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, Technologiepark 125, 9052 Gent, Belgium

* Corresponding author: Kevin.VanGeem@UGent.be, Technologiepark 125, 9052 Gent, Belgium;

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

Abstract:

By combining machine learning with design of experiments, so-called active machine learning, more efficient and cheaper research can be conducted. Machine learning algorithms are more flexible, and are better at investigating the processes spanning all length scales of chemical engineering. While the active machine learning algorithms are maturing, its applications are lacking behind. Three types of challenges faced by active machine learning are identified and ways to overcome them are discussed: the convincing of the experimental researcher, the flexibility of data creation, and the robustness of the active machine learning algorithms. A bright future lies ahead for active machine learning in chemical engineering thanks to increasing automation and more efficient algorithms to drive novel discoveries.

1. Introduction

Performing experiments at well-defined conditions and first-principles based calculations constitute the basis of engineering research. In chemical engineering, these activities are aimed at e.g. 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. Design of experiments (DoE) can help by extracting maximal information with a minimum of effort [2, 3], making sure that the 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, in particular for high-dimensional and highly-nonlinear phenomena, a more effective selection of experimental conditions [4].

In this "Perspective", we discuss the potential of combining DoE and machine learning, i.e. active machine learning. Olsson defines active machine learning as a supervised machine learning technique in which the learner, being the machine learning model, is in control of the data from which it learns [5]. With active machine learning, machine learning algorithms are used to iteratively determine new experimental data, so-called training data, based on uncertainty criteria. Note that "experimental" can also refer to computationally expensive high-level simulations e.g. high level ab initio calculations of molecular properties or large eddy simulations of reactive flow with Computational Fluid Dynamics (CFD) codes [6]. 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 of "experiments" to ensure accurate predictions over the entire design space [7]. Bayesian

optimization is essentially a machine learning-based optimization strategy, where iteratively new experimental data is selected to find the experiment which optimizes the objective [8].

1.1. The Basic Principles of Active Machine Learning

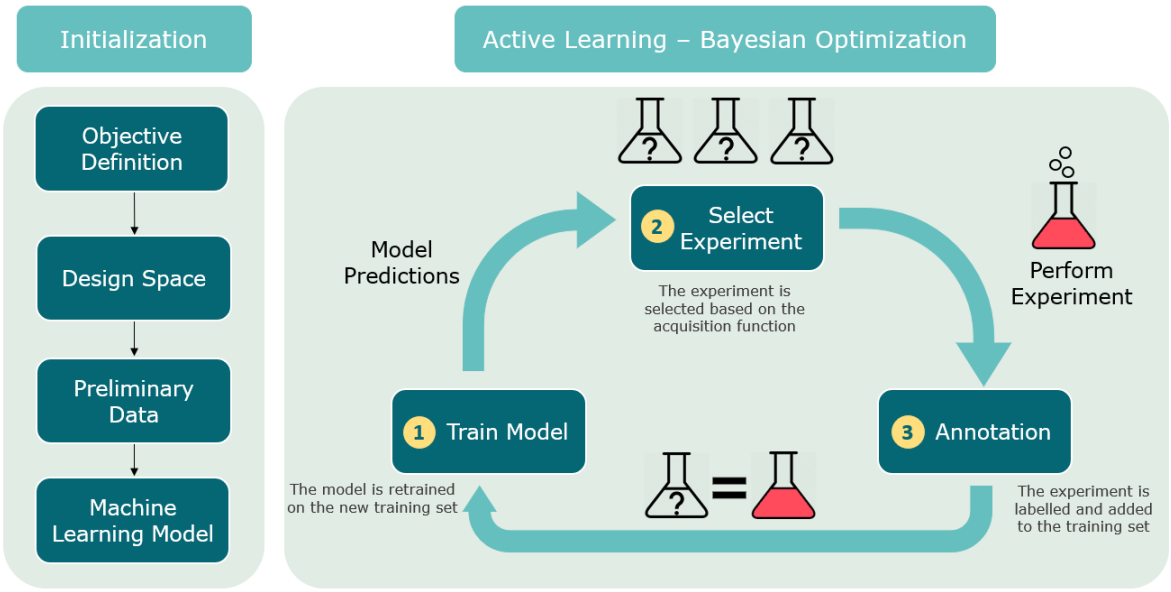

Figure 1. Overview of the general active machine learning workflow, depicting the initialization and the iterative query selection (based on [9]).

Figure 1 illustrates the general workflow of active machine learning algorithms with first the initialization followed by an iterative loop consisting of three phases. The critical first step 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, whereas with Bayesian optimization the goal is to find the optimal reaction conditions to maximize this conversion. In both cases a design space is set up which 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, being experiments of which the outcome is known, stemming from literature, previous experiments, or newly performed experiments. Generally, 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 is essentially the same, and 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 as accurate as possible predictions of the design space. Oppositely, Bayesian optimization balances both exploration and exploitation for finding the optimum in the design space and treats every iteration as the potentially final one. Exploitation investigates areas with a high objective value to find an optimum nearby whereas exploration discovers areas on which the predictions are unknown and therefore uncertain. Exploration requires a measure of uncertainty on the predictions to identify which areas of the design space remain unexplored [10]. Therefore, popular machine learning models for active machine learning are Gaussian processes [11-14] and Bayesian neural networks [15-17] as these allow an uncertainty estimation of their predictions. 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 [18-20].

After initialization, the active machine learning procedure consists of three phases being the training of the machine learning model, the selection of new experiments, and the execution and annotation of these experiments (Figure 1). The active machine learning query (phase 2) is

86 determined by a so-called acquisition function which is a measure of potential informativeness
87 or optimality. The model needs the most informative next data point, and thus the point where
88 the acquisition function is maximal for the selected query. The query is performed and new data
89 is gathered (phase 3) after which the machine learning model is retrained (phase 1) and now
90 can make improved predictions. This loop is sequentially iterated until an optimum (Bayesian
91 optimization) is found or a sufficiently accurate model (active learning) is obtained.

92 1.2. Active Machine Learning in Chemical Engineering

93 The applications of active machine learning span all length scales of chemical engineering from
94 ab initio calculations [12, 13, 21], material, molecule and catalyst design [22-31], reaction
95 design [32-39] up to reactor design [40-42]. For example, the design of catalysts is an important
96 asset in achieving carbon neutrality as catalysts can enable more sustainable processes, and
97 increase the energy efficiency of chemical processes in general. However, nowadays their
98 design is still deemed an art, mainly relying on high-throughput screening and limited
99 theoretical relations such as the Sabatier principle and linear scaling relations [43-46]. This
100 makes catalyst design prone to human bias as researchers tend to exploit catalyst designs which
101 are known to work but this hampers real breakthroughs [47, 48]. With active machine learning,
102 this human bias is removed and a substantially larger fraction of the catalyst space can be
103 studied. Currently, the applications of active machine learning in catalysis only consider a
104 limited design space varying only the catalyst composition while maintaining the catalyst
105 structure. E.g., Zhong et al. performed Bayesian optimization on DFT calculations to identify
106 and synthesize promising electrocatalysts for the reduction of CO_2 [49], whereas Nugraha et al.
107 determined the optimal composition of the most active Pt/Pd/Au-catalyst to electrocatalytically
108 oxidize methanol [50].

In reaction or process design, the goal with Bayesian optimization is to determine the optimal operating conditions to maximize the product yields, minimize the emissions per product, achieve the highest energy efficiency, etc. Optimization of reaction conditions has been demonstrated multiple times, from multi-objective reaction optimization with both discrete and continuous variables which makes it probably the most well-developed field of active machine learning in chemical engineering [33-35]. Shields et al. applied Bayesian optimization to optimize the reaction conditions for a Mitsunobu reaction, and found an optimal yield (>99%) for several non-intuitive reaction conditions after 40 experiments, beating the standard reaction yield of 60% [37]. With active learning the goal is to acquire reaction knowledge which can be used for reactor and catalyst design, process control or retrosynthesis. Eyke et al. 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 [20]. Recently, a DoE-tool for the study of chemical reactions has been developed and validated on the catalytic pyrolysis of plastic waste by Ureel et al [9].

Computational fluid dynamics (CFD) has become an important tool for reactor, optimization and trouble shooting. Bayesian optimization allows to find an optimal reactor configuration with a minimum of computationally intensive CFD simulations. Park et al. demonstrated the power of multi-objective Bayesian optimization by maximizing the gas-holdup and minimizing the power consumption of a stirred tank reactor [41]. 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 its use is not yet widespread and there are some hurdles to overcome before it is a trusted asset in the chemical engineer’s toolkit.

133 In this “Perspective” we focus on active machine learning as a DoE technique for an
134 experimentalist and how to popularize it. We identify three types of thresholds: the convincing
135 of the experimental researcher, the flexibility of data creation, and the robustness of the active
136 machine learning algorithms (Figure 2). In the following sections we will discuss each of these
137 challenges and how they can be overcome.

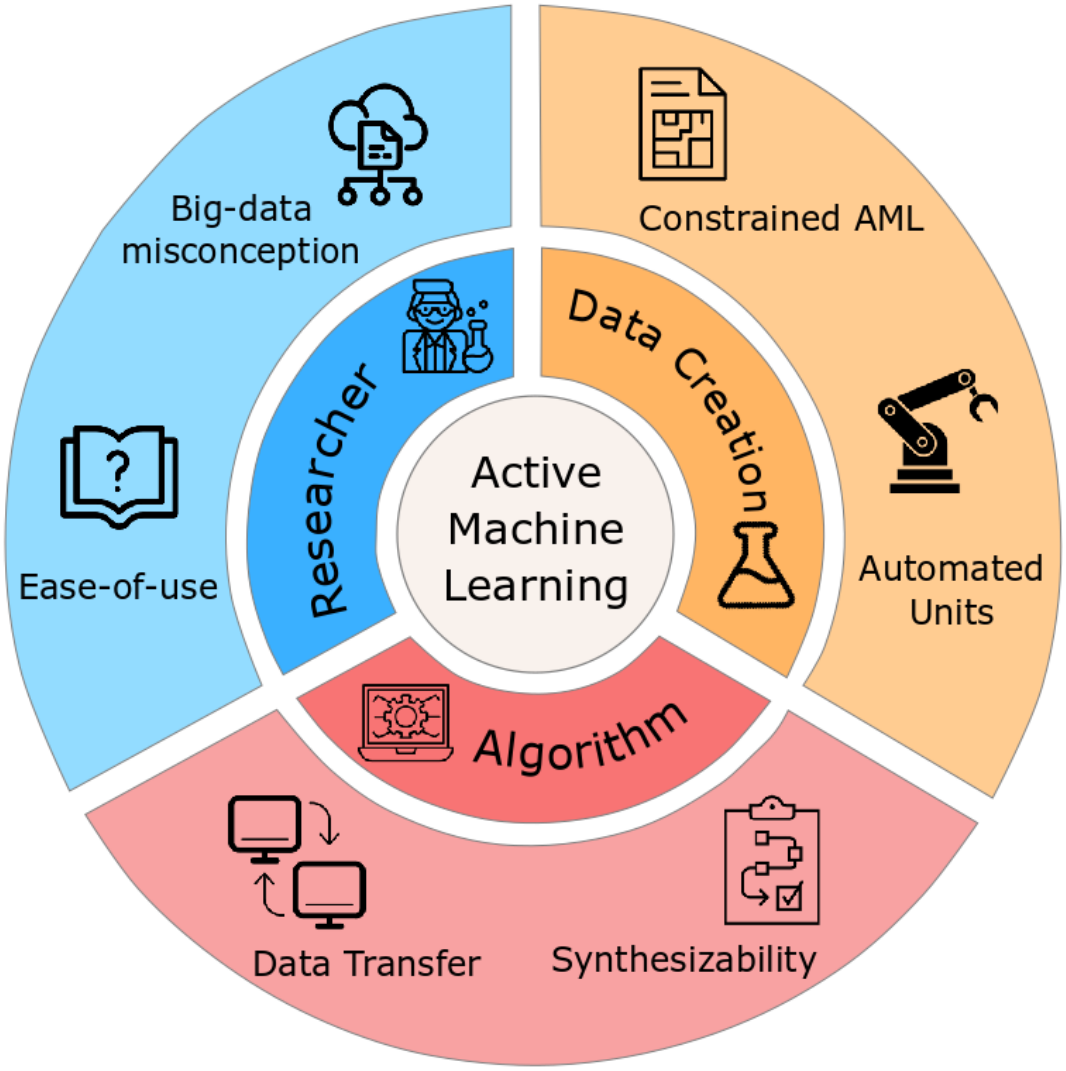

Figure 2. Three different types of thresholds for the breakthrough of active machine learning (AML).

138

2. Convincing the Researcher

2.1. Big-data Misconception

Currently, there exists a knowledge gap between the experimentalist community and the machine learning experts [51]. This is at the origin of active machine learning not yet being systematically applied by the former. First, there is a misconception that for active machine learning “big data” is mandatory and an enormous experimental campaign is required to make it feasible. Nugraha et al. found an optimal catalyst composition performing only 47 from a total of 5151 possible experiments [50]. Similarly, Schweidtmann et al. identified their pareto-front after 68 experiments for a four dimensional reaction optimization [33]. Moreover, Ureel et al. showed that active learning strategies are already beneficial for experimental campaigns consisting of as little as 18 experiments [9]. These examples illustrate that both active learning and Bayesian optimization are already feasible for smaller datasets.

A second issue is less related to the experimental researcher but more to the intrinsic algorithms. Initially, all active machine learning algorithms explore the entire design space which can result in counter-intuitive 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 to both the human bias and the perception of these algorithms by their users, and to the absence of preliminary knowledge within these models. The incorporation of preliminary knowledge into active machine learning models will be discussed in section 4.1.

2.2. Ease-of-use

With active learning strategies, multiple factors are varied at a 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 effect of the factors is not isolated. As a result, a statistical analysis is required to draw conclusions from the experimental campaign [52]. These tools are incorporated in regular DoE software but not in the active machine learning packages that are available as of today. This is closely related to another issue that limits the applicability, namely its ease-of-use. The current active machine learning packages require programming skills to be used and have no graphical user interface (GUI). The absence of a GUI hampers the usage of these methodologies as programming expertise is required before they can be applied. There is at present a substantial time investment of the researchers needed to use Active Machine Learning. This "activation barrier" is for many too high, in particular because of the required ability to code.

3. Flexibility and Inflexibility of Data Creation

3.1. Constrained Active Machine Learning

The development of active machine learning algorithms is often done on simulated data where there are no practical limitations on the data creation side [27, 31, 53]. 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 time to stabilize, which slows down the generation of a new datapoint when different temperatures are selected by the algorithm. Additionally, 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 to make them applicable to “Real World applications” [54]. In the example above it is often easier to heat up an experimental unit than to cool it, therefore an extra constraint should be added to the algorithm which prefers to select experiments which increase in temperature rather than decrease in temperature.

Next to constraints resulting from how the experimental equipment operates, these are also important for simulations [40, 42]. Consider the case when optimizing a reactor in silico with CFD. When defining the reactor geometry for CFD it is not trivial that every type of geometry is feasible to simulate nor that it can be properly meshed or that the results are mesh independent. 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 [40].

Another example of constrained experimental units are high-throughput experimental campaigns which are for example 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 on the batch selection of these experiments as variables need to be fixed for all selected queries. To tune the active machine learning algorithms according to their application, a close collaboration between the machine learning expert and experimentalist is thus required. In this way, the benefits of applying active machine learning are also available to less flexible experimental units.

This symbiosis between experimentalist and 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 experimentalist and machine learning expert helps in informing experimental researchers and remove the currently existing biases on active machine learning.

3.2. Automation

In an ideal case, active machine learning is coupled with a flexible automated experimental unit or are even equipped by a robot [33, 35, 55]. In this way control and optimization of the performance of the experiments can become optimal, and thus saving valuable time and effort. Automated experimental units are being increasingly applied for molecular synthesis and chemical engineering but these units are not yet commonplace [56-58]. One requirement of automated robotic units is that they should be reconfigurable [59]. They moreover should have a broad application range and not be limited to the investigation of a single reaction type or narrow temperature range. The use of automated units is of course not self-evident as these often are expensive and currently not well-suited for every problem. For example, despite past efforts [60] the automated synthesis and testing of catalysts is a challenging task, definitely when studying a broad design space [61]. By coupling these systems with active machine learning techniques, a huge time saving is expected for experimental campaigns, which will speed up reaction and catalyst optimization, and the acquisition of scientific knowledge. A last threshold of these automated units is of course the question of safety of these units. By expanding the catalyst or reaction design space, safety concerns rise as this increases the probability of undesired reactions to occur. Therefore, a good chemical knowledge is still required when employing these units to identify and incorporate the safety constraints. The definition of safety constraints again requires a close collaboration between experimental experts and machine learning scientists.

4. Robustness of Algorithms

4.1. Data Transfer

When performing experiments, it is advantageous when these experiments are widely applicable and serve multiple purposes. The information gathered in experiments should be made available according to the FAIR-principles and can then be of value for other researchers [62]. However, with active machine learning one 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 output of interest but this limits the information on other properties such as yields or selectivities. In the worst case, the yields are not measured and no information is gathered, on the other hand when these yields would be measured no guarantee is given that all trends are considered in this example. As the goal of the active machine learning was to model conversions, it ignores the behavior of interesting reaction yields which can result in trends to remain hidden. With Bayesian optimization this does not pose an issue as the goal here is to optimize an objective, which makes the data per definition less generally applicable. Multi-objective Bayesian optimization techniques exist while for active learning only single objective strategies are possible, meaning that all interesting outputs should be incorporated within the single active learning objective [33, 38, 41]. Therefore, to ensure the reusability of the gathered data, it is important that during experiments not only the modeled output is measured but also the potential other relevant outputs.

After creating data that is of wide interest, it is also important to be able to incorporate that knowledge in active machine learning tools. When pretraining an active machine learning model on literature data, an improved initial experimental selection is achieved which resolves

the issue of the earlier mentioned suboptimal initial selection [63]. The incorporation of literature data is trivial when the experimental uncertainty is similar to the newly gathered data. However, when the literature data is of better or inferior quality than the gathered data, it is important that the machine learning model can make a distinction between both. Heteroscedastic machine learning models exist [53], but these do not necessarily allow 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 allow to employ widely abundant low-quality data for an accurate pretraining of the active machine learning model [64, 65]. These methods have been developed based on simulated "experimental" data only, but are very promising for improving the performance of active machine learning tools when applied to real experimental data.

Data that is closely related, but not similar in nature, can also serve as initialization of active machine learning models [66]. For example, when modeling reactions with one type of catalyst and literature data on another catalyst is available, this might still contain valuable information for an active learning model [67]. 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. In this way, rudimentary physical knowledge is introduced in the machine learning model which again improves the initial experimental selection. This methodology has been proven to work on reaction yield classification of cross-coupling reactions, by pretraining a machine learning model on reactions with different nucleophiles [67].

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 a further development on algorithms is crucial for making it applicable within all domains of chemical engineering.

4.2. Synthesizability

Active machine learning determines the optimal query for the either optimization or modeling purposes. However, for certain problems it is not evident that these queries are executable. For instance in catalyst or molecule design, novel compounds are proposed to synthesize and test on the property of interest. Here, the representation of the catalyst or molecule is crucial for the synthesizability of the queries. Synthesizability is defined as the feasibility of the proposed queries, referring to whether the proposed catalysts or molecules can be synthesized. Often, a simple representation of a catalyst is a vector containing the catalyst composition [50, 68]. This guarantees the synthesizability of the catalyst but limits the design space explored by the active machine learning algorithm as only the composition is varied but no structural or geometrical properties are considered. Ideally, one considers the complete catalyst space for every problem by for example considering the complete 3D-geometry as a representation for the catalyst site or molecule. However, not every imaginable catalyst or molecule 3D-geometry can be synthesizable, which makes that there is trade-off between the magnitude of the design space, so-called creativity, and synthesizability.

As illustrated by the problem of synthesizability this essentially boils down to a problem of representation on which constraints are added. 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, presence of ion exchange or impregnation, can be used to represent a catalyst. In this way, the synthesizability of the queries is guaranteed, as every proposed recipe is executable. However, this representation does not

necessarily ensure an easy mapping to the property of interest, which might require an increased amount of data to model this relation.

Next to this intuitive approach, learned machine learning representations allow to create a continuous representation which ensures the validity of the proposed queries [69, 70]. 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, a so-called latent space, can be developed which guarantees the synthesizability of the proposed queries [69, 71, 72]. Upon this representation additional constraints on the catalyst or molecule can be enforced according to the application [26].

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

5. Conclusions and Perspectives

Active machine learning is excellently suited for chemical engineering researchers to speed up experimental campaigns ranging from molecule and catalyst design, up to reaction and reactor design. However, among experimental researchers active machine learning is less known and many active machine learning applications are not user-friendly today. A better collaboration between machine learning experts and chemical engineers can overcome these barriers. This interaction also helps to tune active machine learning algorithms depending on the applied (automated) experimental units and procedures, which improves the performance of these algorithms. To fully profit from the creativity of active machine learning, improvements on the machine learning methods related to data transfer and synthesizability are still required. By harmonizing synthesizability and creativity, active machine learning is bound to make

significant advances in the fields of molecule and catalyst synthesis. The recent promising breakthroughs will allow active machine learning to become an essential tool for the chemical engineer and further facilitate autonomous and efficient scientific discoveries which will contribute to a more sustainable chemical industry in the future.

Acknowledgements:

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

6. References

[1] The global chemical industry: Catalyzing growth and addressing our world's sustainability challenges. Oxford Economics; 2019:29.
[2] Lazic ZR. Design of experiments in chemical engineering: a practical guide. First ed. Weinheim: Wiley-VCH Verlag; 2006.
[3] Franceschini G, Macchietto S. Model-based design of experiments for parameter precision: State of the art. Chemical Engineering Science 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. Proceedings of the National Academy of Sciences 2018;115(6):1221-6.
[5] Olsson F. A literature survey of active machine learning in the context of natural language processing. 2009.
[6] Marin GB, Galvita VV, Yablonsky GS. Kinetics of chemical processes: From molecular to industrial scale. Journal of Catalysis 2021;404:745-59.
[7] Settles B. Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 2012;18:1-111.
[8] Frazier PI. A Tutorial on Bayesian Optimization. 2018(Section 5):1-22.
[9] 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.
[10] Thrun S. Exploration in active learning. Handbook of Brain Science and Neural Networks 1995:381-4.
[11] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. the MIT Press, Massachusetts Institute of Technology; 2006.
[12] Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. Computational Materials Science 2017;140:171-80.
[13] 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 Computational Materials 2020;6(1):1-11.
[14] Riis C, Antunes FN, Hüttel FB, Azevedo CL, Pereira FC. Bayesian Active Learning with Fully Bayesian Gaussian Processes. arXiv preprint arXiv:220510186 2022.

[15] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight Uncertainty in Neural Networks. 32nd International Conference on Machine Learning, ICML 2015 2015;2:1613-22.
[16] Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. 34th International Conference on Machine Learning, ICML 2017 2017;3:1923-32.
[17] Hafner D, Tran D, Lillicrap T, Irpan A, Davidson J. Noise contrastive priors for functional uncertainty. PMLR:905-14.
[18] Zhang Y, Lee AA. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. Chemical Science 2019;10(35):8154-63.
[19] Núñez M, Vlachos DG. Multiscale Modeling Combined with Active Learning for Microstructure Optimization of Bifunctional Catalysts. Industrial & Engineering Chemistry Research 2019;58(15):6146-54.
[20] Eyke NS, Green WH, Jensen KF. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. Reaction Chemistry & Engineering 2020;5(10):1963-72.
[21] 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 Computational Materials 2020;6(1):104.
[22] Reker D, Schneider P, Schneider G, Brown JB. Active learning for computational chemogenomics. Future Medicinal Chemistry 2017;9(4):381-402.
[23] Brown KA, Brittman S, Jariwala D, Celano U. Machine Learning in Nanoscience: Big Data at Small Scales. Nano Lett 2020;20(2):7-.
[24] Hansen MH, Antonio J, Torres G, Jennings PC, Wang Z, Boes JR, et al. An Atomistic Machine Learning Package for Surface Science and Catalysis. 2019.
[25] Griffiths R-R, Hernández-Lobato JM. Constrained Bayesian Optimization for Automatic Chemical Design. 2017.
[26] Griffiths R-R, Hernández-Lobato JM. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. Chemical science 2020;11(2):577-86.
[27] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. Nature Catalysis 2018;1(9):696-703.
[28] 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. Nature Communications 2020;11(1):5966.
[29] Bassman L, Rajak P, Kalia RK, Nakano A, Sha F, Sun J, et al. ARTICLE Active learning for accelerated design of layered materials.
[30] Kitchin JR. Machine learning in catalysis. Nature Catalysis 2018;1(4):230-2.
[31] Jablonka KM, Melpatti Jothiappan G, Wang S, Smit B, Yoo B. Bias Free Multiobjective Active Learning for Materials Design and Discovery. Nature communications 2021;12(1):1-10.
[32] Zhang C, Amar Y, Cao L, Lapkin AA. Solvent Selection for Mitsunobu Reaction Driven by an Active Learning Surrogate Model. Organic Process Research & Development 2020;24(12):2864-73.
[33] 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. Chemical Engineering Journal 2018;352:277-82.
[34] Amar Y, Schweidtmann AM, Deutsch P, Cao L, Lapkin AA. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. Chemical Science 2019;10(27):6697-706.
[35] 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. Chemical Engineering Journal 2020;384:123340-.
[36] Clayton AD, Manson JA, Taylor CJ, Chamberlain TW, Taylor BA, Clemens G, et al. Algorithms for the self-optimisation of chemical reactions. *Reaction Chemistry and Engineering*. 4. Royal Society of Chemistry; 2019:1545-54.
[37] 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.
[38] Felton KC, Rittig JG, Lapkin AA. Summit: Benchmarking Machine Learning Methods for Reaction Optimisation. ChemRxiv; 2020.
[39] Felton K, Wigh D, Lapkin AA. Multi-task Bayesian Optimization of Chemical Reactions. 2020.
[40] 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. Computer Methods in Applied Mechanics and Engineering 2019;347:827-52.
[41] Park S, Na J, Kim M, Lee JM. Multi-objective Bayesian optimization of chemical reactor design using computational fluid dynamics. Computers & Chemical Engineering 2018;119:25-37.
[42] Morita Y, Rezaeiravesh S, Tabatabaei N, Vinuesa R, Fukagata K, Schlatter P. Applying Bayesian optimization with Gaussian process regression to computational fluid dynamics problems. Journal of Computational Physics 2022;449:110788.
[43] Sabatier P. La catalyse en chimie organique. 1920.
[44] Ichikawa S. Harmonious optimum conditions for heterogeneous catalytic reactions derived analytically with Polanyi relation and Bronsted relation. Journal of Catalysis 2021;404:706-15.
[45] Landau R, Korre S, Neurock M, Klein M, Quann R. Hydrocracking phenanthrene and 1-methyl naphthalene: Development of linear free energy relationships. Catalytic hydroprocessing of petroleum and distillates. CRC Press; 2020, p. 421-32.
[46] Vijay S, Kastlunger G, Chan K, Nørskov JK. Limits to scaling relations between adsorption energies? The Journal of Chemical Physics 2022;156(23):231102.
[47] Hong X, Chan K, Tsai C, Nørskov JK. How doped MoS2 breaks transition-metal scaling relations for CO2 electrochemical reduction. Acs Catalysis 2016;6(7):4428-37.
[48] Pérez-Ramírez J, López N. Strategies to break linear scaling relationships. Nature Catalysis 2019;2(11):971-6.
[49] Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh C-T, et al. Accelerated discovery of CO2 electrocatalysts using active machine learning. Nature 2020;581(7807):178-83.
[50] 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. Journal of Materials Chemistry A 2020;8(27):13532-40.

[51] 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.
[52] Symoens SH, Aravindakshan SU, Vermeire FH, De Ras K, Djokic MR, Marin GB, et al. QUANTIS: data quality assessment tool by clustering analysis. International Journal of Chemical Kinetics 2019;51(11):872-85.
[53] Griffiths R-R, Aldrick AA, Garcia-Ortegon M, Lalchand V. Achieving robustness to aleatoric uncertainty with heteroscedastic Bayesian optimisation. Machine Learning: Science and Technology 2021;3(1):015004.
[54] Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A. Bayesian optimization with known experimental and design constraints for chemistry applications. arXiv preprint arXiv:220317241 2022.
[55] Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature 2020;583(7815):237-41.
[56] 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. Journal of Medicinal Chemistry 2018;61(13):5719-32.
[57] Bédard A-C, 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.
[58] Mateos C, Nieves-Remacha MJ, Rincón JA. Automated platforms for reaction self-optimization in flow. Reaction Chemistry & Engineering 2019;4(9):1536-44.
[59] Eyke NS, Koscher BA, Jensen KF. Toward machine learning-enhanced high-throughput experimentation. Trends in Chemistry 2021;3(2):120-32.
[60] 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. Chemical Engineering Journal 2002;89(1-3):119-25.
[61] Oh KH, Lee H-K, Kang SW, Yang J-I, Nam G, Lim T, et al. Automated synthesis and data accumulation for fast production of high-performance Ni nanocatalysts. Journal of Industrial and Engineering Chemistry 2022;106:449-59.
[62] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 2016;3(1):1-9.
[63] Wang Z, Dahl GE, Swersky K, Lee C, Mariet Z, Nado Z, et al. Pre-training helps Bayesian optimization too. arXiv preprint arXiv:220703084 2022.
[64] Greenman KP, Green WH, Gómez-Bombarelli R. Multi-fidelity prediction of molecular optical peaks with deep learning. Chemical science 2022;13(4):1152-62.
[65] Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Computational Materials Science 2017;129:156-63.
[66] 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.
[67] Shim E, Kammeraad JA, Xu Z, Tewari A, Cernak T, Zimmerman PM. Predicting reaction conditions from limited data through active transfer learning. Chemical Science 2022;13(22):6655-68.

[68] Kim M, Ha MY, Jung W-B, Yoon J, Shin E, Kim I-d, et al. Searching for an Optimal Multi-Metallic Alloy Catalyst by Active Learning Combined with Experiments. Advanced Materials 2022;34(19):2108900.
[69] 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 Central Science 2018;4(2):268-76.
[70] 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.
[71] 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.
[72] 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 Central Science 2021;7(5):858-67.