Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 14, No. 2, 2024

Data-Driven Approaches for Predicting Catalyst Performance in CO2 Hydrogenation

Yuqi Liang [1], Yan Zhang [2]

[1] School of Computer Science, Shandong University, China
[2] Department of Computer Science, Nanjing University, China

Abstract: This paper explores the critical role of data collection and preparation in leveraging machine learning techniques for predicting catalyst performance in CO2 hydrogenation processes. As the global community seeks sustainable solutions to mitigate carbon emissions, understanding the efficiency of catalysts in converting CO2 into valuable chemicals becomes increasingly important. The study discusses various types of data utilized in this context, including both experimental and simulation data, while highlighting their significance in comprehensively understanding key factors such as reaction rates, selectivity, and catalyst stability. For instance, the ability of a catalyst to selectively produce desired products over others can significantly impact the overall economic viability of CO2 hydrogenation processes. Furthermore, stability data sheds light on the longevity and durability of catalysts, revealing insights into deactivation mechanisms that can occur due to factors like sintering, poisoning, or leaching of active sites. On the other hand, simulation data generated from advanced computational methods such as density functional theory (DFT) and molecular dynamics (MD) provides a deeper understanding of the electronic and structural properties of catalysts. These computational techniques allow researchers to predict reaction pathways, activation energies, and the behavior of intermediate species, thereby complementing experimental findings and guiding the design of more effective catalysts. The paper also emphasizes the importance of utilizing publicly available databases and collaborative research datasets, which serve to enhance data accessibility and foster scientific collaboration among researchers in the field. By pooling resources and sharing findings, the scientific community can accelerate the discovery and optimization of novel catalysts. Additionally, the study outlines essential steps in the data preparation process, including rigorous data cleaning, preprocessing, and feature selection, all of which are crucial for ensuring the quality and reliability of the data used in machine learning models. The paper discusses the implementation of cross-validation techniques and performance metrics that help evaluate and validate model predictions, ensuring that the developed models generalize well to unseen data.

Keywords: Catalyst Performance; CO2 Hydrogenation; Machine Learning.

1. Introduction

The increasing concentration of carbon dioxide (CO2) in the atmosphere has become a pressing global concern, primarily due to its significant role in climate change and global warming [1]. CO2 hydrogenation, a process that converts CO2 into valuable hydrocarbons and alcohols using hydrogen, presents a promising solution to mitigate greenhouse gas emissions while simultaneously producing useful chemicals [2]. This transformation not only provides a pathway for carbon capture and utilization (CCU) but also offers a means to store renewable energy in chemical form [3].

Catalysts are central to the efficiency and selectivity of CO2 hydrogenation reactions. They facilitate the conversion of reactants into products by lowering activation energy and providing alternative reaction pathways [4]. The choice of catalyst significantly influences the reaction kinetics, product distribution, and overall process feasibility [5]. However, the identification and optimization of effective catalysts remain challenging due to the complex interplay of various factors, including catalyst composition, structure, and reaction conditions [6].

Recent advancements in data science and machine learning (ML) have opened new avenues for catalyst discovery and optimization [7]. Data-driven methodologies involve the use of large datasets generated from experimental and computational studies to develop predictive models for catalyst performance [8]. Leveraging these datasets can significantly enhance our understanding of catalyst behavior and facilitate the identification of promising candidates for CO2 hydrogenation [9].

This paper aims to explore how machine learning can be employed to predict catalyst performance in CO2 hydrogenation by utilizing large datasets derived from both experimental and simulation results. We will discuss the methodologies, challenges, and opportunities associated with data-driven approaches in this context, highlighting their potential to accelerate catalyst discovery and optimize CO2 conversion processes.

2. Literature Review

The quest for effective catalysts for CO2 hydrogenation has a rich history, dating back to the early 20th century with the development of metal catalysts [10-15]. Early studies focused on the use of noble metals such as Pt and Ru, which demonstrated high activity but suffered from high costs and limited availability [16-20]. Over time, researchers have explored various catalyst systems, including transition metals, metal oxides, and supported catalysts, to improve performance and reduce costs [21].

Recent advancements have led to the development of novel catalyst formulations, including bimetallic catalysts and metal-organic frameworks, which have shown enhanced activity and selectivity [22-26]. Additionally, the incorporation of promoters and the optimization of support materials have been explored to fine-tune catalyst properties [27-30].

Traditional modeling approaches, such as density

functional theory (DFT) and empirical models, have been extensively used to understand catalyst behavior and predict performance [31-33]. DFT provides insights into the electronic structure of catalysts and their interaction with reactants, enabling the prediction of reaction pathways and energetics [34,35]. However, these methods can be computationally intensive and may not always capture the complexity of real-world catalytic systems [36].

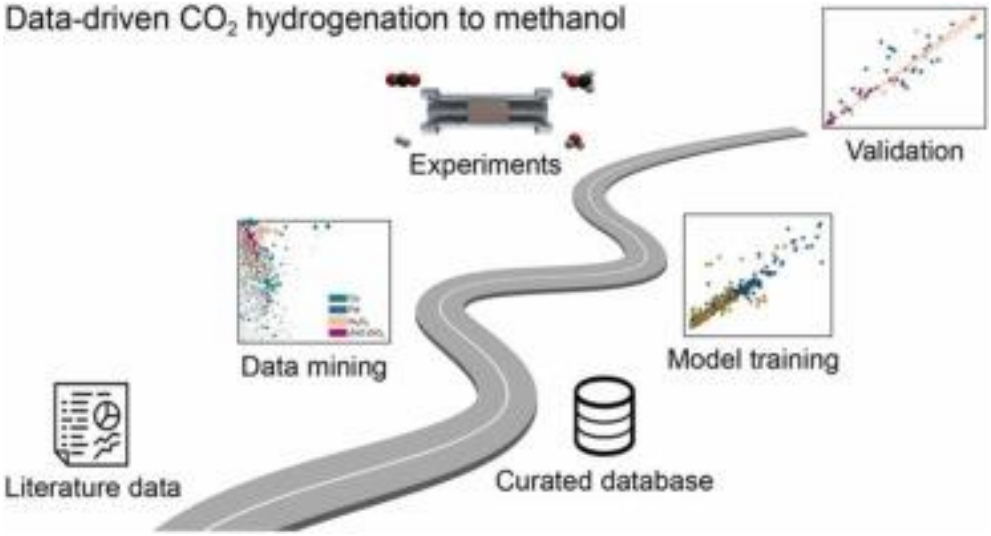

Figure 1. Data-driven CO, hydrogenation to methanol

Despite their utility, traditional modeling approaches often face limitations in terms of scalability and generalizability. The reliance on specific catalyst systems and reaction conditions can hinder the transferability of findings to other contexts [37,38]. Furthermore, the need for extensive computational resources can restrict the exploration of vast chemical spaces [39-41].

Machine learning has emerged as a powerful tool in materials science, enabling researchers to analyze large datasets and uncover patterns that traditional methods may overlook [42-44]. In catalysis, ML techniques have been applied to predict properties such as catalytic activity, selectivity, and stability based on compositional and structural features [45].

Several studies have demonstrated the successful application of machine learning in predicting catalyst performance. For instance, a study by [46] employed a random forest model to predict the activity of metal catalysts for CO2 hydrogenation, achieving high accuracy with minimal computational cost. Similarly, [47,48] utilized deep learning techniques to develop a predictive model for catalyst selectivity, showcasing the potential of ML in accelerating catalyst discovery.

3. Data Collection and Preparation

3.1. Types of Data Utilized

3.1.1. Experimental Data

Experimental data is crucial for understanding catalyst performance in CO2 hydrogenation. This data typically encompasses a wide range of metrics, including reaction rates, selectivity, and stability. Reaction rates can be measured under various conditions, such as temperature, pressure, and reactant concentrations, allowing researchers to assess the catalytic activity of different materials comprehensively. For instance, varying the temperature can help identify optimal conditions for maximizing reaction rates, while adjusting reactant concentrations can reveal insights into the kinetics of the reaction.

Selectivity data is particularly important in multi-product reactions, where a catalyst may produce several different products from CO2 hydrogenation. Understanding a catalyst's ability to favor one product over another is vital for optimizing processes, as it can significantly impact the efficiency and economic viability of the overall reaction. For example, a catalyst that selectively produces methanol over other hydrocarbons may be more desirable for specific applications.

Stability data is essential for evaluating the longevity and durability of catalysts under operational conditions. This data provides insights into deactivation mechanisms, which can result from factors such as sintering, poisoning, or leaching of active sites. By understanding how catalysts behave over time, researchers can develop strategies to enhance their stability, leading to more sustainable and cost-effective catalytic processes.

3.1.2. Simulation Data

Simulation data, often derived from computational methods such as density functional theory and molecular dynamics, provides valuable insights into the electronic and structural properties of catalysts. DFT calculations are particularly useful for predicting reaction pathways, activation energies, and intermediate species in CO2 hydrogenation processes. These computational methods allow researchers to explore a wide range of catalyst materials and configurations, providing a wealth of information that can inform experimental designs.

MD simulations can model the dynamic behavior of catalysts at the atomic level, offering a complementary perspective to experimental observations. By simulating the interactions between catalyst particles and reactants over time, researchers can gain insights into the mechanisms of catalytic reactions, including how the structure of the catalyst influences its performance. These simulations generate large datasets that can be utilized in machine learning models to enhance predictive accuracy, bridging the gap between theoretical predictions and experimental results.

3.2. Data Sources

3.2.1. Publicly Available Databases

Several databases compile experimental and computational

data relevant to catalysis, providing researchers with valuable resources for catalyst discovery and optimization. For instance, Catalysis-Hub and the Materials Project are two prominent repositories that offer extensive datasets, including information on catalyst properties, reaction conditions, and performance metrics. These platforms not only facilitate data sharing among researchers but also promote collaboration within the scientific community, enabling the pooling of knowledge and resources to accelerate catalyst development.

Additionally, platforms like the Computational Materials Repository and the Catalytic Performance Database offer curated datasets that can be directly applied to machine learning models. By providing access to high-quality, standardized data, these databases help reduce redundancy in research efforts and streamline the process of identifying promising catalyst candidates.

3.2.2. Collaborative Research Datasets

Collaborative research initiatives often generate unique datasets that combine experimental and simulation results. These datasets can be more comprehensive and tailored for specific research questions, enhancing the relevance and applicability of the data. For example, the CO2 Utilization Collaborative focuses on developing datasets that encompass various aspects of CO2 hydrogenation, facilitating the application of machine learning techniques to predict catalyst performance. Such collaborations are instrumental in addressing complex challenges in catalysis, as they bring together diverse expertise and resources.

Collaborative datasets may also incorporate data from multiple research institutions, providing a broader perspective on catalyst performance across different experimental setups and conditions. This diversity can enrich the dataset, allowing for more robust machine learning models that can generalize well to new scenarios.

3.3. Data Cleaning and Preprocessing

3.3.1. Handling Missing Values and Outliers

Data cleaning is essential to ensure the integrity of the dataset used for machine learning applications. Missing values can arise from experimental limitations, data collection errors, or inconsistencies in reporting. Techniques such as imputation—where missing values are estimated based on the available data—or removal of incomplete records can be employed to address this issue. Care must be taken in choosing the appropriate method for handling missing data, as improper handling can introduce biases or distort the underlying patterns in the data.

Outliers, which may result from experimental anomalies or measurement errors, should be identified and treated appropriately to avoid skewing the model's predictions. Various statistical methods, such as Z-score analysis or the interquartile range method, can be applied to detect outliers. Once identified, outliers can be removed or transformed, depending on their nature and impact on the dataset.

Table 1. The list of descriptors used for the modelling study.

Input category	Descriptors	Units
Catalyst properties	Metal content (active phase)	wt%
	Covalent radius of metal*	pm
	Promoter 1 content	wt%
	Electronegativity of promoter 1*	–
	Promoter 2 content	wt%
	Electronegativity of promoter 2*	–
	Molecular weight of support	g mol^{-1}
	S_{BET}	m^2 g^{-1}
Synthesis conditions	Calcination temperature	K
	Calcination time	h
Reaction conditions	*T*	K
	P	MPa
	GHSV	cm^3 h^{-1} g_{cat}^{-1}
	H_2/CO_2	–

*The covalent radius of the metal and the electronegativity of the promoters were included as pseudo-variables to assign unique identifier to the catalyst families. For instance, covalent radius of 132 pm, would indicate Cu-based catalyst, while covalent radius of 142 pm implied In_2O_3-based catalyst.

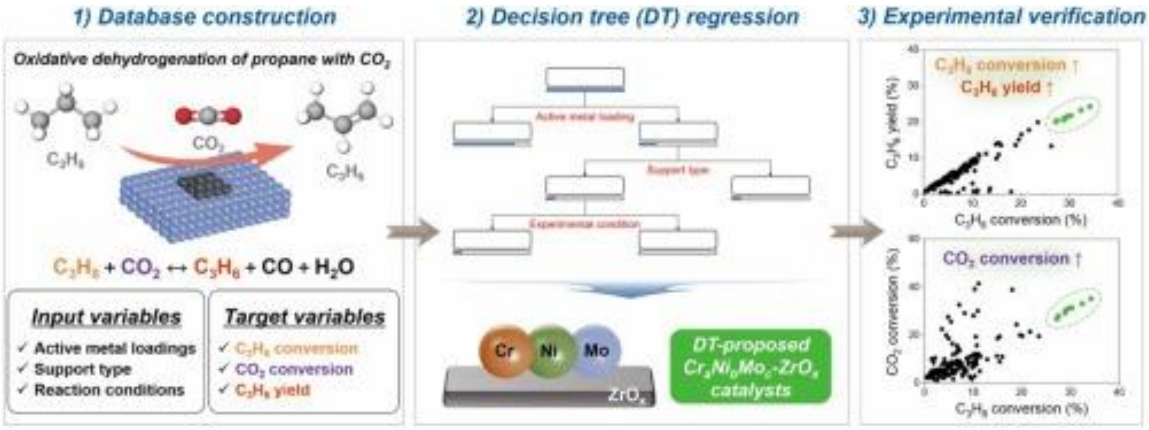

Figure 2. Data Cleaning and Preprocessing

3.3.2. Normalization and Feature Extraction Techniques

Normalization is a critical step in preparing data for machine learning models, as it ensures that features are on a comparable scale. This is important because features with larger ranges can disproportionately influence the model's performance, leading to biased predictions. Common normalization techniques include min-max scaling and z-score normalization, which transform the data to a standard range or distribution, respectively.

Feature extraction techniques can also enhance the quality of the dataset. In catalysis, molecular descriptors—such as electronic properties, geometric features, and chemical compositions—are valuable inputs for machine learning models. By calculating these descriptors, researchers can capture essential characteristics of the catalysts that may influence their performance in CO2 hydrogenation.

Moreover, dimensionality reduction techniques, such as principal component analysis, can help reduce the complexity

of the dataset while retaining the most relevant information. This not only improves computational efficiency but also aids in visualizing the data, allowing researchers to identify trends and relationships among different catalyst features.

4. Machine Learning Models for Catalyst Performance Prediction

4.1. Overview of Machine Learning Techniques

4.1.1. Supervised Learning

Supervised learning techniques are widely used for predicting catalyst performance based on labeled datasets. These techniques involve training a model on a dataset where the input features are known, along with the corresponding output labels. Regression algorithms, such as linear regression and support vector regression, can be employed to predict continuous outcomes like reaction rates. These models learn the relationship between the input features and the target variable, allowing for accurate predictions of catalyst performance under various conditions.

Classification algorithms, such as decision trees and random forests, can be used to categorize catalysts based on their performance metrics. For example, a classification model may predict whether a given catalyst will exhibit high or low selectivity for a specific product. The flexibility of supervised learning allows researchers to tailor their models to specific research questions, making it a powerful tool in catalyst discovery and optimization.

4.1.2. Unsupervised Learning

Unsupervised learning techniques, such as clustering and dimensionality reduction, can help identify patterns in the data without predefined labels. Clustering algorithms, like k-means and hierarchical clustering, can group similar catalysts based on their features, aiding in the identification of promising candidates for further study. For instance, clustering can reveal distinct groups of catalysts that share similar characteristics, allowing researchers to focus on specific clusters for targeted optimization efforts.

Dimensionality reduction techniques, such as PCA, can also be employed to visualize high-dimensional data in a lower-dimensional space, facilitating the identification of trends and correlations among different catalyst features. This approach can provide valuable insights into the relationships between catalyst properties and performance, guiding future experimental designs.

4.1.3. Reinforcement Learning Applications

Reinforcement learning has emerged as a novel approach for optimizing catalyst performance. In this paradigm, an agent learns to make decisions by interacting with an environment, receiving feedback in the form of rewards or penalties based on its actions. RL can be particularly useful in optimizing reaction conditions and catalyst formulations dynamically. For example, an RL agent could explore different combinations of temperature, pressure, and catalyst composition, learning from each trial to identify the optimal conditions for maximizing product yield.

The adaptability of RL makes it well-suited for complex catalytic systems, where traditional optimization methods may struggle to navigate the vast parameter space. By continuously learning from feedback, RL algorithms can refine their strategies over time, leading to improved catalyst performance and process efficiency.

4.2. Feature Selection and Engineering

4.2.1. Importance of Selecting Relevant Features

The success of machine learning models hinges on the selection of relevant features that significantly influence catalyst performance. Irrelevant or redundant features can lead to overfitting, where the model performs well on training data but poorly on unseen data. This is particularly problematic in the context of catalysis, where the relationships between features and performance can be complex and nonlinear. Therefore, feature selection techniques, such as recursive feature elimination and LASSO regression, can be employed to identify the most impactful features. By focusing on the most relevant features, researchers can improve model performance and enhance interpretability.

4.2.2. Techniques for Feature Engineering

Feature engineering involves creating new features from existing data to improve model performance. In catalysis, molecular descriptors such as electronic properties, geometric features, and chemical compositions can serve as valuable inputs for machine learning models. For example, features such as the electronegativity of metal atoms, coordination numbers, and surface area can provide insights into a catalyst's reactivity and selectivity.

Additionally, domain knowledge can guide the creation of features that capture the underlying chemistry of the catalytic processes. For instance, researchers may develop features that quantify the presence of specific active sites or the arrangement of atoms within a catalyst structure. This tailored approach to feature engineering can significantly enhance the predictive power of machine learning models, leading to more accurate and reliable predictions of catalyst performance.

4.3. Model Training and Validation

4.3.1. Splitting Datasets into Training, Validation, and Test Sets

To effectively evaluate the performance of machine learning models, it is crucial to systematically divide the dataset into three distinct subsets: training, validation, and test sets. The training set is the foundation of the model's learning process, as it is used to train the model by allowing it to learn the complex relationships between input features and output labels. This set typically comprises the majority of the data, providing the model with ample examples to identify patterns and make predictions.

The validation set serves a different yet equally important purpose. It is used to fine-tune hyperparameters, which are the settings that govern the learning process of the model. By evaluating the model's performance on the validation set, researchers can adjust these hyperparameters to optimize the model's performance and mitigate the risk of overfitting. Overfitting occurs when a model learns the training data too well, capturing noise and fluctuations rather than the underlying relationships, which can lead to poor generalization to new, unseen data.

Finally, the test set is a critical component of the model evaluation process. This set consists of data that has not been seen by the model during training or validation, providing an unbiased assessment of the model's performance. By evaluating the model on this separate test set, researchers can gauge its predictive capabilities in real-world scenarios, ensuring that the model is not only accurate but also robust and reliable. This systematic approach to dataset splitting is

essential for developing machine learning models that can accurately predict catalyst performance, paving the way for effective applications in CO2 hydrogenation processes and beyond.

4.3.2. Cross-Validation Techniques

Cross-validation techniques, particularly k-fold cross-validation, play a vital role in enhancing the robustness and reliability of machine learning models. In k-fold cross-validation, the dataset is divided into k equally sized folds or subsets. The model is then trained and validated k times, with each fold serving as the validation set once while the remaining k-1 folds are used for training. This iterative process allows for a comprehensive evaluation of the model's performance across different subsets of the data.

One of the primary advantages of k-fold cross-validation is its ability to mitigate the risk of overfitting. By training and validating the model on multiple subsets, researchers can ensure that the model's performance is consistent and not overly reliant on any single portion of the data. This technique provides a more reliable estimate of the model's predictive capabilities, as it assesses how well the model generalizes to various data distributions.

Furthermore, by averaging the performance metrics obtained from each fold, researchers gain a comprehensive understanding of how well the model is likely to perform on unseen data. This approach not only helps in identifying potential weaknesses in the model but also aids in selecting the most suitable model architecture and hyperparameters. Ultimately, the use of cross-validation techniques is a best practice in machine learning that significantly contributes to the development of robust models capable of accurately predicting catalyst performance in CO2 hydrogenation and other applications.

4.3.3. Performance Metrics

Evaluating model performance is a critical step in the machine learning workflow, requiring the selection of appropriate metrics tailored to the specific task at hand. For regression tasks, where the goal is to predict continuous outcomes, commonly used metrics include root mean square error and R-squared. RMSE quantifies the average deviation of predicted values from actual values, providing a clear indication of the model's accuracy. A lower RMSE value signifies better predictive performance, while R-squared offers insights into how well the model explains the variability in the data. It indicates the proportion of variance in the dependent variable that can be attributed to the independent variables, with higher values suggesting a better fit.

In contrast, for classification tasks, where the objective is to categorize data into discrete classes, a different set of performance metrics is employed. Metrics such as accuracy, precision, recall, and F1 score are commonly used to assess the model's classification capabilities. Accuracy measures the overall correctness of the model's predictions, providing a straightforward assessment of performance. However, in scenarios where class imbalances exist—such as when one class is significantly more prevalent than others—accuracy alone may be misleading.

To address this, precision and recall are utilized to provide more nuanced insights into the model's performance in identifying specific classes. Precision measures the proportion of true positive predictions relative to the total positive predictions made by the model, while recall assesses the proportion of true positives identified out of the total actual positives. The F1 score, which combines precision and recall into a single metric, serves as a balanced measure that accounts for both false positives and false negatives, making it particularly useful in imbalanced class scenarios.

In summary, the careful collection, preparation, and analysis of data are critical steps in leveraging machine learning for predicting catalyst performance in CO2 hydrogenation. By utilizing a combination of experimental and simulation data, employing rigorous data cleaning and preprocessing techniques, and selecting appropriate machine learning models and evaluation metrics, researchers can significantly enhance the accuracy and reliability of their predictions. This comprehensive approach not only accelerates the discovery of effective catalysts but also contributes to the broader goal of developing sustainable chemical processes for CO2 utilization. Through these methods, the research aims to facilitate advancements in catalysis and environmental sustainability, addressing the pressing challenges posed by climate change.

5. Conclusion

In this paper, we explored the potential of data-driven approaches, particularly machine learning, to predict catalyst performance in CO2 hydrogenation. This exploration is grounded in the urgent need to address climate change by reducing CO2 emissions and converting them into valuable products. The significance of CO2 hydrogenation as a transformative process that not only mitigates environmental impacts but also provides a pathway for sustainable chemical synthesis cannot be overstated. By adopting machine learning frameworks, we can harness the wealth of data generated from both experimental and computational studies to enhance our understanding of catalyst behavior and performance.

We highlighted the importance of large datasets derived from both experimental and simulation results, emphasizing the need for robust data collection and preparation methodologies. The quality and comprehensiveness of the data used in machine learning models are paramount, as they directly influence the accuracy and reliability of predictions. By utilizing diverse datasets that encompass various catalyst types, reaction conditions, and performance metrics, researchers can train models that are not only predictive but also generalizable across different systems.

Machine learning techniques, including supervised, unsupervised, and reinforcement learning, were discussed, showcasing their applicability in optimizing catalyst performance. Supervised learning methods have shown remarkable success in predicting reaction rates and selectivity based on labeled datasets, while unsupervised learning techniques can uncover hidden patterns in data that may not be immediately apparent. Reinforcement learning, although still in its nascent stages in the field of catalysis, presents exciting opportunities for dynamically optimizing reaction conditions and catalyst formulations based on real-time feedback.

The integration of data-driven approaches in catalyst research represents a paradigm shift in how catalysts are discovered and optimized. Traditional methods, which often relied heavily on trial-and-error experimentation and theoretical modeling, are now being complemented—and in some cases, replaced—by data-driven insights. This shift allows researchers to explore vast chemical spaces more efficiently, reducing the time and resources required to identify effective catalysts. By leveraging large datasets and

advanced machine learning techniques, researchers can accelerate the identification of promising catalysts and optimize reaction conditions, ultimately contributing to more sustainable chemical processes.

As the field of catalysis continues to evolve, we encourage researchers and practitioners to embrace data-driven methodologies and collaborate in sharing datasets to enhance the predictive capabilities of machine learning models. The collaborative sharing of data not only enriches the available resources but also fosters a collective intelligence that can drive innovation. Initiatives aimed at creating open-access databases for catalytic data can significantly enhance the research landscape, providing a foundation for future advancements.

Moreover, interdisciplinary collaboration between chemists, data scientists, and engineers is essential for maximizing the potential of machine learning in catalysis. By combining domain expertise with advanced computational techniques, researchers can develop more sophisticated models that account for the complexities of catalytic systems. This collaborative approach can lead to the discovery of novel catalysts and the optimization of existing ones, ultimately advancing the field toward practical applications.

In conclusion, the integration of machine learning and data-driven approaches into catalyst research is not merely a trend; it is a fundamental transformation that holds the promise of accelerating innovation in sustainable chemistry. By fostering a culture of collaboration and innovation, we can advance the development of effective catalysts for CO2 hydrogenation and other critical chemical transformations, paving the way for a more sustainable future. As we face the challenges of climate change and resource scarcity, the need for efficient and effective catalytic processes has never been more critical. Embracing these modern methodologies will enable us to tackle these challenges head-on, ensuring a cleaner and more sustainable planet for generations to come.

In summary, the potential for machine learning to revolutionize catalyst discovery and optimization is immense. By continuing to refine our methodologies, share our findings, and work together across disciplines, we can unlock new possibilities in the field of catalysis, ultimately contributing to a more sustainable and environmentally friendly chemical industry.

References

[1] Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. J Materiomics. 2017;3(3):159-177.

[2] Gao W, Chen Y, Li Y, et al. Automated catalyst design for CO2 electroreduction. Nature. 2022;610(7931):287-293.

[3] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. Nat Catal. 2018;1(9):696-703.

[4] Zhong M, Tran K, Min Y, et al. Accelerated discovery of CO2 electrocatalysts using active machine learning. Nature. 2020;581(7807):178-183.

[5] Schlexer Lamoureux P, Winther KT, Garrido Torres JA, et al. Machine Learning for Computational Heterogeneous Catalysis. ChemCatChem. 2019;11(16):3581-3601.

[6] Li Z, Wang S, Chin WS, et al. Machine learning-guided discovery and optimization of catalysts for CO2 electroreduction. Nat Catal. 2022;5(10):858-868.

[7] Asif M, Yao C, Zuo Z, Bilal M, Zeb H, Lee S, Kim T. Machine learning-driven catalyst design, synthesis and performance prediction for CO2 hydrogenation. Journal of Industrial and Engineering Chemistry. 2024 Sep 21.

[8] Gu GH, Noh J, Kim S, et al. Practical deep-learning representation for fast heterogeneous catalyst screening. NPJ Comput Mater. 2019;5(1):47.

[9] Jia X, Lynch A, Huang Y, et al. Anthropogenic CO2 reduction: from materials design to flow reactors. Mater Today. 2019; 24:15-47.

[10] Medford AJ, Kunz MR, Ewing SM, et al. Extracting knowledge from data through catalysis informatics. ACS Catal. 2018;8(8):7403-7429.

[11] Xie T, Grossman JC. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. Phys Rev Lett. 2018;120(14):145301.

[12] Grajciar L, Heard CJ, Bondarenko AA, et al. Towards operando computational modeling in heterogeneous catalysis. Chem Soc Rev. 2018;47(22):8307-8348.

[13] Ulissi ZW, Tang MT, Xiao J, et al. Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO2 Reduction. ACS Catal. 2017;7(10):6600-6608.

[14] Ma X, Li Z, Achenie LEK, Xin H. Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening. J Phys Chem Lett. 2015;6(18):3528-3533.

[15] Boes JR, Mamun O, Winther K, Bligaard T. Graph Theory Approach to High-Throughput Surface Adsorption Structure Generation. J Phys Chem A. 2019;123(11):2281-2285.

[16] Back S, Tran K, Ulissi ZW. Toward a Design of Active Oxygen Evolution Catalysts: Insights from Automated Density Functional Theory Calculations and Machine Learning. ACS Catal. 2019;9(9):7651-7659.

[17] Jinnouchi R, Asahi R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. J Phys Chem Lett. 2017;8(17):4279-4283.

[18] Li Z, Ma X, Xin H. Feature engineering of machine-learning chemisorption models for catalyst design. Catal Today. 2017; 280:232-238.

[19] Gu GH, Choi C, Lee Y, et al. Progress in computational and machine-learning methods for heterogeneous small-molecule activation. Nat Rev Mater. 2020;5(8):605-621.

[20] Gusarov S, Stoyanov SR, Siahrostami S, Kovalenko A. Deep Learning for Molecular Design and Discovery: From Chemical Reactions to Materials. ACS Cent Sci. 2021;7(6):902-915.

[21] Pegis ML, Roberts JAS, Wasylenko DJ, et al. Standard reduction potentials for oxygen and carbon dioxide couples in acetonitrile and N,N-dimethylformamide. Inorg Chem. 2015;54(24):11883-11888.

[22] Gasper R, Shi H, Ramasubramaniam A. Adsorption of CO2 on Pt-Cu Nanoalloys: Density Functional Theory and Machine Learning Approaches. J Phys Chem C. 2017;121(10):5612-5619.

[23] Zhang Y, Xu Z, Wang T, et al. Reinforcement learning-enabled autonomous optimization for CO2 hydrogenation. Nat Commun. 2020;11(1):5812.

[24] Kitchin JR. Machine learning in catalysis. Nat Catal. 2018;1(4):230-232.

[25] Wan Y, Zhang Z, Xu X, et al. Engineering active sites of polyoxometalate-based metal-organic frameworks for boosting CO2 electroreduction. Nat Commun. 2021;12(1):2870.

[26] Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic Chemical Design Using a Data-Driven Continuous

Representation of Molecules. ACS Cent Sci. 2018;4(2):268-276.

[27] Andersen M, Levchenko SV, Scheffler M, Reuter K. Beyond Scaling Relations for the Description of Catalytic Materials. ACS Catal. 2019;9(4):2752-2759.

[28] Rosen AS, Iyer SM, Ray D, et al. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. Matter. 2021;4(5):1578-1597.

[29] Guo Y, Li J, Alxneit I, et al. Machine Learning Assisted Engineering of Hydrogels with Physicochemical Properties for Biomedical Applications. Adv Sci. 2021;8(20):2100931.

[30] Artrith N, Butler KT, Coudert FX, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-508.

[31] Schleder GR, Padilha ACM, Acosta CM, et al. From DFT to machine learning: recent approaches to materials science–a review. J Phys Mater. 2019;2(3):032001.

[32] Melander M, Kuisma MJ, Christensen TE, Honkala K. Grand-canonical approach to density functional theory of electrocatalytic systems: Thermodynamics of solid-liquid interfaces at constant ion and electrode potentials. J Chem Phys. 2019;150(4):041706.

[33] Janet JP, Kulik HJ. Machine Learning in Inorganic Chemistry. J Phys Chem A. 2021;125(39):8561-8568.

[34] Cui C, Han J, Zhu X, et al. Promotional effect of surface hydroxyls on electrochemical reduction of CO2 over SnO2/Sn electrode. J Catal. 2016; 343:257-265.

[35] Tao H, Stach EA, Goodman ED, et al. Accelerating Catalysts Discovery through Machine Learning and High-Throughput Experimentation. ACS Catal. 2022;12(5):2962-2968.

[36] Sun T, Yang J, Li J, Chen J, Liu M, Fan L, Wang X. Enhancing Auto Insurance Risk Evaluation with Transformer and SHAP. IEEE Access. 2024 Aug 19.

[37] Lin Y, Fu H, Zhong Q, Zuo Z, Chen S, He Z, Zhang H. The influencing mechanism of the communities' built environment on residents' subjective well-being: A case study of Beijing. Land. 2024 Jun 4;13(6):793.

[38] Li J, Fan L, Wang X, Sun T, Zhou M. Product Demand Prediction with Spatial Graph Neural Networks. Applied Sciences. 2024 Aug 9;14(16):6989.

[39] Liu M, Ma Z, Li J, Wu YC, Wang X. Deep-Learning-Based Pre-training and Refined Tuning for Web Summarization Software. IEEE Access. 2024 Jul 4.

[40] Chen X, Liu M, Niu Y, Wang X, Wu YC. Deep-Learning-Based Lithium Battery Defect Detection via Cross-Domain Generalization. IEEE Access. 2024 Jun 3.

[41] Wang X, Wu YC, Ji X, Fu H. Algorithmic discrimination: examining its types and regulatory measures with emphasis on US legal practices. Frontiers in Artificial Intelligence. 2024 May 21; 7:1320277.

[42] Wang X, Wu YC, Ma Z. Blockchain in the courtroom: exploring its evidentiary significance and procedural implications in US judicial processes. Frontiers in Blockchain. 2024 Apr 12; 7:1306058.

[43] Ma Z, Chen X, Sun T, Wang X, Wu YC, Zhou M. Blockchain-Based Zero-Trust Supply Chain Security Integrated with Deep Reinforcement Learning for Inventory Optimization. Future Internet. 2024 May 10;16(5):163.

[44] Wang X, Wu YC, Zhou M, Fu H. Beyond surveillance: privacy, ethics, and regulations in face recognition technology. Frontiers in big data. 2024 Jul 3;7:1337465.

[45] Wang X, Wu YC. Balancing innovation and Regulation in the age of generative artificial intelligence. Journal of Information Policy. 2024 Jul 2;14.

[46] Zuo Z, Niu Y, Li J, Fu H, Zhou M. Machine Learning for Advanced Emission Monitoring and Reduction Strategies in Fossil Fuel Power Plants. Applied Sciences. 2024 Sep 19;14(18):8442.