JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

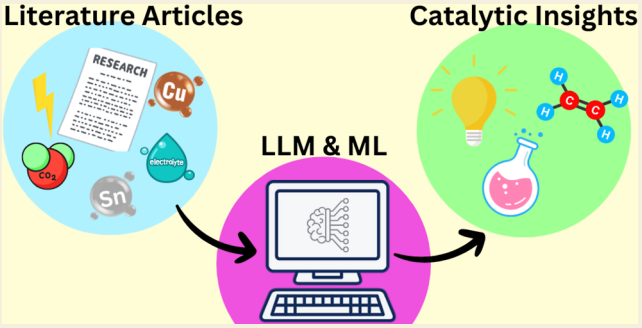
pubs.acs.org/jacsau
Article

Accelerating Catalysis Understanding via Large Language Model Data Extraction and Shallow Machine Learning Techniques

Brianna R. Farris and Kevin C. Leonard*

Cite This: JACS Au 2025, 5, 5578−5589

ABSTRACT: Catalysis is inherently complex. The lack of precise knowledge available to experimental researchers about the microenvironment, catalytic sites, mechanisms, and changes that occur under reaction conditions has hindered the effectiveness of deep-learning artificial intelligence algorithms to predict catalyst behavior under reaction conditions. Given the type and quality of data available in the scientific literature, there are still open questions on how machine learning can be used by experimentalists working in the field of catalysis to accelerate catalyst design. Here, we present a framework that leverages large language models to extract textual data from known and trusted sources to automatically generate large, but relatively low-fidelity, experimental catalysis data sets across many research groups. We also show that instead of using deep-learning models, which require high-quality data, shallow learning models with posthoc interpretability can extract valuable information about experimental catalytic systems from these low-fidelity data sets. The innovation of this work lies not in the model development but in the prompt engineering, data encoding, and question architectures employed to extract meaningful information. We applied this framework to two different model reactions: the electrocatalytic reduction of carbon dioxide and the electrocatalytic oxygen reduction reaction. We showcase that this framework has the ability to uncover known and established facts within the catalysis community, such as the catalytic properties of Cu, as well as novel insights, including the critical role of voltages above a certain threshold in producing multicarbon products from CO_2. We anticipate that this proposed framework will serve as an entryway for experimental catalytic researchers to utilize machine learning to rapidly process literature, generate novel hypotheses, and design experiments to accelerate catalyst development.



KEYWORDS: *machine learning, large language models, shallow learning, interpretable machine learning, electrocatalysis, CO_2 reduction*

1. INTRODUCTION

There is a growing need within the catalysis community to gain a deeper understanding of how machine learning can be utilized to guide the design of novel catalysts and catalytic systems.[1−5] Currently, there are some areas where generalization of machine learning models has proven to be very effective in catalysis. For example, machine learning has demonstrated success in predicting the synthesis, structure, or other physical properties of materials employing a diverse range of machine learning techniques.[6−18] However, predicting how a catalyst will perform (e.g., the rate of reaction, selectivity, energy efficiency, etc.) under a given set of operating conditions (e.g., temperature, pressure, applied voltage, etc.) is much more difficult.[2,19−22] The challenge is that the complexity of catalysis is so extensive that it is nearly impossible to measure, control, or describe all the factors that influence a catalyst's performance. For example, there are certain properties that catalytic research can measure and control, such as the elemental composition of the catalysts, which influences the adsorption of key intermediates, the surface area of the catalysts, which affects the turnover frequency, and the porosity and tortuosity of a catalyst, which affects the mass transfer of reactants and products to and from the catalyst's surface.[23−27] In contrast, variables such as impurities in the reactant feed streams or catalytic synthesis, defects in the catalytic surface, or structural changes that occur during the reaction's operation are either too challenging to measure and control or simply overlooked and not reported in the literature.[28,29] The aforementioned challenges render it virtually impossible to construct the high-fidelity data sets of experimental catalysis obtained from diverse research groups that are suitable for deep-learning algorithms. Moreover, artificial intelligence and machine learning are developing so

Received: August 22, 2025
Revised: September 26, 2025
Accepted: September 29, 2025
Published: October 21, 2025

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

rapidly that it is unclear exactly how this lack of high-fidelity data will impact the use of large language models and artificial neural networks.[3,30] Some researchers have started to experiment in using large language models, some for data extraction and some for creating chat bots within scientific fields.[31−34] Thus, the open question is how experimental researchers should utilize machine learning to accelerate catalyst design.

To bridge this gap, we propose a novel artificial intelligence framework that starts with questions pertaining to a catalytic system. This framework automatically extracts information from predetermined, reliable sources selected by the user and uncovers insights related to the specific questions. The framework consists of three separate components. The first component automatically obtains large data sets by applying large language models as a textual data extraction tool on existing literature. We show that Generative Pretraining Transformer (GPT) models offer a high ease of use and sufficient extraction capabilities. We also show that while the textual data extracted from LLMs alone may not capture all the intricate parameters required for modeling catalytic systems, it can still provide valuable, albeit low-fidelity, data sets that can provide insights and trends. In addition, utilizing LLMs trained on only user-selected trusted sources offers an alternative to other opaque chatbot-style large models.

The second component of our framework is the encoding of data, ensuring its compatibility with machine-learning algorithms for processing. Encoding is frequently overlooked in the machine-learning process. In this study, we demonstrate that appropriately selecting either label encoding or a one-hot encoding style can reveal distinct types of information pertaining to catalytic processes.

The third component of our framework is the type of machine-learning model that we utilized to be of the most utility to experimental catalytic researchers. In this work, we focus on shallow learning models such as decision trees and random forest rather than deep learning approaches, as they offer greater interpretability between the extracted parameters. Unlike neural networks, which function as a black box, shallow models identify feature importance and relationships within the data. The transparency of relationships provided by shallow learning models is crucial in catalysis research, where understanding the underlying factors driving selectivity can be as important as making accurate predictions. In addition, we show that these shallow-learning models work well on the types of low-fidelity data sets that can be automatically extracted from GPT text models.

As a first model reaction, we investigated what information this artificial intelligence framework could glean from the electrocatalytic CO_2 reduction reaction (CO_2RR). The CO_2RR has several properties that make it an ideal test case for this framework. First and foremost, the CO_2RR is a highly researched field, with many publications, due to the demand for novel technologies to convert CO_2 into value-added products, which is crucial for mitigating the impacts of anthropogenic climate change. The majority of these studies focus on catalyst development and improving the selectivity of CO_2 conversion, providing extensive catalytic data that are key to understanding the underlying reaction mechanisms. However, the sheer volume of published research makes it nearly impossible for any individual researcher to stay fully informed. Thus, it leaves room for crucial catalytic insight gaps for researchers in the field. Moreover, despite the large amount of publications and the many review articles, there are currently no repositories with large enough databases within this field to perform insightful machine learning.

In this study, we obtained the key parameters that control (1) whether the electron transfer process was low (2 e^- transfers) or high (more than 2 e^- transfers), (2) whether the reaction produced multicarbon or single-carbon products, (3) whether CO or formate/formic acid was the favored product, and (4) whether CO or ethylene was the favored product. We then demonstrated that this AI framework independently determined some known parameters about the CO_2RR (such that Cu is a catalyst that produces ethylene, and In and Bi produce formic acid), and some more complex interactions such as voltages more negative than −0.43 V vs RHE are critical for producing multicarbon products. We anticipate that this proposed framework will serve as the entryway for experimental catalytic researchers to utilize machine learning to rapidly process the literature and generate novel hypotheses. We also anticipate that this framework, which incorporates posthoc interpretable models, will provide catalytic researchers with additional information that will facilitate extrapolation and the development of novel catalytic systems.

As a second model reaction, and to demonstrate reproducibility of this extraction framework, we took another reaction, the oxygen reduction reaction (ORR), and performed the same procedure of feature extraction and ML predictions. For this demonstration, we took a smaller data set spanning 5 years of literature from 2020 to 2024. Using this framework, we identified the most influential features on the model's prediction to be Pt, nanoparticle structure, Fe, KOH as the electrolyte, and Pd. If the reaction conditions contained any of those features, the model was more likely to predict that the half-wave potential was higher than the median.

2. METHODS

2.1. Data Collection, Processing, and Encoding

A collection of peer-reviewed publications was obtained using Google Scholar (https://scholar.google.com) to serve as our starting library. These articles were from a 10-year period of 2013−2023 relating to electrochemical CO_2 reduction.

To extract textual data from the manuscript library, a commercial Generative Pretraining Transformer (GPT) API: ChatPDF API (https://www.chatpdf.com/) was used. Given a prompt, the API collected low-fidelity textual data on electrochemical CO_2 catalyst and reaction parameters from the literature articles. The prompt was created by repetitively adjusting the prompt wording on a few known papers to extract data and compare the output to the known features. It was then expanded to larger sets of data, and if a feature requested for extraction was consistently not being extracted by the LLM, it was removed from the prompt, and the resulting papers were checked to see if they contained the feature. The prompt given to API was "can you extract the relevant electrochemical CO_2 catalyst data with columns including: metal, voltage, units of voltage (mv or V), reference electrode, electrolyte, product, catalyst surface area, a yes or no on if it is a gas diffusion electrode, and if the catalyst structure one of the following: nanoparticle, polycrystalline, foil, thin film, porous, oxide derived, or amorphous in table format?". The response was then cleaned and put into table formatting for postprocessing (Supporting Information Figure S1). A manual check of 20% of each year's articles was performed to ensure they were being extracted properly, and a qualitative assessment showed that the automated text extraction was very similar to human extraction.

The post-processing of the data included evaluating, cleaning, and handling missing data points. For values within the other categories that the GPT API could not find, the data point was assigned as "not

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

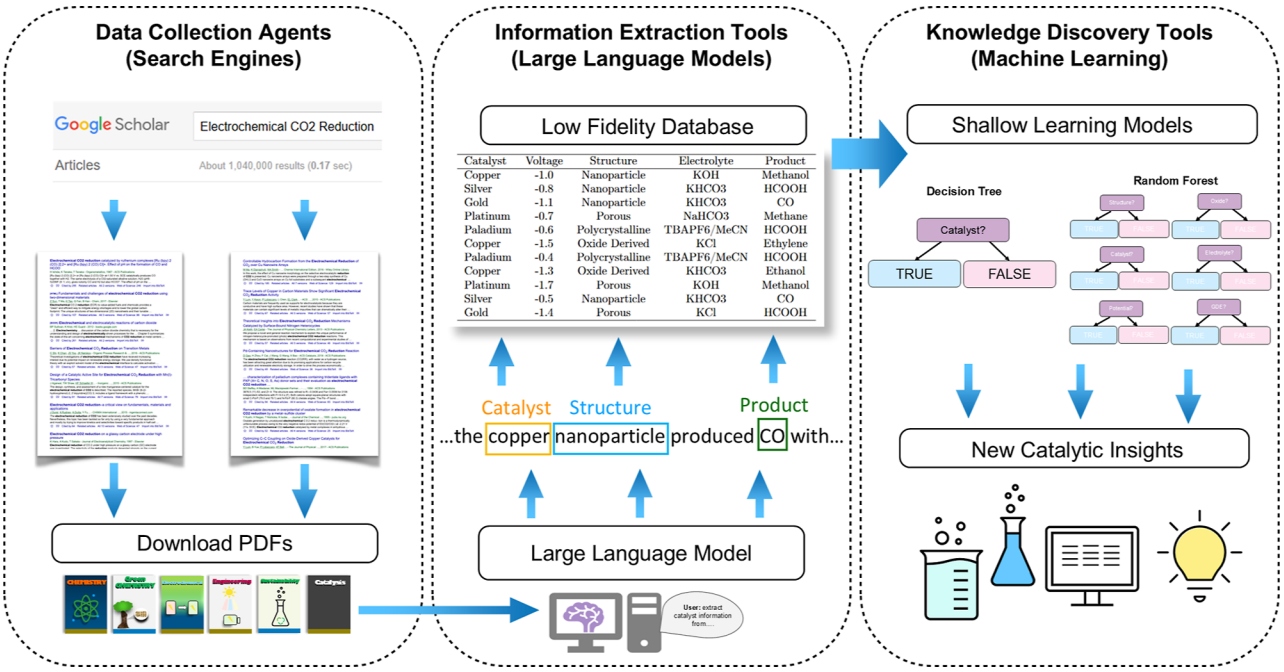


Figure 1. Flow diagram for the integrated LLM-shallow learning framework. The framework starts with a topic search query to produce a collection of PDF-formatted manuscripts. An LLM is then used on the manuscript collection to automatically produce a low-fidelity database of reaction conditions and catalytic properties. From this database, shallow-learning models produce catalytic insights without significant model tuning.

specified". If a data point is being predicted a specific way and has a "not specified," it shows that there is enough data in the data set that the model recognized the pattern in the other features and was able to predict the corresponding question. Some of the columns needed cleaning; for example, the voltage was collected with a unit of V or mV. The unit was separated into its own column and used to convert all voltages in mV to V. The voltage was then standardized using the reference electrode column of the table, so that every point references the RHE. If the pH was not reported, we used the standard conditions for the electrolyte given. Additionally, for all the columns, minor variations of spelling and names were standardized.

Once the database was in a clean and usable format, a mixed encoding strategy was implemented where the numerical values were left continuous, the catalyst metals were one-hot encoded, and the other categorical columns were label encoded (Supporting Information Table S1 and S2). For the label-encoded features, they were strategically given numerical values based on general size and likeness to each other. For example, the ionic radius of Na is smaller than that of K; therefore, the sodium-based electrolytes were grouped together and given smaller values than the potassium-based electrolytes.

2.2. Machine Learning

The processed and encoded database was then used as the input for the binary classification machine learning models. For each of the proposed questions, the database only utilized the data points relevant to the specific prediction (i.e., for CO vs formate only data points with CO or formate from the database were used). Therefore, each of the predictions used a different number of data points from the database.

The models utilized were scikit-learn version 1.3.2 Python package for the decision tree and random forest classifiers.[35] The training of the models used a training/test/validation split of 0.8/0.1/0.1 for each of the proposed questions. Evaluation of the models utilized validation versus training data to determine the optimal layers of the decision trees and the random forest. For the evaluation of the results for the random forest model, two methods were used. First, the built-in feature importance from scikit-learn with label-encoded features. Second SHAP (SHapley Additive exPlanations) force plots were used with one-hot encoded features for further catalytic understanding.[36]

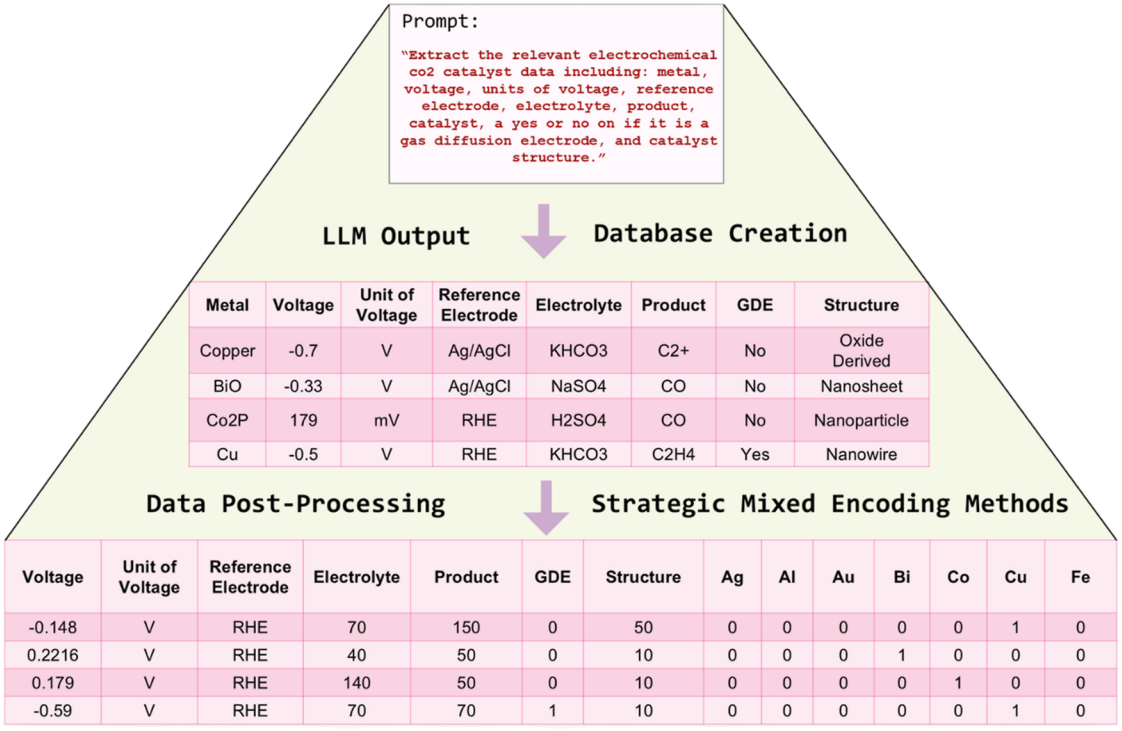
3. RESULTS AND DISCUSSION

The flow diagram for the integrated LLM-shallow learning framework, which starts with a topic search query and subsequently generates valuable catalytic insights, is depicted in Figure 1. The initial step involves utilizing a commercial scientific search engine to generate a list of digital object identifiers of peer-reviewed publications pertaining to the topic query: "electrochemical CO_2 reduction." By restricting the search to a 10 year period, spanning from 2013 to 2023, 2302 peer-reviewed PDF-formatted manuscripts on the topic of electrochemical CO_2 reduction were identified. To ensure the relevance of the data, a thorough review of the collected PDFs was conducted to remove articles unrelated to CO_2 reduction, verify that all of the manuscripts contained experimental data, and exclude review papers. Review papers were excluded as they may potentially introduce multiple repetitive data points, potentially biasing machine learning models.

Following the procurement of the PDF-formatted manuscript collection, a commercial Generative Pretrained Transformer Large Language Model (GPT LLM) was employed to read and automatically extract textual data (see Methods and Supporting Information for more details).

To extract textual data from the scientific literature, the GPT LLM model must be capable of handling the inconsistency in the terminology and completeness of the reported data, which can vary significantly across different authors and journals. These inconsistencies necessitate the implementation of effective prompt engineering as a critical component of LLMs for data extraction purposes. Engineered prompts guide the model to retrieve the most relevant information, despite variations within the publications. Here, we engineered a single prompt that would extract specific textual information regarding the nature of the catalyst, the electrochemical reaction conditions [specifically, the catalyst material, catalyst structure, applied voltage, reference electrode, electrolyte, and

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589



Metal	Voltage	Unit of Voltage	Reference Electrode	Electrolyte	Product	GDE	Structure
Copper	-0.7	V	Ag/AgCl	KHCO3	C2+	No	Oxide Derived
BiO	-0.33	V	Ag/AgCl	NaSO4	CO	No	Nanosheet
Co2P	179	mV	RHE	H2SO4	CO	No	Nanoparticle
Cu	-0.5	V	RHE	KHCO3	C2H4	Yes	Nanowire

Voltage	Unit of Voltage	Reference Electrode	Electrolyte	Product	GDE	Structure	Ag	Al	Au	Bi	Co	Cu	Fe
-0.148	V	RHE	70	150	0	50	0	0	0	0	0	1	0
0.2216	V	RHE	40	50	0	10	0	0	0	1	0	0	0
0.179	V	RHE	140	50	0	10	0	0	0	0	1	0	0
-0.59	V	RHE	70	70	1	10	0	0	0	0	0	1	0

Figure 2. Schematic showing the workflow starting with the search query to obtain a library of .pdf documents (left), followed by a GPT LLM prompt applied to all manuscripts in the collection to an unencoded database (middle). This database is then encoded so that it is suitable for interpretable shallow-learning to obtain new catalytic insights (right).

whether the system utilized a gas-diffusion electrode (GDE) setup], and the major product from each manuscript.

As illustrated in Figure 2, the prompt that yielded the most effective results for our electrocatalytic data extraction was: "extract the relevant electrochemical CO_2 catalyst data with columns including: metal, voltage, units of voltage (mv or V), reference electrode, electrolyte, product, catalyst surface area, a yes or no on if it is a gas diffusion electrode, and if the catalyst structure one of the following: nanoparticle, polycrystalline, foil, thin film, porous, oxide derived, or amorphous in table format with no other wording?" Examining this prompt shows several key aspects of prompt engineering specifically tailored for catalytic data extraction. First, the user must clearly define the specific features to be extracted and verify if different units are reported (e.g., in the voltage feature). Second, binary categorical features (e.g., yes or no for GDE systems) should be explicitly stated to provide a binary output. Third, complex multiclass categorical features (e.g., the catalyst structure) can be prebinned during prompt engineering. This step is crucial for subsequent machine learning phases. Having excessive classes for any one feature can hinder the learning of shallow-learning models. Therefore, prebinning the features into user-specified classes using the GPT LLM facilitates the extraction of the most information from shallow-learning models.

Running this prompt on all manuscripts in the collection produced an unencoded database as presented in Figure 2. Following the creation of the GPT LLM database, post-processing was performed to encode and standardize the data set. First, the database was analyzed to ensure all categories had enough data to be effectively used in the subsequent machine-learning studies. After examining the database, we observed that some information, such as the surface area of the catalyst and the faradaic efficiency of the major product, was not consistently reported. Consequently, this information was deemed insufficient for subsequent machine-learning studies and was not included in the final database.

Second, there was the further binning of certain categorical values, such as structure and electrolyte, into more concise categories, which was done with the GPT-prompt engineering binning. Refining the bins improves model performance by preventing overfitting to minor variations and allowing the algorithms to learn generalizable patterns rather than memorizing specific cases. For instance, the category "structure" could include items such as nanoparticle, nanowire, nanotube, and nanosized. Binning these into an overarching class of "nanoparticle" ensures that the model captures underlying trends in catalytic behavior rather than being influenced by slight naming variations across publications. Another example of binning was for the products of the reaction. If the literature article had multiple products listed under the same conditions, the LLM would combine multiple products into the data point. To address this issue, we categorized these examples into three categories: (1) single carbon products, if all the products listed were single carbons, (2) multicarbon products, if all the products listed were multicarbon, and (3) mixed carbon products, if the data point contained both single and multicarbon products. Note, those data points that contained multiple products were only used for predicting the carbon−carbon coupling ability of the catalysts.

The third step is encoding the data set. Typically, categorical data are either labeled encoded (by assigning each class a number) or one-hot encoded, where each class in a category is separated into its own feature. However, we observed that a strategic mixed encoding method, where a combination of label and one-hot encodings is used for different features

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

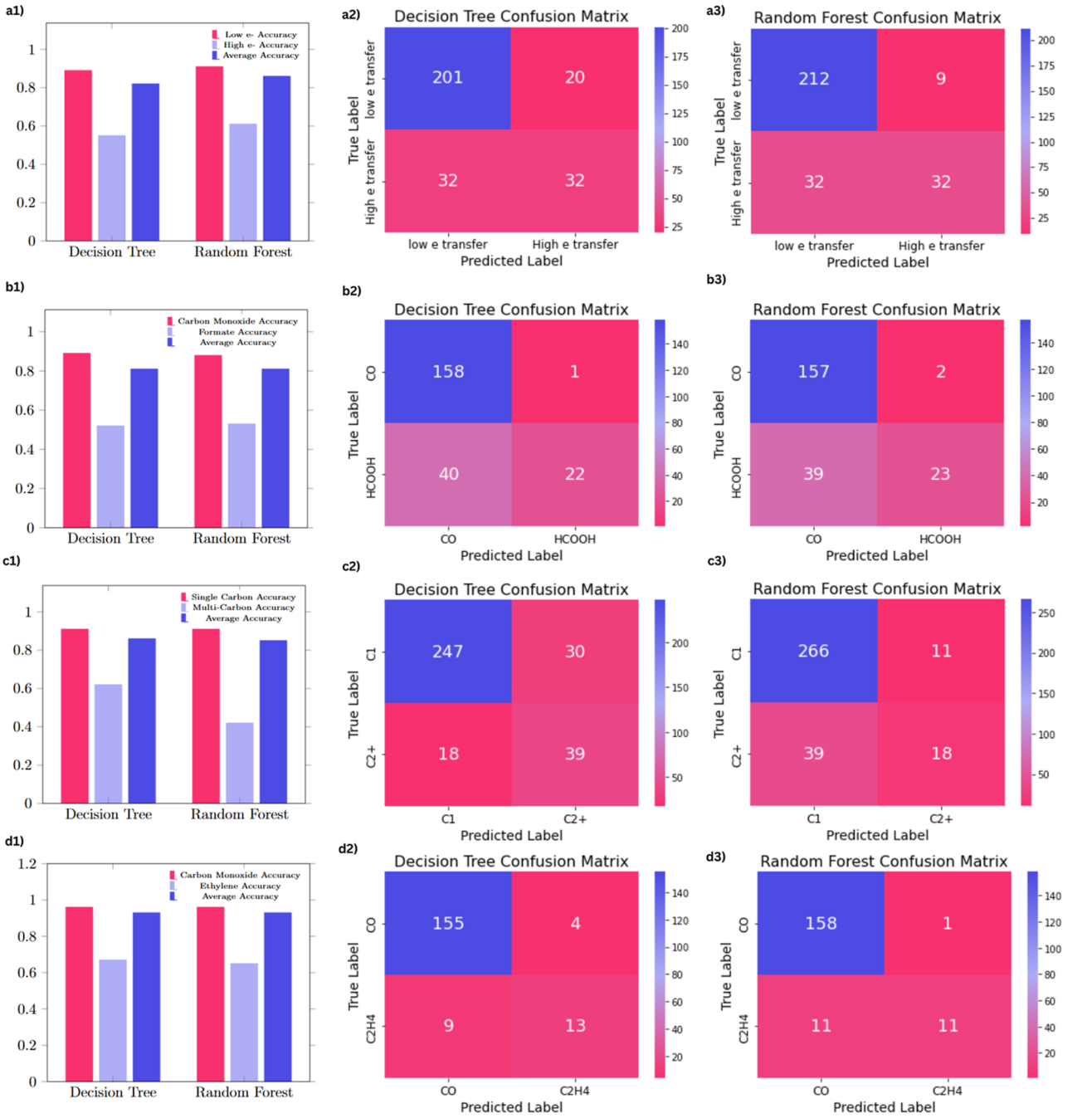


Figure 3. Model accuracies (1) and confusion matrices from the results of the decision tree (2) and random forest (3) models for each of the four questions asked to the database. (a) Whether the electron transfer process was low (2 e^- transfers) or high (more than 2 e^- transfers). (b) Whether CO or formate/formic acid was the major product, (c) whether the reaction produced multicarbon or single-carbon products, (d) whether CO or ethylene was the major product.

depending on the nature of the categorical data, can extract additional knowledge from shallow-learning models. Label encoding was used for the features of electrolyte, structure, and product because the numerical values help preserve the meaningful order relationships. For instance, the electrolyte was encoded to group similar cations together and separate organic versus aqueous electrolytes. In contrast, one-hot encoding was used for the catalyst metals to ensure that relationships between catalyst compositions were captured by the model. The objective of the mixed encoding strategy was to maintain the dimensionality of the data set relatively low by utilizing only one-hot encoding for one feature while preserving the numerical relationships of the label-encoded features. This produced a strategically encoded, low fidelity database suitable for shallow machine learning models to answer questions on the intricacies of the CO_2RR (Figure 2).

Once the database is created, we employed shallow learning models with native or posthoc interpretability to gain insights from the machine learning models. For instance, decision trees are inherently interpretable because they make predictions based on a series of decision points where each decision best separates the data. Features at the top of the tree are the most important and have the greatest impact on decision-making. Therefore, analyzing the decisions at the top of the decision

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

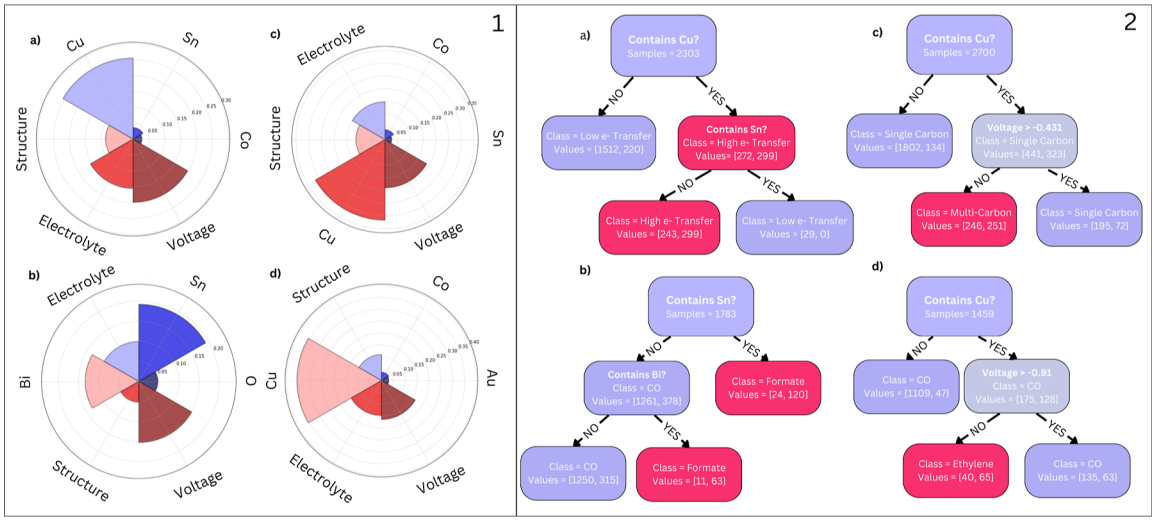


Figure 4. Feature importance (1) and decision trees (2) for (a) whether the electron transfer process was low (2 e^- transfers) or high (more than 2 e^- transfers). (b) Whether CO or formate/formic acid was the major product, (c) whether the reaction produced multicarbon or single-carbon products, and (d) whether CO or ethylene was the major product.

tree can provide valuable insights to the experimental researcher regarding properties that are more likely to influence the outcome of the catalytic reaction. Random forest models, which are an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting, are also highly interpretable. First, by measuring how each tree improves decision splits, random forests can provide insights into which features are most important. Random forests are also highly receptive to posthoc interpretability techniques such as SHAP, which quantifies the impact of each feature on a machine learning model's predictions. Moreover, shallow learning models can provide this comprehensive understanding of a data set with very minimal tuning. In both decision trees and random forests, the primary parameter that influences overfitting versus underfitting is the tree depth, which is very easy to tune. Therefore, we opted to analyze this data set using both decision trees and random forest models to gain maximum insight from the low fidelity data, with minimum tuning required.

The next step to gaining human understanding using machine learning is to ask the machine learning models the correct questions for the type of data that you have. Here, our goal was to determine how the interactions of the different catalysts and reaction parameters that the GPT LLM model extracted affected the outcome of the CO_2RR. To achieve this, we designed a series of binary classification questions relating to CO_2 electroreduction. Specifically, we asked the models to classify: (1) whether the electron transfer process was low (2 e^- transfers) or high (more than 2 e^- transfers), (2) whether CO or formate/formic acid was the major product, (3) whether the reaction produced multicarbon or single-carbon products, and (4) whether CO or ethylene was the major product. These questions were chosen as they are of interest within the CO_2 electroreduction community. Moreover, choosing only binary classification questions helps with the model interpretability, and understanding which features affect the outcome of these binary classification questions helps humans understand these complex data sets.

For question 1 (whether the electron transfer process was low or high) and question 3 (whether the reaction produced multicarbon or single-carbon products), the full data set was used for machine learning. For questions 2 (whether CO or formate/formic acid was the major product) and 4 (whether CO or ethylene was the major product), a subset of the data set was used that only contained examples where one of the two products of interest was produced. Each data set was split into a training set to train the model, a validation set to select a tree depth that limits over- and under-fitting, and a final testing of data that the model has not seen during training. For each tree depth, the accuracy of the model on the training set and the validation set was calculated (Supporting Information Figures S5−S12). To prevent over- and under-fitting, the smallest tree depth that produced maximum accuracy on the validation set was used. Once the tree depth was chosen, the accuracy of the testing data was calculated for each model, with no additional tuning.

The accuracy and confusion matrices of each testing set for both the decision tree and random forest models across the four classification tasks are shown in Figure 3. In many machine learning studies, the goal is to obtain a generalization model that can achieve high accuracy on the data set (both training and testing) collected. Here, the goal is to use the model accuracy of the shallow learning models to quantify the quality and amount of data collected by the GPT LLM automated data extraction. Figure 3 shows that relatively high shallow-learning accuracies can be achieved from data automatically collected from GPT LLM techniques. Specifically, for predicting electron transfer as low (2 e^-) or high (>2 e^-), the decision tree model achieved an average accuracy of 0.82 with a 5-fold cross-validation of 0.80, while the random forest model performed slightly better at 0.86 with a cross-validation of 0.82 (Figure 3a). The prediction of whether CO or formate/formic acid was the favored product resulted in an average accuracy of 0.81 for both models (Figure 3b). The 5-fold cross-validation for the decision tree was 0.80, and the random forest was 0.81. The classification of multicarbon versus single-carbon products yielded similar results, with the decision tree reaching 0.86 average accuracy and the random forest slightly lower at 0.85 (Figure 3c). The 5-fold cross-validation was 0.83 for the decision tree and 0.85 for the random forest model. The highest accuracy was observed in

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

the classification of CO versus ethylene selectivity, where both models achieved 0.93 average accuracy with a 5-fold cross-validation score of 0.90 for both models, suggesting that there could be a clearer distinction or separation of features within the data set (Figure 3d).

The relatively high accuracy of the shallow learning models suggests the effectiveness of the GPT LLM data collection procedure. If there were large inaccuracies with the automated data collection of the features, then we would have expected a very low accuracy in the testing. Additionally, significant inaccuracies in the extraction of the products would result in either highly unbalanced data sets or poor predictive performance. Since the accuracies were high, we can have high confidence that the GPT LLM model is able to extract the relevant information from a single prompt. Moreover, analysis of the confusion matrices in Figure 3 shows whether the data set is balanced or imbalanced. Typically, in machine learning models, one desires a highly balanced data set to improve the prediction power. However, when data from the experimental literature are collected, the data might be inherently imbalanced due to the actual results and findings from experimental researchers. Here, we can see that despite the desire of the CO_2RR community to produce high-value multicarbon products, the major products of most studies are biased toward single-carbon, low electron transfer products.

Beyond classification accuracy, the interpretability of these models provides valuable insights into the underlying trends governing the CO_2 electroreduction. Here, we show that combining the analysis of feature importance from the random forest model with the analysis of the first few splits of the decision trees can provide great insights into key parameters of a catalytic reaction. For example, starting with the low or high electron classification in Figure 4-1a, the random forest feature importance found that the two most important contributors to the difference between high and low electron transfers were the use of Cu catalysts and the voltage. Now connecting that to the decision tree (Figure 4-2a), the first split of the decision tree was whether the catalyst contains Cu. The tree determined that if the catalyst contained Cu, the electron transfer rate would be high, with over half, 299 out of the 519 high electron samples, using Cu in the catalysts. If the catalyst did not contain copper, the model predicted that it would have low electron transfer, with 85% of the low electron samples not containing copper. Moreover, the decision tree shows that 29 samples used catalysts that contained both Cu and Sn, and 100% of those samples underwent low electron transfers. This is an interesting observation because it demonstrates that combining Cu with Sn favors the typical Sn-only mechanism compared to the Cu-only mechanism.

The two major products for samples that undergo low electron transfers are carbon monoxide (CO) or formate/formic acid. Since both of these products are single carbon and exhibit low electron transfers, we expected different trends highlighted by both the random forest and decision tree models than the earlier questions. As shown in Figure 4-1b, the random forest model identified the presence of Sn in the catalyst as the most important feature, followed by voltage and the presence of Bi in the catalyst. This finding was further supported by the decision tree model (Figure 4-2b), which showed that both Sn and Bi produce formate/formic acid. For those researchers already experienced with electrocatalytic CO_2 reduction, it is known that both Sn and Bi are catalysts that primarily form formate/formic acid. Nevertheless, the fact that the GPT LLM machine learning framework was able to independently arrive at these conclusions lends credibility to the overall process and the other findings of the model.

For the case of single versus multicarbon product formation, the feature importance noted by the random forest model was overwhelmingly the presence of Cu in the catalyst, followed by voltage (Figure 4-1c). This trend is similar to the electron transfer classification, as the underlying reaction pathways largely overlap, except those of methane and methanol, which require a higher electron transfer, shift from the high electron transfer category to the single carbon category. The dependence on copper previously seen by the electron transfer classification and the random forest model is reiterated in the decision tree model, with the root node containing Cu. However, only having a catalyst that contains Cu does not fully determine if a multicarbon product is made. Instead, while 80% of single carbon data points lack a Cu presence in the catalyst, the formation of multicarbon products is only observed by the model when Cu is present and the voltage is more negative than −0.431 V vs RHE, as seen by the second split in the decision tree (Figure 4-2c). This tie between the applied voltage and the catalysts for promoting C−C coupling is an important observation and one that might not be obvious to those who even work in the field of electrocatalytic CO_2 reduction. These machine learning insights are consistent with trends we have seen in DFT literature for C−C bond formations. It has been demonstrated that C−C bonds are formed at −0.4 V vs RHE through a CO dimer,[37] which is similar to our −0.431 V prediction for C−C coupling.

For the final proposed question, we evaluated how machine learning models predict the formation of CO versus ethylene. We asked this as it is similar to both the electron transfer and the carbon−carbon coupling, but on a more precise scale. As expected, the random forest returned that Cu was the most important feature (Figure 4-1d). The decision tree also confirmed the importance of Cu as an important feature shown in (Figure 4-2d). If the catalyst did not contain copper, it separated 86% of the CO data points from the majority of the ethylene production data. However, interestingly, 58% of the samples whose catalyst contained Cu produced CO over ethylene. Analyzing the decision tree further shows that a voltage more negative than −0.91 V vs RHE was required to classify the sample as ethylene. Here, 62% of samples whose catalysts contained Cu and voltage was more negative than −0.91 V vs RHE produced ethylene over CO. Comparing this to the carbon−carbon coupling predictions, the voltage to produce ethylene with a Cu catalyst −0.91 V, was relatively higher than separating single carbon from multicarbon products −0.431 V vs RHE, suggesting that forming ethylene requires a more negative potential compared to other multicarbon products. These findings are consistent with DFT studies for C−C coupling versus ethylene formation. The DFT study mentioned before found that ethylene formation needs −1.0 V vs RHE, which was more than their finding of −0.4 for C−C coupling.[37]

While feature importance for random forest has proven itself to be an important tool for identifying key catalytic parameters, it does not provide direct insight into how these features influence the reaction. To further investigate the impact of the features on the reaction, we used SHAP to quantify the contribution of individual features to the model's predictions. SHAP works by connecting optimal credit allocations with local explanations. For binary classification, SHAP will tell you

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

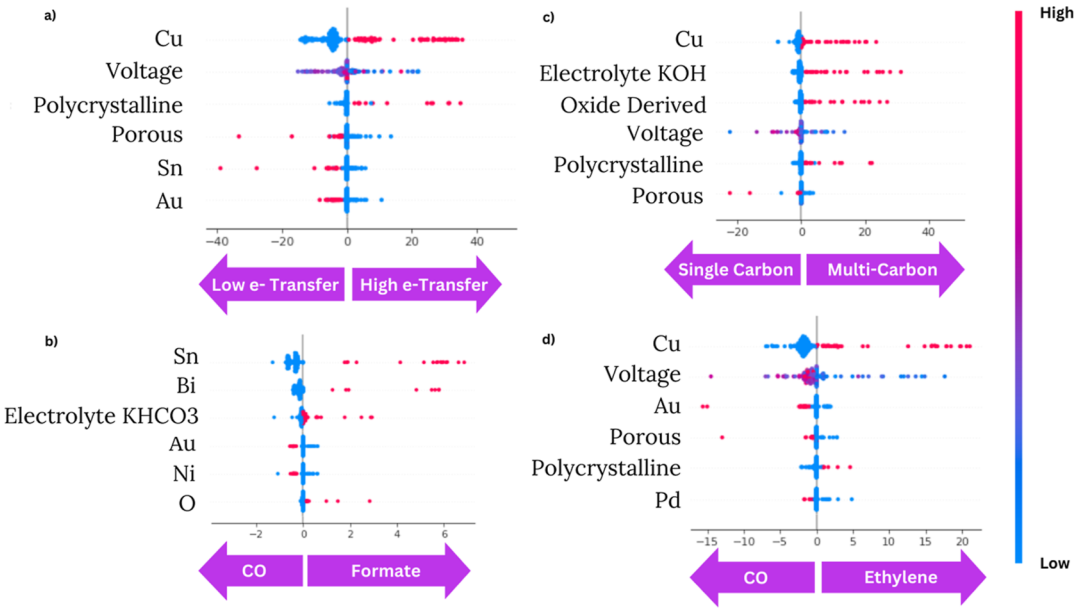


Figure 5. SHAP force plots including the top 6 influential features for each of the four questions asked to the database. (a) Whether the electron transfer process was low (2 e^- transfers) or high (more than 2 e^- transfers). (b) Whether CO or formate/formic acid was the major product, (c) whether the reaction produced multicarbon or single-carbon products, (d) whether CO or ethylene was the major product.

if higher or lower values indicate a specific prediction by the model, making it a powerful method for interpreting machine-learning models. By structuring the complicated problem of CO_2 electroreduction as a series of binary classification questions, we can extract more detailed insights into underlying reaction trends. To truly understand how each individual feature influenced the reaction, we decided to one-hot encode the previously label-encoded features. For example, instead of stating that "electrolyte" influences a reaction a certain way, it breaks down each electrolyte in the database and determines how the reaction is influenced by the individual electrolytes.

For each SHAP analysis, we took the six most relevant and distinguishing features and displayed them in Figure 5. In the case of predicting low or high electron transfer (Figure 5a), the features chosen were Cu, voltage, polycrystalline structure, porous structure, and Sn- and Au-containing catalysts. The SHAP analysis displays a strong distinction between catalysts containing Cu and those that do not. Catalysts containing Cu overwhelmingly contribute to high electron transfer, as indicated by the red data points in the plot, while catalysts without Cu present overwhelmingly contribute to low electron transfer. This reinforces the well-known connection that Cu catalysts have within the reaction and shows that machine learning models can pick up on catalytic trends within the data.

Since the reported voltages are continuous values rather than binary, their SHAP values display a gradient from blue to red rather than distinct classifications. In the case of the electron transfer, we see that lower values (i.e., more negative potentials) are most prominent in creating high electron transfer. Interestingly, some higher values (i.e., less negative) potentials were still able to facilitate high electron transfers, indicating that the voltage was not the sole determining factor. Additionally, two structures were also shown as promising features by SHAP analysis with opposing influences on the reaction. Polycrystalline catalysts were found to be more likely to facilitate high electron transfer, while porous catalysts were more likely to produce low electron transfer. Circling back to catalyst metals, the model was able to correctly identify that both Sn and Au are known to favor low electron transfer products, which further reinforces known catalytic trends.

Moving the case of CO versus formate/formic acid production (Figure 5b), the key features identified by the SHAP analysis were almost all relating to catalyst composition. The model determined that the presence of Sn, Bi, or oxygen in the catalyst strongly favored formate/formic acid formation, whereas catalysts containing Au or Ni were more likely to favor CO production. Additionally, the last feature that was the most significant in the SHAP analysis was the electrolyte $KHCO_3$, which favored formate/formic acid production. Comparing these results to the feature importance in Figure 4, we see that while voltage was ranked as the second most important feature in the overall classification, it was not as influential in determining the favored product. This suggests that while voltage plays an important role in the reaction, its effect on product distribution may be more complex and catalyst composition dependent rather than a primary factor. This is a benefit to using SHAP in congruence with the feature importance within a random forest model to understand which features are important and how they can interact to influence specific outcomes.

Moving to the single carbon versus multicarbon classification (Figure 5c), we expected to observe similar trends in features as the electron transfer again due to their similarity in products produced, which we do see, with the Cu catalyst, voltage, polycrystalline, and porous structures all having the same overarching trends as the electron transfer SHAP had shown. However, two new features were highlighted with regard to carbon−carbon coupling. The use of KOH as the electrolyte and the catalyst being oxide-derived strongly favored the production of multicarbon products. Although oxide-derived catalysts are well-known for producing multicarbon products, the utilization of KOH for C−C coupling may not be immediately apparent to individuals working in the field of electrocatalytic CO_2 reduction.

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

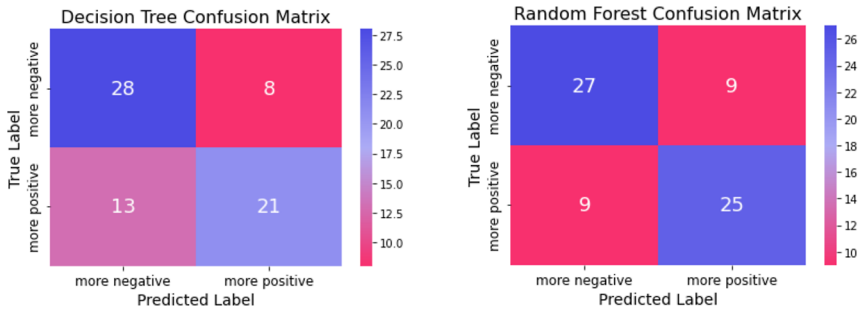


Figure 6. Confusion matrices from the results of the decision tree (left) and random forest (right) for the Oxygen Reduction Reaction (ORR) dataset to predict if the half-wave potential was more positive or more negative than the median half-wave potential.

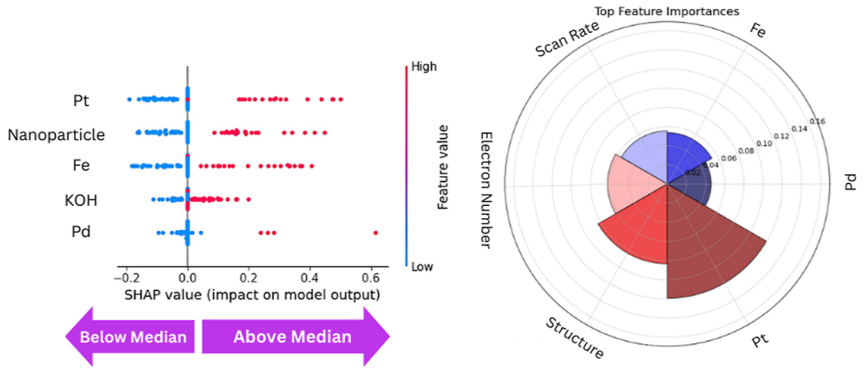


Figure 7. SHAP force plots (left) and feature importance (right) for the Oxygen Reduction Reaction (ORR) dataset that shows which features are most important in predicting if the half-wave potential was above or below the median half-wave potential.

Lastly, for the case of CO versus ethylene production (Figure 5d), we observe a combination of the features previously identified by the electron transfer and multicarbon classifications. These features include Cu, voltage, Au, and polycrystalline and porous. As expected, Cu in the catalyst favors the production of ethylene, as ethylene requires multiple electron transfers and carbon−carbon coupling. It is also observed that more negative voltage values favor ethylene, while more positive and intermediate values favor CO production. Similar to previous findings in both the electron transfer and the CO versus formate questions, SHAP determined that catalysts containing Au favor CO. Additionally, SHAP reinforced that the porous and polycrystalline structures promote opposite effects on product production, as polycrystalline favors ethylene and porous favors CO. Lastly, we see that catalysts containing Pd favor CO.

To demonstrate reproducibility of this extraction framework, we took another reaction, the ORR, and performed the same procedure of feature extraction and ML predictions. For this demonstration, we took a smaller data set spanning 5 years of literature from 2020 to 2024. Since ORR has different features than the CO_2RR the prompt was updated to the following: "extract the relevant electrochemical ORR catalyst data with columns including: metal, if the catalyst is an oxide yes or no, half wave potential, Tafel slope, overpotential, reference electrode, electrolyte, the catalyst support material, pH, scan rate, rpm, average electron number, and if the catalyst structure one of the following: nanoparticle, polycrystalline, foil, thin film, porous, oxide derived, or amorphous". The resulting database was evaluated and cleaned using the same procedure as the CO_2 database, removing features that did not have enough data points for predictions.

Unlike the CO_2RR, the ORR does not have as many products to use for predictions. Therefore, a new question was proposed that needed to be answered. For the ORR, the half-wave potential is an important indicator of catalytic efficiency and was found to be reported in many ORR papers. We decided from the extracted literature articles to predict if the half-wave potential was higher or lower than the median half-wave potential over the 5-year span of articles. The median from the extracted papers was found to be 0.86 V vs RHE. Data points higher than the 0.86 V were marked as above median for training the machine learning models. We chose to predict if the reaction parameters were above the median half-wave potential, as those would be the more favorable reaction conditions. Additionally, we chose the half-wave prediction as it is still a binary classification problem, the same as the CO_2 predictions. The data set used for the machine learning consisted of 693 data points, of which 352 were below the median and 341 were above the median half-wave potential.

Starting with decision trees, the binary classification of predicting above the median produced an accuracy of 0.70 with a max depth of 6. For the smaller ORR data set, this is considered good as there are fewer data points for model training and is better than a 50/50 random guessing that could occur with low amounts of data. The random forest model on the ORR data produced an accuracy of 0.74 with a maximum depth of 5. The comparison of both models' predictions can be seen in (Figure 6). The feature importance from the random forest gave Pt as the most important feature with an importance of 0.12. This was closely followed by the catalyst structure being a nanoparticle as the next most important feature, with an importance of 0.08. To understand how the individual features influenced the prediction, SHAP was performed on the random forest model. SHAP identified the

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

most influential features on the model predictions to be Pt, nanoparticle structure, Fe, KOH as the electrolyte, and Pd. If the reaction conditions contained any of those features the model was more likely to predict that the half-wave potential was higher than the median (Figure 7).

With both the ORR and CO_2 data sets, there are important catalytic descriptors that we believe could have made the accuracy of the models significantly higher. However, they are not consistently reported within the text of the literature. These parameters were left out of the database, including faradic efficiency, surface area, and catalyst support material. Since LLMs are currently limited to reading text and some of these features are commonly reported within figures as data points on a graph, they are difficult to extract. Advancements in computer vision and large language models are needed to extract these important catalytic parameters in their current form. Additionally, there is no standard way to report specific data features within literature articles. Some articles lacked important features such as catalyst surface area within the paper as a whole, which leads to significant gaps when doing machine learning predictions.

As a test to determine if non-catalytic parameters influenced the model as a feature, we added the year of publication to the ORR predictions. The year of publication is good to know within the data set to keep track of trends within the literature. However, we wanted to determine if it would influence the model decision making, as it is not directly related to the catalytic reaction parameters. For the decision tree model, it did not affect the model accuracy as it remained at 0.70 at a max depth of 6. For the random forest model, it only slightly increased the accuracy to 0.77 with an additional 2 correctly predicted data points at a max depth of 6. Thus, adding the year as a feature does not significantly increase the accuracy of the models.

4. CONCLUSION

There are still open questions regarding how machine learning models trained on experimental data taken from across research groups can be utilized to design novel catalysts, given the limitations in terms of data depth and quality. Nevertheless, the framework presented showcases several key applications of machine learning that will be beneficial to numerous experimental researchers engaged in catalysis.

First, we show that commercial GPT LLM models are accurate enough to construct low-fidelity databases automatically with a single prompt, given that the prompt is engineered properly. Second, we show that a mixed label and one-hot encoding system of this GPT LLM database is necessary to extract maximum information from the low-fidelity database. Third, we demonstrate that, due to the low fidelity of the automated database, applying shallow-learning models with posthoc interpretability can provide human researchers with a greater understanding of the complexity of the catalytic system compared to simply running a "black-box" deep learning model that provides a prediction without any indication of the reasoning behind it. Furthermore, these shallow learning models can be utilized with minimal tuning, enhancing their accessibility compared with more intricate models. Finally, we demonstrate that instead of using the database to make a complex multiclass prediction, dividing the problem into several binary classification problems provides the experimental researcher with more insights from the interpretable algorithms.

To demonstrate these findings, we applied this framework to the electrocatalytic reduction of carbon dioxide as a model reaction due to the challenges in understanding product selectivity. We observed that this framework was able to independently obtain information that is known by the CO_2 electrocatalysis community, such that Sn and Bi produce formate/formic acid, Au and Ni produce CO, and Cu undergoes both high electron-transfer reactions and C−C coupling. The fact that these well-known relations were identified by the machine-learning framework gives credence to the machine-learning model. In addition, the framework was able to identify more complex interactions that could be useful to experimental researchers. For example, the framework indicates that to produce multicarbon products, oxide-derived Cu catalysts in KOH with a voltage more negative than ca. −0.45 V vs RHE will likely create multicarbon products. To produce ethylene specifically, the model suggests that polycrystalline Cu at a voltage more negative than ca. −0.9 V vs RHE is preferred.

We believe that using this type of analysis gives experimental researchers more confidence in the outcomes and more understanding than a "chat bot" style GPT model, where no information is given on why or how the decisions are being made. Moreover, we believe that as artificial intelligence and large language models continue to advance, they will play a critical role in helping catalytic researchers understand these complex systems more effectively.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.5c01087.

Data encoding tables, additional decision tree graphs, max depth determination figures, additional full SHAP analysis figures, and database of extracted properties (PDF)

Sample Python code (PDF)

Information on metals, voltage, units of voltage, reference electrode, electrolyte, surface area, GDE, structure, and molarity (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Kevin C. Leonard − *Department of Chemical & Petroleum Engineering, The University of Kansas, Lawrence, Kansas 66045, United States; Center for Environmentally Beneficial Catalysis, The University of Kansas, Lawrence, Kansas 66045, United States;* orcid.org/0000-0002-0172-3150; Email: kcleonard@ku.edu

Author

Brianna R. Farris − *Department of Chemical & Petroleum Engineering, The University of Kansas, Lawrence, Kansas 66045, United States; Center for Environmentally Beneficial Catalysis, The University of Kansas, Lawrence, Kansas 66045, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.5c01087

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

■ ACKNOWLEDGMENTS

This work was funded by the U.S. National Science Foundation Research Traineeship (NRT) grant through Award DGE-1922649. We also acknowledge support by the U.S. Army DEVCOM ARL Army Research Office (ARO) Energy Sciences Competency, Electrochemistry Program award # W911NF-22-1-0293. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army or the U.S. Government.

■ REFERENCES

(1) Sun, J.; Liu, C. What and how can machine learning help to decipher mechanisms in molecular electrochemistry? *Curr. Opin. Electrochem.* **2023**, *39*, 101306.
(2) Mou, L. H.; Han, T. T.; Smith, P. E.; Sharman, E.; Jiang, J. Machine Learning Descriptors for Data-Driven Catalysis Study. *Adv. Sci.* **2023**, *10*, 2301020.
(3) Behr, A. S.; Chernenko, D.; Koßmann, D.; Neyyathala, A.; Hanf, S.; Schunk, S. A.; Kockmann, N. Generating knowledge graphs through text mining of catalysis research related literature. *Catal. Sci. Technol.* **2024**, *14*, 5699−5713.
(4) Yang, W.; Fidelis, T. T.; Sun, W. H. Machine Learning in Catalysis, from Proposal to Practicing. *ACS Omega* **2020**, *5*, 83−88.
(5) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260−2297.
(6) Hickey, K.; Feinstein, J.; Sivaraman, G.; MacDonell, M.; Yan, E.; Matherson, C.; Coia, S.; Xu, J.; Picel, K. Applying machine learning and quantum chemistry to predict the glass transition temperatures of polymers. *Comput. Mater. Sci.* **2024**, *238*, 112933.
(7) Mohan, M.; Smith, M. D.; Demerdash, O. N.; Simmons, B. A.; Singh, S.; Kidder, M. K.; Smith, J. C. Quantum Chemistry-Driven Machine Learning Approach for the Prediction of the Surface Tension and Speed of Sound in Ionic Liquids. *ACS Sustainable Chem. Eng.* **2023**, *11*, 7809−7821.
(8) Zhang, S. Q.; Xu, L. C.; Li, S. W.; Oliveira, J. C.; Li, X.; Ackermann, L.; Hong, X. Bridging Chemical Knowledge and Machine Learning for Performance Prediction of Organic Synthesis. *Chem.—Eur. J.* **2023**, *29*, No. e202202834.
(9) AlQuraishi, M. Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* **2021**, *65*, 1−8.
(10) Kouba, P.; Kohout, P.; Haddadi, F.; Bushuiev, A.; Samusevich, R.; Sedlar, J.; Damborsky, J.; Pluskal, T.; Sivic, J.; Mazurenko, S. Machine Learning-Guided Protein Engineering. *ACS Catal.* **2023**, *13*, 13863−13895.
(11) Zhang, W.; Wang, Y.; Ren, S.; Hou, Y.; Wu, W. Novel Strategy of Machine Learning for Predicting Henry's Law Constants of CO2 in Ionic Liquids. *ACS Sustainable Chem. Eng.* **2023**, *11*, 6090−6099.
(12) Huang, C.; Jia, X.; Zhang, Z. A modified back propagation artificial neural network model based on genetic algorithm to predict the flow behavior of 5754 aluminum alloy. *Materials* **2018**, *11*, 855.
(13) Cao, Z.; Dan, Y.; Xiong, Z.; Niu, C.; Li, X.; Qian, S.; Hu, J. Convolutional Neural Networks for Crystal Material Property Prediction Using Hybrid Orbital-Field Matrix and Magpie Descriptors. *Crystals* **2019**, *9*, 191.
(14) Valli Priyadharshini, K.; Vijay, A.; Swaminathan, K.; Avudaiappan, T.; Banupriya, V. Materials property prediction using feature selection based machine learning technique. *Mater. Today: Proc.* **2022**, *69*, 710−715.
(15) Don-Tsa, D.; Mohou, M. A.; Amouzouvi, K.; Maaza, M.; Beltako, K. Predictive models for inorganic materials thermoelectric properties with machine learning. *Mach. Learn.: Sci. Technol.* **2024**, *5*, 035067.
(16) Tamtaji, M.; Chen, S.; Hu, Z.; Goddard, W. A.; Chen, G. A Surrogate Machine Learning Model for the Design of Single-Atom Catalyst on Carbon and Porphyrin Supports towards Electrochemistry. *J. Phys. Chem. C* **2023**, *127*, 9992−10000.
(17) Rittig, J. G.; Felton, K. C.; Lapkin, A. A.; Mitsos, A. Gibbs−Duhem-informed neural networks for binary activity coefficient prediction. *Digital Discovery* **2023**, *2*, 1752−1767.
(18) Saquer, N.; Iqbal, R.; Ellis, J. D.; Yoshimatsu, K. Infrared spectra prediction using attention-based graph neural networks. *Digital Discovery* **2024**, *3*, 602−609.
(19) Farris, B. R.; Niang-Trost, T.; Branicky, M. S.; Leonard, K. C. Evaluation of Machine Learning Models on Electrochemical CO2Reduction Using Human Curated Datasets. *ACS Sustainable Chem. Eng.* **2022**, *10*, 10934−10944.
(20) Rosser, D. A.; Farris, B. R.; Leonard, K. C. Predictive machine learning models trained on experimental datasets for electrochemical nitrogen reduction. *Digital Discovery* **2024**, *3*, 667−673.
(21) Leonard, K. C.; Hasan, F.; Sneddon, H. F.; You, F. Can Artificial Intelligence and Machine Learning Be Used to Accelerate Sustainable Chemistry and Engineering? *ACS Sustainable Chem. Eng.* **2021**, *9*, 6126−6129.
(22) Singh, S.; Sunoj, R. B. Molecular Machine Learning for Chemical Catalysis: Prospects and Challenges. *Acc. Chem. Res.* **2023**, *56*, 402−412.
(23) Al Abdulghani, A. J.; Turizo-Pinilla, E. E.; Fabregas-Angulo, M. J.; Hagmann, R. H.; Ibrahim, F.; Jansen, J. H.; Agbi, T. O.; Bhat, S.; Sepúlveda-Pagán, M.; Kraimer, M. O.; Queen, C. M.; Sun, Z.; Nikolla, E.; Pagán-Torres, Y. J.; Hermans, I. Realizing synergy between Cu, Ga, and Zr for selective CO2 hydrogenation to methanol. *Appl. Catal., B* **2024**, *355*, 124198.
(24) Hebisch, K. L.; Chmielniak, P. A.; Cutler, V. A.; Watson, R. B.; Gołąbek, K.; Sievers, C. Promoting role of residual organic structure directing agent in skeletal butene isomerization over ferrierite. *J. Catal.* **2025**, *441*, 115847.
(25) Vos, R. E.; Koper, M. T. Nickel as Electrocatalyst for CO(2) Reduction: Effect of Temperature, Potential, Partial Pressure, and Electrolyte Composition. *ACS Catal.* **2024**, *14*, 4432−4440.
(26) Leofanti, G.; Padovan, M.; Tozzola, G.; Venturelli, B. Surface area and pore texture of catalysts. *Catal. Today* **1998**, *41*, 207−219.
(27) Mayrhofer, K. J.; Blizanac, B. B.; Arenz, M.; Stamenkovic, V. R.; Ross, P. N.; Markovic, N. M. The impact of geometric and surface electronic properties of Pt-catalysts on the particle size effect in electrocatalysis. *J. Phys. Chem. B* **2005**, *109*, 14433−14440.
(28) Abdelmaqsoud, K.; Shuaibi, M.; Kolluru, A.; Cheula, R.; Kitchin, J. R. Investigating the error imbalance of large-scale machine learning potentials in catalysis. *Catal. Sci. Technol.* **2024**, *14*, 5899−5908.
(29) Kolluru, A.; Shuaibi, M.; Palizhati, A.; Shoghi, N.; Das, A.; Wood, B.; Zitnick, C. L.; Kitchin, J. R.; Ulissi, Z. W. Open Challenges in Developing Generalizable Large-Scale Machine-Learning Models for Catalyst Discovery. *ACS Catal.* **2022**, *12*, 8572−8581.
(30) Schilling-Wilhelmi, M.; Ríos-García, M.; Shabih, S.; Gil, M. V.; Miret, S.; Koch, C. T.; Márquez, J. A.; Jablonka, K. M. From Text to Insight: Large Language Models for Materials Science Data Extraction. *arXiv* **2024**, arXiv:2407.16867.
(31) Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525−535.
(32) Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **2024**, *15*, 1569.
(33) Gougherty, A. V.; Clipp, H. L. Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. *NPJ Biodiversity* **2024**, *3*, 13.
(34) Suvarna, M.; Vaucher, A. C.; Mitchell, S.; Laino, T.; Pérez-Ramírez, J. Language models and protocol standardization guidelines for accelerating synthesis planning in heterogeneous catalysis. *Nat. Commun.* **2023**, *14*, 7964.
(35) Pedregosa, F.; et al. *Scikit-Learn: Machine Learning in Python*, Tech. Rep., 2011; pp 2825−2830.

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589

(36) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.0787.
(37) Goodpaster, J. D.; Bell, A. T.; Head-Gordon, M. Identification of Possible Pathways for C-C Bond Formation during Electrochemical Reduction of CO2: New Theoretical Insights from an Improved Electrochemical Model. *J. Phys. Chem. Lett.* **2016**, *7*, 1471−1477.

https://doi.org/10.1021/jacsau.5c01087
JACS Au 2025, 5, 5578−5589