Machine Learning in Process Systems Engineering: Challenges and Opportunities

Prodromos Daoutidis[a,*], Jay H. Lee[b,**], Srinivas Rangarajan[c], Leo Chiang[d], Bhushan Gopaluni[e], Artur M. Schweidtmann[f], Iiro Harjunkoski[g], Mehmet Mercangöz[h], Ali Mesbah[i], Fani Boukouvala[j], Fernando V. Lima[k], Antonio del Rio Chanona[h], Christos Georgakis[l]

[a]*Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN*
[b]*Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, CA*
[c]*Department of Chemical & Biomolecular Engineering, Lehigh University, Bethlehem PA*
[d]*The Dow Chemical Company, Lake Jackson TX*
[e]*Department of Chemical Engineering, University of British Columbia, Vancouver BC*
[f]*Department of Chemical Engineering, Delft University of Technology, Delft, The Netherlands*
[g]*Hitachi Energy Research, Mannheim, Germany / Aalto University, Espoo, Finland*
[h]*Department of Chemical Engineering, Imperial College, London, UK*
[i]*Department of Chemical & Biomolecular Engineering, University of California, Berkeley CA*
[j] *School of Chemical & Biomolecular Engineering, Georgia Instotute of Technology, Atlanta GA*
[k]*Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown WV*
[l]*Department of Chemical and Biological Engineering, Tufts University,Boston MA*

Abstract

This "white paper" is a concise perspective of the potential of machine learning in the process systems engineering (PSE) domain, based on a session during FIPSE 5, held in Crete, Greece, June 27-29, 2022. The session included two invited talks and three short contributed presentations followed by extensive discussions. This paper does not intend to provide a comprehensive review on the subject or a detailed exposition of the discussions; instead its aim is to distill the main points of the discussions and talks, and in doing so, highlight open problems and directions for future research. The general conclusion from the session was that machine learning can have a transformational impact on the PSE domain enabling new discoveries and innovations, but research is needed to develop domain-specific techniques for problems in molecular/material design, data analytics, optimization, and control.

Keywords: machine learning, molecule discovery, process monitoring, control, optimization, modeling

1. Introduction

Machine learning (ML), artificial intelligence (AI) and more generally data science are attracting tremendous attention across science and technology fields. The increasing availability of data and computing power, and significant algorithmic advances have resulted in several breakthroughs in image and video processing, natural language processing, voice recognition, and game playing. Deep learning, reinforcement learning, and other ML algorithms have been central to these breakthroughs. ML techniques are also being rapidly adopted by the process industries due to the realization that they can be key enablers of innovation and efficiency in the discovery and engineering of new products, automation, as well as the management of supply chains and company operations. Chemical engineering as a discipline is similarly impacted by these developments (Venkatasubramanian, 2019; Schweidtmann et al., 2021; Pistikopoulos et al., 2021) in areas such as cheminformatics, bioinformatics, materials design, and process systems engineering (PSE). In PSE specifically, ML techniques can find potential applications in multiple areas (Lee et al., 2018; Chiang et al., 2017; Zavala, 2023), such as in:

- Flowsheet analysis
- Surrogate modeling for simulation and optimization
- Integrated planning and scheduling
- Supply chain design and operation
- Process monitoring and fault diagnosis
- Real time optimization and control

At the same time, data generated in chemical engineering applications tend to be heterogeneous (discrete or continuous, time-series or static), high-dimensional, noisy, biased, and are typically constrained by physical laws (Thebelt et al., 2022). This hinders the direct adoption of

[*]Corresponding author
[**]Corresponding author
Email addresses: daout001@umn.edu (Prodromos Daoutidis), jlee4140@usc.edu (Jay H. Lee)

existing data-driven inference and prediction methods for learning models. From a process operations point-of-view, whereas "observational" data may be abundant, truly "informational" data are limited and hard to obtain since process plants are typically run at the same conditions for long times. For optimization and control, exploratory data are needed but the exploration space is typically limited by safety or operational constraints and on-line learning must be carried out safely with minimal impact to on-going production. Off-line learning is also challenged as available data tend to be confined to specific operation conditions and realistic simulators are seldom available. These challenges in turn raise non-trivial questions regarding the most effective utilization of data in process optimization and control. Finally, the materials design and PSE communities do not interact as closely, despite the fact that PSE offers an abundance of methods and tools that could be invaluable to materials discovery and design tasks.

The goal of this FIPSE session was to shed light on the afore-mentioned challenges which also represent opportunities for further research in PSE. The following two keynote talks anchored the session:

- **Machine Learning Challenges and Opportunities for Catalysis and Materials Design**, Srinivas Rangarajan, Lehigh University
- **Industrial Perspective of Machine Learning and AI Challenges in PSE**, Leo Chiang, Dow Chemical

These talks were complementary, offering both an academic and an industrial perspective, and covering both materials design and process operations. In addition to these, the following contributed talks were presented:

- **Big Data are Not Necessarily Good Data: What are Good Data Anyways?**, Bhushan Gopaluni and Richard D. Braatz, University of British Columbia and Massachusetts Institute of Technology
- **Using ML and AI to Speed-Up Large-Scale Optimization Problems**, Iiro Harjunkoski, Hitachi Energy and Aalto University
- **What is the Future of Systems Modeling?**, Mehmet Mercangöz, Imperial College London

These talks raised critical questions on the role of data-driven models across PSE applications and the degree to which the Big Data revolution can have a major impact in the process industries like it has in other sectors. Following these talks, there were extensive discussions organized along the following themes:

1. How can data science and PSE contribute to the design and discovery of new chemicals and materials? *Discussion leader:* Antonio del Rio Chanona
2. What additional advantages can modern ML techniques offer over the existing approaches to process monitoring? How can we create the educated workforce and culture to incorporate these techniques into industrial operation? *Discussion leader:* Leo Chiang
3. How can ML and AI aid in the solution of large-scale optimization problems? *Discussion leader:* Artur Schweidtmann
4. Which models among first-principles, data-driven, or hybrid ones are best suited for control? *Discussion leader:* Fernando V. Lima

The goal of this paper is to distill and summarize the views expressed in these presentations and discussions. The next section focuses on catalysis and materials emphasizing the PSE challenges and opportunities, and the subsequent one on industrial data analytics, control, and optimization.

2. Machine learning and PSE in catalysis and materials design

Data science and ML have become a mainstay in catalysis and materials design. While sophisticated data-driven techniques are increasingly being employed in these fields, several methodological challenges remain. In what follows, we identify several topics with substantial opportunities for PSE and foundational ML experts to make significant contributions in terms of methodological developments that, in turn, lead to deployable frameworks and software for use by catalysis and materials modelers. For each topic a brief description of the problem/challenges is provided followed by opportunities for PSE experts.

2.1. Learning kinetic models from data

A classic application of the ubiquitous PSE tool, i.e., optimization, is the parameter estimation problem of learning the kinetic/thermodynamic parameters of a physics-based model, e.g., a microkinetic model, from experimental data. Such problems are typically non-convex optimization problems, often subjected to stiff ordinary differential equations as constraints (representing the reaction model) and addressed using sequential approaches (Rangarajan et al., 2017; Matera et al., 2019). However, opportunities emerge for PSE experts to enable the mainstream use of physics-informed ML to solve differential equations (Karniadakis et al., 2021; Gusmão et al., 2022) and automatic differentiation to compute analytical derivatives to train physics-based models with data (Andersson et al., 2019).

Data-driven models can also serve as surrogates for physics-based models (Döppel and Votsmeier, 2022) when the latter models are too time-consuming to execute in a larger simulation effort (e.g., reactor or process simulation). One may also derive purely data-driven models from experimental kinetics data when ab initio inputs are limited or the reaction system is too complex to readily model at the atomic scale (Lejarza et al., 2023). In both of these cases, data-driven models may be grounded in domain knowledge (set either in the formulation of the model or enforced through constraints) so that fundamental laws (e.g., mass balance) are always satisfied.

2.2. Optimization of computationally expensive functions

The design of catalysts and materials usually involves large-scale screening (Zhong et al., 2020; Gómez-Bombarelli et al., 2016) of the plausible material space, often aided by machine learned models. However, catalyst/material design can be formulated as a nonlinear constrained optimization problem. The objectives and constraints may, among others, include energy functions to estimate the stability of one phase over competing ones, quantum chemical calculations to compute the kinetics of a rate-determining step, molecular simulations to compute equilibrium properties, etc. Evaluation of such functions is often computationally expensive. Further, decision variables can be discrete, e.g., whether or not a specific atom should be present at a location or if a bond between a pair of atom exists. Finally, the material space tends to be very large, defying any kind of exhaustive search for a global solution. Standard gradient-based optimization formulations can naturally not be applied in such cases; methods that identify the optimal value of the function while minimizing the functional evaluations are particularly valuable. In this context, while many examples exist (e.g., (Hanselman et al., 2019; Isenberg et al., 2020; Yoon et al., 2021; Lan and An, 2021; Sun et al., 2021)) there remains a large scope for developing and applying techniques such as mixed-integer linear/nonlinear programming, Bayesian optimization, derivative-free optimization, reinforcement learning, and evolutionary algorithms to tackle highly nonlinear and expensive to evaluate functions. Key here is to be able to accommodate large-scale nonlinear constraints that are motivated from physics. For instance, a worthy problem in this context is designing alloy nanoparticles (the shape, size, and the composition) for maximizing the reaction rate for a catalytic reaction (Jinnouchi and Asahi, 2017) wherein the energetic stability of the nanoparticle is a metric that needs to be computed using density functional theory (DFT). Such methods may also be able to accommodate the fact that a constraint function could be computed via two different methods of differing accuracy and cost (e.g., a cheap but less accurate model vs. expensive but accurate model or experiment) and an optimal solution needs to be obtained within an overall computing budget.

Related to the previous two subsections, such optimization methods may also be employed to learn the parameters of costly first principles models from experimental data. For instance, kinetic Monte Carlo simulations that solve the stochastic chemical master equation pertaining to reactions on surfaces are high fidelity solutions of reaction kinetics, however, they are also expensive (often a factor 1000 or more compared to deterministic ordinary differential equations) and computing sensitivities requires numerical differentiation. Bayesian optimization or other derivative-free approaches that can accommodate constraints would be a numerically efficient way to estimate the parameters of such kinetic models (Gao et al., 2018).

2.3. Better training methods for deep learning

Neural networks, especially graph neural networks (GNNs), are among the most popular surrogate functions in catalysis and materials science because they directly relate structure (atom identity, connections, and positions) to energy or bulk material property. The graph convolutional layers learn the underlying embedding of these materials thereby obviating the need for the user to handcraft data representations (Xie and Grossman, 2018). While these models have been shown to be highly flexible in training potential energy functions for molecular and material properties, computing a universal GNN for computing the binding energies of small adsorbates on alloy surfaces has proven to be particularly challenging. One unexplored area in GNNs in general and in the context of catalysis, in particular, is the adoption of better training algorithms for such models. Specifically, stochastic gradient descent techniques such as Adam are commonly employed to train these models; however, these methods do not include the second-order derivative (Hessian) information. Hessian-augmented techniques (Yao et al., 2020; Jahani et al., 2021) techniques have recently been shown to be promising for neural networks, however, their performance on GNNs for molecular and material applications remains to be systematically analyzed.

Physically motivated constraints can be incorporated into machine learned potentials trained on first principles data, either in the design of the network itself (such as translational/rotational invariance, equivariance, or physically motivated fingerprints) or as constraints (e.g., force is equal to the negative of the derivative of the energy). PSE techniques can play a major role in innovations in the latter scenario, where constraints are explicitly incorporated (Berahas et al., 2021) or approximated via rigorous mathematical programming techniques (Fioretto et al., 2021), rather than *weakly* embedded via hyperparameter tuning of regularization terms in the objective.

2.4. Uncertainty quantification

Data driven models and physics-based models often provide a point estimate of a property, e.g., the binding energy of a species on a surface without quantifying the uncertainty of that estimate. As a result, the reliability of the prediction is not clear in advance. In this context, there still remains tremendous scope for PSE experts to develop efficient methods to: (i) carry out Bayesian inference of first principles models to learn posterior distributions of parameters from data (Savara and Walker, 2020) and (ii) quantify the uncertainty of machine learned models, e.g., neural networks (Hirschfeld et al., 2020). A challenge in Bayesian inference is the ability to incorporate domain constraints during Monte Carlo sampling moves. For instance, Bayesian inference of kinetic models needs to preserve thermodynamic consistencies, which often result in linear inequalities that must be satisfied. Further, computationally efficient schemes to enable the identification of the multiple modes of the posterior of the kinetic parameters (i.e., multiple peaks, each corresponding to a distinct local optimum) are also needed (Galagali and Marzouk, 2015).

2.5. Learning interpretable governing equations from data

The ability to acquire high-resolution temporally and spatially varying data such as concentration profiles, temporal analysis of products (TAP), operando spectroscopy data under reaction conditions, and in situ or in operando microscopy data of the evolution of a material system enhances the possibility of learning the governing equations directly from experiments rather than invoking approximate physical models (Chen et al., 2022b). Such governing equations have to be explainable, i.e., the individual terms must be easy to ascribe to some expected phenomena (e.g., linking a term to a plausible reaction) and domain informed (satisfy mass balance, laws of thermodynamics, etc.). Several methods such as SINDy (Brunton et al., 2016), Eureka (Schmidt and Lipson, 2009), SISSO (Ouyang et al., 2018), AI-DARWIN (Chakraborty et al., 2021), ALAMO (Cozad et al., 2014) etc. have been proposed, but there remains tremendous scope for developing methods and software for learning domain-informed data-driven models from high-dimensional, noisy data from disparate sources (of differing fidelity).

2.6. Handling data imperfections

Neural networks can serve as excellent data-driven surrogate models when data are plentifully available from a single source. However, often catalysis and materials problems suffer from imperfect data, i.e., data that are sparse, have differing fidelity, and originate from disparate sources; typical approaches of training a neural network or any other flexible ML model would lead to overfitting in such cases. For instance, computing accurate binding energies of adsorbates on the catalyst surface or calculating electronic properties of materials such as band gaps correctly requires methods beyond the common functionals of density functional theory. Building accurate data-driven models of such properties is challenging because the underlying datasets themselves are either sparse or large, but inaccurate. In such cases, concepts such as transfer learning and multitask learning can be employed on (i) fused datasets of the same property measured/computed with differing resolution or accuracy or (ii) datasets of related properties so that essential features can be learned and transferred between models. However, several challenges remain that PSE experts are well-positioned to address: (i) What is the best strategy for transfer learning, e.g., model control or parameter control? (ii) How related should the datasets be so that transfer learning does not result in negative learning or overfitting? (iii) How to balance the training of different tasks in multitask learning? (iv) How to acquire data at different levels to minimize training data requirements?

2.7. Generative modeling

As mentioned above, the search space for most material discovery/design problems is huge. This can be true even for small molecules, depending on how many atoms are involved, how many different elements are considered, and what kinds of bonds are allowed. Naturally, the problem is much worse for large molecules. In addition, the design of crystalline materials, e.g., zeolites, MOFs, is more complicated as such materials have more parameters (e.g., angles) and large choices of atoms as well as branching structures. It is further complicated because of periodicity in crystals and the non-uniqueness of the unit cell selections. No matter how many samples are provided, it is likely that they represent only a miniscule fraction of the possible choices. For ML to lead to the discovery of truly innovative materials, it needs to suggest samples that are outside of the given data domain and yet are plausible. Such ML falls in the category of generative modeling.

Generative modeling Anstine and Isayev (2023) is a kind of unsupervised learning that learns the regularities or patterns (the probability distribution) in the input data such that the model can be used to generate new samples that plausibly could have been drawn from the original dataset. This is in contrast to ML-based predictive modeling where the goal is to discover the probability distributions of given samples and relationships among variables. Among generative modeling methods, Generative Adversarial Networks (GANs) use deep learning methods, such as convolutional neural networks and frame the unsupervised learning task as a supervised learning problem with two sub-models: the generator model that generates new examples and the discriminator model that classifies a generated sample as either real or fake. As the two models are trained, the generator model becomes increasingly clever and is able to "fool" the discriminator, i.e., it generates data that are indistinguishable from the real data by the discriminator. The training continues until the generator is able to generate samples that are indistinguishable by the discriminator from the real samples - in other words, the probability of the discriminator determining that a true sample is true is 0.5. GANs have been used across a range of problem domains, most notably in image-to-image translation and photos of fake objects and people. They have also seen applications in material design, for example in the design of crystalline materials such as zeolites (Kim et al., 2020). While they provide some exciting new tools to discover materials that have not been conceived by humans, further learning through more applications and follow-up research is needed. For example, many materials generated by such methods may not be synthesized in labs and therefore cannot be tested beyond simulation. Incorporating the consideration of physical synthesizability into the generator or discriminator model is an interesting open problem. Recently, large language models that utilize revolutionary encoder-decoder-based transformer models have disrupted the machine learning field. This transformer technology has also shown promising results in the context of reaction/synthesis prediction (Schwaller et al., 2019; Mann and Venkatasubramanian, 2021). Such tools have also been used in material property prediction (Kang et al., 2023), protein structure prediction (Jumper et al., 2021), data extraction from scientific publications (Polak et al., 2023), etc.

2.8. Automated/high throughput experimentation

As high throughput experiments and automated synthesis set ups are becoming more common in catalysis and materials design, data-driven algorithms are required to (i) decide what experiments to perform, (ii) learn models from the experiments, and (iii) make decisions in a closed-loop fashion to maximize a desired property. PSE ideas from the design of experiments and active learning along with concepts such as Bayesian optimization (Shields et al., 2021; Gonzlez and Zavala, 2023) and reinforcement learning (Bennett and Abolhasani, 2022) will play an important role here. In such a context, opportunities remain to design sampling strategies that balance exploration of the experimental space vs. exploiting the currently available data while deciding the next set of experiments. In particular, opportunities arise in automating multiple types of experimentation (and computation) to generate and leverage multimodal data (of differing accuracy and cost) to minimize the overall cost of search campaigns. One could extend such an idea to search for materials/reaction conditions to optimize process-level metrics (e.g., overall cost or carbon footprint) rather than a specific material property (e.g., product selectivity).

3. Machine learning in PSE

Data Science, including ML and AI, are becoming widely adopted in all areas of PSE research and industrial applications, including data analytics, process control, process design, multiscale modeling, and optimization. Many of the topics discussed in the previous section are indeed subjects of research in the PSE community. The following subsections summarize the prevalent themes that emerged from the discussion of ML and AI Challenges in PSE from an industrial data analytics and process control and optimization perspective.

3.1. Industrial data analytics

Several successful industrial case studies were presented at the conference, within the broad theme of Chemometrics, i.e. the combination of analytical chemistry and chemical engineering with data science methods rooted in AI and statistics (Qin and Chiang, 2019). These case studies include the use of deep neural networks for image classification, reinforcement learning, natural language processing, hybrid modeling, predictive formulation, and sensor fusion for monitoring. It was noted that the application of linear methods such as PCA/PLS for real-time process monitoring has pockets of successes for certain unit operations, yet less so for plantwide monitoring. Neural networks offer an alternative but they require more data and involve more parameters. Very recently, transformer-based models have been used for the autocompletion of Process Flow Diagrams (PFDs) (Vogel et al., 2023) and the translation of PFDs to Process and Instrumentation Diagrams (P&IDs) (Hirtreiter et al., 2023). In addition, historical operating data can play an important role in bridging and connecting the different time scales in multiscale, integrated decision making (e.g., the integration of scheduling/control and planning/scheduling) (Tsay and Baldea, 2019). Multiscale modeling in general involves the integration of information and processes across different spatial and temporal scales. ML can play a useful role in bridging scales in multiscale modeling by providing efficient and accurate representations of complex systems. For example, ML models, such as neural networks or Gaussian processes, can be trained to approximate complex and computationally expensive simulations. These surrogate models can be used to replace detailed simulations at certain scales, making the overall multiscale model more computationally efficient. ML methods can also be employed to develop algorithms that upscale or downscale information from one scale to another. For instance, if detailed information is available at a fine scale, ML models can be trained with the data to predict the behavior or properties at a coarser scale. For a comprehensive overview of this topic see (Ingolfsson et al., 2023).

Common misconceptions on the benefits of big data were also discussed. These include: (i) higher dimensional data are always better (truth: they may exhibit counter-intuitive phenomena); (ii) with more data we should be able to get better models (truth: not necessarily, since data can be noisy, with missing values, outliers, etc. and highly localized to one or few operating points); (iii) ML algorithms with big data will outperform algorithms with small data (truth: not necessarily so, especially when there are no high quality big data available). The bases of these misconceptions can be traced to the following facts: (i) historical data in industry have outliers and other anomalies, (ii) they lack persistent excitation, a pre-requisite for estimating reliable models, (iii) data collected over several months or years are not necessarily representative of the same process conditions, and (iv) often, the data corresponding to events of interest are an extremely small percentage of the total data. For example, fault detection and diagnosis and predictive analytics problems are typically classification problems; for best results, the relevant classification algorithms need a similar number of data samples for each class. However, industrial data are usually highly imbalanced, with faults appearing in an extremely small portion ($<< 1\%$) of the data. It is either too expensive or impossible to generate faulty data in real time. If such data are not adequately pre-processed, ML models could provide high accuracy during training and fail during testing, rendering the models impractical. From these considerations, several challenges and opportunities were pointed out:

- Obtain "informative" or persistently excited data in a plant environment. For a specific target application, what additional experiments and sensors are needed to this end? How can we characterize and quantify data quality and then translate this information to model uncertainty?

- Develop hybrid models (combining first principles and data) for a broad range of domains/scales (R&D, Manufacturing, Supply Chain, etc.).

- Explore new methods for system identification and process monitoring at a plantwide scale.

- Safety and reliability concerns for end users and concerns related to liability and reputation loss for technology providers remain a challenge for the adoption of deep learning solutions in plant environments, particularly for closed-loop solutions. Currently, the acceptance of the models depends on how well the model predictions are aligned with domain knowledge and their criticality to operations and profit.

- Countering the previous point, gaining trust to validate, implement, and sustain deep and reinforcement learning models in plant environments as well as interpreting deep learning results, for example by leveraging the concept of explainable AI and the associated methods and tools, are potential opportunities.

In addition to these technical challenges and opportunities, "cultural" ones were also emphasized.

- Data science education in chemical engineering at current time is limited at best. There is an urgent need to incorporate data science into the chemical engineering curriculum, exposing the students to foundational statistics, ML, and programming concepts, as well as their domain applications (Proctor and Chiang, 2023).

- There is also a need to develop the corporate AI workforce and culture. There is a scarcity of data science literate engineers in the industrial workforce. It is also difficult for upper level managers to fully appreciate the scope and potential of the application of data science in their companies. Plant leadership needs to communicate this potential clearly to process engineers. Data scientists need to motivate plant personnel to implement and adopt new ML tools. There needs to be "continuing education" opportunities for both company executives and data scientists in industry. PSE faculty can play an important role in developing the appropriate platforms and mechanisms and also transferring data science tools and experiences to industry.

- The ML, AI and data science landscape of commercial tools, vendors and applications is exploding. Partnering with the appropriate partners is a major challenge for companies and universities alike.

- Finally, it is important to mention the relevance of large language models, such as ChatGPT, in relation to the use of ML in the industry. Although the use of these generative AI tools promises to significantly increase productivity, allowing the ingestion of sensitive data for further training of open-access models creates risks related to the loss of intellectual property and competitiveness for industrial players. This is likely to drive either in-house custom developments on top of open-source models by industrial players themselves or the offering of closed solutions by the developers of the open-access models to individual industrial users. The success of the latter option will depend on establishing the necessary levels of trust between the parties.

3.2. Control

The conference generated many discussions on the role of data and ML in control, a subject of rich research activity in recent years (Tang and Daoutidis, 2022). Efficient use of data in control can be a key enabler of a transition from automation to autonomy in the process industries. The most direct way of incorporating learning and data in control is in the dynamic modeling of a process (Esche et al., 2022). Standard system identification methods are essentially data-driven methods but their industrial use are usually limited to linear dynamics. Neural networks (especially recurrent neural networks and neural ODEs) as universal approximators offer the potential to capture nonlinear functions and could in principle be incorporated in model predictive control (MPC) algorithms (Ren et al., 2022; Chen et al., 2022a; Lanzetti et al., 2019). Moreover, new transformer architectures have shown promising results for learning dynamic systems (Sitapure and Kwon, 2023). The central challenge however is that there is no theoretical basis to guide the amount of data needed to learn the behavior of a nonlinear dynamic system (Van Waarde et al., 2020), and indeed such data, i.e., data that sufficiently cover the dynamic operating range of the nonlinear system which neural networks represent, may be difficult or even impossible to obtain in practice. It is also challenging to obtain stability and performance guarantees or establish physical interpretability in the models and the control actions. Physics informed neural networks and regularization approaches to improve interpretability are possible avenues to mitigate these challenges. Gaussian processes as nonparametric statistical models, often combined with Bayesian optimization, are an alternative approach that can address the intricate balance between exploration and exploitation towards stability guarantees and closed-loop performance improvement (Makrygiorgos et al., 2022; del Rio Chanona et al., 2021; Bradford et al., 2020).

Another emerging application of learning in control, especially for optimization-based control methods that typically yield an implicitly-defined control law, e.g., nonlinear model predictive control, is approximating controllers that are computationally expensive to evaluate in real-time Mesbah et al. (2022). The key notion of these approaches is to learn an explicit and cheap-to-evaluate control law using open- or closed-loop simulation data generated by solving the original control law. As such, approximate controllers can be useful for control of large-scale systems Kumar et al. (2021), or embedded control applications for fast-sampling systems Karg and Lucia (2020). An important open challenge in this direction is how to achieve efficient approximate controllers that can readily adapt to changing situations.

A very different, less investigated approach involves not learning the dynamic model itself, but rather some control relevant information that is simpler than the entire process model. Examples include learning the optimal value function, policy in a reinforcement learning (RL) framework (Shin et al., 2019; Nian et al., 2020; Spielberg et al., 2019), transfer learning and batch process optimization (Petsagkourakis et al., 2020; Yoo et al., 2021), or learning dissipativity functions from input/output data (Tang and Daoutidis, 2021). For RL, in particular, policy search or actor-critic methods look to directly optimize parameters of an explicitly parametrized control policy such as a (deep) neural network. Yet, such approaches also rely on exciting the plant in an active manner, and the amount of data needed to guarantee sufficiently dense sampling and accurate learning again needs to be characterized. In this sense, high-fidelity first principles models may be critical to ensure sufficient off-line learning spanning the operating space before taking it to on-line for further adjustment. In addition, by embedding physics into the parametrization of a control policy to be learned, the size of the policy search space can be greatly reduced, while devising easier to implement and more interpretable control policies in a data efficient manner (Paulson et al., 2023).

3.3. Data representations

Data representations are fundamental as they serve as a bridge between raw data and ML models to inject domain knowledge into otherwise black-box frameworks. The PSE community is uniquely positioned to address suitable data representations due to specific-domain knowledge and knowledge in mathematical modeling (Schweidtmann et al., 2021). The choice of representation encapsulates crucial information about the inherent characteristics of the objects under consideration. Some notable examples are SMILES strings, molecular graphs, or three-dimensional coordinates, for example, molecular representations (Wigh et al., 2022; David et al., 2020), string/graph-representations of flowsheets (Gao and Schweidtmann, 2023), representations for infinite-dimensional optimization (Pulsipher et al., 2022), prior and kernel functions on Bayesian optimization (del Rio Chanona et al., 2021; Deshwal et al., 2020), and (convolutional) neural networks for feature extraction in manufacturing (Jiang et al., 2022).

For applications, where human interpretability is important, the data objects should be readable to both the human and machine. For example, let us say we want to predict solubility between two substances if we get a high solubility but we encode the object as a hashed fingerprint or as information bits, the human will not understand why the model made the prediction. If however, we encoded the molecule as a group of functional groups, interpretability methods could be used to see if the model identifies the functional group as a major contribution (Schwaller et al., 2021). Thanks to the data representation, the human can understand the prediction and trust the model. As ML is more and more intertwined with chemical and process systems engineering, data representations will become not only more important but an engineering necessity.

Moreover, when dealing with time series data, the manner in which data is organized for training becomes pivotal, impacting not just performance but also the interpretability of the model. Take, for instance, time series data derived from cyclic operations used in monitoring and prediction. In such cases, structuring the data as an array with distinct dimensions for time and cycle number, akin to the approach in multi-way PCA, proves beneficial. Recent advancements in time-series data representation, exemplified in the context of predicting the remaining useful life of lithium-ion batteries, showcase the advantages of arranging voltage, current, and temperature profiles from early charge/discharge cycles in this manner. This configuration enables the training of a 2-D CNN or a hybrid model combining CNN with recurrent neural network (RNN), demonstrating marked improvements and benefits in model interpretability and data requirement (Lee and Lee, 2023).

3.4. Optimization

It is generally recognized that there are many decision-making problems that cannot be solved reliably with current optimization methods within reasonable time constraints. The underlying causes can include the shear size of the problems, the number and type (integer, continuous) of decision variables, the need to consider decision variables and constraints occurring at multiple timescales, and the presence of nonlinear/non-convex terms. Examples of such problems include plantwide optimization, planning, scheduling and their integration, and more generally enterprise-wide optimization.

For such problems, ML can be used to train surrogate models and embed them in the optimization algorithms (Sansana et al., 2021; Bradley et al., 2022; Schweidtmann and Mitsos, 2019; Bhosekar and Ierapetritou, 2018). Challenges to this end include: (i) the interoperability of different models, (ii) the ability to generate these models automatically, (iii) quantifying the extrapolation capabilities of these models, (iv) determining what parts of a model should be replaced with a surrogate, (v) addressing the need to adapt/update model when we change the process, (vi) developing stochastic surrogate models for optimization under uncertainty, (vii) developing surrogate models for global and bilevel optimization, (viii) ensuring constraint satisfaction during training, and (ix) choosing the most suitable surrogate model. Computational tools to accelerate embedding of trained ML models within optimization formulations, such as the OMLT framework (Ceccon et al., 2022), can expedite solutions and enable comparative studies to answer many of the identified open challenges.

A different avenue for using ML in optimization is to use it to accelerate or improve the computational performance of existing solution algorithms (e.g., optimizing heuristics to accelerate genetic algorithms, tuning the parameters of optimization solvers using Bayesian optimization, finding optimal decompositions, generating high quality cuts, initializing in an optimal way, pre-fixing or eliminating binary variables, finding global solutions, etc.) and to determine which algorithm among several is the most suitable one for a given optimization problem (see e.g., (Bengio et al., 2021; Chen et al., 2021; Cappart et al., 2023) as well as (Harjunkoski et al., 2020; Mitrai and Daoutidis, 2023b,c,a)). Important problems to be addressed to this end include: (i) representations of the optimization problems and their features that enable using them as inputs to ML models, (ii) automation of the solution and learning methods to allow for efficient screening and learning, (iii) availability of large numbers of benchmark problems to be used for training and testing, and (iv) interpretability of the realized improvements.

ML can also be used with classical PSE tools (e.g. mathematical programming) to tackle problems that are previously beyond reach. One of the outstanding problems in PSE is integrating decisions occurring at different layers of the vertical decision hierarchy. For example, planning and scheduling occur over different time scales and time horizon of very different lengths, but are inherently linked. The usual practice of coarse-graining the fast time-scale layer and incorporating the coarsened model into the optimization of the upper layer can result in mismatched production plans that cannot be executed. Uncertainties and unexpected disturbances aggravate this further. To address such problem, reinforcement learning can be combined with traditional optimization (e.g. mathematical programming). An illustrative approach was shown in (Shin et al., 2017; Shin and Lee, 2019) where the upper layer capacity decisions are made through reinforcement learning with data obtained by simulating the lower layer at its time scale (e.g., hourly) over the time horizon set by the capacity planning layer (e.g., years or decades). The lower layer decisions are made by linear programming, which becomes a part of the simulation to generate the data needed for reinforcement learning in the upper layer. The upper layer adopts a Markov Decision Process description of the overall system and therefore has the flexibility of accommodating various stochastic uncertainty descriptions (e.g., Markov processes). Uncertainties occurring in the lower layer at the faster time-scale can be handled by scenarios or Monte-Carlo simulation to allow recourse actions.

Finally, optimization itself lies at the heart of ML methods. This brings up an opportunity for the PSE community to contribute to better learning methods based on optimization rather than heuristics, especially for small data problems that arise in chemical engineering. Additional opportunities include using optimization formulations to enforce constraints (such as physics-based) and training models that have features that facilitate their subsequent use in optimization.

4. Conclusions

The main conclusion from the FIPSE session, as summarized in this paper, was that ML in the context of PSE can have a transformational impact on catalysis and materials design, as well as on process operations and automation. Numerous domain-specific challenges need to be overcome to this end. Whereas domain knowledge is essential to guide method and software development, data science expertise is also necessary to deal with the ever increasing complexity of data structures and algorithms. These considerations suggest an exciting opportunity for the PSE community - both in academia and in industry - to lead in meeting these outstanding challenges. For industrial practitioners there are scaling and end-to-end deployment challenges, which are mostly practical in nature and outside the scope of the academic community, whereas the academic community is focusing mostly on new methods and algorithms and their properties, which is beyond the interest of industrial practitioners. Since ML relies on data, focusing on industrial data could bring industry and academia together and foster closer collaborations.

CRediT Authorship Contribution Statement

Prodromos Daoutidis: conceptualization, funding acquisition, project administration, supervision, writing – original draft, writing – review & editing. **Jay H. Lee**:

conceptualization, funding acquisition, project administration, supervision, writing – original draft, writing – review & editing. **Srinivas Rangarajan**: conceptualization, writing – original draft, writing – review & editing. **Leo Chiang**: conceptualization, writing – review & editing. **Bhushan Gopaluni**: conceptualization, writing – review & editing. **Artur M. Schweidtmann**: conceptualization, writing – review & editing. **Iiro Harjunkoski**: conceptualization, writing – review & editing. **Mehmet Mercangöz**: conceptualization, writing – review & editing. **Ali Mesbah**: conceptualization, writing – review & editing. **Fani Boukouvala**: conceptualization, writing – review & editing. **Fernando V. Lima**: conceptualization, writing – review & editing. **Antonio del Rio Chanona**: conceptualization, writing – review & editing. **Christos Georgakis**: conceptualization, writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

There is no data reported or used in this manuscript.

Acknowledgments

PD would like to acknowledge financial support by NSF-CBET.

Andersson, J. A., Gillis, J., Horn, G., Rawlings, J. B., and Diehl, M. (2019). Casadi: a software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11:1–36.

Anstine, D. M. and Isayev, O. (2023). Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society*, 145(16):8736–8750.

Bengio, Y., Lodi, A., and Prouvost, A. (2021). Machine learning for combinatorial optimization: a methodological tour d?horizon. *European Journal of Operational Research*, 290(2):405–421.

Bennett, J. A. and Abolhasani, M. (2022). Autonomous chemical science and engineering enabled by self-driving laboratories. *Current Opinion in Chemical Engineering*, 36:100831.

Berahas, A. S., Curtis, F. E., O'Neill, M. J., and Robinson, D. P. (2021). A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient jacobians. *arXiv preprint arXiv:2106.13015*.

Bhosekar, A. and Ierapetritou, M. (2018). Advances in surrogate based modeling, feasibility analysis, and optimization: A review. *Computers & Chemical Engineering*, 108:250–267.

Bradford, E., Imsland, L., Zhang, D., and del Rio Chanona, E. A. (2020). Stochastic data-driven model predictive control using Gaussian processes. *Comput. % Chem. Eng.*, 139:106844.

Bradley, W., Kim, J., Kilwein, Z., Blakely, L., Eydenberg, M., Jalvin, J., Laird, C., and Boukouvala, F. (2022). Perspectives on the integration between first-principles and data-driven modeling. *Computers & Chemical Engineering*, page 107898.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937.

Cappart, Q., Chatelat, D., Khalil, E. B., Lodi, A., Morris, C., and Velickovic, P. (2023). Combinatorial optimization and reasoning with graph neural networks. *Journal of Machine Learning Research*, 24(130):1–61.

Ceccon, F., Jalving, J., Haddad, J., Thebelt, A., Tsay, C., Laird, C. D., and Misener, R. (2022). Omlt: Optimization & machine learning toolkit. *J. Mach. Learn. Res.*, 23(1).

Chakraborty, A., Sivaram, A., and Venkatasubramanian, V. (2021). Ai-darwin: A first principles-based model discovery engine using machine learning. *Computers & Chemical Engineering*, 154:107470.

Chen, S. W., Wang, T., Atanasov, N., Kumar, V., and Morari, M. (2022a). Large scale model predictive control with neural networks and primal active sets. *Automatica*, 135:109947.

Chen, T., Chen, X., Chen, W., Heaton, H., Liu, J., Wang, Z., and Yin, W. (2021). Learning to optimize: A primer and a benchmark. *arXiv preprint arXiv:2103.12828*.

Chen, Y.-Y., Kunz, M. R., He, X., and Fushimi, R. (2022b). Recent progress toward catalyst properties, performance, and prediction with data-driven methods. *Current Opinion in Chemical Engineering*, 37:100843.

Chiang, L., Lu, B., and Castillo, I. (2017). Big data analytics in chemical engineering. *Annu. Rev. Chem. Biomol. Eng.*, 8:63–85.

Cozad, A., Sahinidis, N. V., and Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6):2211–2227.

David, L., Thakkar, A., Mercado, R., and Engkvist, O. (2020). Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12(1):56.

del Rio Chanona, E. A., Petsagkourakis, P., Bradford, E., Graciano, J. A., and Chachuat, B. (2021). Real-time optimization meets bayesian optimization and derivative-free optimization: A tale of modifier adaptation. *Comput. & Chem. Eng.*, 147:107249.

Deshwal, A., Belakaria, S., and Doppa, J. R. (2020). Mercer features for efficient combinatorial bayesian optimization. *CoRR*, abs/2012.07762.

Döppel, F. A. and Votsmeier, M. (2022). Efficient machine learning based surrogate models for surface kinetics by approximating the rates of the rate-determining steps. *Chemical Engineering Science*, 262:117964.

Esche, E., Weigert, J., Rihm, G. B., Göbel, J., and Repke, J.-U. (2022). Architectures for neural networks as surrogates for dynamic systems in chemical engineering. *Chem. Eng. Res. Des.*, 177:184–199.

Fioretto, F., Van Hentenryck, P., Mak, T. W. K., Tran, C., Baldo, F., and Lombardi, M. (2021). Lagrangian duality for constrained deep learning. In Dong, Y., Ifrim, G., Mladenić, D., Saunders, C., and Van Hoecke, S., editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pages 118–135, Cham. Springer International Publishing.

Galagali, N. and Marzouk, Y. M. (2015). Bayesian inference of chemical kinetic models from proposed reactions. *Chemical Engineering Science*, 123:170–190.

Gao, H., Waechter, A., Konstantinov, I. A., Arturo, S. G., and Broadbelt, L. J. (2018). Application and comparison of derivative-free optimization algorithms to control and optimize free radical polymerization simulated using the kinetic monte carlo method. *Computers & Chemical Engineering*, 108:268–275.

Gao, Q. and Schweidtmann, A. M. (2023). Deep reinforcement learning for process design: Review and perspective. *arXiv preprint arXiv:2308.07822*.

Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A., Chae, H. S., Einzinger, M., Ha, D.-G., Wu, T., et al. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials*, 15(10):1120–1127.

Gonzlez, L. D. and Zavala, V. M. (2023). New paradigms for exploiting parallel experiments in bayesian optimization. *Computers & Chemical Engineering*, 170:108110.

Gusmão, G. S., Retnanto, A. P., Da Cunha, S. C., and Medford, A. J.

(2022). Kinetics-informed neural networks. *Catalysis Today*.

Hanselman, C. L., Zhong, W., Tran, K., Ulissi, Z. W., and Gounaris, C. E. (2019). Optimization-based design of active and stable nanostructured surfaces. *The Journal of Physical Chemistry C*, 123(48):29209–29218.

Harjunkoski, I., Ikonen, T., Mostafaei, H., Deneke, T., and Heljanko, K. (2020). Synergistic and intelligent process optimization: First results and open challenges. *Industrial and Engineering Chemistry Research*, 59(38):16684–16694.

Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., and Coley, C. W. (2020). Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780.

Hirtreiter, E., Schulze Balhorn, L., and Schweidtmann, A. M. (2023). Toward automatic generation of control structures for process flow diagrams with large language models. *AIChE Journal*, page e18259.

Ingolfsson, H. I., Bhatia, H., Aydin, F., Oppelstrup, T., Lopez, C. A., Stanton, L. G., Carpenter, T. S., Wong, S., Di Natale, F., Zhang, X., Moon, J. Y., Stanley, C. B., Chavez, J. R., Nguyen, K., Dharuman, G., Burns, V., Shrestha, R., Goswami, D., Gulten, G., Van, Q. N., Ramanathan, A., Van Essen, B., Hengartner, N. W., Stephen, A. G., Turbyville, T., Bremer, P.-T., Gnanakaran, S., Glosli, J. N., Lightstone, F. C., Nissley, D. V., and Streitz, F. H. (2023). Machine learning-driven multiscale modeling: Bridging the scales with a next-generation simulation infrastructure. *Journal of Chemical Theory and Computation*, 19(9):2658–2675. PMID: 37075065.

Isenberg, N. M., Taylor, M. G., Yan, Z., Hanselman, C. L., Mpourmpakis, G., and Gounaris, C. E. (2020). Identification of optimally stable nanocluster geometries via mathematical optimization and density-functional theory. *Molecular Systems Design & Engineering*, 5(1):232–244.

Jahani, M., Rusakov, S., Shi, Z., Richtárik, P., Mahoney, M. W., and Takáč, M. (2021). Doubly adaptive scaled algorithm for machine learning using second-order information. *arXiv preprint arXiv:2109.05198*.

Jiang, S., Qin, S., Pulsipher, J. L., and Zavala, V. M. (2022). Convolutional neural networks: Basic concepts and applications in manufacturing. *arXiv preprint arXiv:2210.07848*.

Jinnouchi, R. and Asahi, R. (2017). Predicting catalytic activity of nanoparticles by a dft-aided machine-learning algorithm. *The journal of physical chemistry letters*, 8(17):4279–4283.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kang, Y., Park, H., Smit, B., and Kim, J. (2023). A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nature Machine Intelligence*, 5(3):309–318.

Karg, B. and Lucia, S. (2020). Efficient representation and approximation of model predictive control laws via deep learning. *IEEE Transactions on Cybernetics*, 50(9):3866–3878.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.

Kim, B., Lee, S., and Kim, J. (2020). Inverse design of porous materials using artificial neural networks. *Science advances*, 6(1):eaax9324.

Kumar, P., Rawlings, J. B., and Wright, S. J. (2021). Industrial, large-scale model predictive control with structured neural networks. *Comput. & Chem. Eng.*, 150:107291.

Lan, T. and An, Q. (2021). Discovering catalytic reaction networks using deep reinforcement learning from first-principles. *Journal of the American Chemical Society*, 143(40):16804–16812.

Lanzetti, N., Lian, Y. Z., Cortinovis, A., Dominguez, L., Mercangöz, M., and Jones, C. (2019). Recurrent neural network based mpc for process industries. In *2019 18th European Control Conference (ECC)*, pages 1005–1010. IEEE.

Lee, J. H., Shin, J., and Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Comput. & Chem. Eng.*, 114:111–121.

Lee, J. W. and Lee, J. H. (2023). Simultaneous extraction of intra- and inter-cycle features for predicting lithium-ion battery's knees using convolutional and recurrent neural networks. *ChemRxiv.Cambridge: Cambridge Open Engage.*

Lejarza, F., Koninckx, E., Broadbelt, L. J., and Baldea, M. (2023). A dynamic nonlinear optimization framework for learning data-driven reduced-order microkinetic models. *Chemical Engineering Journal*, 462:142089.

Makrygiorgos, G., Bonzanini, A. D., Miller, V., and Mesbah, A. (2022). Performance-oriented model learning for control via multi-objective bayesian optimization. *Comput. & Chem. Eng.*, 162:107770.

Mann, V. and Venkatasubramanian, V. (2021). Predicting chemical reaction outcomes: A grammar ontology-based transformer framework. *AIChE Journal*, 67(3):e17190.

Matera, S., Schneider, W. F., Heyden, A., and Savara, A. (2019). Progress in accurate chemical kinetic modeling, simulations, and parameter estimation for heterogeneous catalysis. *Acs Catalysis*, 9(8):6624–6647.

Mesbah, A., Wabersich, K. P., Schoellig, A. P., Zeilinger, M. N., Lucia, S., Badgwell, T. A., and Paulson, J. A. (2022). Fusion of machine learning and mpc under uncertainty: What advances are on the horizon? In *American Control Conference*, pages 342–357. IEEE.

Mitrai, I. and Daoutidis, P. (2023a). Computationally efficient solution of mixed integer model predictive control problems via machine learning aided benders decomposition. *arXiv preprint arXiv:2309.16508*.

Mitrai, I. and Daoutidis, P. (2023b). A graph classification algorithm to determine when to decompose optimization problems. In *Proceedings of the 33rd European Symposium on Computer Aided Process Engineering (ESCAPE33)*, pages 655–661.

Mitrai, I. and Daoutidis, P. (2023c). Learning to initialize generalized benders decomposition via active learning. In *Proceedings of FOCAPO/CPC, San Antonio, Texas*.

Nian, R., Liu, J., and Huang, B. (2020). A review on reinforcement learning: Introduction and applications in industrial process control. *Comput. & Chem. Eng.*, page 106886.

Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., and Ghiringhelli, L. M. (2018). Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):083802.

Paulson, J. A., Sorourifar, F., and Mesbah, A. (2023). A tutorial on derivative-free policy learning methods for interpretable controller representations. In *American Control Conference (ACC)*, pages 1295–1306. IEEE.

Petsagkourakis, P., Sandoval, I., Bradford, E., Zhang, D., and del Rio-Chanona, E. (2020). Reinforcement learning for batch bioprocess optimization. *Computers & Chemical Engineering*, 133:106649.

Pistikopoulos, E. N., Barbosa-Povoa, A., Lee, J. H., Misener, R., Mitsos, A., Reklaitis, G. V., Venkatasubramanian, V., You, F., and Gani, R. (2021). Process systems engineering - the generation next? *Computers Chemical Engineering*, 147:107252.

Polak, M. P., Modi, S., Latosinska, A., Zhang, J., Wang, C.-W., Wang, S., Hazra, A. D., and Morgan, D. (2023). Flexible, model-agnostic method for materials data extraction from text using general purpose language models. *arXiv preprint arXiv:2302.04914*.

Proctor, M. and Chiang, L. (2023). Data science and digitalisation for chemical engineers. *IChemE The Chemical Engineers (TCE) magazine*, May issue:36–40.

Pulsipher, J. L., Zhang, W., Hongisto, T. J., and Zavala, V. M. (2022). A unifying modeling abstraction for infinite-dimensional optimization. *Computers & Chemical Engineering*, 156:107567.

Qin, S. J. and Chiang, L. H. (2019). Advances and opportunities in machine learning for process data analytics. *Comput. & Chem. Eng.*, 126:465–473.

Rangarajan, S., Maravelias, C. T., and Mavrikakis, M. (2017). Sequential-optimization-based framework for robust modeling and design of heterogeneous catalytic systems. *The Journal of Physi-*

cal Chemistry C, 121(46):25847–25863.

Ren, Y. M., Alhajeri, M. S., Luo, J., Chen, S., Abdullah, F., Wu, Z., and Christofides, P. D. (2022). A tutorial review of neural network modeling approaches for model predictive control. *Comput. & Chem. Eng.*, page 107956.

Sansana, J., Joswiak, M. N., Castillo, I., Wang, Z., Rendall, R., Chiang, L. H., and Reis, M. S. (2021). Recent trends on hybrid modeling for industry 4.0. *Computers & Chemical Engineering*, 151:107365.

Savara, A. and Walker, E. A. (2020). Chekipeuq intro 1: Bayesian parameter estimation considering uncertainty or error from both experiments and theory. *ChemCatChem*, 12(21):5385–5400.

Schmidt, M. and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. (2019). Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

Schwaller, P., Probst, D., Vaucher, A. C., Nair, V. H., Kreutter, D., Laino, T., and Reymond, J.-L. (2021). Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152.

Schweidtmann, A. M., Esche, E., Fischer, A., Kloft, M., Repke, J.-U., Sager, S., and Mitsos, A. (2021). Machine learning in chemical engineering: A perspective. *Chemie Ingenieur Technik*, 93(12):2029–2039.

Schweidtmann, A. M. and Mitsos, A. (2019). Deterministic global optimization with artificial neural networks embedded. *Journal of Optimization Theory and Applications*, 180(3):925–948.

Shields, B. J., Stevens, J., Li, J., Parasram, M., Damani, F., Alvarado, J. I. M., Janey, J. M., Adams, R. P., and Doyle, A. G. (2021). Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96.

Shin, J., Badgwell, T. A., Liu, K.-H., and Lee, J. H. (2019). Reinforcement learning – overview of recent progress and implications for process control. *Comput. Chem. Eng.*, 127:282–294.

Shin, J. and Lee, J. H. (2019). Multi-timescale, multi-period decision-making model development by combining reinforcement learning and mathematical programming. *Computers & Chemical Engineering*, 121:556–573.

Shin, J., Lee, J. H., and Realff, M. J. (2017). Operational planning and optimal sizing of microgrid considering multi-scale wind uncertainty. *Applied energy*, 195:616–633.

Sitapure, N. and Kwon, J. S.-I. (2023). Exploring the potential of time-series transformers for process modeling and control in chemical systems: an inevitable paradigm shift? *Chemical Engineering Research and Design*, 194:461–477.

Spielberg, S., Tulsyan, A., Lawrence, N. P., Loewen, P. D., and Gopaluni, R. B. (2019). Toward self-driving processes: A deep reinforcement learning approach to control. *AIChE J.*, 65(10):e16689.

Sun, S., Tiihonen, A., Oviedo, F., Liu, Z., Thapa, J., Zhao, Y., Hartono, N. T. P., Goyal, A., Heumueller, T., Batali, C., et al. (2021). A data fusion approach to optimize compositional stability of halide perovskites. *Matter*, 4(4):1305–1322.

Tang, W. and Daoutidis, P. (2021). Dissipativity learning control (DLC): theoretical foundations of input–output data-driven model-free control. *Systems & Control Letters*, 147:104831.

Tang, W. and Daoutidis, P. (2022). Data-driven control: Overview and perspectives. In *2022 American Control Conference (ACC)*, pages 1048–1064. IEEE.

Thebelt, A., Wiebe, J., Kronqvist, J., Tsay, C., and Misener, R. (2022). Maximizing information from chemical engineering data sets: Applications to machine learning. *Chemical Engineering Science*, 252:117469.

Tsay, C. and Baldea, M. (2019). 110th anniversary: Using data to bridge the time and length scales of process systems. *Industrial & Engineering Chemistry Research*, 58(36):16696–16708.

Van Waarde, H. J., Eising, J., Trentelman, H. L., and Camlibel, M. K. (2020). Data informativity: a new perspective on data-driven analysis and control. *IEEE Trans. Autom. Control*, 65(11):4753–4768.

Venkatasubramanian, V. (2019). The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE J.*, 65(2):466–478.

Vogel, G., Balhorn, L. S., and Schweidtmann, A. M. (2023). Learning from flowsheets: A generative transformer model for autocompletion of flowsheets. *Computers & Chemical Engineering*, 171:108162.

Wigh, D. S., Goodman, J. M., and Lapkin, A. A. (2022). A review of molecular representation in the age of machine learning. *WIREs Computational Molecular Science*, 12(5):e1603.

Xie, T. and Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301.

Yao, Z., Gholami, A., Keutzer, K., and Mahoney, M. W. (2020). Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE.

Yoo, H., Byun, H. E., Han, D., and Lee, J. H. (2021). Reinforcement learning for batch process control: Review and perspectives. *Annual Reviews in Control*, 52:108–119.

Yoon, J., Cao, Z., Raju, R. K., Wang, Y., Burnley, R., Gellman, A. J., Farimani, A. B., and Ulissi, Z. W. (2021). Deep reinforcement learning for predicting kinetic pathways to surface reconstruction in a ternary alloy. *Machine Learning: Science and Technology*, 2(4):045018.

Zavala, V. M. (2023). Outlook: How i learned to love machine learning (a personal perspective on machine learning in process systems engineering). *Industrial & Engineering Chemistry Research*, 62(23):8995–9005.

Zhong, M., Tran, K., Min, Y., Wang, C., Wang, Z., Dinh, C.-T., De Luna, P., Yu, Z., Rasouli, A. S., Brodersen, P., et al. (2020). Accelerated discovery of co2 electrocatalysts using active machine learning. *Nature*, 581(7807):178–183.