Catalysis Today xxx (xxxx) xxx

Contents lists available at ScienceDirect

Catalysis Today

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

Machine learning in experimental materials chemistry

Balaranjan Selvaratnam, Ranjit T. Koodali *

Department of Chemistry, University of South Dakota, Vermillion, SD, 5769, USA

ARTICLE INFO

Keywords:
Machine learning
Catalysis
Materials Chemsitry

ABSTRACT

The development of advanced materials is an important aspect of modern life. However, the discovery of novel materials involves searching the vast chemical space to find materials with desired properties. Recent developments in the applications of Machine Learning (ML) in materials chemistry show promise to accelerate the material discovery process. In this perspective article, we highlight the importance of ML in materials chemistry. We discuss few examples of ML applications in synthesis, characterization, and predicting activities of materials. Finally, we discuss the challenges in this field and how the progress in ML in chemistry is leveraged together with advanced robotics to perform automated optimization of material discovery.

1. Introduction

The rapid development of information technology has led to the creation of extremely large datasets across various disciplines such as business, healthcare, science, and so forth. These datasets are too big to be processed by humans in a reasonable timeframe in order to make informed decisions and predictions. Hence, various algorithms and tools are being developed to identify patterns and to create models. These tools and algorithms belong to a subset of Artificial Intelligence (AI) called Machine Learning (ML) and is defined as the study of algorithms and models that enables computer systems to perform a task by learning from the data rather than explicitly programmed. ML tools excel at analyzing multi-dimensional data and enable one to create surrogate models that describe the large datasets. ML technology has led to several breakthrough innovations involving high dimensional data such as self-driving vehicles, image classification, speech recognition, natural language processing, *etc.* [1].

The input data to the ML model is known as descriptor/feature and the target property is known as label/ground truth. ML can be classified into supervised, unsupervised, and reinforcement learning. In supervised learning, both data and their labels will be given, and the model will learn the mapping between the input data and the expected output. In unsupervised learning, only the input will be provided, and the model will learn the similarities and differences between the input data and then group the data accordingly. Finally, in reinforcement learning, an agent (typically a model) will take actions based on the current state and the environment will respond with reward or penalty for the action. Over time, the agent will learn how to maximize reward and achieve the tasked goal in its environment.

Traditionally, research in chemistry uses the Edisonian approach where discovery is progressed through domain knowledge and trial-and-error experimentations and/or calculations [2]. For instance, more than 10,000 experiments led to the development of an industrially significant iron-based catalyst for ammonia production by Haber and Bosch [3]. This iterative process is often both expensive and laborious because of the large chemical space and experimental parameter space, which makes it difficult to find better (catalyst) materials and optimum experimental conditions [4]. Considering that the number of possible drug-like molecules is larger than 10^{60} and that of materials is greater than 10^{100}, the discovery problems in chemistry are similar to a needle in a haystack that is larger than the number of atoms in the observable universe [5,6]. This makes it impossible to use brute-force explorations for anything but smaller design spaces. Hence, it is necessary to use data-driven ML methods to traverse the chemical space in search of novel materials and molecules.

The applications of these data-driven tools in materials science are known as the fourth paradigm [7,8]. In these applications, the input descriptors are be made of known physicochemical parameters or encodings of positions and identities of atoms in the system. Using various descriptors and models, several applications have been developed to predict various physicochemical properties that are expensive to obtain using experiments or *ab initio* calculations. These models assist the discovery process by accelerating several tasks in research including the search of molecules and materials with desired properties [9], development of machine learning potentials [10,11], rapid prediction of various electronic and physicochemical properties [12–14],

* Corresponding author.
E-mail addresses: ktranjit@gmail.com, ranjit.koodali@wku.edu (R.T. Koodali).

https://doi.org/10.1016/j.cattod.2020.07.074
Received 16 April 2020; Received in revised form 13 July 2020; Accepted 30 July 2020
Available online 21 August 2020

optimization of reaction conditions [15], understanding hidden insights [16], synthesis planning [17], *etc.*

An introduction to this topic is provided by Rothenberg [18]. A step-by-step tutorial on ML in chemistry introducing the ML workflow including pre-processing and hyperparamter optimization using Python programming language is reported by Palkovis [19]. The applications of ML for materials discovery and optimization can be found in several review articles [6,20,21]. Applications of ML in catalysis research is summarized by Toyao et al. [22]. For more in-depth reviews on this emerging field, we direct the readers to recent reviews [4,5,21–26] and perspectives [27].

The discovery of novel materials is an important requirement in modern industrialized world. Various materials such as polymers, porous materials, two-dimensional materials, semiconductors, have been developed for applications in catalysis, electronics, automotive industry, information technology, *etc*. Several recent studies show that ML can be used to accelerate the identification of new of materials by aiding in the efficient search of the materials space by building surrogate predictive models [28–30]. In this perspective, we look into these applications of ML in three important areas of materials development: synthesis, characterization, and catalytic applications. In the first section, the applications of ML in materials synthesis is discussed.

2. Machine learning in materials synthesis

The synthesis of new materials often involves an intensive trial-and-error exploratory approach to optimize the composition and experimental parameters. Moreover, in most cases, only successful attempts are reported in the literature, rendering some of the initial explorations to be performed again during the synthesis of new material. The exploration attempts (consisting of successful and failed reactions) contain information about the reaction parameter surface. This data can be interpreted to identify regions of the parameter space that result in the formation of desired product. Such an analysis is feasible for reactions with few parameters, however, with more parameters, it can be exhausting. For these problems, ML provides several models to analyze high-dimensional data and they can be utilized to identify the product forming region in the parameter space.

Raccuglia et al. used a ML approach to identify crystallization of templated vanadium selenites using experimental parameters and features of the reagents as input data [31]. For this model, they used synthesis parameters and their results of 3,955 unique examples composed of failed experimental attempts for the crystallization of templated vanadium selenites collected from laboratory notebooks. The descriptors for each entry include various physicochemical properties of reactants, atomic properties, and experimental conditions. The models were trained to classify the outcome for the given entry as one of the four allowed outcomes: no solid product, amorphous solid product, polycrystalline product, and single crystals with an average dimension of 0.01 mm. After testing Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), logistic regression, and k-Nearest Neighbors (KNN), the SVM model was found to provide the highest accuracy of 78 %. Then, this model was used to predict outcomes for synthesis with new organic building units that were not present in the original dataset. The performance of the model was compared with decisions made by experienced chemists in the field. The results indicate that the ML model predicts crystallization of new vanadium selenites with a success rate of 89 %. Furthermore, the model was interpreted by building a decision tree model to reveal chemical insights. This decision tree analysis leads to a rationale hypothesis for further experimentations.

The ML based approach can also be used to understand the chemical space of materials that are synthesized using kinetically controlled pathways where the precise control of the experimental parameters is often the key to obtain the desired products. Muraoka et al. used ML driven navigation of synthesis parameter space to understand the physicochemical, structural, heuristic insights, and the relationship between structure and synthesis parameters of zeolites. In the synthesis of zeolites, kinetic control by means of reactant ratio, amount of water, base, temperature, heating time, choice of Organic Structure Directing Agent (OSDA) are required to get the desired zeolite [32]. In order to understand the relationship between the synthesis parameters and the structure, they collected experimental data for 686 OSDA free syntheses of aluminosilicate zeolites. Using 80 % of the data, various ML models were trained and upon testing, extreme gradient boosting (XGBoost) and RF models performed better with accuracies of 75–80 %. Then the XGBoost model was further analyzed by building a decision tree. The decision tree was useful in elucidating the chemical knowledge behind zeolite synthesis (Fig. 1). For example, the model could correctly identify the boundary at which the Si/Al ratio of 1 corresponding to the Löwenstein's rule could be obtained.

In addition to the presented examples discussed, several works in the literature show the utility of applying machine learning based methods in the synthesis of metal organic frameworks [33], polymers [34], oxides [35], *etc*.

The examples discussed in this section shows that ML can assist in the synthesis of various materials using datasets of relatively low to modest in size. Further, they also show that using experimental and/or literature data, ML models can be trained to predict trends that are valuable to researchers. In the next section, recent examples of the use of ML to augment materials characterization is discussed.

3. Machine learning in materials characterization

Characterization of materials using diffraction, electron microscopy, spectroscopy, and other techniques are performed to understand the physicochemical properties of materials. Molecular ordering and crystal structures of materials are studied using diffraction and imaging methods and ML methods can be used for rapid interpretation of these data. For instance, localized molecular ordering is an important factor that determines the electronic properties of semicrystalline polymer poly(3-hexylthiopene) (P3HT) used for organic electronic applications. To identify the crystalline domains, Masters et al. used Non-negative Matrix Factorization (NMF) of Secondary Electron (SE) spectra [36]. The SE spectra decomposed by NMF resulted with two components with component one resembling the features of amorphous regions and component two resembling crystalline features. By analyzing the composition of these components at each pixel in a SE hyperspectral imaging map, they were able to map the crystalline regions on 60 nm^2 area (of P3HT) in about two minutes.

Powder X-ray diffraction (PXRD) is used to study the phase and composition of materials, packing in organic crystals, and ordered pores in porous materials. Once the diffractogram is obtained, a trained user can perform the analysis using available appropriate software, relatively in short time for most samples. Performing this analysis in a high-throughput setting, however, can be a bottleneck. Sohn and coworkers used a Convolutional Neural Network (CNN) to identify phases and compositions of materials by training on simulated XRD patterns [37]. The training data was acquired by simulating the XRD of structures containing Sr, Al, Li, and O. The CNN model was trained to identify phases as well as composition. Then, they tested the models against test sets from simulated and real data, and for simulated test data, the accuracies were greater than 99.5 %, and for test sets on real data, the classification accuracies were greater than 97.3 %. Then, the CNN model was used to predict the percentage of each phase by transforming the regression problem into a classification problem. This was achieved by predicting classes corresponding to bins of percentages of phase composition as target properties for the CNN model and this gives an approximate estimate of the fraction of the components of the phases in an unknown mixture. Various CNN configurations were analyzed and among them CNN-6 F (6 refers to CNN's with deep architecture consisting on 6 layers) performed with highest accuracy of 98.13 % for the simulated test data and 82.67 % for test data set from real samples.

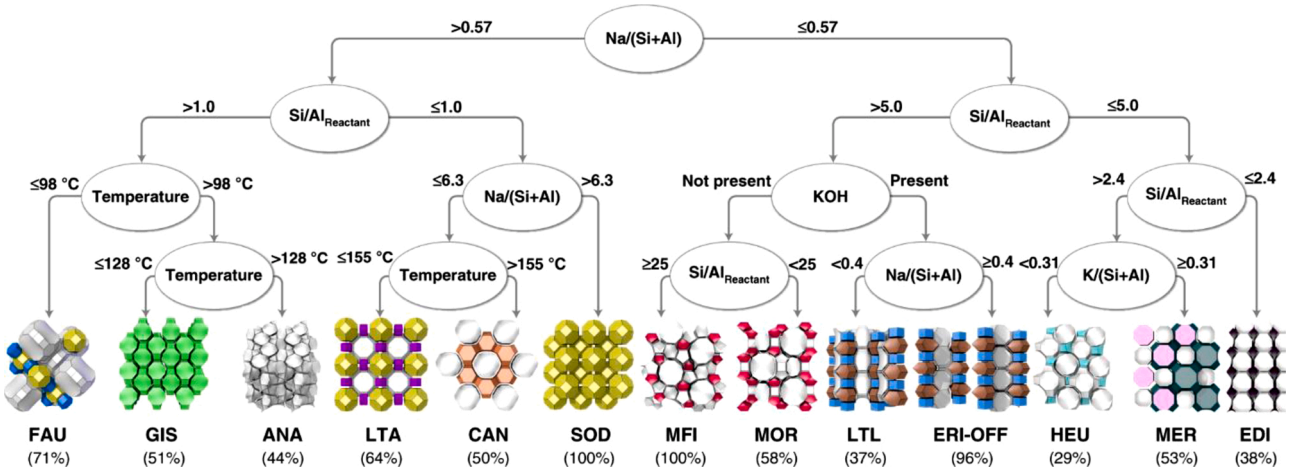

Fig. 1. Decision tree constructed from the trained model with the highest accuracy of the XGBoost and this model was interpreted as a decision tree shown here with a depth of 4. Reproduced with permission from Springer Nature.

For some catalytic applications, fine structure of the active species needs to be characterized. Frenkel and coworkers used Small Angle X-ray Scattering (SAXS) and X-ray Absorption Near-Edge Structure (XANES) and Neural Network (NN) models for structure elucidation of Cu nanoclusters and nanoparticles under catalytic conditions[38]. Their NN model was trained using simulated XANES spectra of copper nanoparticles with various sizes and shapes at various sites. The average coordination numbers of first few shells and nearest neighbor distance were used as targets for the NN model. The validation of the model using test data demonstrated that the model can accurately predict the desired target properties. Then, Cu_4, Cu_{12}, and Cu_{20} clusters were synthesized and loaded on ZrO_2 and ZnO substrates and heated under CO_2 and H_2 in He to study the changes in copper nanoparticles during catalytic CO_2 conversion reaction. *In situ* XANES and SAXS spectra were obtained during the reaction and the coordination number of the Cu nanoclusters were predicted by the NN model using the XANES data (Fig. 2). The comparison of the cluster sizes estimated using NN model and XANES and SAXS techniques provided complementary structure information and a better understanding of the formation of assembly of Cu nanoclusters and their agglomeration to form nanoparticles. This ML and SAXS analysis is helpful in understanding the catalysis of metal supported systems. The recent developments in this topic is summarized by Timoshenko and Frenkel [39].

Rational design of materials using predictive rules can be useful in iterating materials toward desired properties. However, in some chemical spaces such as intermetallics, simple rule-based predictions traditionally yield less fruitful results. Hence for a given composition of metals, it is difficult to predict which structure will form as the product. To this end, Mar and coworkers have demonstrated that various ML methods can be used to predict the crystal structure of intermetallics using readily available descriptors [30]. They used a SVM model to predict crystal structure of equiatomic ternary compositions using the properties of costituent elements as descriptors [40]. A total of 990 physicochemical features including quantum numbers, electronegativities, thermodynamic quantities were used as descriptors. Analysis of the predictions made by the model revealed that 10 % of the validation were misclassified. Then a subset of features was selected by adopting Cluster Resolution–Feature Selection (CR-FS) method and this reduced the number of features to 113. The SVM model trained with the new subset of features predicted crystal structures with 93 % accuracy. Furthermore, the feature importance predicted by the CR-FS method revealed that size, electronegativity, group, period, and valence electrons were important parameters for the prediction of the crystal structure of intermetallics.

In this section, we have discussed some examples of ML applications showing how ML can assist in the interpretation of data from materials characterization and their utility in *apriori* prediction of material structures.

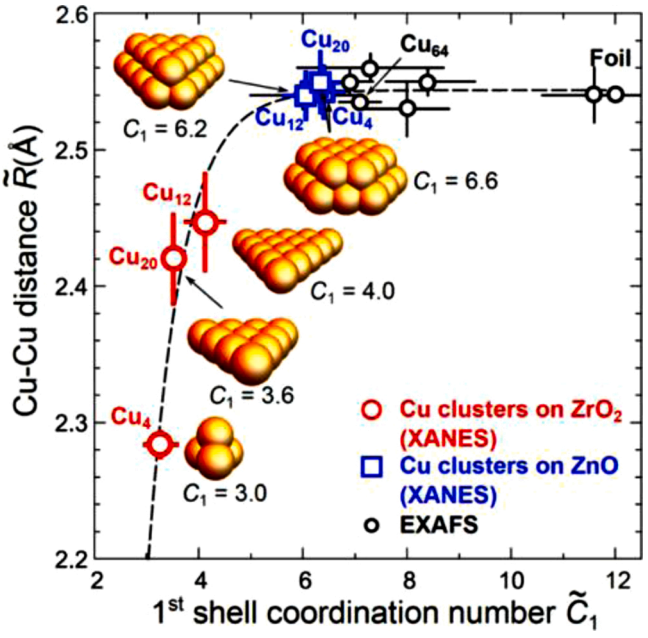

Fig. 2. First-shell CNs and effective nearest-neighbor distances for size-selected clusters on ZrO_2 and ZnO, extracted by NN method from high-temperature (375 °C) XANES data. Reproduced with permission from American Chemical Society.

4. Machine learning for prediction of catalytic reactions

Materials applications require the optimization of materials toward specific property and certain activity threshold. Hence, the ability to predict the properties and activities ahead of experimental evaluation is invaluable in materials chemistry. Research works have been reported towards developing ML models to predict materials properties and activities to obtain a structure-activity relationship since 1990s. Hattori and Kito used NN models to predict properties and catalytic activities of materials such as acid strength of materials, activity of lanthanide oxidation for butane oxidation, and selectivity in the oxidative dehydrogenation of ethylbenzene [41]. Soleimani et al. used a decision tree

model to predict the carbon dioxide solubility in polymers using experimental data collected from 515 published works [42]. The solubility of CO_2 is an important factor in the cost-effective design of fluid-polymer processes such as polymer grafting and foaming. Hazama and coworkers used 800 data points from high throughput experiments to predict the lifetime of the afterglow in phosphorescent materials [43]. These long-persistent phosphorescence materials are used in products such as clock dials, exit signs, and toys that glow in the dark. In this work, they analyzed the effect of various semiconductor thin film on $SrAl_2O_4$ based phosphorescent material. Several semiconducting oxides were deposited at different thickness and calcined at different temperature and atmospheres using high throughput experiments. Using the data, they trained a RF model and then used it to predicted the activity for each unary oxide at temperatures up to 1000 °C, with varying film thickness up to 140 nm. From this, they identified that MgO improves the afterglow lifetime at a thickness of 100 nm and calcination at 400−600 °C. They also predicted the activity for combinations of oxides and their model predicted MgO-Ga_2O_3 and MgO-ZnO to improve the lifetime of phosphorescence.

Takahashi and coworkers used machine learning on selectivity of products from 196 experimental Oxidative Coupling of Methane (OCM) reactions to (i) build predictive models for the selectivities of CO, CO_2, C_2H_4, C_2H_6, and H_2 and (ii) to understand reaction network of OCM reaction [44]. They used Mn-Na_2WO_4/SiO_2 as the catalyst and collected data by varying weight of the catalyst, amount of $CH_4 + O_2$, total flow, CH_4/O_2 ratio, and temperature. Using the collected data, they initially built six different machine learning models: Linear Regression (LR), SVM, Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest Regressor (RFR), and Extra Tree Regressor (ETR). After hyperparameter optimization and cross-validation analysis, RFR showed the best performance for the prediction task. Further analysis was performed to understand the importance of the descriptors for the selectivity predictions. This revealed that temperature plays an important role in the selectivity of CO and H_2. The selectivity of CO_2 strongly depends on the amount of $CH_4 + O_2$. However, the selectivity towards C_2H_4 and C_2H_6 shows significant dependence on several experimental parameters, as shown in Fig. 3.

To understand the OCM reaction network, several RFR models were trained to predict the selectivity of each product using all experimental parameters and selectivity of one product as input features. The cross-validation scores for the different RFR models were used to build the reaction route map. This map was constructed by calculating the cross-validation scores for the selectivity prediction of Y = (one of CO, CO_2, C_2H_4, C_2H_6, and H_2) using experimental conditions and X as features where experimental condition refers to "weight", "$CH_4 + O_2$", "total flow", "CH_4/O_2", "temperature, and X refers to selectivity of one product or none (X = none, CO, CO_2, C_2H_4, C_2H_6, and H_2), X ≠ Y. A higher cross validation score obtained from this analysis for the selectivity of one product, Y, can be interpreted as follows. In addition to experimental condition, if selectivity of X is also directly connected to the selectivity of Y, then it is possible that both X and Y may form in the same step in the reaction network. When X = none, it suggests that Y is independent of selectivity any other products hence it can be formed as the initial product in the reaction network. The reaction route map (Fig. 4) shows that C_2H_6 and CO rely more on experimental conditions suggesting that they are products of initial reaction step. Furthermore, using this map, they also suggest that dehydrogenation of C_2H_6 results in the formation of C_2H_4, and H_2 forms from the dehydrogenation of CH_4. Moreover, since the selectivities of C_2H_4, C_2H_6, and H_2 have high cross-validation score for the formation of CO, it be the final product of OCM [44]. This study demonstrates the importance of ML models in delineating

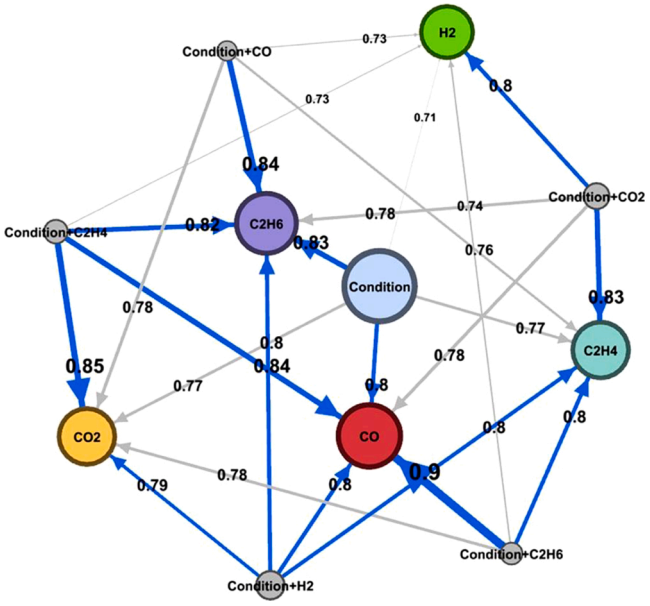

Fig. 4. OCM reaction route map created by score in cross validation of random forest regressor machine learning. CO, CO_2, C_2H_4, C_2H_6, and H_2 represent selectivity, while condition + X (X = none, CO, CO_2, C_2H_4, C_2H_6, and H_2 selectivity) represents the descriptor variable. Reproduced with permission from American Chemical Society.

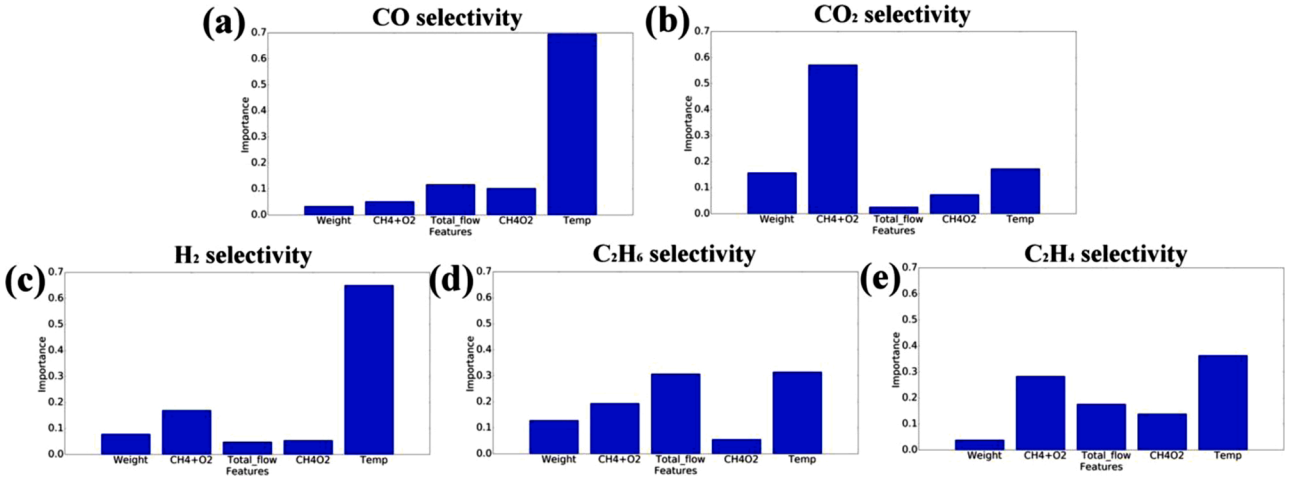

Fig. 3. Importance of descriptors within the random forest regressor model against (a) CO selectivity, (b) CO_2 selectivity, (c) H_2 selectivity, (d) C_2H_6 selectivity, and (e) C_2H_4 selectivity. Reproduced with permission from American Chemical Society.

catalytic reaction mechanisms.

In an earlier study, Takahashi and coworkers used data derived from 156 OCM reactions on Mn-Na_2WO_4/SiO_2 catalyst and tested various ML models to predict the C_2 yield. In that work, they concluded that extreme tree regressor was the best model for the task and that with a dataset with only 45 points, this model can sufficiently predict the yields to guide experiments [45].

In another study on catalysts for OCM reaction, Nguyen et al. screened 59 catalysts including reference materials under 219 experimental conditions to create an experimental dataset composed of 12,708 points [46]. They used this dataset to train ML models for the prediction of C_2 yield and the selectivity of CO, CO_2, C_2H_4, and C_2H_6. Atomic numbers and mol percentages of the metals, type of support, temperature, flow rates of Ar, CH_4, and O_2 were used as a feature to build both linear and non-linear ML regression models. After hyperparameter optimization and cross-validation analysis, a RF model was found to be the best for the prediction of C_2 yield. Among the features, the reaction temperature showed more significance in determining the activity than other features. Besides, they also demonstrated that including data with poor activity is important for the optimal performance of the ML model. Further analysis revealed that the trained ML model can be used to optimize the experimental conditions. This was achived by interpolating the C_2 yield activity with respect to temperature and the flow of CH_4 and O_2 using the trained model to identify the trends. The predicted surface by ML was much smoother than the surface generated using experimental data only (see Fig. 5).

Although most ML studies use datasets with a large number of entries, studies have demonstrated that ML can assist the search with a relatively fewer number of data points. For instance, Williams et al. used a dataset containing ammonia conversion activity of three catalysts at three different temperatures to build a random forest model to predict the NH_3 conversion activity using catalyst composition, operating conditions, synthesis variables, and elemental features available in the literature as descriptors [47]. The composition of the catalysts used for the initial dataset are RuCaK, RuMnK, and RuInK with weight percentages of 3, 1, and 12 for Ru, secondary metal (Ca, Mn, and In), and the promoter, K, respectively. Then, the model was used to predict the activity for catalysts with other secondary metals (M = Mg, Sc, Ti, V, Cr, Fe, Co, Ni, Cu, Zn, Sr, Y, Zr, Nb, Mo, Rh, Pd, Ag, Cd, Sn, Hf, Ta, W, Re, Os, Ir, Pt, Au, Pb, and Bi). The experimental validation of the predictions (Fig. 6) revealed that the model was able to correctly predict the best and top three catalysts at 3% secondary metal loading. Furthermore, the influence of the secondary metal loading was evaluated at 1 and 2 wt percentages. In this exploration, by building models with the sequential addition of data with 1 and 2 wt. %, they demonstrated that the inclusion of 10 additional data points to the full 3 wt. % dataset is sufficient to efficiently explore the design space they choose. This corresponds to 20 % of their parameter space and further addition of data gave diminishing returns. Although the prediction error was decreasing with more data,

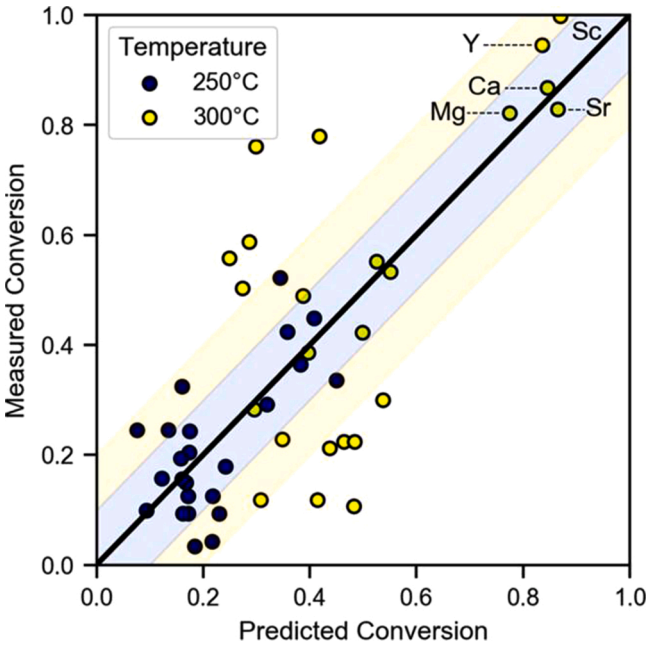

Fig. 6. Predicted ammonia conversion for secondary element-substituted catalysts. ML predictions compared to measured conversion using a leave-one-out cross-validation strategy for the entire 3,1,12 RuMK catalyst dataset. The parity line indicates a perfect prediction, the blue band indicates catalyst predictions within 10 % error, and the yellow band indicates catalyst predictions within 20 % error. The secondary element for the best catalysts at 300 °C is annotated. Reproduced with permission from American Chemical Society. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

the decrease was small with the addition more data. This can be due to several factors such as limitation of the model, not enough descriptors to explain the difference in the observed activities. Using this approach, they identified RuYK with 1, 3, 12 wt percentage as the best forming catalyst and comparison with state-of-the-art catalyst, Cs/Ru based catalyst showed higher conversion rates while using less than 7 times Ru content.

Asymmetric catalysis is an industrially important research area due to its applications in pharmaceuticals and agrochemicals where the synthesis of high-purity chiral compounds is required. The catalysts used need to be optimized for the desired chiral configuration as well as high enantiomeric excess (%ee) due to the economic significance of the products. To accelerate this, Singh et al. used a RF model to predict the %ee using molecular descriptors derived from quantum chemical calculations [48]. The descriptors include various lengths and angles, sterimol parameters, vibrational frequencies and intensities, chemical

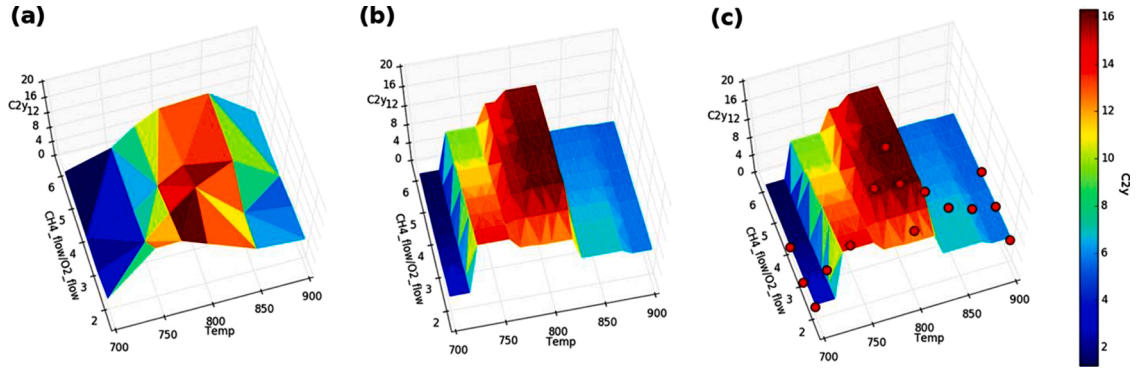

Fig. 5. Surface plot of the CH_4/O_2 ratio and temperature against the C_2 yield for Mn–Na_2WO_4/SiO_2. (a) Actual data points, (b) interpolation filling by random forest regressor, and (c) validation with separate experimental data points. Reproduced with permission from American Chemical Society.

shifts, charges, and electronic properties such as energies of highest occupied molecular orbital, lowest unoccupied molecular orbital, dipole moment, polar surface area, *etc*. The dataset was composed of 368 substrate-catalyst combinations spanning five different classes (BINOL-phosphite [L1], BINOL-phosphoramidite [L2], BINAP [L3], BINAP-O [L4], and BINOL-phosphoric acid [L5]) and their respective % ee curated from the literature. Using 80 % of this data, they trained a RF model to predict the experimentally obtained %ee. Their best model predicted the %ee for the 20 % test set with a Root Mean Squared Error (RMSE) of 8.4 ± 1.8. Then, to analyze the generalizability of their model, they trained the model on the entire dataset and tested on 43 samples with different axially chiral catalysts. The RMSE of the prediction on this previously unseen data was 8.5 and this shows the utility of the model in assisting the design of asymmetric catalysts. Furthermore, a decision tree analysis was also performed to reveal chemical insights (Fig. 7). This analysis revealed that the vibrational intensity of the substrate (VI12-16-S) plays an important role in %ee and the authors note that this can be tuned by using various substituents on the alkene/imine moiety. Further, it also shows that C═C, C═N stretching frequency intensities (VI12-16-S) higher than 14.1 and ^{13}C NMR shifts less than 39.6 is likely to yield %ee greater than 92 %. Also, they trained a model with three classes (L1, L2, and L3) of compounds and tested on the remaining two (L4 and L5). In both cases, the prediction RMSE was higher, but comparable the other results of that study. However, it is not clear if the descriptor values of these test sets are out of the training sets.

Another example of a prediction of optical properties by machine learning was performed by Chen et al. to predict the specific optical rotations of chiral fluorinated molecules [49]. They collected data of 88 chiral fluorinated molecules and their optical rotations from the literature and trained Multi Layer Perceptron (MLP), RF, and Multi Linear Regression (MLR) models to predict the specific optical rotation using Physicochemical Atomic Stereo (PAS) descriptors. Using PAS and descriptors based on the structural features that are common to all entities in their dataset, the random forest model was able to predict the optical rotation with a mean absolute error of 9.8°.

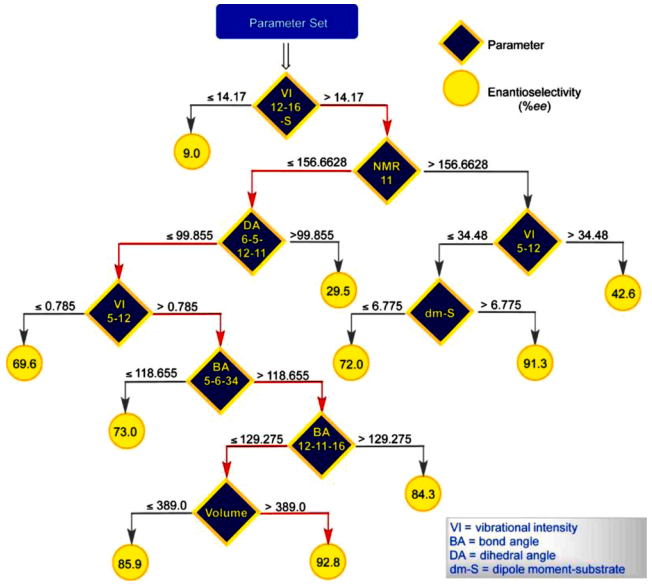

Fig. 7. Decision tree analysis for 368 reactions. The discriminating attribute at the higher level in a decision tree has a more pronounced impact on the outcome while the lower attributes tend to exert differing influences depending on the preceding set of attributes in that branch. The paths shown in red convey the combination of descriptors for the most promising substrate–catalyst combinations. Reproduced with permission from National Academy of Sciences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

So far, we have discussed the utility of ML in three areas of experimental materials chemistry: synthesis, characterization, and activity measurement. These works show how ML can aid the material discover by expediting experiments. By combining these together with modern robotics platforms, *i.e.* testing the synthesis conditions predicted by ML models using a robotic synthesis platform and validating the predicted material properties and activity using automated experimental workflows, the iterations of the material discovery cycle can be performed at an accelerated pace. This AI/ML-robotics based closed-loop approach enables the incorporation of AI/ML models at multiple stages of the discovery cycle and it requires communication and coordination with various robotics-based synthesis, characterization, and activity measurement platforms. Toward this goal, Aspuru-Guzik and coworkers have developed modular, versatile software package called ChemOS that bridges AI/ML models, robotics, and human researchers to perform the research operations [50,51]. It incorporates software modules for AI algorithms, automation and robotics, communication with characterization equipment, interaction with databases, interaction with researchers, and analysis of results. Using ChemOS, and an in-house developed robotic modular laboratory platform to handle thin films (named "Ada"), they recently demonstrated the automated material optimization using AI/ML-robotics approach. In that work, they optimized the hole mobility of spiro-OMe-TAD, an organic hole transport material used in perovskite solar cells. Since the measurement of hole mobility can be time consuming, they opted for pseudomobility (measured using four-point probe conductivity and UV–vis-NIR spectroscopy measurement), which is proportional to the hole mobility. Using this platform, they optimized the dopant concentration and annealing time. This platform was able to carry out precursor solution preparation, spin coating, annealing, imaging, measurement of optical spectra and conductance, calculate the pseudomobility, and plan the next experiment. The optimization run with 35-samples was completed under 30 h, and two different such runs lead to 0.4 mol equivalent for dopant concentration and 75 s annealing as the optimum values. One potential issue in this work is that it used a specialized robotic hardware platform, development of which can be expensive initially. This can be reduced by adapting existing robotic systems that can be modified to work in a laboratory, as demonstrated by Burger et al. [52]. They used an existing mobile robotic platform (KUKA Mobile Robot mounted on a KUKA Mobile Platform base) to work with most of standard laboratory equipment including liquid dispensing, sonication, and gas chromatograph. Using this, they attempted to optimize the parameters that maximize the photocatalytic hydrogen produced using P10 (a conjugated polymer) photocatalyst. For this optimization, they considered the parameter space encompassing five hypotheses (for each, the modulated parameters are given in squared brackets): (i) dye sensitization [Rhodamine B, Acid Red 87, and Methylene Blue], (ii) pH – [addition of NaOH], (iii) ionic strength – [addition of NaCl], (iv) wettability of the catalyst – [surfactants - sodium dodecyl sulfate and polyvinylpyrrolidone], and (v) hydrogen bonding –[addition of sodium disilicate]. For these five hypotheses, they considered the synergistic effects as well leading to a search space containing 98 million points. To explore this space, they coupled the robotic platform with an inhouse developed algorithm that performs Bayesian optimization based on Gaussian process regression and parallel search strategy. The baseline Hydrogen Evolution Rate (HER) was 3.36 $\mu mol\ h^{-1}$ for P10 – L-cysteine mixture. In the course of automated analysis, their robot-AI platform performed 688 experiments in 8 days. At the end, it identified the best catalyst formulation mixture of NaOH, L-cysteine, sodium disilicate and P10, giving a HER of 21.05 $\mu mol\ h^{-1}$ which is six-fold higher than the baseline. In this exploration, based on the parameters selected by the robot-AI platform, they infer that their platform discovered the following trends. Addition of NaCl improves the HER slightly, dye sensitization and addition of surfactants are detrimental to this photocatalyst, and addition of sodium disilicate promotes the HER. One of the

important aspects of this work is that only the vial capping station and photolysis station were tailored to be used by the robot, and thanks to the high precision positioning of the used robot, standard laboratory equipment used by human researchers was utilized for other tasks. This also means that it can be adapted perform additional characterization and experiments using standard laboratory instruments.

5. Challenges

Despite the progress made in this field, there challenges that need to be overcome to get the most out of ML applications in experimental chemistry. The main challenge is the access to training dataset that is machine accessible. This requires a community effort, especially for data gathered through experiments. In order to leverage the potential of AI/ML in chemistry the data shared should adhere to the F.A.I.R. guidelines: findability, accessibility, interoperability, and repurposability [53]. Deep learning models are typically trained using massive datasets that can be expensive to obtain in some domains such as medical imaging, experimental chemistry, *etc.* In the absence of large datasets, techniques such as transfer learning, and stacked auto encoders can be utilized. In transfer learning, a pretrained model is used as starting network and then the weights of the final few layers will be retrained while keeping the rest of the network frozen. For example, in the case of medical imaging, the starting pretrained model is usually an image classification model that was trained using a large dataset [54]. During transfer learning, the last few layers will be modified for the current problem and their weights will be optimized using the small dataset.

Feng et al. used stacked auto encoders approach to overcome the issue posed by small dataset [55]. They used a Deep Neural Network (DNN) model to predict the solidification cracking susceptibility. Their DNN model was initialized with weights obtained using stacked auto-encoder method and then trained (fine-tuned) and validated using a small dataset containing 487 entries. In this approach, instead of training the weights of the entire DNN at once, individual variational auto encoders (VAE) with one hidden layer will be trained to get the weights for each hidden layer of DNN. VAEs are trained to predict their input as the ouput with hidden layers smaller than the input. During training, VAE learns to encode the input to a smaller dimension in the hidden layer and then decodes to strating dimension as output. In the staked auto encoder approach, for a DNN with 21-(10−5)-1 structure, first a VAE with 21-(10)-21 will be trained and the optimized weights from this VAE will be used to calculate the activations of first hidden layer of DNN. Then for second layer, a VAE with 10-(5)-10 will be trained using the activation of DNN's first layer as input and output and the optimized weights will be used as the weights for the DNN's second layer. This process will be repeated for the subsequent layers to optimize the weights. Using this method, they tested multiple DNN configurations and found 21-(6-5-4-3)-1 as the optimum structure and this network showed better generalizability compared to shallow NN.

The other challenges are related with software framework, and the adoption of these frameworks, and workflow by the research community. Most of the algorithms and data processing tools used for ML are open-source and available for free. *e.g.* Scikit-learn [56], TensorFlow [57], PyTorch [58], *etc*. However, the choice of ML model is dependent on the problem in hand. In general, Support Vector Machine (SVM) model for example, works best where the number of dimensions is greater than the number of samples and when there is a clear margin of separation between classes. On the contrary, SVM model is not particularly helpful for large data sets and when the target classes overlap (*i.e.* data has more noise). Random Forest models are widely used because of their versatility. Classification and regression tasks can be well handed by RF model and it is easy to view the relative importance of the input features. The disadvantage of the RF model is that if a large number of decision trees are used, then the algorithm may be slow and thus ineffective of real-time predictions. Thus, in general RF models are quick to train, but relatively slow to predict for applications that where run-time performance is important for real-world applications. The choice of ML model can automated by using Auto-ML tools such as Tree based Pipeline Optimization Tool (TPOT) [59], which enables the automated search for best ML pipeline (includes methods for data selection, data transformation, model selection and hyperparameter optimization). Although these tools enhance the wide adoption, caution must be exercised as ML models can give superior prediction performance by unintentionally exploiting experimental artifacts or confounding variables [60]. Hence, emphasis should be made on when training the models and using predictions made by the models. Regularization and cross-validation can be performed when training the models. In terms of interpretation, some ML models offers ways to interpret the predictions for example, feature coefficients in linear models, feature importance in random forests, *etc.* For other models, model agnostic methods such as feature ablation, individual conditional expectation, permutation feature importance, Shapley values, and so forth [60,61].

6. Summary

The examples discussed in this perspective show the utility of machine learning in experiment-driven materials research. The applications show that ML can be used to predict various properties and catalytic activities using descriptors that are readily available and/or obtained by calculations. In addition, they also show that surrogate models built using literature data can be used to select experiments/catalysts for further studies. Some of the results discussed in this study indicate promise in using limited set of data to screen and identify new materials for applications including catalysis. However, caution must be exercised in using ML models as they can overfit to training data and may exploit confounding variables. Hence validation and interpretation of models need to be performed in order to understand the models. For ML to reach a stage of maturity and guide towards the accelerated discovery of new materials, advances in machine learning and integration with robotics is useful. It also seems prudent to test various ML models. Platforms integrating AI/ML with robotics are emerging and they aim to automate the discovery and optimization of novel materials with required characteristics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are thankful to the University of South Dakota-Neuroscience, Nanotechnology, and Networks (USD-N3) program, funded by the National Science Foundation, DGE-1633213 grant.

References

[1] X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, Concepts of artificial intelligence for computer-assisted drug discovery, Chem. Rev. 119 (2019) 10520–10594.
[2] B. Cao, L.A. Adutwum, A.O. Oliynyk, E.J. Luber, B.C. Olsen, A. Mar, J.M. Buriak, How to optimize materials and devices via design of experiments and machine learning: demonstration using organic photovoltaics, ACS Nano 12 (2018) 7434–7444.
[3] A.J. Medford, M.R. Kunz, S.M. Ewing, T. Borders, R. Fushimi, Extracting knowledge from data through catalysis informatics, ACS Catal. 8 (2018) 7403–7429.
[4] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J. Mater. 3 (2017) 159–177.
[5] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, Npj Comput. Mater. 5 (2019) 83.
[6] T.C. Le, D.A. Winkler, Discovery and optimization of materials using evolutionary approaches, Chem. Rev. 116 (2016) 6107–6132.

[7] L. Himanen, A. Geurts, A.S. Foster, P. Rinke, Data-driven materials science: status, challenges, and perspectives, Adv. Sci. 6 (2019), 1900808.
[8] A. Agrawal, A. Choudhary, Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science, APL Mater. 4 (2016), 053208.
[9] K. Kim, S. Kang, J. Yoo, Y. Kwon, Y. Nam, D. Lee, I. Kim, Y.-S. Choi, Y. Jung, S. Kim, W.-J. Son, J. Son, H.S. Lee, S. Kim, J. Shin, S. Hwang, Deep-learning-based inverse design model for intelligent discovery of organic molecules, NPJ Comput. Mater. 4 (2018) 67.
[10] T. Mueller, A. Hernandez, C. Wang, Machine learning for interatomic potential models, J. Chem. Phys. 152 (2020), 050902.
[11] V.L. Deringer, M.A. Caro, G. Csányi, Machine learning interatomic potentials as emerging tools for materials science, Adv. Mater. 31 (2019), 1902765.
[12] V. Venkatraman, S. Evjen, H.K. Knuutila, A. Fiksdahl, B.K. Alsberg, Predicting ionic liquid melting points using machine learning, J. Mol. Liq. 264 (2018) 318–326.
[13] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, Predicting the band gaps of inorganic solids by machine learning, J. Phys. Chem. Lett. 9 (2018) 1668–1673.
[14] A.J. Chowdhury, W. Yang, K.E. Abdelfatah, M. Zare, A. Heyden, G.A. Terejanu, A multiple filter based neural network approach to the extrapolation of adsorption energies on metal surfaces for catalysis applications, J. Chem. Theory Comput. 16 (2020) 1105–1114.
[15] M. Fujinami, J. Seino, T. Nukazawa, S. Ishida, T. Iwamoto, H. Nakai, Virtual reaction condition optimization based on machine learning for a small number of experiments in high-dimensional continuous and discrete variables, Chem. Lett. 48 (2019) 961–964.
[16] H.I. Labouta, N. Asgarian, K. Rinker, D.T. Cramb, Meta-analysis of nanoparticle cytotoxicity via data-mining the literature, ACS Nano 13 (2019) 1583–1594.
[17] C.W. Coley, W.H. Green, K.F. Jensen, Machine learning in computer-aided synthesis planning, Acc. Chem. Res. 51 (2018) 1281–1289.
[18] G. Rothenberg, Data mining in catalysis: separating knowledge from garbage, Catal. Today 137 (2008) 2–10.
[19] S. Palkovits, A primer about machine learning in catalysis - A tutorial with code, ChemCatChem (2020) n/a.
[20] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Quantitative structure–Property relationship modeling of diverse materials properties, Chem. Rev. 112 (2012) 2889–2919.
[21] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S.P. Ong, A critical review of machine learning of energy materials, Adv. Energy Mater. 10 (2020), 1903242.
[22] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, K.-i. Shimizu, Machine learning for catalysis informatics: recent applications and prospects, ACS Catal. 10 (2020) 2260–2297.
[23] J.E. Gubernatis, T. Lookman, Machine learning in materials design and discovery: examples from the present and suggestions for the future, Phys. Rev. Mater. 2 (2018), 120301.
[24] K.F. Jensen, C.W. Coley, N.S. Eyke, Autonomous Discovery in the Chemical Sciences Part I: Progress, 2020.
[25] B.R. Goldsmith, J. Esterhuizen, J.-X. Liu, C.J. Bartel, C. Sutton, Machine learning for heterogeneous catalyst design and discovery, AIChE J. 64 (2018) 2311–2323.
[26] K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai, J. Ohyama, T.N. Nguyen, S. Nishimura, T. Taniike, The rise of catalyst informatics: towards catalyst genomics, ChemCatChem 11 (2019) 1146–1152.
[27] N.E. Jackson, M.A. Webb, J.J. de Pablo, Recent advances in machine learning towards multiscale soft materials design, Curr. Opin. Chem. Eng. 23 (2019) 106–114.
[28] K. Tran, Z.W. Ulissi, Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution, Nat. Catal. 1 (2018) 696–703.
[29] F. Ren, L. Ward, T. Williams, K.J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments, Sci. Adv. 4 (2018) eaaq1566.
[30] A.O. Oliynyk, A. Mar, Discovery of intermetallic compounds from traditional to machine-learning approaches, Acc. Chem. Res. 51 (2018) 59–68.
[31] P. Raccuglia, K.C. Elbert, P.D.F. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A.J. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature 533 (2016) 73–76.
[32] K. Muraoka, Y. Sada, D. Miyazaki, W. Chaikittisilp, T. Okubo, Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials, Nat. Commun. 10 (2019) 4459.
[33] S.M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K.C. Stylianou, B. Smit, Capturing chemical intuition in synthesis of metal-organic frameworks, Nat. Commun. 10 (2019) 539.
[34] J.N. Kumar, Q. Li, Y. Jun, Challenges and opportunities of polymer design with machine learning and high throughput experimentation, MRS Commun. 9 (2019) 537–544.
[35] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, Materials synthesis insights from scientific literature via text extraction and machine learning, Chem. Mater. 29 (2017) 9436–9444.
[36] R.C. Masters, N. Stehling, K.J. Abrams, V. Kumar, M. Azzolini, N.M. Pugno, M. Dapor, A. Huber, P. Schäfer, D.G. Lidzey, C. Rodenburg, Mapping polymer molecular order in the SEM with secondary Electron hyperspectral imaging, Adv. Sci. 6 (2019), 1801752.
[37] J.-W. Lee, W.B. Park, J.H. Lee, S.P. Singh, K.-S. Sohn, A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns, Nat. Commun. 11 (2020) 86.
[38] J. Timoshenko, A. Halder, B. Yang, S. Seifert, M.J. Pellin, S. Vajda, A.I. Frenkel, Subnanometer substructures in nanoassemblies formed from clusters under a reactive atmosphere revealed using machine learning, J. Phys. Chem. C 122 (2018) 21686–21693.
[39] J. Timoshenko, A.I. Frenkel, "Inverting" X-ray absorption spectra of catalysts by machine learning in search for activity descriptors, ACS Catal. 9 (2019) 10192–10211.
[40] A.O. Oliynyk, L.A. Adutwum, B.W. Rudyk, H. Pisavadia, S. Lotfi, V. Hlukhyy, J. J. Harynuk, A. Mar, J. Brgoch, Disentangling structural confusion through machine learning: structure prediction and polymorphism of equiatomic ternary phases ABC, J. Am. Chem. Soc. 139 (2017) 17870–17881.
[41] T. Hattori, S. Kito, Neural network as a tool for catalyst development, Catal. Today 23 (1995) 347–355.
[42] R. Soleimani, A.H. Saeedi Dehaghani, A. Rezai-Yazdi, S.A. Hosseini, S.P. Hosseini, A. Bahadori, Evolving an accurate decision tree-based model for predicting carbon dioxide solubility in polymers, Chem. Eng. Technol. 43 (2020) 514–522.
[43] H. Hazama, S. Sobue, S. Tajima, R. Asahi, Phosphorescent material search using a combination of high-throughput evaluation and machine learning, Inorg. Chem. 58 (2019) 10936–10943.
[44] I. Miyazato, S. Nishimura, L. Takahashi, J. Ohyama, K. Takahashi, Data-driven identification of the reaction network in oxidative coupling of the methane reaction via experimental data, J. Phys. Chem. Lett. 11 (2020) 787–795.
[45] J. Ohyama, S. Nishimura, K. Takahashi, Data driven determination of reaction conditions in oxidative coupling of methane via machine learning, ChemCatChem 11 (2019) 4307–4313.
[46] T.N. Nguyen, T.T.P. Nhat, K. Takimoto, A. Thakur, S. Nishimura, J. Ohyama, I. Miyazato, L. Takahashi, J. Fujima, K. Takahashi, T. Taniike, High-throughput experimentation and catalyst informatics for oxidative coupling of methane, ACS Catal. 10 (2020) 921–932.
[47] T. Williams, K. McCullough, J.A. Lauterbach, Enabling catalyst discovery through machine learning and high-throughput experimentation, Chem. Mater. 32 (2020) 157–165.
[48] S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan, R. B. Sunoj, A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation, PNAS 117 (2020) 1339–1345.
[49] M. Chen, T. Wu, K. Xiao, T. Zhao, Y. Zhou, Q. Zhang, J. Aires-de-Sousa, Machine learning to predict the specific optical rotations of chiral fluorinated molecules, Spectrochim. Acta A: Mol. Biomol. Spectrosc. 223 (2019), 117289.
[50] L.M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L.P.E. Yunker, J.E. Hein, A. Aspuru-Guzik, ChemOS: Orchestrating autonomous experimentation, Sci. Robot. 3 (2018) eaat5559.
[51] L.M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L.P.E. Yunker, J.E. Hein, A. Aspuru-Guzik, ChemOS: An orchestration software to democratize autonomous discovery, PLoS One 15 (2020), e0229862.
[52] B. Burger, P.M. Maffettone, V.V. Gusev, C.M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R.S. Sprick, A.I. Cooper, A mobile robotic chemist, Nature 583 (2020) 237–241.
[53] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M. E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (2016), 160018.
[54] H. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (2016) 1285–1298.
[55] S. Feng, H. Zhou, H. Dong, Using deep neural network with small dataset to predict material defects, Mater. Des. 162 (2019) 300–310.
[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[57] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow: a System for Large-scale Machine Learning, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.
[58] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic Differentiation in Pytorch, 2017.
[59] R.S. Olson, J.H. Moore, TPOT: a tree-based pipeline optimization tool for automating machine learning. Workshop on Automatic Machine Learning, 2016, pp. 66–74.
[60] K.V. Chuang, M.J. Keiser, Adversarial controls for scientific machine learning, ACS Chem. Biol. 13 (2018) 2819–2821.
[61] C. Molnar, Interpretable Machine Learning, Lulu. Com, 2020.