Wiley Online Library

ChemBioEng Reviews

ML@ChemE: Past, Present, and Future of Machine Learning in Chemical Engineering

Pınar Özdemir[1], **Ramazan Yıldırım**[1],*

Abstract

This paper aims to review the machine learning (ML) applications in chemical engineering (ChemE) and provide perspectives for the future. First, the evolution of ML, data structures, and ML applications in ChemE were reviewed; then, the current state of the art in ML and its ChemE applications were summarized. Finally, a perspective for the future developments, including recently popularized tools like generative artificial intelligence (AI) and large language models (LLMs), as well as major challenges and limitations, was provided. Although the initial applications were mainly on fault detection, signal processing, and process modeling, the focus had been extended to other fields involving material development, property estimation, and performance analysis in later years with the use of more complex models and datasets. In future, new developments like LLMs will likely spread more; the other new applications like automated ML, physics-informed ML, and transfer learning, as well as field-specific databases, will also get more attention. ML applications in ChemE-related fields, like new energy technologies, environmental issues, and new material discovery, are expected to grow further.

Keywords: Artificial intelligence, Chemical engineering, Generative AI, Large language models, Machine learning

Received: February 21, 2025; *revised:* February 21, 2025; *accepted:* May 15, 2025

DOI: 10.1002/cben.70012

1 Introduction

In the last few decades, we have been witnessing a digital revolution that is comparable to but much faster than the industrial revolution. Powerful computers and digital devices, advanced algorithms and software codes, and the internet have dramatically changed and mostly improved every aspect of life, including academia; we can process massive amounts of data or make very complex calculations with a speed and accuracy that were not even conceivable a few decades ago. The growing popularity of *artificial intelligence* (AI) adds another dimension to change; machine learning (ML), which is a subdivision of AI, has already become an indispensable tool for almost all fields of scientific research, including those in the chemical engineering (ChemE) discipline. It seems that we are entering to a new phase of change with the rise of generative AI tools in last few years.

ML uses certain algorithms and statistics to create and optimize programs with the capability of learning from known examples or past data. Due to the versatility of its functions (like prediction, classification, and clustering) and algorithms (from simple linear regression to recently popularized deep learning networks), ML can be implemented in any field, especially if there are large datasets to be analyzed, as ML is often associated with *data mining*, *data analytics*, or *big data* trends [1]. More recent tools, like large language models (LLMs), can also perform human-like functions like inference. Clearly, *bigness* of data is a relative term that may change from discipline to discipline or industry to industry; however, the premises are the same: (1) An increasingly larger amount of data is created in every field of life, including research publications, patents, business transactions, and social media, (2) it is getting much easier to store and retrieve large datasets as most of the data are in digital form, (3) accessibility of data is continuously increasing both in terms of technical capabilities and willingness for data sharing, and (4) we have tools to process large datasets and extract hidden information effectively. As the popular saying goes, "*The data is the new oil*."

The early applications of AI in ChemE mostly involved the use of expert systems between the early 1980s and mid-1990s even though various ML tools (as we call them today), especially artificial neural networks (ANN), have been also

[1] Pınar Özdemir https://orcid.org/0000-0003-0967-1188, Prof. Ramazan Yıldırım https://orcid.org/0000-0001-5077-5689 (yildirra@boun.edu.tr)
Department of Chemical Engineering, Boğaziçi University, Istanbul, Bebek 34342, Turkey.

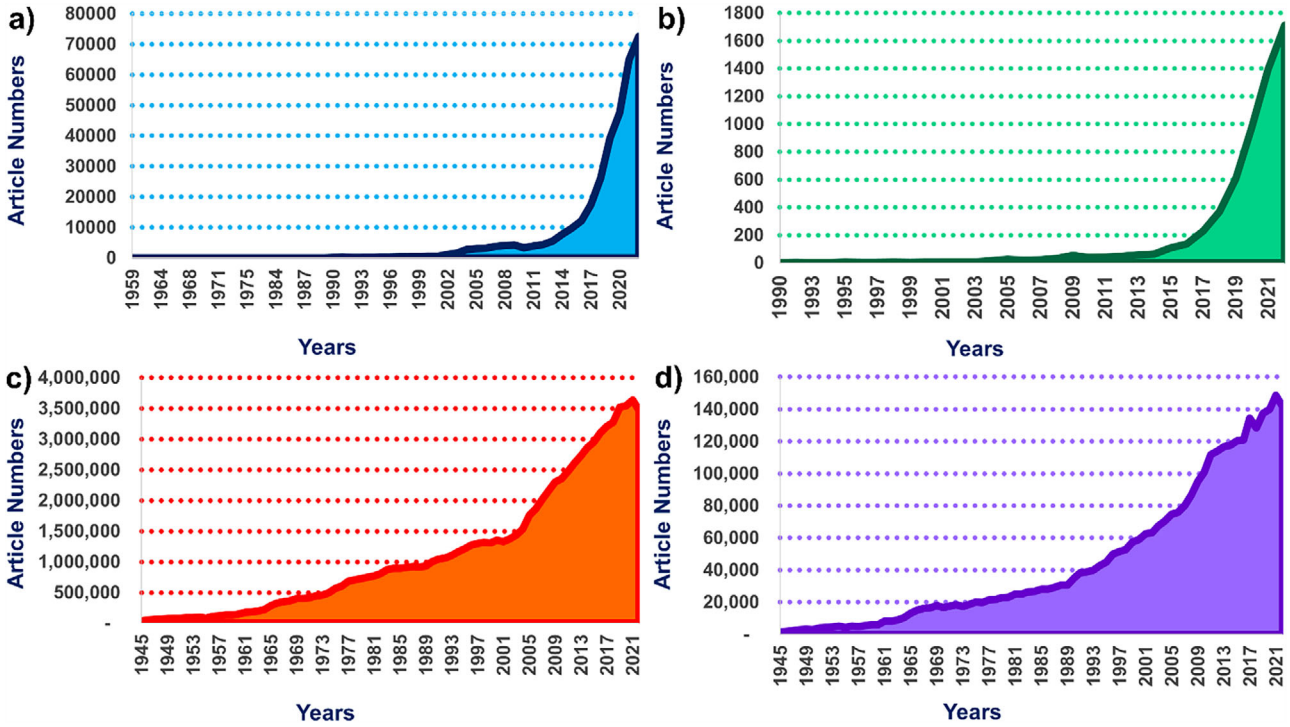

Figure 1. Number of publications indexed by the Web of Science database through the years (a) on machine learning, (b) on machine learning in chemical engineering and chemistry, (c) all publications, and (d) chemical engineering and chemistry publications.

employed occasionally. However, the use of expert systems did not spread further, probably due to the difficulties and high costs of establishing, maintaining, and updating an expert system in a continuously changing scientific and technological environment. ML applications, on the other hand, had grown with remarkable speed and diffused into all fields of ChemE, from molecular to system-level applications. Apparently, ML is better suited to the dynamic and evolutionary nature of life, as it also involves self-learning from examples. This is also evident from the exponential increase in the number of published papers indexed by the Web of Science (WOS) database as indicated by Fig. 1a (with data until July 25, 2023, https://mjl.clarivate.com/search-results); the actual number should be higher considering that not all publications are indexed by WOS. This is also true for the publications involving ML applications in chemistry and ChemE publications (Fig. 1b) (we could not separate chemistry and ChemE keywords and decided to take them together). We also put the total number of WOS publications (Fig. 1c) and publications involving chemistry and ChemE (Fig. 1d) to allow the comparison of the numbers and trends in publication; although the number of ML applications is still relatively small, its exponential growth in recent years is remarkable.

Various reviews have been also published on the field starting from 1990s as summarized in Sect. 2.3. Each of these, especially the recent ones, which reflects the current state of art better, are providing a different perspective by looking at a different aspect of the field. For example, Venkatasubramanian discussed the issue based on a historical perspective through the phases [2], whereas Schweidtmann et al. focused on ML challenges in ChemE [3]. Chiang and Castillo, on the other hand, covered the industrial applications of ML through various ChemE-related areas, like chemical process industry, energy, semiconductors, pharmaceuticals, and food industry [4]. Finally, Venkatasubramanian and Chakraborty reviewed the LLMs and discussed the limitations and the potential ways of their utilization in ChemE field [5].

In this work, we reviewed ML applications in ChemE and provided our perspectives for the future. We adopted a historical viewpoint to identify the major trends and turning points from the 1990s to the present, including recently popularized generative AI tools; we have done this for various aspects of subjects, including structure and availability of data, algorithms and tools used, and the applications in ChemE fields. Although we were summarizing the critical issues in the past experience, we also tried to reflect on the opportunities and challenges, as well as the potential applications, for the future. We think that such holistic approach will make significant novel contributions to the utilization of ML in ChemE field.

2 A Brief History: Evolution of ML and Its ChemE Applications

Starting with the major switch of attention from expert systems to ML algorithms, especially to ANNs, in the late 1990s, ML has been evolving continuously in many respects, including algorithms and tools, software codes and libraries, data structure,

Figure 2. Some concepts and problems appeared in ML applications through the years.

and databases; meanwhile, ML-related concepts have been used and discussed by increasingly larger number of people, including non-experts, creating a significant amount of confusion, misperceptions, and chaos, as we tried to illustrate in Fig. 2. For better comprehension of what we have today, we will summarize the major developments in the field under the subtitles of *ML models and tools, data structure, and ML applications in* ChemE.

2.1 ML Models: From Individual Algorithms to ML Framework

Although ML has been used in literature since the 1950s (with varying meanings depending on source and context), the concept, as we understand it today, is relatively new; it was mostly popularized in the 2000s and started to appear in scientific publications. On the other hand, techniques such as ANN, support vector machines (SVM), and decision trees (DT), which are considered tools under the ML umbrella today, have been used in ChemE (and in other fields) for decades without referring to ML; in fact, *data mining concept* was more popular in the early 2000s. Why, then, we suddenly needed to label our works as ML? Why it was not sufficient to use ANN or DT itself as we used to do so? How come a *good old* multiple regression suddenly becomes an ML tool? One possible reason is that the size and complexity of the data increased significantly through the years, and the part played by the machines became much more important. Better algorithms and tools were needed to handle large and complex datasets, whereas the data collection, storage, and processing were integrated as a single process, which is mostly automated (as if run by *machines*). Developments of *complete packages*, including data management software, ML techniques, and workflow management programs, rightfully implied a series of in silico activities, making the use of ML term more meaningful. Indeed, although the early works mostly focus on the algorithm used, the more recent publications cover the entire *framework* in a holistic manner, including the steps taken before and after running the ML algorithm as we will discuss later.

2.2 Data Structure: From Small In-House Datasets to External Big Data

In early applications of ML, relatively small datasets generated in the same laboratory were more common. The size of the dataset was not even the main focus; small or large, we needed to analyze our dataset. Even if we desired to have large datasets, it was not easy to construct; the experimental facilities were limited, and

molecular-level modeling like density functional theory (DFT) simulations was not as fast and accurate as it is today. Having the data from external sources was not practical either because the dissemination of knowledge was through printed media only; it was hard to reach data, retrieve it, and use it for ML with limited computational power. Then, starting in the late 1990s (or early 2000s), the structure and our perception of data were changed completely.

First of all, the research and development capabilities as well as the amount of effort in the world have increased considerably; consequently, the amount of scientific data that needs to be analyzed has increased in terms of *volume*, *velocity*, and *variety* as the common measures of big data [4]. High-throughput experimentation has become much easier thanks to the developments in robotics, automation, and algorithms to support them, whereas molecular-level modeling, especially DFT simulations, has become much easier and more accurate, allowing the generation of large amounts of data, especially for material research. Potential benefits of data sharing or utilizing external data have been also realized more; digital publications enhanced not only the dissemination of knowledge but also the accessibility to data. Almost all papers, patents, data repositories, and databases are accessible by most researchers with no or very small cost, and they will be easier to access in the future thanks to the continuously growing open access trend. Today, although *in-house* generated data are still invaluable, especially those generated with high-throughput experiments or computations, the use of data from external databases is becoming more important every day. The number of such databases has increased in recent years, whereas the existing databases are enhancing their content and format for automated data retrieval.

The technological developments that enabled the generation of large datasets also improved our data management capabilities; the storage, transfer, retrieval, and processing of data, including ML tools, also have been improved continuously, allowing us to process large amounts of data in a short time and extract knowledge that would remain hidden without such capabilities. Consequently, the data become an invaluable source as it is also evident from the recent popularity of concepts such as *big data, data science, data analytics, and data mining*.

2.3 ChemE Applications: From Simple Models to In-Depth Analysis of ChemE Problems

As clearly seen in the review papers that started to appear in the early 1990s, some ML tools have been used much earlier than ML concept in ChemE research; for example, Zupan and Gasteiger reviewed the applications of artificial neural networks in the fields of chemistry in 1991 [6] and 1993 [7], and they indicated that ANNs can be applied to spectroscopy, potentiometry, structure/activity relationships, protein structure, process control, and chemical reactivity, showing the examples starting from 1980s. Similarly, Burns and Whitesides reviewed the early applications of ANN in chemistry-related fields such as protein structure and functions, DNA sequences, spectroscopy, sensors, and QSAR [8]. Himmelblau also published a paper in 2000 to review the application of ANN in ChemE, especially the use of ANN in fault detection, signal processing, process modeling, and control [9].

Venkatasubramanian labels the growing popularity of ANN in the early 1990s as *Phase two*, replacing the early efforts that were mostly concentrated on expert systems (named *Phase one*); he reported the use of ANN in various fields such as modeling, fault diagnosis, control, and product design [2]. The other ML techniques have been also used for a long time, even though they were not as popular as the ANN. For example, DT was used for the classification of petroleum pollutants as early as 1977 [10], whereas it was also used for the classification and prediction of octane numbers of a set of 230 hydrocarbons in 1995 [11]. Similarly, it was used for the classification of the color of Barbaresco wine samples in 1995 [12]. Other ML techniques, like SVMs, were also occasionally utilized. Nevertheless, ANN seemed to dominate the 1990s as it is still the most common technique.

Today, the initial fields of ChemE applications, such as fault detection, signal processing, process modeling, and control, as stated by Himmelblau in 2000 for ANN, are still relevant with the addition of new algorithms; in fact, such applications became much easier thanks to the automation and digitalization of processes. However, ML applications have been spread over much larger areas of ChemE; catalysis [13], bioinformatics [14], solar cells [15], Li batteries [16], biofuels [17], reaction network analysis [18], and thermodynamic property estimation [19] can be given as some representative examples (not an exhaustive list). Fig. 3 summarizes the areas in which ML has been applied more frequently through the years. We drew data from author keywords of ML papers in ChemE (we searched on WOS databases by a combination of two keywords; "Chemical Engineering" and the name of common ML tools) but organized the data ourselves in two dimensions because some of the keywords (like optimization) were not sufficient to describe the ChemE area involved (rising the question of *optimization of what?*). In Fig. 3, the beams represent keywords (as taken from the papers), whereas the balls with different colors indicate different ChemE areas collected in four commonly used groups in the field; the circles in the radial direction, on the other hand, show the change in total numbers in last four decades, whereas the decade-wise number of keyword appearances is in the left-hand side of the figure (we normalized 2020–2023 data to a decade).

Fig. 3 shows the ChemE research areas in which ML was implemented together with some general trends. First, as expected, the number of keywords increases with time as the number of ML papers related to ChemE increases. Second, the data distributed in larger research areas (as evident from the distribution of balls) in recent decades. Third, ML has been used most frequently in the process and reactor design area at all times; this can be attributed to that most of the traditional (and still continuing to be important) applications of ML (like process control and identification, optimization, and fault detection) are accumulated in this category if they are not directly related to one of the remaining three fields.

As an example, to see the diversity of ML applications (in terms of aim, data, and tools), we can take a closer look to the new energy technologies as some of them are directly related to ChemE discipline. ML is used to analyze market-level data such as supply/demand forecasting [20] and grid optimization [21]. It can be also used in the area related to individual energy

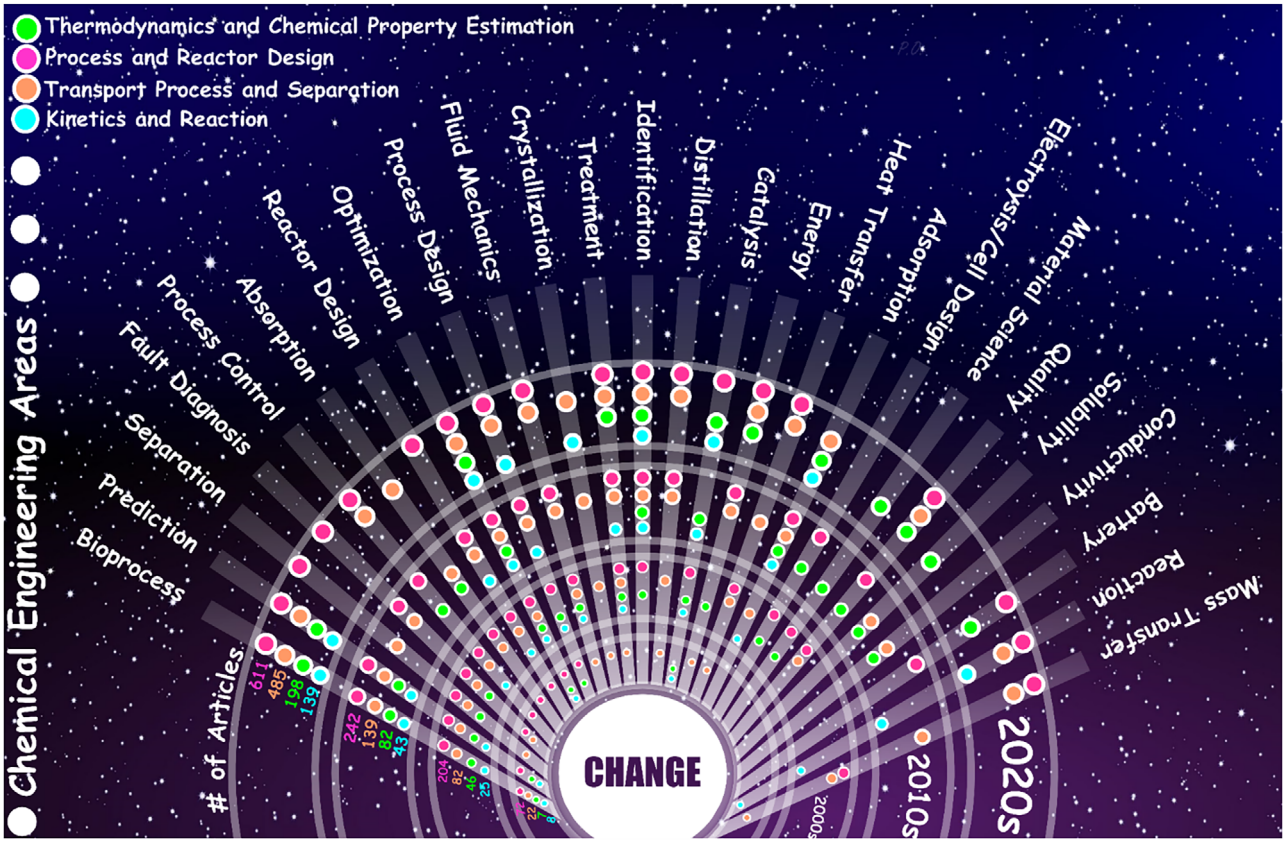

Figure 3. ML applications in chemical engineering areas. Beams show the author keywords in ML publications in chemical engineering fields, whereas color balls indicate chemical engineering fields. Data of 2020s were obtained by normalizing 2020–2023 data to 10 years.

technologies; for example, ML has been utilized for various solar energy technologies such as concentrated solar power (CSP) generations [22], photovoltaic (PV) [23], production of solar fuels [24], and desalination [25]. The same diversity also exists in other renewable energy technologies such as biofuels [26], which are also quite diverse in feedstock and technologies used; this is also true for the batteries [27], which are indispensable companions of renewable energy technologies; variety of battery technologies have been investigated in recent years with the help of ML tools. In each of these energy technologies, one can find ML applications for various tasks such as material screening, performance assessment, and optimization and control.

Up to this point, we discussed the application of ML in ChemE from an academic perspective; we can also look at the subject from an industrial point of view briefly. Chiang et al. reviewed the use of big data in ChemE-related fields such as the chemical process, energy, semiconductors, pharmaceuticals, and food industries. They pointed out that 88 % of the executives in the field recognize the need for big data [4]; apparently, the increasing volume of the data generated makes the initial/traditional applications (like fault detection, process control, and optimization) more meaningful. However, it is also not surprising that energy (both on the supply and demand side) has been receiving significant attention for ML use as well, because a large number of factors and past data should be considered in planning the generation and distribution of energy, whereas the new energy systems (like batteries and PVs) also require ML-assisted searches for the new materials, including semiconductors, electrolytes, and other ingredients. As another example, the food industry can benefit from indirect applications of ML involving weather, climate change, and agriculture as well as direct applications such as improving the process steps and laboratory analysis, which is becoming more complex (like spectroscopic measurement of trace materials in foods) with increasing health concerns and tighter regulations. These examples should be sufficient to illustrate that ML can make a significant contribution to industrial applications as well.

3 Current State of Art in ML

In recent years, ML is implemented as an integrated framework covering various steps systematically as we presented in Fig. 4. We may also add *step zero* to describe *the clarification of the objective* as all the remaining steps in the process depend on it; if the aim is not clear at the beginning, as happens frequently, one may end up selecting the most popular but not suitable applications and producing results that may not have practical use for that specific problem. In this section, we will briefly go over the steps and summaries in ML applications.

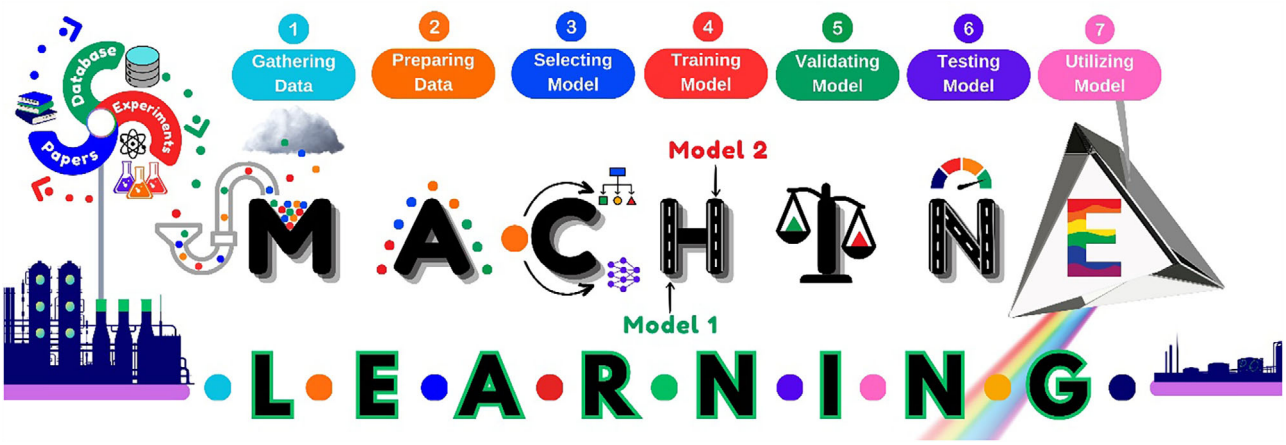

Figure 4. Implementation steps of machine learning.

3.1 Dataset Construction

As the first requirement, the data should contain information to establish meaningful relations between the *input variables* (like properties of material or characteristics of a process) *with some response (output) variable* (like other properties or function material, the performance of a process). The input variables, which are also called *descriptors, features, fingerprints, and so on* (we will use descriptors), may be categorical (like material type) or continuous (like concentration of material); they may be set by the user (like intended mole fraction of an ingredient) or measured properties (like surface area); similarly, they may be real or derived from the others (like principal components).

Although the final set of descriptors is decided in the second step when we are preparing the data for analysis, as we will discuss below, we need to have an initial list to determine the structure of the dataset and collect data. We may do that by using our knowledge and experience on the subject or inspecting similar works in the literature; there are also works that analyze the potential descriptors for specific fields. For example, Ward et al. stated that they identified 148 candidate descriptors (as they classified as *stoichiometric attributes, elemental property statistics, electronic structure attributes*, and *ionic compound attributes*) that may be used for the prediction of material properties [28]. It will not be practical to incorporate all potential descriptors in the dataset (often not necessary either); we need to consider only those relevant for the output variables. The selection of descriptors also depends on the way we establish a relationship with the output variables; we can do that experimentally (like performance experiments of processes or characterization tests of materials) or computationally (like property estimation with DFT), requiring different sets of descriptors.

The dataset can be generated experimentally or computationally, and this can be done in-house or extracted from external sources. Because the ML analysis uses a large number of data, in-house data generation usually involves high-throughput experimentation [29] and computations mostly by using DFT [30]. However, the extraction of data from external sources becomes more popular and effective as the number of external data sources increases. For example, databases like Pauling File Database [31], Inorganic Crystal Structure Database (ICSD) [32], Pearson Crystal Data [33], Cambridge Structural Database [34], Crystal Open Database [35], and PubChem [36], which mostly contain experimental data for crystal structures, are used as the source of data in ML applications. There are also databases that contain DFT computed properties for various materials; examples are The Materials Project [37], Automatic FLOW for Materials Discovery Library (AFLOWLIB) [38], The Computational Materials Repository [39], Open Quantum Materials Data (OQMD) [40], and Novel Materials Discovery (NOMAD) Repository [41]. Apparently, in addition to the technical enablers making these databases possible, the success of Human Genome Project provided additional motivation as indicated by the Materials Genome Initiative (https://www.mgi.gov/;) and other similar projects that use the concept of *genome* in their names. The data extracted from published papers, patents, industrial processes, business transactions, and social media communication also became a valuable resource for ML [18, 42].

3.2 Data Preprocessing

Before ML application, the data should be organized first in a machine-readable format and homogenized in units. The missing data (empty cells) must be completed using a suitable way; depending on the character of the missing information, the missing data may be imputed using the mean, mode, or other representative values for that variable. For instance, it is very likely that the absence of pressure data represents the atmospheric pressure, whereas the lack of information for a pretreatment (like washing) probably means that there was no such treatment. In more complex situations, the missing data could be obtained with the help of similar structures (like similar molecular formula, similar crystal structure, and similar pore structure) or can be computed using alternative tools, including to construct ML models using the data points with complete information and to use that model to predict the incomplete data. As an example, we often complete the bandgap data for photocatalysis by training an ML model with known bandgap values as the function of molecular properties of semiconductors

and using that model to predict the missing bandgaps in the dataset [43, 44].

Then, the data are processed further with normalization, standardization, and transformation if it is necessary. Lastly, the descriptors list can be analyzed and finalized by removing those that are not necessary for the analysis performed. For example, we may not need all the physical or chemical properties of a material for all problems to be solved; there may be simply ineffective or there may be strong linear correlations making some of them redundant and causing misinterpretation of the results. In any case, the smaller models (i.e., models with a smaller number of descriptors with respect to a number of data points) are generally more robust and reliable [45]; hence, we may need to reduce the descriptor lists by analyzing their importance for the model. This is called *feature engineering* by some investigators to stress the importance of *selecting the right descriptors* and *dimensionality reduction* to point out that the number of descriptors should be as small as possible. This is usually performed through two tasks. The first is *feature selection*, which involves decreasing the number of descriptors through some tools like correlation analysis (to remove highly correlated descriptors) or input important analysis (by identifying the descriptors that do not have a significant effect on the outcome using a method like Boruta analysis). The second is *feature extraction*, which creates a smaller number of new descriptors from the original features (like principal component analysis).

3.3 Model Selection

In this step, an ML algorithm suitable for the analysis and structure of data is selected. ML algorithms are generally grouped as *supervised* (uses labeled datasets to train the algorithms, like ANNs), unsupervised (learns from unlabeled data, like *k*-means clustering), and *semi-supervised* (uses both labeled and unlabeled data); they can be also grouped according to their functions such as *description, clustering, classification, estimation/prediction*, and *association* [37]. *Description* or *exploratory data analysis (or metaanalysis)* [46] is used to show the basic characteristics of data in graphical forms; it may be argued that there is no learning to present the data in graphical form; however, this technique is usually used for the pre-analysis of data, and it may be quite useful to understand the structure of the dataset [47]. The algorithms under clustering aim to group the data using the similarity in features (not the output); again, it may be a beneficial tool for pre-analysis. *k-means clustering or hierarchical clustering* are the common clustering techniques. *The classification*, on the other hand, groups in terms of the range (for continuous variables) and categories (for categorical variables) of the output variable (like hot/cold; material A, B, or C; faster than 10 km/h; between 10 and 50 km/h). DT, *k*-nearest neighbor (KNN), logistic regression (LOR), Bayesian classification (BC), SVM, and ANN are the common classification techniques used. *Estimation/prediction, which is the most common ML task*, creates a model describing the relation between descriptors and output (dependent) variables to make predictions; MR, ANN, RT, RF, GBR, and SVM are the common estimation tools used in ML analysis. *Association rule mining (ARM)* is used to identify the hidden relations among the features, which appear together (such as one may be promoting the other). Apriori is one of the simplest and most common algorithms used for this purpose [28]. As we will discuss in more detail later, we can also add *inference* capabilities of newly popularized generative AI tools to the list of functions that can be performed.

3.4 Training and Testing of Model

After the data are prepared and the ML algorithm is selected, one can *train the model* with the data (i.e., determine the optimum model hyperparameters representing the data best). In common practice, the data are divided into training (large portion like 80 %) and testing (small portion like 20 %) subsets. Then ML model is built using a *training set* mostly through a procedure called *k*-fold cross-validation. In this approach, the testing set is further divided into k subsets; the model is built using $k-1$ subsets (training) and validated with the remaining one in rotation under various values of the model hyperparameters (by employing parameter screening procedures like grid search). After the model parameters are determined (usually those resulting in the smallest average validation errors), they are tested against the *testing set*, which is separated in the first step and, hence, not seen by the model during training.

4 Recent Applications in ChemE

Chemical engineers, both in academia and industry, must deal with materials or process-related tasks involving the analysis of large and complex datasets; this has become especially important in recent years with the automation and digitization of chemical processes. Consequently, ML has been used in ChemE extensively with an increasing level of acceptance and diversity in application areas. In this section, we will summarize the most common ML applications in ChemE today in three *categories: material discovery and property estimation, performance analysis/improvement of chemical processes,* and *process monitoring, control, and optimization;* this classification should cover majority of applications and provide some generalization to assess the current state, even though there may also be applications that do not fall into one of these categories.

4.1 Material Discovery and Property Estimation

In the last two decades, one of the most popular ML application types in physical sciences has been the prediction of *material properties* [48, 49] to anticipate the performance of the existing materials for a specific use or to discover new materials with desired properties or functions [50]. Similar to chemistry, physics, material science, and medicine/health sciences, such use of ML is quite relevant to ChemE processes requiring property estimation or new material design; examples can be found in various fields such as catalysts [51], membranes [52], MOFs with specific functions [53], battery materials [54], and semiconductors for photoelectrochemical and PV systems [55]. The popularity of ML-assisted material search or property estimation is likely to increase in the future because it mostly relies on

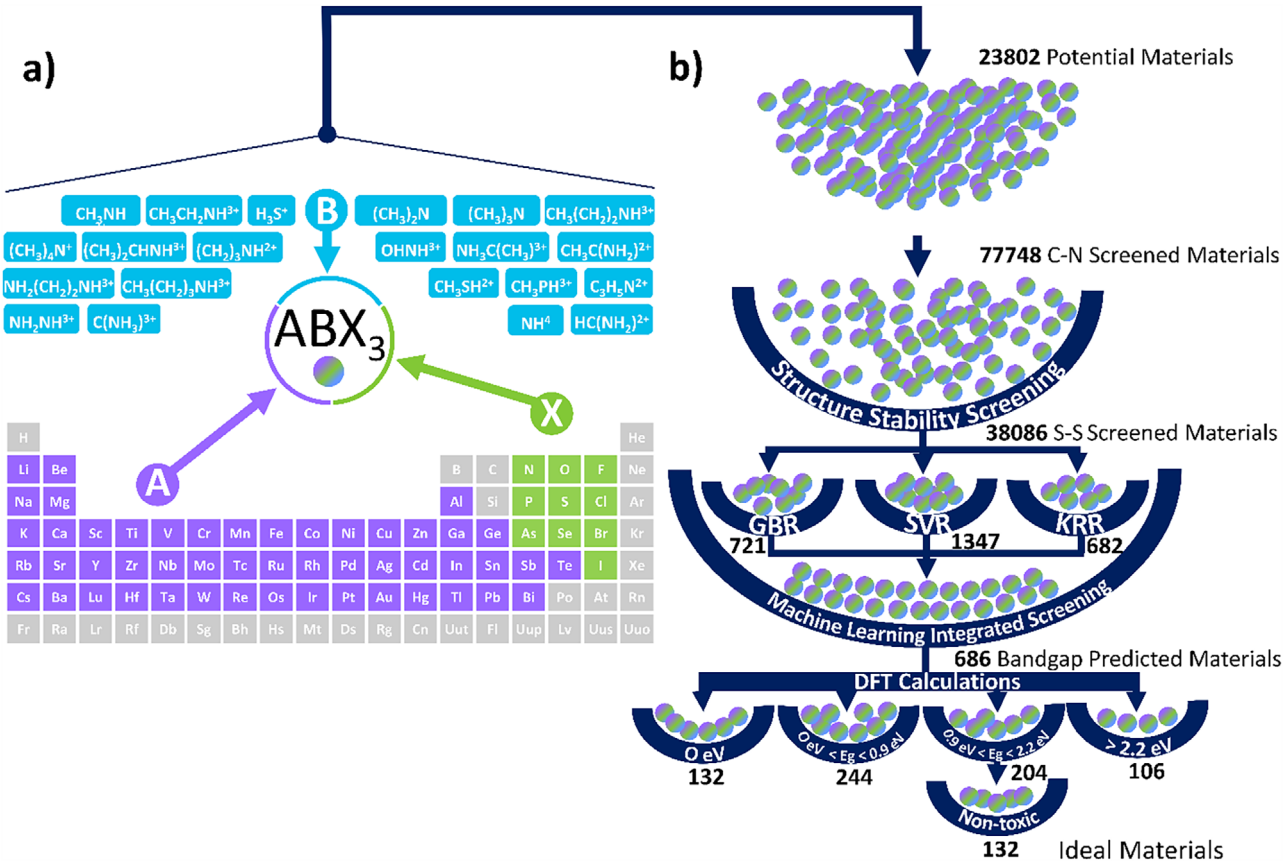

Figure 5. Screening for novel organic halide perovskite, like materials using ML and DFT: (a) alternatives used for A, B, and X; (b) screening steps, and a number of candidates passed those steps. C-N; charge neutrality, S-S; structure stability. *Source:* Adapted from Ref. [56].

high-throughput computations (especially the use of DFT) and high-resolution experimental characterization; the progress in both areas is remarkable, whereas the databases containing such data are also growing continuously.

In material discovery or property estimation applications, computed or measured molecular properties are used as descriptors to estimate the function or performance of a material in a specific use or some other properties that may be linked to the intended function or performance. This can be illustrated using the study performed by Wu et al. to screen the halide perovskite-like materials, which are quite popular in solar cell and photocatalytic applications in recent years [56]. The authors constructed a large dataset containing 230 808 potential materials from the combinations of 21 monovalent organic molecular cations (A-site), 50 metallic cations (B-site), and 10 typical anions from the periodic table (Fig. 5a). In the first step of screening, they used charge neutrality as the selection criterion and reduced the number of potential materials to 77 748 (Fig 5b). Next, they screened the new dataset for stability using the Goldschmidt tolerance factor (T_f) and octahedral factor (O_f) reaching a new list of 38 086 materials. Then, they predicted the bandgap of these materials using ML (gradient boosting regression, support vector machine, and kernel ridge regression—KRR). To do that, they trained their ML models with DFT-generated dataset of 1346 species (having computed bandgap values) using 32 molecular descriptors, including ionic radii, packing factor, tolerance factor, octahedral factor, and electronegativity, together with the elemental properties of A-, B-, and X-sites. The band gap of 686 materials, which were found to be suitable based on ML predictions, was also verified by DFT calculations, and 204 ideal materials (132 of them were non-toxic) with proper band gaps were selected for the final list.

4.2 Performance Analysis/Improvement of Chemical Processes

Performance analysis of chemical processes by ML, often followed by optimization or re-determining the values of descriptors, is another common and probably the easiest application of ML in ChemE. Such analysis involves the development of models to describe the performance measures of the process (may be described in terms of conversion, selectivity, efficiency, purity, and so on) from the descriptors representing the process (like temperature, pressure, feed composition, loading, formulation, and properties of the materials used in the process). Then the model is used to predict the process performance and improve by optimizing or re-configuring the descriptors; it may

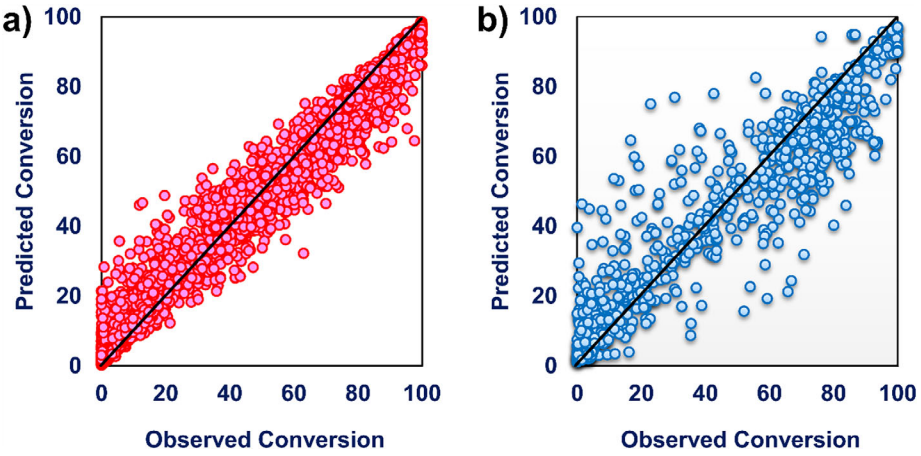

Figure 6. Predicted vs. real conversion results from random forest model for (a) training and (b) testing sets. *Source:* Adapted from Ref. [63].

also help to understand the effects of individual descriptors on the outcome to plan future studies. The data could be created in-house or could be from external sources like papers, patents, or databases. Examples of such work were reported in various fields of ChemE research, like catalysis [57], biofuel production [58], batteries [59], hydrogen storage [60], distillation [61], and PV devices [62]. We can illustrate this using one of our works involving the ML analysis of catalytic CO_2 methanation [63]. The dataset consists of 4049 data points extracted from 526 experiments reported in 100 papers published in 2012–2022 and selected as the result of relevance sorting in the WOS database with the keywords of CO_2 methanation and catalytic "CO_2 hydrogenation" (in April 2022). Total 23 descriptors, including information related to catalyst contents (support, active metal type and composition, promoter's type and content, and so on), preparation and pretreatment conditions (like calcination time and temperature), and reaction conditions (temperature, feed flowrate, catalyst loading, and feed composition), were used to predict CO_2 conversion using random forest algorithm. The experimental versus predicted CO_2 conversion plots for training and testing are presented in Fig. 6, showing a reasonably successful model to represent the experimental works published on this subject in recent years (overpredictions near to zero and underpredictions near to 100 % are inevitable because lower than zero and higher than 100 % conversions are forbidden).

4.3 Process Monitoring, Control, and Optimization

ML has also been implemented for monitoring, controlling, and optimizing the control systems since the 1990s; in fact, they may be the first applications in the ChemE area, and they still constitute an important part of present works. Examples range from fault detection and diagnosis [64] to scheduling [65], and from monitoring and control [66] to real-time optimization of the chemical processes [67]. For instance, Taqvi et al. reviewed various ML techniques such as autoencoders, Bayesian networks, DT, random forests, and artificial neural networks that have been utilized for fault detection and diagnosis of ChemE processes such as Tennessee Eastman process, batch semi-batch reactor systems, distillation column, biochemical wastewater treatment plant, and industrial gas turbines [68]; the same group reviewed the AI-based controllers in petrochemical processes as well [69]. They covered various AI-based controllers such as fuzzy logic control, ANN, reinforcement learning (RL)-based controllers, genetic algorithm-based approaches, and SVM; they also reviewed applications in literature, whereas they discussed the applications in vinyl chloride monomer process in more detail. In another review paper, Melo et al. discussed the tools and industrial applications of data-driven process monitoring and fault diagnosis tools as well as future perspectives and challenges [70].

The new developments in ML, like deep learning and RL, also seemed to create new opportunities for the monitoring, control, and optimization of chemical processes. Deep learning algorithms, which are quite effective in identifying complex patterns in data, are often used for fault diagnosis in chemical processes, especially through the Tennessee Eastman process; for example, Zhang et al. used deep belief network [71], whereas Hao and Jinsong [72] used convolutional neural network for fault diagnosis in the same process. The deep learning algorithms were also used for the fault diagnosis in other processes like reactive distillation [73], fed-batch reactors [74], and battery systems [75].

RL, which is considered the third category of ML together with supervised and unsupervised learning, is quite suitable and therefore utilized extensively in process monitoring and control problems as well. In RL, the model dynamically interacts with the environment, takes actions, and learns through trial and error in a sequential manner (involving the rewarding and punishing the desired and undesired behaviors, respectively) to achieve the desired outcome. It is highly effective for automatic control, scheduling, planning, and logistics, which require sequential decision-making [76]. Sheen et al. suggest that RL can be utilized in process control in three ways: *direct replacement of existing control technologies, integrating RL with model predictive control, and the use of RL in managing the control systems* [77]. Indeed, RL is extensively used in various chemical processes such as batch processes [78], bioprocesses [79], and wastewater treatment [80].

5 A Look Into the Future

In last two decades, ML concept has grown in many ways, including the speed, variety, and effectiveness of algorithms as well as the area of applications from the material research to customer relations and health. Although the progress was remarkable, there were also some critics, including low interpretability and explainability due to its black box nature of ML models, overreliance on data and disregard of physical laws and limits, and suitability for numerical data only while most of the human experience is stored in the form of text in natural human languages. The recent developments indicate that all these shortcomings will be reduced with new tools, as ML is one of the fastest growing fields of scientific research, and it can be expected to grow further in the future. For example, the amount and accuracy of data will increase continuously as the speed and resolution of scientific instruments as well as the hardware and software for computational data generation (like DFT) will continue to improve. The same is also true for data storage and management capabilities, whereas the FARE (findability, accessibility, interoperability, and reusability) [81] data sharing will likely be encouraged more with the increasing pressure for open-access publications; the same trend can be expected for the open ML tools and sources as very large number of open ML libraries are already available in both R and Phyton environment. We can also expect to see new and more effective ML algorithms and data management tools, as evidence form the remarkable success of LLMs in recent years, and predict that ML will be more efficient, easier, cheaper, and have wider acceptance in the future. New applications like transfer learning that allow learning from similar systems [82], algorithms that can be used for small datasets (where the data generation is more difficult), improved interpretability and explainability (against the criticism of the black box nature of ML models) [83], and LLMs based on natural language processing will likely spread more. All these developments will also influence the ChemE field. However, everything comes with some limitations and price to pay. In this section, we will summarize the major developments that can be expected in future as well as challenges and limitations that we may face.

5.1 Toward Field-Specific Databases Involving Chemical Engineering Applications

As we discussed above, a large number of material databases have been developed in recent years. Even though these databases are expected to retain their importance in the future (especially for the initial screening of materials), they will not be sufficient for the solution of complex ChemE problems. To begin with, the properties that can be determined from all-purpose databases are limited; especially the databases created using DFT cannot provide most of the properties (like macro structures, bulk properties of materials, and properties of mixtures) that are important for ChemE applications like catalysis, adsorption/absorption, solutions, and membranes. Furthermore, as a discipline that aims to implement the basic scientific and engineering principles to develop practical products, processes, and solutions, the ChemE problems usually involve processing of the material; hence, processing conditions (like temperature, pressure, concentrations, and contact time) are as influential as the materials. For example, the performance (activity, selectivity, stability, etc.) of a catalyst is not just a function of molecular properties; macro-structural properties (surface area, porosity, surface properties, etc.) and operational conditions (temperature, pressure, time on stream, and so on) are also equally important. Consequently, application-specific databases, containing macro-structural properties and process/system-level information, should also be developed; indeed, some initiatives to construct such databases have already started in various fields such as catalysis [84, 85], perovskite solar cells [86], and batteries [87]. Needless to say that the development of such databases will not be easy due to the nonstandard nature of macro and system-level properties (even the needs are not standard); however, the necessity will likely force the researchers and industrial organizations for large-scale collaborations, which is the basic requirement to develop such complex databases.

5.2 Hybrid/Physics-Informed ML, Transfer Learning, and Automated ML

One of the major weaknesses of purely data-driven ML models is that we can never be sure whether our results are physically plausible and consistent. For example, how an ML model will know the thermodynamic limits of a reaction and will not predict higher conversions than these limits if it uses performance data only. Similarly, how are we going to be sure that our ML model related to momentum or energy transfer does not suggest solutions that violate the conservation of mass principle? There is no guarantee for that, and the ChemE field has a variety of such limits. A hybrid or *physics-informed*-ML approach can help to avoid such inconsistencies by integrating the data and mathematical equations describing the physical laws in a way that the ML model developed will satisfy the data and related physical laws together [88]. Such an approach, allowing first principles knowledge and data complement each other, will also improve our effectiveness for the solution of complex ChemE problems as also discussed by Schweidtmann et al. [89]; it will also improve the *interpretability* of the model and may reduce the common criticism of ML models being *black-box* that cannot provide much insight into the physics involved. Sharma and Liu reviewed the ChemE applications of hybrid science-guided ML; they divided the approach in two categories as "ML complements science" and "science compliments ML" based on the weights of contributions of each side, and discussed various models in each category with the ChemE examples. [2].

Multidimensional nature of engineering problems will also require hybrid approaches involving the use of multiple ML algorithms/tools employing data from multiple sources. For instance, the catalysis can be considered an example for ChemE processes utilizing complex materials. The ML work in the field usually involves the initial screening of potential material candidates first; such task mostly relies on computational and experimental material databases or model structures like pure materials, single or perfect crystals, or smooth surfaces and corners [90]. However, the properties of such simplified structures are not sufficient to explain the functions and performance of

actual catalysts; the macroscopic properties such as surface area, defects, surface groups, pore structure, and metal dispersion are as important as the molecular properties, and they have to be determined from the experimental resources with some assistance from computational tools [91]; meanwhile, the kinetic information, mostly experimental with some computational contribution, has to be also obtained (often the solution of transport equations is needed to account the mass/heat transfer limitations). Finally, the performance of catalyst in the final form must be determined experimentally. All these works, which are the unavoidable tasks in catalysis research, can be assisted by ML algorithms, but not in a single step; instead, various tools, approaches, and data sources should be combined in a well-designed framework [92, 93].

The use of transfer learning, which transfers information from a reliable model to analyze a new but similar field with data scarcity, may also suit well for the ChemE problems that have similarity like the well-established analogies among the momentum, mass, and heat transfer processes. These processes are generally described by complex differential equations with no analytical solutions; consequently, they are also among the fields in which the physic-assisted hybrid ML models are implemented frequently [82]. Transfer learning principles may also be combined with hybrid approaches and may improve the effectiveness of ML models in such complex ChemE fields.

The use of ML will likely be required by many professions, including those that do not have expertise in ML or background to gain such expertise as the result of its rapid expansion into every aspect of life. This will make the concept of automated ML, which aims to automate the implementation steps of ML in a way that even non-experts can utilize it, will be more important in the future [94]. Automation of ML, together with data and code sharing trends discussed above, will increase the accessibility to ML tools by larger groups of people and institution.

5.3 Times of Transition: Rise of Generative AI

Although the ML models involving the analysis (mostly regression) of large datasets continue to be relevant with new improvements (like explainable ML and physics-informed ML), the concept of *generative AI*, referring to the AI tools that can create content themselves, has emerged in last few years and become an instant success. Even though generative AI is not restricted to LLMs, it is mostly popularized by LLMs, especially GPT (generative pre-trained transformer) or ChatGPT, as it is generally referred to. Most of the current (popular) use of these programs involves writing papers and proposals, carrying out literature surveys, or solving student homework and is discussed mostly in ethical and legal grounds; however, their future potential for research and development, which already started to appear, seems to be undeniable considering their success in human-like performance, which is improving in an astonishing speed. For instance, although the number of parameters increased from 1.5 billion to 175 billion from GPT2 to GPT3, it is estimated to exceed one trillion in GPT4 [5]; such models offer great opportunities for research by providing waste amount of knowledge, significant inference capability, and fast connections with web search engines and various calculation tools [95]. Indeed, the number of papers reporting or discussing the applications of LLMs in science and engineering, including ChemE and related fields, has been increased significantly in last few years. Examples are developing synthesis recipes/procedures for thermoelectrical materials [96] and MOFs [97], studying reaction mechanisms [96], optimization of molecular structures [98], automatic generation of control structures for process flow diagrams [99], and autocorrection of chemical process flowsheets [100]; there are also works that suggest the incorporation of LLMs with other approaches and tools [101, 5, 102].

In another group of works, the researchers have been testing the capabilities and limits of LLMs by using them for the solution of known problems or comparing the results with those obtained with other methods. For example, Hatakeyama-Sato et al. evaluated the performance of GPT4 in foundational chemistry knowledge, cheminformatics, data analysis, problem prediction, and proposal abilities [103]. They found that although GPT4 performs very well in understanding the textbook-level chemistry knowledge and related tasks, it fell short in tasks related to specialized knowledge areas and more specific methods. Guo et al. also investigated the performance of GPT4 in eight chemistry-related tasks: (1) name prediction, (2) property prediction, (3) yield prediction, (4) reaction prediction, (5) retrosynthesis (prediction of reactants from products), (6) text-based molecule design, (7) molecule captioning, and (8) reagent selection; they reported that GPT4 outperformed all other models in most of the tasks with few exceptions [104]. In a similar analysis, Deb et al. argued that ChatGPT can be utilized in the tasks like material discovery, structure–property relationships, materials design and optimization, predictive modeling, material characterization, and literature review and knowledge discovery; they also assigned some simple tasks and showed the capabilities of ChatGPT with some minor errors and inconsistencies [105].

Interestingly, there are also significant number of papers suggesting or demonstrating the use of LLM in science and engineering education. In one of such publication, Honig et al. argued that customized ChatGPT can be used for ChemE safety education by integrating the experiences of experienced professional engineers aligned with case studies, replacing traditional presentations with comprehensive engineering meetings, and using generative AI to enhance first two [106]. There are also works that report the use of ChatGPT in designing an industrial dryer in a unit operation course [107], suggesting the use of ChatGPT as an assistant to the students in problem-solving [108] or to the instructors for generating and scoring exams [109].

All these examples reported for the research and education demonstrate that LLMs indeed have significant potential for the future; it is very likely that the situation is also the same in industry even though the experience is not fully reported for obvious reasons. However, it is also clear that there are significant concerns to be resolved in two areas. First, the variety and size of sources feeding LLMs creates some concerns for the reliability of the knowledge created; this is also evidenced by the tests showing much better performance in well-known textbook-level issues, whereas there are more errors and uncertainties in the results obtained in more specific and newly researched areas. Second, there are ethical and legal issues to be clarified to make sure the fair use of intellectual properties and achieve accountability and responsibility of the outcomes.

5.4 ML Application of Chemical Engineering in the Future

We can assume that all the ML-related developments, like the construction of new libraries and databases as well as the progress in new approaches and algorithms, including generative AI tools, will also be applicable to the scientific research and industrial problems that chemical engineers deal with. Additionally, it is easy to predict that ML will be utilized more in the future as well; the researchers and practitioners of ChemE will use ML in a growing number of fields considering the potential contributions of ML to areas like bioinformatics, healthcare, catalysts, batteries, solar cells, fuel cells, polymers, and solvents, which also draw strong attractions from chemical engineers. As we discussed above, the strong tendency to develop field-specific databases and physics-informed ML will also provide great benefit for the complex ChemE problems that involve material and process-related variables together with constraints induced by the physical laws.

What about ChemE itself; will not that change as well? It certainly will, and the ML contribution in the future will be also depended on the trajectory of ChemE. The future of ChemE research was discussed in a recent publication by Torrente-Murciano et al. [110]; various fields, like decarbonization, green energy and other environmental issues, biochemical, bioprocesses, and natural products, processes, operational safety, the pharmaceutical and human health, as well as the process diagnosis, modeling, and control using ML, were underlined as important fields of research in future. All these fields are overlapping significantly with the current fields of ML applications, indicating that there is an alignment with the direction of ChemE research and ML applications. This is also evident from the recent report, *New Directions for Chemical Engineering*, by the National Academy of Science, Engineering, and Medicine [111], which covers the discipline from a broader perspective (not just research). The report seems to represent various issues that have been discussed in recent years and put them together in a comprehensive manner with a future perspective. Briefly, the report lists the following key challenges facing the world and discusses the potential role of ChemE in addressing those challenges: *decarbonization of energy systems, sustainable engineering solutions for environmental systems, engineering targeted and accessible medicine, novel and improved materials for the 21st century, flexible manufacturing and the circular economy, and tools to enable the future of* ChemE. If we just excluded the last item (which already contains ML among the tools that ChemE will be used in the future as suggested in the report), all the other fields listed above are already benefiting from ML extensively. For example, ML has been applied in various energy-related problems like the search for new energy materials (for PVs, photocatalysis, batteries, and so on), performance and state of health studies (in batteries), and demand/supply forecasting and modeling/analyzing environmental pollution [112]. This trend is likely to continue in the future because most of the new energy technologies involve the discovery/development of new material, which is the most common application area for ML in scientific research. This can be extended to other novel materials, as the report listed as other important fields (like catalysts, sorbents, solvents, and membrane materials) for chemical engineers in the future. *Engineering targeted and accessible medicine* has also been investigated significantly for many years by various disciplines, including ChemE. We can expect more progress in the field with further development of *physics-informed* ML (or similar new approaches), because the development of targeted medicine, in principle, requires the consideration of mass, heat, and momentum transfer limitations, which are governed by fundamental conservation laws, in the vessels, tissues, and cells in a human body.

As the initial indicators show, LLMs may also have significant role in the discovery and development of materials for the critical areas such as health, energy, and food supply. Clearly, the success in new material searches will be highly depend on the accessibility to wider and richer domains and effectiveness of the strategies and tools of the search. As we briefly mentioned above, even today, LLMs can reach almost all digitally accessible data, whereas they can utilize very effective tools, including the search engines; furthermore, they can learn and improve their effectiveness fast.

The influence of generative AI on the ChemE education should not be overlook, as well. Apart from the benefits of numerous intended and controlled implementation, as we summarized in the previous section, the LLMs are major players for the life of any college students today; LLMs have a great influence on education through their use in homework and projects, mostly raising ethical and educational concerns.

6 Challenges and Limitations

6.1 Challenges in Data

Most of the ML algorithms rely on statistical learning; hence, as in any other field, the availability of high-quality data is one of the biggest challenges in ML applications in ChemE. Although LLMs seem to achieve the data availability, the quality is still a big concern, especially for the newly researched fields, as was also evident from the examples discussed above; besides, the lack of knowledge about the source of data may also create some questions about the reliability of outcomes of these models. The creation of field-specific databases, as described above, may improve data availability, but their contribution will be also limited for two reasons. First, these datasets will mostly contain academic data with limited contribution from industry due to the trade secrecy. Second, most of the ChemE fields are complex and case dependent; hence, it may not be possible to collect all relevant data in a single database. For instance, almost all material databases (both experimental and computational) contain pure materials and usually rely on the molecular properties. However, many ChemE-related materials (like catalysts, membranes, and solvents) have complex structures made from multiple materials with different functions; besides, the macroscopic properties (such as particle size, pore structure, and surface), which are quite hard to generalize, are also relevant to the task that will be performed. Complexity will increase further when the operational conditions of the processes are considered. Consequently, even though new databases will help to improve

ML models, the data in ChemE and similar fields have to come from multiple sources.

Another problem about the data is that its acquisition from different sources in the same standard/format, which is key in many ML models, is almost impossible due to the complexity of the ChemE materials and processes. For instance, in experimental photocatalysis, the results are not even qualitatively comparable in some reactions. To begin with, the light sources used in the experiments are generally non-standard; the frequency distribution and the intensity of even the same light source (e.g., 300 W xenon light that is usually used to imitate the sunlight) may change from brand to brand (even model to model in the same brand). The orientation and distance of light from the reaction medium, the type and thickness of transparent reactor material, and light adsorption and transmittance properties of reaction medium make the things worse [113]. One can also add the other source of uncertainties, such as sweep gas flowrate and interfacial surface area between the solid–liquid and liquid–gas phases in dead volume [114]. Development of standard testing and reporting protocols may reduce this risk as done in some fields; however, it will always be an issue for the industrial data because the nonstandard (and often incomplete) nature of industrial processes and practices comes with the territory.

6.2 Challenges in Implementation

There are also limits of ML and some pitfalls that should be avoided in implementation. To begin with, most of the ML models rely on past data and can provide solutions within the boundaries of already known spectrum. As the results of their human-like inference capabilities, generative AI tools seem to have the potential to do beyond that; however, they cannot (at least not always) suggest breakthrough solutions, which usually come from unexplored areas, for every problem they encounter (as humans cannot do that either). Even the signs of new developments started to be appeared within the border of data or in searched sources; they may be overlooked because of initial low performance or lower frequency of appearance as statistical inference mostly relies on large data sizes even though new approaches and algorithms are more effective handling the rare appearance as well [115]. Hence, the management of expectations may be another challenge in the future, considering that the use of AI is getting much easier for everyone, including those who do not have much knowledge on its working principles and limitations; we see the evidence of this from the overreliance of college students on the LLM-assisted homework and projects (even papers), which are submitted without checking at least for obvious mistakes stemming from prompt writing. There are also problems that may emerge as the result of wrong or ineffective implementation of ML models. For instance, ML algorithms are often selected based on their popularity at that time instead of the expected benefit or structure of the data, which may be requiring the use of completely different algorithm; furthermore, occasionally, the train-validate-test scheme is not implemented correctly, resulting in models with very low generalizability. Another problem related to implementation is the unintentional use of over-fitted models because the test of such risk is not well defined in some ML algorithms. Finally, the scalability of ML model, referring to the capability of the ML model in handling the large size without compromising the performance, may be a problem for some applications, specially for the data obtained by continuously monitoring the complex industrial processes.

6.3 Legal, Ethical, Societal, and Environmental Issues

Ethical and legal issues are also among the challenges that need to be resolved, especially after LLMs. Under normal conditions, extracting knowledge from any open source should not be unethical or illegal if it is not under legal protection like patents, industrial designs, or trade secrets; the copyright protection, which is the main protection mechanism for the publications, covers the artistic creation (drawings and narrative expressions), not the results reported in a paper. The use of LLMs creates various ambiguities among the border between the fair use of literature and violating intellectual property rights because it also involves the analysis and construction of narrative expressions, which may be partially obtained from another source (meaning they belong to someone else) or created by LLMs (meaning they belong to the owner of LLMs). Another problem is that the user has no control (or even information) on the data source, which may have the potential risk of the use of protected data; furthermore, not only the intellectual rights but also the privacy of the individual have to be protected in the new era of AI tools that can easily reach and exploit any digitally stored data, including those belonging to the people or organizations with limited protection capabilities.

Even if the legal issues are solved completely, there is still problem of uncertainty for the contributions of research and LLMs to the outcome. Normally, the researchers design and control the workflow in research studies, and sometimes, the uniqueness of their approach is the main aspect of novelty for their contribution. Apparently, LLMs will be able to do this in a near future, again creating uncertainties about the contributions of researchers and institutions.

Another important challenge is the high cost of building large AI models and infrastructures, which has to be eventually paid by the users; this may also worsen the inequalities among different regions of the world, as not all institutions and societies have the technological and financial resources to build and maintain such high-tech and high-cost systems. Even the cost of energy required to run such systems can be a problem for many societies. It is reported by the International Energy Agency that the data centers are currently using about 1–1.3 % of world electricity consumptions (please note that energy for cryptocurrency mining is excluded), whereas they are responsible for about 1 % of energy-related GHG emission [116]. These numbers are likely to increase significantly in the future considering the speed of developments in generative AI, and they may also be considered the indirect cost of such systems to entire humanity, whereas one may argue that these models have much higher benefits to justify themselves. As an example, Tomlinson et al. argued that one-page AI-generated text or image actually consumes less energy than those generated by humans [117].

7 Conclusion

One of the first conclusions we can draw is that ML algorithms have been used in ChemE for more than three decades, whereas the use of the ML concept is relatively new. Initial applications were usually in areas such as fault detection, signal processing, and process modeling; after 2000, however, the use of ML spread to more areas, whereas the datasets became larger and more complex. Meanwhile, the ML applications became more comprehensive with an increasing focus on dataset construction, descriptor selection, and pre- and post-analysis. It seems that we are in the onset of bigger changes.

Although the current approach continuously grows in evolutionary manner by more effective algorithms, bigger and richer databases, and new approaches like transfer learning and hybrid models combining data and physical law-based approaches (accounting for the physical laws and limits), more revolutionary developments, like LLMs, are also emerging and spreading fast. The LLMs and other generative AI tools clearly outperform more traditional approaches in terms of number of functions, including human-like inference, number of parameters they use, and number and strength of resources they can employ. However, as with any revolution, they also bring some new problems and uncertainties, like legal and ethical considerations; their potential contributions are also not yet to be fully seen.

ML applications in some ChemE-related areas, such as new energy technologies, environmental issues, new materials developments, and targeted drug developments, will likely grow further while ML becomes an indispensable tool for the future chemical engineers. It is also interesting to note that the new approaches like LLMs seem to also find some roles and influences in the ChemE education as evident from the initial publications, of which the significant portion are on the use LLMs in some aspect of education.

Pınar Özdemir is a Ph.D. candidate in Chemical Engineering Department at Boğaziçi University in Istanbul and recently working with Prof. Dr. Ramazan Yıldırım at Boğaziçi University in Yıldırım Research Group-SOLCAT & Machine Learning laboratory. Her research interests are sustainable energy production by solar energy; photocatalytic water splitting and glycerol reforming, photocatalytic CO_2 reduction, data mining, and machine learning algorithms for sustainable energy production.

Ramazan Yıldırım is a professor in Department of Chemical Engineering at Boğaziçi University. He has a PhD degree from University of California, Los Angeles. His current research areas are photocatalysis, photovoltaics and machine learning applications in renewable energy fields including solar fuels and bioenergy as well as energy storage systems.

Symbols Used

Sub- and Superscripts

O_f	octahedral factor
T_f	Goldschmidt tolerance factor

Abbreviations

AFLOWLIB	Automatic FLOW for Materials Discovery Library
AI	artificial intelligence
ANN	artificial neural networks
ARM	association rule mining
BC	Bayesian classification
ChemE	chemical engineering
C-N	charge neutrality
CSP	concentrated solar power
DT	decision tree
DFT	density functional theory
FARE	findability, accessibility, interoperability, reusability
GBR	gradient boosting regression
GPT	generative pre-trained transformer
ICSD	inorganic crystal structure database
KNN	*k*-nearest neighbor
KRR	kernel ridge regression
LLM	large language model
LOR	logistic regression
ML	machine learning
MR	multiple regression
MOF	metal organic frameworks
NOMAD	novel materials discovery
OQMD	Open Quantum Materials Data
PV	photovoltaic
RF	random forest
RL	reinforcement learning

RT	regression tree
S-S	structure stability
SVM	support vector machine
WOS	Web of Science

References

[1] Z. A. Al-Sai, M. H. Husin, S. M. Syed-Mohamad, R. M. S. Abdin, N. Damer, L. Abualigah, A. H. Gandomi, *BDCC* **2022**, *6* (*4*), 157. DOI: https://doi.org/10.3390/bdcc6040157

[2] V. Venkatasubramanian, *AlChE J.* **2019**, *65*, 466–478. DOI: https://doi.org/10.1002/aic.16489

[3] A. M. Schweidtmann, E. Esche, A. Fischer, M. Kloft, J. U. Repke, S. Sager, A. Mitsos, *Chem. Ing. Tech.* **2021**, *93* (*12*), 2029–2039. DOI: https://doi.org/10.1002/cite.202100083

[4] L. Chiang, B. Lu, I. Castillo, *Annu. Rev. Chem. Biomol. Eng.* **2017**, *8*, 63–85. DOI: https://doi.org/10.1146/annurev-chembioeng-060816-101555

[5] V. Venkatasubramanian, A. Chakraborty, *Comput. Chem. Eng.* **2025**, *192*, 108895. DOI: https://doi.org/10.1016/j.compchemeng.2024.108895

[6] J. Zupan, J. Gasteiger, *Anal. Chim. Acta* **1991**, *248* (*1*), 1–30. DOI: https://doi.org/10.1016/S0003-2670(00)80865-X

[7] J. Gasteiger, J. Zupan, *Angew. Chem., Int. Ed.* **1993**, *32*, 503–527. DOI: https://doi.org/10.1002/anie.199305031

[8] J. A. Burns, G. M. Whitesides, *Chem. Rev.* **1993**, *93* (*8*), 2583–2601. DOI: https://doi.org/10.1021/cr00024a001

[9] D. M. Himmelblau, *Korean J. Chem. Eng.* **2000**, *17* (*4*), 373–392. DOI: https://doi.org/10.1007/BF02706848

[10] J. S. Mattson, C. S. Mattson, M. J. Spencer, F. W. Spencer, *Anal. Chem.* **1977**, *49* (*3*), 500–502. DOI: https://doi.org/10.1021/ac50011a041

[11] E. S. Blurock, *Comput. Chem.* **1995**, *19* (*2*), 91–99. DOI: https://doi.org/10.1016/0097-8485(95)00001-9

[12] I. E. Frank, *Chemom. Intell. Lab. Syst.* **1995**, *27*, 1–19. DOI: https://doi.org/10.1016/0169-7439(95)80003-R

[13] M. Erdem Günay, R. Yıldırım, *Catal. Rev.: Sci. Eng.* **2021**, *63*, 120–164. DOI: https://doi.org/10.1080/01614940.2020.1770402

[14] C. S. Greene, J. Tan, M. Ung, J. H. Moore, C. Cheng, *J. Cell. Physiol.* **2015**, *229* (*12*), 1896–1900. DOI: https://doi.org/10.1002/jcp.24662

[15] B. Yılmaz, Ç. Odabaşı, R. Yıldırım, *Energy Technol.* **2022**, *10* (*3*), 2100948. DOI: https://doi.org/10.1002/ente.202100948

[16] Y. Liu, Q. Zhou, G. Cui, *Small Methods* **2021**, *5* (*8*), 2100442. DOI: https://doi.org/10.1002/smtd.202100442

[17] A. Coşgun, M. E. Günay, R. Yıldırım, *Green Chem.* **2023**, *25* (*9*), 3354–3373. DOI: https://doi.org/10.1039/D3GC00389B

[18] S. Rangarajan, A. Bhan, P. Daoutidis, *Comput. Chem. Eng.* **2012**, *45*, 114–123. DOI: https://doi.org/10.1016/j.compchemeng.2012.06.008

[19] M. Liao, F. Wu, X. Yu, L. Zhao, H. Wu, J. Zhou, *J. Solution Chem.* **2023**, *52* (*4*), 487–498. DOI: https://doi.org/10.1007/s10953-023-01247-6

[20] Z. Eddaoudi, Z. Aarab, K. Boudmen, A. Elghazi, M. D. Rahmani, *Procedia Comput. Sci.* **2023**, *236* (*C*), 33–40. DOI: https://dl.acm.org/doi/10.1016/j.procs.2024.05.001

[21] B. I. Oladapo, M. A. Olawumi, F. T. Omigbodun, *Atmosphere* **2024**, *15* (*10*), 1250. DOI: https://doi.org/10.3390/atmos15101250

[22] J. Segarra-Tamarit, E. Pérez, E. Moya, P. Ayuso, H. Beltran, *Comput. Simul.* **2021**, *184*, 306–318. DOI: https://doi.org/10.1016/j.matcom.2020.02.007

[23] Ç. Odabaşı, R. Yıldırım, *Nano Energy* **2019**, *56*, 770–791. DOI: https://doi.org/10.1016/j.nanoen.2018.11.069

[24] M. E. Günay, N. A. Tapan, G. Akkoç, *Int. J. Hydrogen Energy* **2022**, *47* (*4*), 2134–2151. DOI: https://doi.org/10.1016/j.ijhydene.2021.10.191

[25] V. K. Chauhan, S. K. Shukla, J. V. Tirkey, P. K. Singh Rathore, *J. Clean Prod.* **2021**, *284*, 124719. DOI: https://doi.org/10.1016/j.jclepro.2020.124719

[26] V. G. Sharmila, S. P. Shanmugavel, J. R. Banu, *Biomass Bioenergy* **2024**, *180*, 106997. DOI: https://doi.org/10.1016/j.biombioe.2023.106997

[27] A. Chen, P. K. Sen, Advancement in battery technology: A state-of-the-art review, 2016 IEEE Industry Applications Society Annual Meeting, Portland, OR, USA, **2016**, pp. 1–10, DOI: https://doi.org/10.1109/IAS.2016.7731812.

[28] D. T. Larose, C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd ed., JohnWiley & Sons, NewJersey **2014**.

[29] T. Williams, K. McCullough, J. A. Lauterbach, *Chem. Mater.* **2020**, *32* (*1*), 157–165. DOI: https://doi.org/10.1021/acs.chemmater.9b03043

[30] Z. Li, S. Wang, W. S. Chin, L. E. Achenie, H. Xin, *J. Mater. Chem. A* **2017**, *5* (*46*), 24131–24138. DOI: https://doi.org/10.1039/C7TA01812F

[31] P. Villars, N. Onodera, S. Iwata, *J. Alloys Compd.*, **1998**, *279*, 1–7, DOI: https://doi.org/10.1016/S0925-8388(98)00605-7.

[32] G. Bergerhoff, R. Hundt, R. Sievers, I. D. Brown, *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 66–69. DOI: https://doi.org/10.1021/ci00038a003

[33] https://www.crystalimpact.com/pcd/ (May 07, 2024)

[34] F. H. Allen, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2002**, *58*, 380–388. DOI: https://doi.org/10.1107/S0108768102003890

[35] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail, *J. Appl. Crystallogr.* **2009**, *42*, 726–729. DOI: https://doi.org/10.1107/S0021889809016690

[36] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, *44*, D1202-D1213. DOI: https://doi.org/10.1093/nar/gkv951

[37] A. Jain, G. Hautier, S. P. Ong, K. Persson, *J. Mater. Res.* **2016**, *31* (*8*), 977–994. DOI: https://doi.org/10.1557/jmr.2016.80

[38] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **2012**, *58*, 227–235. DOI: https://doi.org/10.1016/j.commatsci.2012.02.002

[39] T. R. Munter, D. D. Landis, F. Abild-Pedersen, G. Jones, S. Wang, T. Bligaard, *Comput. Sci. Discovery* **2009**, *2*, 015006. DOI: https://doi.org/10.1088/1749-4699/2/1/015006

[40] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *npj Comput. Mater.* **2015**, *1*, 15010. DOI: https://doi.org/10.1038/npjcompumats.2015.10

[41] https://nomad-repository.eu/ (Accessed on February 28, 2024)

[42] B. Yılmaz, R. Yıldırım, *Nano Energy* **2021**, *80*, 105546. DOI: https://doi.org/10.1016/j.nanoen.2020.105546
[43] D. Saadetnejad, B. Oral, E. Can, R. Yıldırım, *Int. J. Hydrogen Energy* **2022**, *47* (*45*), 19655–19668. DOI: https://doi.org/10.1016/j.ijhydene.2022.02.030
[44] B. Oral, E. Can, R. Yildirim, *Int. J. Hydrogen Energy* **2022**, *47* (*45*), 19633–19654. DOI: https://doi.org/10.1016/j.ijhydene.2022.01.011
[45] E. Alpaydın, *Introduction to Machine Learning*, 3rd ed., The MIT Press, Cambridge, MA **2014**.
[46] J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Boston, MA **1977**.
[47] R. Schmack, A. Friedrich, E. V. Kondratenko, J. Polte, A. Werwatz, R. Kraehnert, *Nat. Commun.* **2019**, *10* (*1*), 67. DOI: https://doi.org/10.1038/s41467-019-08325-8
[48] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **2013**, *3*, 2810. DOI: https://doi.org/10.1038/srep02810
[49] K. Stergiou, C. Ntakolia, P. Varytis, E. Koumoulos, P. Karlsson, S. Moustakidis, *Comput. Mater. Sci.* **2023**, *220*, 112031. DOI: https://doi.org/10.1016/j.commatsci.2023.112031
[50] J. Cai, X. Chu, K. Xu, H. Li, J. Wei, *Nanoscale Adv.* **2020**, *2* (*8*), 3115–3130. DOI: https://doi.org/10.1039/D0NA00388C
[51] P. Schlexer Lamoureux, K T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, *ChemCatChem* **2019**, *11* (*16*), 3581–3601. DOI: https://doi.org/10.1002/cctc.201900595
[52] J. Yang, L. Tao, J. He, J. R. McCutcheon, Y. Li, *Sci. Adv.* **2022**, *8* (*29*), eabn9545. DOI: http://doi.org/10.1126/sciadv.abn9545
[53] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, *Matter* **2021**, *4* (*5*), 1578–1597. DOI: https://doi.org/10.1016/j.matt.2021.02.015
[54] A. Kilic, B. Oral, D. Eroglu, R. Yildirim, *J. Energy Storage* **2023**, *73*, 109057. DOI: https://doi.org/10.1016/j.est.2023.109057
[55] Q. Tao, P. Xu, M. Li, W. Lu, *npj Comput. Mater.* **2021**, *7* (*1*), 23. DOI: https://doi.org/10.1038/s41524-021-00495-8
[56] T. Wu, J. Wang, *Nano Energy* **2019**, *66*, 104070. DOI: https://doi.org/10.1016/j.nanoen.2019.104070
[57] F. M. Cavalcanti, M. Schmal, R. Giudici, R. M. Brito Alves, *J. Environ. Manage.* **2019**, *237*, 585–594. DOI: https://doi.org/10.1016/j.jenvman.2019.02.092
[58] A. Coşgun, M. E. Günay, R. Yıldırım, *Fuel* **2022**, *315*, 122817. DOI: https://doi.org/10.1016/j.fuel.2021.122817
[59] A. Kilic, R. Yildirim, D. Eroglu, *Int. J. Energy Res.* **2022**, *46* (*15*), 21716–21726. DOI: https://doi.org/10.1002/er.8611
[60] H Vo Thanh, S. Ebrahimnia Taremsari, B. Ranjbar, H. Mashhadimoslem, E. Rahimi, M. Rahimi, A. Elkamel, *Energies* **2023**, *16* (*5*), 2348. DOI: https://doi.org/10.3390/en16052348
[61] H. Kwon, K. C. Oh, Y. Choi, Y. G. Chung, J. Kim, *Int. J. Intell. Syst.* **2021**, *36* (*5*), 1970–1997. DOI: https://doi.org/10.1002/int.22368
[62] G. Li, Z. Su, M. Li, H. K. H Lee, R. Datt, D. Hughes, C. Wang, M. Flatken, H. Köbler, J. J. Jerónimo-Rendon, R. Roy, F. Yang, J. Pascual, Z. Li, W. C. Tsoi, X. Gao, Z. Wang, M. Saliba, A. Abate, *Adv. Energy Mater.* **2022**, *12* (*48*), 2202887. DOI: https://doi.org/10.1002/aenm.202202887
[63] B. Yılmaz, B. Oral, R. Yıldırım, *Int. J. Hydrogen Energy* **2023**, *48* (*64*), 24904–24914. DOI: https://doi.org/10.1016/j.ijhydene.2022.12.197
[64] S. A. A. Taqvi, H. Zabiri, F. Uddin, M. Naqvi, L. D. Tufa, M. Kazmi, S. Rubab, S. R. Naqvi, A. S. Maulud, *Energy Sci. Eng.* **2022**, *10* (*3*), 814–839. DOI: https://doi.org/10.1002/ese3.1058
[65] G. Campos, N. H. El-Farra, A. Palazoglu, *Ind. Eng. Chem. Res.* **2022**, *61* (*24*), 8443–8461. DOI: https://doi.org/10.1021/acs.iecr.1c04984
[66] J. C. B. Gonzaga, L. A. C. Meleiro, C. Kiang, R. Maciel Filho, *Comput. Chem. Eng.* **2009**, *33* (*1*), 43–49. DOI: https://doi.org/10.1016/j.compchemeng.2008.05.019
[67] B. K. M. Powell, D. Machalek, T. Quah, *Comput. Chem. Eng.* **2020**, *143*, 107077. DOI: https://doi.org/10.1016/j.compchemeng.2020.107077
[68] S. A. A. Taqvi, H. Zabiri, L. D. Tufa, F. Uddin, S. A. Fatima, A. S. Maulud, *ChemBioEng Rev.* **2021**, *8* (*3*), 239–259. DOI: https://doi.org/10.1002/cben.202000027
[69] T. S. Ansari, S. A. A. Taqvi, *ChemBioEng Rev.* **2023**, *10* (*6*), 884–906. DOI: https://doi.org/10.1002/cben.202300017
[70] A. Melo, M. M. Câmara, J. C. Pinto, *Processes* **2024**, *12* (*2*), 251. DOI: https://doi.org/10.3390/pr12020251
[71] Z. Zhang, J. Zhao, *Comput. Chem. Eng.* **2017**, *107*, 395–407. DOI: https://doi.org/10.1016/j.compchemeng.2017.02.041
[72] H. Wu, J. Zhao, *Comput. Chem. Eng.* **2018**, *115*, 185–197. DOI: https://doi.org/10.1016/j.compchemeng.2018.04.009
[73] X. Ge, B. Wang, X. Yang, Y. Pan, B. Liu, B. Liu, *Comput. Chem. Eng.* **2021**, *145*, 107172. DOI: https://doi.org/10.1016/j.compchemeng.2020.107172
[74] D. Hematillake, D. Freethy, J. McGivern, C. McCready, P. Agarwal, H. Budman, *Ind. Eng. Chem. Res.* **2022**, *61* (*13*), 4625–4637. DOI: https://doi.org/10.1021/acs.iecr.1c04534
[75] J. Zhang, Y. Wang, B. Jiang, H. He, S. Huang, C. Wang, Y. Zhang, X. Han, D. Guo, G. He, M. Ouyang, *Nat. Commun.* **2023**, *14* (*1*), 5940. DOI: https://doi.org/10.1038/s41467-023-41226-5
[76] J. H. Lee, J. Shin, M. J. Realff, *Comput. Chem. Eng.* **2018**, *114*, 111–121. DOI: https://doi.org/10.1016/j.compchemeng.2017.10.008
[77] J. Shin, T. A. Badgwell, K. H. Liu, J. H. Lee, *Comput. Chem. Eng.* **2019**, *127*, 282–294. DOI: https://doi.org/10.1016/j.compchemeng.2019.05.029
[78] H. Yoo, H. E. Byun, D. Han, J. H. Lee, *Annu. Rev. Control* **2021**, *52*, 108–119. DOI: https://doi.org/10.1016/j.arcontrol.2021.10.006
[79] P. Petsagkourakis, I. O. Sandoval, E. Bradford, D. Zhang, E. A. del Rio-Chanona, *Comput. Chem. Eng.* **2020**, *133*, 106649. DOI: https://doi.org/10.1016/j.compchemeng.2019.106649
[80] H. C. Croll, K. Ikuma, S. K. Ong, S. Sarkar, *Crit. Rev. Environ. Sci. Technol.* **2023**, *53* (*20*), 1775–1794. DOI: https://doi.org/10.1080/10643389.2023.2183699
[81] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **2016**, *3*, 160018. DOI: https://doi.org/10.1038/sdata.2016.18

[82] X. Li, Y. Dan, R. Dong, Z. Cao, C. Niu, Y. Song, S. Li, J. Hu, *Appl. Sci.* **2019**, *9* (*24*), 5510. DOI: https://doi.org/10.3390/app9245510

[83] X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, T. Y. J. Han, *npj Comput. Mater.* **2022**, *8* (*1*), 204. DOI: https://doi.org/10.1038/s41524-022-00884-7

[84] K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich, T. Bligaard, *Sci. Data* **2019**, *6* (*1*), 75. DOI: https://doi.org/10.1038/s41597-019-0081-y

[85] P. S. F. Mendes, S. Siradze, L. Pirro, J. W. Thybaut, *ChemCatChem* **2021**, *13* (*3*), 836–850. DOI: https://doi.org/10.1002/cctc.202001132

[86] T. J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A. G. Ricciardulli, A. Vijayan, A. Kulkarni, A. Y. Anderson, B. P. Darwich, B. Yang, B. L. Coles, C. A. R. Perini, C. Rehermann, D. Ramirez, D. Fairen-Jimenez, D. Di Girolamo, D. Jia, E. Avila, E. J. Juarez-Perez, F. Baumann, F. Mathies, G. S. A. González, G. Boschloo, G. Nasti, G. Paramasivam, G. Martínez-Denegri, et al., *Nat. Energy* **2022**, *7* (*1*), 107–115. DOI: https://doi.org/10.1038/s41560-021-00941-3

[87] L. Ward, S. Babinec, E. J. Dufek, D. A. Howey, V. Viswanathan, M. Aykol, D. A. C. Beck, B. Blaiszik, B. R. Chen, G. Crabtree, S. Clark, V. De Angelis, P. Dechent, M. Dubarry, E. E. Eggleton, D. P. Finegan, I. Foster, C. B. Gopal, P. K. Herring, V. W. Hu, N. H. Paulson, Y. Preger, D. Uwe-Sauer, K. Smith, S. W. Snyder, S. Sripad, T. R. Tanim, L. Teo, *Joule* **2022**, *6* (*10*), 2253–2271. DOI: https://doi.org/10.1016/j.joule.2022.08.008

[88] R. J. Chimentão, B. C. Miranda, J. Szanyi, C. Sepulveda, J. B. O. Santos, J. V. S. Correa, J. Llorca, F. Medina, *Mol. Catal.* **2017**, *435*, 49–57. DOI: https://doi.org/10.1016/j.mcat.2017.03.023

[89] A. M. Schweidtmann, E. Esche, A. Fischer, M. Kloft, J. U. Repke, S. Sager, A. Mitsos, *Chem. Ing. Tech.* **2021**, *93* (*12*), 2029–2039. DOI: https://doi.org/10.1002/cite.202100083

[90] M. Erdem Günay, R. Yıldırım, *Catal. Rev.* **2021**, *63*, 120–164. DOI: https://doi.org/10.1080/01614940.2020.1770402

[91] L. Luo, X. Liu, J. Zhang, J. Yao, B. Liu, J. Zhang, H. Wang, S. Lu, Y. Xiang, *Chem. Eng. Sci.* **2025**, *302* (*A*), 120830. DOI: https://doi.org/10.1016/j.ces.2024.120830

[92] S. Özsoysal, B. Oral, R. Yıldırım, *J. Mater. Chem. A* **2024**, *12*, 5748–5759. DOI: https://doi.org/10.1039/D3TA07001H

[93] J. Sun, R. Tu, Y. Xu, H. Yang, T. Yu, D. Zhai, X. Ci, W. Deng, *Nat. Commun.* **2024**, *15*, 6036. DOI: https://doi.org/10.1038/s41467-024-50417-7

[94] M. Baratchi, C. Wang, S. Limmer, J. N. van Rijn, H. Hoos, T. Bäck, M. Olhofer, *Artif. Intell. Rev.* **2024**, *57*, 122. DOI: https://doi.org/10.1007/s10462-024-10726-1

[95] K. Hatakeyama-Sato, N. Yamane, Y. Igarashi, Y. Nabae, T. Hayakawa, *Sci. Technol. Adv. Mater.* **2023**, *3* (*1*), 2260300. DOI: https://doi.org/10.1080/27660400.2023.2260300

[96] G. S. Na, *Chem. Mater.* **2023**, *35* (*19*), 8272–8280. DOI: https://doi.org/10.1021/acs.chemmater.3c01834

[97] Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, O. M. Yaghi, *J. Am. Chem. Soc.* **2023**, *145* (*32*), 18048–18062. DOI: https://doi.org/10.1021/jacs.3c05819

[98] Z. Wu, O. Zhang, X. Wang, L. Fu, H. Zhao, J. Wang, H. Du, D. Jiang, Y. Deng, D. Cao, C.-Y. Hsieh, T. Hou, *Nat. Mach. Intell.* **2024**, *6*, 1359–1369. DOI: https://doi.org/10.1038/s42256-024-00916-5

[99] E. Hirtreiter, L. Schulze Balhorn, A M. Schweidtmann, *AlChE J.* **2023**, *70* (*1*), e18259. DOI: https://doi.org/10.1002/aic.18259

[100] L. S. Balhorn, M. Caballero, A. M. Schweidtmann, *Comput. Aided Chem. Eng.* **2023**, *53*, 3109–3114. DOI: https://doi.org/10.1016/B978-0-443-28824-1.50519-6

[101] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A D. White, P. Schwaller, *Nat. Mach. Intell.* **2024**, *6*, 525–535. DOI: https://doi.org/10.1038/s42256-024-00832-8

[102] D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, *Nature* **2023**, *624*, 570–578. DOI: https://doi.org/10.1038/s41586-023-06792-0

[103] Z. Wu, O. Zhang, X. Wang, L. Fu, H. Zhao, J. Wang, H. Du, D. Jiang, Y. Deng, D. Cao, C.-Y. Hsieh, T. Hou, *Nat. Mach. Intell.* **2024**, *6*, 1359–1369. DOI: https://doi.org/10.1038/s42256-024-00916-5

[104] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. V. Chawla, O. Wiest, X. Zhang, *arXiv* **2023**, 1. DOI: https://doi.org/10.48550/arXiv.2305.18365

[105] J. Deb, L. Saikia, K. D. Dihingia, G. N. Sastry, *J. Chem. Inf. Model.* **2024**, *64* (*3*), 799–811. DOI: https://doi.org/10.1021/acs.jcim.3c01702

[106] C. D. F. Honig, A. Desu, J. Franklin, *Educ. Chem. Eng.* **2025**, *49*, 55–66. DOI: https://doi.org/10.1016/j.ece.2024.09.001

[107] B. Ramos, R. Condotta, *J. Chem. Educ.* **2024**, *101* (*8*), 3246–3254. DOI: https://doi.org/10.1021/acs.jchemed.4c00244

[108] M.-L. Tsai, C. W. Ong, C.-L. Chen, *Educ. Chem. Eng.* **2023**, *44*, 71–95. DOI: https://doi.org/10.1016/j.ece.2023.05.001

[109] A. A. Fernández, M. López-Torres, J. J. Fernández, D. Vázquez-García, *J. Chem. Educ.* **2024**, *101*, 3780–3788. DOI: https://doi.org/10.1021/acs.jchemed.4c00231

[110] L. Torrente-Murciano, J. B. Dunn, P. D. Christofides, J. D. Keasling, S. C. Glotzer, S. Y. Lee, K. M. Van Geem, J. Tom, G. He, *Nat. Chem. Eng.* **2024**, *1* (*1*), 18–27. DOI: https://doi.org/10.1038/s44286-023-00017-x

[111] National Academies of Sciences, Engineering and Medicine, *New Directions for Chemical Engineering*, The National Academies Press, Washington, DC **2022**, DOI: https://doi.org/10.17226/26342.

[112] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S. P. Ong, *Adv. Energy Mater.* **2020**, *10* (*8*), 1903242. DOI: https://doi.org/10.1002/aenm.201903242

[113] M. Melchionna, P. Fornasiero, *ACS Catal.* **2020**, *10* (*10*), 5493–5501. DOI: https://doi.org/10.1021/acscatal.0c01204

[114] E. Can Özcan, D. Uner, R. Yildirim, *Int. J. Hydrogen Energy* **2024**, *75*, 540–546. DOI: https://doi.org/10.1016/j.ijhydene.2024.03.218

[115] C. Shyalika, R. Wickramarachchi, A. P. Sheth, *ACM Comput. Surv.* **2024**, *57* (*3*), 1–39. DOI: https://doi.org/10.1145/3699955

[116] https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks (Accessed on February 19, 2025)

[117] B. Tomlinson, R. W. Black, D. J. Patterson, A. W. Torrance, *Sci. Rep.* **2024**, *14*, 3732. DOI: https://doi.org/10.1038/s41598-024-54271-x

Although the initial machine learning (ML) applications were mainly on fault detection, signal processing, and process modeling, they extended to new areas like property estimation and material screening in later years; energy technologies, environmental issues, health, and new materials will likely be more important in future with the use of larger databases and new approaches like physics-informed ML and generative AI.

ML@ChemE: Past, Present, and Future of Machine Learning in Chemical Engineering

Pınar Özdemir,
Ramazan Yıldırım*

ChemBioEng Rev. **2025**, *00 (0)*, **e70012**

DOI: 10.1002/cben.70012