Available online at www.keaipublishing.com/en/journals/journal-of-magnesium-and-alloys/

ScienceDirect

Journal of Magnesium and Alloys 14 (2026) 101858

www.elsevier.com/locate/jma

Full Length Article

# From LLM to Agent: A large-language-model-driven machine learning framework for catalyst design of $MgH_2$ dehydrogenation

Tongao Yao[a,b], Yang Yang[c], Jianghao Cai[a,b], Rui Liu[c], Zhaoyan Dong[c], Xiaotian Tang[a,b], Xuqiang Shao[c], Zhengyang Gao[a,b], Guangyao An[a,b,*], Weijie Yang[a,b,*]

[a] *Department of Power Engineering, North China Electric Power University, Baoding, 071003, Hebei, China*

[b] *Hebei Key Laboratory of Energy Storage Technology and Integrated Energy Utilization, North China Electric Power University, Baoding, 071003, Hebei, China*

[c] *Department of Computer Science, North China Electric Power University, Baoding, 071003, Hebei, China*

Received 1 June 2025; received in revised form 3 August 2025; accepted 13 August 2025
Available online 22 October 2025

**Abstract**

Magnesium hydride ($MgH_2$), a promising high-capacity hydrogen storage material, is hindered by slow dehydrogenation kinetics. AI-driven catalyst discovery to address this is often hampered by the laborious extraction of data from unstructured literature. To overcome this, we introduce a transformative "LLM to Agent" framework that synergistically integrates Large Language Models (LLMs) for automated data curation with Machine Learning (ML) for predictive design. We automatically constructed a comprehensive database of 809 $MgH_2$ catalysts (6555 data rows) with high fidelity and an ~40-fold acceleration over manual methods. The resulting ML models achieved high accuracy (average $R^2 > 0.91$) in predicting dehydrogenation temperature and activation energy, subsequently guiding a Genetic Algorithm (GA) in an exploratory inverse design that autonomously uncovered key design principles for high-performance catalysts. Encouragingly, a strong alignment was found between these AI-discovered principles and the design strategies of recently reported, state-of-the-art experimental systems, providing substantial evidence for the validity of our approach. The framework culminates in Cat-Advisor, a novel, domain-adapted multi-agent system. Cat-Advisor translates ML predictions and retrieval-augmented knowledge into actionable design guidance, demonstrating capabilities that surpass those of general-purpose LLMs in this specialized domain. This work delivers a practical AI toolkit for accelerated materials discovery and advances the emerging Agent-based paradigm for designing next-generation energy technologies.



*Keywords:* $MgH_2$ dehydrogenation; Large language model; Machine learning; Genetic algorithm; Catalyst design; Hydrogen storage.

## 1. Introduction

Hydrogen, with its exceptional energy density (142 MJ/kg [1]) and clean water byproduct upon combustion [2], is increasingly vital as a sustainable energy vector. Its remarkable versatility in conversion to both electricity and heat renders it an ideal medium for large-scale renewable energy integration and flexible energy deployment [3]. As the hydrogen economy advances, efficient and cost-effective hydrogen storage remains a critical bottleneck for widespread adoption. Solid-state hydrogen storage, particularly magnesium hydride ($MgH_2$), stands out as a promising solution due to its inherent safety and high volumetric capacity. $MgH_2$, characterized by a high theoretical gravimetric capacity (7.6 wt%), earth-abundant constituents, and low cost [4], is compelling for practical hydrogen storage applications.

However, the practical utilization of $MgH_2$ is significantly limited by thermodynamic and kinetic barriers. High dehydrogenation temperatures (typically >300 °C) and sluggish hydrogen absorption/desorption kinetics impede its operation under ambient conditions [5]. Catalytic modification has emerged as an effective strategy to overcome these limitations

Peer review under the responsibility of Chongqing University
* Corresponding authors.
*E-mail addresses:* angy@ncepu.edu.cn (G. An), yangwj@ncepu.edu.cn (W. Yang).

https://doi.org/10.1016/j.jma.2025.08.021

by reducing the dehydrogenation temperature and enhancing sorption kinetics [6]. Nevertheless, rational catalyst design remains a formidable challenge due to an ever-expanding parameter space. This space encompasses not only traditional catalyst compositions, structures, and morphologies, but also novel material systems at the frontier of materials science. Furthermore, catalyst activity (i.e., kinetics) is not the sole determinant of practical viability; achieving long-term cycling stability is an equally critical challenge. However, modeling this aspect is notoriously difficult, largely due to the heterogeneous and often qualitative nature of stability data reported in the literature.

For instance, the development of Mg-based high-entropy alloys (HEAs) offers exciting new possibilities for catalysis. As explored in publications such as the Journal of Magnesium and Alloys, these materials, often synthesized via high-energy mechanical milling, can form unique structures like amorphous or complex intermetallic phases, providing abundant active sites and enhanced structural stability [7]. Understanding the fundamental principles of Mg-containing HEAs is crucial for designing new catalysts [8]. The sheer complexity of these emerging systems, combined with recent work in JMA highlighting the use of machine learning to accelerate advanced magnesium alloy design [9], underscores the urgent need for powerful, data-driven design strategies like the one we propose.

Machine learning (ML) has rapidly emerged as a transformative, data-driven strategy to accelerate materials discovery, revolutionizing the design and optimization of advanced materials for energy applications and offering a powerful alternative to traditional trial-and-error approaches. Early ML studies have demonstrated their potential in predicting hydrogen storage properties for metal hydrides and alloys [10,11], but its application has since expanded dramatically. Today, ML is instrumental across the entire materials science landscape, for example, in the rational design of novel high-entropy alloys [12] and high-performance titanium alloys [13]; the prediction of mechanical properties in advanced composites [14,15]; and the optimization of advanced manufacturing techniques, such as laser powder bed fusion [16] and other material processing parameters [17]. Within the energy sector specifically, these techniques have been equally impactful. For instance, in hydrogen storage, ML has identified critical features like MeanIonicChar and Fe content to guide the development of high-performance materials [18]. Similarly, neural networks have facilitated the design of highly conductive alkaline anion exchange membranes (AEMs) by optimizing molecular structures for fuel cells and water electrolyzers [19,20]. Beyond hydrogen energy, ML has accelerated electrocatalyst development for hydrogen evolution reactions [21], enabled efficient material screening for perovskite solar cells [21], and streamlined the discovery of multimetallic catalysts for oxygen evolution reactions [22]. These advancements underscore the broad applicability and transformative potential of ML in materials science. However, the efficacy of conventional ML models is often limited by the availability of large, high-quality, and structured datasets, particularly for catalytic materials like those needed for $MgH_2$. While computational databases are growing, capturing the complexities of catalytic systems necessitates experimental data, and traditional ML methods struggle to fully utilize the wealth of unstructured knowledge embedded in the scientific literature. This data and knowledge gap currently restricts the full potential of ML to expedite catalyst discovery, especially for $MgH_2$, where current datasets remain limited in size and scope, hindering the development of robust, predictive models.

Recent advancements in Artificial Intelligence (AI), particularly Large Language Models (LLMs), offer transformative tools for scientific inquiry. LLMs, exemplified by models like ChatGPT [23], Claude [24], Gemini [25], Deepseek [26], and Qwen [27], exhibit remarkable capabilities in natural language processing, knowledge extraction, and pattern recognition from extensive textual data. These capabilities are exceptionally relevant to data-driven materials science, where LLMs can be leveraged for large-scale literature mining [28–34], novel materials design [32,35–43], and experimental workflow optimization [32–35, 44–48]. Specifically within the field of magnesium alloys, recent studies have demonstrated this potential by developing domain-specific models like MagBERT for extracting material properties from text [49], and expert systems such as PDGPT to streamline the retrieval of phase diagram information [50].

While the synergy of LLMs and machine learning (ML) is emerging as a promising paradigm in materials informatics [51], a critical bottleneck persists: the gap between unstructured scientific literature and the high-quality, structured datasets required for robust predictive modeling. This is especially true for $MgH_2$ catalyst design, where decades of experimental data are locked in varied textual formats, hindering accelerated, AI-driven discovery. To bridge this gap, this study proposes a structured "LLM to Agent" framework. The process begins with LLM1 (Data Curation), where a state-of-the-art model (GPT-4o) performs automated, high-fidelity extraction of a comprehensive database from the scientific literature. This structured data then serves as the foundation for our predictive ML models and a genetic algorithm. The framework culminates in LLM2 (Expert Advisory), a multi-agent system named Cat-Advisor. Powered by a domain-adapted, fine-tuned LLM, Cat-Advisor functions as an interactive AI Agent by integrating ML predictions with a retrieval-augmented knowledge base to deliver context-aware recommendations. This end-to-end framework is characterized by its functional integration of data extraction, predictive modeling, and an interactive multi-agent advisory system, offering a methodological template for addressing data-driven challenges in materials science.

## 2. Methods

### 2.1. Data collection and processing

The construction of our dataset began with a systematic literature search on the Web of Science using the keywords "catalyst" and "$MgH_2$," which yielded 759 relevant publica-

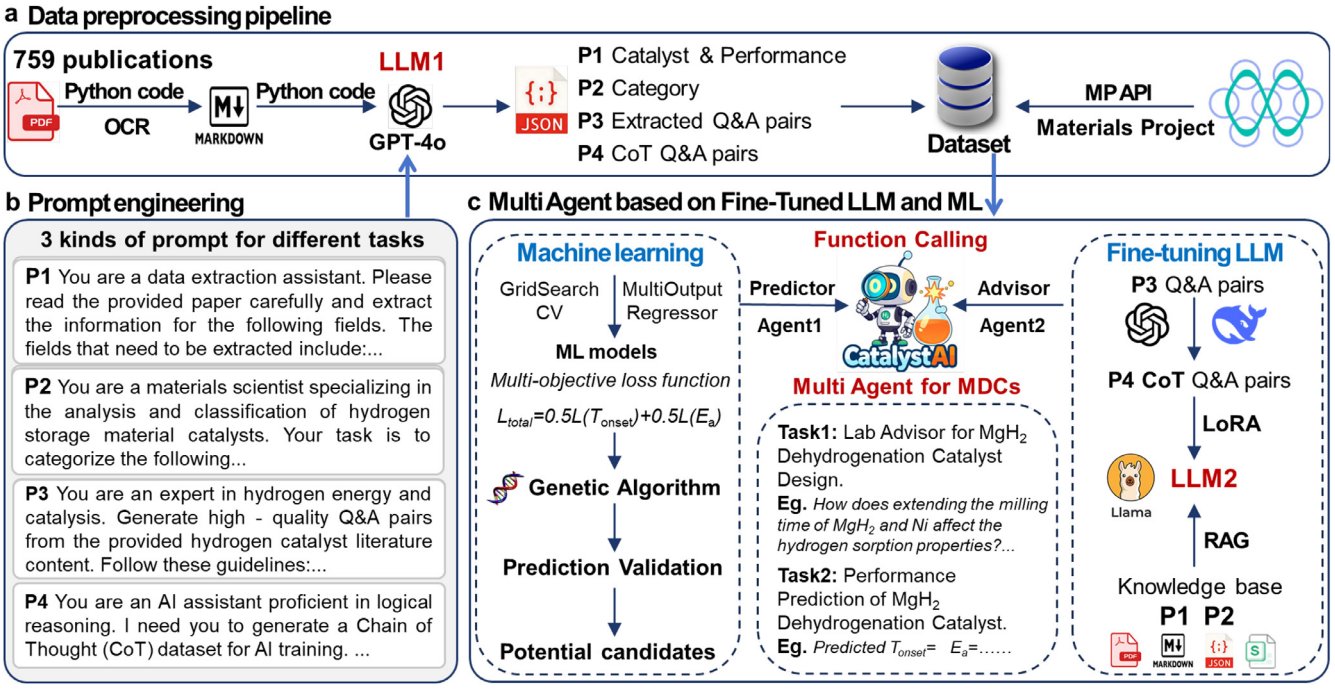


Fig. 1. Schematic of a Large Language Model (LLM)-Driven Machine Learning Framework for $MgH_2$ Dehydrogenation Catalyst Design. a) Data Preprocessing Pipeline Employing LLM Technology. This panel depicts the data preprocessing pipeline, where P1 (catalyst and performance data), P2 (catalyst category), P3 (extracted Question & Answer [Q&A] pairs), and P4 (Chain-of-Thought [CoT] Q&A pairs) represent information types extracted using prompts (see Figs. S2–S5 in the Supporting Information for prompt details). These labels (P1-P4) also correspond to the prompt engineering steps outlined in panel (b). b) Prompt Engineering. This panel details the prompt engineering stages: P1—extraction of catalyst experimental parameters from publications; P2—classification of extracted catalysts; P3—extraction of Q&A pairs from publications; and P4—addition of a Chain-of-Thought (CoT) process to Q & A pairs. c) Multi-Agent System Development using Fine-Tuned LLMs and ML. Panel (c) shows the development of a multi-agent system integrating: machine learning models (Agent1, ML) trained on the processed dataset from panel (a), utilizing a Genetic Algorithm and prediction validation for candidate identification; and a fine-tuned LLM (Agent2, LLM2) incorporating Chain-of-Thought (CoT) datasets (P3 & P4) and a Retrieval-Augmented Generation (RAG)-enhanced knowledge base (P1 & P2). Function-calling techniques connect Agent1 and Agent2 to create the multi-agent system for magnesium-based dehydrogenation catalysts (MDCs).

tions. These publications form the corpus for our data extraction workflow, which proceeds in four key steps:

1. **PDF to markdown conversion**: to optimize the source text for LLM parsing, all 759 publications were first converted from PDF to a structured Markdown format using the Nougat [52] package. This step enhances the ability of the model to recognize document structure, such as tables and sections.
2. **Prompt-Driven data extraction**: we developed a series of highly specific prompts to instruct the GPT-4o model to act as a domain expert [53]. As illustrated in Fig. 1b, these prompts fall into four categories (P1-P4). The full details and exemplary prompts for each category are provided in the Supporting Information (Figs. S2–S5). The primary data extraction prompt (P1) commanded the model to read each Markdown file and extract 16 predefined experimental parameters. Crucially, the prompt enforced the output to be a structured JSON object, a critical step for ensuring data consistency and facilitating automated processing.
3. **Data aggregation**: the individual JSON files generated for each publication were then programmatically parsed and aggregated into a single master database.
4. **Validation and curation**: finally, this raw database was subjected to the comprehensive, semi-automated validation process detailed below to correct errors and ensure the highest possible data fidelity for model training.

We acknowledge that automated data extraction from complex scientific texts presents challenges, such as LLM "hallucinations" causing unit inconsistencies (e.g., extracting temperature in Kelvin instead of Celsius) or parsing complex notations (e.g., "1.3 wt% at 300 °C, 3.7 wt% at 350 °C"). To address these issues robustly and ensure scalability, we developed a semi-automated validation workflow instead of relying on purely manual correction.

First, after the initial LLM extraction, an automated post-processing script performs a series of sanity checks. For instance, it flags temperature values outside a physically plausible range for $MgH_2$ dehydrogenation (e.g., >500), presumes them to be in Kelvin, and performs an automatic conversion to Celsius. This script also standardizes varied notations where possible.

Second, all automatically flagged or modified data points are then passed to an expert-in-the-loop verification stage. In this step, a domain expert performs a final, rapid review to confirm the corrections, ensuring the high data fidelity required for model training. This two-step process significantly enhances the efficiency and reliability of our data curation

pipeline, making it more robust and scalable for future applications.

### 2.2. Text mining evaluation

To assess the performance of this method for extracting parameters from catalyst literature, we utilized precision, recall, and F1-score as our primary metrics. Precision quantifies the accuracy of the extracted information, confirming that the identified data points accurately represent the intended parameters. Recall evaluates the completeness of the extraction process, measuring the proportion of relevant data points successfully retrieved. The F1-score offers a balanced metric combining precision and recall, providing a comprehensive evaluation of the effectiveness of the method.

Our evaluation demonstrates the robustness of the text mining approach while identifying areas for improvement in managing complex or inconsistently formatted data. This thorough assessment is essential for refining the extraction methodology, thus improving the reliability and utility of the resulting dataset for subsequent analyses. The insights derived from this evaluation will inform ongoing efforts to enhance the integration of machine learning models with domain-specific knowledge, ultimately advancing the field of magnesium-based hydrogen storage catalysts.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

In the text mining task aimed at extracting design parameters for catalysts, each retrieved parameter is classified into one of three categories: True Positives (TP), indicating accurately extracted parameters; False Positives (FP), representing incorrectly extracted parameters or irrelevant information; and False Negatives (FN), denoting parameters that were not successfully retrieved. Precision quantifies the accuracy of the method in retrieving catalyst parameters, confirming that the identified data points accurately reflect the intended parameters. Recall evaluates the completeness of the extraction process, measuring the proportion of relevant parameters successfully retrieved. The F1-score, which balances precision and recall, provides a comprehensive measure of the overall performance of the method.

### 2.3. GPT APIs and machine learning models

In this study, we employed the GPT API provided by OpenAI for multiple tasks, with all operations performed in a Python 3.9.19 environment using the openai package, version 1.55.0. For converting PDF documents into Markdown format, we utilized the Nougat package from Meta, version 0.1.17. Data extraction and classification were conducted using the GPT-4o model, ensuring thorough and accurate information retrieval. All parameter settings conformed to the default configurations specified in the openai Python package. For predictive modeling, we implemented an Extreme Gradient Boosting (XGBoost) model from the xgboost package, version 2.1.3, alongside other machine learning algorithms using the scikit-learn package, version 1.5.2. All relevant code is publicly available in the GitHub repository (https://github.com/Weijie-Yang/cat_advisor).

### 2.4. Predictive modeling framework

We developed four machine learning regression models: Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), and Extreme Gradient Boosting (XGBoost) to predict two interdependent target properties: the onset dehydrogenation temperature ($T_{onset}$) and activation energy ($E_a$) of MDCs($MgH_2$ dehydrogenation catalysts). These models were implemented within a MultiOutputRegressor framework from the scikit-learn package (version 1.5.2) in Python 3.9.19, enabling simultaneous prediction of both targets while accounting for potential correlations. The MultiOutputRegressor treats each target as an independent regression task, fitting a separate base estimator (e.g., RF, XGBoost) to $T_{onset}$ and $E_a$, with predictions defined as:

$$\hat{y} = \left[\hat{y}_{T_{\text{onset}}}, \hat{y}_{E_a}\right] = \left[f_{T_{\text{onset}}}(X), f_{E_a}(X)\right]) \tag{4}$$

where $X$ is the input feature matrix, and $f_{T_{\text{onset}}}$ and $f_{E_a}$ are the trained regression functions for each target. To ensure balanced learning across both properties, we introduced an implicit multi-objective loss function within the MultiOutputRegressor framework. For a given base estimator, the total loss $L_{total}$ is computed as a weighted sum of individual losses:

$$L_{\text{total}} = w_T \cdot L_{T_{\text{onset}}}\left(\hat{y}_{T_{\text{onset}}}, y_{T_{\text{onset}}}\right) + w_E \cdot L_{E_a}\left(\hat{y}_{E_a}, y_{E_a}\right) \tag{5}$$

where $L_{T_{\text{onset}}}$ and $L_{E_a}$ are the mean squared error (MSE) losses for $T_{onset}$ and $E_a$, respectively, $y_{T_{\text{onset}}}$ and $y_{E_a}$ are the true values, and $w_T$ and $w_E$ are weights (set to 0.5 each) to enforce equal contribution of both targets during optimization. This approach mitigates bias toward one property and enhances overall predictive accuracy.

### 2.5. Machine learning model training and inverse design via genetic algorithm

The dataset was split into 80% training and 20% testing sets. Hyperparameter optimization was conducted using GridSearchCV and RandomizedSearchCV, paired with five-fold cross-validation (=5), to ensure model generalization. For example, in the XGBoost model (xgboost package, version 2.1.4), key hyperparameters such as learning rate, maximum depth, and number of estimators were tuned systematically. Cross-validation provided robust performance estimates

by averaging the loss across folds:

$$L_{CV} = \frac{1}{k}\sum_{i=1}^{k} L_{total,i} \quad (6)$$

Following model training, we employed a guided Genetic Algorithm (GA) to perform exploratory inverse design, leveraging the trained XGBoost model as a fitness evaluator to identify novel catalyst compositions with optimal performance. The algorithm explores a high-dimensional feature space including a constrained set of high-potential elements, hierarchical catalyst architecture, and process parameters. The detailed implementation of the GA, including its non-linear, Gaussian-based reward function and evolutionary operations, are provided in the Supporting Information (Supplemental Notes 1, Supporting Information).

### 2.6. LLM optimization and fine-tuning

As introduced earlier, the second stage of our framework involved creating the Cat Advisor multi-agent system. To power this agent, we optimized the open-source DeepSeek-R1-Distill-Llama-8B model for the specific requirements of catalyst research using parameter-efficient fine-tuning through the optimized Low-Rank Adaptation (LoRA) methodology from the Unsloth framework. This method, applied to improve task-specific performance, selectively updates rank-decomposed weight matrices, reducing trainable parameters by >90% while maintaining the inherent generalization capabilities of the base mode. Importantly, we incorporated the R1 reasoning module during fine-tuning to enhance the structured logical inference and multi-hop reasoning capabilities of the model, which are essential for iterative analysis and knowledge synthesis in complex scientific contexts. Fine-tuning was performed using an RTX A6000 Ada graphics card in an Ubuntu 22.04.5 LTS environment. All operations were conducted in Python 3.10, using Unsloth (version 2025.3.1), CUDA (version 12.4), PyTorch (version 2.6.0 + cu124), and Triton (version 2.2.0). Additionally, four-bit quantization was applied to reduce the GPU memory footprint to 7 GB, enabling efficient experimentation. The Chain-of-Thought (CoT) dataset (https://huggingface.co/datasets/Yy245/cot_2000) was used as the training corpus for fine-tuning. The fine-tuned DeepSeek-R1-Distill-Llama-8B model weights are publicly available on Hugging Face at https://huggingface.co/Yy245/Cat-Advisor. All relevant code is publicly available in the GitHub repository (https://github.com/Weijie-Yang/cat_advisor).

## 3. Results and discussion

### 3.1. LLM-Enhanced data acquisition and processing for $mgh_2$ dehydrogenation catalysts

Developing robust machine learning models for predicting magnesium hydride dehydrogenation catalyst (MDC) performance critically depends on readily accessible, high-quality, structured experimental data. Manual extraction of such data from scientific literature is laborious and inefficient, especially for MDC research where performance is intricately linked to numerous experimental parameters across diverse publications (Fig. S1, Supporting Information). To address this, we developed an LLM-enhanced framework to systematically and efficiently acquire experimental data from 759 Web of Science publications concerning "catalyst" and "$MgH_2$."

Our framework uses a multi-stage preprocessing pipeline (Fig. 1a) to optimize data extraction. Initially, Nougat [52] converted PDFs to Markdown. This Markdown conversion, with its simplified syntax and enhanced structural clarity compared to PDFs, significantly improved LLM parsing and comprehension of scientific text, particularly facilitating accurate identification of key experimental parameters and their relationships. Subsequently, iterative prompt [53] optimization (Fig. 1b), combining domain expertise with GPT-4o natural language processing, progressively enhanced the accuracy of targeted parameter extraction (see Figs. S2–S5 in the Supporting Information for prompt details). This iterative, human-machine collaborative approach refined prompts and yielded valuable insights for optimizing LLM performance in materials science literature processing.

Batch processing of 759 Markdown publications using the GPT-4o API and optimized prompts systematically extracted a dataset containing 2360 valid experimental data points for MDCs, derived from 809 unique catalyst entries. This dataset includes 16 critical experimental parameters (Table S1, Supporting Information) characterizing catalyst composition, synthesis, conditions, and key performance metrics like onset dehydrogenation temperature and activation energy. To ensure data consistency and enrich the dataset, facilitating both machine learning and multi-modal analysis, we constrained the LLM output to JSON and integrated materials property data from the Materials Project database [54] (Table S2, Supporting Information), along with publication metadata (DOI, year). This resulting structured database, with 6555 total data rows, significantly advances data availability for $MgH_2$ dehydrogenation catalyst research, dramatically accelerating data acquisition compared to manual methods while maintaining comparable accuracy.

Building on this comprehensive, high-fidelity dataset, which accelerates data availability for $MgH_2$ dehydrogenation catalyst research and greatly improves data acquisition efficiency over manual methods while maintaining comparable accuracy, we demonstrated its utility by developing advanced computational tools for MDC design. Specifically, to fully utilize this extensive MDCs database for computational catalyst design, we constructed machine learning models (Agent1) using extracted catalyst parameters (P1), catalyst classifications (P2), and Materials Project data. Agent1 employs a MultiOutputRegressor framework with a multi-objective loss function to optimize onset dehydrogenation temperature and activation energy concurrently, and incorporates a Genetic Algorithm for efficient candidate identification. Furthermore, to create a more interactive and knowledge-rich system, we fine-tuned a Large Language Model (LLM2) using Chain-of-Thought

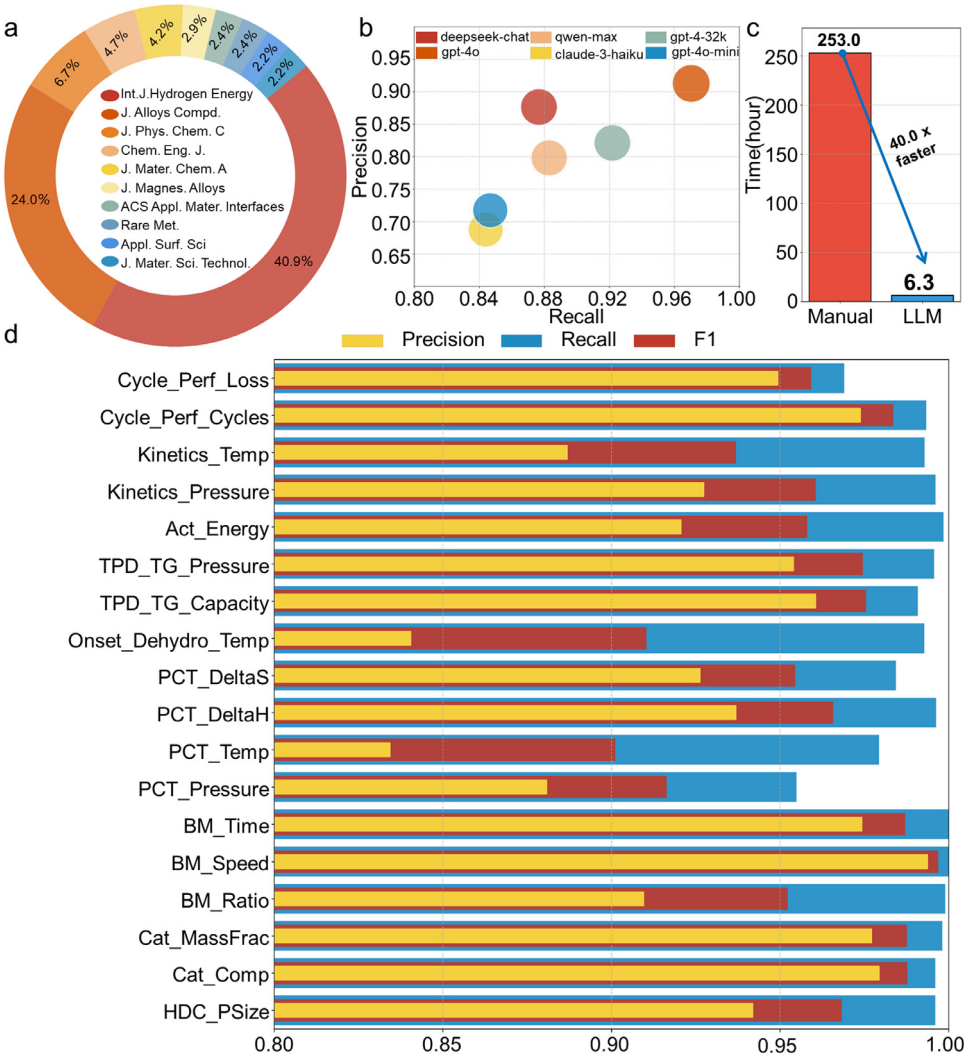


Fig. 2. Data Characterization and Performance Evaluation of Our Workflow. a) Journal Distribution of Publications. Distribution of retrieved publications across leading journals in hydrogen energy and materials science. b) LLM Performance Comparison for Information Retrieval. Comparative evaluation of mean precision, recall, and F1-score for DeepSeek, Qwen-max, and some other models, assessed using 78 randomly selected publications. Axes indicate recall and precision; node size represents dataset size rank. c) Comparison of the automated pipeline method vs. average manual execution time for 759 publications. d) GPT-4o Performance for Parameter Extraction. Mean precision, recall, and F1-score for each of the 16 extracted parameters, based on comparison with human annotations.

datasets (P3 & P4), while incorporating an external Retrieval-Augmented Generation (RAG)-enhanced knowledge base (P1 & P2). Function-calling techniques then integrated the machine learning models (Agent1) with the knowledge-enhanced LLM2 (Agent2), resulting in a robust multi-agent system for magnesium-based dehydrogenation catalysts (MDCs), whose functionalities and applications are detailed in subsequent sections.

### *3.2. Performance benchmarking of LLM-Driven text mining*

To provide a bibliometric context for magnesium-based hydrogen storage catalyst research, an initial analysis of journal distributions over the past 2 decades (Fig. 2a) indicated a notable concentration of publications in preeminent hydrogen energy and materials science journals, notably the International Journal of Hydrogen Energy and the Journal of Alloys and Compounds (full titles provided in Table S3, Supporting Information). This distribution underscores the role of these journals as critical platforms for disseminating substantial advancements in this field.

To rigorously benchmark our LLM-driven text mining approach, we established a test set of 78 randomly selected publications from our corpus of 759. Performance was evaluated by comparing information retrieval from DeepSeek, Qwen-max, GPT-4o and other models (model versions are listed in Table S4, Supporting Information), using consistent prompts, against a human-annotated gold standard. Standard metrics,precision, recall, and F1-score,were calculated. As shown in Fig. 2b, GPT-4o exhibited superior performance, achieving the highest F1-score and thus was selected as our primary information retrieval model.

Further comprehensive validation, involving manual assessment of 2360 catalyst data entries (∼10,000 parameters),

confirmed the high accuracy of GPT-4o across the full dataset (Fig. 2d). Consistently achieving F1-scores above 0.83 for all 16 key parameters (detailed metrics in Table S5, Supporting Information), GPT-4o demonstrated robust and reliable information extraction.

To further validate our methodology, we quantitatively compared its efficiency to manual parameter extraction by experienced researchers. Our automated approach achieved approximately a 40-fold acceleration in data extraction speed (Fig. 2c), reducing the estimated manual extraction time for 759 publications from approximately 253 h to just 6.3 h (A detailed breakdown of this efficiency comparison is provided in Fig. S6 and Supplemental Notes 3, Supporting Information). This dramatic reduction in time (from over a month of full-time work to less than a day of computation) highlights the transformative potential of our approach.

Building on the demonstrated efficiency, we next evaluated LLM performance for catalyst classification, a related task crucial for broader applicability and research trend analysis. Using prompt engineering (exemplary prompts in Fig. S3, Supporting Information), we employed GPT-4o to automatically categorize catalysts in our database into 12 categories [55] (Table S6, Supporting Information). Comparison against human-annotated classifications revealed high accuracy for LLM-driven catalyst categorization (Fig. S7, Supporting Information). Following manual verification, we integrated this classification data into our database, providing a robust foundation for subsequent research trend analysis.

### *3.3.1. Evolution of catalyst material trends and performance landscapes*

To understand the evolving research landscape in magnesium-based hydrogen storage catalysts (MDCs), we analyzed publication trends from 759 relevant articles over 2 decades, charting publication year against catalyst material categories (Fig. 3a). The temporal distribution reveals fluctuating growth in MDC publications since the early 2000s, with a recent peak around 2024, likely reflecting the cyclical nature of research interest and stages of technological advancement. Notably, early research (pre-2019) predominantly focused on binary and multi-metallic alloys, along with metal oxides, indicating initial efforts centered on conventional metal catalysts for $MgH_2$ dehydrogenation. However, in recent years, research has diversified significantly, with an increasing emphasis on composite, metal-carbon composite, and transition metal-based catalysts. This diversification, visualized in Fig. 3a, suggests an evolving trend towards more complex catalytic systems, including multi-component synergistic catalysis, nanocomposite materials, and strategies for precise active site modulation to enhance performance. Metal oxides, however, remain a consistently investigated category, likely due to their cost-effectiveness, facile synthesis, and chemical tunability [56].

This observed evolution in catalyst material trends, coupled with the persistent challenge of data scarcity in the field (as highlighted in the Introduction), directly motivated the construction of a comprehensive and high-quality dataset for machine learning applications. Building upon established methodologies, we assembled a dataset by integrating two distinct sources: (1) experimental parameters for 809 unique catalyst compositions extracted from 759 publications via our LLM pipeline, and (2) relevant computational materials properties (e.g., formation energy, band gap) retrieved from the Materials Project database [54] for these compositions.

To enhance model robustness and account for chemical reality where a single composition can exhibit multiple crystal structures, we implemented a data augmentation strategy. For each of the 809 compositions, we queried the Materials Project for all corresponding stable crystal structures (polymorphs). This one-to-many mapping resulted in a total of 6555 unique structure-property data rows. To simulate minor experimental variations and prevent overfitting on repeated performance values, a small random perturbation ($\pm 5\%$) was applied to the $T_{onset}$ and $E_a$ for these augmented entries.

This fusion of experimental performance data with structure-specific computational properties creates a rich, high-dimensional feature set for robust model training. As depicted in Fig. 3b, the resulting size of our dataset substantially surpasses those reported in comparable studies, addressing a critical limitation for data-driven catalyst design. The overall performance landscape captured by our dataset reveals a median onset dehydrogenation temperature of 240.00 °C and a median activation energy of 93.06 kJ/mol, with significant variance that highlights the complexity of the catalytic system (see Figs. S8 and S9, Supporting Information).

To further explore the performance landscape and the relationship between catalyst material type, onset dehydrogenation temperature ($T_{onset}$), and activation energy ($E_a$), we developed a Sankey diagram (Fig. 3c). This diagram effectively visualizes the distribution of catalyst categories across 10 discrete intervals for both $T_{onset}$ and $E_a$, balancing data granularity with sufficient sample sizes within each interval. The Sankey diagram (Fig. 3c) reveals a complex interplay between catalyst material, $T_{onset}$, and $E_a$. While the distribution of most catalyst types spans multiple $T_{onset}$ and $E_a$ ranges—confirming that material type alone is not a deterministic predictor of performance and highlighting the critical influence of factors such as microstructure, component synergy, and preparation methods—discernible trends still emerge. Specifically, metal-carbon composite catalysts (MCCs) and, to a lesser extent, metal oxides (MOx), show a notably higher proportion within the lower $E_a$ (<75.72 kJ/mol) and $T_{onset}$ ranges. For catalysts exhibiting $T_{onset}$ below 250 °C, the counts are: MCC (46), MOx (36), and bimetallic/multimetallic alloys (BMA, 20). Similarly, for catalysts with $E_a$ below 75.72 kJ/mol, the counts are: MOx (17), MCC (16), and BMA (13). These data, visualized in Fig. 3c, suggest that both MCCs and MOx demonstrate enhanced potential for achieving lower dehydrogenation temperatures and activation energies, although this trend appears more pronounced for MCCs, particularly regarding activation energy reduction.

For MCCs, this enhanced performance is plausibly attributable to the high surface area of carbon supports, which facilitates active component dispersion, improves electron

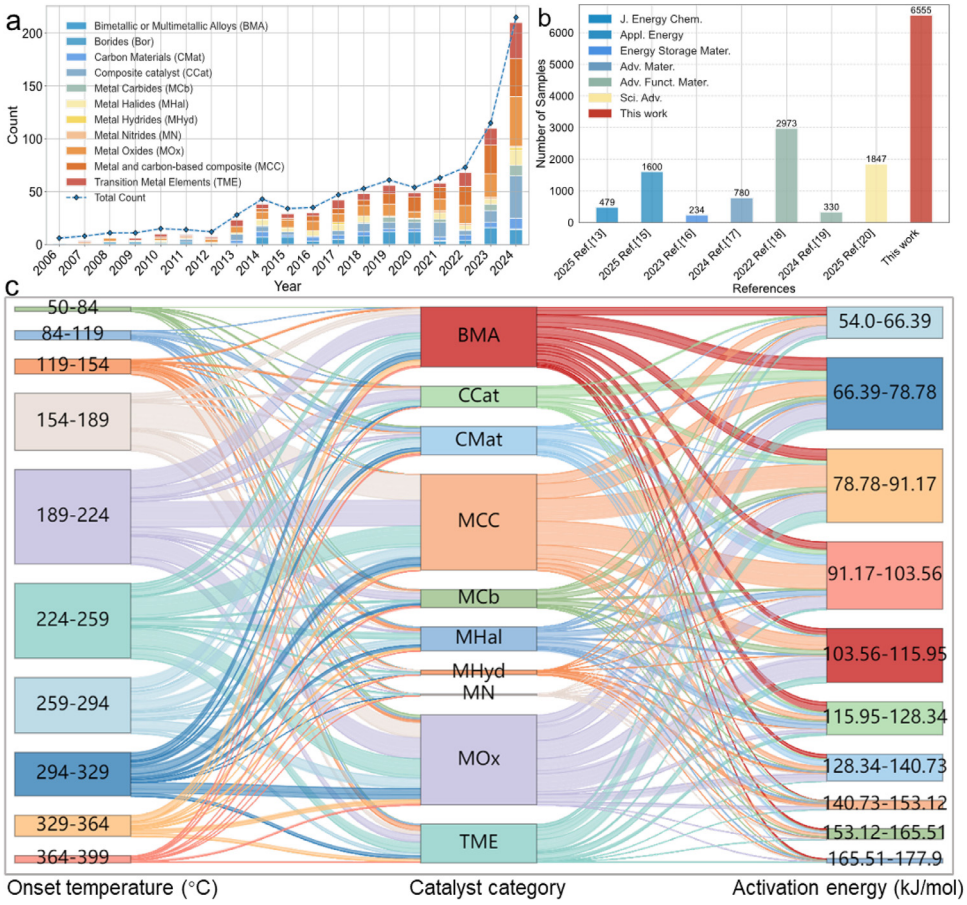


Fig. 3. Analysis of Catalyst Material Trends and Performance Data. a) Temporal Trends in Catalyst Material Categories. Stacked histogram illustrating the distribution of catalyst material categories across 759 publications over 2 decades (2004–2024). b) Dataset Size Comparison. Bar chart comparing the number of samples in the dataset compiled in this work with representative datasets from the literature in the fields of hydrogen storage and catalysis. c) Sankey Diagram of Catalyst Material Categories, Onset Dehydrogenation Temperature ($T_{onset}$), and Activation Energy ($E_a$). Sankey diagram visualizing the distribution and relationships between catalyst material categories and discretized ranges of $T_{onset}$ and $E_a$ (each divided into 10 intervals).

transport within $MgH_2$, and mitigates $MgH_2$ agglomeration [55]. Analogous to MCCs, metal oxides are also believed to enhance $MgH_2$ dehydrogenation kinetics through a combination of mechanisms, including reducing the activation energy, optimizing interfacial reactions, forming new catalytic phases, and enhancing hydrogen diffusion [57,58].

The superior performance of these two catalyst families is quantitatively supported by the detailed distribution analysis presented in the raincloud plots (Figs. S10 and S11, Supporting Information). These plots clearly show that Metal-Carbon Composites (MCCs) and Metal Oxides (MOx) possess some of the lowest median onset temperatures and activation energies among all categories. For instance, the median $T_{onset}$ for MCCs is visibly lower than that for categories like Metal Halides (MHal) or Bimetallic Alloys (BMA). Similarly, the distribution of $E_a$ for MCCs and MOx is skewed towards lower values, confirming their catalytic advantage. This observation is also consistent with the reported high performance of specific MCC examples like nanocrystalline $Mg_2Ni$/carbon and $Mg_2NiH_4$/carbon [59], further strengthening the evidence for the beneficial role of carbon supports in $MgH_2$ dehydrogenation catalysts, and highlighting the diverse catalytic mechanisms through which metal oxides, such as $Nb_2O_5$, $CeO_2$, $Fe_3O_4$, and $Co_3O_4$, can also contribute to enhanced performance [57,58].

In conclusion, while catalyst material significantly impacts dehydrogenation, the broad $T_{onset}$ and $E_a$ distributions within material types (Fig. 3c) emphasizes that effective catalyst design requires a holistic approach beyond material selection alone. Synergistic optimization of parameters like particle size, milling ratio, and catalyst loading is crucial for rationally designing and precisely controlling high-performance magnesium-based hydrogen storage catalysts. Future research should prioritize such multi-parameter optimization strategies.

### *3.3.2. Model performance prediction and feature engineering*

To enable accurate and efficient prediction of $MgH_2$ dehydrogenation catalyst (MDC) performance, we integrated catalyst material descriptors from the Materials Project with catalyst design parameters and experimental data from LLM-based text mining (prompt engineering and validation details in Figs. S2 and S3, Supporting Information). This integration yielded a comprehensive dataset (6555 samples, 61 raw descriptors). To imbue the model with chemical intuition and

Table 1
Performance Comparison of Machine Learning Models With Hierarchical Features.

| Model | XGBoost | GB | RF | DT |
|---|---|---|---|---|
| Average $R^2$ | 0.916 | 0.915 | 0.909 | 0.899 |
| Average MAE | 0.032 | 0.032 | 0.034 | 0.033 |
| Average RMSE | 0.082 | 0.082 | 0.086 | 0.089 |

enhance its interpretability, we developed a novel hierarchical feature engineering strategy. This approach deconstructs the identity of the catalyst into three distinct, chemically meaningful categories: 1) the active component (e.g., Metal Oxide, Bimetallic Alloy), 2) the support material (e.g., Carbon-based, MXene), and 3) the synthesis form (e.g., Supported, Core-Shell). These were combined with elemental presence features derived from the catalyst formula (e.g., Elem_Ti), key process parameters, and calculated physicochemical properties, resulting in a robust feature set for model training (see feature engineering details in Supplemental Notes 4, Supporting Information).

Using an 80:20 training-test split, the data was preprocessed to handle missing values, which were prevalent due to unreported parameters in the literature. We employed a K-Nearest Neighbors (KNN) imputer, an advanced method that estimates missing values based on the most similar data points in the feature space. Following imputation, the features were scaled using a RobustScaler to handle outliers effectively. We then developed and compared four established machine learning regression models (with detailed explanations provided in Supplemental Notes 2, Supporting Information), namely Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), and Extreme Gradient Boosting (XGBoost), within a MultiOutputRegressor framework to simultaneously predict MDC onset dehydrogenation temperature and activation energy, accounting for potential interdependencies. To ensure balanced learning, we employed an implicit multi-objective loss function [60]. GridSearchCV and RandomizedSearchCV, with five-fold cross-validation, systematically optimized hyperparameters [61], enhancing model generalization and reliability.

Performance benchmarking of the four models, now trained on the rich hierarchical feature set, is summarized in Table 1. To provide a comprehensive assessment of both model fit to the training data and generalization to unseen data, the metrics were averaged across both the training and test sets. The results indicate that all ensemble models achieved excellent overall performance, with XGBoost exhibiting the highest Overall Average $R^2$ of 0.916, marginally outperforming Gradient Boosting ($R^2$= 0.915). The consistently high performance across the top models highlights the effectiveness of our hierarchical feature engineering in creating a highly predictive feature space. Given its leading $R^2$ score, the XGBoost model was selected for all subsequent inverse design tasks. Detailed performance metrics for the training and test sets are provided separately in the Supporting Information (Table S7) for a granular analysis of model generalization. For a comprehensive visual evaluation, the prediction performance plots for the GB, DT, and RF models are also available in the Supporting Information (Figs. S12–S17).

To rigorously assess model generalization and diagnose potential overfitting, we generated learning curves for all four models (see Figs. S18–S25 in the Supporting Information). The learning curves for our best-performing model, XGBoost, are representative (see Figs. S18–S19 in the Supporting Information). They show that as the training set size increases, the $R^2$ score on the test set consistently improves and begins to plateau, while the gap between the training ($R^2$ = 0.964) and test ($R^2$=0.867) scores narrows. This behavior is characteristic of a well-generalized model that is benefiting from more data and not suffering from high variance (overfitting). It suggests that while a small generalization gap exists: typical for complex datasets, the performance of the model is robust and predictive.

Furthermore, SHAP (SHapley Additive exPlanations) analysis and feature importance assessments (Fig. 4c–d, Figs. S26–S32, Supporting Information) provided mechanistic insights and improved model interpretability. This analysis, now empowered by our hierarchical feature engineering, highlights not just catalyst composition in general, but specific elemental contributions (e.g., the crucial role of Titanium), catalyst architecture (e.g., the encoded active component), and process parameters (e.g., particle size and ball milling) as critical descriptors influencing MDC performance. This aligns with fundamental materials science principles and offers more granular, actionable guidance for targeted catalyst design.

To elucidate the underlying chemical and physical drivers of performance, we employed SHAP analysis [62] on the trained XGBoost model. The results (Fig. 4c–d) provide unprecedented, fine-grained insights thanks to our hierarchical feature engineering.

For onset temperature (Fig. 4c), process parameters like ball milling speed (BM_Speed) and time (BM_Time) remain critically important, underscoring the role of nanostructuring. Crucially, the elemental feature Elem_Ti (presence of Titanium) emerges as a top-ranking descriptor, with its presence (high feature value, red dots) strongly pushing the SHAP value to the left, indicating a significant contribution to lowering the onset temperature. Furthermore, the encoded active component (Active_Comp_Enc) is identified as a key factor, confirming that the intrinsic nature of the catalyst class is vital.

For activation energy (Fig. 4d), a similar trend is observed. The presence of Titanium (Elem_Ti) is again the most impactful feature, consistently driving the activation energy down. This provides strong, data-driven evidence for the exceptional catalytic role of Titanium in $MgH_2$ systems. Other important factors include process parameters (BM_Time, BM_Speed) and catalyst loading (Cat_MassFrac). The SHAP analysis indicates that our hierarchical and elemental features capture key patterns, transforming the model into a more interpretable tool for scientific inquiry.

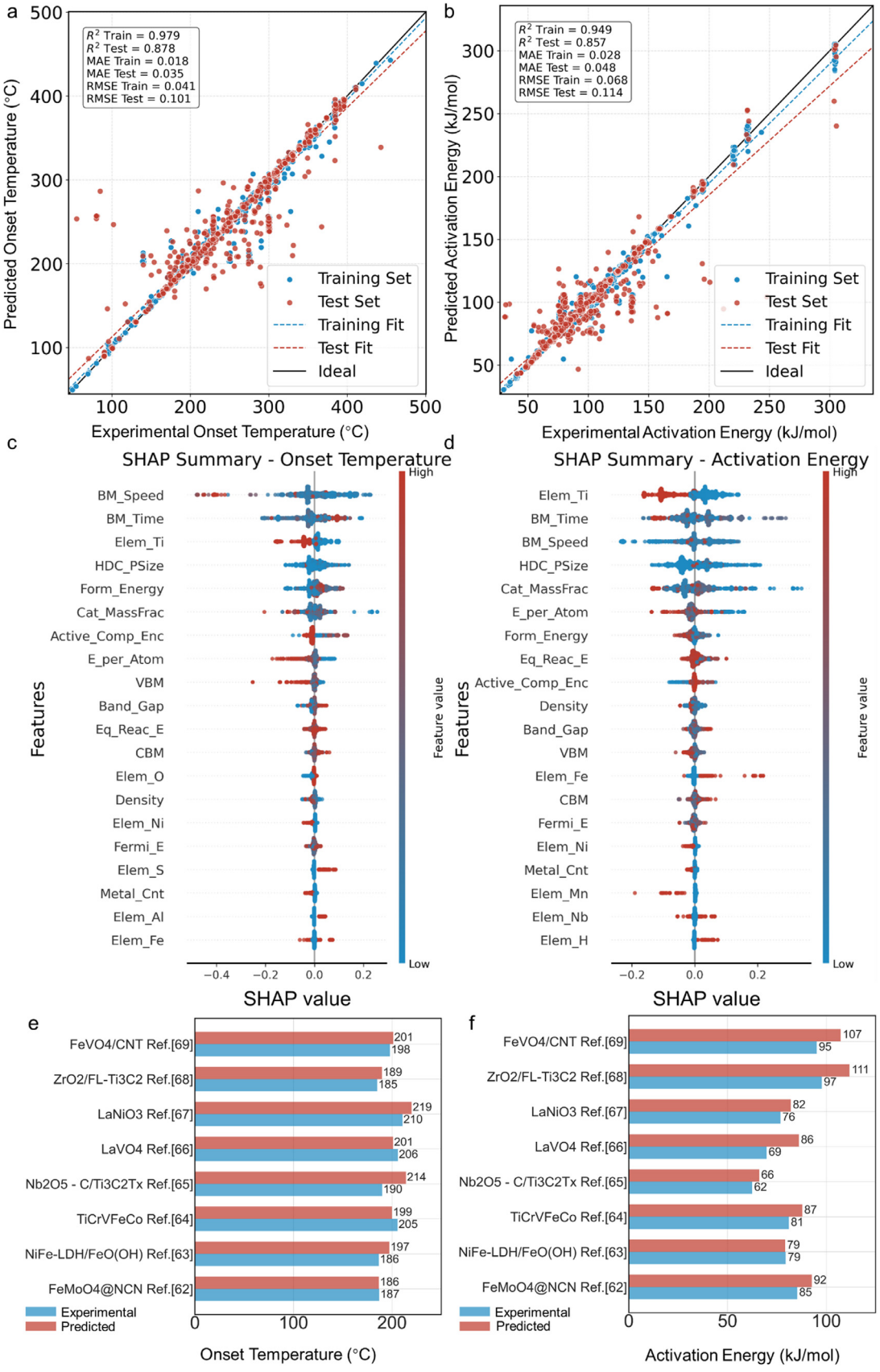


Fig. 4. XGBoost Model Performance and Feature Importance. a) XGBoost model performance for predicting $T_{onset}$. b) XGBoost model performance for predicting $E_a$. c) SHAP summary plot of feature importance for XGBoost model predictions of $T_{onset}$. d) SHAP summary plot of feature importance for XGBoost model predictions of $E_a$. e) Comparison of predicted vs. actual $T_{onset}$ for eight unseen catalysts. f) Comparison of predicted vs. actual $E_a$ for eight unseen catalysts.

*3.3.3. Inverse design and prospective validation of next-generation catalysts*

Having established a highly accurate and interpretable predictive model, we proceeded to the ultimate test of our data-driven framework: its ability to perform inverse design and its predictive power on completely unseen materials. To this end, we implemented a comprehensive, multi-faceted validation strategy using a curated set of 8 diverse, state-of-the-art catalyst systems from very recent literature (2025) [63–70]. These materials, including complex composites, high-entropy alloys, and LDH-derivatives, were entirely absent from our training data.

First, we assessed the forward predictive power of the trained model. The results, presented in Fig. 4e and f, compare the direct predictions from the model against the experimental values for the unseen catalysts. This excellent predictive capability across a challenging set of novel materials confirms that the model has learned the fundamental structure-property relationships and can accurately extrapolate to new regions of the chemical space, establishing a solid foundation for its use in inverse design.

Next, to guide the inverse design process towards chemically fertile and relevant spaces, we performed a statistical analysis to identify key elemental drivers of performance (Fig. 5a and b). This analysis revealed a core set of "effective elements" (e.g., Ti, V, Ni, Mn, C) consistently associated with improved performance. To bridge the knowledge of the model with the rapidly evolving experimental frontier, we augmented this data-driven element pool with additional elements featured prominently in the 2025 literature (e.g., Fe, Nb, Mo, Zr, La). This combined set of 16 high-potential elements formed the constrained chemical search space for our GA, ensuring a focused yet exploratory search (see Supplemental Notes 1 for the full list of elements, Supporting Information).

We then employed this guided GA to perform exploratory inverse design, aiming to identify novel catalyst compositions with optimal performance. The GA iteratively evolves a population of candidate catalysts, leveraging the trained XGBoost model as a fitness evaluator. To guide the search toward promising and physically realistic candidates, a nonlinear fitness function was implemented to reward catalyst designs with lower predicted $T_{onset}$ and $E_a$ (see Table S8 and Supplemental Notes 1 for details, Supporting Information). The evolution of the mean $T_{onset}$ and $E_a$ of the population over the generations is visualized in Fig. 5c and d, respectively. These plots demonstrate the effectiveness of the GA, showing a clear convergence towards lower temperatures and activation energies as the optimization progresses. The interquartile range (shaded area) also narrows, indicating the population is consistently evolving towards high-performance regions.

To validate the rationality of the search process, we visualized the high-dimensional feature space using t-distributed Stochastic Neighbor Embedding (t-SNE), as shown in Fig. 5e. The plot confirms that the GA-identified solutions are not outliers but are well-embedded within the data-supported manifold of the training set, alongside the literature targets. A comparative t-SNE visualization highlighting the feature space overlap is provided in the Supporting Information (Fig S33). The final performance distribution of the top 20 GA-identified candidates is compared against the training set in Fig. 5f. This plot vividly shows that the GA has successfully navigated the performance landscape to identify a cluster of candidates in the highly desirable low- $T_{onset}$, low- $E_a$ region.

The results of this exploratory inverse design are presented in Table 2, which details the top 20 highest-performing catalyst design blueprints discovered by the GA. Each blueprint consists of a predicted architecture (active component, support, form) and an associated elemental base. For use in the ML model, these three categorical architectural features were converted into numerical format using one-hot encoding, as detailed in the feature engineering section of the Supporting Information (Supplemental Notes 4, Supporting Information). This list showcases a diverse range of promising designs. A recurring theme is the prevalence of multi-metallic systems, often combined with advanced supports like MXenes or carbon-based materials.

The most powerful validation of our framework lies in demonstrating the alignment of these computationally generated blueprints with the forefront of experimental research. To this end, we conducted a systematic comparison between the findings in Table 2 and our curated set of state-of-the-art catalysts. Remarkably, the elemental bases and, in many cases, the predicted architectures of the six top-ranked GA-discovered blueprints show a strong correspondence to six distinct, high-performance catalyst families from our 2025 validation set. For instance, the top-ranked candidate (a blueprint for a Metal Oxide on a Carbon-based support with a C-Fe-O-V elemental base) corresponds precisely to the conceptual recipe for the $FeVO_4$/CNT catalyst. Similarly, the third-ranked candidate perfectly matches the architectural and elemental profile of a TiCrVFeCo high-entropy alloy. A detailed, side-by-side comparison of these six successfully aligned systems is provided in the Supporting Information (Table S9).

This successful "rediscovery" of multiple, diverse, and top-performing catalyst systems serves as a robust, data-driven alternative to direct experimental synthesis for validating our computational framework. Crucially, the remaining 14 top-ranked candidates, which do not correspond to any known systems in our validation set, represent novel, high-potential design hypotheses. These computationally-derived blueprints, marked as "Unreported" in Table 2, offer new avenues for future research and provide concrete, actionable insights for experimentalists aiming to synthesize next-generation catalysts. While the GA-predicted performances represent theoretical optima, the fact that the GA independently identified and prioritized the exact chemical combinations at the forefront of experimental research is a powerful testament to the ability of the framework to learn underlying chemical principles. This suggests that our framework is more than a fitting tool and can provide valuable insights to guide rational materials design , thereby supporting the credibility of the GA-generated hypotheses.

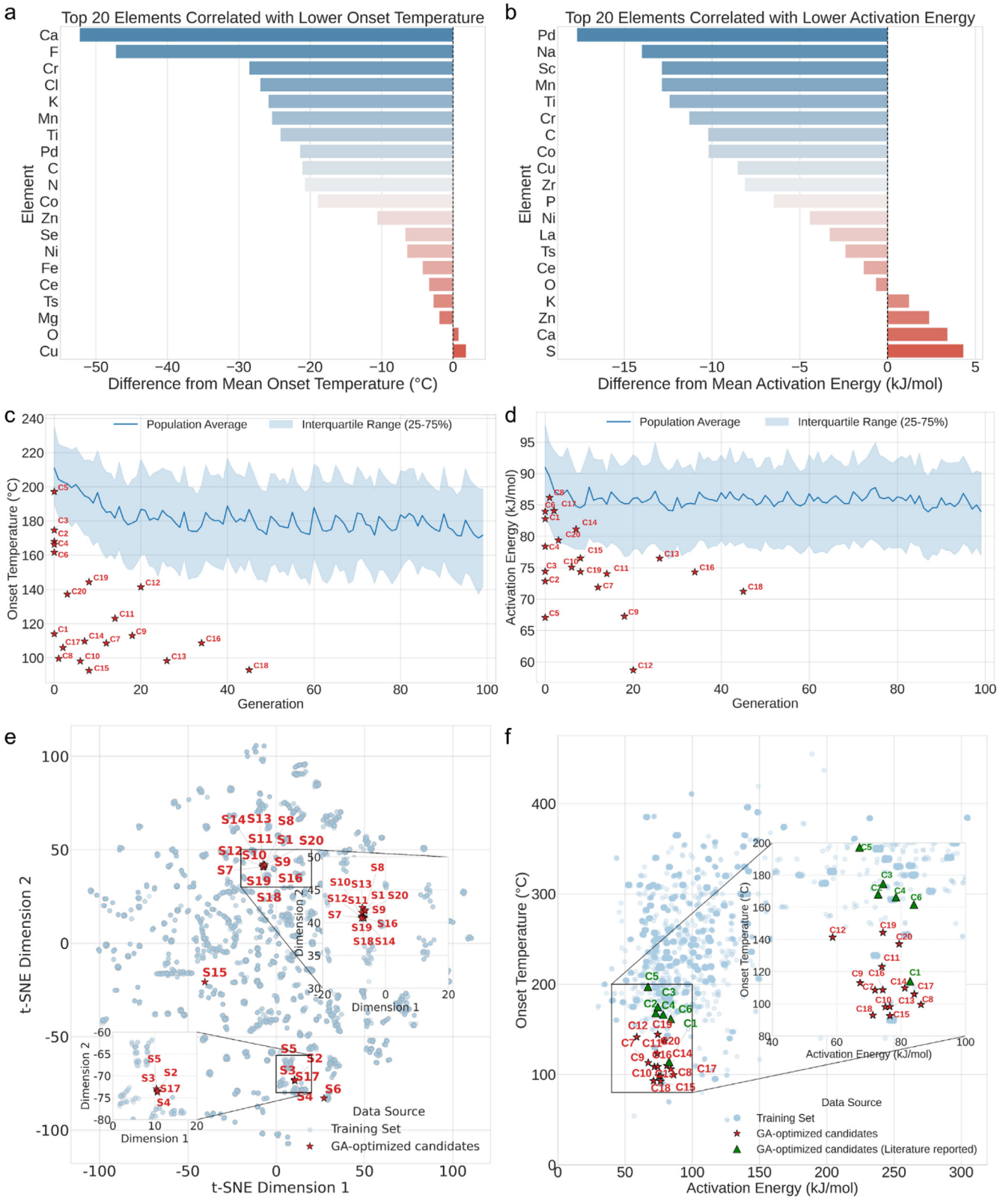


Fig. 5. Guided Inverse Design Framework and Validation. a) Statistical analysis identifying the top 20 elements whose presence most strongly correlates with a reduction in $T_{onset}$. b) A corresponding statistical analysis highlighting the top 20 elements most correlated with a reduction in $E_a$. c) Convergence plot illustrating the evolution of the mean $T_{onset}$ of the GA population, showing the average (solid line), interquartile range (shaded area), and discovery generations of top candidates (red stars). d) Convergence plot for the mean $E_a$ of the GA population, with metrics analogous to (c). e) A t-SNE visualization of the high-dimensional feature space, confirming that rediscovered literature targets and GA-identified solutions are well-embedded within the training set manifold. f) Performance landscape plot comparing the distribution of the top 20 GA-identified candidates against the training set, demonstrating the successful identification of a high-performance solution cluster by the GA.

Table 2
Top 20 Catalyst Candidates Identified by the Guided Genetic Algorithm.

| Number | GA-Identified optimum | | GA-Discovered elemental base | Predicted active component | Predicted support material | Predicted synthesis form | Notes |
|---|---|---|---|---|---|---|---|
| | $T_{onset}$ ( °C) | $E_a$ (kJ/mol) | | | | | |
| 1 | 114.01 | 82.79 | C-Fe-O-V | Metal oxide | Carbon-based | Supported | Reported [70] |
| 2 | 168.06 | 72.87 | Co-Cr-Fe-Ti-V | Composite/Mixture | No support | Alloyed | Reported [66] |
| 3 | 174.61 | 74.41 | C-Nb-O-Ti | Composite/Mixture | MXene | Supported | Reported [65] |
| 4 | 166.25 | 78.4 | C-Fe-Mo-N-O | Metal oxide | Carbon-based | Supported | Reported [63] |
| 5 | 197.18 | 67.08 | C-O-Ti-Zr | Composite/Mixture | No support | Self-supported | Reported [69] |
| 6 | 161.62 | 83.97 | La-O-V | Metal oxide | No support | Self-supported | Reported [67] |
| 7 | 108.54 | 71.89 | C-Cr-Fe-O-V | Metal oxide | Carbon-based | Supported | Unreported |
| 8 | 99.56 | 86.18 | C-Fe-Nb-O-V | Metal oxide | Carbon-based | Supported | Unreported |
| 9 | 113.05 | 67.28 | C-Fe-Mn-O-V | Metal oxide | Carbon-based | Supported | Unreported |
| 10 | 98.09 | 75.09 | C-Fe-O-V-Zr | Metal oxide | Carbon-based | Supported | Unreported |
| 11 | 123.06 | 74.04 | C-Fe-O-Ti-V | Metal oxide | Carbon-based | Supported | Unreported |
| 12 | 141.35 | 58.72 | C-Ca-Fe-O-V | Composite/Mixture | MXene | Supported | Unreported |
| 13 | 98.25 | 76.5 | C-Fe-Mg-O-V | Transition metal elements | Carbon-based | Supported | Unreported |
| 14 | 109.78 | 81.16 | C-Fe-O-Ts-V | Metal oxide | Carbon-based | Supported | Unreported |
| 15 | 92.55 | 76.52 | C-Fe-La-O-V | Metal oxide | Carbon-based | Core-shell | Unreported |
| 16 | 108.65 | 74.32 | C-Fe-K-O-V | Transition metal elements | Others | Supported | Unreported |
| 17 | 106 | 84.09 | C-Fe-Ni-O-V | Bimetallic/Multimetallic alloy | MOF | Self-supported | Unreported |
| 18 | 92.92 | 71.24 | C-Fe-V | Others | Carbon-based | Supported | Unreported |
| 19 | 144.3 | 74.36 | C-Nb-O-Ti-V | Composite/Mixture | No support | Core-shell | Unreported |
| 20 | 137.16 | 79.4 | C-Fe-O-Se-V | Metal oxide | Carbon-based | Supported | Unreported |

#### *3.3.4. Cat-advisor: a multi-agent system for $mgh_2$ dehydrogenation catalyst design*

While LLMs demonstrate broad text comprehension, their generalist nature can limit efficacy in specialized scientific domains like catalyst design, which demand deep domain expertise and high factual accuracy, often leading to issues like "hallucinations" [71]. To overcome these limitations and accelerate $MgH_2$ dehydrogenation catalyst research, we developed Cat-Advisor (https://cat-advisor.cpolar.top/chat/share?shareId=wfz7t90zohng7vaej6b1dxgd) (Fig. 6a), a multi-agent system specifically tailored for this field. Cat-Advisor integrates predictive modeling, knowledge retrieval, and advanced LLM reasoning capabilities to provide specialized guidance for catalyst design.

The architecture of Cat-Advisor comprises two primary agents:

- Agent 1 (Predictor): employs ML models trained on a structured database. This database combines P1 (catalyst and performance data) and P2 (catalyst category) extracted using Prompts 1 & 2 with information from the Materials Project. Agent 1 predicts key performance metrics, namely the onset dehydrogenation temperature and activation energy, for diverse catalysts. The applicability of these predictions is supported by t-SNE visualization, which confirms that the input feature combinations lie within the distribution density of the training data.
- Agent 2 (Advisor): leverages Retrieval-Augmented Generation (RAG) for context-aware literature retrieval and utilizes advanced reasoning capabilities. Its core LLM functionality integrates a specifically fine-tuned DeepSeek/Llama-architecture model (DeepSeek-R1-Distill-Llama-8B) with GPT-4, employing a multi-routing strategy to optimize performance across different tasks [72], such as recommendation generation and complex Q&A. The fine-tuning corpus included P3 (extracted question-answer (Q&A) pairs) and P4 (Chain-of-Thought [CoT] Q&A pairs) extracted using Prompts 3 & 4. Agent 2 utilizes function-calling to interact dynamically with Agent 1 (for predictive data) and external knowledge bases, enabling it to generate targeted experimental design recommendations and provide informed scientific answers.

The development involved domain-specific fine-tuning of the open-source DeepSeek-R1-Distill-Llama-8B model (8B parameters), selected considering factors like performance and adaptability. Fine-tuning was performed using the Unsloth framework with Low-Rank Adaptation (LoRA) (details in Supplemental Notes 1 and Figs. S34–S38, Supporting Information). The fine-tuning dataset was derived from a corpus of 759 $MgH_2$ dehydrogenation publications, generating a CoT dataset (2225 instances) and a self-awareness Q&A dataset (25 pairs; prompt details in Figs. S4 and S5, Supporting Information). This process significantly enhanced the domain-specific question-answering capabilities of the model (Fig. 6b; further details in Figs. S39 and S40, Supporting Information), bolstering its advisory functions.

To further enhance reliability and mitigate potential LLM hallucinations, Cat-Advisor incorporates a multi-layered RAG strategy [73] (details in Fig. S41, Supporting Information). A hybrid retrieval approach combining semantic vector search with full-text search identifies relevant literature passages.

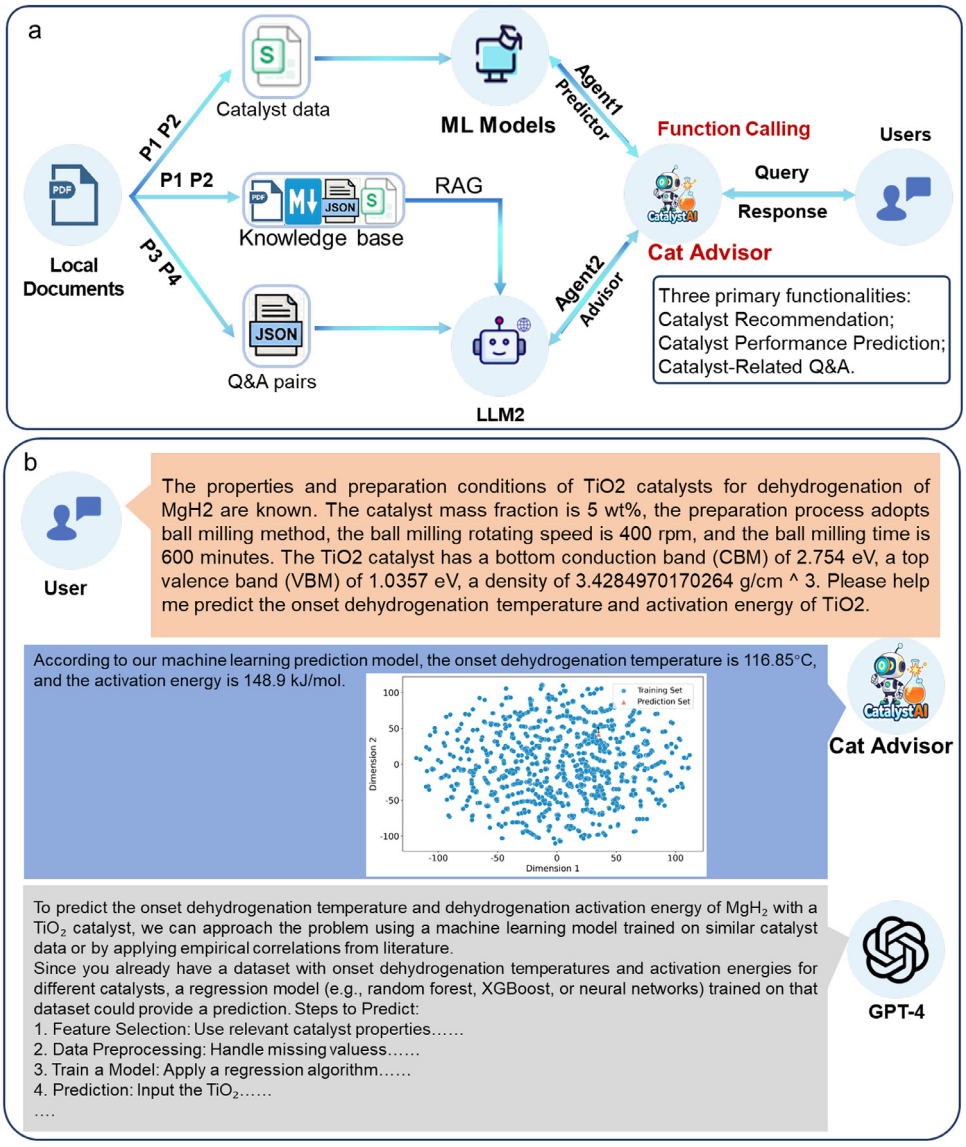


Fig. 6. Cat-Advisor Multi-Agent System for $MgH_2$ Dehydrogenation Catalyst Design. a) Cat-Advisor System Architecture. Diagram illustrating the multi-agent system architecture, integrating Agent 1 (Predictor) and Agent 2 (Advisor) with RAG-enhanced knowledge retrieval. b) Cat-Advisor Performance Benchmarking. Comparison of Cat-Advisor vs. general-purpose LLMs (native GPT-4) in domain-specific question-answering tasks related to $MgH_2$ dehydrogenation catalysts.

These passages are then prioritized using Reciprocal Rank Fusion (RRF) and a re-ranking model before being integrated into the context of the LLM for generating informed responses.

The seamless integration between the predictive power of Agent 1 and the reasoning of Agent 2 is enabled by a function-calling mechanism. This allows Agent 2, during a conversation with a user, to programmatically invoke Agent 1 as if it were an external tool. For example, when a user requests a performance prediction for a specific catalyst, the reasoning process of Agent 2 identifies the need for a quantitative prediction. It then formats the user input into the feature vector required by the machine learning model in Agent 1 and executes the predict performance function call. Agent 1 runs the prediction and returns the numerical output (e.g., "$T_{onset}$ = 90 °C, $E_a$ = 77 kJ/mol"). Agent 2 subsequently receives this output and dynamically embeds it into its natural language response to the user. This real-time integration of predictive data into the generative process is what allows Cat-Advisor to provide quantitatively grounded advice. As our evaluations demonstrate, this integrated system significantly enhances accuracy and contextual relevance, allowing Cat-Advisor to outperform general-purpose LLMs like native GPT-4 on domain-specific tasks (Fig. 6b).

Cat-Advisor offers researchers three primary functionalities (Practical usage illustrated in Fig. S42 and Video, Supporting Information):

- **Catalyst recommendation:** suggests potential catalysts or composite materials tailored to user-defined performance

criteria (e.g., target ranges for $T_{onset}$ and $E_a$). Agent 2 utilizes its specialized knowledge and RAG capabilities to identify promising candidates and associated synthesis strategies. These suggestions can then be evaluated by Agent 1 for predicted performance.
  - *Illustrative Example:* a user specifies a target of $T_{onset}$ <100 °C (373 K) and $E_a$ <80 kJ/mol. Agent 2 might propose $Mg(AlH_4)_2$ based on its domain knowledge about complex hydrides potentially acting as catalysts [74]. Agent 1 then predicts the performance for this material under specified conditions (e.g., predicting $T_{onset}$ = 90 °C and $E_a$ = 77 kJ/mol based on its ML model). This predicted performance meets the criteria specified by the user. (Note: The reliability of the prediction by Agent 1 is implicitly supported by the validation methods, including t-SNE, described previously).
- **Catalyst performance prediction:** predicts $T_{onset}$ and $E_a$ for a specific catalyst based on user-provided properties and preparation conditions. Input can be supplied through structured input fields or natural language descriptions. Agent 1 performs the prediction using its trained ML models. (Note: As above, the reliability of the prediction is supported by prior validation).
  - *Example Natural Language Input:* "Predict the onset dehydrogenation temperature and activation energy for a $TiO_2$ catalyst (5 wt%) prepared by ball milling (400 rpm, 600 min) with known properties: CBM = 2.754 eV, VBM = 1.0357 eV, density = 3.428 g/cm³."
- **Catalyst-Related Q&A:** addresses scientific inquiries regarding $MgH_2$ dehydrogenation catalysts, covering aspects like reaction mechanisms, structure-property relationships, and the impact of synthesis parameters or additives, leveraging the RAG-enhanced LLM capabilities of Agent 2.
  - *Example User Query:* "how does the carbon shell structure influence the catalytic activity of Ni nanoparticles in $MgH_2$ composites?"

In summary, Cat-Advisor, as a multi-agent system, represents a notable advance in applying AI to the complex challenge of catalyst discovery. This hybrid predictive-advisory approach can accelerate the screening process and help streamline experimental workflows. The underlying multi-agent framework offers potential as a generalizable AI research platform adaptable to other scientific domains requiring integrated prediction and knowledge synthesis.

### 3.4. Framework limitations and future outlook

While our "LLM to Agent" framework demonstrates a powerful new paradigm for materials discovery, it is crucial to acknowledge its current limitations, which in turn define clear avenues for future enhancement.

A primary methodological concern is the inherent risk of LLM "hallucination," that is, the generation of factually incorrect information. To ensure data integrity across the pipeline, we implemented a multi-layered mitigation strategy including: (i) **Strict Prompt Engineering** with a predefined JSON output format; (ii) **Automated Post-Processing** for validation and unit conversion; and (iii) **Expert-in-the-Loop Validation** of a significant data subset. For the Cat-Advisor multi-agent system, these principles were further strengthened by its RAG architecture.

From a scientific scope perspective, another significant limitation of the framework is its current focus on catalyst activity($T_{onset}$ and $E_a$), while neglecting the equally critical aspect of catalyst stability and recyclability. This omission was a deliberate and carefully considered choice, driven by the profound challenges associated with the available literature data for these metrics. Unlike temperature and activation energy, which are typically reported as precise numerical values, our comprehensive literature survey revealed that stability data are often presented in: (i) non-standardized formats (e.g., "capacity loss of 5% after 20 cycles" vs. "retains 90% capacity after 10 h at 300 °C"); (ii) qualitative descriptions (e.g., "good stability," "excellent recyclability"); and (iii) highly variable experimental conditions that hinder direct comparison. This heterogeneity makes it extremely challenging to reliably extract and structure these metrics into a consistent, machine-learnable format without introducing significant noise and uncertainty.

These limitations, however, point directly to exciting future research directions and highlight the potential of our framework. A key goal is to develop more sophisticated LLM extraction techniques, leveraging the advanced parsing capabilities of models like GPT-4o to interpret, parse, and standardize these complex, often narrative-based stability reports. Successfully curating a reliable stability dataset would, in turn, enable a true multi-objective optimization that simultaneously targets high activity and long-term durability, a critical step towards designing practically viable catalysts.

Furthermore, the robustness of the framework could be significantly enhanced by exploring more advanced methodologies. One key direction is LLM ensembling, where outputs from multiple top-performing models (e.g., GPT-4o, Claude 4, Gemini 2.5) are systematically compared. By developing a sophisticated consensus algorithm to resolve discrepancies, this approach could further mitigate model-specific biases and create a more reliable "self-correcting" data pipeline, building upon the strong, expert-validated baseline established in this work.

Finally, the greatest potential of the framework lies in its generalizability. Its modular design, which encompasses data extraction, predictive modeling, and agent-based reasoning, is not specific to magnesium hydride and can be readily adapted to accelerate discovery in other data-rich chemical domains. This includes creating new databases and predictive models for challenges such as designing electrocatalysts, screening next-generation battery materials, or discovering novel high-performance polymers. This would extend the impact of our work far beyond the specific system studied here, establishing a versatile and powerful template for AI-driven materials science.

## 4. Conclusion

This study establishes an AI-driven framework, embodying an "LLM to Agent" paradigm, for catalyst discovery in magnesium-based hydrogen storage. By leveraging a foundational LLM (GPT-4o) for the automated curation of an extensive chemical dataset, we constructed a robust foundation for data-driven design. High-fidelity ML models, built upon this dataset, subsequently enabled accurate performance prediction and served as the core of a guided Genetic Algorithm.

A key finding of this work is that the framework can move beyond performance prediction to generate scientific insights. The GA-driven inverse design process identified a set of promising design principles for next-generation catalysts, such as the importance of multi-metallic synergy and the strategic use of advanced supports. The credibility of these computationally derived principles was supported by their alignment with the design strategies employed in recent (2025) experimental literature. This demonstrates that the framework has the potential to provide useful, high-level guidance by generating viable catalyst design blueprints, each comprising an elemental base and a hierarchical architecture.

A notable component of this framework is Cat-Advisor, a domain-adapted multi-agent system. This system represents an approach to translate static predictions into an interactive format. Fine-tuned with a purpose-built Chain-of-Thought (CoT) dataset, Cat-Advisor synergizes ML predictions with a RAG-enhanced knowledge base to deliver context-aware chemical design recommendations. This work explores the potential of specialized AI Agents to assist in complex scientific reasoning, aiming to reduce reliance on empirical trial-and-error and aid in intelligent chemical interpretation.

In summary, our integrated "LLM to Agent" framework presents a systematic approach to how chemical knowledge can be extracted, understood, and utilized for discovery. This work provides a practical blueprint for developing AI tools that can navigate the complex chemical literature and data landscape. We believe this approach holds potential to catalyze innovation not only in $MgH_2$ catalyst development but also in broader fields such as electrocatalysis and advanced battery materials, contributing to the overall pace of scientific research and development.

## Data Availability

The comprehensive dataset supporting the findings of this study is publicly available through the repository at *Digital Hydrogen-S* (http://digital-hydrogen.com/storage/). Additionally, supplementary information relevant to this study, encompassing detailed methodologies and additional analytical data, is provided in the Supporting Information. The associated CoT dataset is publicly available on Hugging Face at https://huggingface.co/datasets/Yy245/cot_2000. The fine-tuned DeepSeek-R1-Distill-Llama-8B model weights are publicly available on Hugging Face at https://huggingface.co/Yy245/Cat-Advisor. All relevant code is publicly available in the GitHub repository (https://github.com/Weijie-Yang/cat_advisor). An online web application demo of Cat-Advisor is available at (https://cat-advisor.cpolar.top/chat/share?shareId=wfz7t90zohng7vaej6b1dxgd).

## Declaration of competing interest

The authors declare no competing interests.

## CRediT authorship contribution statement

**Tongao Yao:** Writing – original draft, Software, Methodology, Data curation. **Yang Yang:** Software, Data curation. **Jianghao Cai:** Methodology, Data curation. **Rui Liu:** Visualization, Software. **Zhaoyan Dong:** Visualization, Software, Data curation. **Xiaotian Tang:** Visualization, Formal analysis. **Xuqiang Shao:** Writing – review & editing. **Zhengyang Gao:** Writing – review & editing. **Guangyao An:** Writing – review & editing. **Weijie Yang:** Writing – review & editing, Resources, Methodology, Funding acquisition.

## Acknowledgments

This research was also supported by the Natural Science Foundation of Hebei Province (E2023502006) and Fundamental Research Fund for the Central Universities (2025MS131).

## Supplementary materials

The Supporting Information associated with this article provides comprehensive methodological details, supplementary data, and additional analyses. This includes: detailed descriptions of the LLM prompt engineering and data curation workflow; the mathematical formulations for the machine learning models and the Genetic Algorithm; a quantitative breakdown of the data extraction efficiency; a step-by-step guide to our hierarchical feature engineering; and a complete set of supplementary tables and figures, including full model performance metrics, learning curves, SHAP analyses.

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jma.2025.08.021.

## References

[1] L. Schlapbach, A. Züttel, Nature 414 (2001) 353–358.
[2] T. Sadhasivam, et al., Renew Sustain Energy Rev 72 (2017) 523–534.
[3] H. Wang, et al., J Alloys Compd 658 (2016) 280–300.
[4] C. Pistidda, et al., J Power Sources 270 (2014) 554–563.
[5] H. Guan, J. Liu, X. Sun, D. Li, Z. Zhou, F. Li, et al., Adv Mater 36 (24) (2025) 2500178.
[6] H.Y. Wang, H. Li, J. Wei, et al., Adv Funct Mater 34 (42) (2024) 2406639.
[7] P. Sharma, et al., J Magnes Alloys 12 (5) (2024) 1792–1798.
[8] P. Sharma, N.S. Rohila, A. Tiwari, Magnesium Alloys Structure and Properties, IntechOpen, London, 2022.
[9] X. Qin, Q. Wang, X. Zhao, et al., J Magnes Alloys (2025) https://doi.org/10.1016/j.jma.2025.06.005.
[10] P.P. Zhou, P. Zhou, Q. Xiao, et al., Adv Mater 37 (6) (2025) 2413430.
[11] S.Y. Dong, et al., Int J Hydrogen Energy 48 (97) (2023) 38412–38424.
[12] H. Kurban, M. Kurban, Comput Mater Sci 195 (2021) 110490.

[13] S.Y. An, S. Li, K. Zhu, et al., Front Mater 11 (2024) 1364572.
[14] H. Kurban, M. Kurban, M.M. Dalkilic, Sci Rep 12 (1) (2022) 14403.
[15] H. Kurban, et al., Key Eng Mater 880 (2021) 89–94.
[16] K. Li, et al., J Mater Process Technol 318 (2023) 117997.
[17] C. Polat, M. Kurban, H. Kurban, Mach Learn Sci Technol 5 (4) (2024) 045062.
[18] P.P. Zhou, P. Xiao, X. Zhu, et al., Energy Storage Mater 63 (2023) 102964.
[19] Q.H. Zhang, Q. Yuan, Y. Zhang, et al., Adv Mater 36 (36) (2024) 2404981.
[20] L.T. Chen, L. Tian, Y. Hu, et al., Adv Funct Mater 32 (47) (2022) 2208418.
[21] X. Zhang, X. Ding, B. Wang, et al., Adv Funct Mater 34 (30) (2024) 2314529.
[22] R. Ding, R. Liu, J. Hua, et al., Sci Adv 11 (14) (2025) eadr9038.
[23] OpenAI, et al., GPT-4 Technical Report. arXiv preprint arXiv:2303.08774, (2023).
[24] A. Priyanshu, Y. Maurya, Z. Hong, AI Governance and Accountability: An Analysis of Anthropic's Claude. arXiv preprint arXiv:2407.01557, (2024).
[25] G. Team, et al., Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805, (2023).
[26] DeepSeek-AI, et al., DeepSeek-V3 Technical Report. arXiv preprint arXiv:2412.19437, (2024).
[27] Qwen, et al., Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115, (2024).
[28] Z.L. Zheng, et al., J Am Chem Soc 145 (32) (2023) 18048–18062.
[29] W. Zhang, et al., Chem Sci 15 (27) (2024) 10600–10611.
[30] S.X. Leong, et al., Chem Sci 15 (43) (2024) 17881–17891.
[31] Z. Zheng, Z. He, O. Khattab, et al., Digit Discov 3 (3) (2024) 491–501.
[32] K. Chen, et al., Chemist-X: Large Language Model-empowered Agent for Reaction Condition Recommendation in Chemical Synthesis. arXiv preprint arXiv:2311.10776, (2023).
[33] A.M. Bran, M. Bran, A. Cox, O. Schilter, et al., Nat Mach Intell 6 (5) (2024) 525–535.
[34] Z.L. Zheng, Z. Florit, F. Jin, et al., Angew Chem Int Ed 64 (6) (2025) e202418074.
[35] S. Tao, T. Song, M. Luo, X. Zhang, et al., J Am Chem Soc 147 (15) (2025) 12534–12545.
[36] Z.L. Zheng, et al., J Am Chem Soc 145 (51) (2023) 28284–28295.
[37] H. Wang, et al., Efficient evolutionary search over chemical space with large language models. arXiv preprint arXiv:2406.16976, (2024).
[38] J.M. Parrilla-Gutiérrez, et al., Nat Comput Sci 4 (3) (2024) 200–209.
[39] J.T. Li, et al., IEEE Trans Knowl Data Eng 36 (11) (2024) 6071–6083.
[40] Y. Kang, J. Kim, Nat Commun 15 (1) (2024) 4705.
[41] N. Janakarajan, et al., Language models in molecular discovery. arXiv preprint arXiv:2309.16235, (2023).
[42] A.D. McNaughton, et al., ACS Omega 9 (46) (2024) 46563–46573.
[43] H.W. Sprueill, et al., ChemReasoner: Heuristic search over a Large Language Model's knowledge space using quantum-chemical feedback. arXiv preprint arXiv:2402.10980, (2024).
[44] Z.L. Zheng, R. Zheng, N. Rampal, et al., Angew Chem Int Ed 62 (46) (2023) e202311983.
[45] Z.L. Zheng, et al., ACS Cent Sci 9 (11) (2023) 2161–2170.
[46] D.A. Boiko, et al., Nature 624 (7992) (2023) 570–578.
[47] N. Yoshikawa, N. Skreta, M. Darvish, et al., Robots 47 (8) (2023) 1057–1086.
[48] M. Livne, Z. Miftahutdinov, E. Tutubalina, et al., nach0: Multimodal natural and chemical languages foundation model. arXiv preprint arXiv:2311.12410, (2023).
[49] S. Kumar, et al., J Magnes Alloys 12 (8) (2024) 3216–3228.
[50] Z.Y.A.H.L. Zeng, Z. Wang, et al., Mater Genome Eng Adv 2 (4) (2024) e77.
[51] Q. Zhang, Q. Hu, Y. Yan, et al., Adv Mater 36 (32) (2024) 2405163.
[52] L. Blecher, et al., Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418, (2023).
[53] L. Wang, L. Chen, X. Deng, et al., Npj Digit Med 7 (1) (2024) 41.
[54] A. Jain, S.P. Ong, G. Hautier, et al., Apl Materials 1 (1) (2013) 011002.
[55] J. Zhang, S. Yan, H. Qu, Int J Hydrogen Energy 43 (3) (2018) 1545–1565.
[56] H. Ishaq, I. Dincer, C. Crawford, Int J Hydrogen Energy 47 (62) (2022) 26238–26264.
[57] S.T. Sabitu, A.J. Goudy, Metals (Basel) 2 (3) (2012) 219–228.
[58] J.H. Wang, et al., Int J Hydrogen Energy 70 (2024) 61–70.
[59] X. Lu, X. Lu, X. Zhang, L. Zheng, et al., J Alloys Compd 905 (2022) 164169.
[60] X. Pengcheng, et al., J Mater Inform 5 (2) (2025) 26.
[61] L. Ziliang, et al., J Mater Inform 4 (4) (2024) 19.
[62] S.M. Lundberg, et al., Explainable AI for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610, (2019).
[63] J.Q. Zhang, L. Bian, N. Zhang, et al., J Energy Storage 130 (2025) 117418.
[64] Y.P. Chen, et al., Int J Hydrogen Energy 121 (2025) 326–336.
[65] J. Deng, et al., Int J Hydrogen Energy 137 (2025) 83–94.
[66] C. Peng, X.X. Chen, Q.A. Zhang, J Alloys Compd 1014 (2025) 178709.
[67] M.H. Wu, et al., J Magnes Alloys 13 (2) (2025) 613–625.
[68] Y.A. Liu, et al., J Alloys Compd 1032 (2025).
[69] F.Q. Bu, et al., J Mater Chem A 13 (21) (2025) 16102–16111.
[70] R.Y. Zhang, R. Zhang, H. Sun, et al., J Alloys Compd 1029 (2025) 180767.
[71] L. Huang, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232, (2023).
[72] C. Varangot-Reille, et al., Implementing routing strategies in large language model-based systems: An extended survey. arXiv preprint arXiv:2502.00409, (2025).
[73] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401, (2020).
[74] Y. Wang, et al., Int J Hydrogen Energy 39 (31) (2014) 17747–17753.