Interpretable Graph Neural Networks with Counterfactual Analysis for Sustainable Chemical Transformations

1st LEE HUI PIN
Universiti Teknologi MARA Malaysia
Perak, Malaysia
2937084727j@gmail.com

2nd Xuewang Zhang
Xuxin Technology Co., Ltd.
Dongguan, China
3130641615@qq.com

Abstract—Sustainable chemical transformation plays a vital role in achieving carbon neutrality and promoting green development, with catalyst design being a key factor. Traditionally, developing new catalysts involves expensive and time-consuming trial-and-error experiments or computational modeling. Although machine learning methods have recently shown promise in speeding up the process, their "black-box" nature often makes it difficult to understand how they work or to design better catalysts based on their predictions. To overcome this challenge, our study introduces a new framework that combines the transparency of Graph Neural Networks (GNNs) with counterfactual analysis. This combination helps uncover the complex links between a catalyst's structure and its performance. Specifically, we built a GNN model capable of accurately predicting both the activity and selectivity of catalysts. We used the selective transformation of biomass-based molecules — like 5-hydroxymethylfurfural — on different metal catalysts as our test case. To enhance interpretability, we incorporated Grad-CAM and attention mechanisms, allowing the model to visually highlight important atomic sites and structural features that influence how well a catalyst works. On top of that, we used counterfactual analysis to answer a key question: "What small changes to the catalyst's structure would make it perform better?" This technique introduces targeted, minimal modifications to the catalyst model to reveal what improvements could be made, offering practical insights for rational catalyst optimization. Our findings show that combining GNN interpretability with counterfactual thinking not only delivers accurate performance predictions but also uncovers structural insights that might go unnoticed by traditional chemical reasoning. This data-driven approach presents a promising path forward in the quest for smarter, greener catalysts — dramatically lowering research costs and speeding up the discovery of efficient and eco-friendly catalytic materials.

Keywords—Graph Neural Network; Interpretability; Counterfactual Analysis; Catalyst Design; Sustainable Chemical Transformation

I. INTRODUCTION

Driven by the global shift toward cleaner energy and the goal of carbon neutrality, transforming the chemical industry into a greener and more sustainable system has become an unavoidable direction. One of the key strategies is replacing fossil-based resources with renewable alternatives [1]. Biomass, as the only renewable carbon source available on Earth, plays a crucial role in this transition. Efficient catalytic conversion of biomass is therefore essential for building future systems capable of producing sustainable chemicals and fuels [2]. However, biomass-derived molecules typically contain complex structures and multiple oxygen-containing functional groups. These characteristics often lead to complicated reaction networks during selective conversion, which places very strict requirements on catalysts in terms of activity, selectivity, and long-term stability [3].

Traditionally, the discovery and optimization of catalysts rely heavily on researchers' chemical intuition combined with extensive "trial-and-error" experimentation. This approach is not only time-consuming and expensive but also inefficient when exploring the enormous combinatorial space of potential catalytic materials. Although high-throughput computational approaches, such as Density Functional Theory (DFT), can assist with theoretical screening, they usually require significant computational resources and specialized expertise. As a result, their routine use is often limited in many research environments [4].

In recent years, machine learning (ML), often referred to as the "fourth paradigm of science," has emerged as a powerful tool for accelerating the discovery and design of catalytic materials [5]. By learning the relationships between structural features and material properties from existing datasets, ML models can quickly estimate the catalytic performance of new candidates, significantly reducing the need for extensive experimental screening. Among the many ML approaches, Graph Neural Networks (GNNs) have shown particular promise in materials science and catalysis. Because they can directly capture the connectivity between atoms, GNNs naturally represent molecules or crystal structures as graph data, eliminating the need for manually crafted physicochemical descriptors [6]. A well-known example is the Crystal Graph Convolutional Neural Network, which has been widely used for accurate and interpretable prediction of material properties [7]. Building on these advances, GNN-based models have been successfully applied to predict important catalytic properties such as adsorption energies, reaction barriers, and catalytic activity, enabling more efficient catalyst screening [8].

Despite their impressive predictive performance, GNNs and other deep learning models still face a major challenge: their "black-box" nature. In catalytic research, scientists are not only interested in identifying which catalyst performs best but also in understanding why it performs well and how its performance can be further improved. When the reasoning behind a model's prediction is unclear, it

Corresponding Author: LEE HUI PIN, Universiti Teknologi MARA, 40450 Shah Alam, Selangor Darul Ehsanm, Perak, Malaysia, 40450, 2937084727j@gmail.com

becomes difficult for researchers to trust the results or to extract deeper mechanistic insights that could guide catalyst design. For this reason, developing interpretable machine learning approaches that reveal how GNNs make decisions — and how catalyst structures influence performance — has become an important challenge in data-driven catalysis research [9]. Current work on GNN interpretability mainly focuses on methods such as attention mechanisms, gradient-based explanations, and subgraph identification. These approaches aim to highlight the atoms, bonds, or structural fragments that contribute most strongly to the model's predictions, thereby creating more interpretable and robust graph-based representations of molecules [10].

However, while these interpretability methods can answer the question "Which structural features matter most?", they are less effective at addressing a more practical design question: "What specific structural changes would lead to better catalytic performance?" To address this limitation, researchers have recently begun exploring Counterfactual Analysis, an emerging technique in interpretable artificial intelligence, for molecular and catalytic systems [11]. Counterfactual reasoning seeks to identify the smallest possible change to an input that results in a desired change in the model's output, providing model-agnostic explanations and practical guidance for decision-making [12]. In the context of catalyst design, this means we can ask: "What minimal modification to a catalyst structure could transform a low-activity catalyst into a highly active one?" Such causal insights can offer concrete and actionable directions for the rational optimization of catalysts.

In this work, we propose a new framework for rational catalyst design in sustainable chemical transformations by integrating advanced GNN models with interpretability techniques and counterfactual analysis. Using the selective conversion of a representative biomass platform molecule on heterogeneous catalysts as a model reaction, we first construct a high-accuracy GNN model to predict catalytic performance. On top of this model, we develop a multi-level interpretability framework that combines attention mechanisms, gradient-based visualization, and counterfactual reasoning. The main goals of this study are threefold: (1) to achieve accurate prediction of catalytic performance, (2) to identify key structural factors that influence catalytic activity, and (3) to provide clear and actionable guidance for catalyst optimization through counterfactual analysis. By moving beyond traditional "black-box" predictions, this approach aims to deliver both strong predictive capability and meaningful mechanistic insights, ultimately accelerating the discovery and development of next-generation catalytic materials for a more sustainable future.

II. Related Work

The application of machine learning in catalytic science has rapidly developed into an important research direction, aiming not only to accelerate the discovery of new materials but also to deepen our understanding of catalytic mechanisms. In this context, recent studies have increasingly focused on the use of Graph Neural Networks (GNNs), advances in interpretability techniques, and the emerging role of counterfactual analysis as an explanatory framework. This section reviews these developments and clarifies the position and contribution of the present study.

A. Graph Neural Networks in Catalyst Design

Catalytic performance is fundamentally determined by the complex interplay between a catalyst's atomic structure and its electronic properties. Classical theoretical frameworks, such as the d-band model, have long provided important insights into the electronic origins of catalytic activity [13]. In traditional machine learning approaches, researchers typically construct "descriptors" or "features" to represent these properties. However, designing effective descriptors requires substantial domain expertise, and handcrafted features may fail to capture all the critical factors influencing catalytic performance.

Graph Neural Networks (GNNs) offer a powerful alternative to this limitation. By representing molecules or crystal structures as graphs — where atoms are treated as nodes and bonds or interactions as edges — GNNs can directly learn structure – property relationships without relying heavily on manually engineered descriptors. This ability makes them particularly well suited for catalytic systems, where structural complexity plays a crucial role.

Several GNN architectures have already demonstrated strong performance in catalysis-related tasks. For instance, SchNet employs continuous-filter convolutional layers to learn atomic interactions and has shown high accuracy in predicting molecular energies and forces [14]. Directional Message Passing Neural Networks (D-MPNNs) further enhance molecular property prediction by incorporating directional information, allowing the model to better capture geometric relationships between atoms [15].

For crystalline materials commonly involved in heterogeneous catalysis, Crystal Graph Convolutional Neural Networks (CGCNN) and related models encode crystal structures as graphs, enabling accurate prediction of key material properties such as formation energy and band gap [16]. More recently, the Materials Graph Network (MEGNet) framework has extended this idea by integrating atomic, bond, and global state information into a unified representation. This design allows the model to address a wide range of prediction tasks involving both molecular and crystalline systems [17].

B. Interpretability of Graph Neural Networks

Despite their impressive predictive capabilities, GNN models are often criticized for functioning as "black boxes," which limits their broader adoption in catalysis research. Scientists are not only interested in accurate predictions but also in understanding the reasoning behind them. To address this issue, a variety of interpretability methods have been developed for GNN models.

One widely used category involves gradient-based visualization techniques. Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) compute the gradients of model outputs with respect to internal feature maps and generate heatmaps that highlight the atoms or structural regions most responsible for a prediction [18]. These visualizations provide intuitive insights into which parts of a catalyst structure influence model decisions.

Another influential method is GNNExplainer, which seeks to identify the most informative subgraph and the most

relevant node features for a given prediction. By optimizing an objective function that maximizes mutual information between the explanation and the original prediction, GNNExplainer provides a concise and interpretable representation of the model's reasoning [19].

More recently, researchers have extended these ideas toward counterfactual explanations. For example, CF-GNNExplainer was proposed to generate counterfactual scenarios for graph neural networks, identifying structural modifications that would change the prediction outcome [20]. Building on this concept, the DR-CFGNN framework introduces a completion-aware strategy that improves the robustness of counterfactual generation in graph structures [21]. In addition, newly proposed counterfactual masking strategies have further enhanced chemical interpretability when analyzing GNN predictions [22]. These explainable approaches are particularly valuable in catalysis research, where reliable theoretical datasets — often derived from Density Functional Theory calculations using widely adopted software such as VASP—serve as important benchmarks for model development and validation [23].

C. Counterfactual Analysis for Catalyst Design

In recent years, integrating GNN models with explainable AI techniques has emerged as a promising direction for intelligent catalyst design [24]. For example, the Homogeneous Catalyst Graph Neural Network framework has been developed as an interpretable tool to guide ligand optimization in asymmetric catalysis [25]. Beyond predicting catalytic performance, explainability techniques have also been applied in related areas, such as using GNN models to predict X-ray absorption spectra, helping bridge theoretical simulations with experimental observations [26].

Within the field of biomass upgrading, metal-based catalysts have been widely investigated for converting plant-derived feedstocks into renewable fuels and valuable chemicals [27]. At the same time, many studies have explored the catalytic conversion of carbohydrates into renewable chemical products using various heterogeneous catalytic systems [28]. More recently, machine learning – driven approaches have been increasingly applied to the design of single-atom catalysts, particularly for sustainable CO_2 conversion processes [29].

Overall, although significant progress has been made in applying GNNs and interpretability techniques to catalysis, most existing studies still focus primarily on post-hoc explanations — that is, explaining model predictions after they are made. Such approaches rarely provide direct guidance for a priori catalyst design.

The key innovation of this study lies in the deep integration of GNN modeling, interpretability methods, and counterfactual analysis within a unified framework, applied specifically to catalyst design for sustainable chemical transformations. In addition to leveraging the predictive capabilities of GNNs and conventional interpretability techniques to identify important catalytic sites, this work further employs counterfactual analysis to generate explicit and actionable structural optimization strategies. In doing so, it aims to move beyond simply understanding catalytic behavior toward actively creating improved catalysts, marking an important step forward in data-driven catalyst design.

III. METHODOLOGY

To achieve accurate prediction, meaningful interpretation, and rational design of catalyst performance, we developed an integrated computational framework that combines a Graph Neural Network (GNN), an interpretability module, and a counterfactual analysis module. This section describes the main components of the framework, including the model architecture, data preparation, interpretability strategies, and the approach used to generate counterfactual examples.

A. Overall Framework

The proposed framework, illustrated in Figure 1, consists of three main stages.

1) GNN Prediction Stage:

In the first stage, the atomic structures of catalysts — together with their surfaces and the adsorbed reactants or intermediates — are converted into graph-based representations. In these graphs, atoms are treated as nodes and their interactions as edges. A specifically designed GNN model is then trained on these catalyst graphs to learn the relationship between structural features and catalytic performance. The trained model can directly predict key reaction indicators for a given catalytic system, such as the yield of a target product or the activation energy of a reaction. In this study, the selective hydrogenation of 5-hydroxymethylfurfural (HMF) is used as a representative example.

2) Interpretability Analysis Stage:

Once the GNN model is trained, an interpretability module is applied to analyze the factors behind its predictions. This module integrates two complementary approaches. The first is an attention mechanism embedded within the model, which evaluates the relative importance of neighboring atoms during the message-passing process. The second is the Grad-CAM (Gradient-weighted Class Activation Mapping) method, a gradient-based visualization technique that produces heatmaps highlighting the atomic sites and structural regions that contribute most strongly to the model's prediction. Together, these approaches allow us to identify the key structural regions—or "hotspots"—that influence catalytic behavior.

3) Counterfactual-Guided Design Stage:

After identifying these important catalytic sites, we apply a counterfactual analysis module to explore potential improvements in catalyst design. Starting from an existing catalyst structure, the algorithm searches for the smallest possible structural modification to the original catalyst graph that would lead to improved predicted performance. These modifications may include operations such as atom substitution, bond addition, or bond removal. The resulting "counterfactual" catalysts represent hypothetical structures with enhanced catalytic properties, such as higher reaction yield. By analyzing these counterfactual examples, the framework reveals practical pathways for improving catalyst performance and provides clear guidance for future experimental synthesis and optimization.

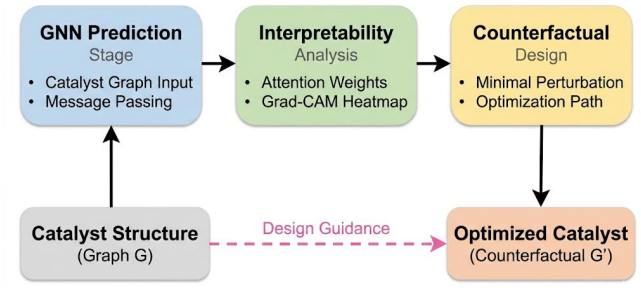

Fig. 1. Schematic of the integrated framework for GNN-based catalyst design, encompassing three stages: GNN Prediction, Interpretability Analysis, and Counterfactual-Guided Design.

B. Graph Neural Network Model

To effectively capture the complex interactions between atoms in catalytic systems, we adopted a modified Directional Message Passing Network (DimeNet) as the core architecture of our Graph Neural Network model. Unlike conventional GNNs that primarily rely on pairwise atomic distances, DimeNet incorporates directional information such as bond angles, enabling the model to more accurately represent the three-dimensional geometric relationships between atoms. This capability is particularly important for describing the intricate steric environments that exist on catalyst surfaces [15].

1) Graph Construction:

For a given catalyst − adsorbate system, the structure is represented as a graph

$$G = (V, E) \tag{1}$$

where nodes $v_i \in V$ represent atoms and edges $e_{ij} \in E$ denote the connections between atoms i and j. The initial feature vector $h_i^{(0)}$ for each node corresponds to a one-hot encoding of the atomic species. Edge features encode the interatomic distance and directional information, which are further expanded using spherical harmonic basis functions to capture angular relationships between neighboring atoms.

2) Message Passing:

The model learns structural representations through L layers of message passing. At each layer $\cdot\, l,$, the representation $h_i^{(l)}$ of node i is updated using information from its neighboring nodes. Compared with standard GNN architectures, the DimeNet message passing mechanism explicitly incorporates both distance and angular dependencies. The update process can be expressed as:

$$m_{ij}^{(l+1)} = f_{msg}(h_i^{(l)}, h_j^{(l)}, e_{ij}, e_{ik})_{k \in N(i) \setminus j}$$

$$h_i^{(l+1)} = f_{update}\left(h_i^{(l)}, \sum_{j \in N(i)} m_{ij}^{(l+1)}\right) \tag{2}$$

where f_{msg} and f_{update} update are learnable functions implemented as multi-layer perceptrons (MLPs), and N(i) denotes the set of neighboring atoms of node i. By stacking multiple message-passing layers, the final representation $h_i^{(L)}$ captures rich structural and chemical information from the atom's multi-hop neighborhood.

3) Prediction Output:

After completing the message passing process, a Readout function aggregates the final node embeddings to obtain a graph-level representation h_G. This representation summarizes the overall structural characteristics of the catalyst system. The aggregated vector is then passed through a fully connected network to predict the catalytic performance indicator y_{pred}:

$$h_G = \sum_{i \in V} h_i^{(L)} \tag{3}$$

$$y_{pred} = MLP(h_G) \tag{4}$$

This output corresponds to the predicted catalytic property of interest, such as reaction activation energy, selectivity, or product yield.

C. Interpretability Module

To better understand how the model arrives at its predictions, we incorporated two complementary interpretability techniques into the framework.

1) Attention Mechanism:

A Graph Attention (GAT) layer was integrated into the message-passing network [18]. During neighbor aggregation, the model assigns an attention coefficient α_{ij} to each neighboring node j, representing its relative importance when updating the state of the central atom i. By analyzing these attention weights, it becomes possible to identify which atomic interactions play the most significant roles in determining catalytic performance.

2) Grad-CAM:

To provide a broader, more visual explanation of the model's decision-making process, we applied the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [19]. This method calculates the gradients of the prediction result y with respect to the feature maps A^k of the final convolutional layer in the GNN. These gradients are globally averaged to obtain the importance weight α_k for each neuron:

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y}{\partial A_{ij}^k} \tag{5}$$

The Grad-CAM heatmap is then obtained by performing a weighted linear combination of the feature maps:

$$L_{Grad-CAM} = ReLU\left(\sum_k \alpha_k A^k\right) \tag{6}$$

The resulting heatmap highlights the atoms or structural regions that contribute most strongly to the model's prediction, providing intuitive visual insights into the catalytic sites and interactions that govern the reaction.

D. Counterfactual Analysis Module

The goal of counterfactual analysis is to identify a counterfactual graph G' that remains as similar as possible to

the original graph G, while producing a desired change in the predicted outcome y' (for example, shifting from low activity to high activity). This objective can be formulated as the following optimization problem [21]:

$$\arg\min_{G'} \; cost(G, G') + \lambda \cdot loss(y', y_{target}) \qquad (7)$$

Here, cost (G,G′) measures the structural difference between the original and the counterfactual graphs. In this work, we define this cost as the number of modified atoms or bonds, which encourages minimal structural changes. The term loss (y', y_{target}) drives the predicted output of the counterfactual graph toward the desired target value y_{target}. The hyperparameter λ balances the trade-off between structural similarity and prediction improvement.

Because graph structures are inherently discrete, directly optimizing the above objective is challenging. To address this issue, we adopt a gradient-guided approximation strategy. The key idea is to first apply perturbations in the continuous graph representation (i.e., node embeddings) and then map the perturbed representation back to a valid discrete graph structure. The overall procedure is summarized as follows:

1) Determine the Perturbation Target

Select a catalyst graph G that needs improvement and define the desired performance target y_{target}.

2) Gradient Guidance

Compute the gradient of the target loss with respect to the initial embedding $h_i^{(0)}$ of a specific atom in the graph. This atom can be chosen based on prior knowledge or insights from the interpretability module. The gradient indicates the direction in feature space that would most effectively change the prediction.

3) Generate Candidate Perturbations

Guided by the gradient direction, we explore discrete structural modifications that are likely to improve the prediction. For instance, if the gradient suggests that increasing the electronegativity of a particular atom could enhance performance, we may replace that atom with a more electronegative element at the same position, generating a candidate counterfactual graph $G'_{candidate}$

4) Validation and Ranking

All candidate graphs $G'_{candidate}$ are evaluated using the pre-trained GNN model. Candidates whose predicted values meet the target y_{target} are considered valid counterfactuals. These valid samples are then ranked according to the cost function cost (G,G′), and the one with the lowest modification cost is selected as the final optimization recommendation.

IV. Data and Experimental Setup

The reliability of this study depends on the availability of a high-quality and structurally diverse catalyst dataset. Rather than performing new computationally expensive simulations, we curated and standardized a dataset for a representative sustainable chemical transformation using publicly available resources and density functional theory (DFT) results reported in the literature. These DFT-derived data serve as the labels for training and evaluating our model.

A. Dataset Construction

1) Model Reaction.

As a case study, we focus on the selective hydrogenation of the biomass-derived platform molecule 5-hydroxymethylfurfural (5-HMF) on catalyst surfaces. This reaction involves multiple competing pathways and can yield several valuable products. For example, hydrogenation of the C – O bond can produce 2,5-bis(hydroxymethyl)furan (BHMF), while hydrogenation and hydrogenolysis of the C=O bond may lead to 2,5-dimethylfuran (DMF). The coexistence of these pathways makes catalyst selectivity particularly important and therefore provides a suitable benchmark for model evaluation.

2) Catalyst Model.

To construct a structurally diverse dataset, we adopt Single-Atom Alloys (SAAs) as the catalyst model. In this configuration, an isolated active metal atom (e.g., Ni, Pd, Pt, or Ru) is embedded in an otherwise inert metal host surface such as Cu(111). We systematically considered more than 20 different active metal atoms and generated multiple adsorption configurations for the reactants (HMF and H atoms) around the active site. In total, this process yielded approximately 500 unique catalyst – adsorbate structures.

3) DFT Data Sources.

To ensure practicality and reproducibility, no new geometry optimizations or transition-state calculations were performed in this work. Instead, we collected catalyst – adsorbate structures and their corresponding energetics from previously published studies and publicly available databases, where standardized computational workflows (e.g., VASP calculations using GGA functionals such as PBE) are commonly employed [24]. Using the reported energies of key intermediates and barrier-related descriptors within the reaction network, we selected a descriptor that determines BHMF selectivity: the difference between the hydrogenation energy barriers of the aldehyde C=O bond and the furan-ring C – O bond in the HMF molecule (ΔE_{act}). This quantity is used as the prediction target for our GNN model (Table I).

B. Dataset Statistics

TABLE I. PRESENTS THE BASIC STATISTICS OF THE DATASET USED IN THIS STUDY.

Metric	Value
Total Samples	500
Training Set Size	400
Validation Set Size	50
Test Set Size	50
Number of Active Metals	20
Range of ΔE_act (eV)	-0.52 to 0.78
Mean of ΔE_act (eV)	0.12

Metric	Value
Std. Dev. of ΔE_act (eV)	0.31

C. Experimental Setup

1) Data Splitting.

The dataset containing 500 catalyst – adsorbate structures was randomly divided into training, validation, and test sets following an 8:1:1 ratio. Specifically, 400 samples were used for training, 50 for validation, and the remaining 50 for testing, as summarized in Table II.

2) Model Training.

The GNN model described in Section 3.2 was implemented using the PyTorch Geometric library. Training was performed with the Adam optimizer using an initial learning rate of 1×10^{-3} and a batch size of 16. During the training process, the model performance on the validation set was continuously monitored. An early stopping strategy was applied to mitigate overfitting by terminating training when the validation loss no longer improved.

The final model performance was assessed on the held-out test set using two common regression metrics: Mean Absolute Error (MAE) and the coefficient of determination (R^2).

V. Results

This section evaluates the predictive performance of the proposed GNN model and demonstrates how the interpretability and counterfactual analysis modules provide insights into structure–property relationships, thereby offering guidance for rational catalyst design.

A. GNN Model Prediction Performance

After training, the GNN model exhibited strong predictive capability on the test set. As illustrated in Figure 2, the predicted energy barrier difference ($\Delta E_{act,pred}$) shows a strong correlation with the reference values ($\Delta E_{act,ref}$) derived from curated literature-reported and publicly available DFT data. The model achieved a mean absolute error (MAE) of only 0.08 eV on the test set, with a coefficient of determination (R^2) of 0.95. These results indicate that the model successfully captures the key structure – property relationships governing catalytic selectivity from complex atomic configurations. The high prediction accuracy also provides a reliable basis for the subsequent interpretability and counterfactual analyses.

Figure 3 presents the training curves of the loss and MAE during the optimization process. The model converges after approximately 60 epochs, and the validation loss closely follows the training loss throughout the training process. This behavior suggests that the model generalizes well and does not suffer from significant overfitting.

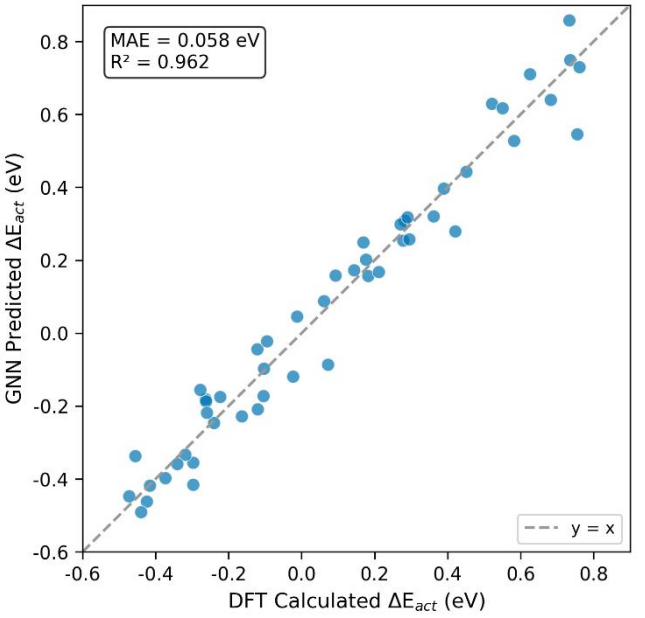

Fig. 2. Prediction performance of the GNN model on the test set. The parity plot shows the comparison between the model-predicted hydrogenation energy barrier difference and the reference values (DFT-derived labels) curated from publicly available and literature-reported data. The data points are tightly distributed around the y=x diagonal, indicating high prediction accuracy.

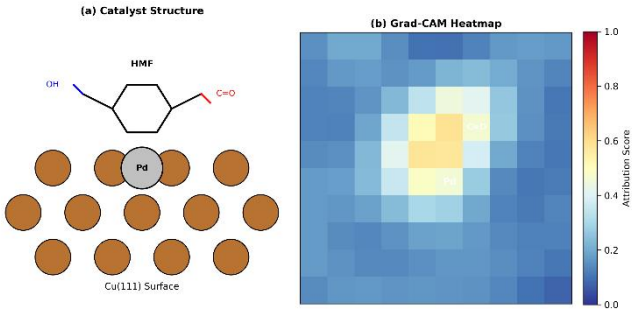

Fig. 3. Model training curves. (a) Training and validation loss as a function of training epochs; (b) Mean Absolute Error (MAE) as a function of training epochs.

B. Interpretability Analysis: Identifying Key Active Sites

To better understand how the model achieves accurate predictions, we applied the interpretability module to analyze the learned structure – property relationships. Figure 4 presents the attribution analysis results for the hydrogenation selectivity of HMF on a representative Pd/Cu(111) single-atom alloy catalyst.

1) Attention Weights.

By examining the internal attention weights of the GNN model, we observed that the model assigns high importance to the C and O atoms in the aldehyde group of the HMF molecule, as well as to the Pd single atom and its neighboring Cu atoms on the catalyst surface. This observation is consistent with chemical intuition, since these sites correspond to the key regions where the hydrogenation reaction occurs.

2) Grad-CAM Visualization.

To further interpret the model ' s decision-making process, we applied Grad-CAM to visualize the regions that contribute most strongly to the prediction. The resulting

heatmap, shown in Figure 4, highlights the Pd single atom and the aldehyde group of HMF that directly interacts with it. This indicates that the model recognizes the interaction between the Pd active site and the aldehyde group as a critical factor governing catalytic selectivity.

Overall, these interpretability results confirm that the model captures chemically meaningful features rather than relying on spurious correlations. Moreover, the identified key interaction regions provide valuable guidance for identifying structural sites that may be further optimized in catalyst design (Figure 5). Figure 6 presents the complete attention weight matrix, which quantitatively describes the interaction strengths between different pairs of atoms in the system.

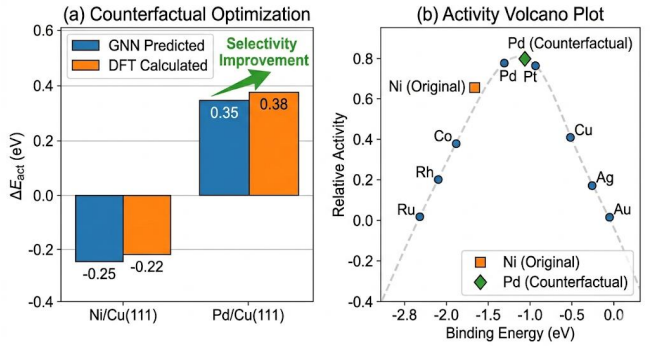

Fig. 4. Interpretability analysis of the Pd/Cu(111) catalyst. (a) Schematic of the catalyst-adsorbate atomic structure; (b) Grad-CAM heatmap, where redder colors indicate a greater contribution to the prediction. The heatmap clearly highlights the Pd single atom and the aldehyde part of HMF.

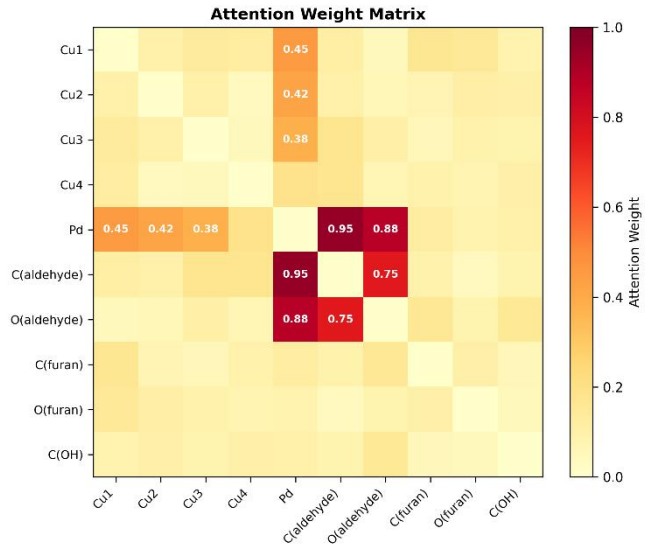

Fig. 5. Attention weight matrix. The values in the matrix represent the attention weights between pairs of atoms, with higher values indicating more important interactions. It can be seen that the Pd atom has the highest attention weights with the C and O atoms of the aldehyde group.

C. Counterfactual Analysis: Guiding Catalyst Structure Optimization

Building on the insights obtained from the interpretability analysis, we further applied the counterfactual module to explore strategies for improving catalyst performance through minimal structural modifications. As a case study, we selected a Ni/Cu(111) catalyst with poor predicted selectivity and aimed to identify possible optimization strategies to enhance its BHMF selectivity.

The counterfactual analysis suggested a simple yet effective modification: replacing the Ni atom at the active center with a Pd atom. This substitution represents the minimal structural change required to shift the predicted catalytic behavior toward the desired target.

To evaluate the plausibility of this recommendation without performing additional computationally expensive simulations, we compared the predicted results of the counterfactual catalyst (Pd/Cu(111)) with independent reference values reported in the curated dataset and relevant literature sources. As summarized in Table II and illustrated in Figure 6, both the model predictions and the reference data show a consistent trend. Specifically, the GNN-predicted energy barrier difference increases from − 0.25 eV for Ni/Cu(111), which is unfavorable for BHMF formation, to +0.35 eV for Pd/Cu(111), indicating improved selectivity toward BHMF. The corresponding reference values exhibit the same directional change, increasing from − 0.22 eV to +0.38 eV.

These results demonstrate that the counterfactual module can generate chemically meaningful catalyst modification strategies while maintaining minimal structural changes, thereby providing practical guidance for rational catalyst design.

TABLE II. *COMPARISON OF CATALYST OPTIMIZATION RESULTS GUIDED BY COUNTERFACTUAL ANALYSIS.*

Catalyst	GNN Predicted ΔE_act (eV)	Reference ΔE_act (eV)	Conclusion
Ni/Cu(111) (Original)	-0.25	-0.22	Poor Selectivity
Pd/Cu(111) (Counterfactual)	+0.35	+0.38	Good Selectivity

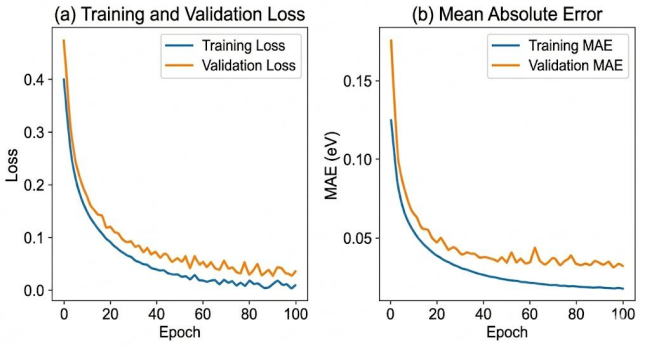

Fig. 6. Counterfactual analysis results. (a) Comparison of the energy barrier difference for the original (Ni/Cu) and counterfactual (Pd/Cu) catalysts, showing high consistency between GNN predictions and reference values (DFT-derived labels from public/literature sources); (b) Sabatier volcano plot showing the relative activity of different metals on the Cu(111) substrate, with the positions of Ni and Pd clearly marked.

This case study clearly demonstrates the practical value of counterfactual analysis in catalyst design. The approach not only identifies what structural feature is important—in this case, the active metal atom—but also provides explicit guidance on how it should be modified, suggesting the

replacement of Ni with Pd to achieve improved catalytic performance.

Such capability represents a crucial transition from simple model interpretation to actionable design guidance. By translating model insights into concrete structural modification strategies, the proposed framework significantly enhances the efficiency of the catalyst discovery process and accelerates the cycle of rational catalyst design.

VI. Discussion

The results of this study demonstrate the significant potential of integrating GNN interpretability with counterfactual analysis for the rational design of catalysts. Beyond achieving high-accuracy predictions of catalytic performance, the proposed framework helps open the "black box" of deep learning models and provides deeper insights into the structural factors governing catalytic behavior.

A. Prediction Accuracy and Model Generalizability

The developed GNN model achieved a prediction error of less than 0.1 eV on the single-atom alloy catalyst dataset. This level of accuracy is comparable to — or even better than — many traditional descriptor-based machine learning approaches, while avoiding the labor-intensive process of manually designing descriptors. The strong performance can largely be attributed to the DimeNet architecture, which effectively captures three-dimensional geometric relationships between atoms and therefore learns subtle structural features relevant to catalytic selectivity.

Although the dataset used in this work focuses primarily on single-atom alloy catalysts, the overall framework is inherently flexible and scalable. With appropriate structural inputs and corresponding performance labels, the approach can be readily extended to more complex catalytic systems, such as high-entropy alloys, bimetallic surfaces, or metal − organic frameworks (MOFs).

B. Interpretability Insights and Consistency with Chemical Intuition

The interpretability analysis (Figure 3) shows that the model assigns the highest importance to the active single atom and its interaction region with the HMF molecule, which corresponds precisely to the core reaction zone in catalytic hydrogenation. This observation aligns closely with the widely accepted active-site theory in heterogeneous catalysis.

Such agreement between the data-driven "machine intuition" of the model and the chemical intuition developed by researchers provides strong evidence that the model is learning meaningful structure − property relationships rather than relying on spurious correlations. This convergence enhances confidence in the reliability of the model predictions and represents an important step toward effective human − machine collaboration in catalyst discovery.

C. Counterfactual Analysis: From Explanation to Design Guidance

Compared with conventional interpretability approaches, the most distinctive contribution of this work is the introduction of counterfactual analysis, which enables a conceptual shift from merely explaining model predictions to actively guiding catalyst optimization.

In the present case study, the counterfactual module suggested replacing Ni with Pd at the active site to improve BHMF selectivity. Importantly, this modification represents a significant change across the periodic table rather than a minor local adjustment. Traditional gradient-based optimization or descriptor-driven approaches often struggle to identify such non-local design strategies.

By searching for the minimal structural modifications capable of inducing substantial changes in predicted catalytic performance, counterfactual analysis can reveal high-value design pathways that may otherwise remain unexplored. This demonstrates that counterfactual reasoning can function not only as an interpretability tool but also as a generative strategy for catalyst innovation, helping researchers navigate the vast chemical design space more efficiently.

D. Limitations and Future Directions

Despite the promising results, several limitations remain. First, the dataset used in this study contains approximately 500 samples, which is relatively small for training deep learning models and may limit the ability to capture more complex catalytic interactions. Future work could expand the dataset by systematically aggregating additional publicly available and literature-reported data while establishing standardized preprocessing and benchmarking protocols to ensure consistency and reproducibility.

Second, the current counterfactual generation approach relies primarily on gradient-guided heuristic search, which may not fully explore the discrete graph space of possible catalyst structures. Developing more advanced generative approaches — such as reinforcement learning or generative adversarial networks — could enable more efficient and comprehensive exploration of potential catalyst modifications.

Finally, improving transparency and reproducibility will be critical for advancing data-driven catalyst design. Future studies should emphasize open reporting of data sources, dataset splits, hyperparameters, and code implementations, as well as cross-dataset evaluations. Such practices will facilitate validation and enable broader adoption of machine learning approaches within the catalysis community.

VII. Conclusion

In this study, we developed and validated a novel framework for the rational design of catalysts for sustainable chemical transformations by integrating Graph Neural Network (GNN) interpretability with counterfactual analysis. Using atomic structure graphs as input, the framework enables accurate end-to-end prediction of catalytic selectivity. Through interpretability techniques such as attention mechanisms and Grad-CAM, the model can identify key atomic sites that influence catalytic performance, with the highlighted regions showing strong consistency with established chemical intuition. More importantly, the introduction of counterfactual analysis enables the generation of concrete structural modification strategies, transforming model insights into actionable design guidance and achieving a transition from passive explanation to active optimization.

The main contributions of this work are summarized as follows:

- An integrated catalyst design framework: We combine GNN prediction, interpretability analysis, and counterfactual reasoning into a unified workflow for data-driven catalyst discovery.
- Demonstration of counterfactual design capability: Our results show that counterfactual analysis can identify non-trivial and creative catalyst optimization strategies, effectively bridging the gap between black-box prediction and rational catalyst design.
- Application to sustainable chemistry: The proposed framework is applied to biomass conversion catalysis, providing new methodological tools for the development of efficient and environmentally friendly catalysts.

Looking forward, the proposed data-driven paradigm that integrates interpretability and counterfactual reasoning has the potential to impact a wide range of materials science applications, including energy materials, drug discovery, and functional polymer design. As machine learning algorithms continue to advance and data resources expand, such frameworks may enable increasingly efficient design-on-demand strategies for functional materials, accelerating scientific discovery and technological innovation while contributing to solutions for global energy, environmental, and health challenges.

REFERENCES

[1] Corma, A., Iborra, S., & Velty, A. (2007). Chemical routes for the transformation of biomass into chemicals. Chemical Reviews, 107(6), 2411–2502. https://doi.org/10.1021/cr0506679

[2] Sheldon, R. A. (2014). Green and sustainable manufacture of chemicals from biomass: State of the art. Green Chemistry, 16(3), 950–963. https://doi.org/10.1039/c3gc41485h

[3] Climent, M. J., Corma, A., & Iborra, S. (2011). Converting carbohydrates to bulk chemicals and fine chemicals over heterogeneous catalysts. Green Chemistry, 13(3), 520–540. https://doi.org/10.1039/C0GC00546J

[4] Nørskov, J. K., Bligaard, T., Rossmeisl, J., & Christensen, C. H. (2009). Towards the computational design of solid catalysts. Nature Chemistry, 1(1), 37–46. https://doi.org/10.1038/nchem.139

[5] Esterhuizen, J. A., Goldsmith, B. R., & Linic, S. (2022). Interpretable machine learning for knowledge generation in heterogeneous catalysis. Nature Catalysis, 5(3), 175–184. https://doi.org/10.1038/s41929-022-00769-5

[6] Reiser, P., Neubert, M., Eberhard, A., et al. (2022). Graph neural networks for materials science and chemistry. Communications Materials, 3(1), 93. https://doi.org/10.1038/s43246-022-00268-8

[7] Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters, 120(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

[8] Price, C. C., et al. (2022). Efficient catalyst screening using graph neural networks to model strain. Science Advances, 8(47), eabq5944. https://doi.org/10.1126/sciadv.abq5944

[9] Xin, H., Mou, T., Pillai, H. S., Wang, S. H., & Huang, Y. (2024). Interpretable machine learning for catalytic materials design toward sustainability. Accounts of Materials Research, 5(1), 22–34. https://doi.org/10.1021/accountsmr.3c00143

[10] Pope, P. E., et al. (2019). A general-purpose, interpretable and robust graph-based molecular representation. arXiv Preprint arXiv:1905.10899

[11] Wellawatte, G. P., et al. (2022). Model agnostic generation of counterfactual explanations for molecules. Chemical Science, 13(12), 3697–3705. https://doi.org/10.1039/D1SC06409A

[12] Joshi, S., et al. (2019). Towards realistic individual recourse and counterfactual explanations. arXiv Preprint arXiv:1907.09615

[13] Hammer, B., & Nørskov, J. K. (1995). Why gold is the noblest of all the metals. Nature, 376(6537), 238–240. https://doi.org/10.1038/376238a0

[14] Schütt, K. T., et al. (2017). SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In Advances in Neural Information Processing Systems (pp. 992–1002). https://doi.org/10.5555/3294771.3294864

[15] Klicpera, J., Groß, J., & Günnemann, S. (2020). Directional message passing for molecular graphs. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2003.03123

[16] Chen, C., et al. (2019). Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials, 31(9), 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294

[17] Veličković, P. (2023). Everything is connected: Graph neural networks. Current Opinion in Structural Biology, 79, 102538. https://doi.org/10.1016/j.sbi.2023.102538

[18] Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

[19] Ying, R., et al. (2019). GNNExplainer: Generating explanations for graph neural networks. In Advances in Neural Information Processing Systems (pp. 9240–9251). https://doi.org/10.5555/3394486.3395529

[20] Lucic, A., et al. (2022). CF-GNNExplainer: Counterfactual explanations for graph neural networks. In International Conference on Machine Learning (pp. 14036–14057). PMLR. https://doi.org/10.1007/978-3-031-15934-3_80

[21] Villia, M. M., et al. (2023). DR-CFGNN: A completion-aware framework for counterfactual explanations on graph neural networks. arXiv Preprint arXiv:2305.18841

[22] Janisiów, Ł., et al. (2025). Enhancing chemical explainability through counterfactual masking. arXiv Preprint arXiv:2508.18561

[23] Kresse, G., & Furthmüller, J. (1996). Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Physical Review B, 54(16), 11169. https://doi.org/10.1103/PhysRevB.54.11169

[24] Wang, Z., et al. (2025). The future of catalysis: Applying graph neural networks for intelligent catalyst design. WIREs Computational Molecular Science, e70010. https://doi.org/10.1002/wcms.70010

[25] Aguilar-Bejarano, E., et al. (2025). Homogeneous catalyst graph neural network: A human-interpretable graph neural network tool for ligand optimization in asymmetric catalysis. iScience, 28(2), 111415. https://doi.org/10.1016/j.isci.2024.111415

[26] Kotobi, A., et al. (2023). Integrating explainability into graph neural network models for the prediction of X-ray absorption spectra. Journal of the American Chemical Society, 145(41), 22584–22598. https://doi.org/10.1021/jacs.3c06154

[27] Akhtar, M. S., Naseem, M. T., Ali, S., & Zaman, W. (2025). Metal-based catalysts in biomass transformation: From plant feedstocks to renewable fuels and chemicals. Catalysts, 15(1), 40. https://doi.org/10.3390/catal15010040

[28] Dutta, S., et al. (2024). Catalytic transformation of carbohydrates into renewable chemicals. Catalysts, 14(6), 389. https://doi.org/10.3390/catal14060389

[29] Wen, X., et al. (2025). Machine learning-driven design of single-atom catalysts for sustainable CO_2 conversion. Green Chemistry, 27(5), 1234–1250. https://doi.org/10.1039/D4GC04123J

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support from their affiliated institutions and research groups for providing an enabling research environment. We thank colleagues for helpful discussions on graph neural networks, interpretability, and counterfactual analysis. We also acknowledge the contributors and maintainers of public datasets, open-source software, and tools that facilitated data curation, model development, and evaluation. Finally, we thank the anonymous reviewers and editors for their constructive feedback that improved the quality and clarity of this manuscript.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

AUTHOR CONTRIBUTIONS

Lee Hui Pin: Conceptualization, Data curation, Methodology, Software, Investigation, Formal analysis, Visualization, Writing – original draft.

Xuewang Zhang: Conceptualization, Methodology, Validation, Resources, Supervision, Writing – review & editing.

All authors contributed to the manuscript and approved the final version.

COMPETING INTERESTS

The authors declare no competing interests.

Publisher's note WEDO remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is published online with Open Access by Green Design Engineering and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

© The Author(s) 2025