Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

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

Research
Smart Process Manufacturing toward Carbon Neutrality—Perspective

Active Machine Learning for Chemical Engineers: A Bright Future Lies Ahead!

Yannick Ureel, Maarten R. Dobbelaere, Yi Ouyang, Kevin De Ras, Maarten K. Sabbe, Guy B. Marin, Kevin M. Van Geem *

Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Ghent University, Ghent 9052, Belgium

ARTICLE INFO

Article history:
Received 9 September 2022
Revised 7 December 2022
Accepted 28 February 2023
Available online xxxx

Keywords:
Active machine learning
Active learning
Bayesian optimization
Chemical engineering
Design of experiments

ABSTRACT

By combining machine learning with the design of experiments, thereby achieving so-called active machine learning, more efficient and cheaper research can be conducted. Machine learning algorithms are more flexible and are better than traditional design of experiment algorithms at investigating processes spanning all length scales of chemical engineering. While active machine learning algorithms are maturing, their applications are falling behind. In this article, three types of challenges presented by active machine learning—namely, convincing the experimental researcher, the flexibility of data creation, and the robustness of active machine learning algorithms—are identified, and ways to overcome them are discussed. A bright future lies ahead for active machine learning in chemical engineering, thanks to increasing automation and more efficient algorithms that can drive novel discoveries.

1. Introduction

Experiments performed under well-defined conditions and calculations based on first principles constitute the basis of engineering research. In chemical engineering, these activities are aimed at, for example, the development and optimization of catalysts, reaction conditions, and reactor configurations. In the chemical industry, 51 billion USD was spent in 2017 on research and development [1]. This illustrates the importance of high-quality data; however, obtaining accurate data is tedious and error prone. The design of experiments (DoE) can help in extracting the maximal information with a minimum of effort [2,3], making sure that time and resources are spent efficiently. By integrating machine learning with DoE, a more flexible and efficient DoE is achieved. This so-called "active machine learning" allows a more effective selection of experimental conditions, particularly for high-dimensional and highly nonlinear phenomena [4].

Machine learning can facilitate the automation of the whole experimental cycle, from experimental selection to model building and data analysis [5]. While the most common field of application in machine learning is model building and data analysis, the focus of this article is on the potential of combining DoE with machine learning for active machine learning. Olsson [6] defined active machine learning as a supervised machine learning technique in which the learner—that is, the machine learning model—is in control of the data from which it learns. In active machine learning, machine learning algorithms are used to iteratively determine new experimental data, the so-called training data, based on uncertainty criteria. It should be noted that "experimental" can also refer to computationally expensive high-level simulations, such as high-level *ab initio* calculations of molecular properties or large eddy simulations of reactive flow with computational fluid dynamics (CFD) codes [7]. Active machine learning consists of two branches with two different purposes: active learning and Bayesian optimization. Active learning aims to explore and model a process with a minimum number of "experiments" to ensure accurate predictions over the entire design space [8]. Bayesian optimization is essentially a machine learning-based optimization strategy, where iteratively new experimental data is selected to find an experiment that optimizes the objective [9]. Either active learning or Bayesian optimization can be employed for experimental selection, depending on whether the goal is to model a process and acquire process knowledge or to optimize an objective.

1.1. Basic principles of active machine learning

Fig. 1 [10] illustrates the general workflow of active machine learning algorithms, starting with the initialization followed by an iterative loop consisting of three phases. The critical first step

* Corresponding author.
E-mail address: Kevin.VanGeem@UGent.be (K.M. Van Geem).

https://doi.org/10.1016/j.eng.2023.02.019
2095-8099/

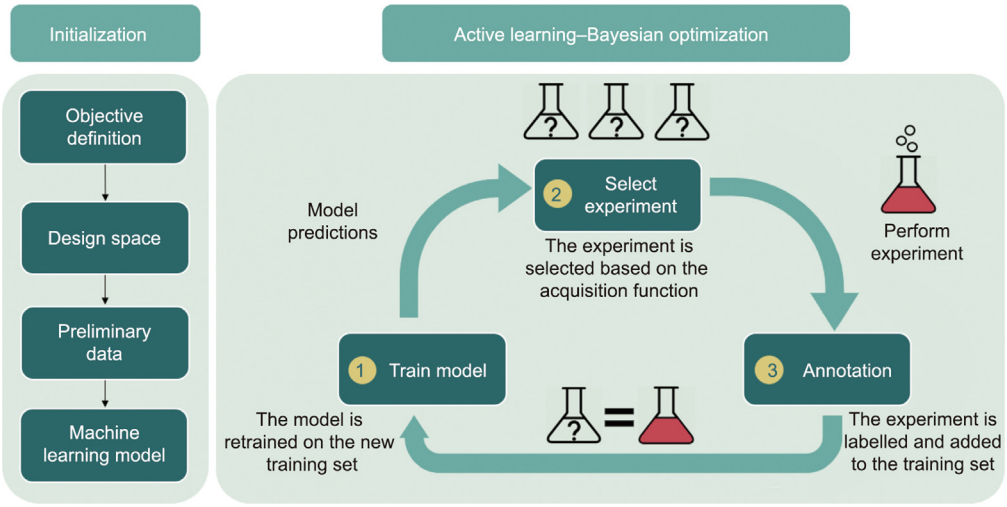

Fig. 1. Overview of the general active machine learning workflow, depicting initialization and iterative query selection. Reproduced from Ref. [10] with permission.

of initialization consists of clearly defining the research problem as either the modeling of an output (active learning) or the optimization of an objective (Bayesian optimization). An example of active learning is the investigation of the effect of reaction conditions, such as temperature and pressure, on the conversion [10,11]. With Bayesian optimization, the goal is to find the optimal reaction conditions to maximize this conversion [12–14]. In both cases, a design space is set up that defines the ranges of the studied variables by considering the objectives and the intrinsic limitations of the experimental tools. A machine learning model is then initialized and trained using a small sample of labeled data, which comes from experiments whose outcomes are known, stemming from literature, previous experiments, or newly performed experiments. In general, the amount of preliminary labeled data is very low.

After initial training, the machine learning model is able to make rudimentary predictions in the design space. The model can vaguely estimate where an optimum could be situated for Bayesian optimization, or which experiment—the so-called query—is most informative for active learning. While the definition and initialization of both active learning and Bayesian optimization are essentially the same (and are not even too different from a classic experimental campaign), the main differences and advantages are found in the model training.

Active learning is purely based on exploration, to enable predictions of the design space that are as accurate as possible. Conversely, Bayesian optimization balances both exploration and exploitation in order to find the optimum in the design space, treating every iteration as the potentially final one. Exploitation investigates areas with a high objective value to find an optimum nearby, whereas exploration discovers areas for which the predictions are unknown and therefore uncertain. Exploration requires a measure of uncertainty in the predictions to identify which areas of the design space remain unexplored [15]. Therefore, popular machine learning models for active machine learning are Gaussian processes [16–19] and Bayesian neural networks [20–22], as these allow an uncertainty estimation of their predictions. Another advantage of Gaussian processes is that they deal very well with noisy measurements, which are inherent in real-life experiments. By adding a noise term to the Gaussian process kernel, the machine learning model can estimate the experimental uncertainty and allow optimal performance of the active machine learning method [16,23]. Neural networks can also be employed for active machine learning purposes, but approximative methods such as Monte Carlo dropout or model ensembling are required to estimate the model uncertainty [11,24,25].

After initialization, the active machine learning procedure consists of three phases: the training of the machine learning model, the selection of new experiments, and the execution and annotation of these experiments (Fig. 1). The active machine learning query (phase 2) is determined through a so-called acquisition function, which is a measure of potential informativeness or optimality. The model needs the most informative subsequent data point, which is the point where the acquisition function is maximal for the selected query. The query is performed and new data is gathered (phase 3), after which the machine learning model is retrained (phase 1) and can now make improved predictions. This loop is sequentially iterated until an optimum (Bayesian optimization) is found or a sufficiently accurate model (active learning) is obtained.

To further illustrate the workflow, we present the example of a researcher examining the performance of a new catalyst for a chemical process. The researcher either aims to investigate (with active learning) or optimize (with Bayesian optimization) the effect of reaction variables (design space), such as the temperature, pressure, and reactant concentrations, on the desired product yield (objective). First, initial experiments must be performed at a number of random combinations of temperature, pressure, and reactant concentrations. Next, the researcher initiates the active machine learning loop by training the machine learning model on these randomly picked experimental data points, after which the model proposes a new experiment. When using active learning, this experiment is the most informative one; when optimizing with Bayesian optimization, this experiment is the most likely experiment to improve upon the desired product yield. The researcher performs the experiment and retrains the machine learning model, which now makes improved predictions. The experimental selection continues until the desired number of experiments is performed and an optimal machine learning model or process condition is obtained.

1.2. Active machine learning in chemical engineering

The applications of active machine learning span all the length scales of chemical engineering, from *ab initio* calculations [17,18,26] to material, molecule, and catalyst design [27–36], reaction design [12–14,37–42], and reactor design [43–45]. For example, the design of catalysts is an important asset in achieving carbon neutrality, as catalysts can enable more sustainable processes and can increase the energy efficiency of chemical processes in general [46]. However, catalyst design is still deemed an art nowadays, as it mainly relies on high-throughput screening and

limited theoretical relations, such as the Sabatier principle and linear scaling relations [47–50]. This makes catalyst design prone to human bias, as researchers tend to exploit catalyst designs that are known to work, which hampers real breakthroughs [51,52]. With active machine learning, this human bias is removed, and a substantially larger fraction of the catalyst space can be studied. Currently, the applications of active machine learning in catalysis only consider a limited design space, varying only the catalyst composition while maintaining the catalyst structure [53,54]. For example, Zhong et al. [53] performed Bayesian optimization on density functional theory (DFT) calculations to identify and synthesize promising electrocatalysts for the reduction of CO_2, whereas Nugraha et al. [54] determined the optimal composition of the most active PtPdAu catalyst to electrocatalytically oxidize methanol.

In reaction or process design, the goal of Bayesian optimization is to determine the optimal operating conditions in order to maximize the product yields, minimize the emissions per product, achieve the highest energy efficiency, and so forth. Optimization of reaction conditions has been demonstrated multiple times, including multi-objective reaction optimization with both discrete and continuous variables, likely making this the most well-developed field of active machine learning in chemical engineering [12–14]. Shields et al. [39] applied Bayesian optimization to optimize the reaction conditions for a Mitsunobu reaction and obtained an optimal yield (> 99%) for several non-intuitive reaction conditions after 40 experiments, thereby overcoming the standard reaction yield of 60%. With active learning, the goal is to acquire reaction knowledge that can be used for reactor and catalyst design, process control, or retrosynthesis. Eyke et al. [11] demonstrated the potential of active learning for DoE in reaction design by predicting reaction yields for combinations of catalysts and solvents with a minimum of available data. Recently, a DoE tool for the study of chemical reactions was developed and validated on the catalytic pyrolysis of plastic waste by Ureel et al. [10].

CFD has become an important tool for reactors, optimization, and trouble shooting. Bayesian optimization makes it possible to find an optimal reactor configuration with a minimum of computationally intensive CFD simulations. Park et al. [44] demonstrated the power of multi-objective Bayesian optimization by maximizing the gas holdup and minimizing the power consumption of a stirred tank reactor. Clearly integrating active machine learning in CFD allows for a faster and more efficient reactor design.

This survey shows that chemical engineering is a broad and diverse research field with a whole spectrum of possible active machine learning applications. Nevertheless, the use of active machine learning is not yet widespread, and there are some hurdles to overcome before it can become a trusted asset in the chemical engineer's toolkit. In this perspective article, we focus on active machine learning as a DoE technique for an experimentalist and how to popularize it. We identify three types of thresholds: convincing the experimental researcher, the flexibility of data creation, and the robustness of active machine learning algorithms (Fig. 2). In the following sections, we discuss each of these challenges and how they can be overcome.

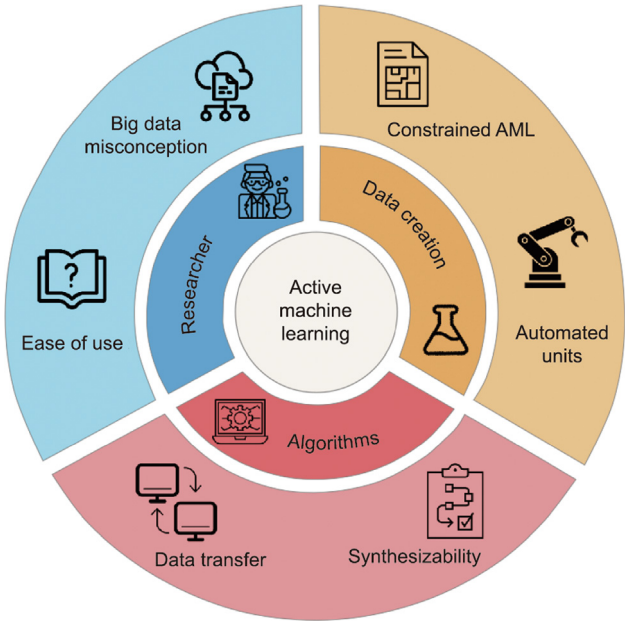

Fig. 2. Three different types of thresholds for the breakthrough of active machine learning (AML).

2. Convincing the researcher

2.1. Big data misconception

At present, a knowledge gap exists between the experimentalist community and machine learning experts [55]. This knowledge gap is the fundamental reason why active machine learning is not yet being systematically applied by experimentalists. First, there is a misconception that big data is mandatory for active machine learning and that an enormous experimental campaign is required to make it feasible. Nugraha et al. [54] reported on an optimal catalyst composition performing only 47 of a total of 5151 possible experiments, as shown in Fig. 3. In their work, Bayesian optimization was employed to determine the optimal PtPdAu catalyst composition for the electrocatalytic oxidation of methanol. Similarly, Schweidtmann et al. [12] identified their Pareto front after 68 experiments for a four-dimensional reaction optimization. Moreover, Ureel et al. [10] showed that active learning strategies are already beneficial for experimental campaigns consisting of as few as 18 experiments. These examples illustrate that both active learning and Bayesian optimization are already feasible for smaller datasets.

A second issue is related less to the experimental researcher and more to the intrinsic algorithms. Initially, all active machine learning algorithms explore the entire design space, which can result in counterintuitive or trivial queries. Consequently, the experimentalist loses confidence in the machine learning tool. The initial selection of experiments does not rely on any preliminary or physical knowledge within the machine learning models. Therefore, this issue is related both to human bias and the perception of these algorithms by their users, and to the absence of preliminary knowledge within these models. Integrating process knowledge beforehand in the machine learning model is the most powerful methodology to alleviate this problem. Such knowledge can be incorporated via two different approaches: either through the design of the machine learning model, such as a Gaussian process kernel [56], or through training on literature or simulation data [57]. The incorporation of preliminary knowledge into active machine learning models will be discussed in Section 4.1.

2.2. Ease of use

In active learning strategies, multiple factors are varied at the same time, whereas regular DoE strategies often vary a single factor at a time. This makes the post-processing of the experiments less trivial, as the effects of the factors are not isolated. As a result, a statistical analysis is required to draw conclusions from an experimental campaign using an active learning strategy [58]. These

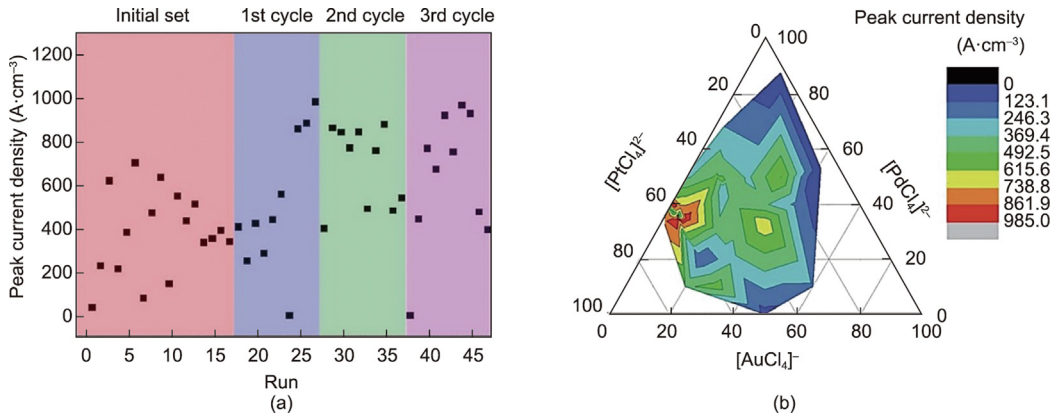

Fig. 3. (a) Nugraha et al. [54] determined the optimal PtPdAu catalyst composition for the electrocatalytic oxidation of methanol by performing only 47 experiments, with a higher peak current density denoting a better catalyst. (b) Contour plot of the effect of catalyst composition on peak current density, as determined by the 47 performed experiments. Reproduced from Ref. [54] with permission.

tools are incorporated in regular DoE software but not in the active machine learning packages that are currently available. This problem is closely related to another issue that limits the applicability of active learning—namely, its ease of use. Many different active machine learning packages exist these days, such as Gryffin [59], Phoenics [60], and BayesianOptimization [61] for Bayesian optimization, and Gaussian *N*-dimensional active learning framework (GandALF) [10] or general and efficient active learning (GEAL) [62] for active learning. However, most of the current active machine learning packages must be configured with Python, except for GandALF, which uses a csv spreadsheet. The use of these active machine learning tools requires programming skills, as they offer no graphical user interface (GUI), hampering the usage of these methodologies. Thus, at present, researchers that wish to use active machine learning must make a substantial time investment. This "activation barrier" is too high for many researchers, particularly because of the required ability to code.

3. Improving the flexibility of data creation

3.1. Constrained active machine learning

Active machine learning algorithms are often developed on simulated data, where there are no practical limitations on the data creation side [32,36,63]. However, in real life, experimental units or procedures do not allow this flexibility. For example, even a completely automated experimental unit often needs to heat up or cool down, or requires time to stabilize, which slows down the generation of a new data point when different temperatures are selected by the algorithm. In addition, experiments are often performed in parallel (e.g., in high-throughput units), as opposed to the algorithms, which assume a sequential selection of experiments. Therefore, active machine learning strategies should be constrained to the unit on which they are used, to allow for an optimal experimental efficiency that will make them applicable to real-world applications [64]. In the example above, it is often easier to heat an experimental unit than it is to cool it; therefore, an extra constraint should be added to the algorithm to make it preferable to select experiments that increase rather than decrease in temperature.

Next to constraints resulting from how the experimental equipment operates, constraints can also be important for simulations [43,45]. Let us consider a case that involves optimizing a reactor *in silico* using CFD. When defining the reactor geometry for CFD, it is not trivial that every type of geometry is feasible to simulate, nor that the geometry can be properly meshed or the results are mesh independent [65]. When these constraints are non-trivial, a separate machine learning model can be trained to learn the constraints and enforce the viability of the simulations [43].

Another example with constrained experimental units is a high-throughput experimental campaign that is used to screen different catalytic materials. Within these units, several experimental variables, such as temperature and pressure, are often fixed for every type of experiment per batch. This requires another constraint for the batch selection of these experiments, as the variables must be fixed for all selected queries. To tune active machine learning algorithms according to their application, a close collaboration between the machine learning expert and the experimentalist is thus required. In this way, the benefits of applying active machine learning are also available for less flexible experimental units.

Symbiosis between the experimentalist and the machine learning scientist will benefit both parties. First of all, it will extend the fields of application for active machine learning as researchers become more aware of the benefits of active machine learning. This close collaboration will help in identifying useful features within these active machine learning algorithms, such as blocking or automatic post-processing. More practical constraints might be added to the experimental selection, such as the time or cost required for a proposed experiment. Lastly, this collaboration between the experimentalist and the machine learning expert assists in informing experimental researchers and removing the currently existing biases against active machine learning.

3.2. Automation

In an ideal case, active machine learning is coupled with a flexible automated experimental unit or is even equipped by a robot [12,14,66]. Thus, the control and optimization of the performance of the experiments can become optimal, saving valuable time and effort. Automated experimental units are increasingly being applied in molecular synthesis and chemical engineering, although these units are not yet commonplace [67–69]. One requirement of automated robotic units is that they should be reconfigurable [70]. Moreover, they should have a broad application range and should not be limited to the investigation of a single reaction type or a narrow temperature range. Of course, the use of automated units is not self-evident, as they are often expensive and are currently not well-suited for every problem. For example, despite past efforts [71], the automated synthesis and testing of catalysts is a challenging task, especially when studying a broad design space [72]. By coupling these systems with active machine learning techniques, enormous time saving is expected for experimental campaigns,

as this will speed up reaction and catalyst optimization, as well as the acquisition of scientific knowledge. A last threshold of these automated units is the question of the safety of these units. By expanding the catalyst or reaction design space, safety concerns increase, as doing so increases the probability that undesired reactions will occur. Therefore, good chemical knowledge is still required when employing these units in order to identify and incorporate safety constraints. Here, the definition of safety constraints again requires close collaboration between experimental experts and machine learning scientists.

4. Algorithm robustness

4.1. Data transfer

When performing experiments, it is advantageous for the experiments to be widely applicable and to serve multiple purposes. The information gathered in experiments should be made available according to the FAIR guiding principles (i.e., findability, accessibility, interoperability, and reusability) and can then be of value for other researchers [73]. However, with active machine learning, a single objective is chosen, which determines the experimental selection. This hampers the applicability of the experiments, as only one experimental output is well-studied. For example, when investigating reactions, the conversion is typically selected as the output of interest; however, this limits the information on other properties, such as yields or selectivity. In the worst-case scenario, the yields are not measured and no information is gathered; contrarily even when these yields are measured, it cannot be guaranteed that all trends are considered in the example. As the goal of active machine learning is to model conversions, this method ignores the behavior of interesting reaction yields, which can result in trends remaining hidden. With Bayesian optimization, this does not pose an issue, as the goal is to optimize an objective, which makes the data per definition less generally applicable. Multi-objective Bayesian optimization techniques exist, whereas only single objective strategies are possible for active learning, meaning that all interesting outputs should be incorporated within a single active learning objective [12,40,44]. Therefore, to ensure the reusability of the gathered data, it is important that not only the modeled output but also other potential relevant outputs are measured during experiments.

After creating data that is of wide interest, it is important to be able to incorporate that knowledge into active machine learning tools. Fig. 4 summarizes the different data sources and modeling strategies that can be employed to achieve this. When an active machine learning model is pretrained on literature data, an improved initial experimental selection is achieved that resolves the issue of suboptimal initial selection that was mentioned earlier [57]. The incorporation of literature data is trivial when the experimental uncertainty is similar to that of the newly gathered data. However, when the literature data is of better or inferior quality than the gathered data, it is important for the machine learning model to be able to make a distinction between the two. Heteroscedastic machine learning models exist [63], but they do not necessarily permit the incorporation of two separate noise factors, as the variation in noise is dependent on the variable in heteroscedastic models. Conversely, multi-fidelity active machine learning strategies make it possible to employ widely abundant low-quality data for accurate pretraining of the active machine learning model [74–76]. These methods have been developed based on simulated "experimental" data only, but they are very promising for improving the performance of active machine learning tools when applied to real experimental data. Moreover, these multi-fidelity models can also be used for the incorporation of data from a mechanistic model into the machine learning model. When the uncertainty of the mechanistic model predictions is known, an appropriate distinction can be made between experimental data and modeled data, both with their respective uncertainties, in the multi-fidelity model. In this way, additional mechanistic information can be incorporated into a machine learning model, which improves the experimental selection.

Data that is closely related—but not similar in nature—can also serve as an initialization for active machine learning models [77]. For example, when modeling reactions with one type of catalyst and literature data on another catalyst are available, this data might still contain valuable information for an active learning model [78]. With active transfer learning, the goal is to leverage this knowledge from nearly similar data to obtain a machine learning model with an improved perception of the examined problem. Active transfer learning is the combination of the two main methods of active machine learning and transfer learning to make machine learning less data intensive. With transfer learning, (abundantly) available low-quality data is used to pretrain a machine learning model, which is then refined with a limited amount of high-quality data. In this way, rudimentary physical knowledge is introduced into the machine learning model, which again improves the initial experimental selection. This methodology has been proven to work on the reaction yield classification of cross-coupling reactions, by pretraining a machine learning model on reactions with different nucleophiles [78].

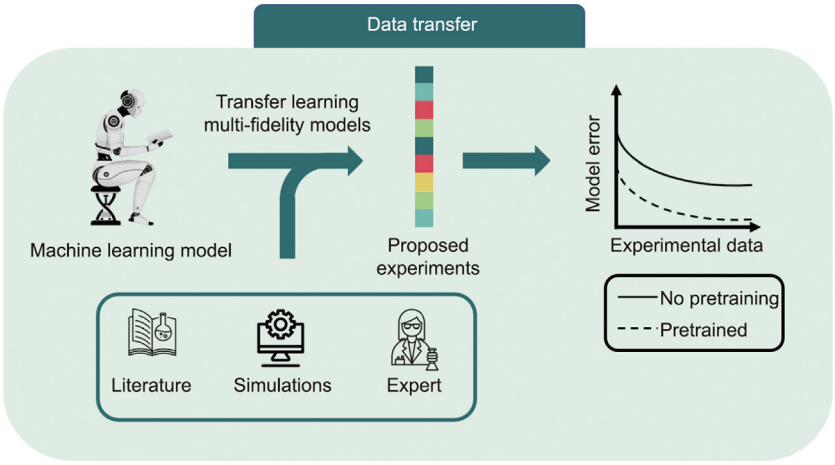

Fig. 4. The incorporation of data from the literature, simulations, or expert knowledge into machine learning models via transfer learning or multi-fidelity models improves the active machine learning performance.

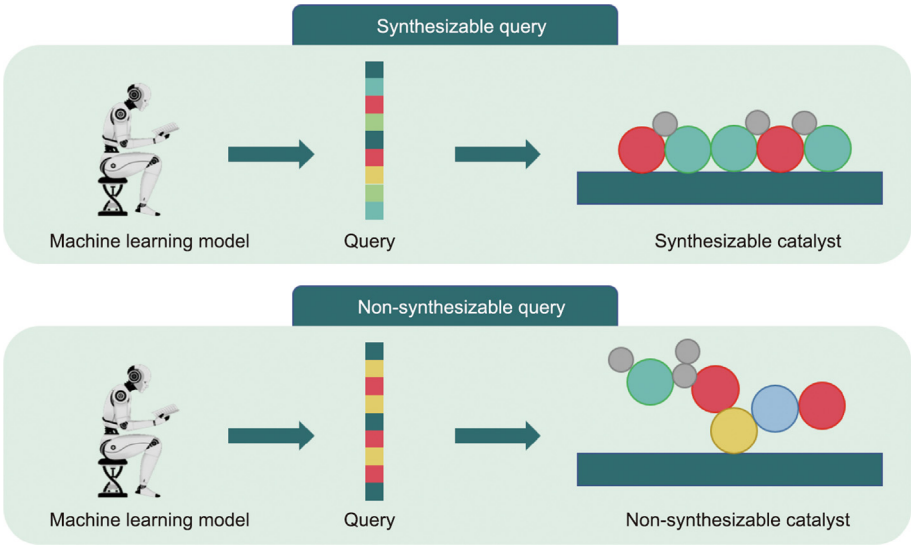

Fig. 5. An illustration of synthesizability. A machine learning model proposes a query, which is essentially a vector representation of the catalyst. This query corresponds with a catalyst, which can either be realistic and synthesizable (top) or unrealistic and non-synthesizable (bottom).

The reuse of literature data within active machine learning applications will further enhance the performance of these tools. The first active transfer learning approaches are being developed within chemical engineering, but further development of algorithms is crucial to make active transfer learning applicable within all domains of chemical engineering.

4.2. Synthesizability

Active machine learning can be used to determine the optimal query for either optimization or modeling purposes. However, for certain problems, it is not evident that these queries are executable. For example, in catalyst or molecule design, novel compounds are proposed to synthesize and test the property of interest. Here, the representation of the catalyst or molecule is crucial for the synthesizability of the queries. Synthesizability, which is defined as the feasibility of the proposed queries, refers to whether the proposed catalysts or molecules can be synthesized, as illustrated in Fig. 5. Often, a vector containing the catalyst composition is a simple representation of a catalyst [54,79]. This ensures the synthesizability of the catalyst but limits the design space explored by the active machine learning algorithm, as only the composition is varied and no structural or geometrical properties are considered. Ideally, the complete catalyst space is considered for every problem by, for example, considering the complete three-dimensional (3D) geometry as a representation of the catalyst site or molecule. However, not every imaginable catalyst's or molecule's 3D geometry can be synthesizable, so there is tradeoff between the magnitude of the design space, so-called creativity, and synthesizability.

As illustrated by the previous example, the problem of synthesizability essentially boils down to a problem of the machine learning representation upon which constraints are added to enforce synthesizability. One intuitive approach is to use the synthesis process of the catalyst or molecule as the machine learning representation. A vector containing the catalyst composition, calcination temperature and time, and presence of ion exchange or impregnation can be used to represent a catalyst. In this way, the synthesizability of the queries is ensured, as every proposed recipe is executable. However, this representation does not necessarily ensure an easy mapping to the property of interest, and an increased amount of data might be required to model this relation.

Aside from this intuitive approach, learned machine learning representations make it possible to create a continuous representation, which ensures the validity of the proposed queries [80,81]. By training recently developed methodologies such as variational auto-encoders or generative adversarial neural networks on a set of synthesizable molecules or catalysts, a learned machine learning representation—that is, a so-called latent space—can be developed, ensuring the synthesizability of the proposed queries [80,82,83]. Upon this representation, additional constraints on the catalyst or molecule can be enforced, according to the application [31].

Finding an adequate representation is always important in machine learning problems. For active machine learning, this representation is essential in order to harmonize both synthesizability and creativity.

5. Conclusions and perspectives

Active machine learning is extremely well suited for use by chemical engineering researchers to speed up experimental campaigns ranging from molecule and catalyst design to reaction and reactor design. However, active machine learning is not well-known among experimental researchers, and many active machine learning applications are not currently user friendly. Better collaboration between machine learning experts and chemical engineers can overcome these barriers. Such interactions will also help to tune active machine learning algorithms, depending on the applied (automated) experimental units and procedures, which will improve the performance of these algorithms. A key barrier here is the suboptimal initial experimental selection, which can be overcome by integrating transfer learning and active learning with the aid of multi-fidelity models. Moreover, the application domain of active machine learning can be significantly extended by adapting general active machine learning algorithms to obtain "tailor-made" algorithms, depending on the setup constraints. While the algorithms should be customized, the data should be generally usable, such that performed experiments can serve multiple purposes. By harmonizing synthesizability and creativity, active machine learning is bound to make significant advances in the fields of molecule and catalyst synthesis. Recent promising breakthroughs will allow active machine learning to become an essential tool for the chemical engineer and will further facilitate autonomous and

efficient scientific discoveries, which will contribute to a more sustainable chemical industry in the future.

Acknowledgments

Yannick Ureel, Maarten R. Dobbelaere, and Kevin De Ras respectively acknowledge financial support from the Fund for Scientific Research Flanders (FWO Flanders) through the doctoral fellowship grants (1185822N, 1S45522N, and 3F018119). The authors acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (818607).

Compliance with ethics guidelines

Yannick Ureel, Maarten R. Dobbelaere, Yi Ouyang, Kevin De Ras, Maarten K. Sabbe, Guy B. Marin, and Kevin M. Van Geem declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Oxford Economics Ltd. The global chemical industry: catalyzing growth and addressing our world's sustainability challenges. Oxford: Oxford Economics Ltd.; 2019.
[2] Lazić ŽR. Design of experiments in chemical engineering: a practical guide. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA; 2006.
[3] Franceschini G, Macchietto S. Model-based design of experiments for parameter precision: state of the art. Chem Eng Sci 2008;63(19):4846–72.
[4] Melnikov AA, Poulsen Nautrup H, Krenn M, Dunjko V, Tiersch M, Zeilinger A, et al. Active learning machine learns to create new quantum experiments. Proc Natl Acad Sci USA 2018;115(6):1221–6.
[5] Duong-Trung N, Born S, Kim JW, Schermeyer MT, Paulick K, Borisyak M, et al. When bioprocess engineering meets machine learning: a survey from the perspective of automated bioprocess development. Biochem Eng J 2023;190:108764.
[6] Olsson F. A literature survey of active machine learning in the context of natural language processing. Kista: Swedish Institute of Computer Science; 2009.
[7] Marin GB, Galvita VV, Yablonsky GS. Kinetics of chemical processes: from molecular to industrial scale. J Catal 2021;404:745–59.
[8] Settles B. Active learning. Cham: Springer Nature Switzerland AG; 2012.
[9] Frazier PI. A tutorial on Bayesian optimization. 2018. arXiv:1807.02811v1.
[10] Ureel Y, Dobbelaere MR, Akin O, Varghese RJ, Pernalete CG, Thybaut JW, et al. Active learning-based exploration of the catalytic pyrolysis of plastic waste. Fuel 2022;328:125340.
[11] Eyke NS, Green WH, Jensen KF. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. React Chem Eng 2020;5(10):1963–72.
[12] Schweidtmann AM, Clayton AD, Holmes N, Bradford E, Bourne RA, Lapkin AA. Machine learning meets continuous flow chemistry: automated optimization towards the Pareto front of multiple objectives. Chem Eng J 2018;352:277–82.
[13] Amar Y, Schweidtmann AM, Deutsch P, Cao L, Lapkin A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. Chem Sci 2019;10(27):6697–706.
[14] Clayton AD, Schweidtmann AM, Clemens G, Manson JA, Taylor CJ, Niño CG, et al. Automated self-optimisation of multi-step reaction and separation processes using machine learning. Chem Eng J 2020;384:123340.
[15] Thrun S. Exploration in active learning. In: Arbib MA, editor. The handbook of brain theory and neural networks. Cambridge: MIT Press; 1995. p. 381–4.
[16] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge: MIT Press; 2006.
[17] Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. Comput Mater Sci 2017;140:171–80.
[18] Vandermause J, Torrisi SB, Batzner S, Xie Y, Sun L, Kolpak AM, et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. NPJ Comput Mater 2020;6(1):20.
[19] Riis C, Antunes F, Hüttel FB, Azevedo CL, Pereira FC. Bayesian active learning with fully Bayesian Gaussian processes. 2022. arXiv:2205.10186.
[20] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. In: Proceedings of the 32nd International Conference on Machine Learning; 2015 Jul 7–9; Lille, France; 2015. p. 1613–22.
[21] Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia; 2017. p. 1183–92.
[22] Hafner D, Tran D, Lillicrap T, Irpan A, Davidson J. Noise contrastive priors for functional uncertainty. In: Proceedings of the 35th Uncertainty in Artificial Intelligence Conference; 2019 Jul 22–25; Tel Aviv, Israel; 2020. p. 905–14.
[23] McHutchon A, Rasmussen C. Gaussian process training with input noise. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, editors. Proceedings of the 24th International Conference on Neural Information Processing Systems; 2011 Dec 12–14; Granada, Spain; 2011. p. 1341–9.
[24] Zhang Y, Lee AA. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. Chem Sci 2019;10(35):8154–63.
[25] Núñez M, Vlachos DG. Multiscale modeling combined with active learning for microstructure optimization of bifunctional catalysts. Ind Eng Chem Res 2019;58(15):6146–54.
[26] Sivaraman G, Krishnamoorthy AN, Baur M, Holm C, Stan M, Csányi G, et al. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. NPJ Comput Mater 2020;6(1):104.
[27] Reker D, Schneider P, Schneider G, Brown JB. Active learning for computational chemogenomics. Future Med Chem 2017;9(4):381–402.
[28] Brown KA, Brittman S, Maccaferri N, Jariwala D, Celano U. Machine learning in nanoscience: big data at small scales. Nano Lett 2020;20(1):2–10.
[29] Hansen MH, Torres JAG, Jennings PC, Wang Z, Boes JR, Mamun OG, et al. An atomistic machine learning package for surface science and catalysis. 2019. arXiv:1904.00904.
[30] Griffiths RR, Hernández-Lobato JM. Constrained Bayesian optimization for automatic chemical design. 2017. arXiv:1709.05501.
[31] Griffiths RR, Hernández-Lobato JM. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. Chem Sci 2020;11(2):577–86.
[32] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nat Catal 2018;1(9):696–703.
[33] Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun 2020;11(1):5966.
[34] Oftelie LB, Rajak P, Kalia RK, Nakano A, Sha F, Sun J, et al. Active learning for accelerated design of layered materials. NPJ Comput Mater 2018;4(1):74.
[35] Kitchin JR. Machine learning in catalysis. Nat Catal 2018;1(4):230–2.
[36] Jablonka KM, Jothiappan GM, Wang S, Smit B, Yoo B. Bias free multiobjective active learning for materials design and discovery. Nat Commun 2021;12(1):2312.
[37] Zhang C, Amar Y, Cao L, Lapkin AA. Solvent selection for Mitsunobu reaction driven by an active learning surrogate model. Org Process Res Dev 2020;24(12):2864–73.
[38] Clayton AD, Manson JA, Taylor CJ, Chamberlain TW, Taylor BA, Clemens G, et al. Algorithms for the self-optimisation of chemical reactions. React Chem Eng 2019;4:1545–54.
[39] Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JIM, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature 2021;590(7844):89–96.
[40] Felton KC, Rittig JG, Lapkin AA. Summit: benchmarking machine learning methods for reaction optimisation. Chem–Methods 2021;1(2):116–22.
[41] Felton K, Wigh D, Lapkin A. Multi-task Bayesian optimization of chemical reactions. 2020. ChemRxiv: 13250216.v1.
[42] Dogu O, Eschenbacher A, Varghese RJ, Dobbelaere M, D'Hooge DR, Van Steenberge PHM, et al. Bayesian tuned kinetic Monte Carlo modeling of polystyrene pyrolysis: unraveling the pathways to its monomer, dimers, and trimers formation. Chem Eng J 2023;455:140708.
[43] Tran A, Sun J, Furlan JM, Pagalthivarthi KV, Visintainer RJ, Wang Y. pBO-2GP-3B: a batch parallel known/unknown constrained Bayesian optimization with feasibility classification and its applications in computational fluid dynamics. Comput Methods Appl Mech Eng 2019;347:827–52.
[44] Park S, Na J, Kim M, Lee JM. Multi-objective Bayesian optimization of chemical reactor design using computational fluid dynamics. Comput Chem Eng 2018;119:25–37.
[45] Morita Y, Rezaeiravesh S, Tabatabaei N, Vinuesa R, Fukagata K, Schlatter P. Applying Bayesian optimization with Gaussian process regression to computational fluid dynamics problems. J Comput Phys 2022;449:110788.
[46] Friend CM, Xu B. Heterogeneous catalysis: a central science for a sustainable future. Acc Chem Res 2017;50(3):517–21.
[47] Sabatier P. La catalyse en chimie organique. Paris: Hachette Livre; 1920. French.
[48] Ichikawa S. Harmonious optimum conditions for heterogeneous catalytic reactions derived analytically with Polanyi relation and Bronsted relation. J Catal 2021;404:706–15.
[49] Landau RN, Korré SC, Neurock M, Klein MT, Quann RJ. Hydrocracking phenanthrene and 1-methyl naphthalene: development of linear free energy relationships. In: Oballa M, editor. Catalytic hydroprocessing of petroleum and distillates. Boca Raton: CRC Press; 2020. p. 421–32.
[50] Vijay S, Kastlunger G, Chan K, Nørskov JK. Limits to scaling relations between adsorption energies? J Chem Phys 2022;156(23):231102.
[51] Hong X, Chan K, Tsai C, Nørskov JK. How doped MoS_2 breaks transition-metal scaling relations for CO_2 electrochemical reduction. ACS Catal 2016;6(7):4428–37.
[52] Pérez-Ramírez J, López N. Strategies to break linear scaling relationships. Nat Catal 2019;2(11):971–6.
[53] Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh CT, et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. Nature 2020;581(7807):178–83.

[54] Nugraha AS, Lambard G, Na J, Hossain MSA, Asahi T, Chaikittisilp W, et al. Mesoporous trimetallic PtPdAu alloy films toward enhanced electrocatalytic activity in methanol oxidation: unexpected chemical compositions discovered by Bayesian optimization. J Mater Chem A 2020;8(27):13532–40.
[55] Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering 2021;7(9):1201–11.
[56] Duvenaud DK. Automatic model construction with Gaussian processes [dissertation]. Cambridge: University of Cambridge; 2014.
[57] Wang Z, Dahl GE, Swersky K, Lee C, Mariet Z, Nado Z, et al. Pre-training helps Bayesian optimization too. 2022. arXiv:220703084.
[58] Symoens SH, Aravindakshan SU, Vermeire FH, De Ras K, Djokic MR, Marin GB, et al. QUANTIS: data quality assessment tool by clustering analysis. Int J Chem Kinet 2019;51(11):872–85.
[59] Häse F, Aldeghi M, Hickman RJ, Roch LM, Aspuru-Guzik A. Gryffin: an algorithm for Bayesian optimization of categorical variables informed by expert knowledge. Appl Phys Rev 2021;8(3):031406.
[60] Häse F, Roch LM, Kreisbeck C, Aspuru-Guzik A. Phoenics: a Bayesian optimizer for chemistry. ACS Cent Sci 2018;4(9):1134–45.
[61] Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, editors. Proceedings of the 25th International Conference on Neural Information Processing Systems; 2012 Dec 3–6; Lake Tahoe, NV, USA. Red Hook: Curran Associates Inc.; 2012. p. 2951–9.
[62] Xie Y, Tomizuka M, Zhan W. Towards general and efficient active learning. 2021. arXiv:211207963.
[63] Griffiths RR, Aldrick AA, Garcia-Ortegon M, Lalchand V, Lee AA. Achieving robustness to aleatoric uncertainty with heteroscedastic Bayesian optimisation. Mach Learn Sci Technol 2021;3(1):015004.
[64] Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A. Bayesian optimization with known experimental and design constraints for chemistry applications. Digit Discov 2022;1:732–44.
[65] Habashi WG, Dompierre J, Bourgault Y, Ait-Ali-Yahia D, Fortin M, Vallet MG. Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD. Part I: general principles. Int J Numer Meth Fluids 2000;32(6):725–44.
[66] Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature 2020;583(7815):237–41.
[67] Hoffer L, Voitovich YV, Raux B, Carrasco K, Muller C, Fedorov AY, et al. Integrated strategy for lead optimization based on fragment growing: the diversity-oriented-target-focused-synthesis approach. J Med Chem 2018;61(13):5719–32.
[68] Bédard AC, Adamo A, Aroh KC, Russell MG, Bedermann AA, Torosian J, et al. Reconfigurable system for automated optimization of diverse chemical reactions. Science 2018;361(6408):1220–5.
[69] Mateos C, Nieves-Remacha MJ, Rincón JA. Automated platforms for reaction self-optimization in flow. React Chem Eng 2019;4(9):1536–44.
[70] Eyke NS, Koscher BA, Jensen KF. Toward machine learning-enhanced high-throughput experimentation. Trends Chem 2021;3(2):120–32.
[71] Hahndorf I, Buyevskaya O, Langpape M, Grubert G, Kolf S, Guillon E, et al. Experimental equipment for high-throughput synthesis and testing of catalytic materials. Chem Eng J 2002;89(1–3):119–25.
[72] Oh KH, Lee HK, Kang SW, Yang JI, Nam G, Lim T, et al. Automated synthesis and data accumulation for fast production of high-performance Ni nanocatalysts. J Ind Eng Chem 2022;106:449–59.
[73] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3(1):160018.
[74] Greenman KP, Green WH, Gómez-Bombarelli R. Multi-fidelity prediction of molecular optical peaks with deep learning. Chem Sci 2022;13(4):1152–62.
[75] Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci 2017;129:156–63.
[76] Folch JP, Lee RM, Shafei B, Walz D, Tsay C, van der Wilk M, et al. Combining multi-fidelity modelling and asynchronous batch Bayesian optimization. Comput Chem Eng 2023;172:108194.
[77] Mao S, Wang B, Tang Y, Qian F. Opportunities and challenges of artificial intelligence for green manufacturing in the process industry. Engineering 2019;5(6):995–1002.
[78] Shim E, Kammeraad JA, Xu Z, Tewari A, Cernak T, Zimmerman PM. Predicting reaction conditions from limited data through active transfer learning. Chem Sci 2022;13(22):6655–68.
[79] Kim M, Ha MY, Jung WB, Yoon J, Shin E, Kim ID, et al. Searching for an optimal multi-metallic alloy catalyst by active learning combined with experiments. Adv Mater 2022;34(19):2108900.
[80] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018;4(2):268–76.
[81] Shang C, You F. Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. Engineering 2019;5(6):1010–6.
[82] Sanchez-Lengeling B, Outeiral C, Guimaraes GL, Aspuru-Guzik A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). 2017. ChemRxiv: 5309668.v3.
[83] Jensen Z, Kwon S, Schwalbe-Koda D, Paris C, Gómez-Bombarelli R, Román-Leshkov Y, et al. Discovering relationships between OSDAs and zeolites through data mining and generative neural networks. ACS Cent Sci 2021;7(5):858–67.