likelihood that the ML model prediction will lie in the given range. The resulting RMSE value lies within 1.2 to 1.5, with the standard deviation of melt index data equals 5.1. In the figure, we see that the prediction interval is the area between the two black lines represented by the upper quantile (95\textsuperscript{th} percentile) and the lower quantile (5\textsuperscript{th} percentile). From the figure, we see a larger prediction interval that means a higher uncertainty in prediction for time less than 100 hours compared to the later stage because of a more appreciable change in MI in that interval. Thus, uncertainty quantification (UQ) helps in making better process decisions after knowing the error estimate of the model.
Figure 12. Uncertainty quantification of melt Index prediction of a slurry HDPE process

4.8 | Hybrid SGML Modeling to Aid in Discovering Scientific Laws Using ML

One way in which ML can help science-based modeling is by discovering new scientific laws which governs the system. There is a growing application of ML in physics to rediscover or discover physical laws mainly by data-driven discovery of partial differential equations. ML can be used to develop an empirical correlation which can be used as a scientific law in a science-based model, or ML can be used to solve the partial differential equation defining scientific laws as illustrated in Figure 13.
Rudy et. al. showcase the discovery of physical laws like the Navier-Stokes equation and the reaction-diffusion equation in chemical processes by a sparse regression method governing the PDE by using a system of time series measurements. Langley et. al. present the applications of ML in rediscovering some of the chemistry laws, such as the law of definite proportions, law of combining volumes, determination of atomic weights and many others.

Another important application of ML is to discover some of the thermodynamic laws which can be useful in defining the phase equilibrium and critical for an accurate science-based process model. Nentwich et. al. use data-based mixed adaptive sampling strategy to calculate the phase composition, instead of the complex equation-of-state models. Thus, ML application can have promising use in discovering more accurate physical and chemistry laws that govern the chemical process. This methodology can be used to obtain the functional form of scientific laws as well as the estimation of the parameters of existing laws. Brunton et al. demonstrate a novel
framework to discovering governing equations underlying a dynamic system simply from data measurements, leveraging advances in sparsity techniques and machine learning. These scientific laws calculated by ML-based models can then be utilized in first-principles model to improve accuracy as well as reduce model complexity.

5 | SCIENCE COMPLIMENTS ML

Referring to Figure 1, we can also improve ML models using scientific knowledge. We can improve the generalization or extrapolation capability and reduce the scientific inconsistency of ML models by using scientific knowledge in designing the ML models. The scientific knowledge can also help in improving the architecture of the data-based ML model or the learning process of the ML model and even with the final post-processing of the ML model results.

5.1 | Science-Guided Design

In science-guided design, we choose the model architecture based on scientific knowledge. For a neural network, we can decide the intermediate variables expressed as hidden layers based on scientific knowledge of the system. This helps in improving the interpretative ability of the models. Figure 14 illustrates a neural network model whose architecture like the number of neurons, hidden layers, activation layers etc. can be decided by prior scientific knowledge.
In a bioprocess application, Rodriguez-Granrose et al.\textsuperscript{39} use the design of experiments (DOE) to create and evaluate a neural network architecture. They use DOE to evaluate activation functions and neurons on each layer to optimize the neural network. In their recent study, Wang et al.\textsuperscript{107} design their theory-infused neural networks based on adsorption energy principles for interpretable reactivity prediction. The use of the novel neural differential equation\textsuperscript{108} to solve a first-principles dynamic system represents a hybrid SGML approach, where the architecture of ML model is influenced by the system and finds applications in continuous time series models and scalable normalizing flows. The derivative of hidden state is parameterized using a neural network and the output of the network is computed using a differential equation solver. In a recent study, Jaeger et al.\textsuperscript{109} use the neural differential equation to predict the dynamic behavior of electro-dialysis fouling under varying process conditions. In a recent application of this theme in chemical process for model predictive control, Wu et al.\textsuperscript{110} use prior process knowledge to design the recurrent neural network (RNN) structure\textsuperscript{9}. They showcase a methodology to design the RNN structure using prior scientific knowledge of the system and also
employ weight constraints in the optimization problem of the RNN training process. Reis et. al.\textsuperscript{111} discuss the concept of incorporation of process-specific structure to improve process fault detection and diagnosis.

Fuzzy artificial neural networks (ANN) is a class of neural networks which utilize prior scientific knowledge of the system to formulate rules mapped on to the structure of the ANN\textsuperscript{9,112}. The weights of the ANN connecting the process input to output can be connected to physical process variables\textsuperscript{64}. Apart from making the models more scientifically consistent with prior knowledge, they also reduce computational complexity and provides interpretable results. The use of prior knowledge also makes them suitable for extrapolation. Fuzzy ANN have been particularly useful for applications in process control\textsuperscript{113}. Simutis et. al. use fuzzy ANN system for industrial bioprocess monitoring and control\textsuperscript{114-115}. They also illustrate the application of fuzzy ANN process control expert to perform appropriate control actions based on process trends for bioprocess optimization and control\textsuperscript{116}.

Sparse Identification of Nonlinear Dynamics (SINDy) is another data-based modeling method that utilizes scientific knowledge for improving the model performance with the algorithms\textsuperscript{128}. Bhadiraju et. al.\textsuperscript{129} have used the SINDy algorithm to identify the Non Linear Dynamics of a chemical process system(CSTR). They used sparse regression in combination with feature selection to identify accurate models in an adaptive model identification methodology which requires much less than data that current methods. In a similar study Bhadiraju et. al.\textsuperscript{130} have a modified adaptive SINDy approach that is helpful in cases of plant model mismatch and does not require retraining and hence computationally less expensive.
5.2 | Science-Guided Learning

Here, we make use of the scientific principles to improve the scientific consistency of data-based models by modifying the machine learning process. We do this by modifying the loss function, constraints and even the initialization of ML models based on scientific laws. Specifically, in order to make the ML models physically consistent we make the loss function of neural network model incorporate physical constraints \(^2\). A loss function in ML measures how far an estimated value is from its true value. A loss function maps decisions to their associated costs. Loss functions are not fixed, they change depending on the task in hand and the goal to be met. We can define a loss function (based on the mean squared error, MSE) of the ML model (Loss\(_M\)) for regression to calculate the difference between the true value (\(Y_{true}\)) and the model predicted value (\(Y_{pred}\)). Likewise, we can define a loss function for a science-based model (Loss\(_SC\)), which is a function of the model predicted value (\(Y_{pred}\)) consistent with science-based loss. We include a weighting factor \(\lambda\) to express the relative importance of both loss terms. We write the overall loss function (Loss) as:

\[
Loss = Loss_M(Y_{true} - Y_{pred}) + \lambda Loss_{SC}(Y_{pred}) \quad (1)
\]

Figure 15 illustrates the concept of science-guided loss function.

A science-guided initialization helps in deriving an initial choice of parameters before a model is trained so that it improves model training and also prevents from reaching a local minimum, which is the concept of transfer learning. Thus, we can use the data from a science-based model to pre-train a ML model based on this concept of initialization \(^1,2,7\). This concept has been utilized in chemical process model in the form of process similarity and developing new process models.
through migration. In particular, Lu et. al. introduce the concept of process similarity, and classify it into attribute-based and model-based similarities. They present a model migration strategy to develop a new process model by taking advantage of an existing base model, and process attribute information. Adapting existing process models can allow using fewer experiments for the development of a new process model, resulting in a saving of time, cost, and effort. They apply the concept to predict the melt-flow-length in injection molding and obtain satisfactory results.

Figure 15. Science-guided loss function representation
In another study on the similar concept, Yan et al.\textsuperscript{118} use a Bayesian method for migrating a ML Gaussian process regression model. They showcased an approach of an iterative model migration and process optimization for an epoxy catalytic reaction process.

Recently, Kumar et. al.\textsuperscript{119} try to optimize the Non-Newtonian fluid flow for industrial processes like crude oil transportation using a physics-based loss function for the shear stress calculation for more accurate flow predictions. In another study on the similar principle, Pun et. al.\textsuperscript{120} apply physics-informed neural networks for more accurate and transferable atomistic modeling of materials.

5.3 | An illustrative Example of Science-Guided Learning

We showcase the application of the science-guided loss function in the slurry HDPE process for the industrial HDPE process described in Section 2.1.4. The goal is to predict the melt index of the polymer. The plant only measures the polymer melt index as the quality output, but we also want the data-based ML model to predict the scientifically consistent polymer density values.

We express polymer density as a function of the melt index using some empirical correlations and modify the loss function (based on the mean squared error, MSE) to consider density as well. See Eq. (2) below. We then train a deep learning neural network model to predict the melt index of the polymer. Figure 16 illustrates that the SGML hybrid model calculates the melt index, resulting in a RMSE of the melt Index that is slightly higher (RMSE = 0.8) (standard deviation of data= 5) compared to a standalone ML model. In addition to predicting the melt index values, the hybrid SGML model is simultaneously predicting the polymer density correctly within the
physically consistent range of 0.94-0.97 g/c. By contrast, the density estimates by the ML model alone result in density values greater than 1, which is physically inconsistent.

\[ \text{Loss} = \text{Loss}_M(MI_{true} - MI_{pred}) + \lambda \text{Loss}_{SC}(\rho(MI_{true}) - \rho(MI_{pred})) \] (2)

Figure 16. Melt index and polymer density prediction with a ML model with a science-guided loss function

5.4 | Science-Guided Refinement

By science-guided refinement, we mean the post-processing of ML model results based on scientific principles. This post-processing of results of the ML model using science-based models can be useful to the design and prediction of material structure\textsuperscript{113}. Thus, the discovery of materials forms the basis of chemical process development from which the manufacturing
process of any compound can be designed. This is different than the serial direct hybrid model discussed in Section 3.1.2. In particular, we use the science-based model to merely test the scientific consistency of the ML model results. Hautier et. al.\textsuperscript{114} use first-s models based on density functional model to refine the results of probabilistic ML models to discover ternary oxides. Figure 17 illustrates the science-guided refinement framework.

![Science-guided refinement framework](image)

Figure 17. Science-guided refinement framework

Another application for science-guided learning is for data generation. ML techniques like generalized adversarial networks (GAN) are useful for generating data in an unsupervised learning. GANs do have a problem of high sample complexity\textsuperscript{2} which can be reduced by incorporating some science-based constraints and prior knowledge. Cang et al.\textsuperscript{115} apply ML models to predict the structure and properties of materials and use the results of the ab initio calculations to refine the ML model results. They generate more imaging data for property prediction using a convolution neural network and introduce a morphology constraint form scientific principles, while training of the generative models so that it improves the prediction of the structure- property model.
Thus, some of these methodologies of having science complimenting ML have much potential for future applications to bioprocessing and chemical engineering.

6 | CHALLENGES AND OPPORTUNITIES OF HYBRID SGML APPROACH FOR MODELING CHEMICAL PROCESSES

Along with all the merits of using the SGML methodology there are challenges as well. Incorrect fundamental knowledge and the assumptions of the science-based first-principles model will lead to inaccurate hybrid model, so it is important for the scientific model to be very accurate. Lack of engineer/scientists having expertise in both domain knowledge and machine learning. Computation infeasibility in certain modeling approaches like inverse modeling. Data cleaning, preprocessing, feature engineering maybe difficult in certain cases but may be imperative in science-based model parameter estimation hence in these cases the hybrid models may increase the complexity compared to a stand-alone ML models like Neural Network which may not require feature engineering. Model predictions must not only be accurate but also with lower uncertainty which may be difficult for certain hybrid model methods.

There is a lot of scope of using hybrid SGML methodologies in chemical process modeling, summarizing here some of the opportunities and areas where they can be beneficial. As we have seen Hybrid SGML models are useful for extrapolation and predicting beyond operating range, hence it will be particularly useful for processes development. Process fault diagnosis and anomaly detection is one such area where data-based methods have been used extensively, thus there is opportunity to combine scientific knowledge as well to make the anomaly detection process more scientifically consistent.
Table 1 summarizes all the hybrid SGML models and their requirements, advantages, limitations and potential applications.

Table 1. Summary of hybrid SGML approach

| Hybrid SGML Modeling | Science-based model /knowledge | ML model | Advantages | Limitations | Potential applications |
|----------------------|--------------------------------|----------|------------|-------------|------------------------|
| ML compliments science (base model: science-based) | | | | | |
| Direct Hybrid modeling | Series | Science-based model (SBM) | Regression | Extrapolation; Parameter estimation; data augmentation | Limited by data for parameter estimation, Interpolation, Scientific Knowledge dominated | Kinetic estimation 77, 78 Soft sensor 86 Process optimization 75 Process design 51 Process modeling 29, 76 |
| | | | | | |
| Parallel | SBM | Regression | Improved accuracy of prediction, Interpolation | Scientific consistency depends on SBM, Extrapolation; Data-dominated | Process scale-up 41 Process optimization 33, 68 Process control 36, 71-72 Soft sensor 33 Process monitoring 68 Predictive maintenance 66, 74 |
| Series-parallel | SBM | Regression | Higher Accuracy, Interpolation | Increased model complexity, Data-dominated | Process optimization 79, 80 Process monitoring and control 58 Plant-model mismatch 17 |
| Inverse modeling | SBM | Probabilistic, Regression | Computationally cheaper inverse problem | Lower generality of the model | Product design and development 84-85 Polymer grade change, Material design 87 |
| Reduced-order models | SBM | Regression | Fast online deployment, reduce model complexity | Higher Bias, limited by SBM accuracy | Process optimization at plant scale 45, 94 Dynamic modeling 93, 95 Soft sensor 90 Feasibility analysis 91-92 |
7 | CONCLUSION

We present a broad perspective of hybrid modeling with a science-guided machine learning (SGML) approach and its application in bioprocessing and chemical engineering. We give a detailed review and exposition of the hybrid SGML modeling approach and its applications, and classify the approach into two categories. The first refers to the case where a data-based ML model compliments and makes the first-principles science-based model more accurate in prediction, and the second corresponds to the case where scientific knowledge helps make the ML model more scientifically consistent. We point out some of the areas of SGML which have not been explored much in chemical process modeling and have potential for further use like in the
areas where Science can help improve the data-based model by improving the model design, learning and refinement. We also illustrate some of these applications of the hybrid SGML methodologies for industrial polymer/chemical process improvement.

Thus, based on our review, we recommend that the use of hybrid models will perform better than standalone ML for applications like process development, since they are better at extrapolation while standalone ML models which can be adequate for prediction in a steady running plant.

**Acknowledgement**

We gratefully acknowledge Aspen Technology, Inc., for their support of the Center of Excellence in Process System Engineering in the Department of Chemical Engineering at Virginia Tech since 2002. We thank Mr. Antonio Pietri, CEO, Mr. Willie K. Chan, Chief Technology Officer, Dr. Steven Qi, Vice President for Training and Customer Support, and Mr. Daniel Clenzi, Director of University Programs of Aspen Technology, Inc. for their strong support. We would like to thank Professor Anuj Karpatne for sharing his knowledge of theory-guided data science through his course, physics-guided machine learning, at Virginia Tech.
References

1. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, Shekhar S, Samatova N, Kumar V. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on Knowledge and Data Engineering. 2017;29(10):2318-31.

2. Willard J, Jia X, Xu S, Steinbach M, Kumar V. Integrating physics-based modeling with machine learning: A survey. arXiv preprint arXiv:2003.04919. 2020;1(1):1-34.

3. Muralidhar N, Bu J, Cao Z, He L, Ramakrishnan N, Tafti D, Karpatne A. PhyNet: Physics Guided Neural Networks for Particle Drag Force Prediction in Assembly. In Proceedings of the 2020 SIAM International Conference on Data Mining 2020 (pp. 559-567). Society for Industrial and Applied Mathematics.

4. Bode M, Gauding M, Lian Z, Denker D, Davidovic M, Kleinheinz K, Jitsev J, Pitsch H. Using physics-informed super-resolution generative adversarial networks for subgrid modeling in turbulent reactive flows. arXiv preprint arXiv:1911.11380. 2019.

5. Schütt KT, Kindermans PJ, Sauceda HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. arXiv preprint arXiv:1706.08566. 2017.

6. Faghmous JH, Kumar V. A big data guide to understanding climate change: The case for theory-guided data science. Big Data. 2014;2(3):155-63.

7. Karpatne A, Watkins W, Read J, Kumar V. Physics-guided neural networks (pgnn): An application in lake temperature modeling. arXiv preprint arXiv:1710.11431. 2017.
8. Lee D, Jayaraman A, Kwon JS. Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling. PLoS Computational Biology. 2020;16(12):e1008472.

9. Baughman DR, Liu YA. Neural networks in bioprocessing and chemical engineering. Academic Press; 1995.

10. Psichogios DC, Ungar LH. A hybrid neural network-first principles approach to process modeling. AIChE Journal. 1992;38(10):1499-511.

11. Thompson ML, Kramer, MA., Modeling Chemical Process Using Prior Knowledge and Neural Networks. AIChE Journal, 1994; 40(8),1328-40.

12. Agarwal M. Combining neural and conventional paradigms for modelling, prediction and control. International Journal of Systems Science. 1997;28(1):65-81.

13. Asprion N, Böttcher R, Pack R, Stavrou ME, Höller J, Schwientek J, Bortz M. Gray-Box Modeling for the Optimization of Chemical Processes. Chemie Ingenieur Technik. 2019;91(3):305-13.

14. Bohlin TP. Practical grey-box process identification: theory and applications. Springer Science & Business Media; 2006.

15. Yang S, Navarathna P, Ghosh S, Bequette BW. Hybrid modeling in the era of smart manufacturing. Computers & Chemical Engineering. 2020;140:106-874.

16. Sansana J, Joswiak MN, Castillo I, Wang Z, Rendall R, Chiang LH, Reis MS. Recent trends on hybrid modeling for Industry 4.0. Computers & Chemical Engineering. 2021;11:107-365.
17. Chen Y, Ierapetritou M. A framework of hybrid model development with identification of plant-model mismatch. AIChE Journal. 2020;66(10):e16996.

18. Von Stosch M, Oliveira R, Peres J, de Azevedo SF. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Computers & Chemical Engineering. 2014;60:86-101.

19. Qin SJ, Chiang LH. Advances and opportunities in machine learning for process data analytics. Computers & Chemical Engineering. 2019;126:465-73.

20. Qin SJ, Guo S, Li Z, Chiang LH, Castillo I, Braun B, Wang Z. Integration of process knowledge and statistical learning for the Dow data challenge problem. Computers & Chemical Engineering. 2021; 153:107451.

21. O’Brien CM, Zhang Q, Daoutidis P, Hu WS. A hybrid mechanistic-empirical model for in silico mammalian cell bioprocess simulation. Metabolic Engineering. 2021; 66:31-40.

22. Pinto J, de Azevedo CR, Oliveira R, von Stosch M. A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development. Bioprocess and Biosystems Engineering. 2019 ;42(11):1853-65.

23. Chopda V, Gyorgypal A, Yang O, Singh R, Ramachandran R, Zhang H, Tsilomelekis G, Chundawat SP, Ierapetritou MG. Recent advances in integrated process analytical techniques, modeling, and control strategies to enable continuous biomanufacturing of monoclonal antibodies. Journal of Chemical Technology & Biotechnology. 2021; http://doi.org/10.1002/jctb.6765
24. Zhang D, Del Rio-Chanona EA, Petsagkourakis P, Wagner J. Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization. Biotechnology and Bioengineering. 2019;116(11):2919-30.

25. Al-Yemni M, Yang RY. Hybrid neural-networks modeling of an enzymatic membrane reactor. Journal of the Chinese Institute of Engineers. 2005;28(7):1061-7.

26. Chabbi C, Taibi M, Khier B. Neural and hybrid neural modeling of a yeast fermentation process. International Journal of Computational Cognition. 2008;6(3).

27. Corazza FC, Calsavara LP, Moraes FF, Zanin GM, Neitzel I. Determination of inhibition in the enzymatic hydrolysis of cellobiose using hybrid neural modeling. Brazilian Journal of Chemical Engineering. 2005;22(1):19-29.

28. Azarpour A, Borhani TN, Alwi SR, Manan ZA, Mutalib MI. A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors. Chemical Engineering Research and Design. 2017;117:149-67.

29. Luo N, Du W, Ye Z, Qian F. Development of a hybrid model for industrial ethylene oxide reactor. Industrial & Engineering Chemistry Research. 2012;51(19):6926-32.

30. Zahedi G, Lohi A, Mahdi KA. Hybrid modeling of ethylene to ethylene oxide heterogeneous reactor. Fuel Processing Technology. 2011;92(9):1725-32.

31. Simon LL, Fischer U, Hungerbühler K. Modeling of a three-phase industrial batch reactor using a hybrid first-principles neural-network model. Industrial & engineering chemistry research. 2006;45(21):7336-43.
32. Bellos GD, Kallinikos LE, Gounaris CE, Papayannakos NG. Modelling of the performance of industrial HDS reactors using a hybrid neural network approach. Chemical Engineering and Processing: Process Intensification. 2005;44(5):505-15.

33. Chang JS, Lu SC, Chiu YL. Dynamic modeling of batch polymerization reactors via the hybrid neural-network rate-function approach. Chemical Engineering Journal. 2007;130(1):19-28.

34. Hinchliffe M, Montague G, Willis M, Burke A. Hybrid approach to modeling an industrial polyethylene process. AIChE Journal. 2003;49(12):3127-37.

35. Madar J, Abonyi J, Szeifert F. Feedback linearizing control using hybrid neural networks identified by sensitivity approach. Engineering Applications of Artificial Intelligence. 2005;18(3):343-51.

36. Simutis R, Lübbert A. Hybrid approach to state estimation for bioprocess control. Bioengineering. 2017;4(1):21-9.

37. Cubillos F, Callejas H, Lima EL, Vega MP. Adaptive control using a hybrid-neural model: application to a polymerisation reactor. Brazilian Journal of Chemical Engineering. 2001;18(1):113-20.

38. Doyle III FJ, Harrison CA, Crowley TJ. Hybrid model-based approach to batch-to-batch control of particle size distribution in emulsion polymerization. Computers & Chemical Engineering. 2003;27(8-9):1153-63.

39. Rodriguez-Granrose D, Jones A, Loftus H, Tandeski T, Heaton W, Foley KT, Silverman L. Design of experiment (DOE) applied to artificial neural network architecture enables
rapid bioprocess improvement. Bioprocess and Biosystems Engineering. 2021;44(6):1301-8.

40. Brendel M, Marquardt W. Experimental design for the identification of hybrid reaction models from transient data. Chemical Engineering Journal. 2008;141(1-3):264-77.

41. Bollas GM, Papadokonstadakis S, Michalopoulos J, Arampatzis G, Lappas AA, Vasalos IA, Lygeros A. Using hybrid neural networks in scaling up an FCC model from a pilot plant to an industrial unit. Chemical Engineering and Processing: Process Intensification. 2003;42(8-9):697-713.

42. Von Stosch M, Hamelink JM, Oliveira R. Hybrid modeling as a QbD/PAT tool in process development: an industrial E. coli case study. Bioprocess and Biosystems Engineering. 2016;39(5):773-84.

43. Iwama R, Kaneko H. Design of ethylene oxide production process based on adaptive design of experiments and Bayesian optimization. Journal of Advanced Manufacturing and Processing. 2021;3(3):e10085.

44. Zhang S, Wang F, He D, Jia R. Batch-to-batch control of particle size distribution in cobalt oxalate synthesis process based on hybrid model. Powder technology. 2012; 224:253-9.

45. Kumar A, Baldea M, Edgar TF. Real-time optimization of an industrial steam-methane reformer under distributed sensing. Control Engineering Practice. 2016;54:140-53.

46. Zhou T, Gani R, Sundmacher K. Hybrid data-driven and mechanistic modeling approaches for multiscale material and process design. Engineering. 2021;7(9):1231-38.
47. Cardillo AG, Castellanos MM, Desaily B, Dessoy S, Mariti M, Portela RM, Scutella B, von Stosch M, Tomba E, Varsakelis C. Towards in silico process modeling for vaccines. Trends in Biotechnology. 2021;39(11):1120-30.

48. McBride K, Sanchez Medina EI, Sundmacher K. Hybrid Semi-parametric Modeling in Separation Processes: A Review. Chemie Ingenieur Technik. 2020;92(7):842-55.

49. Safavi AA, Nooraii A, Romagnoli JA. A hybrid model formulation for a distillation column and the on-line optimisation study. Journal of Process Control. 1999;9(2):125-34.

50. Mahalec V. Hybrid Modeling of Petrochemical Processes. In Hybrid Modeling in Process Industries, 2018 : 129-165. CRC Press.

51. Mahalec V, Sanchez Y. Inferential monitoring and optimization of crude separation units via hybrid models. Computers & Chemical Engineering. 2012;45:15-26.

52. Peroni CV, Parisi M, Chianese A. Hybrid modelling and self-learning system for dextrose crystallization process. Chemical Engineering Research and Design. 2010 ;88(12):1653-8.

53. Xiong Z, Zhang J. A batch-to-batch iterative optimal control strategy based on recurrent neural network models. Journal of Process Control. 2005;15(1):11-21.

54. Zhang S, Chu F, Deng G, Wang F. Soft sensor model development for cobalt oxalate synthesis process based on adaptive Gaussian mixture regression. IEEE Access. 2019 ;7:118749-63.

55. Nentwich C, Winz J, Engell S. Surrogate modeling of fugacity coefficients using adaptive sampling. Industrial & Engineering Chemistry Research. 2019;58(40):18703-16.
56. Kunde C, Keßler T, Linke S, McBride K, Sundmacher K, Kienle A. Surrogate Modeling for Liquid–Liquid Equilibria Using a Parameterization of the Binodal Curve. Processes. 2019;7(10):753.

57. Côte M, Grandjean BP, Lessard P, Thibault J. Dynamic modelling of the activated sludge process: improving prediction using neural networks. Water Research. 1995;29(4):995-1004.

58. Hwang TM, Oh H, Choi YJ, Nam SH, Lee S, Choung YK. Development of a statistical and mathematical hybrid model to predict membrane fouling and performance. Desalination. 2009;247(1-3):210-21.

59. Piron E, Latrille E, Rene F. Application of artificial neural networks for crossflow microfiltration modelling: “black-box” and semi-physical approaches. Computers & Chemical Engineering. 1997;21(9):1021-30.

60. Piron E, Latrille E, Rene F. Application of artificial neural networks for crossflow microfiltration modelling: “black-box” and semi-physical approaches. Computers & Chemical Engineering. 1997;21(9):1021-30.

61. Zbiciński I, Strumiłło P, Kamiński W. Hybrid neural model of thermal drying in a fluidized bed. Computers & chemical engineering. 1996; 20: S695-700.

62. Cubillos FA, Alvarez PL, Pinto JC, Lima EL. Hybrid-neural modeling for particulate solid drying processes. Powder Technology. 1996;87(2):153-60.

63. Venkatasubramanian V. The promise of artificial intelligence in chemical engineering: Is it here, finally. AIChE Journal. 2019;65(2):466-78.

64. Glassey J, Von Stosch M, editors. Hybrid modeling in process industries. CRC Press; 2018.
65. Kahrs O, Marquardt W. Incremental identification of hybrid process models. Computers & Chemical Engineering. 2008;32(4-5):694-705.

66. Chan WK, Fischer B, Varvarezos D, Rao A, Zhao H, inventors; Aspen Technology Inc, assignee. Asset Optimization Using Integrated Modeling, Optimization, and Artificial Intelligence. United States patent application US 16/434,793. 2020.

67. Beck, R.; Munoz, G. Hybrid Modeling: AI and Domain Expertise Combine to Optimize Assets., 2020. https://www.aspentech.com/en/resources/white-papers/hybrid-modeling-ai-and-domain-expertise-combine-to-optimize-assets/?src=blog-global-wpt

68. Galvanauskas V, Simutis R, Lübbert A. Hybrid process models for process optimization, monitoring and control. Bioprocess and Biosystems Engineering. 2004;26(6):393-400.

69. Tian Y, Zhang J, Morris J. Modeling and optimal control of a batch polymerization reactor using a hybrid stacked recurrent neural network model. Industrial & Engineering Chemistry Research. 2001;40(21):4525-35.

70. Su HT, McAvoy TJ, Werbos P. Long-term predictions of chemical processes using recurrent neural networks: A parallel training approach. Industrial & Engineering Chemistry Research. 1992;31(5):1338-52.

71. Hermanto MW, Braatz RD, Chiu MS. Integrated batch-to-batch and nonlinear model predictive control for polymorphic transformation in pharmaceutical crystallization. AIChE Journal. 2011;57(4):1008-19.

72. Ghosh D, Hermonat E, Mhaskar P, Snowling S, Goel R. Hybrid modeling approach integrating first-principles models with subspace identification. Industrial & Engineering Chemistry Research. 2019;58(30):13533-43.
73. Ghosh D, Moreira J, Mhaskar P. Model Predictive Control Embedding a Parallel Hybrid Modeling Strategy. Industrial & Engineering Chemistry Research. 2021;60(6):2547-62.

74. Hanachi H, Yu W, Kim IY, Liu J, Mechefske CK. Hybrid data-driven physics-based model fusion framework for tool wear prediction. The International Journal of Advanced Manufacturing Technology. 2019;101(9):2861-72.

75. Babanezhad M, Behroyan I, Nakhjiri AT, Marjani A, Rezakazemi M, Shirazian S. High-performance hybrid modeling chemical reactors using differential evolution based fuzzy inference system. Scientific Reports. 2020;10(1):1-1.

76. Kripp M, Dürauer A, Duerkop M. Hybrid modeling of cross-flow filtration: Predicting the flux evolution and duration of ultrafiltration processes. Separation and Purification Technology. 2020; 248:117064.

77. Mantovanelli IC, Rivera EC, Da Costa AC, Maciel Filho R. Hybrid neural network model of an industrial ethanol fermentation process considering the effect of temperature. Applied biochemistry and biotechnology. 2007;137(1):817-33.

78. Sharma N, Liu YA. 110th Anniversary: An Effective Methodology for Kinetic Parameter Estimation for Modeling Commercial Polyolefin Processes from Plant Data Using Efficient Simulation Software Tools. Industrial & Engineering Chemistry Research. 2019;58(31):14209-26.

79. Bhutani N, Rangaiah GP, Ray AK. First-principles, data-based, and hybrid modeling and optimization of an industrial hydrocracking unit. Industrial & Engineering Chemistry Research. 2006;45(23):7807-16.
80. Song W, Du W, Fan C, Yang M, Qian F. Adaptive Weighted Hybrid Modeling of Hydrocracking Process and Its Operational Optimization. Industrial & Engineering Chemistry Research. 2021;60(9):3617-32.

81. Lima PV, Saraiva PM, Group GP. A semi-mechanistic model building framework based on selective and localized model extensions. Computers & Chemical Engineering. 2007;31(4):361-73.

82. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.

83. Savkovic-Stevanovic J. Neural net controller by inverse modeling for a distillation plant. Computers & chemical engineering. 1996; 20:5925-30.

84. Tomba E, Barolo M, García-Muñoz S. In-silico product formulation design through latent variable model inversion. Chemical Engineering Research and Design. 2014;92(3):534-44.

85. Tomba E, Barolo M, García-Muñoz S. In-silico product formulation design through latent variable model inversion. Chemical Engineering Research and Design. 2014;92(3):534-44.

86. Bayer B, von Stosch M, Striedner G, Duerkop M. Comparison of Modeling Methods for DoE-Based Holistic Upstream Process Characterization. Biotechnology Journal. 2020;15(5):1900551.

87. Raccuglia P, Elbert KC, Adler PD, Falk C, Wenny MB, Mollo A, Zeller M, Friedler SA, Schrier J, Norquist AJ. Machine-learning-assisted materials discovery using failed experiments. Nature. 2016;533(7601):73-6.
88. Liao TW, Li G. Metaheuristic-based inverse design of materials—A survey. Journal of Materiomics. 2020;6(2):414-30.

89. Zhou ZH. Ensemble Methods: Foundations and algorithms. Chapman & Hall/CRC Machine Learning, 2012.

90. MacGregor JF, Skagerberg B, Kiparissides C. Multivariate statistical process control and property inference applied to low density polyethylene reactors. In Advanced Control of Chemical Processes. IFAC Symposia Series. 1992; 155-159.

91. Rogers A, Ierapetritou M. Feasibility and flexibility analysis of black-box processes Part 1: Surrogate-based feasibility analysis. Chemical Engineering Science. 2015; 137:986-1004.

92. Rogers A, Ierapetritou M. Feasibility and flexibility analysis of black-box processes part 2: Surrogate-based flexibility analysis. Chemical Engineering Science. 2015; 137:1005-13.

93. Abdullah F, Wu Z, Christofides PD. Data-based reduced-order modeling of nonlinear two-time-scale processes. Chemical Engineering Research and Design. 2021; 166:1-9.

94. Agarwal A, Biegler LT, Zitney SE. Simulation and optimization of pressure swing adsorption systems using reduced-order modeling. Industrial & Engineering Chemistry Research. 2009; 48(5):2327-43.

95. Schäfer P, Caspari A, Kleinhans K, Mhamdi A, Mitsos A. Reduced dynamic modeling approach for rectification columns based on compartmentalization and artificial neural networks. AIChE Journal. 2019;65(5): e16568.
96. Zhang J, Yin J, Wang R. Basic Framework and Main Methods of Uncertainty Quantification. Mathematical Problems in Engineering. 2020. Article ID 6068203; https://doi.org/10.1155/2020/6068203.

97. Duong PL, Ali W, Kwok E, Lee M. Uncertainty quantification and global sensitivity analysis of complex chemical process using a generalized polynomial chaos approach. Computers & Chemical Engineering. 2016; 90:23-30.

98. Francis-Xavier F, Kubanek F, Schenkendorf R. Hybrid Process Models in Electrochemical Syntheses under Deep Uncertainty. Processes. 2021;9(4):704.

99. Proppe J, Husch T, Simm GN, Reiher M. Uncertainty quantification for quantum chemical models of complex reaction networks. Faraday discussions. 2017; 195:497-520.

100. Parks HL, McGaughey AJ, Viswanathan V. Uncertainty quantification in first-principles predictions of harmonic vibrational frequencies of molecules and molecular complexes. The Journal of Physical Chemistry C. 2019;123(7):4072-84.

101. Wang S, Pillai HS, Xin H. Bayesian learning of chemisorption for bridging the complexity of electronic descriptors. Nature Communications. 2020;11(1):1-7.

102. Boukouvala F, Ierapetritou MG. Feasibility analysis of black-box processes using an adaptive sampling Kriging-based method. Computers & Chemical Engineering. 2012; 36:358-68.

103. Ghenis, M. Quantile Regression- From Linear Regression to Deep Learning. https://towardsdatascience.com/quantile-regression-from-linear-models-to-trees-to-deep-learning-af3738b527c3, accessed August 5, 2021
104. Rudy SH, Brunton SL, Proctor JL, Kutz JN. Data-driven discovery of partial differential equations. Science Advances. 2017;3(4): e1602614.

105. Langley P, Bradshaw GL, Simon HA. Rediscovering chemistry with the BACON system. In Machine learning, Springer, Berlin, Heidelberg. 1983 ;307-329.

106. Nentwich C, Engell S. Surrogate modeling of phase equilibrium calculations using adaptive sampling. Computers & Chemical Engineering. 2019;126:204-17.

107. Wang SH, Pillai HS, Wang S, Achenie LE, Xin H. Infusing Theory into Machine Learning for Interpretable Reactivity Prediction. arXiv preprint arXiv:2103.15210. 2021.

108. Chen RT, Rubanova Y, Bettencourt J, Duvenaud D. Neural ordinary differential equations. arXiv preprint arXiv:1806.07366. 2018 Jun 19. Chen RT, Rubanova Y, Bettencourt J, Duvenaud D. Neural ordinary differential equations. arXiv preprint arXiv:1806.07366. 2018.

109. De Jaegher B, Larumbe E, De Schepper W, Verliefde A, Nopens I. Colloidal fouling in electrodialysis: A neural differential equations model. Separation and Purification Technology. 2020;249:116939.

110. Wu Z, Rincon D, Christofides PD. Process structure-based recurrent neural network modeling for model predictive control of nonlinear processes. Journal of Process Control. 2020;89:74-84.

111. Reis MS, Gins G, Rato TJ. Incorporation of process-specific structure in statistical process monitoring: A review. Journal of Quality Technology. 2019;51(4):407-21.

112. Hayashi Y, Buckley JJ, Czogala E. Fuzzy neural network with fuzzy signals and weights. International Journal of Intelligent Systems. 1993;8(4):527-37.
113. Brown M, Harris CJ. Neurofuzzy adaptive modelling and control, Prentice Hall, 1994.

114. Simutis R, Havlik I, Schneider F, Dors M, Lübert A. Artificial neural networks of improved reliability for industrial process supervision. IFAC Proceedings Volumes. 1995; 28(3):59-65.

115. Simutis R, Lübert A. Bioreactor control improves bioprocess performance. Journal of Biotechnology Journal. 2015; 10(8):1115-30.

116. Schubert J, Simutis R, Dors M, Havlik I, Lübert A. Bioprocess optimization and control: Application of hybrid modelling. Journal of Biotechnology. 1994; 35(1):51-68.

117. Lu J, Yao K, Gao F. Process similarity and developing new process models through migration. AIChE Journal. 2009; 55(9):2318-28.

118. Yan W, Hu S, Yang Y, Gao F, Chen T. Bayesian migration of Gaussian process regression for rapid process modeling and optimization. Chemical Engineering Journal. 2011; 166(3):1095-103.

119. Kumar A, Ridha S, Narahari M, Ilyas SU. Physics-Guided Deep Neural Network to Characterize Non-Newtonian Fluid Flow for Optimal Use of Energy Resources. Expert Systems with Applications. 2021; 19:115409.

120. Pun GP, Batra R, Ramprasad R, Mishin Y. Physically informed artificial neural networks for atomistic modeling of materials. Nature Communications. 2019; 10(1):1-0.

121. Fischer CC, Tibbetts KJ, Morgan D, Ceder G. Predicting crystal structure by merging data mining with quantum mechanics. Nature Materials. 2006; 5(8):641-6.
122. Hautier G, Fischer CC, Jain A, Mueller T, Ceder G. Finding nature’s missing ternary oxide compounds using machine learning and density functional theory. Chemistry of Materials. 2010;22(12):3762-7.

123. Cang R, Li H, Yao H, Jiao Y, Ren Y. Improving direct physical properties prediction of heterogeneous materials from imaging data via convolutional neural network and a morphology-aware generative model. Computational Materials Science. 2018;150:212-21.

124. Herwig C, Pörtner R, Möller J, editors. Digital Twins: Tools and Concepts for Smart Biomanufacturing. Springer, Berlin, Heidelberg; 2021.

125. Bangi MS, Kwon JS. Deep hybrid modeling of chemical process: Application to hydraulic fracturing. Computers & Chemical Engineering. 2020; 134:106696.

126. Kumari P, Bhadriraju B, Wang Q, Kwon JS. Development of parametric reduced-order model for consequence estimation of rare events. Chemical Engineering Research and Design. 2021; 169:142-52.

127. Narasingam A, Kwon JS. Development of local dynamic mode decomposition with control: Application to model predictive control of hydraulic fracturing. Computers & Chemical Engineering. 2017; 106:501-11.

128. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences. 2016;113(15):3932-7.
129. Bhadriraju B, Narasingam A, Kwon JS. Machine learning-based adaptive model identification of systems: Application to a chemical process. Chemical Engineering Research and Design. 2019; 152:372-83.

130. Bhadriraju B, Bangi MS, Narasingam A, Kwon JS. Operable adaptive sparse identification of systems: Application to chemical processes. AIChE Journal. 2020;66(11):e16980.