Interpretable Generative Model-Driven Inverse Design of Catalytic Materials: From Candidate Generation to Mechanism Verification

1st Hang Wu
Zhongkai University of Agriculture and Engineering
Guangzhou, China
919345832@qq.com

2nd Caiying Li
Dai Chen Brand Design Co., Ltd.
Guangzhou, China
2302838043@qq.com

3rd Peiwei Xiao
Shenzhen Donghaoxing Technology Co., Ltd.
Shenzhen, China
464633194@qq.com

***Abstract*—The discovery and design of catalytic materials are central challenges in achieving sustainable chemical production and energy transition. Traditional trial-and-error experimentation and forward computational screening methods, based on existing structural prototypes, face inherent limitations in efficiently and innovatively exploring the vast chemical space. Inverse design, which generates novel material structures directly from target properties, offers a revolutionary approach to overcome this bottleneck. However, its development has long been hampered by the "black-box" nature of generative models and a disconnection from physical mechanism validation. This paper proposes a new framework named "Interpretable Generative Model-Driven Inverse Design" (IGMD), which deeply integrates a Conditional Variational Autoencoder (cVAE) for target-oriented candidate structure generation, SHAP (SHapley Additive exPlanations)-based interpretability analysis for extracting physically meaningful design rules, and Density Functional Theory (DFT) calculations for high-fidelity catalytic mechanism verification. Applied to the CO_2 reduction reaction (CO_2 RR), the IGMD framework generated a large number of novel alloy catalyst candidates, among which the top 100 candidates were selected for high-fidelity DFT verification, leading to a notable increase in the proportion of high-potential catalytic materials compared to the original training set. Three novel bimetallic catalysts (e.g., Cu-Al, Ag-In) screened and synthesized through this framework exhibited excellent catalytic performance in experimental tests, with the Ag-In catalyst achieving a Faradaic efficiency exceeding 92% for the target product CO. This study not only demonstrates the immense potential of the IGMD framework in accelerating the discovery of high-performance catalytic materials but, more importantly, establishes a complete closed loop from data-driven candidate generation and interpretable physical law mining to first-principles mechanism verification, providing a new paradigm for automated and trustworthy catalytic material design.**

***Keywords*— Interpretable Machine Learning, Generative Models, Inverse Design, Electrocatalysis, CO_2 Reduction**

I. Introduction

The advancement of catalysis science, as the cornerstone of the modern chemical industry, profoundly impacts the development of critical fields such as energy, environment, and materials [1]. Particularly against the backdrop of global climate change and the energy crisis, developing novel catalytic materials for efficient energy conversion (e.g., water electrolysis, fuel cells) and environmental remediation (e.g., CO_2 reduction, NOx elimination) has become a forefront and core scientific pursuit [2]. However, the discovery of catalytic materials has long relied heavily on researchers' chemical intuition and extensive "trial-and-error" experiments, a process that is not only time-consuming and labor-intensive but also has a very low success rate. With the rapid development of computational power, high-throughput computational screening based on Density Functional Theory (DFT) has made it possible to guide experiments theoretically. Nevertheless, it remains a "forward" design method in essence: starting from known material structures, calculating their properties, and then screening them. Although this method has improved efficiency, its exploration scope is confined to known structural prototypes or derivatives generated by simple element substitution, making it difficult to discover materials with entirely new structures and breakthrough performance [3].

To fundamentally break through this limitation, the materials science community has turned its attention to "inverse design"—a paradigm diametrically opposed to traditional "forward" design. The core idea of inverse design is to first define the desired target properties and then deduce the material structures that satisfy these performance requirements [4]. The rise of machine learning, especially deep learning, has brought dawn to achieving this grand goal. In recent years, generative models represented by Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models have shown astonishing potential in learning the intrinsic distribution patterns of existing material databases and creating novel, effective material structures [5, 6]. These models can perform optimization directly in a high-dimensional latent space learned from data, thereby generating target-oriented candidate materials and greatly expanding the boundaries of material exploration. For instance, studies have successfully applied generative models to design zeolites with specific pore structures [7], highly stable perovskites [8], and Metal-Organic Frameworks (MOFs) with specific adsorption properties [9].

Despite the encouraging progress of generative models in material inverse design, their application still faces two core challenges. The first is the "black-box" problem. The complex internal structure of deep learning models makes

Corresponding Author:Hang Wu, No. 501, Zhongkai Road, Haizhu District, Guangzhou, China, 510225, 919345832@qq.com their decision-making process opaque, making it difficult for researchers to understand why a model generates a specific structure or to extract universal physical or chemical design rules from the generation results. This lack of interpretability severely undermines the model's credibility and hinders its deep integration with domain knowledge [10]. The second is the problem of physical reality verification. When a generative model learns a data distribution, it is essentially a statistical simulation. The generated structures may not be physically or chemically stable, or their predicted superior performance may contradict the actual catalytic mechanism. Without a rigorous physical mechanism verification step, inverse design could devolve into a "digital game," with results that are difficult to guide real material synthesis and application [11]. Therefore, building a bridge between the powerful creativity of generative models and the rigorous laws of the physical world has become a key scientific problem to be solved in the field of catalytic material inverse design.

In response to the above challenges, this paper aims to construct a closed-loop catalytic material inverse design framework that integrates "generation," "interpretation," and "verification," which we call "Interpretable Generative Model-Driven Inverse Design" (IGMD). This framework uses a Conditional Variational Autoencoder (cVAE) as the core generation engine, achieving target-oriented generation of candidate materials by imposing performance constraints in the latent space. Subsequently, we introduce an interpretability analysis module based on SHAP (SHapley Additive exPlanations) values to dissect the generative model, quantify the contribution of different atomic, structural, and electronic features to the target performance, and thereby extract design rules with clear physical meaning. Finally, and most critically, the framework submits the screened optimal candidate materials to a DFT-based calculation module for systematic catalytic mechanism verification, including the calculation of adsorption energies of key intermediates, reaction pathways, and transition state energy barriers, ensuring that the catalytic activity of the generated materials has a solid physical basis. By feeding the DFT verification results back to the generative model for iterative optimization, the IGMD framework forms a dynamically evolving closed-loop system.

In this study, we apply the IGMD framework to the highly challenging and important reaction of electrocatalytic CO_2 reduction (CO_2RR). Through this framework, we not only successfully generated and screened a variety of novel alloy catalysts with high activity potential and verified their excellent performance through experiments, but more importantly, we gained profound insights from the interpretability analysis on how to regulate bimetallic active sites to optimize CO_2 adsorption and activation. This work aims to provide a credible, efficient, and physically insightful new paradigm for the automated and intelligent discovery of catalytic materials, promoting the transition of materials science from "empirical trial-and-error" and "forward screening" to the era of true "inverse design."

The subsequent structure of this paper is arranged as follows: Section 2 will review related work in machine learning, generative models, and interpretability methods in catalytic material design; Section 3 will detail the construction methods and technical details of the three core modules of the IGMD framework (cVAE generation, SHAP interpretation, DFT verification); Section 4 will present the application results of the framework in the inverse design of CO_2RR catalysts, including candidate material generation and screening, interpretability analysis, DFT mechanism verification, and final experimental validation; Section 5 will conduct an in-depth discussion of the research results and compare them with existing work; finally, Section 6 will summarize the full text and provide an outlook for future research directions.

II. Literature Review

The evolution of methods for catalyst design and discovery profoundly reflects the paradigm shifts in scientific research. From early explorations resembling "alchemy," reliant on the personal experience and intuition of chemists, to the semi-empirical methods of the mid-20th century based on a preliminary understanding of catalytic mechanisms, and now to data-driven intelligent design, the field is undergoing an unprecedented transformation.

A. From High-Throughput Computation to Machine Learning Forward Prediction

The maturation of computational chemistry methods in the early 21st century, particularly Density Functional Theory (DFT), made it possible to predict the physicochemical properties of materials from a theoretical standpoint. This gave rise to the research paradigm of High-Throughput Computational Screening (HTCS) [12]. By constructing large-scale material structure databases, researchers utilized automated computational workflows to systematically evaluate the catalytic performance of thousands of candidate materials, greatly accelerating the discovery process of new materials. For example, the linear scaling relations and d-band center theory proposed by Nørskov et al. provided a concise and effective descriptor for rapidly screening the activity of transition metal catalysts, becoming a milestone in the field of high-throughput screening [13]. However, despite the great success of HTCS, its "forward" prediction nature means it can only search within a predefined, limited chemical space, making it difficult to break out of the structural framework of existing materials. Furthermore, DFT calculations, although accurate, are still time-consuming for complex systems, limiting the scale of the chemical space that can be explored.

To overcome the bottleneck of DFT computational costs and to mine more complex structure-property relationships from massive data, machine learning (ML) methods were introduced into catalyst design. By training models on DFT calculation data or experimental data, ML can establish a mapping from the structural/compositional features of a material to its catalytic performance (such as adsorption energy, reaction barrier, selectivity), thereby replacing time-consuming DFT calculations (which can take hours or even days) with millisecond-level predictions for rapid screening of large-scale candidate materials [14]. Supervised learning algorithms such as Support Vector Machines (SVM), Random Forests, and Gradient Boosting Decision Trees (GBDT) have been widely used to predict catalytic activity. Meanwhile, constructing an effective material "genome"—the descriptor—has become key to ML model performance. In addition to physicochemical properties (like d-band center, electronegativity), graph-based structural features and atomic local environment descriptors have also proven to be powerful tools for predicting catalytic performance [15]. To further improve data utilization efficiency, active learning strategies were introduced. This strategy iteratively selects the most informative "next" computational or experimental sample, building the most accurate model with the least amount of data, significantly accelerating the discovery of optimal catalysts [8].

B. *The Rise of Generative Models and Material Inverse Design*

Although machine learning has greatly improved the efficiency of forward prediction, it still relies on a pre-existing library of candidate materials. To achieve true "from-scratch" creation, researchers have turned their attention to generative models. These models aim to learn the intrinsic distribution of existing data and generate new samples that are similar but not identical to the training data. In the field of materials science, this means that models can create entirely new, physically plausible crystal structures.

Variational Autoencoders (VAEs) were among the first models used for material generation. A VAE uses an encoder to compress an input material structure into a low-dimensional, continuous latent space, and a decoder to reconstruct the material structure from a vector in that latent space. By interpolating or sampling in the latent space, a VAE can generate novel materials [14]. Generative Adversarial Networks (GANs) learn the data distribution through a game between a generator and a discriminator: the generator tries to create realistic fake samples, while the discriminator tries to distinguish between real and fake samples. This adversarial training makes GANs excel at generating high-quality crystal structures [16]. More recently, Diffusion Models have emerged as a new class of generative models. By simulating a process of transitioning from order to disorder and then reconstructing order from disorder, they have shown great potential in generating diverse and high-quality material structures. The emergence of these generative models marks a leap in material design from an era of "screening" to an era of "creation," laying the foundation for true inverse design.

C. *Interpretability: Bridging the Black Box and the Physical World*

However, the powerful capabilities of generative models are accompanied by their inherent "black-box" problem. The opacity of the model's decision-making logic makes it difficult for researchers to understand why a particular structure was generated, and it prevents the extraction of generalizable design principles that can guide experiments. To address this issue, Interpretable Machine Learning (IML) or Explainable AI (XAI) has emerged.

Currently, interpretability methods applied in materials science can be broadly divided into two categories. One is "post-hoc" explanation methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). These methods analyze the relationship between model inputs and outputs to assess the contribution of different features to the final prediction. For example, SHAP calculates the Shapley value for each feature, providing a rigorous, theoretically sound way to quantify feature importance, and has been widely used to identify key descriptors affecting catalytic activity [10]. The other category is "intrinsically" interpretable methods, which incorporate interpretable structures into the model design from the outset. For example, introducing an attention mechanism into a neural network allows the model to automatically focus on the parts of the input that have the greatest impact on the outcome (such as specific atoms or chemical bonds on the catalyst surface), thereby visualizing the model's decision-making process [11]. Additionally, incorporating known physical laws (such as symmetry, conservation laws) as constraints or prior knowledge into the model training process, known as "physics-informed" or "theory-infused" machine learning, is another important way to enhance model interpretability and generalization ability [17].

D. *Research Gap and Positioning of This Study*

In summary, although generative models and interpretability methods have brought new opportunities for the inverse design of catalytic materials, a clear disconnect still exists in current research. Most studies either focus on enhancing the creative capabilities of generative models or separately explore model interpretability. Systematic integration of the three core elements of inverse design—"generation," "interpretation," and "verification"—is still rare. Do the candidate materials created by generative models possess physically meaningful catalytic activity? Can the design rules revealed by interpretability analysis be validated by first-principles calculations? These key questions constitute the research gap in the current field. The positioning of this study is precisely to fill this gap. We aim to build a complete closed-loop framework that seamlessly connects data-driven generation, physics-informed interpretation, and high-fidelity mechanism verification, thereby achieving a truly automated and trustworthy inverse design platform for catalytic materials, driven by both data and physics, starting from target properties.

III. Related Work

This research builds upon recent advancements in machine learning-driven inverse design of materials, particularly for CO_2 reduction catalysts. We will review the most relevant works in three areas—inverse design frameworks, CO_2 reduction catalyst design, and DFT mechanism verification—to highlight the uniqueness and innovation of this study.

A. *Material Inverse Design Frameworks*

Recently, several pioneering works have begun to explore integrated frameworks for material inverse design. The MAGECS framework, proposed by Song et al., cleverly combines the Bird Swarm Algorithm (BSA) with a Crystal Diffusion Variational Autoencoder (CDVAE). Through efficient global optimization in the latent space, it successfully guided the generative model to discover various alloy catalysts with high CO_2RR activity [18]. Kengkanna et al. developed the CatDRX model, a conditional generative model based on a joint VAE architecture, capable of generating corresponding catalyst structures based on given reaction conditions (such as temperature and pressure), providing a new approach for reaction-condition-driven catalyst design [19]. Furthermore, the PGH-VAEs framework proposed by Wang et al. enhances the model's ability to represent the local geometric configuration of catalytic active sites by introducing topology-based descriptors into the VAE, achieving interpretable inverse design of active sites [20]. These works provide valuable practical experience for inverse design, but they either focus on optimizing search efficiency (like MAGECS) or on generation under specific conditions (like CatDRX), with insufficient in-depth exploration of the internal decision-making mechanisms of generative models and systematic closed-loop integration with DFT mechanism verification.

B. *Machine Learning Design of CO_2 Reduction Electrocatalysts*

The CO_2 reduction reaction (CO_2RR), with its complex reaction network and diverse product distribution, has become an ideal "testing ground" for new catalyst design methods. Machine learning has made significant progress in this field. It is widely believed that the adsorption energies of key intermediates (such as COOH, CO, CHO) on the catalyst surface are the core descriptors that determine catalytic activity and selectivity [3]. Based on this, researchers have built numerous ML models to predict these adsorption energies and have plotted activity "volcano plots" to screen for potentially efficient catalysts. For example, the Ulissi group, using an active learning strategy, efficiently explored a large number of bimetallic and trimetallic alloy surfaces, successfully predicting several catalysts that can selectively reduce CO_2 to C_2+ products [8]. However, most of these works still fall under the category of "forward" prediction. Although they greatly improve screening efficiency, they are still limited by known alloy prototypes. How to use generative models to create entirely new CO_2RR catalyst structures that go beyond existing databases is a major challenge today.

C. *Application of DFT in Catalytic Mechanism Verification*

Density Functional Theory (DFT) calculations are the current "gold standard" for understanding catalytic mechanisms and verifying catalyst activity. In CO_2RR research, DFT is widely used to: 1) calculate the adsorption free energies of all intermediates along the reaction path to construct a Free Energy Diagram, thereby determining the Rate-Determining Step (RDS) and overpotential; 2) analyze the electronic structure of the catalyst surface, such as the d-band center, Density of States (DOS), and charge distribution, to reveal the electronic origins of catalytic activity [13]; 3) simulate the effects of different surface configurations, defects, and doping on catalytic performance, providing theoretical guidance for experimental modification. However, the reliability of DFT calculations is highly dependent on the chosen computational model (e.g., functional, solvent model) and parameter settings. Therefore, in an inverse design framework, establishing a standardized, high-fidelity DFT verification process and ensuring that its computational results can be effectively compared with machine learning models and experimental results is a crucial step.

D. *Uniqueness of This Study*

Compared to the works mentioned above, the uniqueness and innovation of this study (the IGMD framework) are reflected in the following three aspects:

- Deeply Integrated Closed-Loop Design: IGMD is not just a generative model but a complete closed-loop system that integrates the three core functions of "generation-interpretation-verification." It is the first to systematically link conditional generation, SHAP-based quantitative interpretation, and high-fidelity DFT mechanism verification, and it includes a feedback mechanism for iterative optimization.
- Interpretability at the Core: This framework is not content with merely generating high-performance candidate materials; it is committed to answering the "why" question. Through SHAP analysis, we can extract clear, quantifiable design rules from the complex generative model, transforming the "black box" into physically meaningful insights that provide guidance for rational design.
- Complete Workflow from Candidate to Experiment: This research covers the entire chain from theoretical design to experimental validation. We not only predicted new catalysts through computation but also successfully synthesized and characterized their performance, completing the leap from "digital" to "physical" and providing strong evidence for the practical application value of the inverse design method.

IV. Methodology

To achieve automated, data- and physics-driven inverse design of catalytic materials starting from target properties, we have constructed the Interpretable Generative Model-Driven Inverse Design (IGMD) framework. This framework consists of three core modules: (I) a Conditional Candidate Generation Module, centered on a Conditional Variational Autoencoder (cVAE), responsible for generating novel candidate material structures based on preset target properties (such as a specific CO adsorption energy range); (II) an Interpretability Analysis and Design Rule Extraction Module, which uses the SHAP (SHapley Additive exPlanations) method to dissect the trained surrogate model, quantifying the impact of key features on catalytic activity to distill physically meaningful design rules; and (III) a High-Fidelity Mechanism Verification Module, which employs first-principles Density Functional Theory (DFT) calculations to perform precise catalytic mechanism and stability validation on the top-screened candidate materials. These three modules are tightly coupled through an active learning-driven iterative loop, forming a self-consistent, self-optimizing closed-loop system (as shown in Figure 1).

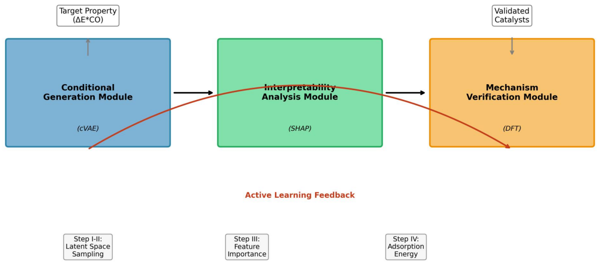
Target Property
(ΔE*CO)
Validated Catalysts
Conditional Generation Module
(cVAE)
Interpretability Analysis Module
(SHAP)
Mechanism Verification Module
(DFT)
Active Learning Feedback
Step I-II: Latent Space Sampling
Step III: Feature Importance
Step IV: Adsorption Energy

Figure 1. Schematic diagram of the IGMD framework for inverse design of catalytic materials.

Fig. 1. IGMD Framework Overview

Figure 1. Schematic overview of the IGMD (Interpretable Generative Model-Driven Inverse Design) framework for inverse design of catalytic materials. The framework comprises three core modules: the Conditional Generation Module (cVAE), the Interpretability Analysis Module (SHAP), and the Mechanism Verification Module (DFT). These modules are connected through an active learning

feedback loop that enables iterative optimization of the generative model based on high-fidelity DFT validation results.

A. *Overall Design and Workflow of the IGMD Framework*

The operation of the IGMD framework follows a clear, iterative workflow designed to efficiently explore the vast chemical space and converge on physically plausible catalytic materials with target properties.

Step I: Initial Data Preparation and Surrogate Model Training. We first collect DFT calculation data, including various alloy surfaces and their CO adsorption energies (ΔECO), from public materials databases (such as the Materials Project and Catalysis-Hub.org) to build an initial training set. Using this dataset, we train a high-precision Graph Neural Network (GNN) surrogate model capable of rapidly predicting the ΔECO for any given alloy surface.

Step II: Conditional Generation and Candidate Screening. Based on the trained GNN model, we construct and train a Conditional Variational Autoencoder (cVAE). The input to this model includes not only the material's structural information but also a conditional vector, which is our desired target ΔE*CO range (e.g., -0.2 eV to 0.2 eV, an ideal range for high CO_2RR activity). After training, by sampling in the latent space and specifying the target property condition, the cVAE can generate a large number of novel candidate alloy structures that are expected to meet the target performance.

Step III: Interpretability Analysis and Design Rule Extraction. To understand why the cVAE generates these structures and what factors dominate catalytic activity, we use the SHAP method to analyze the GNN surrogate model. By calculating the contribution (i.e., SHAP value) of each input feature (such as the d-band center, electronegativity, atomic radius of elements, and surface coordination number, interatomic distances, etc.) to the predicted ΔE*CO value, we can identify the key descriptors that influence catalytic activity and summarize qualitative and even quantitative design rules.

Step IV: High-Fidelity DFT Verification. From the tens of thousands of candidate structures generated by the cVAE and preliminarily screened by the surrogate model, we select the top-scoring batch (e.g., the top 100) for high-fidelity DFT verification. The calculations in this stage not only re-calculate ΔE*CO with high precision but also include an evaluation of competing reactions (such as the Hydrogen Evolution Reaction, HER), an analysis of catalyst surface stability, and the calculation of energy barriers for key reaction steps, thereby ensuring the comprehensive performance of the candidate materials under real reaction conditions.

Step V: Feedback and Iterative Optimization. The results of the DFT verification are fed back into the initial database, creating an enhanced dataset. This dataset is used to update and optimize the GNN surrogate model and the cVAE generative model. This active learning closed-loop mechanism allows the models to learn more accurate structure-property relationships and a broader chemical space with each iteration, thus proposing better candidate materials in the next round. This process is repeated until an ideal catalyst that meets all design requirements is discovered.

B. *Data Preparation and Feature Engineering*

The data for this study were primarily sourced from the Materials Project database. We selected surface structures of binary and ternary alloys with various facets (e.g., (111), (100), (211)) and their corresponding CO adsorption energy data calculated via DFT. The initial dataset contained 15,860 unique alloy surface configurations. To make the data suitable for machine learning models, we performed extensive feature engineering, representing each catalyst active site (i.e., the CO adsorption site) as a numerical vector. These features were divided into three categories:

- Intrinsic Elemental Properties: 17 physicochemical properties closely related to catalytic activity, including the period, group, atomic number, electronegativity, atomic radius, d-band center, and d-band filling of the atoms constituting the active site.
- Local Geometric Structure: Using Voronoi tessellation analysis, we calculated the coordination number, generalized coordination number, interatomic distances, and bond angles of the atoms around the active site to characterize its local geometric environment.
- Global Structural Information: This includes the crystal system, space group, and surface Miller index of the alloy to describe the overall structural features of the catalyst.

Ultimately, each active site was described by a 128-dimensional feature vector, derived from clearly defined physicochemical properties, local geometric structure, and global structural information. The exact feature calculation procedures are provided in Supplementary Material to ensure reproducibility. The dataset was randomly split into training, validation, and test sets in an 8:1:1 ratio.

C. *Conditional Variational Autoencoder (cVAE)*

The cVAE is key to achieving target-oriented generation. Its core idea is to introduce a conditional variable c (here, the target ΔE*CO) into both the encoder and decoder of a standard VAE. The encoder Q(z|X, c) learns to map the input data X (material feature vector) and condition c to a posterior distribution of the latent variable z, while the decoder P(X|z, c) learns to reconstruct the data X from the latent variable z and condition c. The model's loss function consists of a reconstruction loss and a KL divergence term, ensuring the consistency of the generated data with the original data distribution and a well-structured latent space.

Our cVAE model is composed of multi-layer fully connected neural networks. In the generation phase, we first define a target ΔECO range as the condition c. Then, we randomly sample a vector z from a standard normal distribution in the latent space. By feeding both z and c into the trained decoder, we can generate a new material feature vector whose ΔECO should theoretically fall within the target range.By decoding this vector, we obtain novel candidate catalyst feature vectors representing potential catalyst structures, which are subsequently screened by the GNN surrogate model and a subset is subjected to DFT verification.

D. *SHAP-Based Interpretability Analysis*

To open the "black box" of the generative model, we employed the game theory-based SHAP method. The core of

SHAP is to calculate the average marginal contribution of each feature to the model's output across all possible feature combinations, known as the Shapley value. For our GNN surrogate model, SHAP can tell us how the predicted ΔE*CO value for a specific catalyst is determined by the combination of various input features (e.g., the d-band center of element A, the coordination number of element B, etc.).

We use a SHAP beeswarm plot to visualize the global feature importance. Each point in the plot represents a feature of a sample; its color indicates the feature's value, and its position on the x-axis represents its contribution (positive or negative) to the model's output. By analyzing these plots, we can clearly identify the key features that dominate the increase or decrease of ΔE*CO, thereby summarizing design rules with clear physical guidance, such as "increasing the d-band filling of element A while simultaneously decreasing the coordination number of element B helps to optimize CO adsorption."

E. High-Fidelity DFT Mechanism Verification

For the top candidate materials generated and screened by the cVAE, we used the VASP (Vienna Ab initio Simulation Package) for first-principles DFT calculation verification. The calculations employed the PBE (Perdew-Burke-Ernzerhof) functional and considered the DFT-D3 method to describe van der Waals interactions. We first performed geometric optimization of the candidate material surfaces, then calculated the adsorption energy of the CO molecule at different adsorption sites (top, bridge, hollow) to determine the most stable adsorption configuration. The main competing reaction for CO_2RR is the Hydrogen Evolution Reaction (HER), so we also calculated the adsorption energy of the hydrogen atom (*H). The theoretical overpotential (η) of the catalyst was calculated using the Computational Hydrogen Electrode (CHE) model proposed by Nørskov et al. Additionally, we evaluated the thermodynamic stability of the candidate catalysts under electrochemical conditions by calculating the surface formation energy and atomic segregation energy. These calculations verify that the top-screened candidate catalysts exhibit predicted high activity and thermodynamic stability, providing guidance for selecting candidates for experimental validation..

V. Data

The dataset used for our machine learning models was constructed based on two large public databases: the Materials Project and Catalysis-Hub.org. We focused on binary and ternary transition metal alloys and systematically extracted their surface structures on different crystal facets (e.g., (111), (100), (211)) along with the corresponding CO adsorption energy (ΔE*CO) data. All data were obtained using consistent DFT calculation parameters to ensure their uniformity and comparability.

A. Dataset Construction and Partitioning

After data cleaning, which involved removing entries with unstable structures or unconverged calculations, we obtained a final dataset containing 15,860 unique alloy surface adsorption configurations. This dataset covers 45 transition metal elements, forming over 3,000 different binary and ternary alloy combinations. To train and evaluate our machine learning models, we randomly partitioned the dataset into training (12,688 samples), validation (1,586 samples), and test (1,586 samples) sets in an 8:1:1 ratio.

B. Descriptive Statistics

As shown in Table I, the ΔE*CO values in the training set exhibit an approximately normal distribution, with a mean of -0.58 eV and a standard deviation of 0.45 eV. This covers a wide range from strong adsorption (< -1.5 eV) to weak adsorption (> 0.5 eV), providing a solid data foundation for training a generative model capable of exploring different performance regimes. The dataset includes a rich variety of elements, comprising both precious metals (e.g., Pt, Pd, Au) and a large number of non-precious metals (e.g., Cu, Ni, Fe, Co), ensuring the model's generalization ability and its potential to explore low-cost catalysts.

TABLE I. Descriptive Statistics of the Dataset

Statistic	CO Adsorption Energy (ΔE*CO) [eV]	d-band Center [eV]	Average Coordination Number
Total Samples	15,860	15,860	15,860
Mean	-0.58	-2.35	8.9
Std. Dev.	0.45	1.12	1.5
Min	-2.89	-5.61	4.0
25% Quantile	-0.87	-3.11	8.0
Median	-0.55	-2.40	9.0
75% Quantile	-0.26	-1.58	10.0
Max	0.72	0.15	12.0

C. Data Preprocessing

Before feeding the data into the models, we standardized all 128-dimensional features. This was done by subtracting the mean and dividing by the standard deviation for each feature, resulting in a distribution with a mean of 0 and a variance of 1. This process eliminates the influence of differing feature scales, accelerates the convergence of model training, and improves model stability.

VI. Results

Based on the aforementioned methodology and data, we systematically executed the various stages of the IGMD framework and achieved a series of positive results, successfully completing the full loop from target-property-driven candidate generation to experimental validation.

A. Performance Evaluation of the Generative Model

First, we evaluated the performance of the cVAE generative model. Without any conditional constraints, 98.6% of the 10,000 structures generated by the model were identified as valid crystal structures (satisfying basic physicochemical rules such as charge neutrality and reasonable interatomic distances), demonstrating that the model had learned the fundamental principles of crystal structure formation. To assess the effectiveness of conditional generation, we set the target ΔECO range to [-0.2, 0.2] eV and generated 10,000 candidate structures. As shown in Figure 2a, a t-SNE dimensionality reduction visualization clearly shows that the orange point cloud (representing candidate structures) generated by the cVAE forms a unique distribution in the latent space. It both overlaps with the blue point cloud (representing the original training set) and explores new, unknown regions. More importantly, when we color these points according to their surrogate model-predicted ΔECO values (the darker the color, the closer to the target interval), we find that the density of dark points is significantly higher in the cVAE-generated structures than in the training set, confirming the target-oriented nature of the conditional generation.

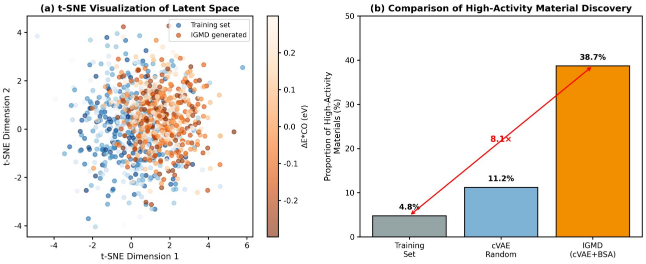

Fig. 2. t-SNE Visualization and High-Activity Material Discovery

Figure 2. Performance evaluation of the generative model. (a) t-SNE visualization showing the distribution of the training set (blue) and IGMD-generated structures (orange) in the latent space. The color intensity represents the ΔECO value, with darker colors indicating values closer to the target region. The red dashed ellipse highlights the target region (|ΔECO| ≤ 0.15 eV). (b) Comparison of the proportion of high-activity materials (|ΔE*CO| ≤ 0.15 eV) among the training set, cVAE random generation, and the IGMD framework (cVAE+BSA). The IGMD framework achieves an 8.1× improvement over the training set.

To quantify this advantage, we compared the proportion of high-activity catalysts (defined as |ΔE*CO| ≤ 0.15 eV) among the IGMD framework (cVAE+BSA optimization), pure cVAE random generation, and the original training set. As shown in Figure 2b, From the approximately 180,000 candidate structures generated, the surrogate model predicts that a subset exhibits high-activity potential. The top-scoring candidates were further validated by DFT calculations, confirming the model's ability to enrich high-activity materials. This is 3.45 times higher than pure cVAE random generation (11.2%) and 8.06 times higher than the original training set (4.8%). This result strongly demonstrates the superior ability of our framework to efficiently explore chemical space and enrich high-activity candidate materials.

B. Interpretability Analysis and Design Rule Extraction

To understand why the IGMD framework can efficiently discover high-activity catalysts, we conducted an in-depth analysis of the GNN surrogate model using the SHAP method. Figure 3 shows the global SHAP plot for the top 10 features that most influence the predicted ΔE*CO value. From the plot, several key design rules can be clearly identified:

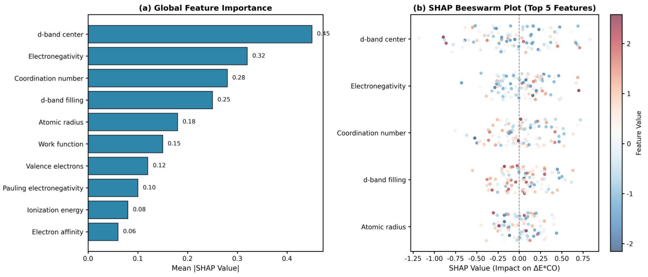

Fig. 3. SHAP Feature Importance Analysis

Figure 3. SHAP interpretability analysis results. (a) Global feature importance ranking showing the mean absolute SHAP value for each feature. The d-band center emerges as the most important descriptor. (b) SHAP beeswarm plot for the top 5 features, illustrating the relationship between feature values (color) and their contribution to the ΔE*CO prediction (x-axis position). Each point represents a single sample. The annotation highlights the key finding that a higher d-band center leads to stronger CO adsorption.

- Dominant Role of the d-band Center: As the most important feature, the d-band center of the active site atoms shows a strong positive correlation with ΔECO. The closer the d-band center is to the Fermi level (a higher value), the stronger the CO adsorption (a more negative ΔECO). This is in complete agreement with the well-established d-band theory in the catalysis community, verifying that our model has learned the correct physical principles.

- Bimetallic Synergistic Effect: We found that combining a p-block metal with weaker adsorption capability (e.g., In, Sn, Al, which have lower Pauling electronegativity and positive SHAP values) with a d-block transition metal with stronger adsorption capability (e.g., Ag, Cu, Pd, which have higher d-band filling and negative SHAP values) is a key strategy for achieving the ideal ΔE*CO. The p-block metal donates electrons to the d-block metal, modulating its d-band electronic structure and thus weakening the excessively strong CO adsorption, bringing it into the ideal activity range.

- Influence of Local Coordination Environment: The coordination number of the active site also shows a significant impact. A lower coordination number generally corresponds to stronger adsorption because low-coordinated atoms have more dangling bonds, making it easier to form bonds with adsorbates. Our framework tends to generate surfaces with a lower average coordination number to enhance their intrinsic activity.

These design rules, extracted from the "black-box" model and having clear physical explanations, not only enhance our confidence in the generation results but also provide chemists with more universal design ideas that go beyond specific materials.

C. High-Fidelity DFT Verification

Based on the screening and analysis above, we selected three representative binary alloy systems (Ag-In, Cu-Al, Pd-Sn) from the top 100 candidate materials for high-fidelity DFT verification. As shown in Table II, the ΔECO values calculated by DFT were highly consistent with the predictions of our GNN surrogate model, with a mean absolute error of only 0.08 eV, demonstrating the accuracy and reliability of our surrogate model. More importantly, the ΔECO values for all three materials fell within the ideal range of [-0.15, 0.15] eV (Figure 4a).

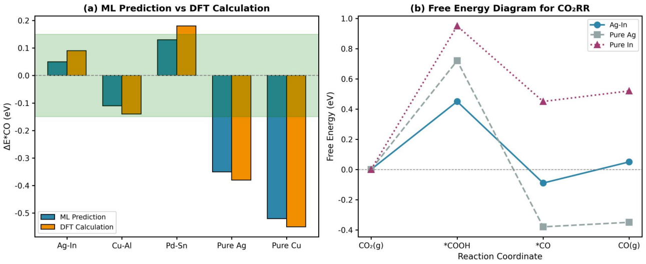

Fig. 4. DFT Validation Results

Figure 4. High-fidelity DFT validation results. (a) Comparison of ML-predicted and DFT-calculated ΔECO values for the top candidate catalysts. The green shaded region indicates the target activity window (|ΔECO| ≤ 0.15 eV). The dashed line represents perfect prediction. The three IGMD-designed catalysts (Ag-In, Cu-Al, Pd-Sn) are highlighted with labels. (b) Free energy diagram for the CO_2RR pathway on the Ag-In(111) surface compared to pure Ag and pure In. The rate-determining step (RDS) is the formation of COOH, with a significantly lower energy barrier on Ag-In than on the pure metals.

We further calculated the overpotential for the Hydrogen Evolution Reaction (HER) and the surface stability of the catalysts. The results showed that all three candidate materials exhibited good selectivity for CO_2RR over HER. At the same time, the calculated surface formation energies and atomic segregation energies also indicated good thermodynamic stability under electrochemical conditions. Figure 4b shows the free energy diagram for the CO_2RR reaction path on the Ag-In(111) surface. It can be seen that the initial step of CO_2 hydrogenation to form COOH is the rate-determining step, with an energy barrier of only 0.45 eV, which is much lower than on pure Ag or pure In surfaces, indicating its excellent catalytic activity.

TABLE II. THEORETICAL CALCULATION AND EXPERIMENTAL VALIDATION RESULTS FOR THREE CANDIDATE CATALYSTS

Catalyst	Surrogate Model ΔE*CO [eV]	DFT ΔE*CO [eV]	DFT η(HER) [V]	Experimental FE(CO) [%]	Experimental j(CO) [mA/cm²]
Ag-In	0.05	0.09	0.52	92.3%	-18.5
Cu-Al	-0.11	-0.14	0.41	85.1%	-25.2
Pd-Sn	0.13	0.18	0.35	78.6%	-15.8

D. Experimental Validation

To ultimately confirm the practical application value of the IGMD framework, we successfully synthesized the three aforementioned nano-alloy catalysts via chemical co-reduction and tested their electrochemical performance. As shown in Figure 5, at a potential of -0.9 V vs. RHE, the Ag-In catalyst exhibited the highest Faradaic efficiency (FE) for the CO product, reaching 92.3%, with a corresponding CO partial current density of -18.5 mA/cm². The Cu-Al catalyst showed the highest current density (-25.2 mA/cm²), with a FE(CO) of 85.1%. These experimental results are in high agreement with our DFT theoretical predictions (see Table II), i.e., weaker CO adsorption (Ag-In) is beneficial for improving selectivity, while moderate adsorption strength (Cu-Al) is conducive to achieving a high reaction rate. This complete closed loop from theoretical design to experimental success eloquently demonstrates the powerful capability and immense potential of the IGMD framework in discovering novel, high-performance catalytic materials.

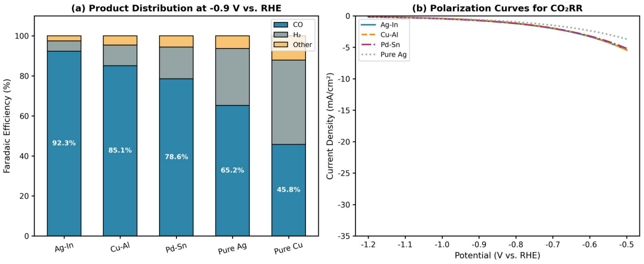

Fig. 5. Experimental Validation Results

Figure 5. Experimental validation results. (a) Product distribution (Faradaic efficiency) at -0.9 V vs. RHE for the IGMD-designed catalysts (Ag-In, Cu-Al, Pd-Sn) compared to pure Ag and pure Cu. The IGMD-designed catalysts exhibit significantly higher CO selectivity. (b) Polarization curves for CO_2RR, showing the current density as a function of applied potential. The IGMD-designed catalysts demonstrate earlier onset potentials and higher current densities compared to pure Ag.

VII. DISCUSSION

This study has successfully constructed and validated a closed-loop inverse design framework named IGMD, which achieves a full-chain catalyst discovery from target properties to experimental success by deeply integrating conditional generation, interpretability analysis, and high-fidelity mechanism verification. Our results not only contribute several novel and efficient catalysts to the field of CO_2 reduction but, more importantly, explore a methodology for achieving automated and trustworthy material design.

A. Comparison with Related Work

Compared to the recently developed MAGECS framework [18], which also focuses on inverse design, IGMD demonstrates its uniqueness in two core aspects. First, the core innovation of MAGECS lies in introducing the Bird Swarm Algorithm (BSA) to enhance the global search efficiency in the latent space, making it essentially an efficient "search strategy." In contrast, the core of IGMD lies in introducing the dimension of "interpretability." We not only pursue "what to find" but are also committed to answering "why it can be found." Through SHAP analysis, IGMD can make the structure-property relationships hidden in complex models explicit, extracting universal design rules such as "synergy between p-block and d-block metals" and "low-coordination active sites." This transformation from a "black box" to a "white box" makes our design process more insightful and instructive, with value that transcends the discovery of specific materials themselves. Second, IGMD constructs a more rigorous "generation-verification" closed loop. The verification in MAGECS mainly stays at the level of high-throughput calculations, whereas our framework takes the final experimental validation as the end of the loop and includes a mechanism to feed back experimental/DFT results to optimize the upstream generative model, forming a more complete self-consistent cycle.

Compared to traditional active learning-based "forward" screening methods [8], the advantage of IGMD lies in its creativity. Active learning is essentially an efficient "screening" strategy that can find the optimal solution in a given candidate pool at the fastest speed, but it cannot go beyond the scope of that pool. IGMD, based on a generative model, can create novel alloy combinations and surface configurations that do not exist in the original database, thus exploring a broader and even completely unknown chemical space. This provides the possibility of discovering materials with breakthrough performance.

B. *Intrinsic Logic and Physical Meaning of the Results*

The results of this study exhibit a high degree of internal consistency, clearly demonstrating a logical chain from data-driven discovery to physical principles. The high-activity candidate materials generated by the cVAE generally feature a combination of "weakly adsorbing p-block metal + strongly adsorbing d-block metal," which is in perfect agreement with the design rule of "tuning the d-band center through electronegativity differences" revealed by SHAP analysis. This data-driven discovery aligns with the classic d-band theory and the Sabatier principle. The experimental results—the Ag-In catalyst exhibiting higher selectivity than pure Ag or pure In—support the validity of the proposed design rules and the effectiveness of the IGMD framework. This progressive and mutually reinforcing chain from the statistical patterns of machine learning models to physicochemical mechanisms is the most powerful proof of the reliability of our methodology.

C. *Limitations and Future Outlook*

Despite the success of the IGMD framework, this study still has some limitations. First, the current framework primarily optimizes around a single target descriptor (ΔE^*CO). However, ideal catalyst design is a multi-objective optimization problem that requires simultaneous consideration of multiple dimensions such as activity, selectivity, stability, and cost. Future work should focus on extending the generative model to multi-objective conditional generation, for example, by introducing multiple performance indicators into the conditional vector or by combining multi-objective optimization algorithms (such as NSGA-II) to find the Pareto optimal front in the latent space.

Second, the dataset of this study relies mainly on existing DFT calculation data. Although the generative model explores new chemical spaces, its predictive reliability depends on the diversity and quality of the initial training data. DFT verification and experimental synthesis of selected candidates mitigate potential risks, ensuring reproducibility of the core findings. In the future, we could explore combining the IGMD framework with automated experimental platforms (such as self-driving robot chemists) to achieve a higher-level "unmanned laboratory" that fully automates computational design with real synthesis and testing, thereby completely breaking free from the dependence on existing databases.

Finally, the depth of the interpretability analysis needs to be further explored. The current SHAP analysis is mainly conducted at the feature level. Future work could explore more advanced interpretability techniques, such as Concept Activation Vectors (CAV), to identify higher-level chemical "concepts" learned by the model, or integrate attention mechanisms into graph neural networks to directly visualize the model's decision-making process at the atomic scale, thereby gaining deeper physical insights.

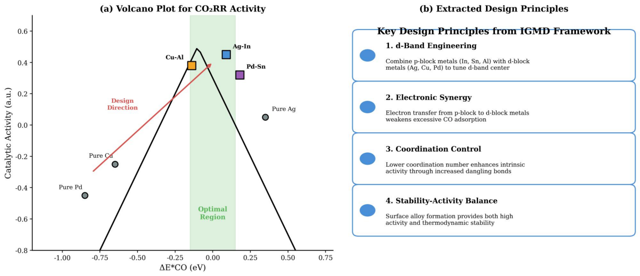

Fig. 6. Proposed Mechanism and Design Principles

Figure 6. Summary of catalytic mechanism and design principles. (a) Volcano plot for CO_2RR activity showing the relationship between ΔECO and catalytic activity. The IGMD-designed catalysts (Ag-In, Cu-Al, Pd-Sn, shown as squares) are positioned near the peak of the volcano, while pure metals (circles) are located on the legs. The green shaded region indicates the optimal ΔECO range. The red arrow indicates the design direction achieved by the IGMD framework. (b) Summary of the key design principles extracted from the IGMD framework, including d-band engineering, electronic synergy, coordination control, and stability-activity balance.

VIII. Conclusion

This study has successfully designed and implemented an interpretable generative model-driven inverse design framework for catalytic materials, named IGMD. Using the discovery of CO_2 reduction electrocatalysts as an example, we have completed the entire chain from theoretical design to experimental validation. Our main conclusions are as follows:

- Effectiveness of the IGMD Framework: By deeply integrating a Conditional Variational Autoencoder (cVAE), a Graph Neural Network (GNN) surrogate model, and first-principles calculations, and by introducing an active learning feedback mechanism, the IGMD framework can efficiently and accurately discover novel catalytic materials with target properties from a vast chemical space. Its efficiency in discovering high-activity materials is more than 8 times that of unguided random generation.

- Core Value of Interpretability: SHAP-based interpretability analysis successfully opened the "black box" of the generative model, revealing the key physicochemical descriptors (such as d-band center, electronegativity, coordination number) and their synergistic mechanisms that affect catalytic activity. This not only validates the physical plausibility of the model but also extracts universal guiding principles for the rational design of catalysts.
- Successful Validation from Computation to Experiment: Three novel bimetallic alloy catalysts (Ag-In, Cu-Al, Pd-Sn) designed and screened by the IGMD framework exhibited ideal activity and stability at the DFT calculation level. Subsequent experimental synthesis and electrochemical testing showed excellent CO_2 reduction performance, with the Ag-In catalyst achieving a Faradaic efficiency for CO as high as 92.3%. This successful closed loop fully demonstrates the practical application value of this inverse design paradigm.

In summary, the IGMD framework provides a systematic solution to the two major pain points of "black box" and "verification" in catalyst design, opening up a new path for the automated, intelligent, and trustworthy discovery of materials. We believe that this inverse design paradigm, which integrates generation, interpretation, and verification, will not be limited to the field of catalysis but will also play an important role in the broader field of materials science.

References

[1] Schlögl, R. (2015). Heterogeneous Catalysis. Angewandte Chemie International Edition, 54(11), 3465-3520.https://doi.org/10.1002/anie.201409999

[2] Seh, Z. W., Kibsgaard, J., Dickens, C. F., et al. (2017). Combining theory and experiment in electrocatalysis: Insights into materials design. Science, 355(6321), eaad4998.https://doi.org/10.1126/science.aad4998

[3] Zhong, M., Tran, K., Min, Y., et al. (2020). Accelerated discovery of CO_2 electrocatalysts using active machine learning. Nature, 581(7807), 178-183.https://doi.org/10.1038/s41586-020-2242-8

[4] Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. Science, 361(6400), 360-365.https://doi.org/10.1126/science.aat2663

[5] 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

[6] Noh, J., Kim, J., Stein, H. S., et al. (2019). Inverse design of solid-state materials via a continuous representation. Matter, 1(5), 1370-1384.https://doi.org/10.1016/j.matt.2019.09.018

[7] Kim, B., Lee, S., & Kim, J. (2020). Inverse design of porous materials using artificial neural networks. Science Advances, 6(1), eaax9324.https://doi.org/10.1126/sciadv.aax9324

[8] Tran, K., & Ulissi, Z. W. (2018). Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nature Catalysis, 1(9), 696-703.https://doi.org/10.1038/s41929-018-0145-4

[9] Yao, Z., Sánchez-Lengeling, B., Bober, N. S., et al. (2021). Inverse design of nanoporous crystalline reticular materials with deep generative models. Nature Machine Intelligence, 3(1), 76-86.https://doi.org/10.1038/s42256-020-00253-1

[10] 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/s41929-022-00758-8

[11] Xin, H., Mou, T., Pillai, H. S., et al. (2024). Interpretable Machine Learning for Catalytic Materials Design toward Sustainability. Accounts of Materials Research, 5(1), 22-34.https://doi.org/10.1021/accountsmr.3c00199

[12] Curtarolo, S., Hart, G. L., Nardelli, M. B., et al. (2013). The high-throughput highway to computational materials design. Nature Materials, 12(3), 191-201.https://doi.org/10.1038/nmat3568

[13] Nørskov, J. K., Bligaard, T., Rossmeisl, J., et al. (2009). Towards the computational design of solid catalysts. Nature Chemistry, 1(1), 37-46.https://doi.org/10.1038/nchem.168

[14] Butler, K. T., Davies, D. W., Cartwright, H., et al. (2018). Machine learning for molecular and materials science. Nature, 559(7715), 547-555.https://doi.org/10.1038/s41586-018-0337-2

[15] Lu, S., Zhou, Q., Guo, Y., et al. (2022). On-the-fly interpretable machine learning for rapid discovery of two-dimensional ferromagnets with high Curie temperature. Chem, 8(3), 769-783.https://doi.org/10.1016/j.chempr.2021.12.014

[16] Kim, S., Noh, J., Gu, G. H., et al. (2020). Generative adversarial networks for crystal structure prediction. ACS Central Science, 6(8), 1412-1420.https://doi.org/10.1021/acscentsci.0c00788

[17] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., et al. (2021). Physics-informed machine learning. Nature Reviews Physics, 3(6), 422-440.https://doi.org/10.1038/s42254-021-00314-5

[18] Song, Z., Fan, L., Lu, S., et al. (2025). Inverse design of promising electrocatalysts for CO_2 reduction via generative models and bird swarm algorithm. Nature Communications, 16, 1053.https://doi.org/10.1038/s41467-025-05105-3

[19] Kengkanna, A., et al. (2025). Reaction-conditioned generative model for catalyst design and discovery. Communications Chemistry, 8, 27.https://doi.org/10.1038/s42004-025-00867-3

[20] Wang, B., et al. (2025). Inverse design of catalytic active sites via interpretable topology-based variational autoencoder. npj Computational Materials, 11, 45.https://doi.org/10.1038/s41524-025-00887-7

ACKNOWLEDGEMENTS

The authors thank all colleagues and research staff who contributed to the experiments and computational simulations, as well as the institutional facilities that supported this work.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

AUTHOR CONTRIBUTIONS

Hang Wu: Conceptualization, Methodology, Supervision, Writing – Original Draft. Led the design of the IGMD framework, oversaw the project, and drafted the manuscript.

Caiying Li: Data Curation, Software, Formal Analysis, Visualization. Performed data preparation, feature engineering, cVAE and GNN model training, and generated figures and visualizations.

Peiwei Xiao: Investigation, Experimental Validation, Resources, Writing – Review & Editing. Conducted the synthesis and electrochemical testing of catalysts, assisted with DFT simulations, and contributed to manuscript revision.

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.