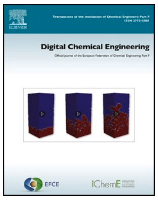
Digital Chemical Engineering 8 (2023) 100111

Contents lists available at ScienceDirect

Digital Chemical Engineering

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



Review Article

Computational applications using data driven modeling in process Systems: A review

Sumit K. Bishnu [a], Sabla Y. Alnouri [b], Dhabia M. Al-Mohannadi [c,*]

[a] Detect Technologies, IIT- Madras Research Park, Chennai 600113, India
[b] Gas Processing Centre, College of Engineering, Qatar University, P.O. Box 2713, Doha, Qatar
[c] Department of Chemical Engineering, Texas A&M University at Qatar, P.O 23874, Education City, Doha, Qatar

ARTICLE INFO

Keywords:
Modeling
Optimization
Machine learning
Data-driven modeling

ABSTRACT

Modeling and optimization of various processes enable more efficient operations and better planning activities for new process developments. With recent advances in computing power, data driven models, such as Machine Learning (ML), are being extensively applied in many areas of chemical engineering topics. Compared to mechanistic models that often do not reflect the realities of field conditions and the high costs associated with them, these techniques are relatively easier to implement. Data-driven models generated via ML techniques can be regularly updated, thereby giving an accurate picture of the system. Due to these inherent benefits, such tools are increasingly gaining a lot of traction in process systems. Even though data-driven models have the potential to be used as a replacement for traditional optimization tools that can be implemented in various process industries, it was found that applications of such models in process systems were quite limited to reactor modeling, molecular design, as well as safety, and relatability. The challenge still exists for data-driven modeling due to the lack of specialized tools tailored for macro systems and scale up. Most datasets were found to be derived from experimental studies which are limited in nature and only fit into microsystems. Hence, this paper provides a state of the art review on recent applications for data driven modeling research in process systems, and discusses the prominent challenges and future outlooks that were observed.

Introduction

Modeling, Control, and Optimization are essential components of any process industry, enabling stakeholders to incorporate changes for improving daily process operations. They also help industries achieve strategic objectives of financial savings and environmental and safety impact. Although the mathematical foundation has been in existence for quite some time, the past decade has provided a real boost in this field due to the vast advancement in computational methods combined with an ease of accessibility for such resources. A lot of recent work has focused on the application of data-driven models for accurately predicting a variety of target variables. Due to the benefits achieved, these models are being adopted for predictive maintenance, as well as for increasing automation and efficiency in daily operations (Cohen, 2021). Compared to mechanistic modeling, machine learning has proved to offer significant advantages in terms of flexibility, cost of computing, and speed of execution. Such benefits along with self-learning abilities that can recognize common patterns in input data set provide a lot of opportunities in the area of real-time optimization, and the modeling of assets (Dobbelaere et al., 2021).

Mechanistic models are generally very common to describe a given process or system. Such models are often dynamic and are developed based on descriptions of processes that exist within a given process or system. As such, they provide a structured mechanism that can generate predictions for process/system behavior and are often constructed using algebraic equations, ordinary differential equations (ODEs), and/or partial differential equations (PDEs). However, unlike mechanistic models that are developed based on a strong foundation in mathematical equations which are developed after detailed study, data driven models are black box models with individual components that interact with each other, without possessing any mathematical knowledge related to how a given system or process operates. Hence, they possess powerful automation features. It should be noted that automation often increases vulnerability to hacking and inappropriate data use. Hence, such tools may become especially risky in the process industry as the impact of any data misuse can be devastating. However, with tremendous

[*] Corresponding author.
E-mail address: dhabia.al-mohannadi@qatar.tamu.edu (D.M. Al-Mohannadi).

https://doi.org/10.1016/j.dche.2023.100111
Received 4 May 2023; Received in revised form 10 July 2023; Accepted 12 July 2023
Available online 13 July 2023

advancements being made in data security, a healthy balance can be achieved in process industries, whereby the benefits of using machine learning can outweigh the risks involved (Zhou et al., 2022).

In recent years, data-driven Machine Learning tools have been extensively used in the area of chemical engineering, especially in System Identification (Gao et al., 2022). Be it Model Predictive Control (MPC), Optimization, or Real Time Online models (RTO), such models were found to be applicable to an extensive range of applications (Pratap and Sardana, 2022). Machine Learning has changed the way data extraction and interpretation are performed, whereby traditional statistical techniques are replaced by automated generic methods and algorithms. In recent literature, Fuentes-Cortés et al. (Fuentes-Cortés et al., 2022) provide a general overview of machine learning algorithms that are used in chemical engineering. Shi et al. (Shi et al., 2021) discuss several applications for Artificial intelligence in process systems engineering. Bogle (Bogle, 2017) discusses the technical challenges that are often encountered by process systems engineers in developing tools and techniques (involving flexibility and uncertainty, responsiveness, agility, robustness, and security), as well as new modeling and mathematics paradigms. The paper aims to highlight the areas of the process industry which have witnessed enhanced usage of data-driven models. Rangel-Martinez et al. (Rangel-Martinez et al., 2021) provide a comprehensive overview on the latest advancements of ML applications within the manufacturing sectors, with particular emphasis onto applications that have a significant influence on sustainability and the environment, specifically in the domains of renewable energies (such as solar, wind, hydropower, and biomass), smart grids, catalysis industry, and power storage and distribution. Guan et al. (Guan et al., 2022) review the latest developments in applying ML to solid heterogeneous catalysis, as well as present many of the notable achievements in this field. Additionally, Guan et al. (Guan et al., 2022) explore the limitations and challenges that ML faces when applied to catalysis. Additionally, we explore prospective avenues for effectively leveraging ML in the design of solid heterogeneous catalysts. Ifaei et al. (Ifaei et al., 2023) review the fundamental principles, significant applications, and existing challenges of machine learning in sustainable energies, as well as delves deeper into these topics, by providing more advanced insights and analyses specifically tailored for experts in the fields of artificial intelligence and sustainable energy. Wu et al. (Wu et al., 2019) reviews MPC systems for nonlinear processes, employing an ensemble of Recurrent Neural Network (RNN) models to forecast nonlinear dynamics. Wu et al. (Wu et al., 2019) also discusses the initial construction of RNN models, which are essentially trained using a dataset derived from extensive open-loop simulations conducted within the desired operating region of the process, so as to ensure that such models accurately capture the process dynamics, and exhibit minimal modeling errors. Stephanopoulos (Stephanopoulos, 1990) reviews the potential benefits of enhanced knowledge-representation schemes and advanced reasoning control strategies in various aspects of process development, design, planning, scheduling, monitoring, analysis, and control. Forootan et al. (Forootan et al., 2022) review the underexplored domain of deep learning algorithms that possess significant problem-solving capabilities. Specifically, it focuses on Deep Learning (DL) algorithms such as RNN, Artificial Neural Network with Fuzzy INference System (ANFIS), Resource Based Network (RBN), Deep Belief Network (DBN), Wavelength Neural Network (WNN), and others, which have received comparatively less attention in previous studies. Their paper leverages knowledge discovery from research databases to gain insights into the current state and future prospects of ML and DL applications in energy systems.

ML plays a pivotal role in process systems engineering by enabling data-driven decision-making, enhanced process modeling, optimal operation, fault detection and diagnosis, control strategies, process optimization, and design. By leveraging ML techniques, PSE practitioners can achieve improved process performance, efficiency, safety, and sustainability. In effect, this paper tries to categorize the work done in the area of the process systems and present a brief highlight of notable achievements. In subsequent sections, we define the generic component of data-driven models and their applications in various fields. Various scientific papers published in this area and new areas that involve data-driven applications in the process systems have also been presented (Rattan et al., 2022). Moreover, an outlook on the observed challenges associated with the use of such models is also provided, while highlighting some of the key findings pertaining to the use of such models in process systems research.

Method of the literature search

Effectively identifying appropriate literature studies from publicly accessible citation databases was a very important step. First off, all relevant keywords that describe the subject matter were identified. After conducting trial keyword searches in citation databases such as Scopus and Google Scholar, which were selected as the citation databases for this study. This was determined based on their ease of use, comprehensive coverage, and quick citation update.

The second step involved narrowing down the keyword search. Machine learning in chemical engineering produced a wide spectrum of results. This included research articles, opinion papers, databases, review articles, and general introductions. Each subfield of chemical engineering from catalysis to supply chain optimization produced work applying machine learning. The objective of this work is to get a review of process systems engineering applications. In order to encompass all pertinent literature, it was necessary to utilize a variety of distinct search terms. In the preliminary searches, using narrowly defined keywords such as "machine learning + process systems engineering" failed to yield the relevant literature. Therefore, a combination of search keywords was utilized. Moreover, a filtering criterion was introduced to find the state of the application of the art and avoid duplication and older applications. Table 1 summarizes the literature criteria that was used for this review.

The next section explains the basics of machine learning algorithms that were deployed. Section 4 shows an overview of the bibliographic analysis obtained from a generalized literature search. Section 5 then summarizes the results of a more classified search that was conducted based on the above-mentioned criteria.

Technical background

The ability of the computer to identify/predict the output using programmed algorithms that receive and perform analysis on input data are essential for any data-driven models (Kotu and Deshpande, 2019). Initially, and before the emergence of such techniques, data analysis mainly relied on trial and error predictions, and those are increasingly impractical especially when large, heterogeneous data sets are involved. As such, automated search algorithms provide a great alternative to trial and error predictions, which becomes particularly useful for large-scale data analysis (Subasi, 2020). This section presents different techniques of Machine Learning which are being applied in the field of PSE. Apart from four broad categorizations, it also discusses the major algorithms

Table 1
Filtering of literature.

Criterion number	Criterion Description
1	English language articles
2	Studies indexed in google scholar and Scopus databases. Exclude Google search.
3	Peer-reviewed journal articles, editorials, proceedings and book chapters are considered, thesis and dissertations are out of scope.
4	Peer-reviewed journal articles, editorials, proceedings and book chapters using machine learning algorithms in process systems applications of reactor modeling, safety, and molecular design are considered.

within them and their applications in tackling various kind of modeling problems.

Before discussing the details of how those algorithms work, it is important to identify the basic components. Three main components that are critical in data-driven models are: (1) data, (2) representation, and (3) model. Such techniques can produce accurate results, due to their ability to generate fast and efficient algorithms and data-driven models for real-time data processing. As such, the input data sets are one of the crucial factors that are necessary for ensuring the success of output predictions. In other words, using the right data is critical to the success of any applied data-driven model, which in turn mainly consists of two different phases. Phase 1 involves the data training stage, whereas Phase 2 involves the prediction phase, as depicted in Fig. 1 (Subasi, 2020).

Data training consists of three different parts: (1) pre-processing, (2) learning, and (3) error analysis. First off, since it is always crucial to enhance the quality of the data used in machine learning, data pre-processing enables the extraction of meaningful data insights from the input data sets. Pre-processing may include activities such as the removal of duplicates, normalization, dimension reduction, and transformation, in addition to data extraction and selection. Following this, the data becomes more organized and structured for use in the learning stage (Cohen, 2021; Bonetto and Latzko, 2020). Depending on the nature of the data sets being used, different pre-processing activities can be applied, such as normalization, dimension reduction, etc. Following this, all pre-processed data are used to generate a trained model, using a certain learning technique. The "learning" technique used a key stage, since four different learning methods exist: (1) supervised, (2) unsupervised, (3) semi-supervised, and (4) reinforcement (Cohen, 2021; Subasi, 2020). Such different categories have been defined based on different system configurations, and will mainly affect how the data is being trained and subsequently how the trained algorithm will identify the mathematical model. A brief description of each learning method is provided below:

Supervised learning

This algorithm trains the machine by example. As indicated by its name, this technique has a supervisor. The supervisor refers to a virtual representation of the entity which enables the entity to analyze and present conclusions to the user. These conclusions can be in form of classification or regression. In classification, the algorithms help in placing any new value in pre-defined classes whereas regression provides numerical predictions. In this technique, the program is trained on well-labeled historical data. This dataset is provided in the form of input and output and the algorithm then finds a correlation between the two sets. It identifies the pattern in data and makes predictions on a set of new data. These predictions contain errors depending on the quality of data and the kind of algorithm being used and are constantly updated by the user to increase the accuracy.

Supervised learning can be used for solving two different classes of problems: (1) classification problems where the output variable is a category or a particular class such as classification of cancer vs non-cancer prognosis based on the input values of the model, and (2) regression problems which involve the prediction of real values for a given set of input e.g. prediction of home price based on input variables. Different supervised learning techniques include: Support Vector Machine, Linear/Polynomial Regression, Logistic Regrression, K-Nearest Neighborhood, Decision Tree, Random Forest and the Naïve Bayes Classifier.

Semi-supervised learning

In semi-supervised learning, the data set consists of both labeled data and unlabeled data. Labeled data is essential information that has meaningful tags so that the algorithm can understand the data, whilst unlabeled data does not have this information. The class of algorithms uses a similar input-output model as supervised learning, but they analyze the hidden information in large amounts of unlabeled data to enhance the accuracy of the supervised learning model constructed with labeled data.

Unsupervised learning

Unsupervised Learning are used to identify patterns in a given dataset. This class of learning does not provide the algorithm with labeled data of the previously known dataset. The algorithm analyzes the features of the input data and endeavors to identify similarities among them. The dataset consists of N examples that are not labeled. The algorithm employs only these input vectors to construct a model capable of uncovering and extracting concealed patterns within the features. These algorithms classify and group the data points given to them without any assisting "supervisor" as was in the case of supervised learning. Here the algorithm will sort the dataset based on their

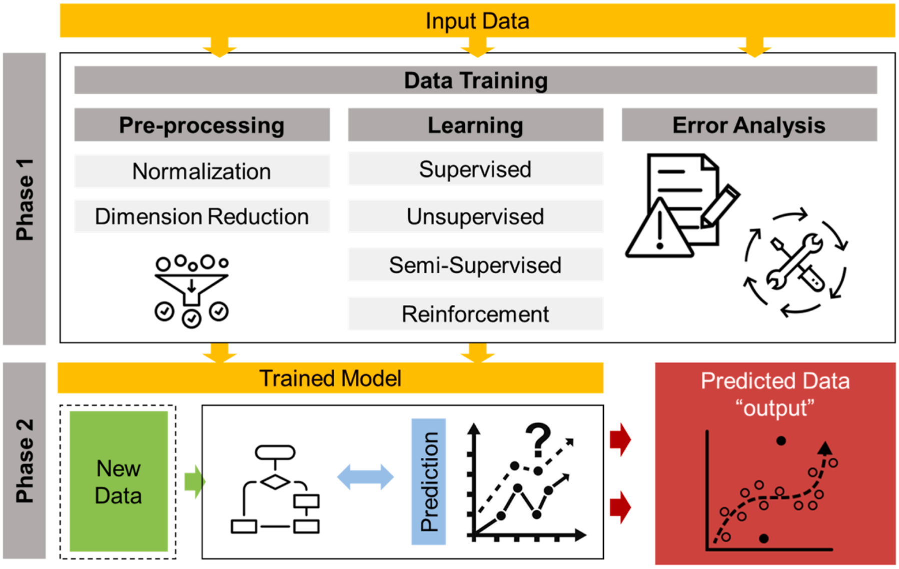


Fig. 1. A general overview of data-driven algorithm execution.

similarities and dissimilarities even though it does not have information about the categories beforehand. The most common unsupervised learning problems are clustering, association, and dimensionality reduction.

Clustering is the process of grouping unlabeled data based on their similarities or differences. This data mining technique is a handy tool for exploratory data analysis and provides insights into data that is not clear for an unlabeled dataset. Clustering algorithms are classified into four classes:

- Exclusive: Groupings formed on the assumption that data points can exist only in one kind of group.
- Overlapping: Allows data points to belong to multiple clusters.
- Hierarchical: Method of clustering technique where a hierarchy of clusters is developed. It can be agglomerative or divisive.
- Probabilistic: Probabilistic clustering helps in solving the density estimation of a dataset. Data points are grouped based on the likelihood that they belong to a particular distribution.

Clustering is beneficial in scenarios where the dataset does not have predefined class labels or when the nature of the data is not well understood. It helps reveal hidden structures, discover inherent similarities or differences, and identify potential subgroups or clusters within the data. Clustering can be applied in various domains, including customer segmentation, document clustering, image segmentation, anomaly detection, and recommendation systems.

By organizing unlabeled data points into clusters based on their similarities or differences, clustering enables exploratory data analysis, provides insights into unlabeled datasets, and serves as a foundation for further analysis and decision-making processes.

Association a method for finding relationships between variables in a given dataset. E.g.: Analysis of market datasets to gain insights into the consumption habits of customers to develop better cross-selling strategies or recommendation engines. The key concept in association analysis is the notion of "itemsets" or sets of items that frequently co-occur together. The most common form of association analysis is known as "frequent itemset mining," which identifies sets of items that appear together frequently in the dataset. The Apriori algorithm is a popular method for finding frequent itemsets efficiently.

The presence of high-dimensionality in process systems poses several modeling challenges. With a large number of variables, the available data points may be sparsely distributed across the feature space. This sparse sampling can lead to unreliable model estimation and poor generalization performance. Increased complexity is always a concern when dealing with high-dimensional data. Moreover, high-dimensional datasets often contain redundant or irrelevant features that do not contribute significantly to the underlying patterns or relationships. Including these features in the modeling process can lead to noise and overfitting, making it difficult to extract meaningful insights or build accurate models. The computational cost associated with modeling and analysis increases exponentially with the number of variables. High-dimensional datasets require substantial computational resources and time for training, optimization, and prediction tasks. Dimensionality reduction helps alleviate this burden by reducing the number of variables and simplifying the modeling process. As such, dimensionality reduction is a method deployed to reduce the dimension of data without any loss of information. It helps in dealing with the cases of overfitting and visualization of data. It reduces the number of data inputs to a smaller size while preserving all the information contained in it. It is a critical step involved in all machine learning problems and is used in the preprocessing stage. Following are some of the dimensionality reduction techniques widely used:

- Principal Component Analysis
- Singular Value Decomposition
- Auto-Encoders

Details regarding the aforementioned dimensionality reduction techniques methods could be found in Velliangiri et al. (Velliangiri et al., 2019).

Reinforcement

Reinforcement Learning enables in solving the class of problem referred to as Markov Decision Process. Markov Decision Process involves solving for sequence of decisions to be taken with an objective of maximizing the reward. This kind of problem is found in Real Time Optimization problems in PSE. Here the objective is to determine the optimal control moves so that the economic objective function can be maximized. It is a focused learning process, where an intelligent agent (computer program) interacts with the environment by moving one state of the system to another using various actions under different policies with a goal is to learn an optimal policy that selects the best action for a given state of the system. The algorithm tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. The algorithm demonstrates dynamic interaction with its environment through the execution of actions designed for varying states of the environment. Subsequently, the algorithm adapts its behavior based on the feedback received, either positive or negative, after each action. It is a trial-and-error method that learns from past experiences and adapts its response to a given situation to achieve the best result. Reinforcement learning differs from other techniques in the following ways:

- Does not require a data collection step as the algorithm learns from directly interacting with the environment.
- Works in dynamic and uncertain environments.

In fact, a very practical application to reinforcement learning was illustrated by Bangi and Kwon (Bangi and Kwon, 2020). Their work employed a reinforcement learning controller that does not rely on explicit models. The controller learns an optimal control policy by actively interacting with the process. The deep reinforcement learning (DRL) controller that was developed by Bangi and Kwon (Bangi and Kwon, 2020) was based on the Deep Deterministic Policy Gradient algorithm, which combines the Deep-Q-network with an actor-critic framework. To expedite the learning process, dimensionality reduction and transfer learning techniques were also employed. Bangi and Kwon (Bangi and Kwon, 2020) demonstrated that the controller successfully learns an optimal policy to achieve a uniform proppant concentration, overcoming the complexities inherent in the process. Furthermore, the controller ensures compliance with various inputs.

Following the implementation of a suitable algorithm from above mentioned methods, the error analysis stage mainly measures how well the trained model predicts the output variables. Hence, an appropriate error analysis technique (e.g. overfitting, cross-validation, etc.) must be utilized to be able to predict the extent of error between the input data sets and the generated “trained model” predictions. Low error predictions are key to ensuring a successfully predicted output in phase 2. Once the trained model is attained from stage 1, new data sets are used to test the trained model’s performance in the prediction phase (phase 2). It is very important to note that the data used to test the model in the prediction stage should be different than the data used in Phase 1. Following this stage, the desired output can then be attained, as depicted in Fig. 1.

According to Fig. 1, any data driven algorithm must first involve input data being fed into the algorithm. Input data can be obtained from various sources, such as sensors, databases, logs, or external data feeds. This data may include process measurements, operational parameters, environmental conditions, or any other information necessary for analysis or decision-making. Following this, and once the input data has been obtained, data training can then be carried out, which typically involves three different stages: (1) Pre-processing, (2) Learning and (3)

Error Analysis (Bonetto and Latzko, 2020). The data preprocessing stage involves collected data being preprocessed, so as to ensure its quality, integrity, and compatibility with the algorithms or models. This step may involve cleaning the data, handling missing values, normalizing or scaling variables, and transforming the data if required (Severson, 2018). Following this, the learning stage involves the extraction of relevant features or variables are extracted from the collected data. Feature selection techniques may be applied to identify the most informative and discriminative features that have the most significant impact on the target variable or desired outcome. Generally speaking, there are four different learning techniques including supervised, unsupervised, semi-supervised and reinforcement learning (Cohen, 2021). The next stage is error analysis, refers to the process of analyzing and understanding the errors or discrepancies between the predictions made by a trained model and the actual ground truth values in the training dataset. It involves examining the types and patterns of errors made by the model and gaining insights into the areas where the model struggles or performs poorly. The purpose of the error analysis stage is to identify the sources of errors and potential issues in the training process, as well as to guide the improvement of the model's performance (Subasi, 2020). Once the model or algorithm is trained, it can be executed using new or unseen data to make predictions, classifications, or other desired outputs. The algorithm leverages the patterns and relationships learned during training to generate predictions or perform specific tasks. The performance of the executed algorithm must then be evaluated using appropriate metrics or criteria (Subasi, 2020). This evaluation helps assess the accuracy, robustness, and generalizability of the algorithm. Based on the evaluation results, the algorithm may be refined, retrained, or adjusted to improve its performance (Bonetto and Latzko, 2020). Any insights, predictions, or outputs generated by the executed algorithm are used to inform decision-making or trigger specific actions in the given domain. These decisions or actions can range from process adjustments, resource allocation, anomaly detection, risk assessment, or any other operational or strategic choices.

Literature analysis results

When it comes to the area of process systems in chemical engineering, PSE is an area that mainly targets the understanding of a "bigger" picture for a given system/process, by breaking large and complex processes/systems into more manageable and well-defined sub-systems. First off, an artificial intelligence tool, VOSviewer, was deployed to trace relevant literature in the area and establish areas of connectivity between published articles (Perianes-Rodriguez et al., 2016; van Eck and Waltman, 2010). The tool was found very helpful in tracking relevant literature from key databases and establishing network visualizations. Network visualizations are very useful for representing connected data, such as graphs, using techniques adapted from graph analytics. Such network graphs usually consist of a set of nodes and edges that illustrate individual data points and display the relationships between them. A node is a single data point, while an edge represents how two given nodes are connected. All of which is then stored in a single graph database.

Network visualizations usually involve numerous nodes and edges that can help understand patterns in a certain dataset and spot any useful trends or anomalies. VOSviewer employs highly efficient graph algorithms that can establish bibliographic coupling, and keyword co-occurrences, and generate co-citation maps that are based on existing bibliographic data. For this, a general search on Scopus was performed, using the keywords "Machine Learning", "Chemical Engineering" and "Process Systems Engineering". Only journal articles were considered in the search. The search resulted in 813 results as follows: articles (423); conference papers (226); reviews (103); book chapters (30); conference reviews (18); books (8); editorials (2); and short surveys (2). All of the search attained results were then imported to the VOS viewer tool, to generate some informative network visualizations that can help us better understand the existing trends between published work, as illustrated in Figs. 2-4 below. Fig. 2 illustrates the co-occurrence of keywords in the published work extracted from the search made above, it is evident that there exist many keywords that have been extracted from the dataset of published sources, many of which utilize "Machine Learning" as a keyword while the occurrence of "Process Systems Engineering" and "Chemical Engineering" as major keywords was less prominent. As for the most cited work, Fig. 3 illustrated the bibliographic coupling network of published work based on the citations of all published work obtained from the search. What is quite noticeable and interesting is the work published by Oztemel et al. in 2020 (Oztemel and Gursev, 2020), which is a relatively recent publication that was equally prominent to older ones such as Kohonen et al. published in 1996 (Kohonen et al., 1996), and Gutierrez-Osuna et al. published in 2002 (Gutierrez-Osuna, 2002). Finally, Fig. 4 presents the bibliographic coupling of published work based on countries, and it is clear that most of the work has been published in countries such as the United States and China.

Computational applications using data driven modeling in process systems

Data-driven modeling in chemical and industrial applications are widespread since many previous studies have utilized such tools for a variety of different purposes. This section provides a glimpse of the application of ML tools in areas of Process Systems Modelling. Various works highlighting the ML application in areas of Reactor Modelling, Computer Assisted Molecular Design & Safety, Reliability & Control are mentioned here. The section first highlights applications in areas of Process Industries and then deep dives into the three sections mentioned above.

In area of Petrochemical Refining, Min et al. (Min et al., 2019) applied ML techniques in the production unit of a petrochemical factory, in which their model was trained via industrial (Internet of Things (IoT) data and used to realize intelligent production control based on real-time data. Steurtewagen and Poel (Steurtewagen and Van den Poel, 2020) used machine learning refinery sensor data to predict catalyst saturation levels in a Fluid Catalytic Cracking Unit (FCCU). To achieve this, Steurtewagen and Poel (Steurtewagen and Van den Poel, 2020) utilized a new soft sensor model in an input mix optimization to continuously optimize the use of the catalyst within the FCCU. Helmiriawan (Helmiriawan, 2018) evaluated the scalability of machine learning techniques for predictive maintenance in an oil refinery. The study involved modeling the normal behavior of the refinery plant, and using the prediction error to detect anomalies that have the potential to result in failures. Helmiriawan (Velliangiri et al., 2019) investigated various methods and learning algorithms to model the normal behavior of multiple components. Harp et al. (Harp et al., 2021) utilized a physics-informed machine learning (PIML) approach is used to manage reservoir pressures. In their study, the effect of the size of the training dataset, on the accuracy and efficiency of the PIML framework was also tested. Severson (Severson, 2018) focused on key algorithmic advances that bridge the gap between data and system insights using a series of hands-on case studies. Schweidtmann et al. (Schweidtmann et al., 2018) presented a novel optimization algorithm for self-optimization using multi-objective machine learning was introduced, which effectively determined a range of optimal conditions representing the trade-off curve (Pareto front) between economic and environmental objectives. They applied this algorithm to two chemical reactions carried out in a continuous flow system to illustrate its efficacy. Petsagkourakis et al. (Petsagkourakis et al., 2020) applied a Policy Gradient method from batch to batch to update a control policy parametrized by a recurrent neural network. Zhou et al. (Zhou et al., 2021) discussed hybrid data-driven and mechanistic modeling computational methods to guide material selection and design.

Quah et al. (Quah et al., 2020) combined an artificial neural network

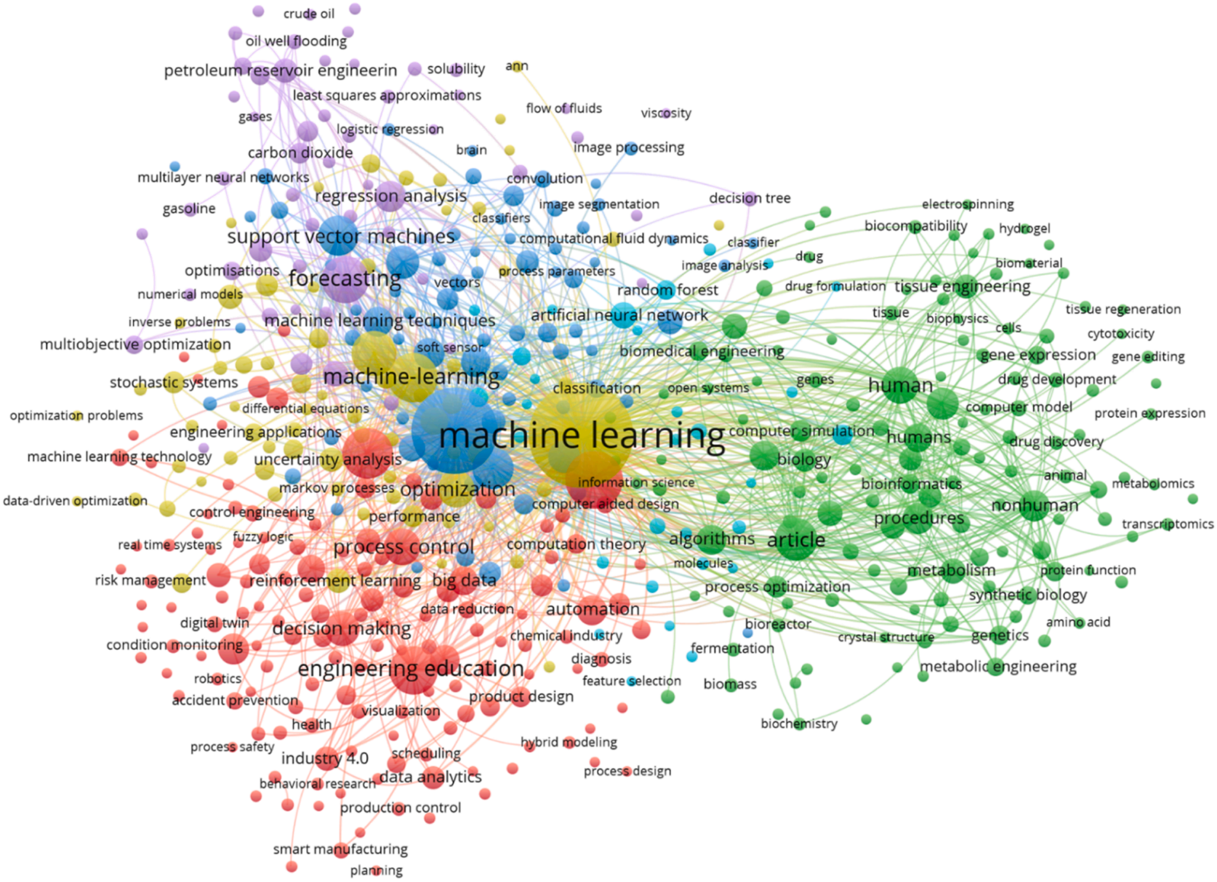


Fig. 2. Co-occurrence of all keywords in published work using VOSviewer network visualization (van Eck and Waltman, 2010).

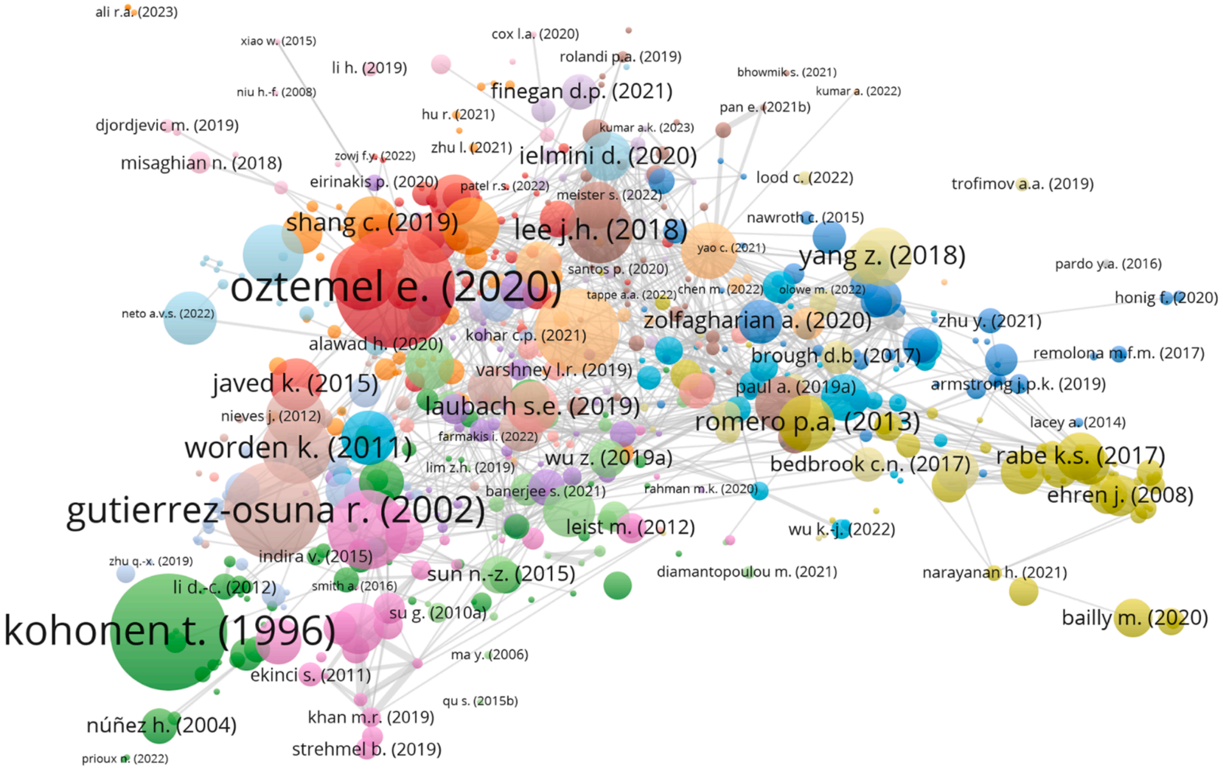


Fig. 3. Bibliographic coupling of published work based on authors using VOSviewer network visualization (van Eck and Waltman, 2010).

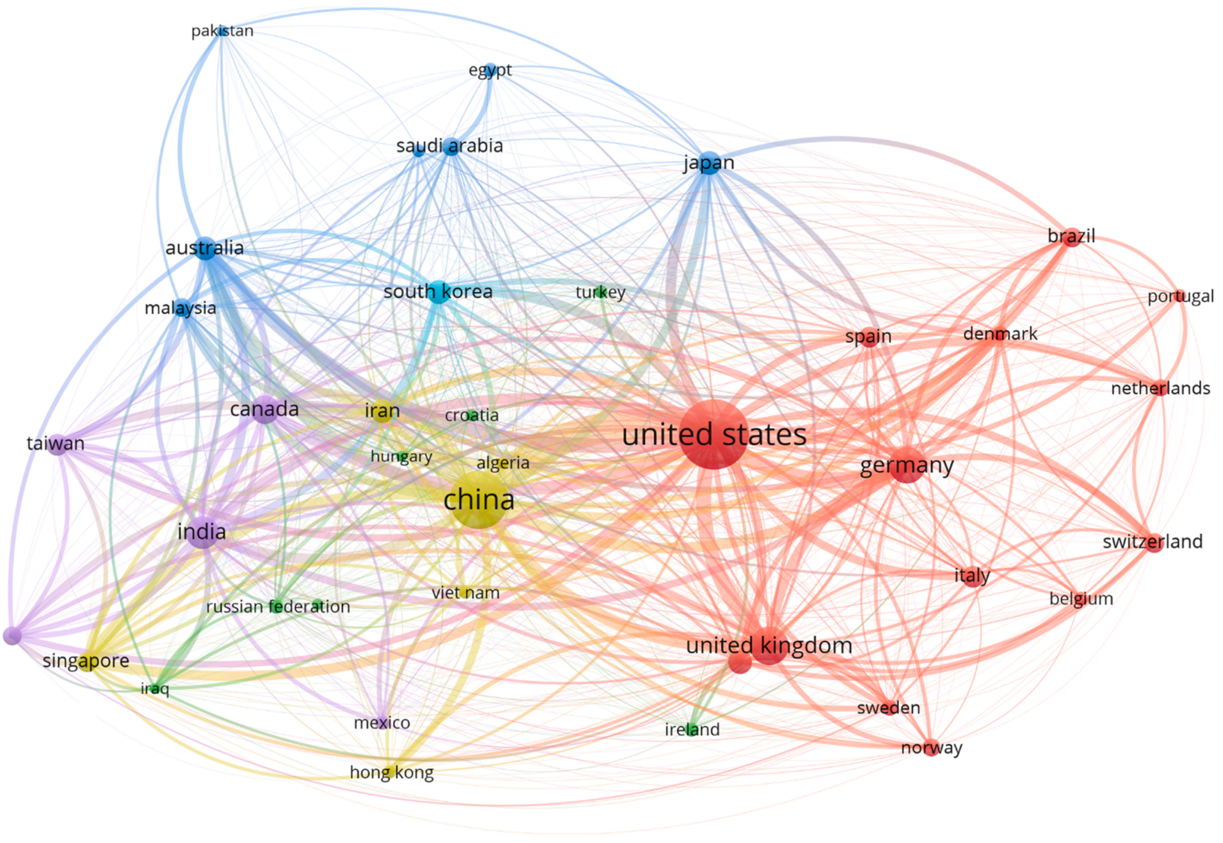


Fig. 4. Bibliographic coupling of published work based on countries using VOSviewer network visualization (van Eck and Waltman, 2010).

(ANN) model together with the particle swarm optimization (PSO) method, into a combined ANN-PSO procedure. The ANN-PSO's performance and applicability were then applied to a steady-state economic optimization of a chemical process, and compared to more conventional procedures. Ma (Ma, 2021) studied reaction modeling, optimization, and control using several deep learning-based approaches. Bogojeski et al. (Bogojeski et al., 2021) adopted machine learning tools to accurately forecast industrial aging processes. Sircar et al. (Sircar et al., 2021) presented a summary of various works on machine learning and artificial intelligence applications and limitations in the upstream oil and gas industry. Wang et al. (Wang et al., 2020) developed a large-scale mixed-integer linear programming model for refinery planning, which was also combined with a deep learning method employed to capture the uncertainties of product prices. Their technique has been proven to be quite efficient, especially when a high-dimensional price is utilized.

When looking more closely at ML tools that are particularly used in PSE, the following three main popular areas have been noted: (1) Modeling of Reactors, (2) Computer-Aided Molecular Design, and (3) Safety, Reliability & Control. Each area has been discussed separately in the following subsections.

Reactor modeling

Optimal reactor design and finding suitable operating conditions is a topic of great importance for many process industries. The benefits achieved are generally very significant compared to the cost involved in finding the optimal parameters. The main drawback of using mechanistic models for such predictions is the lack of flexibility in changing constant parameters in the model with time. These changes are associated with the current state of the reactor and it is quite impossible to account for them if mechanistic models are used. It has been consistently observed that mechanistic models are cost-intensive, time-consuming, and do not factor in non-idealities and degradation aspects of the system while modeling. The lack of such features renders them ineffective for real-time modeling of reactors. Given these drawbacks, reactor design is one of the areas where ML has been effectively used to determine more accurate models (shown in Table 2). Moreover, various predictions regarding the critical operating parameters can be conducted if an ML model is generated using the reactor feed stream information, together with the desired product specifications. For instance, an example of such predictions is the identification of a Weighted Average Bed Temperature (WABT). As such, ML-based models do consider these non-idealities and are constantly updated using a constant stream of fresh data. These models can then ideally be used for design, online optimization, and the control of reactors.

Early work of Bhat et al. (Bhat and McAvoy, 1992) highlighted the benefits of using Neural Networks in solving chemical engineering problems due to their parallel architecture. Bochereau et al. (Bochereau et al., 1991) explored the application of a multilayer artificial neural

Table 2
Summary of the algorithms used in reactor design.

Algorithm	Application
Artificial Neural Networks (ANN)	Modeling batch reactors, catalyst degradation behavior, hydrocracking unit, fluidized catalytic cracking unit, crude fractionation, chemical reactor behavior; predicting polymer quality, data rectification, product properties, yields of various products
Recurrent Neural Networks (RNN)	Modeling semi-generative catalytic reformer
Support Vector Regression (SVR)	Modeling Transportable Fluoride-salt-cooled High-temperature Reactor
Least Absolute Shrinkage and Selection Operator (LASSO)	Variable selection for fluidized catalytic cracking unit
Machine Learning-based predictive model	Feasibility of methanol steam reforming process

network for modeling the dynamic system of batch reactors. Chessari et al. (Chessari et al., 1994) developed an online model for a semi-generative catalytic reformer, which consists of a reactor that converts heavy naphtha to higher octane products. Chessari et al. (Chessari et al., 1994) utilized recurrent neural networks to model the catalyst degradation behavior. Himmelblau (Himmelblau, 2008) describes the characteristics of ANNs including their advantages and disadvantages. Himmelblau (Himmelblau, 2008) mainly focuses on two different types of neural networks, Feedforward and Recurrent. Both types of ANNs discussed have similar architectures, except for the fact that recurrent neural networks incorporate the factor of time. Himmelblau (Himmelblau, 2008) discusses the main elements of such networks and utilizes four different examples in the field of chemical engineering to demonstrate them: fault detection, prediction of polymer quality, data rectification, and modeling and control. Elkamel et al. (Elkamel et al., 1999) used Feed-forward neural networks to model the reactor of a hydrocracking unit using actual plant data. The model developed was used to predict the yields of various products coming out of HCU. Yield predictions enable the refinery to optimize, control, and plan. Long et al. (Long et al., 2019) developed a model that uses the least absolute shrinkage and selection operator (LASSO) method for variable selection and the back-propagation neural network (BPNN) method for predictive model construction for fluidized catalytic cracking units. Fakhr-Eddine et al. (Fakhr-Eddine et al., 1996) developed an ML model for LPCVD Reactors using neural networks. The objective of the work by Fakhr-Eddine et al. (Fakhr-Eddine et al., 1996) was to provide a black box model which can be used to compute online the film thickness.

Chaffart & Ricardez-Sandoval (Chaffart and Ricardez-Sandoval, 2018) developed a hybrid multiscale model for tin film deposition process. This model uses both mechanistic model & ANN for modeling this process. Lee et al. (Lee et al., 2020) developed a hybrid model for a partially known intracellular signaling pathway. Their work combines a mechanistic model and an ANN model to give better prediction than mechanistic models. The ANN model minimizes the discrepancy between the model predictions and available measurements. Nkulikiyinka et al. (Nkulikiyinka et al., 2020) developed a soft sensor model for sorption-enhanced steam methane reforming products. This model aimed for a smooth scale-up of the system as properties of hydrogen and sorption material hinder upscaling the system. Bawazeer and Zilouchian (Bawazeer and Zilouchian, 1997) applied supervised Neural Network architecture in a crude fractionation section of an oil refinery. The main objective of their work was to provide inferential product properties for enhancing the efficiency of the operations of the concerned unit. The properties that were predicted were the Naphtha 95% Cut point, in addition to the Naphtha Reid Vapor pressure. Data for three months has been utilized to develop the ML models and the simulated results for the above-mentioned properties were also analyzed. Bawazeer and Zilouchian (Bawazeer and Zilouchian, 1997) tested several neural network architectures. Wang et al. (Wang et al., 2019) modelled a pilot-scale entrained flow gasifier using ANN. The data driven model was generated from a large data which was generated from a pilot-scale gasifier reduced order model (ROM) and was validated with this model. Their ANN model was able to provide accurate predictions at much lower computational cost compared to ROM model. Bangi & Kwon (Bangi and Kwon, 2020) used deep hybrid modeling of hydraulic fracturing. Bhadriraju et al. (Bhadriraju et al., 2019) used Machine learning-based adaptive model identification of systems for finding a highly non-linear model of Continuous Stirred Tank Reactor (CSTR). Lithoxoidou et al. (Lithoxoidou et al., 2020) developed a Machine Learning based classification model for studying the behavior of chemical reactors. Lithoxoidou et al. (Lithoxoidou et al., 2020) proposed a data-driven methodology for depicting three distinct states of a chemical reactor, (1) normal, (2) warning, and (3) alert, all using ML. Predicting the classification of data input was found helpful in the early prognosis and the prevention of possible malfunctions. The objective of using the three distinct stages was to reveal the number of clusters based on past data, to train normal, warning, and alert behavior models, and finally validate them and third to test and verify their accuracy against real data sets. Mendiola-Rodriguez & Ricardez-Sandoval (Mendiola-Rodriguez and Ricardez-Sandoval, 2022) applied principles of Reinforcement Learning for developing an optimal control scheme for anaerobic digestion. Deep Deterministic Policy Gradient was employed as a learning strategy for single stage and two stage anaerobic digestion to manage Tequila vinasses. Zeng et al. (Zeng et al., 2018) developed a gray-box model for a Transportable Fluoride-salt-cooled High-temperature Reactor (TFHR). The prediction model that was used by Zeng et al. (Zeng et al., 2018) consists of a reactor physics model, a thermal-hydraulic model, and a Support Vector Regression (SVR) model. Several important transient parameters such as the reactivity insertion timings were also studied. Ding et al. (Ding et al., 2021) modelled a plasma-enhanced atomic layer deposition of hafnium oxide thin films. RNN was used for process modeling. This model accurately simulated the deposition process and gas phase transport profile and was computationally less expensive.

Byun et al. (Byun et al., 2021) developed a machine learning-based predictive model to find out the technical, environmental, and economic feasibility of a methanol steam reforming process. The effects of twelve different techno-economic parameters were studied by Byun et al. (Byun et al., 2021) and the predictive model was able to estimate the hydrogen production rate, the amount of carbon dioxide emissions, in addition to unit production costs. Attia et al. (Attia et al., 2020) developed a closed-loop optimization of fast-charging protocols for batteries with machine learning. The aim of this work was to optimize the parameter space specifying the current and voltage profiles of six-step, ten-minute fast charging protocol for maximizing the battery cycle life. Rahnama et al. (Rahnama et al., 2020) utilized machine learning techniques for the modeling of a basic oxygen steelmaking pilot plant. In doing so, Rahnama et al. (Rahnama et al., 2020) were able to obtain several correlations between key input parameters concerning the overall reactor performance. A neural network-based regression model was used to predict the decarburization rate in a basic oxygen steelmaking furnace. The reactor model was assumed to take place in an actual manufacturing plant based on a given lance height and total oxygen flow. Tom et al. (Tom et al., 2022) applied ANN for modeling the atomic layer processes with application in semiconductor industry. Ochoa-Estopier et al. (Ochoa-Estopier and Jobson, 2015) combined an ANN model together with an optimization framework to enhance the operational performance of crude oil distillation units. A new methodology was proposed by Ochoa-Estopier et al. (Ochoa-Estopier and Jobson, 2015) wherein the crude distillation units and the heat exchanger network (HEN) were both optimized using ANN models and a Simulated Annealing (SA) algorithm. The overall outcome was a two-stage process in which the distillation column is first optimized using an objective involving both the product yield and the energy demand. The HEN network was then optimized in the second stage. Abdullah et al. (Abdullah et al., 2021) developed a data driven reduced-order modeling of nonlinear processes that exhibit time-scale multiplicity. Using time-series data from all the state variables of a nonlinear process, an approach that involves nonlinear principal component analysis and neural network function approximators is employed to identify the fast and slow process state variables. Shah et al. (Shah et al., 2022) developed a Deep neural network-based hybrid modeling and experimental validation for an industry-scale fermentation process. Yun et al. (Yun et al., 2022) applied data driven algorithms for modeling and operation of thermal atomic layer etching of aluminum oxide thin films.

Liu et al. (Liu et al., 2021) developed an ANN model to study the effects of chemical additions onto nanoparticles (NPs) on hydrogen yield and hydrogen evolution rate. This included studying changes in the size and concentration of nanoparticles. Based on the results of this model, it was determined that Fe-based nanoparticles are more effective in enhancing the hydrogen yield, unlike Ni-based NPs ones. Gu et al. (Gu et al., 2020) has applied the concepts of deep learning for fast screening of heterogenous catalyst. Zhang et al. (Zhang et al., 2019) integrate neural network models with first principle models in both RTO and MPC for a system involving a continuous stirred tank reactor (CSTR) and a distillation column. The neural network part was used to model the nonlinear reaction rate of the CSTR, which is then combined with a first-principles model in RTO and MPC. The RTO was used to find the optimal reactor operating conditions for which a minimum energy cost within the system is attained. This was all made while ensuring optimum reactant conversion are maintained. The MPC was then used to ensure that the process functions under optimal operating conditions. In a second example, the neural network approach was used to develop a model for phase equilibrium properties, which was then integrated with the first principles model in RTO with aim of maximizing profit and optimal set-point identification. Zhong et al. (Zhong et al., 2020) modelled the Cu-Al electrocatalyst using density functional theory (DFT) & active machine learning. This hybrid model is computationally less expensive and is valuable in guiding the experimental exploration of multi-metallic electrocatalyst system. Choi et al. (Choi et al., 2023)used concepts of unsupervised learning and RNNs to develop a Long Short Term Memory (LSTM) model for multimode chemical process. This model was validated with real world data for distillation column.

Table 2 lists down various algorithms used for modeling various aspects of reactor modeling design. Algorithms like ANN, RNN, Support Vector Regression (SVM), Least Absolute Shrinkage and Selection Operator (LASSO) use principles of Supervised Learning and use labelled data to train the models. In case of lesser availability of labelled training data, Unsupervised learning is utilized for clustering of data and generating labelled data.

Computer-aided molecular design

Computer Aided Molecular Design (CAMD) problems are often geared toward the development and design of new molecules and enhanced molecular structures. Such improved molecules can then be utilized for a variety of unique applications, to attain enhanced process performance. Hence, mathematical formulations in CAMD must identify optimal molecules in a given solution space. CAMD searches through a design space of bond, aromaticity, and other chemical properties. Molecular modeling theories are then combined with thermodynamic properties, to be able to quantify the desired properties of the generated molecular structure. As a result, efficient and robust mathematical optimization techniques must be used to search the molecular design space and identify the optimal structure. The design of molecules may generally either involve improving the current version of molecules, or developing new ones altogether. Solutions to this kind of problem have led to major progress in the areas of power generation, medical treatment, climate change fighting techniques, etc. Machine learning algorithms are used in CAMD, which are summarized in Table 3.

When designing such molecules, Qualitative Structure-Property Relationships (OSPR) are commonly used to describe the properties of the desired molecular design, based on their structure. Group contribution theory, signature descriptors, and topological indices are some examples of QSPR models that could be utilized in this regard. With regards to the modeling and optimization methods that are involved in solving CAMD problems, they become computationally exhaustive and algorithmically very costly when implemented using mechanistic models. This complexity, therefore, has created a much-needed avenue for the utilization of ML techniques within this area. The use of ML techniques in CAMD has proven to be quite efficient, as they can very easily generate the required outputs without going through the complex computational mechanisms that are utilized by conventional mechanistic models. Many of the different Machine Learning techniques that have been applied in the area of CAMD have shown encouraging results. Moreover, the ability of deep learning techniques in handling high dimension data can make them effectively yield robust solutions when used in complex CAMD problems (Fuentes-Cortés et al., 2022). Furthermore, generative models can also very easily be used in conjunction with predictive QSPR models that relate learned feature representations of molecular descriptors to target chemical, physical, or biological properties of structures. For example, Autoencoders, Variational Autoencoders, Recurrent Neural Networks with Reinforcement Learning & Generative Adversarial Networks are some of the ML methods that have been utilized for solving CAMD problems.

Table 3
Summary of the algorithms used in molecular design.

Algorithm	Application
Autoencoders	Generate new molecules and optimize molecular design space
Variational Autoencoders (VAE)	Generate new molecules and optimize molecular design space
Deep Learning Networks	Predict pharmacological properties of drugs
Support Vector Machine (SVM)	Predict pharmacological properties of drugs
Reinforcement Learning (RL)	Generate new molecules and optimize molecular design space
Generative Adversarial Networks (GAN)	Generate new molecules and optimize molecular design space
Adversarial Autoencoder (AAE)	Identify new molecular structures with desired anti-cancer properties
Entangled Conditional Adversarial Autoencoder (ECAAE)	Search across the solution space of molecular structures based on various properties
Probabilistic Model for Gated Graph Neural Networks	Molecular design
Conditional Molecular Design Framework	Generate new molecules with desired properties
Molecular Hypergraph Grammar Variational Autoencoder (MHGVAE)	Molecular design
Chemical Reasoning Model	Predict reaction outcome

Sanchez-Lengeling et al. (Sanchez-Lengeling et al., 2017) utilized Generative Adversarial Network (GAN) and Reinforcement Learning (RL) to generate new molecules with a bias toward certain features. Their framework was based on Objective-Reinforced Generative Adversarial Networks (ORGAN). Sanchez-Lengeling et al. (Sanchez-Lengeling et al., 2017) demonstrated their developed method using several case studies in the area of drug synthesis, and organic photovoltaic material design. Popova et al. (Popova et al., 2018) used deep and reinforcement learning to integrate generative and predictive neural networks to develop novel molecules. Generative models were trained to produce chemically feasible simplified molecular-input line-entry system (SMILES) structure, and predictive models were used to forecast desired properties. Aliper et al. (Aliper et al., 2016) used deep learning networks for predicting the pharmacological properties of drugs. Their model was trained using large transcriptional response data and used to classify various drugs into therapeutic categories. Aliper et al. (Aliper et al., 2016) then compared their deep neural network model to a SVR model and the former was found to perform better. Segler and Waller (Segler et al., 2018) developed a model to mimic chemical reasoning to predict reaction outcomes. Based on the knowledge graph which contains 14.4 million molecules and 8.2 million binary reactions, this model was used to predict reaction outcomes for 180,000 binary reactions and was observed to outperform rule-based systems.

Liu et al. (Liu et al., 2018) developed a probabilistic model for gated graph neural networks into the encoder and decoder of a variational autoencoder (VAE) for molecular design. Constraints for molecular structure generation are then incorporated into this framework for efficiently searching the solution space. Kang & Cho (Kang and Cho, 2019) developed a model for a conditional molecular design framework for efficiently searching the solution space to generate new molecules with desired properties. The model improves the performance of property prediction by exploiting unlabeled molecules and efficiently generates novel molecules fulfilling various target conditions. Kadurin et al. (Kadurin et al., 2017) developed an advanced adversarial autoencoder (AAE) to identify new molecular structures with desired anti-cancer properties. The proposed AAE framework by Kadurin et al. (Kadurin et al., 2017) was compared with the VAE technique. AAE was observed to perform better in terms of flexibility of solution generation, handling enormous molecular datasets, and unsupervised pre-training for the regression model. Polykovskiy et al. (Polykovskiy et al., 2018) developed entangled conditional adversarial autoencoder, that is utilized for searching across the solution space of molecular structures based on various properties. This new model was used to generate a novel inhibitor of Janus kinase 3, which has been found in rheumatoid arthritis, psoriasis, and vitiligo. The discovered molecule was tested in vitro and showed good activity and selectivity. Guimaraes et al. (Guimaraes et al., 2017) proposed a method to guide the structure and quality of samples utilizing a combination of adversarial training and expert-based rewards with reinforcement learning which enables the search to be more effective when applied to a sequence-based generative model. This method is called an objective-reinforced generative adversarial network (ORGAN) and was applied to molecular synthesis problems. Kajino H. (Kajino, 2019) proposed a molecular hypergraph grammar variational autoencoder (MHG-VAE), which uses a single VAE to achieve 100% validity. This inspiration for the work was the fact that the normal VAE and Bayesian Optimization framework employed in solving the molecular synthesis problem has a complex architecture. A graph grammar encoding the hard chemical constraints called molecular hypergraph grammar (MHG) was developed to guide VAE to generate valid molecules. Putin et al. developed a DNN architecture named RANC (Reinforced Adversarial Neural Computer) for designing small-molecule organic structures based on the generative adversarial network (GAN) paradigm and reinforcement learning (RL). This new methodology shows better performance than its DNN-based counterpart objective-reinforced generative adversarial network for inverse-design (ORGANIC). RANC is able to generate structures that match the distributions of the key chemical features and lengths of the SMILES strings in the training data set thereby allowing for a more thorough search for a given amount of time. Putin et al. (Putin et al., 2018) developed another deep neural network-based architecture called Adversarial Threshold Neural Computer (ATNC). This model combines GAN architecture and Reinforcement Learning. To generate more diverse molecules, a new objective reward function named Internal Diversity Clustering (IDC) is also introduced. This framework was compared to ORGANIC and found to perform better in terms of the exhaustiveness of solution space search.

Ikebata et al. (Ikebata et al., 2017) combined machine learning models, Bayes law, the Monte Carlo technique, and natural language processing for generating molecular structures with a desired set of properties at a faster pace. Griffiths et al. (Griffiths and Hernández-Lobato, 2020) proposed an Automatic Chemical Design framework for generating novel molecules with the objective being having optimal specific properties. This method was compared with the original Bayesian Optimization over the latent space of a variational autoencoder. The original Bayesian optimization was modified since the original framework without constraints has been observed to be less efficient when compared to the modified technique. This was attributed to the fact that the original framework generated a lot of infeasible solutions, leading to many invalid molecular structures. Maziarka et al. (Maziarka et al., 2020) developed an improved Mol-CycleGAN, a CycleGAN-based model that generates optimized compounds with high structural similarity to the original ones, using optimized values with specific properties. Ooi et al. (Ooi et al., 2022) utilized ML techniques to identify fragrance molecules with desired product requirements. A case study where the objective was to design fragrance additives used in body lotions was used to demonstrate the effectiveness of the proposed model. Moreover, hyper box classifiers were used to predict the required fragrance properties.

Table 3 provides a list of algorithms used for various application in area of CAMD. Algorithms like Autoencoders, VAE, Generative Adversarial Networks (GAN), AAE, Entangled Conditional Adversarial Autoencoder (ECAAE) and Molecular Hypergraph Grammar Variational Autoencoder (MHGVAE) are Unsupervised Learning algorithms. Neural Network & Support Vector Machine are Supervised Learning Algorithms. Reinforcement Learning has also been used.

Process safety, reliability & control

Process Safety, Reliability & Control have been the areas where ML has been applied to a great extent, especially in problems that are related to regulating inspections and operations within a chemical plant (see Table 4). Moreover, fault detection problems have been extensively used by stakeholders to obtain information for predictive maintenance and improve the overall safety levels of the plant. Early fault detection and the following remedies are very critical for ensuring safe operations in a plant. Applying ML techniques on plant data to identify, isolate and take corrective measures not only enhances the safety levels but also saves cost as proper resources can be allocated to assets that need immediate maintenance. Apart from fault detection, ML has been applied to obtain information about fault prognosis which in turn provides remaining useful life predictions. ML methods like CNN (convolutional neural network), SVM, kNN (k Nearest Neighborhood Algorithm), and RNN have been extensively used on plant data to identify faulty patterns and classify the faults detected.

Kimaev & Ricardez-Sandoval (Kimaev and Ricardez-Sandoval, 2020) used ANNs to develop data-driven models that would enable optimal control of a stochastic multiscale system subject to parametric uncertainty. The system used for the case study was a simulation of thin film formation by chemical vapor deposition. The ANN was seen as a better option for optimization and control of the process as it was computationally less expensive and was accurate. Zhang et al. (Zhang et al., 2017) applied a multi-objective deep belief networks ensemble (MODBNE) method, which employs a multi-objective evolutionary algorithm integrated with the traditional DBN training techniques to obtain remaining useful life estimations in prognostics. Both accuracy and diversity were utilized as the two main conflicting objectives. Evolved deep belief networks were then combined to establish an ensemble model, where combination weights were optimized using a single objective differential evolution algorithm using a task-oriented objective function. The results obtained using the proposed method showed improved process performance when compared to standard methods. Kimaev & Ricardez-Sandoval (Kimaev and Ricardez-Sandoval, 2019) utilized ANN in area of process control. ANN was deployed in this work to develop data-driven models for model predictive control of a

Table 4
Summary of the applications of machine learning algorithms in safety.

Algorithm	Application
Convolutional Neural Network (CNN)	Fault detection
Support Vector Machines (SVM)	Fault detection, time-to-failure, and reliability forecasting
k Nearest Neighborhood Algorithm	Fault detection
Recurrent Neural Network (RNN)	Fault detection
Multi-Objective Deep Belief Networks Ensemble (MODBNE)	Remaining useful life estimations in prognostics
Adaptive Kernel Spectral Clustering (AKSC)	Finding machine anomaly behaviors from multiple degradation features
Deep Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN)	Machine time failure prediction
Deep Belief Networks (EDBN)	Fault analysis
Autoencoder	Fault detection and fault diagnosis
One-Class-SVM	Fault symptoms detection
Multilayer Feedforward Neural Networks based on Multi-Valued Neurons (MLMVN)	Reliability and degradation prediction
Health Index Similarity	Prediction of the remaining useful life (RUL) based on condition-based maintenance
Physics-based Fire Hazard Model	Generating metamodel approximations

computationally intensive stochastic multiscale system of thin film formation by chemical vapor decomposition. Cheng et al. (Cheng et al., 2019) used ML algorithms to handle the complexity of the problem of identifying heterogeneous features pertinent to fault diagnosis. In doing so, a novel ML-based approach was proposed using adaptive kernel spectral clustering (AKSC) and deep long short-term memory recurrent neural networks (LSTM-RNN). A Euclidean distance-based algorithm was then utilized to identify pertinent degradation features, and the AKSC algorithm was used for finding machine anomaly behaviors from multiple degradation features. Lastly, LSTM-RNN was utilized to predict the machine time failure. Nikita et al. (Tomin et al., 2016) developed an automated multi-model approach for online security assessment, in which they tested different state-of-art machine learning techniques and identified the best-performing ones.

Çıtmacı et al. (Çıtmacı et al., 2022)applied ML in area of RTO. The system modelled in this work is an ethylene based electrochemical reactor. The prediction provided by the model is used in a proportional-integral (PI) controller. Wang et al. (Wang et al., 2020) used deep belief networks (EDBN) in fault analysis. When compared to other methods, no loss of any information occurred during feature compression in traditional deep networks, unlike other traditional methods. A dynamic EDBN-based fault classifier was constructed to consider the dynamic characteristics of process data. The data was then tested on a Tennessee Eastman (TE) process for fault classification. Park et al. (Park et al., 2019) developed an integrated learning framework for jointly achieving fault detection and fault diagnosis of rare events in multivariate time series data. Park et al. (Liu et al., 2021) employed an autoencoder trained with offline normal data for detecting anomalies. The predicted faulty data, which was captured by the autoencoder, was then used as input into a LSTM network to classify the types of faults.

Bangi & Kwon (Bangi and Kwon, 2023) developed a Control Lyapunov–Barrier Function-based model predictive controller (CLBF-based MPC) which utilizes a deep hybrid model. The efficacy of the proposed control framework is demonstrated on a continuous stirred tank reactor. Narasingam & Kwon (Narasingam and Kwon, 2019) developed a predictive control scheme where Koopman operator theory with Lyapunov based model predictor control. The feedback control design in this work uses machine learning models. Arunthavanathan et al. (Gu et al., 2020) developed a fault detection framework for multivariable complex systems utilizing a CNN- LSTM approach to forecast system parameters. This approach was found to be effective in identifying fault symptoms in multivariate dynamic systems beforehand, thereby effectively detecting potential fault conditions. Additionally, unsupervised One-class-SVM was utilized for detecting fault symptoms using a forecasted data window. The performance of the proposed method was evaluated using the time series data from the Tennessee Eastman process. Luo et al. (Luo et al., 2022) applied ML based models for electrochemical reactor modeling to handle the data variability and enhancement of the accuracy of empirical models. Fink et al. (Fink et al., 2014) proposed a multilayer feedforward neural network based on multi-valued neurons (MLMVN), a specific type of complex-valued neural network. MLMVN was used for its good performance in extracting complex dynamic patterns from time series data, which resulted in a good performance in such reliability and degradation prediction problems. The performance of those algorithms was first evaluated using a benchmark study that provided railway turnaround data. According to Fink et al. (Fink et al., 2014), MLMVN was able to outperform other machine learning algorithms in terms of prediction precision and is also able to perform multi-step ahead predictions, as opposed to other previously best-performing benchmark studies.

Shah et al. (Shah et al., 2022) used machine learning model in developing optimal control to maximize industry scale fermentation process. A hybrid model has been used for modeling the fermentation process. Son et al. (Son et al., 2022) developed an offset-free Koopman Lyapunov-based model predictive control. mathematical analysis for zero steady-state offset condition considering influence of Lyapunov constraints on equilibrium point was also carried out. Linear model was developed using data driven modeling in this work. Bhadriraju et al. (Bhadriraju et al., 2021) developed Operable Adaptive Sparse Identification of Systems (OASIS) for fault Prognosis of chemical processes. This was application of ML in area of Process Safety. Operable adaptive sparse identification of systems (OASIS) is an adaptive modeling method developed based on sparse identification of nonlinear dynamics (SINDy) and deep learning. SINDy identifies nonlinear system dynamics using measured or simulated data but computationally expensive. OASIS solved this problem by implementing SINDy in real time using deep neural networks. Liu et al. (Liu et al., 2019) proposed a novel prognostic method for condition-based maintenance and the prediction of the remaining useful life (RUL)based on health index similarity. In their work, the nonlinear degradation evolution was revealed by the health index of cutting tools, and both the distance similarity and the spatial direction similarity were considered for similarity matching. The proposed method by Liu et al. (Liu et al., 2019) demonstrated great potential to outperform the LS-SVR method. das Chagas Moura et al. (Moura et al., 2011) used Support Vector Machines (SVMs) and carried a comparative analysis of different advanced learning techniques, including Radial Basis Function, MultiLayer Perceptron model, Box-Jenkins autoregressive-integrated-moving average and Infinite Impulse Response Locally Recurrent Neural Networks. The focus was on forecasting the time-to-failure and reliability of engineered components based on time series data.

Narasingam & Kwon (Narasingam and Kwon, 2019) applied Koopman operator for model-based control of fracture propagation and proppant transport in hydraulic fracturing operation. Narasingam & Kwon (Narasingam and Kwon, 2017)also applied dynamic mode decomposition for model predictive control of hydraulic fracturing. This was achieved by describing the local dynamics of the highly nonlinear process with terprally local reduced order models based on fully resolved data. Worrell et al. (Worrell et al., 2019) explored the application of machine learning for generating metamodel approximations of a physics-based fire hazard model, to improve the modeling realism in probabilistic safety assessments where the computational burden is prohibitive in the development of a high-fidelity model. In their work, Worrell et al. (Worrell et al., 2019) tested twenty-five different metamodel methods ranging in class and complexity were investigated, and kNN model fit the vast majority of calculations. The resulting kNN model was compared to an algebraic model typically used in fire probabilistic safety assessments. Gordon et al. (Gordon et al., 2020) developed and applied a framework to obtain optimal future-failure-aware and safety-conscious production and maintenance schedules, to improve safety and system effectiveness. Nonlinear support ensembles for several vector machine classification models were utilized to predict the time and probability of future equipment failure from equipment condition data. Moreover, a multi-objective optimization was carried out, in which both profit and a safety were used to determine optimal maintenance schedules. Kumari et al. (Kumari et al., 2021)applied ML in process safety and developed a parametric reduced order model for consequence estimation of rare events in process industry.

Wei et al. (Wei et al., 2018) employed variable importance analysis (VIA) and ML to investigate the reliability of structural systems through two novel reliability-based mode importance analysis (MIA) indices. They introduced a learning procedure that combined the multiple response Gaussian process (MRGP) model with Monte Carlo simulation (MCS) to efficiently and adaptively generate surrogate models for system failure surfaces. In the context of ensuring structural system reliability and simplifying reliability-based design problems, it was imperative to quantify the relative importance of random input variables and failure modes. Table 4 provides a list of algorithms used in the area of Process Safety, Control & Reliability. These algorithms belong to Supervised, Unsupervised and Semi Supervised learning methods of Machine Learning.

Challenges and an outlook

This section presents the challenges in modeling and adoption of data driven models in Industries. It also comments on the outlook of application of ML in PSE. Looking back at the application of ML in chemical engineering field, Subramanian (Liu et al., 2018) defines a timeline:

- Phase 0: Early attempts such as the Adaptive Initial Design Synthesizer system, developed by Siirola and Rudd (Siirola and Rudd, 1971)
- Phase 1: Bañares-Alcántara (Bañares-Alcántara et al., 1985) for predicting thermophysical properties of complex fluid mixtures and Stephanopoulos et al. (Stephanopoulos et al., 1990)
- Phase 2: Application of Neural Networks.
- Current Phase: Use of Deep Neural Networks.

Subramanian (Subramanian, 2019) classifies multiple phases of the deployment of AI in PSE. The work concludes with the need to develop domain-specific representations and languages, compilers, ontologies, molecular structure search engines, chemical entities extraction systems. It also argues for the true AI able to developing a theoretical framework or at least can reason using first-principles-based mechanisms as the first step. From the sections presented above we summarize below the key challenges that need to be addressed to enable wide scale applications of machine learning in chemical engineering.

Challenge: data availability

While AI has many potential applications in chemical engineering, chemical processes are often complex and high-quality data is not always readily available or in a format that is easy to use for machine learning (Lee et al., 2018). There exist some databases such as the National Institute of Standards and Technology (NIST) including thermodynamic and transport property data for fluids and solids, as well as reaction kinetics data (Eric et al., 2013). The American Institute of Chemical Engineers (AIChE) has some information on chemical process systems, including data on process dynamics, control systems, and process design (Beck et al., 2017). Nevertheless, there is a lack of standardized datasets that can be used to train and test machine-learning models for chemical engineering applications. These datasets can be used to train and test machine learning models for a variety of applications, such as process control, predictive maintenance, and optimization of chemical reactions. Apart from open sources, industrial data are not readily available due to the confidentiality associated with them. Moreover, a lack of specialized tools tailored for macro systems and scale up was also observed. Most of the datasets that were utilized in state of the art papers were mainly derived from experimental studies, and those were found to be limited in nature and are only fit for microsystem modeling (such as computer-aided molecular design).

Issues with data-driven models

Another challenge was pointed out by Wang et al. (Wang et al., 2020). Their work highlighted that many machine learning models are not easily interpretable, which can make it difficult to understand why a model is making certain predictions. This means that it is essential to understand the underlying physics and chemistry of a process and machine models are yet to be developed at that complexity level.

At the same time, the scalability of machine learning models to industrial-size systems where processes often need to be optimized for large-scale production is a complex challenge that is yet to be addressed. Machine learning algorithms can be used to analyze process data and identify patterns that can be used to improve the control of chemical processes. Similarly, it can be used in predictive maintenance and fault detection, to predict when equipment is likely to fail and schedule maintenance before it occurs and to detect and diagnose faults in chemical processes. However, an issue that can rise is the safety and robustness of machine learning models to changes in the process, making it difficult to train a machine on a wide range of conditions Wei et al. (Wei et al., 2018). Compared to the mechanistic models, data-driven model focus on data fitting to mimic the system. Due to the lack of physics-based models, there is an inertia in adopting these models for operational purposes. The quality of data from sensors is also a suspect as there are a lot of disturbances due to harsh conditions. These issues need to be examined for better adoption of these models.

To summarize, major challenges in data-driven modeling for PSE include the scarcity of available data. In many cases, collecting comprehensive and high-quality process data can be expensive, time-consuming, or even infeasible due to limitations in sensors, data acquisition systems, or process constraints. Insufficient data can lead to poor model performance, limited generalization, and difficulty in capturing complex process dynamics. Techniques such as data augmentation, domain knowledge incorporation, or transfer learning can be employed to mitigate data scarcity issues. Moreover, processes involved in PSE studies often exhibit multimodal behavior, where different modes or operational regimes coexist due to varying process conditions or external influences. Modeling such multimodal behavior poses challenges as traditional modeling techniques may struggle to capture and represent the diverse patterns or responses exhibited by the process. Identifying and characterizing the different modes and developing appropriate models that can adapt to and predict the behavior in each mode is crucial for accurate data-driven modeling in PSE. Process operations can be subject to uncertainties arising from various sources, such as measurement noise, parameter variations, disturbances, or unmeasured variables. Uncertainties can lead to deviations between the modeled and actual process behavior, affecting the performance and reliability of data-driven models. Addressing uncertainties requires robust modeling techniques that can handle noise, account for parameter variations, incorporate uncertainty quantification, and provide reliable predictions even in the presence of uncertain process operations. Hence, overcoming these challenges requires the development and application of advanced data-driven modeling techniques tailored to the specific needs of PSE. For instance, the use of hybrid models may help overcome such limitations since they often combine data-driven modeling approaches with physics-based models or domain knowledge can leverage the strengths of both approaches, improving model accuracy and capturing complex process behavior. Moreover, building ensembles of models that incorporate multiple algorithms, feature representations, or parameterizations can enhance model performance and robustness, especially in the presence of multimodal data or uncertainties.

With the development of better algorithms and computing capabilities, adoption of ML models will be more common in future. Better algorithms will help in dealing in more complicated problems and as computing becomes more accessible and cheaper, these algorithms will be easily implemented. These developments will enable the user to model more uncertain systems in industries which operate in very harsh conditions. Based on industry experience, a comprehensive overhaul of data collection and addressing of sensor reliability issues is required. Resources should be allocated to sensor handling and maintenance so that better data is available for modeling.

Future directions

In conclusion, it has been observed by many researchers that computationally extensive problems are a lot easier to handle using machine learning techniques. They were also found to be quite powerful for assisting optimization algorithms (especially in the case of mechanistic model failure (Zhou et al., 2021; Chaffart and Ricardez-Sandoval, 2018; Lee et al., 2020)). Many have reported that machine learning tools can greatly help increase the performance of their assets using historical data. Applications of data-driven techniques that rely on Reinforcement

Learning in the area of process control are also expected to grow on a larger scale, given the current status of the literature (Rangel-Martinez et al., 2021; Petsagkourakis et al., 2020; Quah et al., 2020; Mendiola-Rodriguez and Ricardez-Sandoval, 2022; Popova et al., 2018). Moreover, it will also likely pave new directions for data-driven applications in novel and underdeveloped subfields within process systems, which also could require the development of more versatile and efficient machine learning algorithms in the future.

With more advanced algorithms being developed and computational power becoming cheaper, data-based models can be used for generating better system identification models. This will enable better process control actions in APC systems (Mendiola-Rodriguez and Ricardez-Sandoval, 2022; Byun et al., 2021; Choi et al., 2023; Kimaev and Ricardez-Sandoval, 2019; Bangi and Kwon, 2023). Data-driven models can also be used for enhancing the pace of optimization solvers by enhancing the speed which can be achieved by looking at past data and then training the solver to search the solution space with greater efficiency. Apart from numerical data generated from sensors, images from catalysts can be used for determining the remaining life of the catalyst. Live video feeds and images can also be used in the area of process safety. The applications of data-driven modeling in process systems are expanding and the field has the potential to grow given that the challenges of data availability and scalability are addressed.

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.

References

Cohen, S., 2021. Chapter 2 - the basics of machine learning: strategies and techniques. In: Cohen, S. (Ed.), Artificial Intelligence and Deep Learning in Pathology. Elsevier, pp. 13–40.

Dobbelaere, M.R., et al., 2021. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering 7 (9), 1201–1211.

Zhou, X., et al., 2022. Chapter 2 - the basics of deep learning. In: Zhou, X., et al. (Eds.), Deep Learning on Edge Computing Devices. Elsevier, pp. 19–36.

Gao, H., et al., 2022. Machine learning and data science in chemical engineering. Ind. Eng. Chem. Res. 61 (24), 8357–8358.

Pratap, A., Sardana, N., 2022. Machine learning-based image processing in materials science and engineering: a review. Mater. Today: Proc. 62, 7341–7347.

Fuentes-Cortés, L.F., Flores-Tlacuahuac, A., Nigam, K.D.P., 2022. Machine learning algorithms used in PSE environments: a didactic approach and critical perspective. Ind. Eng. Chem. Res. 61 (25), 8932–8962.

Shi, T., et al., 2021. Chapter 1 - artificial intelligence in process systems engineering. In: Ren, J., et al. (Eds.), Applications of Artificial Intelligence in Process Systems Engineering. Elsevier, pp. 1–10.

Bogle, I.D.L., 2017. A perspective on smart process manufacturing research challenges for process systems engineers. Engineering 3 (2), 161–165.

Rangel-Martinez, D., Nigam, K.D.P., Ricardez-Sandoval, L.A., 2021. Machine learning on sustainable energy: a review and outlook on renewable energy systems, catalysis, smart grid and energy storage. Chem. Eng. Res. Des. 174, 414–441.

Guan, Y., et al., 2022. Machine learning in solid heterogeneous catalysis: recent developments, challenges and perspectives. Chem. Eng. Sci. 248, 117224.

Ifaei, P., et al., 2023. Sustainable energies and machine learning: an organized review of recent applications and challenges. Energy 266, 126432.

Wu, Z., et al., 2019. Machine learning-based predictive control of nonlinear processes. Part I: Theory 65 (11), e16729.

Stephanopoulos, G., 1990. Artificial intelligence in process engineering—current state and future trends. Comput. Chem. Eng. 14 (11), 1259–1270.

Forootan, M.M., et al., Machine learning and deep learning in energy systems: a review. 2022. 14(8): p. 4832.

Rattan, P., Penrice, D.D., Simonetto, D.A., 2022. Artificial intelligence and machine learning: what you always wanted to know but were afraid to ask. Gastro Hep Adv. 1 (1), 70–78.

Kotu, V., Deshpande, B., 2019. Chapter 10 - deep learning. In: Kotu, V., Deshpande, B. (Eds.), Data Science, 2nd Ed. Morgan Kaufmann, pp. 307–342.

Subasi, A., 2020. Chapter 3 - Machine learning techniques. In: Subasi, A. (Ed.), Practical Machine Learning for Data Analysis Using Python. Academic Press, pp. 91–202.

Bonetto, R., Latzko, V., 2020. Chapter 8 - machine learning. In: Fitzek, F.H.P., Granelli, F., Seeling, P. (Eds.), Computing in Communication Networks. Academic Press, pp. 135–167.

Velliangiri, S., Alagumuthukrishnan, S., Thankumar joseph, S.I., 2019. A review of dimensionality reduction techniques for efficient computation. Proc. Comput. Sci 165, 104–111.

Bangi, M.S.F., Kwon, J.S.-I., 2020. Deep hybrid modeling of chemical process: application to hydraulic fracturing. Comput. Chem. Eng. 134, 106696.

Severson, K.A., 2018. Machine Learning For Applications in Chemical and Biological Engineering. Massachusetts Institute of Technology.

Perianes-Rodriguez, A., Waltman, L., van Eck, N.J., 2016. Constructing bibliometric networks: a comparison between full and fractional counting. J. Informetr. 10 (4), 1178–1195.

van Eck, N.J., Waltman, L., 2010. Software survey: vOSviewer, a computer program for bibliometric mapping. Scientometrics 84 (2), 523–538.

Oztemel, E., Gursev, S., 2020. Literature review of Industry 4.0 and related technologies. J. Intell. Manuf. 31 (1), 127–182.

Kohonen, T., et al., 1996. Engineering applications of the self-organizing map. Proc. IEEE 84 (10), 1358–1384.

Gutierrez-Osuna, R., 2002. Pattern analysis for machine olfaction: a review. IEEE Sens. J. 2 (3), 189–202.

Min, Q., et al., 2019. Machine learning based digital twin framework for production optimization in petrochemical industry. Int. J. Inf. Manage. 49, 502–519.

Steurtewagen, B., Van den Poel, D., 2020. Machine learning refinery sensor data to predict catalyst saturation levels. Comput. Chem. Eng. 134, 106722.

Helmiriawan, H., Scalability Analysis of Predictive Maintenance Using Machine Learning in Oil Refineries. 2018.

Harp, D.R., et al., 2021. On the feasibility of using physics-informed machine learning for underground reservoir pressure management. Expert Syst. Appl. 178, 115006.

Schweidtmann, A.M., et al., 2018. Machine learning meets continuous flow chemistry: automated optimization towards the Pareto front of multiple objectives. Chem. Eng. J. 352, 277–282.

Petsagkourakis, P., et al., 2020. Reinforcement learning for batch bioprocess optimization. Comput. Chem. Eng. 133, 106649.

Zhou, T., Gani, R., Sundmacher, K., 2021. Hybrid data-driven and mechanistic modeling approaches for multiscale material and process design. Engineering 7 (9), 1231–1238.

Quah, T., D. Machalek, and K.M. Powell, Comparing reinforcement learning methods for real-time optimization of a chemical process. 2020. 8(11): p. 1497.

Ma, Y., 2021. Machine Learning Based Applications for Data Visualization, Modeling, Control, and Optimization for Chemical and Biological Systems. Louisiana State University and Agricultural & Mechanical College, Ann Arbor, p. 117.

Bogojeski, M., et al., 2021. Forecasting industrial aging processes with machine learning methods. Comput. Chem. Eng. 144, 107123.

Sircar, A., et al., 2021. Application of machine learning and artificial intelligence in oil and gas industry. Petroleum Research 6 (4), 379–391.

Wang, Y., et al., 2020. A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network. ISA Trans. 96, 457–467.

Bhat, N.V., McAvoy, T.J., 1992. determining model structure for neural models by network stripping. Comput. Chem. Eng. 16 (4), 271–281.

Bochereau, L., et al., Modélisation de réacteurs discontinus à l'aide de réseaux neuronaux. 1991. 5(13): p. 385–391.

Chessari, C., et al., 1994. The application of neural networks in the development of an on-line model for a semi-regenerative catalytic reformer. In: World Congress on Neural Networks.

Himmelblau, D.M., 2008. Accounts of experiences in the application of artificial neural networks in chemical engineering. Ind. Eng. Chem. Res. 47 (16), 5782–5796.

Elkamel, A., Al-Ajmi, A., Fahim, M., 1999. Modeling the hydrocracking process using artificial neural networks. Pet. Sci. Technol. 17 (9–10), 931–954.

Long, J., et al., 2019. Hybrid strategy integrating variable selection and a neural network for fluid catalytic cracking modeling. Ind. Eng. Chem. Res. 58 (1), 247–258.

Fakhr-Eddine, K., et al., 1996. Use of neural networks for LPCVD reactors modelling. Comput. Chem. Eng. 20, S521–S526.

Chaffart, D., Ricardez-Sandoval, L.A., 2018. Optimization and control of a thin film growth process: a hybrid first principles/artificial neural network based multiscale modelling approach. Comput. Chem. Eng. 119, 465–479.

Lee, D., Jayaraman, A., Kwon, J.S., 2020. Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling. PLoS Comput. Biol. 16 (12), e1008472.

Nkulikiyinka, P., et al., 2020. Prediction of sorption enhanced steam methane reforming products from machine learning based soft-sensor models. Energy and AI 2, 100037.

Bawazeer, K., Zilouchian, A., 1997. Prediction of products quality parameters of a crude fractionation section of an oil refinery using neural networks. In: Proceedings of International Conference on Neural Networks (ICNN'97).

Wang, H., Chaffart, D., Ricardez-Sandoval, L.A., 2019. Modelling and optimization of a pilot-scale entrained-flow gasifier using artificial neural networks. Energy 188, 116076.

Bhadriraju, B., Narasingam, A., Kwon, J.S.-I., 2019. Machine learning-based adaptive model identification of systems: application to a chemical process. Chem. Eng. Res. Des. 152, 372–383.

Lithoxoidou, E., et al., 2020. Towards the behavior analysis of chemical reactors utilizing data-driven trend analysis and machine learning techniques. Appl. Soft Comput. 94, 106464.

Mendiola-Rodriguez, T.A., Ricardez-Sandoval, L.A., 2022. Robust control for anaerobic digestion systems of Tequila vinasses under uncertainty: a deep deterministic policy gradient algorithm. Digital Chem. Eng. 3, 100023.

Zeng, Y., et al., 2018. Machine learning based system performance prediction model for reactor control. Ann. Nucl. Energy 113, 270–278.

Ding, Y., et al., 2021. Machine learning-based modeling and operation of plasma-enhanced atomic layer deposition of hafnium oxide thin films. Comput. Chem. Eng. 144, 107148.

Byun, M., et al., 2021. Machine learning based predictive model for methanol steam reforming with technical, environmental, and economic perspectives. Chem. Eng. J. 426, 131639.

Attia, P.M., et al., 2020. Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature 578 (7795), 397–402.

Rahnama, A., Z. Li, and S. Sridhar, Machine learning-based prediction of a BOS reactor performance from operating parameters. 2020. 8(3): p. 371.

Tom, M., et al., 2022. Machine learning-based run-to-run control of a spatial thermal atomic layer etching reactor. Comput. Chem. Eng. 168, 108044.

Ochoa-Estopier, L.M., Jobson, M., 2015. Optimization of heat-integrated crude oil distillation systems. part i: the distillation model. Ind. Eng. Chem. Res. 54 (18), 4988–5000.

Abdullah, F., Wu, Z., Christofides, P.D., 2021. Data-based reduced-order modeling of nonlinear two-time-scale processes. Chem. Eng. Res. Des. 166, 1–9.

Shah, P., et al., 2022. Deep neural network-based hybrid modeling and experimental validation for an industry-scale fermentation process: identification of time-varying dependencies among parameters. Chem. Eng. J. 441, 135643.

Yun, S., et al., 2022. Multiscale computational fluid dynamics modeling of thermal atomic layer etching: application to chamber configuration design. Comput. Chem. Eng. 161, 107757.

Liu, Y., et al., 2021. A review of enhancement of biohydrogen productions by chemical addition using a supervised machine learning method. Energies 14. https://doi.org/10.3390/en14185916.

Gu, G.H., et al., 2020. Practical deep-learning representation for fast heterogeneous catalyst screening. J. Phys. Chem. Lett. 11 (9), 3185–3191.

Zhang, Z., et al., 2019. Real-time optimization and control of nonlinear processes using machine learning. Mathematics 7. https://doi.org/10.3390/math7100890.

Zhong, M., et al., 2020. Accelerated discovery of CO2 electrocatalysts using active machine learning. Nature 581 (7807), 178–183.

Choi, Y., et al., 2023. Data-driven modeling of multimode chemical process: validation with a real-world distillation column. Chem. Eng. J. 457, 141025.

Sanchez-Lengeling, B., et al., 2017. Optimizing Distributions over Molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry. Cambridge Open Engage: Cambridge.

Popova, M., O. Isayev, and A. Tropsha, Deep reinforcement learning for de novo drug design. 2018. 4(7): p. eaap7885.

Aliper, A., et al., 2016. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Mol. Pharm. 13 (7), 2524–2530.

Segler, M.H.S., Preuss, M., Waller, M.P., 2018. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555 (7698), 604–610.

Liu, Q., et al., 2018. Constrained graph variational autoencoders for molecule design. Adv. Neural. Inf. Process. Syst. 31.

Kang, S., Cho, K., 2019. Conditional molecular design with deep generative models. J. Chem. Inf. Model. 59 (1), 43–52.

Kadurin, A., et al., 2017. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Mol. Pharm. 14 (9), 3098–3104.

Polykovskiy, D., et al., 2018. Entangled conditional adversarial autoencoder for de novo drug discovery. Mol. Pharm. 15 (10), 4398–4405.

Guimaraes, G.L., et al., Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. 2017.

Kajino, H., 2019. Molecular hypergraph grammar with its application to molecular optimization. In: International Conference on Machine Learning. PMLR.

Putin, E., et al., 2018. Reinforced adversarial neural computer for de novo molecular design. J. Chem. Inf. Model. 58 (6), 1194–1204.

Ikebata, H., et al., 2017. Bayesian molecular design with a chemical language model. J. Comput. Aided Mol. Des. 31 (4), 379–391.

Griffiths, R.-R. and J.M.J.C.s. Hernández-Lobato, Constrained Bayesian optimization for automatic chemical design using variational autoencoders. 2020. 11(2): p. 577–586.

Maziarka, Ł., et al., 2020. Mol-CycleGAN: a generative model for molecular optimization. J. Cheminform. 12 (1), 2.

Ooi, Y.J., et al., 2022. Design of fragrance molecules using computer-aided molecular design with machine learning. Comput. Chem. Eng. 157, 107585.

Kimaev, G., Ricardez-Sandoval, L.A., 2020. Artificial Neural Networks for dynamic optimization of stochastic multiscale systems subject to uncertainty. Chem. Eng. Res. Des. 161, 11–25.

Zhang, C., et al., 2017. Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics. IEEE Trans. Neural. Netw. Learn. Syst. 28 (10), 2306–2318.

Kimaev, G., Ricardez-Sandoval, L.A., 2019. Nonlinear model predictive control of a multiscale thin film deposition process using artificial neural networks. Chem. Eng. Sci. 207, 1230–1245.

Cheng, Y., et al., 2019. Machine health monitoring using adaptive kernel spectral clustering and deep long short-term memory recurrent neural networks. IEEE Trans. Ind. Inf. 15 (2), 987–997.

Tomin, N.V., et al., 2016. Machine learning techniques for power system security assessment**This work was supported by the Russian scientific foundation under grant No. 14-19-00054 and the 2015 endeavour scholarship and fellowship program. IFAC-PapersOnLine 49 (27), 445–450.

Çıtmacı, B., et al., 2022. Machine learning-based ethylene concentration estimation, real-time optimization and feedback control of an experimental electrochemical reactor. Chem. Eng. Res. Des. 185, 87–107.

Park, P., et al., Fault detection and diagnosis using combined autoencoder and long short-term memory network. 2019. 19(21): p. 4612.

Bangi, M.S.F. and J.S.-I. Kwon, Deep hybrid model-based predictive control with guarantees on domain of applicability. 2023. 69(5): p. e18012.

Narasingam, A. and J.S.-I. Kwon, Koopman Lyapunov-based model predictive control of nonlinear chemical process systems. 2019. 65(11): p. e16743.

Luo, J., et al., 2022. Machine learning-based operational modeling of an electrochemical reactor: handling data variability and improving empirical models. Ind. Eng. Chem. Res. 61 (24), 8399–8410.

Fink, O., Zio, E., Weidmann, U., 2014. Predicting component reliability and level of degradation with complex-valued neural networks. Reliab. Eng. Syst. Saf. 121, 198–206.

Son, S.H., Narasingam, A., Kwon, J.S.-I., 2022. Development of offset-free Koopman Lyapunov-based model predictive control and mathematical analysis for zero steady-state offset condition considering influence of Lyapunov constraints on equilibrium point. J. Process Control 118, 26–36.

Bhadriraju, B., Kwon, J.S.-I., Khan, F., 2021. OASIS-P: operable Adaptive Sparse Identification of Systems for fault Prognosis of chemical processes. J. Process Control 107, 114–126.

Liu, Y., Hu, X., Zhang, W., 2019. Remaining useful life prediction based on health index similarity. Reliab. Eng. Syst. Saf. 185, 502–510.

Moura, M.d.C., et al., 2011. Failure and reliability prediction by support vector machines regression of time series data. Reliab. Eng. Syst. Saf. 96 (11), 1527–1534.

Narasingam, A., Kwon, J.S.-I., 2017. Development of local dynamic mode decomposition with control: application to model predictive control of hydraulic fracturing. Comput. Chem. Eng. 106, 501–511.

Worrell, C., et al., 2019. Machine learning of fire hazard model simulations for use in probabilistic safety assessments at nuclear power plants. Reliab. Eng. Syst. Saf. 183, 128–142.

Gordon, C.A.K., et al., 2020. Data-driven prescriptive maintenance: failure prediction using ensemble support vector classification for optimal process and maintenance scheduling. Ind. Eng. Chem. Res. 59 (44), 19607–19622.

Kumari, P., et al., 2021. Development of parametric reduced-order model for consequence estimation of rare events. Chem. Eng. Res. Des. 169, 142–152.

Wei, P., Liu, F., Tang, C., 2018. Reliability and reliability-based importance analysis of structural systems using multiple response Gaussian process model. Reliab. Eng. Syst. Saf. 175, 183–195.

Siirola, J.J., Rudd, D.F., 1971. Computer-aided synthesis of chemical process designs. From reaction path data to the process task network. Ind. Eng. Chem. Fundament. 10 (3), 353–362.

Bañares-Alcántara, R., Westerberg, A.W., Rychener, M.D., 1985. Development of an expert system for physical property predictions. Comput. Chem. Eng. 9 (2), 127–142.

Stephanopoulos, G., Henning, G., Leone, H., MODEL, L.A., 1990. A modeling language for process engineering—I. The formal framework. Comput. Chem. Eng. 14 (8), 813–846.

Subramanian, V., The promise of artificial intelligence in chemical engineering: is it here, finally? 2019. 65(2): p. 466–478.

Lee, J.H., et al., Machine learning: overview of the recent progresses and implications for the process systems engineering field. 2018. 114: p. 111–121.

Eric, L., Marcia, H., Mark, M., 2013. NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 9.1. Natl Std. Ref. Data Series (NIST NSRDS), National Institute of Standards and Technology, Gaithersburg, MD.

Beck, D., et al., 2017. Data science for chemical engineers. AIChE J.