

Application of machine learning in chemical engineering: outlook and perspectives

Ashraf Al Sharah¹, Hamza Abu Owida², Feras Alnaimat², Mohammad Hassan³, Suhaila Abuowaida⁴
Mohammad Alhaj⁵, Ahmad Sharadqeh¹

¹Department of Electrical Engineering, College of Engineering Technology, Al-Balqa Applied University, Amman, Jordan

²Department of Medical Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan

³Department of Communications and Computer Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan

⁴Department of Computer Science, Faculty of Information Technology, Zarqa University, Zarqa, Jordan

⁵IEEE Member

Article Info

Article history:

Received Apr 9, 2023

Revised Sep 23, 2023

Accepted Nov 6, 2023

Keywords:

Applications

Chemical engineering

Machine learning

Models

Optimization

ABSTRACT

Chemical engineers' formulation, development, and stance processes all heavily rely on models. The physical and economic consequences of these decisions can have disastrous effects. Attempts to employ a hybrid form of artificial intelligence for modeling in various disciplines. However, they fell short of expectations. Due to a rise in the amount of data and computational resources during the previous five years. A lot of recent work has gone into developing new data sources, indexes, chemical interface designs, and machine learning algorithms in an effort to facilitate the adoption of these techniques in the research community. However, there are some important downsides to machine learning gains. The most promising uses for machine learning are in time-critical tasks like real-time optimization and planning that require extreme precision and can build on models that can self-learn to recognize patterns, draw conclusions from data, and become more intelligent over time. Due to their limited exposure to computer science and data analysis, the majority of chemical engineers are potentially vulnerable to the development of artificial intelligence. But in the not-too-distant future, chemical engineers' modeling toolbox will include a reliable machine learning component.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ashraf Al Sharah

Department of Electrical Engineering, College of Engineering Technology, Al-Balqa Applied University
Amman 001962, Jordan

Email: aalsharah@bau.edu.jo

1. INTRODUCTION

For the past 130 years, mathematical modelling has been a crucial tool in chemical engineering, allowing engineers to quickly identify and design chemical processes [1], [2]. Keeping up with the ever-changing demands of today's world is harder than ever. No matter if you're trying to discover and synthesize active pharmaceutical ingredients to treat new diseases or increase process efficiency to conform to stricter environmental legislation, the ability to predict the outcomes of certain events is essential. The efficiency of a chemical interaction, the choice of a reactor, and the regulation of a heat source are all examples. Theories that have been refined over time of several centuries, one can make predictions [3]–[5]. Consequently, for reasonable processes, several of these models can somehow be modeled mathematically and necessitate a great deal of supercomputing capacity to solve numerical results. Because of this limitation, most engineers resort to more elementary models when attempting to explain the world around them. Prandtl's boundary layer model [6] is a notable example of a model from the

past that is still useful today. Scientists and engineers in the field of computational chemistry often compromise precision for the sake of efficiency. It is because of this openness that concentration structural functionalism has become so widely used in place of more advanced theoretical models.

On the other hand, there are many scenarios where greater precision is preferred. Scientists and engineers in the field of chemical technology now has access to a wealth of data collected over many years of modeling, simulation, and experimentation, giving researchers an additional modeling tool in the form of the ability to draw on prior experience to make predictions. To put it simply, machine learning models are a subset of statistical and mathematical models that can "learn" from data by watching their environment and "uncovering" relationship between the data without the usage of predetermined rules. Figure 1 shows the block diagram explains the working of machine learning algorithm.

Machine learning is an AI subfield (AI). Definition of artificial intelligence (AI): the computer's ability to mimic human thought and behavior in certain situations. The study of these topics is not a cutting-edge endeavour. The term "artificial intelligence" was first used in 1956 at a summer meeting at Dartmouth College, USA, attended by mathematicians interested in developing smarter robots.

Efforts to implement AI in the field of chemical engineering didn't begin to gain traction until well after the year 2000 [7]. Rule-based expert systems, one of the earliest and most basic types of AI, saw increased use in the field during the 1980s. The field of machine learning had already begun to expand by that point, but with a few notable exceptions, the chemical engineering community lagged behind by roughly ten years. During the 1990s, there was a rapid uptick in articles on artificial intelligence progress in the field of chemical engineering due to the widespread use of cluster analysis, optimization computation, and, most effectively, artificial neural networks (ANNs) (ANNs). However, this fad did not last, and experts point to the absence of potent modelling and the challenging charge of developing the algorithms as potential reasons. In the past ten years, advancements have been made in deep learning, a branch of machine learning that builds ANNs to simulate the human brain. While ANNs did see increased adoption in the 1990s, the advent of the deep learning era made it possible to develop multi-layered neural networks, or "deep neural networks," which had previously been computationally prohibitive.

Chemical engineers were set off by these innovations, as evidenced by the meteoric rise in related research papers. The question of whether or not artificial intelligence techniques have advanced to the point where they can be considered a typical chemical engineering tool [7]–[9]. In this review article, we will begin by providing an overview of the three major links that currently exist in machine learning as it relates to chemical engineering. We will examine the benefits and drawbacks of machine learning in chemical engineering, presenting a set of hypotheses explaining why machine learning might be useful in the field of chemical engineering, will continue to either be a "hot" topic or become obsolete in the near future.



Figure 1. The block diagram explains the working of machine learning algorithm

2. MACHINE LEARNING; DATA, REPRESENTATIONS, AND MODELS

Machine learning relies on three main components: data, representations, and models. In order to train a machine learning model, it is necessary to first collect the necessary training data. We'll get into how the data ends up being the machine learning process's biggest flaw in its own right later on. All sorts of data, from experiments to theoretical findings to simulation results, can be used to train a model. However, due to the high cost of data collection in massive quantities, it is common practice to employ big data, which involve the use of large databases culled from a variety of existing sources. Because of the high cost of conducting actual experiments, to get such enormous amounts of data, experts use quick algorithms or information extraction of trademarks and publications. Due to researchers' growing comfort with using digital tools, there are now many free and paid databases available to them [10]–[12]. In order to compare the performance of different machine learning models, several benchmark datasets have been created. Standard reference materials for quantum chemical properties include QM9 and Alchemy [13], while standard reference materials for solubility include estimated SOLubility (ESOL) [14], and FreeSolv [15]. Before incorporating a dataset into a model based on machine learning, there are a number of checks and checks and balances that need to be carried out to guarantee adequate system performance. Data curation refers to the process of monitoring and maintaining data quality at every stage, from creation to archiving. When it comes to how they utilize data, machine learning and deep learning approaches differ significantly from more conventional forms of modeling. To begin with, ANNs are

capable of self-learning and training, but this process requires a large amount of data. Therefore, enormous amounts of data points are typical in training datasets. Second, instead of dividing the dataset in half, it is divided in three: the training set, the validation set, and the test set. In contrast to training data, verification data is kept separately used to objectively assess the training phase model's accuracy. The test set is the main indicator of model quality because it evaluates the final model fit with unseen data [9], [10], [16].

A machine learning method's representation of data in the model is also essential. Even when the data is already numeric, the model's performance can be greatly affected by the variables or features chosen to comprise the input. The process of feature selection has been the subject of research in a number of published works [9], [17]. Time and money could be saved by reducing the number of features used in training if the model's accuracy is not compromised. Deep learning methods weigh feature selection less. Thus, convolution layers of basic process variables [18]–[20]. Representing non-numerical data such molecules and reactions is much harder.

Molecules and/or chemical reactions are frequently involved in chemical engineering tasks. Until reliable numerical representations of these datasets are established, they cannot be used. Common methods of representing molecular structure in software include line-based identifiers like the simplified molecular-input lineentry system (SMILES) or the international union of pure and applied chemistry (IUPAC) international chemical identifiers (InChIs) [21], [22] or as three-dimensional (3D) coordinates.

Recently, a molecular string representation tailored to machine learning applications called self-referencing embedded strings (SELFIES) [23] has been developed. As input to a deep neural network or other machine learning model, the molecular data is transformed into a feature vector or tensor. The molecular weight, dipole moment, and dielectric constant are examples of good molecular descriptors that can be used to represent molecules [24], [25].

The 3D geometry of the molecule can also be used as a starting point for generating a feature vector. Examples of geometry-based representations include coulomb matrices [26], bond bags [27], and distance, angle, and dihedral histograms [28]. Nonetheless, many uses don't have access to 3D coordinates or calculated properties. When this is the case, so-called topology-based representations can be built from a molecular graph [29].

The only form of identification possible in topological representations is a line label. Natural language processing (NLP) techniques can be used by some encoders to directly convert the line-based identifier into a representation [30], [31]. This is accomplished by adding some simple characteristics to the linear combination, such as particles and interactions, and then passing data back and forward between them in an incremental way [32].

Some of the earliest molecular representations used in machine learning were circular fingerprints [33], [34] constructed using the Morgan algorithm [35], including the extended-connectivity fingerprint [36]. Due to the fact that they remain unchanged throughout the machine learning model's training process, these fingerprints are known as fixed molecular representations. They continue to be widely used in drug design because of their speed and accuracy in predicting physical, chemical, and biological properties of potential new drugs [37]. Given that the definition of a deep neural network assumes it will learn the important features, a fixed representation vector's use as an input layer seems at odds with this assumption [38], [39], so the focus has shifted from manually engineering the feature vector to learning how to represent a molecule [40]. To aid in this foresight, a model is constructed that incorporates learned molecular representations. Through training, a molecular representation is constructed and refined, beginning with elementary properties of molecules like heavy atoms, bond types, and ring features.

This choice also hints at the fact that there are different molecular representations suitable for different kinds of prediction jobs. Gilmer *et al.* [41] summarize the message-passing neural network framework, which is used to characterize a variety of learned topology-based representations [40]. An important feature of message-passing neural networks is the weighted transfer of atomic and bond information across the molecular graph. Even though many representations exist, their levels of complexity range widely, and no single representation has been developed to work for all types of molecular properties [42], [43]. When compared to molecules, chemical reactions are much more complex. Reactions can be identified using line-based molecular identifiers like reaction SMILES [44] and reaction InChI (RInChI) [45], while reaction mechanisms can be determined using the SMIRKS [44] system. Like molecular interactions, chemical reactions can be vectorised for incorporation into machine learning models. For the quickest and easiest results, begin with the molecular descriptors (such as fingerprints) of the reagents and add, subtract, or concatenate [46], [47]. An alternate approach is to memorize a representation of the reaction that is built around the atoms and bonds that are actually involved in the process. A neural machine translator can be used to translate the names of organic reaction products that have been stored as text (typically InChI) [45], [48].

The final step in any machine learning procedure is a modeling strategy. There are many different kinds of machine learning models. While regression and classification are two of the most common uses for models, there are other classification schemes available, such as those based on various forms of machine learning (unsupervised, supervised, active, or transfer learning) [49], [50]. As commonly understood machine learning can be thought of as any method that implicitly models correlations within datasets. Accordingly, many of the techniques we now refer to as machine learning were in fact employed for some time before the term was coined to describe them. Two such methods are principal component analysis (PCA) and the Gaussian mixture model, both of which emerged in the late 1800s [51] and early 1900s [52]. These two use-cases are now formally represented as unsupervised machine learning algorithms. Many unsupervised clustering

methods exist, including t-distributed stochastic neighbor embedding (t-SNE) [53] and density-based spatial clustering of applications with noise (DBSCAN) [54]. The goal of unsupervised learning is to train a model without giving the algorithm any "solutions" or "labels" to work with as it discovers patterns on its own. Unsupervised learning techniques have found a number of uses in chemical engineering. Palkovits R and S Palkovits [55] used the k-means algorithm [56] to classify groups of catalysts, and t-SNE was used to visualize the resulting high-dimensional representations. t-SNE has been implemented in several different fields outside of catalysis, including chemical process fault diagnosis [57] and reaction-state prediction [58]. Principal component analysis (PCA) is another dimensionality reduction algorithm commonly used by chemical engineers to find the features that best explain the data in the training set [59]. Additional applications of PCA include the detection of outliers [60]. Other algorithms for spotting anomalies include deep belief subspace analysis (DBSCAN) and long short-term memory (LSTM) [61].

Supervised classification methods like decision trees (and, by extension, random forests) can be used [62] when the dataset is labelled, meaning the correct classification of each data point is known. Alternative supervised classification strategies include support vector machines [63]. While initially developed for classifying data, support vector machines have since had their functionality expanded to also perform regression. Although in principle any supervised learning method can be incorporated into an active learning approach, it is necessary to use supervised or active learning techniques for regression problems. ANNs, in their many forms [64], [65] are the method most commonly associated with machine learning. Feed-forward ANNs are used for feature-based classification and regression, while convolutional neural networks are used in image processing, and recurrent neural networks are used in natural language processing (for anomaly detection). It is possible for a chemical engineer to come across ANNs [66], [67] support vector machines [63], or kernel ridge regression [68] used to predict the properties of the representations, as well as convolutional neural networks used to represent molecules [69]. Many applications in catalysis [70], chemical process control [71], and chemical process optimization [72] have used ANNs as a black-box modeling tool. For example, k-nearest neighbors has been implemented in applications such as chemical process monitoring [73] and catalyst clustering [74] because it is effective at classifying data when the labels are already known. Figure 2 shows three main connections between machine learning and chemical engineering.

Recent advances have provided ways to counteract some of the most significant criticisms of machine learning methods, while the methods' much strength open up a wide range of potential applications. Nearly all trained machine learning methods have exceptionally fast execution speeds, making them ideal for use cases where precision and throughput within strict system constraints are paramount. High-frequency, real-time optimization and feed-forward control of processes are two examples of such uses [75], [76]. Although detailed fundamental models are usually not fast enough to avoid computational delays, empirical models are often inaccurate for these applications. By being trained on the same fundamental model, machine learning models can rival the accuracy of empirical models while only requiring a fraction of the processing power. Here, a model is trained using high-level data to estimate the discrepancy between the observed result and the correct one [77], [78]. It has been shown that unsupervised algorithms are superior to supervised ones for spotting anomalies in real-time data, which is useful in process control applications. Better digital twins and improved control could lead to more productive chemical processes if faster, more precise predictions were combined with trustworthy industrial data [79], [80].

The same holds true for multiscale modeling approaches, where phenomena are modeled at multiple scales. This leads to an extremely intricate and tightly coupled system of equations. Machine learning's potential in such contexts is highly context-dependent. Machine learning is not recommended if the goal is to gain fundamental insights into the lower scale phenomena because of its black-box nature. With the incorporation of the lower dimensions into the strategy to create a more appropriate prediction for grander scale processes [81]–[83], machine learning has the potential to substitute the slow core frameworks for the smaller scales without harming the comprehensibility of the larger scale behaviours.

One last chance can be found in fixing machine learning's most serious shortcoming: its inability to be understood by humans. As it turns out, chemical engineering problems aren't the only ones where the problem of interpretable machine learning systems arises [83]–[85]. In the area of catalysis, researchers have tried to put numbers on just what it is that machine learning models pick up on [86]. Despite this effort, no direct interpretation of the model results is provided. The flow process that can be used to describe how and why a certain end result is reached, with a good result from the model, like the correct product from a chemical reaction predictor, should not be blindly accepted without first investigating the model's assumptions. It is helpful to get a sense of the model's confidence in its own decisions by first quantifying the individual prediction uncertainties and then moving on to the model's output [87]–[89]. Ensemble modeling is one easy method for doing so. For decades, meteorologists have relied on this method, which can be used in tandem with practically any model [90], [91]. Several algorithms have also been developed to ascertain the extent to which particular input features affect the output, or to ascertain the training points used by the model to produce

a particular output [91], [92]. Given that human interpretation of the molecular fingerprints used as input to very complex recurrent neural networks is already a formidable challenge, the task of interpreting the output of such a model becomes even more so. In risk management, the as low as reasonably practicable (ALARP) principle is widely applied. A similar "as simple as reasonably possible" principle could be proposed for machine learning models to ensure they are as interpretive as possible [83], [93].

The fields of chemical chemistry and chemical engineering have seen a rise in the popularity of machine learning techniques because they can find trends in data that human researchers miss. Contrary to mathematical model, which are based on clear physical formulas, machine learning models are not intended to tackle a particular issue (resulting from discovered patterns). In contrast, physical models can be thought of as representations of the real world. This indicates that solving classification problems does not necessitate the programming of a single explicitly defined decision function. Accordingly, solving regression problems does not necessitate deriving or parameterizing specific model equations [83], [91]. These advantages allow for efficient upscaling to large systems and datasets without requiring a large amount of computing power. Predicting quantum chemical properties using machine learning has seen a recent uptick in interest.

Calculating the characteristics of an individual atom using traditional ab initio approaches might take many up to several hours. Machine learning algorithm that has been properly trained can make precise predictions in a nanosecond. No doubt, other fast methods that can make accurate predictions have already been developed; however, in comparison to machine learning models, the application range for these methods is quite limited [94]. Machine learning's biggest shortcoming is that it can't extrapolate, but its usefulness can be greatly increased by simply analyzing more data. Using active learning [95], we can increase the scope with comparatively little additional information. This works wonderfully in situations where labels are costly to obtain (such as in quantum chemical calculations [96] or chemical experiments [97]). The amount of new data needed to increase the range can also be decreased with the help of active learning. In addition, preexisting machine learning models like ChemProp [98] and SchNet [99] can be utilized without the need for training or education. The accessibility of machine learning has increased thanks to the development of scikit-learn [100] and TensorFlow [101], as well as Keras [102] and PyTorch [103]. Because of these frameworks, deep learning model training is limited to a manageable number of lines of code. Since these libraries and frameworks already exist, scientists can focus on the implications of their work in the real world, rather than wasting time on developing abstract simulations [83]. Figure 3 shows the use of machine learning for predictive modeling in chemical engineering: opportunities, and benefits.

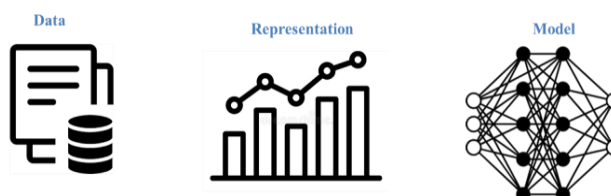


Figure 2. Three main connections between machine learning and chemical engineering, opportunities and benefits

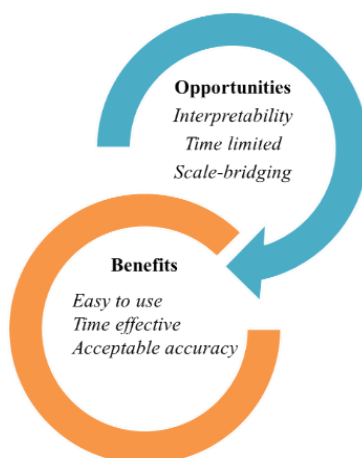


Figure 3. Use of machine learning for predictive modeling in chemical engineering: opportunities, and benefits

3. CHALLENGES AND DRAWBACKS

The scientific community benefits and faces risks from the widespread availability of machine learning models. Machine learning has the potential to benefit anyone with some programming experience, but it also leaves it vulnerable to abuse by those who don't have a thorough understanding of the underlying algorithms [91], [104]. Parameter and hyper parameter space in contemporary machine learning algorithms is vast. Even for the most experienced professionals, machine learning is largely a trial-and-error process. Some people view machine learning as a type of contemporary alchemy because researchers are frequently unwilling to actually identify how one system operates whereas other doesn't really [105]. While the fields of chemistry and chemical engineering are not as vulnerable to a reproducibility crisis as the social sciences [106], growing scepticism in the community may be a result of the increasingly irreproducible use of machine learning. After reaching the peak of exaggerated expectations [107] in Gartner's hype cycle [108], machine learning and deep learning now face the prospect of entering a period of disillusionment where interest is all but dead. Not properly understanding the results of algorithmic analyses is just as risky [83]. It can be very difficult to understand the reasoning behind an algorithm's output because of how opaque the algorithms themselves are. However, sometimes a model will get the right answer for the wrong reasons [87]. This means that researchers employing machine learning must keep in mind a fundamental statistical principle: Relationships, not causation, are what matter most. The misuse of a machine learning model occurs when it is applied to a setting for which it was not intended.

Limitations on applicability result from the information that was used during training. Researchers need to make sure they're covering all of the bases by testing over a wide enough range. A user should be aware that the model's performance will suffer if the points are outside the range [91], [109]. There are open-source programs that employ clustering algorithms for evaluating the data's credibility and its applicability in different settings [110]. The growing skill gap in machine learning (ML) threatens the widespread implementation of ML in chemical engineering study. When applying computer and data science to chemistry and chemical engineering, expertise in both the tool and the process at hand is required. Therefore, it is possible that elementary training in employing machine learning algorithms will be insufficient in the near future [83].

Instead, it will become increasingly important for undergraduates to have a firm grasp of AI and statistical methods in the field of chemical engineering. However, there needs to be more collaboration between IT professionals and other specialists. Researchers who aren't adequately prepared to use computational tools may make mistakes, and computer and data scientists who don't have enough background in the field may have to settle for subpar results. If experts in machine learning and chemistry worked together more often, this period of disillusionment could have been avoided [83], [111], [112].

It's a big problem that many forms of machine learning aren't particularly open to inspection. When a certain set of parameters is provided, all of the procedures yield the same output. A model's statistical performance on a test dataset can provide inferences about the quality of the generated output. Analyzing the model's hyperparameters (such as the ANN's node count) can provide insight into the relationships the model has learned to make, but it can be a time-consuming process. Machine learning models, despite being fast and accurate, are therefore not a good option for modeling in explanatory research [113]–[115].

This lack of interpretability adds complexity to the task of designing an efficient machine learning model. Like any other model, the best machine learning model will have some degree of optimal overfitting and under fitting. The risk of overfitting is typically much higher for machine learning models than the risk of under fitting, with both factors depending on the quality and quantity of the training data and the complexity of the model. There is no way to avoid an over fitted model when attempting to use a polynomial of very high order to fit a (noisy) linear dataset. In deep learning, overfitting typically manifests as overtraining. This means that the model will store away meaningless blips of data rather than actual patterns. Comparing the model's performance on the training data to that on the validation and test data is a sure-fire way to spot overtraining [83], [116]–[118].

In the event that training performance greatly exceeds validation performance, the model may have been overtrained. Estimating the complete number of training iterations can be difficult. Similarly to other optimization problems, machine learning models require stopping criteria to prevent overfitting [119]–[121].

In most cases, machine learning models can perform quite accurately on the training dataset; rather, the difficulty lies in performing well on data that was not used to train the model. As a result, the validation dataset, which contains data that has not been used to train the model, should be used as the stopping criterion. It is standard practice in conventional modeling methods to use a separate, independent dataset called the test dataset to rigorously test the optimized dataset [113]–[115].

One last, often crucial, flaw in machine learning techniques is the data itself. According to the "garbage in-garbage out" (GIGO) principle [122], a network will produce garbage results if the dataset contains too many systematic errors. There are some mistakes that can be easily identified as to their cause or origin, while others, once they have been made, can be extremely challenging to track down. It's possible for outliers

to appear in any statistical technique. When compared to a large dataset, outliers have a greater impact on a model trained on a small dataset. Because of this, machine learning benefits from both high-quality data and large amounts of data. Manually removing these outliers from the dataset is one approach to dealing with systematic errors; alternatively, anomaly detection algorithms like principal component analysis, t-SNE, DBSCAN, and recurrent neural networks (LSTM networks) can be used. Anomaly detection methods based on self-learning unsupervised neural networks have recently been developed [83], [91], [114], [115].

Many years of modeling, simulating, and experimenting have resulted in a massive amount of data for the chemical engineering community, but this information is typically locked away in private archives at universities and private companies. It is possible that even when data is readily available, such as from an internal database, the data is not optimal for machine learning. Text-mining methods used to extract information from scholarly articles and patents yield identical results [123]. For one thing, only positive results from experiments are typically reported, while negative results are often ignored. As a result of the engineer's superior wisdom and training, the chemical process is not subjected to absurd experimental or operating conditions. Machine learning algorithms, however, are unaware of these limitations, and excluding "trivial" data like this could have disastrous results. Figure 4 shows the use of machine learning for predictive modeling in chemical engineering: challenges and drawbacks.

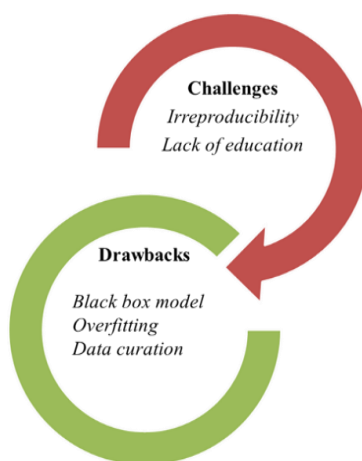


Figure 4. Use of machine learning for predictive modeling in chemical engineering: challenges and drawbacks

4. CONCLUSIONS AND OUTLOOK

In the last decade, machine learning has become an increasingly useful tool in the arsenal of a chemical engineer. There is a significant and expanding interest in machine learning among chemical engineers. The rapid processing times, adaptability, and user-friendliness of machine learning-based applications are all contributing to their growing popularity.

The flip side of machine learning's rising popularity is the chance that it will be misused, or that chemical engineers will incorrectly interpret black-box results, leading to widespread scepticism. When people don't trust one another, it can cause problems on the playing field. Each of the three presented ideas has the potential to make machine learning models more credible and transform them into a more useful and trustworthy modeling strategy.

First off, everyone in the community needs to be able to freely and unrestrictedly access the community's data and models. When researchers have access to high-quality data and open-source models, they are encouraged to use machine learning as a tool because it allows them to focus more on their research topic and less on the programming and data collection required for it. The second point, the creation of understandable models, is intrinsically linked to the first. New models for chemical applications often take their cue from preexisting algorithms, as machine learning is already well-established in other research fields.

Therefore, instead of keeping things as black boxes, the field would benefit most from studying the reasons behind why a certain output is generated from a given input. As a final piece of guidance, consider using some of your resources to acquire a deep understanding of algorithmic theory. Understanding the computer science that lies behind the graphical interface is essential for any modeler, even though chemical engineers typically have very strong mathematical and modeling skills. This is due to the fact that the graphical user interface is employed to symbolize the information being modeled. In addition, this should make it

possible to specify the model's usable domain, which is crucial for figuring out whether the model is interpolating or extrapolating. Without a doubt, the most crucial consideration is the last one. Machine learning models should be trusted, but this is only possible if their credibility is monitored for any instances in which the model is being applied to data that was not included in its training set.

REFERENCES

- [1] C. Boyadjiev, "Theoretical chemical engineering: Modeling and simulation," *Theoretical Chemical Engineering: Modeling and Simulation*, pp. 1–594, 2010, doi: 10.1007/978-3-642-10778-8.
- [2] M. Haghighatlari and J. Hachmann, "Advances of machine learning in molecular modeling and simulation," *Current Opinion in Chemical Engineering*, vol. 23, pp. 51–57, 2019, doi: 10.1016/j.coche.2019.02.009.
- [3] T. G. Dobre and J. G. Marciano, *Chemical engineering: Modeling, simulation and similitude*. John Wiley & Sons, 2007.
- [4] G. Towler and R. Sinnott, "Chemical Engineering Design: Principles, Practice and Economics of Plant and Process Design," *Chemical Engineering Design: Principles, Practice and Economics of Plant and Process Design*, pp. 1–1027, 2021, doi: 10.1016/B978-0-12-821179-3.01001-3.
- [5] D. R. Baughman and Y. A. Liu, "Neural Networks in Bioprocessing and Chemical Engineering," *Academic press*, 2014.
- [6] O. A. Oleinik and V. N. Samokhin, "Mathematical Models in Boundary Layer Theory," *Mathematical Models in Boundary Layer Theory*, 2018, doi: 10.1201/9780203749364.
- [7] T. E. Quantrille and Y. A. Liu, "Artificial Intelligence in Chemical Engineering," 2012.
- [8] L. He *et al.*, "Applications of computational chemistry, artificial intelligence, and machine learning in aquatic chemistry research," *Chemical Engineering Journal*, vol. 426, p. 131810, Dec. 2021, doi: 10.1016/j.cej.2021.131810.
- [9] L. H. Chiang, B. Braun, Z. Wang, and I. Castillo, "Towards artificial intelligence at scale in the chemical industry," *AIChE Journal*, vol. 68, no. 6, Jun. 2022, doi: 10.1002/aic.17644.
- [10] J. Jiang, B. Cui, and C. Zhang, "Basics of Distributed Machine Learning," pp. 15–55, 2022, doi: 10.1007/978-981-16-3420-8_2.
- [11] N. Artrith *et al.*, "Best practices in machine learning for chemistry," *Nature Chemistry*, vol. 13, no. 6, pp. 505–508, Jun. 2021, doi: 10.1038/s41557-021-00716-z.
- [12] A. Pomberger *et al.*, "The effect of chemical representation on active machine learning towards closed-loop optimization," *Reaction Chemistry & Engineering*, vol. 7, no. 6, pp. 1368–1379, 2022, doi: 10.1039/D2RE00008C.
- [13] G. Chen *et al.*, "Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models," 2019, [Online]. Available: <http://arxiv.org/abs/1906.09427>
- [14] J. S. Delaney, "ESOL: Estimating Aqueous Solubility Directly from Molecular Structure," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 3, pp. 1000–1005, May 2004, doi: 10.1021/ci034243x.
- [15] D. L. Mobley and J. P. Guthrie, "FreeSolv: a database of experimental and calculated hydration free energies, with input files," *Journal of Computer-Aided Molecular Design*, vol. 28, no. 7, pp. 711–720, Jul. 2014, doi: 10.1007/s10822-014-9747-x.
- [16] E. Tomelleri, L. Beletti Marchesini, A. Yaroslavtsev, S. Asgharinia, and R. Valentini, "Toward a Unified TreeTalker Data Curation Process," *Forests*, vol. 13, no. 6, p. 855, May 2022, doi: 10.3390/f13060855.
- [17] D. M. Probst *et al.*, "Evaluating Optimization Strategies for Engine Simulations Using Machine Learning Emulators," *Journal of Engineering for Gas Turbines and Power*, vol. 141, no. 9, Sep. 2019, doi: 10.1115/1.4043964.
- [18] N. Buduma, "Fundamentals of Deep Learning," *CEUR Workshop Proceedings*, vol. 1542, pp. 33–36, 2015.
- [19] N. Thuerey, P. Holl, M. Mueller, P. Schnell, F. Trost, and K. Um, "Physics-based Deep Learning," 2021, [Online]. Available: <http://arxiv.org/abs/2109.05237>
- [20] W. Q. Yan, "Computational Methods for Deep Learning," p. 134, 2021, [Online]. Available: <http://link.springer.com/10.1007/978-3-030-61081-4>
- [21] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.
- [22] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev, "InChI - the worldwide chemical structure identifier standard," *Journal of Cheminformatics*, vol. 5, no. 1, p. 7, Dec. 2013, doi: 10.1186/1758-2946-5-7.
- [23] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, vol. 1, no. 4, p. 045024, Dec. 2020, doi: 10.1088/2632-2153/aba947.
- [24] Y. Amar, A. M. Schweidtmann, P. Deutsch, L. Cao, and A. Lapkin, "Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis," *Chemical Science*, vol. 10, no. 27, pp. 6697–6706, 2019, doi: 10.1039/C9SC01844A.
- [25] K. K. Yalamanchi, M. Monge-Palacios, V. C. O. van Oudenhoven, X. Gao, and S. M. Sarathy, "Data Science Approach to Estimate Enthalpy of Formation of Cyclic Hydrocarbons," *The Journal of Physical Chemistry A*, vol. 124, no. 31, pp. 6270–6276, Aug. 2020, doi: 10.1021/acs.jpca.0c02785.
- [26] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning," *Physical Review Letters*, vol. 108, no. 5, p. 058301, Jan. 2012, doi: 10.1103/PhysRevLett.108.058301.
- [27] K. Hansen *et al.*, "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space," *The Journal of Physical Chemistry Letters*, vol. 6, no. 12, pp. 2326–2331, Jun. 2015, doi: 10.1021/acs.jpclett.5b00831.
- [28] F. A. Faber *et al.*, "Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error," *Journal of Chemical Theory and Computation*, vol. 13, no. 11, pp. 5255–5264, Nov. 2017, doi: 10.1021/acs.jctc.7b00577.
- [29] Z. Dong, H. Lin, and J. Chen, "Computational Topology and its Applications in Geometric Design," *Recent Patents on Engineering*, vol. 16, no. 5, 2021, doi: 10.2174/187221215666210901124742.
- [30] S. Liu, M. F. Demirel, and Y. Liang, "N-gram graph: Simple unsupervised representation for graphs, with applications to molecules," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8464–8476, 2019.
- [31] B. Fabian *et al.*, "Molecular representation learning with language models and domain-relevant auxiliary tasks," 2020, [Online]. Available: <http://arxiv.org/abs/2011.13230>
- [32] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of Computer-Aided Molecular Design*, vol. 30, no. 8, pp. 595–608, 2016, doi: 10.1007/s10822-016-9938-8.
- [33] A. Cherkasov *et al.*, "QSAR Modeling: Where Have You Been? Where Are You Going To?," *Journal of Medicinal Chemistry*, vol. 57, no. 12, pp. 4977–5010, Jun. 2014, doi: 10.1021/jm4004285.
- [34] A. Mayr *et al.*, "Large-scale comparison of machine learning methods for drug target prediction on ChEMBL," *Chemical Science*, vol.




- 9, no. 24, pp. 5441–5451, 2018, doi: 10.1039/C8SC00148K.
- [35] H. L. Morgan, “The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service,” *Journal of Chemical Documentation*, vol. 5, no. 2, pp. 107–113, May 1965, doi: 10.1021/c160017a018.
- [36] D. Rogers and M. Hahn, “Extended-Connectivity Fingerprints,” *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, May 2010, doi: 10.1021/ci100050t.
- [37] L. Pattanaik and C. W. Coley, “Molecular Representation: Going Long on Fingerprints,” *Chem*, vol. 6, no. 6, pp. 1204–1207, Jun. 2020, doi: 10.1016/j.chempr.2020.05.002.
- [38] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [39] R. Winter, F. Montanari, F. Noé, and D.-A. Clevert, “Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations,” *Chemical Science*, vol. 10, no. 6, pp. 1692–1701, 2019, doi: 10.1039/C8SC04175J.
- [40] K. Yang *et al.*, “Analyzing Learned Molecular Representations for Property Prediction,” *Journal of Chemical Information and Modeling*, vol. 59, no. 8, pp. 3370–3388, Aug. 2019, doi: 10.1021/acs.jcim.9b00237.
- [41] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” *34th International Conference on Machine Learning, ICML 2017*, vol. 3, pp. 2053–2070, 2017.
- [42] O. A. von Lilienfeld, “First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties,” *International Journal of Quantum Chemistry*, vol. 113, no. 12, pp. 1676–1689, Jun. 2013, doi: 10.1002/qua.24375.
- [43] L. David, A. Thakkar, R. Mercado, and O. Engkvist, “Molecular representations in AI-driven drug discovery: a review and practical guide,” *Journal of Cheminformatics*, vol. 12, no. 1, p. 56, Dec. 2020, doi: 10.1186/s13321-020-00460-5.
- [44] Z. Guo *et al.*, “Graph-based Molecular Representation Learning,” *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2023-August, pp. 6638–6646, 2023, doi: 10.24963/ijcai.2023/744.
- [45] G. Grethe, J. M. Goodman, and C. H. Allen, “International chemical identifier for reactions (RInChI),” *Journal of Cheminformatics*, vol. 5, no. 1, p. 45, Dec. 2013, doi: 10.1186/1758-2946-5-45.
- [46] M. H. S. Segler and M. P. Waller, “Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction,” *Chemistry – A European Journal*, vol. 23, no. 25, pp. 5966–5971, May 2017, doi: 10.1002/chem.201605499.
- [47] B. Sanchez-Lengeling and A. Aspuru-Guzik, “Inverse molecular design using machine learning: Generative models for matter engineering,” *Science*, vol. 361, no. 6400, pp. 360–365, Jul. 2018, doi: 10.1126/science.aat2663.
- [48] N. S. Eyke, W. H. Green, and K. F. Jensen, “Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening,” *Reaction Chemistry & Engineering*, vol. 5, no. 10, pp. 1963–1972, 2020, doi: 10.1039/D0RE00232A.
- [49] G. Mesnil *et al.*, “Unsupervised and Transfer Learning Challenge: a Deep Learning approach,” *JMLR W& CP: Proceedings of the Unsupervised and Transfer Learning challenge and workshop*, vol. 27, pp. 97–110, 2012.
- [50] S. Chakraborty, B. Uzken, K. Ayush, K. Tanmay, E. Sheehan, and S. Ermon, “Efficient Conditional Pre-training for Transfer Learning,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2022-June, pp. 4240–4249, 2022, doi: 10.1109/CVPRW56347.2022.00469.
- [51] K. Pearson, “III. Contributions to the mathematical theory of evolution,” *Proceedings of the Royal Society of London*, vol. 54, no. 326–330, pp. 329–333, Dec. 1894, doi: 10.1098/rspl.1893.0079.
- [52] K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901, [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/14786440109462720>
- [53] V. der M. L. and E. H. G., “Visualizing Data using t-SNE,” *Journal of machine learning research*, vol. 9, pp. 2579–2605, 2008.
- [54] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.
- [55] R. Palkovits and S. Palkovits, “Using Artificial Intelligence To Forecast Water Oxidation Catalysts,” *ACS Catalysis*, vol. 9, no. 9, pp. 8383–8387, Sep. 2019, doi: 10.1021/acscatal.9b01985.
- [56] G. F. Tzortzis and A. C. Likas, “The Global Kernel k\$-Means Algorithm for Clustering in Feature Space,” *IEEE Transactions on Neural Networks*, vol. 20, no. 7, pp. 1181–1194, Jul. 2009, doi: 10.1109/TNN.2009.2019722.
- [57] S. Zheng and J. Zhao, “A new unsupervised data mining method based on the stacked autoencoder for chemical process fault diagnosis,” *Computers & Chemical Engineering*, vol. 135, p. 106755, Apr. 2020, doi: 10.1016/j.compchemeng.2020.106755.
- [58] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen, “Using Machine Learning To Predict Suitable Conditions for Organic Reactions,” *ACS Central Science*, vol. 4, no. 11, pp. 1465–1476, Nov. 2018, doi: 10.1021/acscentsci.8b00357.
- [59] F. H. Vermeire and W. H. Green, “Transfer learning for solvation free energies: From quantum chemistry to experiments,” *Chemical Engineering Journal*, vol. 418, p. 129307, Aug. 2021, doi: 10.1016/j.cej.2021.129307.
- [60] M. Sun *et al.*, “Adsorption mechanism of ammonia nitrogen and phenol on lignite surface: Molecular dynamics simulations and quantum chemical calculations,” *Fuel*, vol. 337, p. 127157, Apr. 2023, doi: 10.1016/j.fuel.2022.127157.
- [61] X. Zhang, Y. Zou, S. Li, and S. Xu, “A weighted auto regressive LSTM based approach for chemical processes modeling,” *Neurocomputing*, vol. 367, pp. 64–74, Nov. 2019, doi: 10.1016/j.neucom.2019.08.006.
- [62] J. A. Richards, *Remote Sensing Digital Image Analysis*. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-030-82327-6.
- [63] M. Tanveer, T. Rajani, R. Rastogi, Y. H. Shao, and M. A. Ganaie, “Comprehensive review on twin support vector machines,” *Annals of Operations Research*, Mar. 2022, doi: 10.1007/s10479-022-04575-w.
- [64] M. Shao, Y. Lin, Q. Peng, J. Zhao, Z. Pei, and Y. Sun, “Learning graph deep autoencoder for anomaly detection in multi-attributed networks,” *Knowledge-Based Systems*, vol. 260, 2023, doi: 10.1016/j.knsys.2022.110084.
- [65] A. Thebelt, J. Wiebe, J. Kronqvist, C. Tsay, and R. Misener, “Maximizing information from chemical engineering data sets: Applications to machine learning,” *Chemical Engineering Science*, vol. 252, p. 117469, Apr. 2022, doi: 10.1016/j.ces.2022.117469.
- [66] C. A. Grambow, L. Pattanaik, and W. H. Green, “Deep Learning of Activation Energies,” *Journal of Physical Chemistry Letters*, vol. 11, no. 8, pp. 2992–2997, 2020, doi: 10.1021/acs.jpclett.0c00500.
- [67] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, “A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting,” *Annals of Data Science*, vol. 10, no. 1, pp. 183–208, 2023, doi: 10.1007/s40745-021-00344-x.
- [68] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, “FCHL revisited: Faster and more accurate quantum machine learning,” *The Journal of Chemical Physics*, vol. 152, no. 4, Jan. 2020, doi: 10.1063/1.5126701.
- [69] K. Ghosh *et al.*, “Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra,” *Advanced Science*, vol. 6, no. 9, May 2019, doi: 10.1002/advs.201801367.
- [70] H. Li, Z. Zhang, and Z. Liu, “Application of Artificial Neural Networks for Catalysis: A Review,” *Catalysts*, vol. 7, no. 10, p. 306, Oct. 2017, doi: 10.3390/catal7100306.

- [71] C. A. Grambow, Y.-P. Li, and W. H. Green, "Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach," *The Journal of Physical Chemistry A*, vol. 123, no. 27, pp. 5826–5835, Jul. 2019, doi: 10.1021/acs.jpca.9b04195.
- [72] A. M. Schweidtmann and A. Mitsos, "Deterministic Global Optimization with Artificial Neural Networks Embedded," *Journal of Optimization Theory and Applications*, vol. 180, no. 3, pp. 925–948, Mar. 2019, doi: 10.1007/s10957-018-1396-0.
- [73] S. Yan and X. Yan, "Using Labeled Autoencoder to Supervise Neural Network Combined with k-Nearest Neighbor for Visual Industrial Process Monitoring," *Industrial & Engineering Chemistry Research*, vol. 58, no. 23, pp. 9952–9958, Jun. 2019, doi: 10.1021/acs.iecr.9b01325.
- [74] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, and S. E. Denmark, "Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning," *Science*, vol. 363, no. 6424, Jan. 2019, doi: 10.1126/science.aau5631.
- [75] Z. Wu, D. Rincon, and P. D. Christofides, "Real-Time Adaptive Machine-Learning-Based Predictive Control of Nonlinear Processes," *Industrial & Engineering Chemistry Research*, vol. 59, no. 6, pp. 2275–2290, Feb. 2020, doi: 10.1021/acs.iecr.9b03055.
- [76] Z. Zhang, Z. Wu, D. Rincon, and P. Christofides, "Real-Time Optimization and Control of Nonlinear Processes Using Machine Learning," *Mathematics*, vol. 7, no. 10, p. 890, Sep. 2019, doi: 10.3390/math7100890.
- [77] A. Das, K. G. Dhal, S. Ray, and J. Gálvez, "Histogram-based fast and robust image clustering using stochastic fractal search and morphological reconstruction," *Neural Computing and Applications*, vol. 34, no. 6, pp. 4531–4554, Mar. 2022, doi: 10.1007/s00521-021-06610-6.
- [78] T. Bikmukhametov and J. Jäschke, "Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models," *Computers & Chemical Engineering*, vol. 138, p. 106834, Jul. 2020, doi: 10.1016/j.compchemeng.2020.106834.
- [79] J. H. Cavalcanti, T. Kovács, and A. Kő, "Production System Efficiency Optimization Using Sensor Data, Machine Learning-based Simulation and Genetic Algorithms," *Procedia CIRP*, vol. 107, pp. 528–533, 2022, doi: 10.1016/j.procir.2022.05.020.
- [80] E. B. Priyanka, S. Thangavel, X.-Z. Gao, and N. S. Sivakumar, "Digital twin for oil pipeline risk estimation using prognostic and machine learning techniques," *Journal of Industrial Information Integration*, vol. 26, p. 100272, Mar. 2022, doi: 10.1016/j.jii.2021.100272.
- [81] M. A. Umer, K. N. Junejo, M. T. Jilani, and A. P. Mathur, "Machine learning for intrusion detection in industrial control systems: Applications, challenges, and recommendations," *International Journal of Critical Infrastructure Protection*, vol. 38, p. 100516, Sep. 2022, doi: 10.1016/j.ijcip.2022.100516.
- [82] I. Pan, L. R. Mason, and O. K. Matar, "Data-centric Engineering: integrating simulation, machine learning and statistics. Challenges and opportunities," *Chemical Engineering Science*, vol. 249, p. 117271, Feb. 2022, doi: 10.1016/j.ces.2021.117271.
- [83] M. R. Dobbelaere, P. P. Plehiers, R. Van de Vijver, C. V. Stevens, and K. M. Van Geem, "Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats," *Engineering*, vol. 7, no. 9, pp. 1201–1211, Sep. 2021, doi: 10.1016/j.eng.2021.03.019.
- [84] D. Gunning and D. W. Aha, "DARPA's Explainable Artificial Intelligence Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, Jun. 2019, doi: 10.1609/aimag.v40i2.2850.
- [85] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and Trajectories for Explainable, Accountable and Intelligible Systems," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Apr. 2018, pp. 1–18, doi: 10.1145/3173574.3174156.
- [86] J. A. Kammeraad, J. Goetz, E. A. Walker, A. Tewari, and P. M. Zimmerman, "What Does the Machine Learn? Knowledge Representations of Chemical Reactivity," *Journal of Chemical Information and Modeling*, vol. 60, no. 3, pp. 1290–1301, Mar. 2020, doi: 10.1021/acs.jcim.9b00721.
- [87] D. P. Kovács, W. McCorkindale, and A. A. Lee, "Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias," *Nature Communications*, vol. 12, no. 1, p. 1695, Mar. 2021, doi: 10.1038/s41467-021-21895-w.
- [88] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner, "Interpretable Deep Learning in Drug Discovery," 2019, pp. 331–345, doi: 10.1007/978-3-030-28954-6_18.
- [89] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, Jan. 2019, doi: 10.1038/s42256-018-0004-1.
- [90] W. S. Parker, "Ensemble modeling, uncertainty and robust predictions," *WIREs Climate Change*, vol. 4, no. 3, pp. 213–223, May 2013, doi: 10.1002/wcc.220.
- [91] A. M. Schweidtmann *et al.*, "Machine Learning in Chemical Engineering: A Perspective," *Chemie Ingenieur Technik*, vol. 93, no. 12, pp. 2029–2039, Dec. 2021, doi: 10.1002/cite.202100083.
- [92] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.
- [93] L. Bruce and G. Popov, "Improving Risk Management," *Professional Safety*, pp. 34–36, 2021.
- [94] K. Han, A. Jamal, C. A. Grambow, Z. J. Buras, and W. H. Green, "An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation," *International Journal of Chemical Kinetics*, vol. 50, no. 4, pp. 294–303, 2018, doi: 10.1002/kin.21158.
- [95] J. Vandermause, Y. Xie, J. S. Lim, C. J. Owen, and B. Kozinsky, "Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt," *Nature Communications*, vol. 13, no. 1, p. 5183, Sep. 2022, doi: 10.1038/s41467-022-32294-0.
- [96] Y.-P. Li, K. Han, C. A. Grambow, and W. H. Green, "Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry," *The Journal of Physical Chemistry A*, vol. 123, no. 10, pp. 2142–2152, Mar. 2019, doi: 10.1021/acs.jpca.8b10789.
- [97] C. Zhang, Y. Amar, L. Cao, and A. A. Lapkin, "Solvent Selection for Mitsunobu Reaction Driven by an Active Learning Surrogate Model," *Organic Process Research & Development*, vol. 24, no. 12, pp. 2864–2873, Dec. 2020, doi: 10.1021/acs.oprd.0c00376.
- [98] E. B. Lenselink and P. F. W. Stouten, "Multitask machine learning models for predicting lipophilicity (logP) in the SAMPL7 challenge," *Journal of Computer-Aided Molecular Design*, vol. 35, no. 8, pp. 901–909, Aug. 2021, doi: 10.1007/s10822-021-00405-6.
- [99] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller, "SchNetPack: A Deep Learning Toolbox For Atomistic Systems," *Journal of Chemical Theory and Computation*, vol. 15, no. 1, pp. 448–455, Jan. 2019, doi: 10.1021/acs.jctc.8b00908.
- [100] Pedregosa Fabian *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, [Online]. Available: <http://scikit-learn.sourceforge.net>.
- [101] S. Ghemawat *et al.*, "TensorFlow: A system for large-scale machine learning," *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, p. 21, 2016, doi: 10.5555/3026877.3026899.
- [102] F. Chollet, "Keras: Deep Learning for humans," *Github*, p. Accessed 4 June 2017, 2015, [Online]. Available: <https://github.com/keras-team/keras>.




- [103] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019, [Online]. Available: <https://www.semanticscholar.org/paper/PyTorch%3A-An-Imperative-Style%2C-High-Performance-Deep-Paszke-Gross/3c8a456509e6c0805354bd40a35e3f2dbf8069b1>
- [104] M. Azeem, A. Haleem, and M. Javaid, "Symbiotic Relationship Between Machine Learning and Industry 4.0: A Review," *Journal of Industrial Integration and Management*, vol. 07, no. 03, pp. 401–433, Sep. 2022, doi: 10.1142/S2424862221300027.
- [105] S. Milivojevic, "Artificial intelligence, illegalised mobility and lucrative alchemy of border utopia," *Criminology and Criminal Justice*, p. 174889582211238, Sep. 2022, doi: 10.1177/17488958221123855.
- [106] A. Goeva, S. Stoudt, and A. Trisovic, "Toward Reproducible and Extensible Research: from Values to Action," *Harvard Data Science Review*, vol. 2, no. 4, Dec. 2020, doi: 10.1162/99608f92.1cc3d72a.
- [107] J. Fenn, M. Raskino, and B. Burton, "Understanding Gartner's Hype Cycles," *Strategic Analysis Report*, no. July 2013, p. 35, 2013.
- [108] S. Shin and J. Kang, "Structural features and Diffusion Patterns of Gartner Hype Cycle for Artificial Intelligence using Social Network analysis," *Journal of Intelligence and Information Systems*, vol. 28, no. 1, pp. 107–129, 2022.
- [109] M. Thombre, Z. Mdoe, and J. Jäschke, "Data-driven robust optimal operation of thermal energy storage in industrial clusters," *Processes*, vol. 8, no. 2, 2020, doi: 10.3390/pr8020194.
- [110] S. H. Symoens *et al.*, "QUANTIS: Data quality assessment tool by clustering analysis," *International Journal of Chemical Kinetics*, vol. 51, no. 11, pp. 872–885, Nov. 2019, doi: 10.1002/kin.21316.
- [111] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.
- [112] S. Vollmer *et al.*, "Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness," *BMJ*, p. l6927, Mar. 2020, doi: 10.1136/bmj.l6927.
- [113] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," in *2017 IEEE Custom Integrated Circuits Conference (CICC)*, Apr. 2017, pp. 1–8. doi: 10.1109/CICC.2017.7993626.
- [114] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine Learning With Big Data: Challenges and Approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017, doi: 10.1109/ACCESS.2017.2696365.
- [115] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, no. none, Jan. 2022, doi: 10.1214/21-SS133.
- [116] H. K. Jabbar and R. Z. Khan, "Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study)," pp. 163–172, 2015, doi: 10.3850/978-981-09-5247-1_017.
- [117] H. Zhang, L. Zhang, and Y. Jiang, "Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems," *2019 11th International Conference on Wireless Communications and Signal Processing, WCSP 2019*, 2019, doi: 10.1109/WCSP.2019.8927876.
- [118] J. Kolluri, V. K. Kotte, M. S. B. Phridviraj, and S. Razia, "Reducing Overfitting Problem in Machine Learning Using Novel L1/4 Regularization Method," *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, pp. 934–938, 2020, doi: 10.1109/ICOEI48184.2020.9142992.
- [119] S. Diksha and K. Neeraj, "A Review on Machine Learning Algorithms, Tasks and Applications," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 6, no. 10, pp. 1548–1552, 2017.
- [120] A. D'Amour *et al.*, "Underspecification Presents Challenges for Credibility in Modern Machine Learning," 2020, [Online]. Available: <http://arxiv.org/abs/2011.03395>
- [121] O. Gheibi, D. Weyns, and F. Quin, "Applying Machine Learning in Self-adaptive Systems," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 15, no. 3, pp. 1–37, Sep. 2020, doi: 10.1145/3469440.
- [122] O. R. P. Bininda-Emonds, K. E. Jones, S. A. Price, M. Cardillo, R. Grenyer, and A. Purvis, "Garbage in, Garbage out," in *Computational biology, Springer Nature (Netherlands)*, 2004, pp. 267–280. doi: 10.1007/978-1-4020-2330-9_13.
- [123] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, and G. A. Landrum, "Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter," *Journal of Medicinal Chemistry*, vol. 59, no. 9, pp. 4385–4402, May 2016, doi: 10.1021/acs.jmedchem.6b00153.

BIOGRAPHIES OF AUTHORS







Ashraf Al Sharah    Ashraf Al Sharah has completed his PhD from Tennessee State University, USA. He was a research associate at cyber vis research lab. And served as an Assistant Professor in the Department of Computer Engineering at Al-Ahliyya Amman University. He is serving now as an assistant professor in electrical engineering department in Al-Balqa Applied University. His research interest includes wireless security, IoT, smart attack, AI, and game theory. He can be contacted at email: aalsharah@bau.edu.jo.







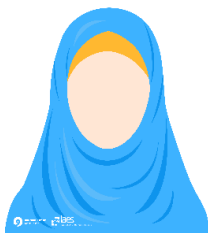
Hamza Abu Owida    Hamza Abu Owida has completed his PhD from Keele university, UK. He was a postdoctoral Research Associate: Developing xeno-free nanofibrous scaffold methodology for human pluripotent stem cell expansion, differentiation and implantation towards a therapeutic product, Keele University, Institute for Science and Technology in Medicine (ISTM), Staffordshire /UK. He is assiosiate professor in medical engineering department in Al-Ahliyya Amman University. He has published more than 30 papers in reputed journals. He can be contacted at email: h.abuowida@ammanu.edu.jo.







Feras Alnaimat     has completed his PhD from University of Birmingham, Birmingham, UK. In 2018, he joined the department of Medical Engineering, Al-Ahliyya Amman University, as an assistant professor. His current research interests include design of artificial disc implant, artificial joints and bio fluid mechanics. He is one of the steering committees of the Innovation and New Trends in Engineering, Science and Technology Education conference. He can be contacted at email: f.alnaimat@ammanu.edu.jo.







Mohammad Hassan     has completed his PhD from Baku State University, Azerbaijan. He is an Associate Professor in the Computer Engineering Department at the Faculty of Engineering at Al-Ahliyya Amman University. Hassan is a member of the Jordanian Engineering Association. He has published numerous research papers in various journals and conferences, covering topics such as machine learning, computer networks, intelligent transportation systems, and mobile learning adaptation models. He can be contacted at email: mhassan@ammanu.edu.jo.







Suhaila Abuowaida     has completed her PhD from USM University, Malaysia. received the B. Sc degrees in computer information system and M. Sc degrees in computer science from AL al-Bayt University, Jordan, in 2012 and 2015, respectively. Her research interests include Deep Learning and Computer vision. She can be contacted at email: suhila@aabu.edu.jo.



Mohammad AlHaj     has completed his PhD in System and Computer Engineering at Carleton University in 2014. He was an Assistant and Associate professor at Al-Ahliyya Amman University from 2016 to 2021. He works now as a consultant at private company in Ottawa, Canada His current research topic is focused on Requirement Engineering, Business Process Management, Artificial Intelligence and Machine. he can be contacted at email: malhaj1971@gmail.com.



Ahmad Sharadqeh     received his PhD Degree in Computer, computing system and networks from National Technical of Ukraine "Kyiv Polytechnic Institute Ukraine in 2007. Since 2009, Dr Ahmed Sharadqeh has been an Associate professor in the Computer Engineering Department, Faculty of Engineering Technology, at Al-Balqa Applied University. His research interests include the Performance of networks, Quality services, security network, IoT, image processing, digital systems design, operating system, and Microprocessors. E-mail: dr.ahmed.sharadqah@bau.edu.jo.