Journal of Materials Chemistry A

Experimental Discovery of Novel Ammonia Synthesis Catalysts via Active Learning

Journal:	*Journal of Materials Chemistry A*
Manuscript ID	TA-ART-09-2023-005939.R1
Article Type:	Paper
Date Submitted by the Author:	16-Nov-2023
Complete List of Authors:	Jayarathna, Rasika; University of South Carolina College of Engineering and Computing, Onsree, Thossaporn; University of South Carolina College of Engineering and Computing, Chemical Engineering Drummond, Samuel; University of South Carolina College of Engineering and Computing Naglic, Jennifer; University of South Carolina College of Engineering and Computing Lauterbach, Jochen; University of South Carolina, Chemical Engineering

ARTICLE

Experimental Discovery of Novel Ammonia Synthesis Catalysts *via* Active Learning

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Rasika Jayarathna,[a] Thossaporn Onsree,[a] Samuel Drummond,[a] Jennifer Naglic,[a] and Jochen Lauterbach*[a]

The importance of ammonia synthesis under mild conditions is increasing due to growing interest in ammonia for large-scale applications of renewable energy storage and utilization. Being one of the most investigated reactions in heterogeneous catalysis, multi-dimensional literature data are available for this reaction as a base to explore new catalysts. Machine learning (ML) can be applied to develop models using existing literature data. However, ML models developed only from literature data may not be able to efficiently predict or suggest new catalyst formulations without additional experimental data. Herein, we present an active learning (AL) framework for accelerating the discovery of novel ammonia synthesis catalysts initiated by literature data to explore a pre-determined search space based on domain knowledge efficiently. This framework generates and selects features for the ML model to capture the effects of catalyst preparation variables, kinetics, thermodynamics, support, and interactions between Ru, promoter, and the support for data mined from literature. Experimental results showed that the AL framework could discover novel catalysts that exceeded the activity of many state-of-the-art catalysts. AL reduced the number of experiments necessary to reach the best catalyst in the search space by 50%, even when no training data related to the best catalyst exists. Furthermore, AL gave insight into the properties of the catalysts that contribute to higher ammonia synthesis activity.

1 Introduction

The traditional material discovery process relies upon humans' domain expertise and intuition based on existing knowledge.[1,2] Here, the potential materials that could meet the target properties are identified based on the analysis of the properties of similar materials that are published in the scientific literature. Then, the most potential material is synthesized and tested to meet the target property. Since the goal is often met for the first time, the material is characterized to identify the reasons for the failure. Based on this knowledge, another potential material is synthesized, tested, and characterized again if it fails to meet the target. Likewise, this process is repeated until the material with the target property is discovered, which is time and resource-consuming due to the time allocated for material characterization and potential biases and failures in human intuition. Materials such as heterogeneous catalysts further increase the time and resources required for this process since the catalytic properties are governed not only by the elemental properties but also by the reaction conditions, which necessitate more extensive testing conditions and characterization.

Artificial intelligence (AI) techniques, such as active learning (AL) (also known as sequential learning, on-the-fly learning, or adaptive design), can assist in material discovery by emulating the aforementioned traditional process but in an accelerated timeframe without the need for time-consuming characterization and analysis.[1,3,4] Here, a machine learning (ML) model is developed from the experimental data mined from the literature or generated in the lab, and the next experiment to run is selected by the model from a pre-determined search space. After the experiments are run, the model is retrained with newly obtained experimental data, and the next experiment to run is suggested iteratively, just as human researchers do until the material with the target property is discovered. AL using experimental data has significantly reduced the number of experiments required to meet the target properties for different types of materials.[5–9] This approach is advantageous over ML approaches utilizing computational data generated via density functional theory (DFT) which are mostly limited to experimentally not validated property predictions for novel materials [10] and especially for heterogeneous catalysts since the inherent complexities of real catalysts are not represented and the lack of control and knowledge over experimental process parameters for AL in the computational data.[9,11]

In this context, data mined from the literature offer real-world catalyst data and domain knowledge about the catalysts

[a] Department of Chemical Engineering, University of South Carolina, 541 Main St., Columbia, SC 29208, USA.
† Electronic Supplementary Information (ESI) available.
See DOI: 10.1039/x0xx00000x

that are generated over decades of time for catalyst discovery. This data is often high-dimensional compared to literature data for other materials due to the vastly different reaction conditions used for catalyst testing and different materials used for different components of the catalyst. These dimensions govern the surface properties of materials, such as activity, selectivity, and stability of the catalyst, instead of the bulk properties, which are interesting to researchers for most materials. Literature data have been used in heterogeneous catalysis to train ML models for novel catalyst exploration.[12–14] However, these models have added a batch of experimental data for the ML model trained on literature data to augment the predictions (since the predictions based only on literature data were not correct) rather than suggesting the next experiments to run as in AL. AL methods in experimental heterogeneous catalysis (also largely in experimental material science) have largely focused on either using only experimental data as the initial training dataset [6,8] or simulated literature data, where the next "experiments" to run are literature data that are not used to train the ML model.[15,16] Little work has been done where literature data is solely used as the initial training dataset for AL focused on truly novel experimental catalyst discovery that is not reported in the literature.

Heterogenous catalysts for ammonia (NH_3) synthesis are a class of catalysts where a considerable amount of literature data has been generated over decades in search of a catalyst with ever greater activity. This is because ammonia is one of the four most important materials in the modern world due to the production of fertilizers (around 80% of produced ammonia) and the subsequent food production for 70% of the world population.[17–19] Recently, due to the requirement of energy storage technologies for renewable energy (e.g., wind, solar), ammonia has also been identified as the only carbon-free strategy that is economically feasible on large scales (MWh to TWh) based on existing technologies.[20] Here, ammonia is a storage material and a carrier for hydrogen (H_2) due to advantageous characteristics of ammonia compared to other hydrogen storage materials, such as high energy density (12.8 GJ m^{-3}), high hydrogen storage capacity (17.7 wt.%), existence as a liquid at 293 K and 0.8 MPa and existing infrastructure for distribution.[21] This enables ammonia to be used as a next-generation fuel in novel engines or fuel cells for energy generation without carbon emissions.[21] The well-known Haber-Bosch (HB) process is employed in large-scale facilities for ammonia synthesis, operating at high pressures (10-25 MPa) and temperatures (723-873 K) using hydrogen derived predominantly from natural gas and nitrogen from the air with the aid of an iron (Fe) catalyst.[22] However, this process is energy-intensive, accounting for about 2% of world energy consumption with a high CO_2 footprint and responsible for 1.6% of the world's anthropogenic CO_2 emissions.[23] This has spurred research and development for sustainable synthesis of "green" ammonia to reduce the dependence on fossil fuels and mitigate climate change.[22,24,25]

Due to the distributed and intermittent nature of renewable energy, small-scale ammonia production must utilize hydrogen generated via water electrolysis for ammonia synthesis. Currently, methane is used as the hydrogen source, and the small-scale HB process is considered uneconomical.[26,27] In this context, catalytic membrane reactors (CMR) enable modular ammonia synthesis by selectively separating ammonia from the reaction mixture, thereby removing the thermodynamic limitation, which increases the NH_3 synthesis rate and reduces the energy of separation for ammonia, compared to condensation that is used in the HB process.[27,28] In addition, this process favors relatively lower reaction conditions of pressures (<5 MPa) and temperatures (573-723 K) to overcome membrane deactivation, energy-intensive pressure ramping, and cumbersome high-pressure installations for hydrogen coming from the electrolyzer.[28–30]

Ruthenium (Ru), being the most active metal for ammonia synthesis, can catalyze the reaction under the milder conditions required for CMRs.[31,32] Current state-of-the-art Ru-based catalysts for this reaction, such as nitrides, hydrides, and electrides suffer from complex synthesis processes and higher sensitivity to air and moisture, prohibiting them from large-scale usage.[33] Ru nanoclusters or nanoparticles on metal oxide supports offer a promising alternative to those catalysts due to the ease of synthesis by wetness impregnation, stability, and durability. Moreover, these catalysts have polydisperse size distributions for Ru nanoparticles, which increase the NH_3 synthesis rates compared to catalysts synthesized using different methods with more uniform size distributions.[34] Particle size and structure control, promoter enhancement, alloying, and optimizing the support effects have been proposed as strategies to increase the activity of these types of catalysts.[32] As one of the most investigated heterogeneous catalytic reactions in the literature, these strategies have been implemented, and various types of novel Ru catalysts containing different promoters and supports under widespread reaction conditions have been developed.[32,33] However, there is still a need to discover cost-effective Ru-based catalysts for ammonia synthesis due to the high cost of Ru.[35]

This work focuses on the experimental search of novel ammonia synthesis catalysts using AL initially based on the literature data, for the first time according to the best knowledge of the authors. Firstly, activity data for thermocatalytic Ru-based ammonia synthesis catalysts are mined from the literature. Then an ML model is developed from features that are engineered and selected for this application so that the effects of catalyst preparation variables, kinetics, thermodynamics, support, and interactions between Ru, promoters, and the support are captured. Based on the domain knowledge of this reaction, a search space is determined for a Ru-based catalyst that has not been explored in the literature. Then AL is used to experimentally explore the search space efficiently and identify what properties of the elements in the catalyst are affecting the model to select the next catalyst to run. This general framework could be employed for heterogeneous catalyst development for reactions other than ammonia synthesis from the data mined

from the literature, combining both domain knowledge and artificial intelligence.

2 Materials and methods

2.1 Data Mining and Selection

The experimental data on conventional thermocatalytic ammonia synthesis catalysts tested on lab-scale fixed bed reactors were mined from existing publications found through online resources, such as Elsevier, Wiley, American Chemical Society, and Springer. Over one hundred publications between 1972 and 2022 were reviewed, and 56 publications were chosen that include catalysts synthesized using wetness impregnation that utilizes Ru as the sole or one of the multiple active metals. Out of these, 24 publications were discarded due to incomplete data (e.g., BET support surface area, space velocity), inconsistencies in data reporting with missing parameters that prevent the data from being converted to the same format (e.g., space velocity, synthesis rate), and outliers (e.g. pretreatment processes and complex supports). Removed outliers included catalysts with pretreatment processes that use CO instead of H_2 for reduction and support pretreatment processes that use H_2 pretreatment of activated carbon instead of pretreatment under an inert atmosphere, which is predominantly used in the literature. The removed complex supports contained oxide mixtures, solid solutions, and oxides with doped elements. The feature engineering for these supports is too complex to be used with other supports, such as single metal oxides and perovskites due to non-uniformity and the local effects that govern the activity of the catalysts compared to single metal oxides and perovskites, which have a more uniform structure. It should also be noted that missing values for pretreatment temperatures and times of the oxide supports were imputed based on values from the literature since these parameters do not vary in an unpredictable wide range like BET support surface area and space velocity. With the remaining 32 publications, 936 data points were extracted as data for ML model development. Most ammonia synthesis data are in the form of images in the literature. Hence, a plot digitizing freeware, 'Web Plot Digitizer' [36] was used to convert the graphical data into numerical data. Tabular data were extracted manually from the publications instead of using natural language processing techniques.

2.2 Ammonia Synthesis Dataset

Different formats have been used for reporting the catalyst design variables (e.g., Ru loading, promoter loading, etc.) and catalyst operational variables (e.g., space velocity, temperature, and pressure). Hence, all those variables were converted to the same format to be used for AL. For instance, reported Ru-to-support and promoter-to-Ru molar ratios were converted to weight percentages (wt%), and operational variables were converted to the same units. The initial features used for data mining are presented in Table S1. As a measure of the activity of the catalysts, the NH_3 synthesis rate was chosen as the target variable for ML since the majority of the publications have opted to use this metric. The NH_3 synthesis rate was re-calculated using Eq. (1) for publications where the ammonia synthesis activity was reported in other units:

$$r = \frac{(x_{out} - x_{in})}{(1 + x_{out})} \cdot \frac{V_0}{m} \cdot \frac{1000}{22.08} \quad (1)$$

where r is the NH_3 synthesis rate (mmol h^{-1} g_{cat}^{-1}), x_{out} is the ammonia concentration in the outlet (mol %), x_{in} is the ammonia concentration in the inlet (mol %), V_0 is the flow rate of the gas (L h^{-1}), m is the mass of the catalyst sample (g), and 22.08 is the molar volume of a stoichiometric mixture of the H_2 and N_2 gas (L mol^{-1}).

The literature data are found to be imbalanced, where the target variable distribution is not similar to a normal distribution. Fig.1 (a) shows the distribution of the reported sythesis rates, which exhibits a highly skewed distribution with a long tail. The majority of ML algorithms have been developed for balanced or normal distributions for the target variable.[37] In other words, ML models developed from imbalanced/skewed data would have poor performance and provide misleading predictions. As a solution, the target NH_3 synthesis rate variable was transformed to a natural log scale before being used for ML model development.[38] This natural

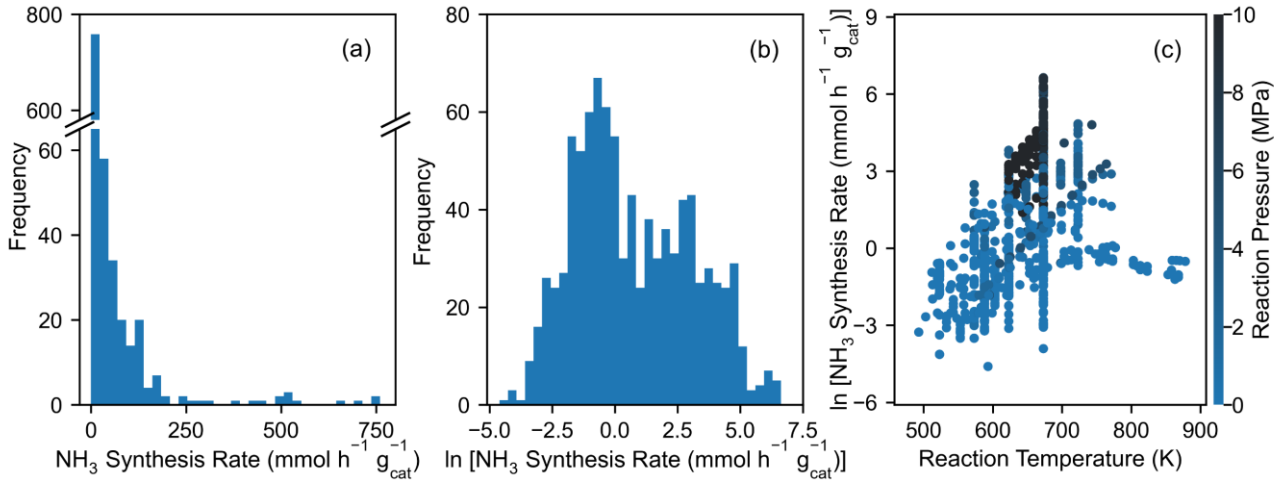

Fig. 1 Data distribution of NH_3 synthesis rate in (a) real and (b) natural log scales and (c) as functions of reaction temperature and pressure.

log transformation of the NH_3 synthesis rate variable results in a near-normal distribution, as shown in Fig. 1 (b). This transformation would ease the application of ML models, but the model accuracy should be evaluated by converting the predicted values (ln y) to real values using inverse transformation (e^y).

Fig. 1 (c) shows the distribution of the natural log (ln) of the NH_3 synthesis rate used for machine learning versus reaction temperature and pressure. It should be noted that these rates correspond to different space velocities. Furthermore, the vast majority of the data points in this plot have a stoichiometric H_2/N_2 ratio of 3 with no initial ammonia concentration. It can be seen that the synthesis rate increases with pressure. Considering the general trend, the increasing temperatures increase the synthesis rates up to 400 °C and then decrease with temperature. These trends conform to Le Chatelier's Principle. Since ammonia synthesis is a mole-reducing reaction: $N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g)$, higher pressures would shift the equilibrium to the right to produce more ammonia. The synthesis rate increases initially with the temperature predominantly due to kinetics. However, when the rate reaches the thermodynamic limit, it is decreased since ammonia synthesis is an exothermic reaction.

Fig. 2 shows the distribution of data points in terms of key reaction conditions, Ru loading, as well as promoters and supports used. The distribution of these parameters is not uniform and is filled with gaps in between. The data primarily consists of lower space velocities, pressures, and Ru loadings suitable for modular ammonia synthesis using CMRs. Fig. 2(e) illustrates that most catalysts have no promoters, and a large number of promoters are unexplored in the literature with only a very small number of doubly promoted catalysts. The most investigated promoters are Cs, Ba, and K used as the single promoter in the catalyst. The remaining promoters include other alkali metals (Li, Na, K, Rb), other alkaline earth metals (Ca, Sr), lanthanoid metals (Ce, La, Nd, Pr, Dy, Gd, Sm), metalloid (B) and transition metals (Co, Rh, Ir). MgO and graphite have the largest share of supports. It should be noted here that graphite is graphitized activated carbon synthesized by thermal treatment, which prevents the carbon hydrogenation to methane that occurs under ammonia synthesis conditions. Moreover, this process removes heteroatoms on the surface of the support and makes support surfaces more uniform for different types of activated carbon. The pretreatment parameters used as features in ML are employed to capture the changes on the surface of the graphite support. The remaining supports include other oxide supports such as CeO_2, α-Al_2O_3, Pr_2O_3, and oxides with two metal ions such as perovskite and spinel-type oxides. The exact amount of data points for every promoter and support combination is shown in Table S2, and the dataset has 57 unique promoter-support combinations. The lack of two-

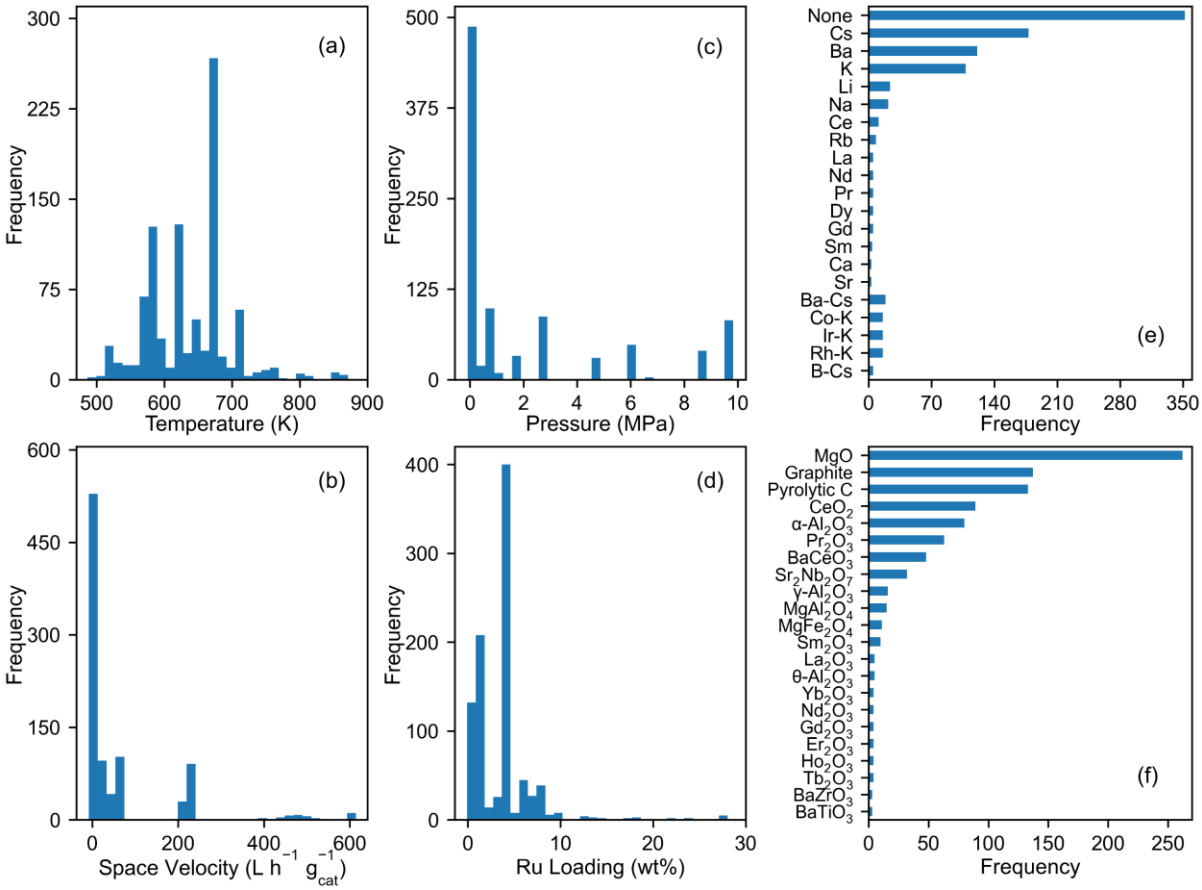

Fig. 2 Data distribution of reaction conditions: (a) temperature, (b) pressure, and (c) space velocity, as well as catalysts: (d) Ru weight loading, and (e) types of promoters and (f) supports.

promoter combinations in the dataset shows a potential for novel doubly promoted catalyst exploration due to the ease of synthesis of catalysts with promoter variations using wetness impregnation. Table S2 shows that the vast majority of doubly promoted Ru catalysts are on carbon supports, which suggests those data should be used in combination with data for catalysts with oxide supports for doubly promoted catalyst exploration using ML models.

2.3 Model Development

2.3.1 Model Selection

Tree ensemble methods have shown higher accuracy for small datasets for predicting catalyst performance.[39] Random forest regressor (RFR), extra trees regressor (ETR), and extra gradient boosting (XGB) ensemble learning algorithms were selected for this study. The first two algorithms were implemented by the Python package scikit-learn (version 1.1.2), and the third algorithm was implemented by the Python package XGBoost (version 1.6.2).

RFR is an advanced algorithm made of multiple decision tree regressors.[40] During the training process, different subsets (i.e., different features) of the training data are generated from the original training data with randomness and replacement, also known as bootstrapping. Then, each decision tree regressor is independently developed using the subset of the training data. The prediction of RFR is an average value from the predictions of all decision tree regressors. Compared to a single decision tree regressor, the randomness of RFR prevents the overfitting of the model while improving the accuracy of the model predictions. Therefore, the RFR is significantly more generalized.

ETR is an extreme version of RFR. [41] Randomness is applied in one more step. ETR randomly selects a candidate feature as a threshold to be split at decision tree nodes, compared to RFR, where the threshold is optimized. Therefore, the variance of the ETR model can be further reduced by slightly increasing biases in the second randomness. In practice, variance reduction is necessary, leading to overall better accuracy of the model predictions.[41] All crucial hyperparameters of ETR are like those of RTR. However, it is essential to note that the bootstrapping of ETR can be canceled. The ETR model is built using the whole training data, while randomness is employed only at the selection of splitting features. This avoids introducing biases from different subsets of the training data generated. Furthermore, as randomized splitting in decision tree nodes, the algorithm is not too strongly dependent on features or patterns in the data. In other words, an ML model developed by ETR could be more generalized than that by RFR.

XGB is a gradient-boosted trees algorithm.[42] It is generally combined with many simple, weak decision tree models. In a learning process, for each boosting iteration, a new decision tree regressor is developed from errors/residuals of previous decision tree regressors, and the new tree regressor is assembled with prior ones to make a prediction. Compared to RFR, where the trees are separately grown, the trees of XGB are done in series. Therefore, the prediction of XGB is from the sum of all the trees' predictions. Each tree prediction can be scaled by 'learning_rate', a hyperparameter, to avoid overfitting the model. The series of XGB ends when either no errors/patterns are to be learned or the trees are not deep enough. It is important to note that XGB optimizes an objective function (consisting of a differentiable convex loss function and a penalty term for model complexity) using second-order Taylor's expansion. This results in a faster calculation compared to other tree-based ML algorithms. Other hyperparameters, mainly related to decision tree regressors, of XGB can be found in the literature.[42]

2.3.2 Feature Engineering

In addition to the initial features in Table S1, additional features were considered. Firstly, features related to the thermodynamics and kinetics of the ammonia synthesis reaction were considered, since the reaction conditions are vastly different within the dataset.

The dissociative adsorption of nitrogen is considered to be the rate-determining step for ammonia synthesis for Ru-based catalysts similar to the Fe-based catalysts.[43] Catalytic kinetic equations based on this basis show that partial pressures of gases (hydrogen, nitrogen, and ammonia) in the reaction mixture and equilibrium constant for ammonia synthesis have a direct relationship with the NH_3 synthesis rate.[44] Hence, the equilibrium constant for ammonia synthesis was calculated according to Gillespie and Beattie.[45,46] As an approximation for the partial pressures of gases in the reaction mixture, initial partial pressures of gases were calculated to be used as features. As a feature of reaction thermodynamics, the thermodynamic limit for the NH_3 synthesis rate was calculated by plugging the maximum ammonia concentration under the reaction condition into Eq. 1. This concentration was calculated by using the equilibrium constant and fugacity coefficients (Newton-Cooper equations [46]) .[43]

The usage of elemental properties (Table S3) for the generation of features has enabled the ML models in heterogeneous catalysis to experimentally explore and discover novel catalyst formulations.[47–49] Features have been generated for the active metals and promoters (for metal-promoter interactions) as well as for the support by taking the weighted mean (Eq. 2) and weighted mean absolute deviation (Eq. 3) of elemental properties (named as elemental features from now on) of the elements in active metals and promoters or support.

$$\bar{x} = \frac{\sum_i^N n_i x_i}{\sum_i^N n_i} \tag{2}$$

$$\hat{x} = \frac{\sum_i^N n_i |x_i - \bar{x}|}{\sum_i^N n_i} \tag{3}$$

where $\bar{x}$ is the weighted mean, $\hat{x}$ is the weighted mean absolute deviation, N is the number of elements present in the

metal and promoter or support, n_i is the mol fraction of the individual element, and x_i is the value of the elemental property for the element.

Elemental features were generated for metal and promoter, support, and metallic elements in the support. Features for the electronic properties of the support were generated by features given for chemical compounds (Table S4) and converting the density of states sourced from the materials project database [50] (material ID numbers given in Table S5) to features using the Python package matminer (version 0.8.0).[51,52]

In heterogeneous catalysts, the interactions between metals, support, and promoters also affect the catalyst's performance. Features describing such interactions are less utilized in ML models found in the literature. Promoter-support interactions can be significant when a large amount of promoters are dispersed on the support surface.[53–57] The effects of these interactions can be captured by promoter-support elemental features. Wang *et al.* demonstrated through DFT calculations that the formation energy of the most thermodynamically stable metal alloy (formed by the metal of the nanoparticle and metal of the oxide) is a descriptor for strong metal support interaction (SMSI).[58] Hence, the most thermodynamically stable metal alloys formed at Ru-support interfaces were sourced from the materials project,[50] and their formation energies and generated elemental features were used as features. Formation energies of the support were used as a feature for metal-support interactions in metal oxides, as illustrated by previous DFT studies.[59,60] In addition, metal precursor components used in wetness impregnation and molar ratios between support, active metals, and promoters were also generated as features.

2.3.3 Feature Selection

The generation of hundreds of features creates a high-dimensional dataset that includes both informative and redundant or non-informative features.[61] These features cause problems in the ML model, such as overfitting, poor predictions, higher computational cost, and harder interpretability.[62] Feature selection methods are used to find a subset of features to remove redundant or non-informative features. Hence, a method known as Boruta [63] implemented in Python package boruta_py (version 0.3) was used for feature selection. Here, pseudo or fake features are created based on the given features, and then the model is trained with both pseudo and given features. Then features are selected by removing features based on a binomial distribution where the probability of removed features having feature importance higher than any of the pseudo features is 0.5%. Although some ensemble learning ML models, such as RFR, ETR, and XGB, inherently selects a subset of features during model training, the Boruta method provides a cut-off value for feature importance based on pseudo features to ease the model interpretation and alleviate the other problems that comes with redundant or non-informative features.

2.3.4 Model Evaluation

Leave one group out cross-validation (LOGOCV) was used to quantify the performance of the ML model. The literature data were divided into groups so that each group has a unique catalyst composition regardless of the weight loadings of the components, pretreatment parameters, and reaction conditions, as shown in Table S2. In each iteration of the model evaluation, one group is used as testing data, while the remaining groups are used as training data. This method has the advantage of evaluating the true extrapolative capability of the ML model, since every time the developed model predicts the activity of a new (unseen) catalyst system,[64] compared to traditional cross-validation methods such as *k*-fold cross-validation, where the whole data is divided randomly into training (*k*-1 fold) and testing (one fold) data. When the dataset is randomly divided, the activity of the same catalyst could be in the training data and testing data at different reaction conditions. The model accuracy would be evaluated mostly based on the interpolation or extrapolation capability of activity based on reaction conditions by the ML model and not on the different types of elemental features of the catalyst.

2.3.5 Catalyst synthesis, Experimental validation, and Active Learning

Details of catalyst synthesis and catalyst testing for experimental validation are found in the supplementary information. The experimental data were used to simulate active learning based on the following pathways. Three possible pathways are used to predict the next experiment for AL as used by Ling *et al* [5] in a pre-determined search space based on the model predictions for every catalyst with feature vector **X**.

(1) Maximum uncertainty (MU): Catalyst with the predicted maximum uncertainty (standard deviation: $\sigma(\mathbf{X})$) for the NH_3 synthesis rate.

(2) Maximum expected improvement (MEI): Catalyst with the predicted maximum value (mean: $\mu(\mathbf{X})$) for NH_3 synthesis rate.

Experiments suggested based on MU are 'exploration' of the search space since the model is most uncertain about these catalysts. MEI-based experiments are 'exploitation' since the model is most certain that these catalysts have the maximum activity. Efficient exploration requires a balance between 'exploration' and 'exploitation' to gather information about unexplored areas and test model predictions for maximum activity simultaneously. Since the catalysts in the search space are predicted to have activities with different means and standard deviations, the next experiment to run for efficient exploration is done based on the calculation of likelihood (L) (Eq. 4) to standardize the activity and identify the outliers with a higher activity that are worthy of exploration.

(3) Maximum likelihood improvement (MLI): Catalyst with the predicted maximum value for likelihood, i.e., the catalyst that is most likely to have a higher NH_3 synthesis rate than the best catalyst experimentally found in this search space in the previous iteration.

$$L = \frac{\mu(\boldsymbol{X}) - y_{max}}{\sigma(\boldsymbol{X})} \quad (4)$$

The parameter, y_{max}, is the maximum experimental NH_3 synthesis rate in the search space.

The means and standard deviations of the machine learning models were calculated by the Python package scikit-optimize (version 0.9.0). For instance, means and standard deviations for the models RFR and ETR are calculated by Eq. 5 and Eq. 6.[65] Here, the mean and standard deviation of the predicted NH_3 synthesis rate for a single catalyst by RFR or ETR model with T regression trees are based on the inherent mean $\mu_i(\mathbf{X})$ and standard deviation $\sigma_i(\mathbf{X})$ of every regression tree.

$$\mu(\boldsymbol{X}) = \frac{1}{T} \cdot \sum_{i=1}^{T} \mu_i(\boldsymbol{X}) \quad (5)$$

$$\sigma^2(\boldsymbol{X}) = \frac{1}{T} \cdot \left(\sum_{i=1}^{T} \mu_i^2(\boldsymbol{X}) + \sigma_i^2(\boldsymbol{X}) \right) - \mu^2(\boldsymbol{X}) \quad (6)$$

3 Results and Discussion

3.1 Model Accuracy and Selection

The three ML models, RFR, ETR, and XGB were evaluated using LOGOCV after feature selection by Boruta. The hyperparameters of the ML models were tuned using grid search and the results are shown in Table S6. Parity plots for the models are shown in Fig. 3. Here, the predicted NH_3 synthesis rates are the predictions of each testing dataset/ group in Table S2 used during LOGOCV. The evaluation metrics R^2, mean absolute error (MAE) and root mean squared error (RMSE) were calculated for both in natural log and real scale and are shown in Table 1.

The ETR model (Table 1) had the best R^2 (0.81), RMSE (32.3), and MAE (11.3) in real scale compared to the RFR and XGB models, where their R^2 and MAE values were less than 0.70 and higher than 12, respectively. From the literature, the ETR algorithm has been reported to perform well in similar applications. For example, Mine *et al.* developed ML models for predicting the yield of C_2 products for oxidative coupling of methane utilizing heterogeneous catalyst data mined from literature, and the ETR model had the highest accuracy (R^2 = 0.736)[15]. Therefore, this model is further used for interpretation and novel catalysts predictions.

3.2 Feature Importance and Interpretation

Boruta method selected 160 features out of 536 generated features, as shown in Table S7. The selection of over 100 features is not surprising since the dataset contains catalysts with many element combinations over a wide range of reaction conditions, and many features are needed to model their behavior. The feature importance for the top 20 selected features is plotted in Fig. 4, along with summations of feature importance for categorized features in the doughnut chart. Feature importance suggests that reaction conditions have the most significant effect on the NH_3 synthesis rate, accounting for around 75% of the total feature importance. This can be explained by the different reaction conditions in the dataset that are spread across a wide range. The remaining features related to the catalyst account for around 25% of the total feature importance. They are categorized according to their roles, also shown in Table S7. The selected features include catalyst pretreatment parameters that affect the performance of any heterogeneous catalyst. The selected features for catalyst preparation included the mols of carbon, CO ligands, NO_X, and chlorine in catalyst precursors, which are consistent with the effects reported in the literature. It is reported that chlorine precursors negatively affect the NH_3 synthesis rate by withdrawing electrons [66,67] and sometimes positively.[68] The Ru precursors with CO ligands (e.g., Ruthenium carbonyl) and carbon groups (e.g., Ruthenium acetylacetonate) affect the formation of nanoparticles due to the ease of decomposition and the subsequent higher dispersion of the nanoparticles.[69] Anionic oxygen in precursors in alkali and alkaline earth promoters determine the decomposition temperatures affecting its behavior.[27] Pretreatment parameters for the support were selected as expected since those would describe the properties of the support especially graphitized carbon as explained earlier.

Table 1 Model evaluation results for the selected ML models (mean and standard deviation of R^2, MAE, and RMSE).

	ML model	RFR	ETR	XGB
Natural log scale	R^2	0.8503 (0.0029)	0.8632 (0.0017)	0.8182 (0.0818)
	MAE	0.6971 (0.0084)	0.6731 (0.0045)	0.7437 (0.0744)
	RMSE	1.0496 (0.0236)	0.8691 (0.0055)	1.2407 (0.0154)
Real scale	R^2	0.6160 (0.0165)	0.8125 (0.0036)	0.6702 (0.0670)
	MAE	14.85 (0.22)	11.30 (0.08)	12.96 (0.13)
	RMSE	43.67 (2.17)	32.32 (0.31)	53.32 (0.37)

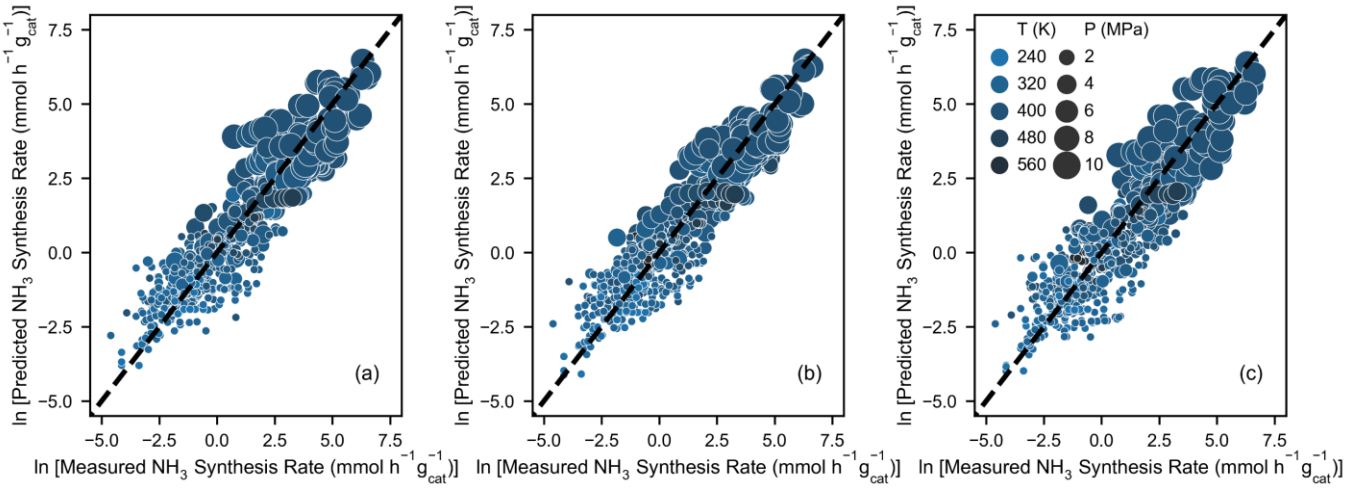

Fig. 3 Predicted versus measured synthesis rates (natural log scale) at pressures and temperatures for (a) RFR, (b) ETR, and (c) XGB models.

The NH_3 synthesis rate of the catalyst is enhanced by the transfer of electron density to the antibonding π-orbitals of adsorbed N_2 on Ru, and the subsequent weakening of the nitrogen triple bond (945 kJ mol^{-1}) and dissociation of N_2.[33] The electronic properties of the promoters and support in the catalyst play a key role in determining the ammonia synthesis activity by either facilitating the transfer or donating extra electron density. Electronic properties, such as 1st ionization energy, 2nd ionization energy, electrical conductivity, and polarizability have been selected as elemental features for metal-promoter interactions. The transfer of electron density from the promoters dispersed on the support to Ru is determined by those electronic properties. Moreover, the molar ratios of promoters (alkali metals and alkaline earth metals) to Ru determine the coverage of Ru nanoparticles by promoters which affect the catalyst activity.[67]

Features selected for the support also include elemental properties related to the electronic properties, such as the ones selected from the density of states, band center, HOMO energy, 1st ionization energy, electronegativity, valence orbital energy, and the number of valence electrons. The higher NH_3 synthesis rates obtained by using basic supports (e.g., Pr_2O_3) over acidic supports (e.g., γ-Al_2O_3) have been attributed to the higher electron density in basic supports.[70] Reducible supports, such as CeO_2, $BaCeO_3$, and $BaTiO_3$ could further increase the NH_3 synthesis rate by donating further electrons.[70–72] These electronic properties and all the other elemental features of the support were used by Karakaya *et al.* to develop an ML model for the experimental discovery of a novel mixed oxide support for a Ru-based ammonia synthesis catalyst using only experimental data.[49] Features related to metal-support interaction can be justified by the enhanced NH_3 synthesis rates reported due to enhanced electron transfer from support to the Ru by strong metal-support interactions.[70]

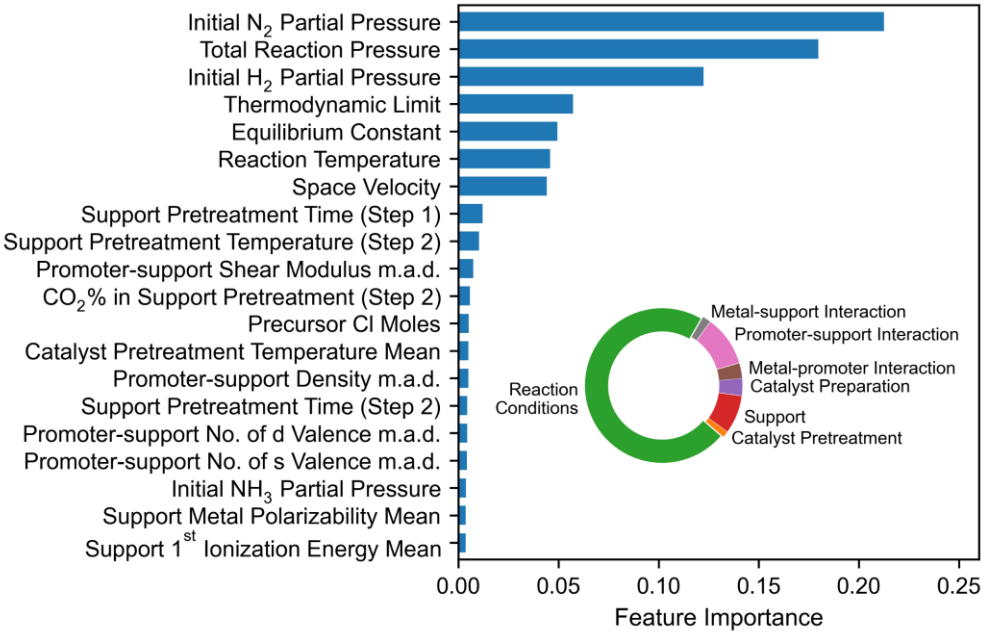

Fig. 4 Top 20-feature importance of the ETR model (bar chart) and the summation feature importance of all features categorized according to their function (doughnut chart).

Most of the selected features are related to promoter-support interactions. It has been experimentally demonstrated that promoters interact with the support, and the dispersion of the promoter is primarily impacted by the properties of the support.[67,73] Sometimes, the promoter reacts with the support and forms compounds on the support surface. For instance, the basicity of the support influences how the alkaline promoter is spread on the catalyst surface.[67,73] A basic oxide does not react with the alkali metal promoters and remains adjacent to the Ru nanoparticle, whereas it reacts with acid sites on an acidic oxide, causing inefficient utilization of the alkali metal promoters.[67,73] The large number of features selected for the promoter-support interaction and their summation of feature importance suggests the higher impact of the promoter and support on the NH_3 synthesis rate of the catalyst.

3.3 Novel Catalyst Discovery via Active Learning

3.3.1 ETR Model Developed from Literature Data

Based on the current domain knowledge of Ru-based ammonia synthesis catalysts, a search space was chosen for exploration using AL. The goal was to increase the activity by promoter enhancement for a Ru-based catalyst on a fixed oxide support due to the long experimental time for a variety of supports. Praseodymium oxide (Pr_2O_3) was selected as a support for the Ru based on the high NH_3 synthesis rates reported in the literature.[74,75] This is due to the alleviation of hydrogen poisoning and results in higher NH_3 synthesis rates at higher pressures.[75] Further increases in activity by the utilization of promoters for Pr_2O_3 are barely investigated in the literature. Moreover, adding multiple promoters and implementing the synergistic effect of promoters is an effective way to achieve a catalyst with far better performance than using just one promoter.[57] The synergistic effect of two promoters across a wide range of elements has not been investigated for Ru-based and Pr_2O_3 supported ammonia synthesis catalysts. Hence, a doubly promoted catalyst was designed by fixing Cs as the primary promoter and 43 other elements were selected to be used as the secondary promoter. Here, Cs was selected as the primary promoter since it is found to be the best electron-donating alkali-metal promoter for supported Ru in ammonia synthesis catalysts.[27,76] The secondary promoters were selected so that they are non-radioactive and not costly (such as Pt, Rh, Pd), with their metal precursors being non-chlorine based (due to the poisoning effect of chlorine) and water-soluble, as shown in Fig. S1. The selected secondary promoters (M) were namely, Li, Na, K, Rb, Mg, Ca, Sr, Ba, Sc, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Al, Ga, Y, Zr, Ag, Cd, In, Sn, Hf, Re, Ir, Pb, Bi, La, Ce, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, and Lu. The weight loadings of the catalysts were fixed so that the final formulation can be written as Ru (1 wt%), Cs (2 wt%), M (2 wt%) / Pr_2O_3. Here, the Ru loading was fixed to 1 wt% regarding the high cost of Ru for large-scale applications.

The activity of these catalyst formulations was predicted by the ML model trained only on literature data at 673 K, 3 MPa, 36000 mL h^{-1} g_{cat}^{-1} at a H_2-to-N_2 ratio of 1:1. The predicted activities of these catalysts are illustrated in Fig. 5 (a) as a violin plot. According to the predictions, the formulation with Zn should have the highest activity in the search space, with Li, V, Cr, and Mg as the next best promoters in descending order. To validate the predictions, these catalysts were synthesized using wetness impregnation and tested in a reactor system. The measured synthesis rates of the top 5 predicted catalysts (Fig. 5(b)) illustrate that the predictions of the ETR ML model developed from literature data are inaccurate. The catalyst with the Zn promoter had no measurable activity even though it was predicted to have the highest activity out of all the promoters. Only the catalyst with Mg as the promoter was predicted with reasonable accuracy. This behavior of predicted

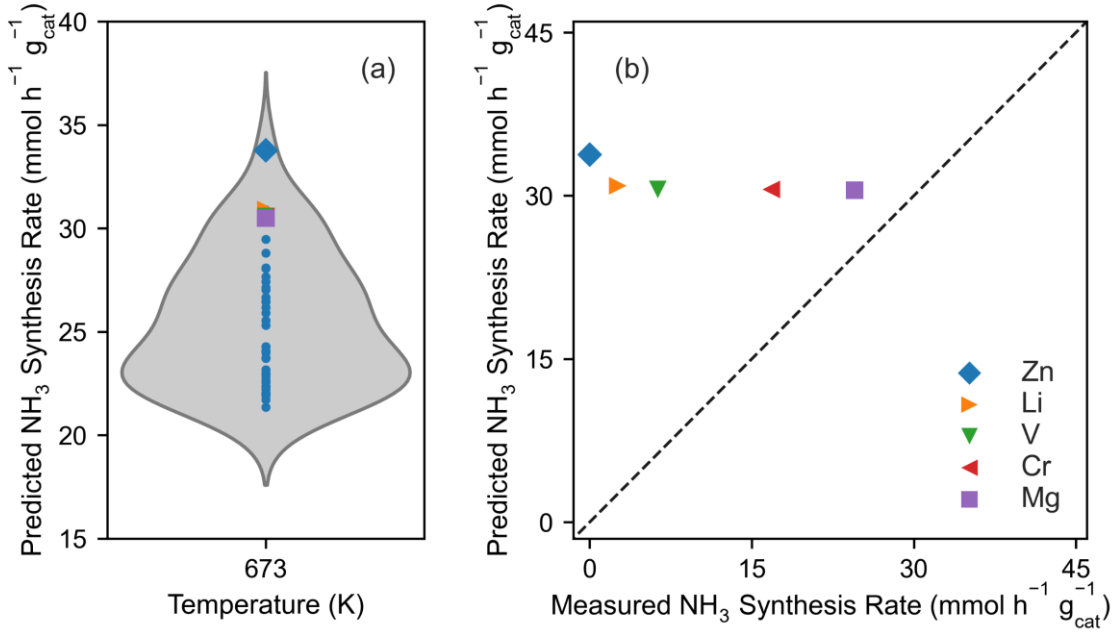

Fig. 5 (a) Model predictions based only on literature data and (b) experimental validation.

catalyst performance, not in agreement with experimental catalyst performance, has been reported before for ML models trained using only literature data.[12,14] The ML model trained by the literature data often does not have any training data in the search space that the model is attempting to predict. This model did not include any training data with promoted Ru supported on Pr_2O_3 due to the lack of data in the literature.

The reasons for the lack of the model's ability to predict novel catalysts can also be seen in the principal component analysis (PCA) of the data, as depicted in Fig. 6. In general, PCA is used to dimensionally reduce the feature space to an information space in dimensions known as 'principal components' so that the maximum amount of information is preserved. The direction with the most variance can be regarded as the 1st principal component. The direction that has the second most variation can be regarded as the 2nd principal component. From Fig. 6 (a), it can be seen that 1st and 2nd principal components contain 92.7 and 6.2% of the variation, respectively. In other words, 98.9% of the variation is contained in just the first two principal components alone. For descriptive purposes, the 1st and 2nd principal components can be used if 80% of the variance can be explained from both.[77] As a result, the literature data and the catalyst search space can be plotted using the first two principal components. Four major clusters of literature data points can be seen in the information space, and it is apparent that there is a largely unexplored area of catalysts in between. It has also been mentioned that the predictions are much more reliable in spaces where the data points are dense and less reliable in spaces where the data points are sparse.[78] Compared to the current literature data, the catalyst search space spreads in empty areas between clusters, implying that this search space has not been explored. Consequently, the predictions of the ETR ML model developed from only the literature data were unreliable and had low accuracy for new catalysts, as depicted in Fig. 5 (b).

3.3.2 Active Learning-assisted Prediction of Novel Catalysts

To find the best catalyst in the search space, more experimental data is required to integrate into the ML model trained on literature data incrementally using AL. Many AL methods in the literature rely on Bayesian methods, such as Bayesian optimization for material discovery.[5] Here, Bayesian statistics are used to calculate *a posteriori* joint probability distribution for the model parameters for the ML model named Gaussian process regressor (GPR) using the model features. The uncertainty estimates inherently given by the GPR model are used to evaluate a function named 'acquisition function' to predict the next experiment to run. The often-used acquisition function named 'Expected Improvement' is designed to efficiently explore the search space. However, Bayesian methods are not recommended for search spaces with more than 20 dimensions, since they often struggle in high-dimensional space in determining the joint probability distribution for the model parameters due to the curse of dimensionality.[79,80] Considering the ML model developed for our application, 53 out of the 160 features/dimensions are related to the variation of the promoter (Table S7), which makes the use of Bayesian methods not suitable for this application. Prediction of the next experiments using uncertainty estimates calculated for ensemble tree methods, such as RFR and ETR, have often outperformed Bayesian methods.[5,15]

The pathways MU and MEI can be followed based on the calculated model predictions. However, for the calculation of MLI to suggest the first experiment to run, y_{max} needs to be selected as the maximum rate in the literature data used to train the ML model. The maximum NH_3 synthesis rate in the literature dataset is 760 mmol h^{-1} g_{cat}^{-1} obtained using a 22 wt% Ru catalyst.[57] The initial setting of y_{max} to this value is not reasonable since the search space is only comprised of catalysts with 1 wt% Ru. Furthermore, since the literature

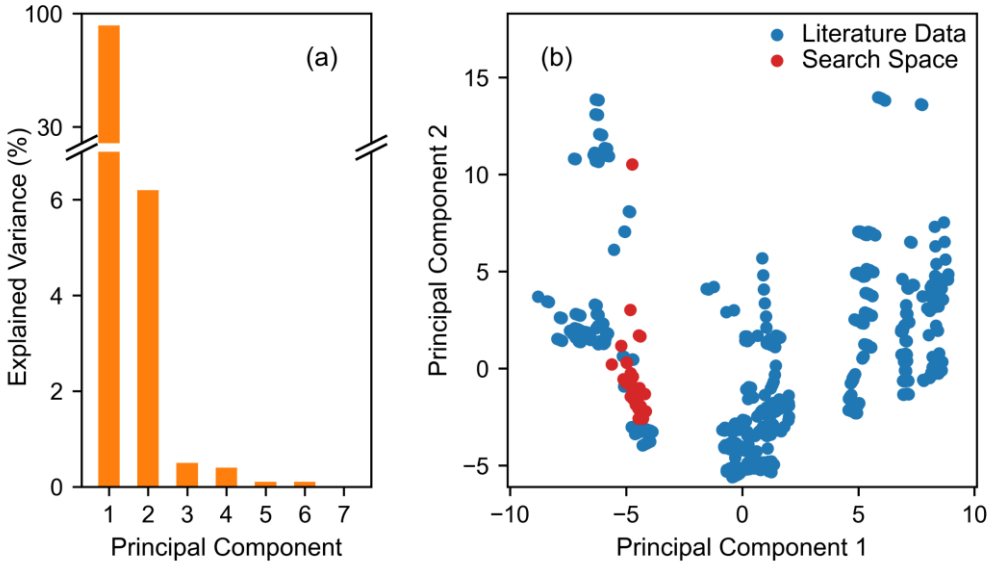

Fig. 6 (a) Percentage variation of the principal components and (b) PCA projection of the literature data and search space.

dataset spans over a wide range of reaction conditions, a catalyst with a lower NH_3 synthesis rate at a different reaction condition has the potential to exceed a catalyst with a higher NH_3 synthesis rate at a different condition. The goal of this search space is to optimize the promoters instead of the reaction conditions, and as an instance, y_{max} was chosen as the maximum NH_3 synthesis rate in the literature training data set for Ru loadings less than or equal to 1 wt%. This value was 24.7 mmol h^{-1} g_{cat}^{-1}, obtained using a 0.7 wt% Ru catalyst supported on graphitized carbon and promoted by Ba and Cs at 673 K, 6.3 MPa, 8.5 vol % initial NH_3 concentration, and over 487000 mL h^{-1} g_{cat}^{-1} space velocity [57]. The next experiments predicted in the MLI pathway will not change the y_{max} unless a higher value is experimentally found in the search space.

The AL simulations for three experimental pathways of MEI, MLI, and MU (with Ba-promoted catalyst data from the literature) are shown in Fig. 7(a) and (b) with the total experimental dataset shown in Table S8. The pathways are plotted as the best experimental NH_3 synthesis rate found at each iteration versus the number of experiments/iterations. The MEI and MLI pathways were more efficient than MU to reach the best catalyst in the search space. This catalyst contained Ba as a promoter with an activity of 48.8 mmol h^{-1} g_{cat}^{-1}, substantially exceeding the best NH_3 synthesis rate in the literature training dataset for Ru loadings less than or equal to 1 wt% (24.7 mmol h^{-1} g_{cat}^{-1}). The MU pathway started with higher experimental rates but could not improve significantly compared to the MLI and MEI pathways and was able to reach the best catalyst after 27 experiments. The MEI pathway outperformed the MLI, reaching the best catalyst in 10 experiments, compared to the 43 experiments that need to be done to confirm that Ba promoted catalyst is the best catalyst in the search space. Hence, AL using the MEI pathway cut down the number of experiments necessary to reach the best catalyst by nearly 75%, reducing the time and resources required. The MEI pathway being better than the MLI pathway (with Ba-containing data) could be due to the inclusion of initial training data from the literature with Ba-containing and doubly promoted (Ba, Cs) catalyst activity data on a different support. Since the model is already trained on this Ba-containing catalyst data, the exploitation pathway (MEI) works better than the exploration (MLI) pathway to reach Ba as the catalyst with the best secondary promoter.

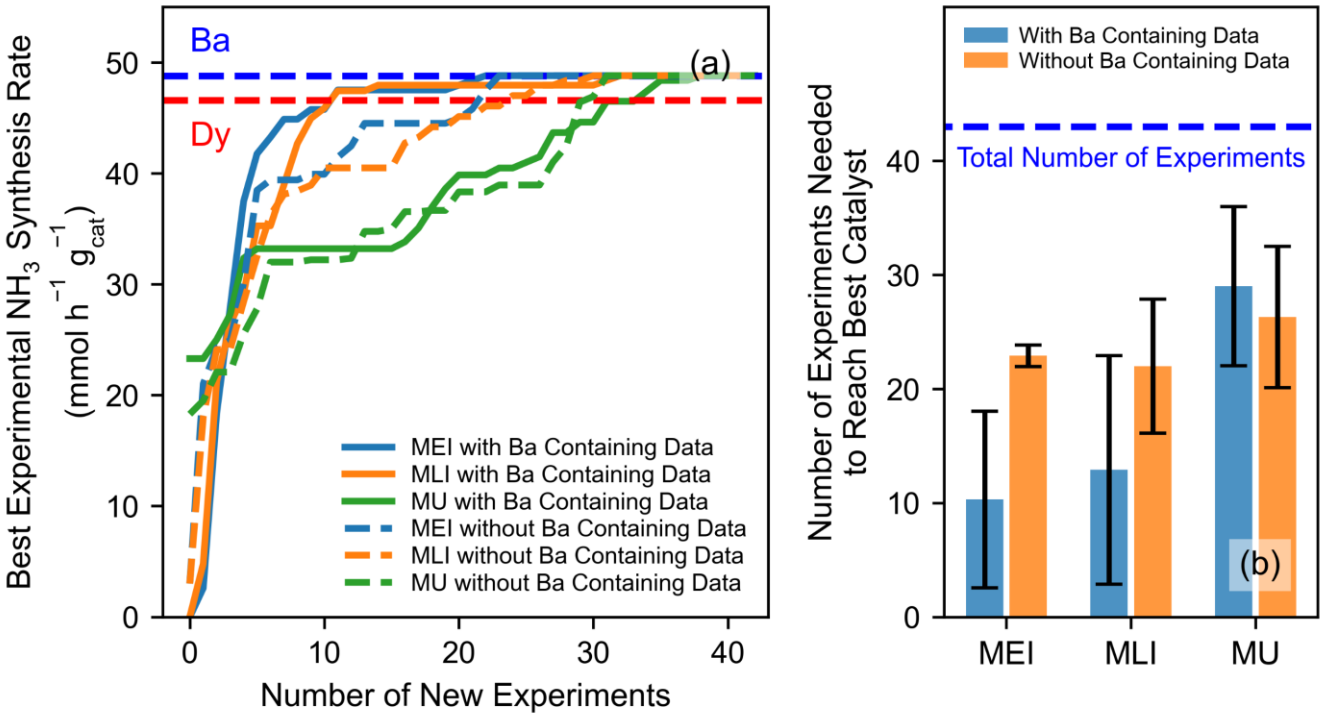

Fig. 7 (a) Active learning pathways by ETR model. Each pathway represents the average over 10 simulations and (b) the number of experiments needed to reach the best catalyst.

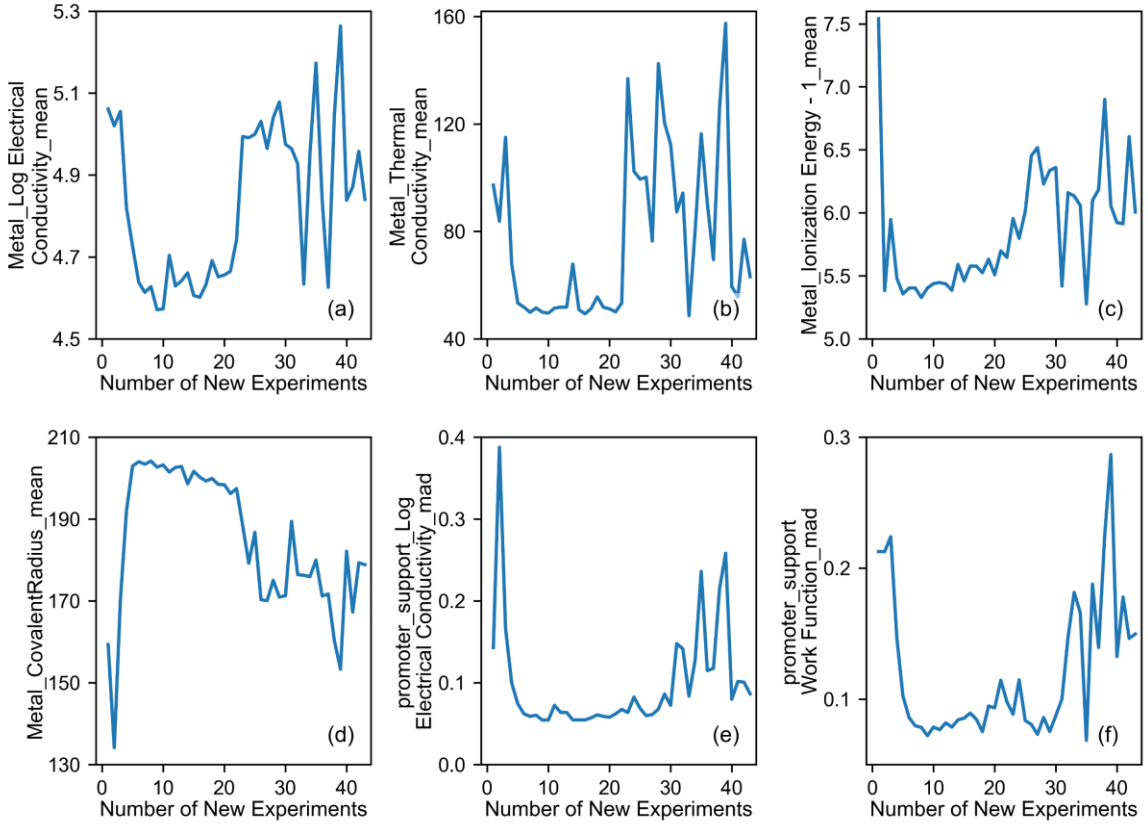

Fig. 8 The variation of selected feature values vs number of experiments during AL for MEI pathway with Ba containing catalyst data.

However, it can be argued that this efficiency came from the initial training data with Ba-promoted catalysts, as shown in Fig. 2(e). Hence, the capability of this framework was reevaluated by removing all the literature data where Ba was used as a promoter and supports that include Ba ($BaCeO_3$, $BaTiO_3$, and $BaZrO_3$). As shown in Fig. 7(a) and (b), the MEI and MLI pathways still outperformed the MU pathway, but it nearly doubled the experiments needed to reach the best catalyst by twice for MEI. Now, the MLI pathway outperforms the MEI pathway by requiring 22 experiments to reach the best catalyst and reducing the number of experiments by nearly 50% in the search space. Since the model does not have initial training data for the Ba-containing catalyst, the AL needs to efficiently explore the search space, resulting in the exploration pathway (MLI) working better than the exploitation pathway (MEI).

This framework led to the discovery of two novel catalyst formulations (Ru, Ba, Cs/Pr_2O_3; Ru, Dy, Cs/Pr_2O_3) with high activities. To the best of our knowledge, these combinations are not yet reported in the literature. The activity of these two catalysts outperformed many state-of-the-art thermocatalytic ammonia synthesis catalysts reported in the literature when compared on grams of Ru basis (mmol h^{-1} g_{Ru}^{-1}), as shown in Table S9. Further, the combination of Dy and Cs is not reported in the literature. However, the higher activity of Ba and Cs-promoted Ru-based catalysts has been reported in the literature for NH_3 synthesis on different types of graphitic carbon supports.[57,81–83] These studies used large amounts of Ba and Cs to reduce the effect of electron-withdrawing and strongly bound oxygen complexes on the carbon support surface.[57] Ba-promoted catalysts increased the thermal stability of the catalyst by diminishing the methanation of the carbon support.[83] The high activity resulting from the synergy of Ba and Cs was ascribed to different roles of the promoters. As explained earlier, the primary role of Cs is attributed to the electronic promotion of Ru to lower the barrier of dissociative adsorption of N_2. Ba is supposed to act as an electronic promoter [84,85] as well as a structural promoter to modify the number [86] and stability [87,88] of the surface sites of Ru. These effects depend on the support. A change of the support varies morphology, dispersion, and electronic structure of the Ru nanoparticles, and the dispersion of the promoters used.[67,88,89] These changes are combined with the change of promoters, which vary the electronic and structural promotional effects for Ru. It appears that the AL framework was able to 'learn' the changes of the complex interactions using the features and the initial set of experimental data integrated into the ML model.

To investigate what the model is 'actively learning' and based on what information the catalyst exploration is done as the experimental data are being added, the variation of feature values related to features of the promoter were plotted for the MEI pathway (with Ba containing data) vs the

number of experiments. The features with discernible patterns are shown in Fig. 8. The feature values for weighted mean or mean absolute deviation of metal-promoter interactions, such as electrical conductivity, thermal conductivity, 1st ionization energy, and covalent radius show notable variations with the number of experiments. These features are calculated as the weighted mean or mean absolute deviation for Ru, Cs, and secondary promoter by Eq. (2) and (3). Since only the secondary promoter is varied, changes in the features of metal-promoter interactions are driven mainly by the secondary promoter.

When the number of experiments is in the range of 5-15 (Fig. 7) the AL was finding the catalysts with the highest activity. According to Fig. 8(c), there is a notable decrease in first ionization energy, which suggests an electron donation from the secondary promoter leading to the higher activity of the catalysts. However, the electrical conductivity (Fig. 8(a)) decreases significantly at the same time. This implies that the secondary promoter should not interfere with the electrical conductivity of Ru since it has a higher electrical conductivity (0.137×10^6 $\Omega^{-1}cm^{-1}$) compared to Cs (0.0489×10^6 $\Omega^{-1}cm^{-1}$).[90] The metal covalent radius (Fig. 8(d)) appears to go through an optimum for promoters with higher ammonia synthesis activity. This could be due to the size of the secondary promoter atoms necessary to have a certain size for adsorption and surface reconstruction on the facets of Ru nanoparticles. In fact, promoters such as Li and Na are known to block stepped sites of Ru due to their smaller atomic radii.[91] Moreover, there is a significant drop in the bulk thermal conductivity (Fig. 8(b)) of the secondary promoter when the catalyst activity is the highest. Thermal conductivity at nanoscale are much lower compared to the bulk which is correlated to the sintering of the metal nanoparticles.[92] Based on these observations, it seems that the best-performing promoters Ba and Dy have both electronic and structural promotion for Ru/Pr_2O_3. The elemental features for promoter-support interaction, such as mean absolute deviations of electrical conductivity and work function (Fig. 8(e) and (f)) suggest that the deviation driven by the secondary promoter should be minimal. These trends suggest that the secondary promoter should not alter the electronic properties of Pr_2O_3 so that it could electronically interact with Ru for higher activity. It can be hypothesized that these minimal electronic interactions with the support are obtained by the secondary promoter being primarily associated with the Ru instead of the support.

4 Conclusions

Data for the activity of Ru-based catalysts for thermocatalytic NH_3 synthesis were mined from the published literature. After data cleaning, 936 data points were selected, and features for the data points were engineered to capture the effects of parameters in a supported catalyst used for ammonia synthesis. Selected features using the Boruta method included the electronic properties of the elements and interactions between elements, conforming to the domain knowledge of the heterogenous catalysts for ammonia synthesis. Out of all ensemble learning algorithms considered, the ETR model resulted in the highest accuracy ($R^2 = 0.86$).

A search space was selected based on the domain knowledge where the secondary promoter was varied in a doubly promoted Ru catalyst supported on Pr_2O_3. However, with experimental validation, the predictions of the ETR ML model developed only on the literature data proved inaccurate. Integration of more experimental data using AL pathways reduced the experimental data needed to reach the best catalyst by nearly 50% when Ba-containing catalyst data was excluded from the literature training data. The MLI pathway, which balances the exploration and exploitation of the search space outperformed the pathways MU and MEI, which are based on pure exploration and exploitation, respectively.

The discovered formulations Ru, Ba, Cs and Ru, Dy, Cs supported on Pr_2O_3 were not reported in the literature, and their activity exceeded that of most state-of-the-art catalysts reported in the literature. Hence, AL initially based on literature data has a great potential to discover novel catalysts, as we showed in this study. Investigation of feature variation of the experimental pathways given by AL suggested that catalysts with higher activities in the search space have secondary promoters that act both as electronic and structural promoters, and do not alter the electronic properties of the Pr_2O_3.

This work suggests that although the literature data for heterogeneous catalysts are high-dimensional, inconsistent, and filled with missing data, careful data selection and pre-processing can make it a great source for AL. Proper feature engineering and selection are still necessary for each application to deliver the most optimal results and gain insights since each catalytic system can have its own processes governing the catalyst's activity, selectivity, and stability. The authors contributed to this work hope that the framework presented in this paper would encourage researchers to utilize literature data to discover novel catalysts experimentally using AL to save time and resources in the future.

Author Contributions

Rasika Jayarathna: conceptualization, investigation, methodology, writing – original draft preparation. Thossaporn Onsree: conceptualization, investigation, writing – review & editing. Samuel Drummond: methodology, experimental data acquisition, writing – review & editing. Jennifer Naglic: methodology, experimental data acquisition, writing – review & editing. Jochen Lauterbach: conceptualization, writing - review & editing, supervision, funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the United States Department of Energy award DE-EE0009409.

References

1 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Comput. Mater.*, , DOI:10.1038/s41524-017-0056-5.

2 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, , DOI:10.1038/s41524-019-0221-0.

3 T. Lookman, P. V Balachandran, D. Xue and R. Yuan, *npj Comput. Mater.*, , DOI:10.1038/s41524-019-0153-8.

4 Z. Rosario, M. Rupp, Y. Kim, E. Antono and J. Ling, *J. Chem. Phys.*, 2020, **153**, 024112.

5 J. Ling, M. Hutchinson, E. Antono, S. Paradiso and B. Meredig, *Integr. Mater. Manuf. Innov.*, 2017, **6**, 207–217.

6 R. Yuan, Z. Liu, P. V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue and T. Lookman, *Adv. Mater.*, 2018, **30**, 1702884.

7 C. Kim, A. Chandrasekaran, A. Jha, R. Ramprasad and M. Science, *MRS Commun.*, 2019, **9**, 860–866.

8 G. Wang, S. Mine, D. Chen, Y. Jing, K. W. Ting, T. Yamaguchi, M. Takao, Z. Maeno, I. Takigawa, K. Matsushita, K.-I. Shimizu and T. Toyao, 2022, chemrxiv-2022-695rj.

9 M. Kim, Y. min Ha, W.-B. Jung, J. Yoon, E. Shin, I. Kim, W. Bo, Y. Kim and H. Jung, *Adv. Mater.*, 2022, **34**, 2108900.

10 Z. Zhou, *J. Mater. Chem. A*, 2021, **9**, 1295–1296.

11 J. A. Esterhuizen, B. R. Goldsmith and S. Linic, *Nat. Catal.*, 2022, **5**, 175–184.

12 S. Nishimura, J. Ohyama, T. Kinoshita, S. Dinh Le and K. Takahashi, *ChemCatChem*, 2020, **12**, 5888–5892.

13 S. Nishimura, S. D. Le, I. Miyazato, J. Fujima, T. Taniike, J. Ohyama and K. Takahashi, *Catal. Sci. Technol.*, 2022, **12**, 2766–2774.

14 M. Suvarna, T. P. Araújo and J. Pérez-Ramírez, *Appl. Catal. B Environ.*, 2022, **315**, 121530.

15 S. Mine, M. Takao, T. Yamaguchi, T. Toyao, Z. Maeno, S. M. A. Hakim Siddiki, S. Takakusagi, K. ichi Shimizu and I. Takigawa, *ChemCatChem*, 2021, **13**, 3636–3655.

16 S. Mine, Y. Jing, T. Mukaiyama, M. Takao, Z. Maeno, K. I. Shimizu, I. Takigawa and T. Toyao, *Chem. Lett.*, 2022, **51**, 269–273.

17 S. Vaclav, *How the World Really Works: A Scientist's Guide to Our Past, Present and Future*, Viking, London, 1st edn., 2022.

18 C. Smith, A. K. Hill and L. Torrente-Murciano, *Energy Environ. Sci.*, 2020, **13**, 331–344.

19 J. N. Galloway, A. R. Townsend, J. W. Erisman, M. Bekunda, Z. Cai, J. R. Freney, L. A. Martinelli, S. P. Seitzinger and M. A. Sutton, *Science (80-.).*, 2008, **320**, 889–892.

20 Institute for Sustainable Process Technology, *Report*, 2017, 51.

21 K. Sato and K. Nagaoka, *Chem. Lett.*, 2021, **50**, 687–696.

22 J. A. Faria, *Curr. Opin. Green Sustain. Chem.*, 2021, **29**, 100466.

23 J. W. Erisman, M. A. Sutton, J. Galloway, Z. Klimont and W. Winiwarter, *Nat. Geosci.*, 2008, **1**, 636–639.

24 C. Li, T. Wang and J. Gong, *Trans. Tianjin Univ.*, 2020, **26**, 67–91.

25 R. Li, W. Ma, Y. Liu, L. Zhang and Z. Zhou, *J. Mater. Chem. A*, 2023, **11**, 18626–18645.

26 R. Schlögl, *Angew. Chemie - Int. Ed.*, 2003, **42**, 2004–2008.

27 Z. Zhang, C. Karakaya, R. J. Kee, J. D. Way and C. A. Wolden, *ACS Sustain. Chem. Eng.*, 2019, **7**, 18038–18047.

28 Z. Zhang, J. D. Way and C. A. Wolden, *AIChE J.*, 2021, **67**, 1–11.

29 K. Masakoto, Y. Akira, T. Yoshioka and T. Oshinori, *AIChE J.*, 2010, **56**, 1204–1212.

30 S. Wu, Y. K. Peng, T. Y. Chen, J. Mo, A. Large, I. McPherson, H. L. Chou, I. Wilkinson, F. Venturini, D. Grinter, P. Ferrer Escorihuela, G. Held and S. C. E. Tsang, *ACS Catal.*, 2020, **10**, 5614–5622.

31 A. J. Medford, A. Vojvodic, J. S. Hummelshøj, J. Voss, F. Abild-Pedersen, F. Studt, T. Bligaard, A. Nilsson and J. K. Nørskov, *J. Catal.*, 2015, **328**, 36–42.

32 H. Fang, D. Liu, Y. Luo, Y. Zhou, S. Liang, X. Wang, B. Lin and L. Jiang, *ACS Catal.*, 2022, **12**, 3938–3954.

33 V. S. Marakatti and E. M. Gaigneaux, *ChemCatChem*, 2020, **12**, 5838–5857.

34 C. Fernández, C. Sassoye, D. P. Debecker, C. Sanchez and P. Ruiz, *Appl. Catal. A Gen.*, 2014, **474**, 194–202.

35 S. Majid, K. Shariati, Y. Suk and J. Lauterbach, *Chem. Eng. J.*, 2023, **467**, 143533.

36 A. Rohatgi, WebPlotDigitizer (Version 4.6), 2022.

37 R. P. Ribeiro and N. Moniz, *Mach. Learn.*, 2020, **109**, 1803–1835.

38 S. Sen, K. P. Singh and P. Chakraborty, *New Astron.*, 2023, **99**, 101959.

39 B. MacQueen, R. Jayarathna and J. Lauterbach, *Curr. Opin. Chem. Eng.*, 2022, **36**, 100781.

40 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.

41 P. Geurts, D. Ernst and L. Wehenkel, *Mach. Learn.*, 2006, **63**, 3–42.

42 T. Chen and C. Guestrin, in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*, 2016, pp. 785–794.

43 I. Rossetti, N. Pernicone, F. Ferrero and L. Forni, *Ind. Eng. Chem. Res.*, 2006, **45**, 4150–4155.

44 I. Chorkendoff and J. W. Niemantsverdriet, *Concepts of Modern Catalysis and Kinetics*, Wiley, Weinheim, 2017, vol. 56.

45 L. J. Gillespie and J. A. Beattie, *Phys. Rev.*, 1930, **36**, 743–753.

46 D. C. Dyson and J. M. Simon, *Ind. Eng. Chem. Fundam.*, 1968, **7**, 605–610.

47 X. Han, C. Zhao, H. Li, S. Liu, Y. Han, Z. Zhang and J. Ren, *Catal. Sci. Technol.*, 2017, **7**, 6042–6049.

48 T. Williams, K. McCullough and J. A. Lauterbach, *Chem. Mater.*, 2020, **32**, 157–165.

49 C. Karakaya, J. Huang, C. Cadigan, A. Welch, J. Kintner, J. Beach, H. Zhu, R. O'Hayre and R. J. Kee, *Chem. Eng. Sci.*, 2022, **247**, 116902.

50 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.

51 L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster and A. Jain, *Comput. Mater. Sci.*, 2018, **152**, 60–69.

52 M. T. Dylla, A. Dunn, S. Anand, A. Jain and G. J. Snyder, *Research*, 2020, **2020**, 1–8.

53 D. Szmigiel, W. Raróg-Pilecka, E. Miśkiewicz, M. Gliński, M. Kielak, Z. Kaszkur and Z. Kowalczyk, *Appl. Catal. A Gen.*, 2004, **273**, 105–112.

54 Z. You, K. Inazu, K. ichi Aika and T. Baba, *J. Catal.*, 2007, **251**, 321–331.

55 Z. H. Zhong and K. I. Aika, *J. Catal.*, 1998, **173**, 535–539.

56 S. M. Yunusov, E. S. Kalyuzhnaya, B. L. Moroz, S. N. Agafonova, V. A. Likholobov and V. B. Shur, *J. Mol. Catal. A Chem.*, 2001, **165**, 141–147.

57 W. Raróg-Pilecka, E. Miśkiewicz, D. Szmigiel and Z. Kowalczyk, *J. Catal.*, 2005, **231**, 11–19.

58 X. Wang, A. Beck, J. A. van Bokhoven and D. Palagin, *J. Mater. Chem. A*, 2021, **9**, 4044–4054.

59 C. T. Campbell and Z. Mao, *ACS Catal.*, 2017, **7**, 8460–8466.

60 S. L. Hemmingson and C. T. Campbell, *ACS Nano*, 2017, **11**, 1196–1203.

61 M. Kuhn and K. Johnson, *Applied predictive modeling*, Springer, New York, 5th edn., 2013.

62 J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang and H. Liu, *Feature selection: A data perspective*, 2017, vol. 50.

63 M. B. Kursa and W. R. Rudnicki, *J. Stat. Softw.*, 2010, **36**, 1–13.

64 B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta and L. Ward, *Mol. Syst. Des. Eng.*, 2018, **3**, 819–825.

65 F. Hutter, L. Xu, H. H. Hoos and K. Leyton-Brown, *Artif. Intell.*, 2014, **206**, 79–111.

66 S. Murata and K. I. Aika, *Appl. Catal. A, Gen.*, 1992, **82**, 1–12.

67 K. ichi Aika, T. Takano and S. Murata, *J. Catal.*, 1992, **136**, 126–140.

68 R. Javaid and T. Nanba, *ChemistrySelect*, 2020, **5**, 4312–4315.

69 Y. Manaka, Y. Nagata, K. Kobayashi, D. Kobayashi and T. Nanba, *Dalt. Trans.*, 2020, **49**, 17143–17146.

70 S. ichiro Miyahara, K. Sato, Y. Kawano, K. Imamura, Y. Ogura, K. Tsujimaru and K. Nagaoka, *Catal. Today*, 2021, **376**, 36–40.

71 W. Li, S. Wang and J. Li, *Chem. - An Asian J.*, 2019, **14**, 2815–2821.

72 Z. Wang, J. Lin, R. Wang and K. Wei, *Catal. Commun.*, 2013, **32**, 11–14.

73 Y. V. Larichev, B. L. Moroz, V. I. Zaikovskii, S. M. Yunusov, E. S. Kalyuzhnaya, V. B. Shur and V. I. Bukhtiyarov, *J. Phys. Chem. C*, 2007, **111**, 9427–9436.

74 K. Sato, K. Imamura, Y. Kawano, S. ichiro Miyahara, T. Yamamoto, S. Matsumura and K. Nagaoka, *Chem. Sci.*, 2016, **8**, 674–679.

75 K. Imamura, S. ichiro Miyahara, Y. Kawano, K. Sato, Y. Nakasaka and K. Nagaoka, *J. Taiwan Inst. Chem. Eng.*, 2019, **105**, 50–56.

76 K. ichi Aika, H. Hori and A. Ozaki, *J. Catal.*, 1972, **27**, 424–431.

77 Interpret the key results for Principal Components Analysis, https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/interpret-the-results/key-results/.

78 A. Smith, A. Keane, J. A. Dumesic, G. W. Huber and V. M. Zavala, *Appl. Catal. B Environ.*, 2020, **263**, 118257.

79 P. I. Frazier, *A Tutorial on Bayesian Optimization*, 2018.

80 S. Shan and G. G. Wang, *Struct. Multidiscip. Optim.*, 2010, **41**, 219–241.

81 L. Forni, D. Molinari, I. Rossetti and N. Pernicone, *Appl. Catal. A Gen.*, 1999, **185**, 269–275.

82 Z. Kowalczyk, M. Krukowski, W. Raróg-Pilecka, D. Szmigiel and J. Zielinski, *Appl. Catal. A Gen.*, 2003, **248**, 67–73.

83 K. N. Iost, V. L. Temerev, N. S. Smirnova, D. A. Shlyapin, V. A. Borisov, I. V. Muromtsev, M. V. Trenikhin, T. V. Kireeva, A. V. Shilova and P. G. Tsyrul'nikov, *Russ. J. Appl. Chem.*, 2017, **90**, 887–894.

84 I. Rossetti, N. Pernicone and L. Forni, *Appl. Catal. A Gen.*, 2001, **208**, 271–278.

85 I. Rossetti, F. Mangiarini and L. Forni, *Appl. Catal. A Gen.*, 2007, **323**, 219–225.

86 T. W. Hansen, P. L. Hansen, S. Dahl and C. J. H. Jacobsen, *Catal. Letters*, 2002, **84**, 7–12.

87 D. Szmigiel, H. Bielawa, M. Kurtz, O. Hinrichsen, M. Muhler, W. Raróg, S. Jodzis, Z. Kowalczyk, L. Znak and J. Zieliński, *J. Catal.*, 2002, **205**, 205–212.

88 H. Bielawa, O. Hinrichsen, A. Birkner and M. Muhler, *Angew. Chemie - Int. Ed.*, 2001, **40**, 1061–1063.

89 C. J. H. Jacobsen, *Chem. Commun.*, 2000, 1057–1058.

90 Angstrom Sciences. Elements electrical conductivity reference table.

91 J. Zheng, F. Liao, S. Wu, G. Jones, T. Y. Chen, J. Fellowes, T. Sudmeier, I. J. McPherson, I. Wilkinson and S. C. E. Tsang, *Angew. Chemie - Int. Ed.*, 2019, **58**, 17335–17341.

92 J. Jeong and Y. Wang, *Addit. Manuf. Lett.*, 2023, **4**, 100114.