A Hybrid Science-Guided Machine Learning Approach for Modeling and Optimizing Chemical Processes

Niket Sharma and Y. A. Liu*, Aspen-Tech Center of Excellence in Process System Engineering, Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.  *Corresponding author: niket@vt.edu;

Abstract

This study presents a broad perspective of hybrid process modeling and optimization 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. The first refers to the case where a data-based ML model compliments and makes the first-principle 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 present a detailed review of scientific and engineering 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 the sub-categories 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 sub-categories of science-guided design, learning and refinement. For each sub-category, we identify its requirements, advantages and limitations, together with their published and potential areas of applications in bioprocessing and chemical engineering.

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, there are too many model parameters to estimate. 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-principle model. However, there are
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. 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. 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.

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-principle 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 chemical engineering applications, like the use of scientific knowledge to help improve the ML model architecture and learning process for more scientifically consistent solutions.

The objective of this paper is to present a comprehensive 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 sciences 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 3-5, presenting a systematic classification and exposition of hybrid SGML methodologies. 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 5.6 summarizes our conclusions.

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-principle model with data-based neural networks. Psichogios and Unger combine a partial first-principle 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 that it 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-principle 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 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-principle 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. 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.

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.

In their 2021 paper, Sansana et al. 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.
model transfer, feasibility analysis and predictive maintenance, apart from the previous mentioned applications like process control, monitoring and optimization.

Von Stosch et. al.\textsuperscript{18} have used the term, \textit{hybrid semi-parametric modeling}, in their 2014 review, and have summarized applications in biochemical engineering for process 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\textsuperscript{19} review the advances in statistical machine learning and process data analytics that can provide efficient tools in developing future hybrid models. In a latest paper, Qin et. al.\textsuperscript{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\textsuperscript{21-27}, chemical and oil and gas process industries\textsuperscript{28-32}, and polymer processes\textsuperscript{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\textsuperscript{35-38}, design of experiments\textsuperscript{39,40}, process development and scale-up\textsuperscript{41,42}, process design\textsuperscript{43} and optimization\textsuperscript{13,44,45}.

In a recent study, Zhou et al.\textsuperscript{46} present a \textit{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{25} 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 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.

A recent patent by Chan et al. presents Aspen Technology’s approach on asset optimization using integrated modeling, optimization and artificial intelligence. In a later white paper, Beck and Munoz describe Aspen Technology’s current focus on hybrid modeling, combining AI and domain expertise to optimize assets. In particular, based on their application experience in chemical industries, Aspen Tech have classified hybrid models into three categories: AI-driven, first-principle driven and reduced-order models. 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-principle driven hybrid model as an existing first-principle 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-principle 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. 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 subcategories 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.
4. ML Complements Science

We can integrate a first-principle scientific model with a data-based model to improve the model accuracy and consistency. In the following, we introduce the subcategories 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-principle 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.

4.1a 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.
Galvanauskas et. al.\textsuperscript{68} combine directly the data-based neural networks for kinetics and viscosity predictions with the first-principle 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.

\textit{Hybrid residual modeling} or \textit{parallel direct hybrid residual model} is a class of the parallel direct hybrid model, where we use a first-principle or science-based process model to quantify the time-dependent prediction error or residual, $Y_{\text{res}}$, 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

Tian et al.\textsuperscript{69} 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-principle 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\textsuperscript{9} trained by process data to predict the residuals of monomer conversion, MWN and MWW of the simplified first-principle 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\textsuperscript{41,43,69-71}.

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-principle 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-principle 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.

4.1b Series Direct Hybrid Model
Figure 4. Serial direct hybrid model

Figure 4 illustrates the serial 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.

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-principle 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 \textit{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-principle 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-principle models for industrial polyolefin processes. Finally, we note that as illustrated in Figure 5.4, we can interchangeably use a science-based model or a ML model first in the hybrid framework, depending on we require to add more features to augment the data set or to estimate model parameters.
4.1c Serial-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 5.3 for prediction. This serial-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.\(^7\) present a definitive study comparing first-principle, data-based and hybrid models applied to an industrial hydrocracking process. In particular, they couple a first-principle 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-principle process model connected in series or correct predictions of the first-principle 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.\(^8\) 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 *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 *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-principle model and the serial 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-principle 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.

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 \textsuperscript{2}. Figure 5.6 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).
One of the earliest applications of inverse modeling for chemical process was by Savkovic-Stevanovic et. al. 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. 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. 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. 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. present a metaheuristic approach to material design that incorporates the inverse modeling framework. Vankatasunramanian also mentions the importance of inverse problem being solved by the application of artificial intelligence in chemical engineering processes.

4.3 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 7. 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.

![Figure 7. Reduced order process modeling framework](image)

The concept of Digital Twin in chemical process industry is also based on the concept of Reduced Order Models of combining multiple models so that they can run multiple processes simultaneously to simulate a virtual plant.

In one of the earliest applications of ROM, MacGregor et. al. 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.

Reduced-order models have also been called *surrogate models* in the context of grey-box
modeling techniques where first-principle 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.

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.

4.4 Hybrid SGML Modeling for Uncertainty Quantification

A science-based model can produce results with some uncertainties which can be quantified by
some 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. 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. 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 8. 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.

Figure 8. Uncertainty quantification modeling framework

4.5 Hybrid HGML 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 (PDE). 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 PDE defining scientific laws as illustrated in figure 9.

Rudy et. al.\textsuperscript{103} 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.\textsuperscript{105} 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.\textsuperscript{107} use data-based mixed adaptive sampling strategy to calculate the phase composition, instead of the complex equation-of-state models. In another novel approach, Hofmann et al.\textsuperscript{108} discovers governing reactions from concentration data. They use a sparse tensor regression method for identification of non-linear dynamics to estimate a complex reaction networks.

Thus, ML application can have promising use in discovering more accurate physical and chemistry laws that govern the chemical process. These scientific laws calculated by ML-based models can then be utilized in first-principle model to improve accuracy as well as reduce model complexity.

![Diagram](image.png)

Figure 9. Discovering scientific laws

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 10 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-principle 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, Jaegher 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 structure. 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 is used to formulate rules mapped on to the structure of the ANN\textsuperscript{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. have used fuzzy ANN system for industrial bioprocess monitoring and control\textsuperscript{114-115}. Simutis et. al. have showcased the application of fuzzy ANN process control expert to perform appropriate control actions based on process trends for bioprocess optimization and control\textsuperscript{116}.

![Figure 10. Science-guided design framework of neural network architecture](image)

**5.2 Science-Guided Learning**

Here, we make use of the scientific principles to improve the scientific consistency of database-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 NN model incorporate physical constraints\textsuperscript{2}. A loss function in ML measures how far an estimated value is from its true value. A loss function maps decision to their associated costs. Loss functions
are not fixed, they change depending on the task in hand and the goal to be met. Figure 11 illustrates the science-guided loss framework.

We can define a loss function of the ML model \((\text{Loss}_M)\) for regression to calculate the difference between the true value \((Y_{\text{true}})\) and the model predicted value \((Y_{\text{pred}})\). Likewise, we can define a loss function for a science-based model \((\text{Loss}_{SC})\), which is a function of the model predicted value \((Y_{\text{pred}})\) 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 \((\text{Loss})\) as:

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

Figure 5.8 illustrates the concept of science-guided loss function.

![Science-guided loss function representation](image)

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.\(^117\) 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.

In another study on the similar concept, Yan et al.\(^118\) 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 Science-Guided Refinement

By science-guided refinement, we mean the post-processing of ML model results based on scientific principle. The raw data without any feature engineering can be used as input to a ML model for prediction and then the science-based model can be used to test the scientific consistency of ML predictions. Figure 12. illustrates the science-guided refinement framework. 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 4.1b. 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-principle models based on density functional model to refine the results of probabilistic ML models to discovery ternary oxides.

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 data. 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.

Figure 12. Science-guided refinement framework

Thus, some of these methodologies of having science complimenting ML have much potential for future applications to bioprocessing and chemical engineering.
6. Conclusion

Table 1 summarizes all the hybrid SGML models and their advantages, limitations and potential applications.

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-principle 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.

Table 1 Summary of hybrid SGML approach

| Hybrid SGML Modeling | Science-/based model/knowledge | ML model | Advantages | Limitations | Potential applications |
|----------------------|--------------------------------|----------|------------|-------------|------------------------|
| Direct Hybrid modeling | Series Science-based model (SBM) | Regression | Parameter estimation, data augmentation | Limited by data for parameter estimation | Kinetic estimation soft sensor \cite{51,86} and process modeling \cite{29,73}, process scale-up \cite{41} |
| Parallel | SBM | Regression | Improved accuracy of prediction | Scientific consistency depends on SBM | Process control \cite{36,43,72} and softsSensor \cite{33} and predictive maintenance \cite{66,74} |
| Series-parallel | SBM | Regression | Higher Accuracy | Increased Model complexity | Process monitoring and control \cite{58} and plant-model mismatch \cite{17} |
| Inverse modeling | SBM | Probabilistic, Regression | Computationally cheaper inverse problem | Lower generality of the model | Product design and development \cite{84,85} and polymer grade change, material design \cite{87} |
Reduced-order models | SBM | Regression | Fast online deployment, reduce model complexity | Higher Bias, limited by SBM accuracy | Process Optimization at plant scale 84,85, dynamic modeling 95,95, soft Sensor 94, feasibility Analysis 91,92

Uncertainty quantification | SBM | Probabilistic | Gives real error estimate and solution space | Limited by SBM assumptions, parameters | Process design and development 97, feasibility Analysis 102

Discovering scientific law | SBM | Regression, probabilistic | System stability interpretability | Limited by data size/availability | Thermodynamics phase equilibrium 105, Reaction Network 106

Science Compliments ML (base model : ML)

Science-guided -design | Laws or SBM | Deep neural network (DNN) Neural Diff. Eqn. | Scientifically consistent and interpretable | Requires deep scientific knowledge of system | Dynamic time-dependent systems 109, process control 110,111

Science-guided -learning | Laws | DNN | Scientifically consistent and interpretable | Possible lower prediction accuracy | Process design and development 100, process Monitoring,

Science-guided refinement | SBM | Probabilistic | Less effort in feature selection | Limited by SBM assumptions, parameters | Process design, discovery of materials 106

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. We would like to thank Dr. 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, J. H.; 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, 2318.

2. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V., Integrating Physics-Based Modeling with Machine Learning: A Survey. 2020, arXiv preprint arXiv:2003.04919v4 [physics.comp-ph].
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, *Proceedings of the 2020 SIAM International Conference on Data Mining*, SIAM: 2020, 559.

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. 2019, arXiv preprint arXiv:1911.11380.

5. Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R., SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. 2017, arXiv preprint arXiv:1706.08566.

6. Faghmous, J. H.; Kumar, V., A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science. *Big Data*. 2014, 2, 155.

7. Karpatne, A.; Watkins, W.; Read, J.; Kumar, V. Physics-Guided Neural Networks: An Application in Lake Temperature Modeling., https://arxiv.org/abs/1710.11431 [cs.LG], accessed August 1, 2021.

8. Yazdani, A.; Lu, L.; Raissi, M.; Karniadakis, G. E., Systems Biology Informed Deep Learning for Inferring Parameters and Hidden Dynamics. *PLoS Computational Biology*. 2020, 16, e1007575.

9. Baughman, D. R.; Liu, Y. A. *Neural Networks in Bioprocessing and Chemical Engineering*. Academic Press, San Didgo, CA, 1995.

10. Psichogios, D. C.; Ungar, L. H., A Hybrid Neural Network-First Principles Approach to Process Modeling. *AIChE Journal*. 1992, 38, 1499.

11. Thompson, M. L.; Kramer, M. A., Modeling Chemical Processes Using Prior Knowledge and Neural Networks. *AIChE Journal*. 1994, 40, 1328.

12. Agarwal, M., Combining Neural and Conventional Paradigms for Modelling, Prediction and Control. *International Journal of Systems Science*. 1997, 28, 65.

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

14. Bohlin, T., *Practical Grey-Box Process Identification*. Springer-Verlag, London, United Kingdom, 2006.

15. Yang, S.; Navarathna, P.; Ghosh, S.; Bequette, B. W., Hybrid Modeling in the Era of Smart Manufacturing. *Computers and Chemical Engineering*. 2020, 140, 106874.

16. Sansana, J.; Joswiak, M. N.; Castillo, I.; Wang, Z.; Rendall, R.; Chiang, L. H.; Reis, M. S., Recent Trends on Hybrid Modeling for Industry 4.0. *Computers and Chemical Engineering*. 2021, 151, 107365.

17. Chen, Y.; Ierapetritou, M., A Framework of Hybrid Model Development with Identification of Plant-Model Mismatch, *AIChE J*. 2020, 66, e16996.

18. Von Stosch, M.; Oliveira, R.; Peres, J.; de Azevedo, S. F., Hybrid Semi-Parametric Modeling in Process Systems Engineering: Past, Present and Future. *Computers and Chemical Engineering*. 2014, 60, 86.
19. Qin, S. J.; Chiang, L. H. Advances and Opportunities in Machine Learning for Process Data Analytics. *Computers and Chemical Engineering*. 2019, 125, 465.
20. Qin, S. J.; Guo, S.; Li, Z.; Chiang, L. H.; Castillo, I.; Braun, B.; Wang, Z., Integration of Process Knowledge and Statistical Learning for the Dow Data Challenge Problem, *Computers and Chemical Engineering* 2021, 107451.
21. O'Brien, C. M.; Zhang, Q.; Daoutidis, P.; Hu, W.-S., A Hybrid Mechanistic-Empirical Model for in silico Mammalian Cell Bioprocess Simulation. *Metabolic Engineering* 2021, 66, 31.
22. Pinto, J.; de Azevedo, C. R.; Oliveira, R.; von Stosch, M., A Bootstrap-Aggregated Hybrid Semi-Parametric Modeling Framework for Bioprocess Development. *Bioprocess and biosystems engineering* 2019, 42, 1853.
23. Chopda, V.; Gyorgypal, A.; Yang, O.; Singh, R.; Ramachandran, R.; Zhang, H.; Tsiolomelekis, G.; Chundawat, S. P.; Ierapetritou, M. G., Recent Advances in Integrated Process Analytical Techniques, Modeling, and Control Strategies to Enable Continuous Biomanufacturing of Monoclonal Antibodies. *Journal of Chemical Technology and Biotechnology*. 2021. https://doi.org/10.1002/jctb.6765, accessed July 23, 2021.
24. Zhang, D.; Del Rio-Chanona, E. A.; Petsagkourakis, P.; Wagner, J., Hybrid Physics-Based and Data-Driven Modeling for Bioprocess Online Simulation and Optimization. *Biotechnology and Bioengineering*. 2019, 116, 2919.
25. Al-Yemni, M.; Yang, R. Y., Hybrid Neural-Networks Modeling of an Enzymatic Membrane Reactor. *Journal of the Chinese Institute of Engineers*. 2005, 28, 1061.
26. Chabbi, C.; Taibi, M.; Khier, B., Neural and Hybrid Neural Modeling of a Yeast Fermentation Process. *International Journal of Computational Cognition*. 2008, 6, 42.
27. Zanin, G., Determination of Inhibition in the Enzymatic Hydrolysis of Cellobiose Using Hybrid Neural Modelling. *Brazilian Journal of Chemical Engineering*. 2005, 22, 19.
28. Azarpour, A.; Borhani, T. N.; Alwi, S. R. W.; Manan, Z. A.; Matalib, M. I. A., A Generic Hybrid Model Development for Process Analysis of Industrial Fixed-Bed Catalytic Reactors. *Chemical Engineering Research and Design*. 2017, 117, 149.
29. Luo, N.; Du, W.; Ye, Z.; Qian, F., Development of a Hybrid Model for Industrial Ethylene Oxide Reactor. *Industrial and Engineering Chemistry Research*. 2012, 51, 6926.
30. Zahed, G.; Lohi, A.; Mahdi, K., Hybrid Modeling of Ethylene to Ethylene Oxide Heterogeneous Reactor. *Fuel Processing Technology*. 2011, 92, 1725.
31. Simon, L. L.; Fischer, U.; Hungerbühler, K., Modeling of a Three-Phase Industrial Batch Reactor Using a Hybrid First-Principles Neural-Network Model. *Industrial and Engineering Chemistry Research*. 2006, 45, 7336.
32. Bellos, G.; Kallinikos, L.; Gounaris, C.; Papayannakos, N., Modelling of the Performance of Industrial HDS Reactors using a Hybrid Neural Network Approach. *Chemical Engineering and Processing: Process Intensification*. 2005, 44, 505.
33. Chang, J.-S.; Lu, S.-C.; Chiu, Y.-L., Dynamic Modeling of Batch Polymerization Reactors via the Hybrid Neural-Network Rate-Function Approach. *Chemical Engineering Journal*. 2007, 130, 19.
34. Hinchliffe, M.; Montague, G.; Willis, M.; Burke, A.,
Hybrid Approach to Modeling an Industrial Polyethylene Process. *AIChE journal*. **2003**, *49*, 3127.

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*, 343.

36. Simutis, R.; Lübbert, A., Hybrid Approach to State Estimation for Bioprocess Control. *Bioengineering*. **2017**, *4*, 21.

37. Cubillos, F.; Callejas, H.; Lima, E.; Vega, M., Adaptive Control Using a Hybrid-Neural Model: Application to a Polymerisation Reactor. *Brazilian Journal of Chemical Engineering*. **2001**, *18*, 113.

38. Doyle III, F. J.; Harrison, C. A.; Crowley, T. J., Hybrid Model-Based Approach to Batch-to-Batch Control of Particle Size Distribution in Emulsion Polymerization. *Computers and Chemical Engineering*. **2003**, *27*, 1153.

39. Rodriguez-Granrose, D.; Jones, A.; Loftus, H.; Tandeski, T.; Heaton, W.; Foley, K. T.; Silverman, L., Design of Experiment (DOE) Applied to Artificial Neural Network Architecture Enables Rapid Bioprocess Improvement. *Bioprocess and Biosystems Engineering*. **2021**, *44*, 1301.

40. Brendel, M.; Marquardt, W., Experimental Design for the identification of hybrid reaction models from transient data. *Chemical Engineering Journal*. **2008**, *141*, 264.

41. Bollas, G.; Papadokonstadakis, S.; Michalopoulos, J.; Arampatjis, G.; Lappas, A.; Vasalos, I.; 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*, 697.

42. von Stosch, M.; Hamelin, J.-M.; Oliveira, R., Hybrid Modeling as a QbD/PAT Tool in Process Development: An Industrial E. coli Case Study. *Bioprocess and Biosystems Engineering*. **2016**, *39*, 773.

43. Iwama, R.; Kaneko, H., Design of Ethylene Oxide Production Process Based on Adaptive Design of Experiments and Bayesian Optimization. *AIChE J.*, **2021**, *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*, 253.45.

45. Kumar, A.; Baldea, M.; Edgar, T. F., Real-Time Optimization of an Industrial Steam-Methane Reformer under Distributed Sensing. *Control Engineering Practice*. **2016**, *54*, 140.

46. Zhou, T.; Gani, R.; Sundmacher, K., Hybrid Data-Driven and Mechanistic Modeling Approaches for Multiscale Material and Process Design. *Engineering* **2021**. [https://doi.org/10.1016/j.eng.2020.12.022](https://doi.org/10.1016/j.eng.2020.12.022), accessed Jul 23, 2021.

47. Cardillo, A. G.; Castellanos, M. M.; Desailly, B.; Dessoy, S.; Mariti, M.; Portela, R. M.; Scutella, B.; von Stosch, M.; Tomba, E.; Varsakelis, C., Towards in silico Process Modeling for Vaccines. *Trends in Biotechnology* **2021**. [https://doi.org/10.1016/j.tibtech.2021.02.004](https://doi.org/10.1016/j.tibtech.2021.02.004), accessed July 23, 2021.

48. McBride, K.; Sanchez Medina, E. I.; Sundmacher, K., Hybrid Semi-Parametric Modeling in Separation Processes: A Review. *Chemie Ingenieur Technik*. **2020**, *92*, 842.
49. Safavi, A.; Nooraii, A.; Romagnoli, J., A Hybrid Model Formulation for a Distillation Column and the On-line Optimisation Study. *Journal of Process Control* **1999**, *9*, 125.
50. Mahalec, V. Hybrid Modeling of Petrochemical Processes. In *Hybrid Modeling in Process Industries*, Glassey, J. and Von Stosch, M., editor, CRC Press, Boca Raton, FL, pp.130-165.
51. Mahalec, V.; Sanchez, Y., Inferential Monitoring and Optimization of Crude Separation Units via Hybrid Models. *Computers and Chemical Engineering*, **2012**, *45*, 15.
52. Peroni, C. V.; Parisi, M.; Chianese, A., Hybrid Modelling and Self-Learning System for Dextrose Crystallization Process. *Chemical Engineering Research and Design* **2010**, *88*, 1653.
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*, 11.
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.
55. Nentwich, C.; Winz, J.; Engell, S., Surrogate Modeling of Fugacity Coefficients Using Adaptive Sampling. *Industrial & Engineering Chemistry Research* **2019**, *58*, 18703.
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*, 753.
57. Côte, M.; Grandjean, B. P.; Lessard, P.; Thibault, J., Dynamic Modelling of the Activated Sludge Process: Improving Prediction Using Neural Networks. *Water Research* **1995**, *29*, 995.
58. Dong, Z.; Wang, R.; Fan, M.; Fu, X., Switching and Optimizing Control for Coal Flotation Process Based on a Hybrid Model. *PLoS ONE* **2017**, *12*, e0186553. 12(10): e0186553. [https://doi.org/10.1371/journal.pone.0186553](https://doi.org/10.1371/journal.pone.0186553). Accessed July 25, 2021.
59. Hwang, T.-M.; Oh, H.; Choi, Y.-J.; Nam, S.-H.; Lee, S.; Choung, Y.-K., Development of a Statistical and Mathematical Hybrid Model to Predict Membrane Fouling and Performance. *Desalination* **2009**, *247*, 210.
60. Piron, E.; Latrille, E.; Rene, F., Application of Artificial Neural Networks for Crossflow Microfiltration Modelling: “Black-Box” and Semi-Physical Approaches. *Computers and Chemical Engineering* **1997**, *21*, 1021.
61. Zbiciński, I.; Strumiło, P.; Kamiński, W., Hybrid Neural Model of Thermal Drying in a Fluidized Bed. *Computers and Chemical Engineering* **1996**, *20*, S695.
62. Cubillos, F. A.; Alvarez, P. I.; Pinto, J. C.; Lima, E. L., Hybrid-Neural Modeling for Particulate Solid Drying Processes. *Powder Technology* **1996**, *87*, 153.
63. Venkatasubramanian, V., The Promise of Artificial intelligence in Chemical Engineering: Is it Here, Finally? *AIChE Journal* **2019**, *65*, 466.
64. Glassy, J.; Von Stosch, M., editors, *Hybrid Modeling in Process Industries*. CRC Press, Boca Raton, FL, **2018**.
65. Kahrs, O.; Marquardt, W., Incremental Identification of Hybrid Process Models. *Computers and Chemical Engineering* **2008**, *32*, 694.
66. Chan, W. K. C.; Fischer, B.; Varvarezos, D.; Rao, A.; Zhao, H., Asset Optimization Using Integrated Modeling, Optimization, and Artificial Intelligence. U.S. patent 20200287818A1, Dec. 10, 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, accessed June 18, 2021.  
68. Galvanauskas, V.; Simutis, R.; Lübbert, A., Hybrid Process Models for Process Optimization, Monitoring and Control. Bioprocess and Biosystems Engineering 2004, 26, 393.  
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 and Engineering Chemistry Research 2001, 40, 4525.
70. Su, H. T.; McAvoy, T. J.; Werbos, P., Long-Term Predictions of Chemical Processes Using Recurrent Neural Networks: A Parallel Training Approach. Industrial and Engineering Chemistry Research 1992, 31, 1338.
71. Hermanto, M. W.; Braatz, R. D.; Chiu, M. S., Integrated Batch-to-Batch and Nonlinear Model Predictive Control for Polymorphic Transformation in Pharmaceutical Crystallization. AIChE journal 2011, 57, 1008.
72. Ghosh, D.; Hermonat, E.; Mhaskar, P.; Snowling, S.; Goel, R., Hybrid Modeling Approach Integrating First-Principles Models with Subspace Identification. Industrial and Engineering Chemistry Research 2019, 58, 13533.
73. Ghosh, D.; Moreira, J. S.; Mhaskar, P., Model Predictive Control Embedding a Parallel Hybrid Modeling Strategy. Industrial and Engineering Chemistry Research 2021, 60, 2547.
74. Hanachi, H.; Yu, W.; Kim, I.Y.; Liu, J.; Mechefske, C.K. Hybrid Data-Driven Physics-Based Model Fusion Framework for Tool Wear Prediction. The International Journal of Advanced Manufacturing Technology, 2019, 101, 2861.
75. Babanezhad, M.; Behroyan, I.; Nakhjiri, A. T.; Marjani, A.; Rezkazemi, M.; Shirazian, S., High-Performance Hybrid Modeling Chemical Reactors Using Differential Evolution Based Fuzzy Inference System. Scientific Reports 2020, 10, 1.
76. Krippel, 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, I. C.; Rivera, E. C.; Da Costa, A. C.; Maciel Filho, R., Hybrid Neural Network Model of an Industrial Ethanol Fermentation Process Considering the Effect of Temperature. Applied Biochemistry and Biotechnology 2007, 137, 817.
78. Sharma, N.; Liu, Y. A., 110th Anniversary: An Effective Methodology for Kinetic Parameter Estimation for Modeling Commercial Polyolefin Processes from Plant Data Using Efficient Simulation Software Tools. Industrial and Engineering Chemistry Research 2019, 58, 14209.
79. Bhutani, N.; Rangaiah, G.; Ray, A., First-Principles, Data-Based, and Hybrid Modeling and Optimization of an Industrial Hydrocracking Unit. Industrial and Engineering Chemistry Research 2006, 45, 7807.
80. Song, W.; Du, W.; Fan, C.; Yang, M.; Qian, F., Adaptive Weighted Hybrid Modeling of Hydrocracking Process and Its Operational Optimization. *Industrial and Engineering Chemistry Research* **2021**, *60*, 3617.

81. Lima, P.V.; Saraiva, P.M.; and Group, G.P. A Semi-Mechanistic Model Building Framework Based on Selective and Localized Model Extensions. *Computers and Chemical Engineering*. **2007**, *31*, 361.

82. Breiman, L. Random Forests. *Machine Learning*. **2001**, *45*, 5.

83. Savkovic-Stevanovic, J., Neural Net Controller by Inverse Modeling for a Distillation Plant. *Computers and Chemical Engineering* **1996**, *20*, S925.

84. Tomba, E., Barolo, M. and García-Muñoz, S., General Framework for Latent Variable Model Inversion for the Design and Manufacturing of New Products. *Industrial and Engineering Chemistry Research* **2012**, *51*, 12886.

85. Tomba, E., Barolo, M. and García-Muñoz, S., In-Silico Product Formulation Design through Latent Variable Model Inversion. *Chemical Engineering Research and Design*. **2014**, *92*, 534.

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*, 1900551.

87. Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J., Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73.

88. Liao, T. W.; Li, G., Metaheuristic-Based Inverse Design of Materials–A Survey. *Journal of Materiomics* **2020**, *6*, 414.

89. Bamberg, A., Urbas, L., Bröcker, S., Bortz, M. and Kockmann, N.,The Digital Twin–Your Ingenious Companion for Process Engineering and Smart Production. *Chemical Engineering & Technology* **2021**, *44*, 954.

90. MacGregor, J.; Skagerberg, B.; Kiparissides, C., Multivariate Statistical Process Control and Property Inference Applied to Low Density Polyethylene Reactors. In *Advanced Control of Chemical Processes 1991*, Elsevier, Amsterdam, Netherlands, **1992**; pp 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.

92. Rogers, A. and Ierapetritou, M., Feasibility and Flexibility Analysis of Black-Box Processes Part 2: Surrogate-based Feasibility Analysis. *Chemical Engineering Science*. 2015, 137, pp.1005-1013.

93. Abdullah, F.; Wu, Z.; Christofides, P. D., Data-Based Reduced-Order Modeling of Nonlinear Two-Time-Scale Processes. *Chemical Engineering Research and Design* **2021**, *166*, 1-9.

94. Agarwal, A.; Biegler, L. T.; Zitney, S. E., Simulation and Optimization of Pressure Swing Adsorption Systems Using Reduced-Order Modeling. *Industrial and Engineering Chemistry Research*. **2009**, *48*, 2327.

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*, 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, P. L. T.; Ali, W.; Kwok, E.; Lee, M., Uncertainty Quantification and Global Sensitivity Analysis of Complex Chemical Process Using a Generalized Polynomial Chaos Approach. *Computers and Chemical Engineering* **2016**, *90*, 23.

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

99. Proppe, J.; Husch, T.; Simm, G. N.; Reiher, M., Uncertainty Quantification for Quantum Chemical Models of Complex Reaction Networks. *Faraday discussions* **2017**, *195*, 497.

100. Parks, H. L.; McGaughey, A. J.; Viswanathan, V., Uncertainty Quantification in First-Principles Predictions of Harmonic Vibrational Frequencies of Molecules and Molecular Complexes. *Journal of Physical Chemistry C* **2019**, *123*, 4072.

101. Wang, S.; Pillai, H. S.; Xin, H., Bayesian Learning of Chemisorption for Bridging the Complexity of Electronic Descriptors. *Nature communications* **2020**, *11*, 1.

102. Boukouvala, F.; Ierapetritou, M. G., Feasibility analysis of Black-Box Processes using an Adaptive Sampling Kriging-based Method. *Computers and Chemical Engineering* **2012**, *36*, 358.

103. Rudy, S. H.; Brunton, S. L.; Proctor, J. L.; Kutz, J. N., Data-Driven Discovery of Partial Differential equations. *Science Advances* **2017**, *3*, e1602614.

104. Langley, P.; Bradshaw, G. L.; Simon, H. A., Rediscovering Chemistry with the Bacon System. In *Machine Learning: An Artificial Intelligence Approach*, Michalski, R. S.; Carbonell, J. G.; Mitchell, T. M., Eds. Springer, Berlin Heidelberg 1983; pp 307-329.

105. Nentwich, C.; Engell, S., Surrogate Modeling of Phase Equilibrium Calculations Using Adaptive Sampling. *Computers and Chemical Engineering* **2019**, *126*, 204.

106. Hoffmann, M., Fröhner, C. and Noé, F., Reactive SINDy: Discovering Governing Reactions from Concentration Data. *The Journal of chemical physics* **2019**, *150*, 025101.

107. Wang, S.-H.; Pillai, H. S.; Wang, S.; Achenie, L. E.; Xin, H., Infusing Theory into Machine Learning for Interpretable Reactivity Prediction. *arXiv:2103.15210v2[physics,chem-ph]*. 2021. [https://arxiv.org/abs/2103.15210](https://arxiv.org/abs/2103.15210). Accessed August 5, 2021.

108. Chen, R. T.; Rubanova, Y.; Bettencourt, J.; Duvenaud, D., Neural Ordinary Differential Equations. *arXiv:1806.07366v5[cs.LG]*, 2018. [https://arxiv.org/abs/1806.07366](https://arxiv.org/abs/1806.07366). Accessed August 5, 2021.

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, Zhe, David Rincon, and Panagiotis D. Christofides., Process Structure-Based Recurrent Neural Network Modeling for Model Predictive Control of Nonlinear Processes. *Journal of Process Control* **2020**, *89*, 74.
111. Reis, Marco S., Geert Gins, and Tiago J. Rato., Incorporation of Process-Specific Structure in Statistical Process Monitoring: A Review., Journal of Quality Technology , 2019, 51, , 407.
112. Hayashi, Y., Buckley, J.J. and Czogala, E., Fuzzy Neural Network with Fuzzy Signals and Weights., International Journal of Intelligent Systems 1993., 8, 527.
113. Brown, M. and Harris, C.J., Neurofuzzy Adaptive Modelling and Control 1994.
114. Simutis, R., Havlik, I., Schneider, F., Dors, M. and Lübbert, A., 1995. Artificial Neural Networks of Improved Reliability for Industrial Process Supervision. IFAC Proceedings Volumes, 28, pp.59.
115. Simutis, R. and Lübbert, A., 2015. Bioreactor Control Improves Bioprocess Performance. Biotechnology journal, 10, pp.1115.
116. Schubert, J., Simutis, R., Dors, M., Havlik, I. and Lübbert, A., 1994. Bioprocess Optimization and Control: Application of Hybrid Modelling. Journal of biotechnology, 35, pp.51.
117. Lu, J.; Yao, K.; Gao, F., Process Similarity and Developing New Process Models through Migration. AIChE journal 2009, 55, 2318.
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, 1095.
119. Kumar, A.; Ridha, S.; Narahari, M.; Ilyas, S. U., Physics-Guided Deep Neural Network to Characterize Non-Newtonian Fluid Flow for Optimal Use of Energy Resources. Expert Systems with Applications 2021, 183, 115409.
120. Pun, G. P.; Batra, R.; Ramprasad, R.; Mishin, Y., Physically Informed Artificial Neural Networks for Atomistic Modeling of Materials. Nature communications 2019, 10, 1.
121. Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G., Predicting Crystal Structure by Merging Data Mining with Quantum Mechanics. Nature Materials 2006, 5, 641.
122. Hautier, G.; Fischer, C. C.; 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, 3762.
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.