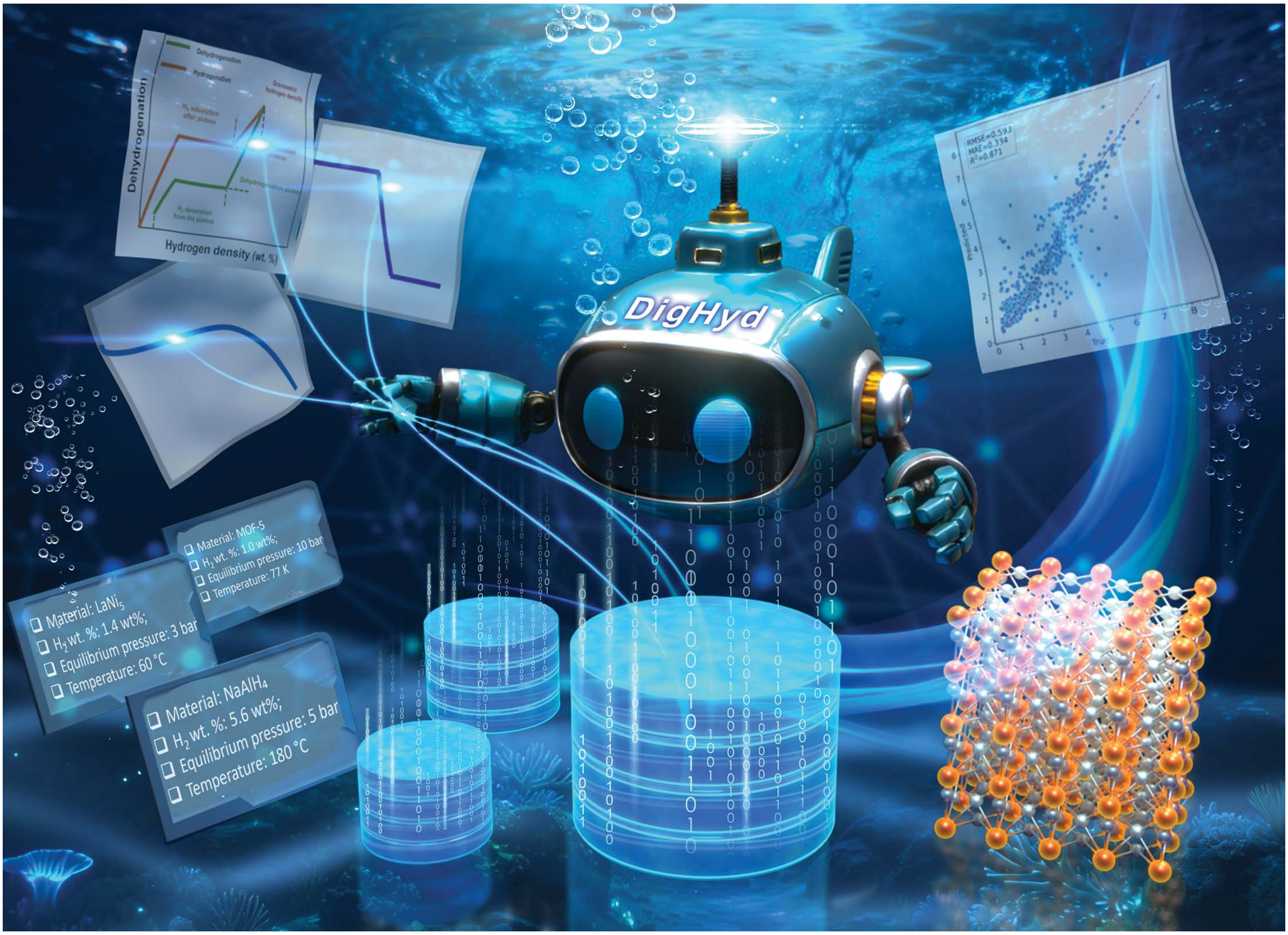


Showcasing research from Advanced Institute for Materials Research (WPI-AIMR), Tohoku University, Japan.

"DIVE" into hydrogen storage materials discovery with AI agents

Autonomous AI-driven materials discovery is still at an early stage. The data-extraction multi-agent system "DIVE" translates literature visuals into descriptive text and structured data to build the Digital Hydrogen Platform (DigHyd), enabling an AI agent to perform H_2-storage inverse design in minutes.

Image reproduced by permission of Hao Li from *Chem. Sci.*, 2026, **17**, 3031.

As featured in:

See Di Zhang, Shin-ichi Orimo, Hao Li *et al.*, *Chem. Sci.*, 2026, **17**, 3031.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

Registered charity number: 207890

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2026, **17**, 3031

All publication charges for this article have been paid for by the Royal Society of Chemistry

"DIVE" into hydrogen storage materials discovery with AI agents

Di Zhang, *[a] Xue Jia, [a] Hung Ba Tran, [a] Seong Hoon Jang, [a] Linda Zhang, [ab] Ryuhei Sato,[c] Yusuke Hashimoto,[b] Toyoto Sato, [e] Kiyoe Konno, [d] Shin-ichi Orimo *[ae] and Hao Li *[a]

Received 18th December 2025
Accepted 18th January 2026

DOI: 10.1039/d5sc09921h

rsc.li/chemical-science

Despite the surge of AI in energy materials research, fully autonomous workflows that connect high-precision experimental knowledge to the discovery of credible new energy-related materials remain at an early stage. Here, we develop the Descriptive Interpretation of Visual Expression (DIVE) multi-agent workflow, which systematically reads and organizes experimental data from graphical elements in scientific literature. Applied to solid-state hydrogen storage materials—a class of materials central to future clean-energy technologies—DIVE markedly improves the accuracy and coverage of data extraction compared to the direct extraction method, with gains of 10–15% over commercial models and over 30% relative to open-source models. Building on a curated database of over 30 000 entries from >4000 publications, we establish a rapid inverse-design AI workflow capable of proposing new materials within minutes. This transferable, end-to-end paradigm illustrates how multimodal AI agents can convert literature-embedded scientific knowledge into actionable innovation, offering a scalable pathway for accelerated discovery across chemistry and materials science.

Introduction

Data-driven approaches are increasingly reshaping the paradigm of materials discovery and design,[1–4] with the integration of large language models (LLMs) and automated workflows opening new frontiers for accelerated chemical innovation.[5–7] A central requirement for realizing this vision is the ability to construct and utilize reliable, high-precision materials knowledge at scale;[8,9] yet much of the experimental information in the literature remains unstructured and underexploited, limiting the full impact of AI in advancing both fundamental science and technological deployment. Moreover, rapidly assembling an effective agent or workflow for specific materials problems also remains a substantial barrier.[10]

The recent surge in LLM applications has greatly enhanced the prospects for automated data mining and reasoning in materials science. Leveraging advanced LLMs, several studies have explored automated extraction of materials data from scientific literature using prompt engineering and conversational interfaces.[11–13] Despite these advances, existing strategies still suffer from limitations in completeness, depth, and precision—especially when extracting key quantitative information from graphical elements, which often encode critical materials properties. Current state-of-the-art multimodal models, while powerful, often require multiple rounds of prompt-based querying and validation, resulting in significant computational cost and inefficient use of token resources. There remains a lack of systematic workflows for one-shot, high-throughput extraction and for rigorous, quantitative benchmarking against human-curated data. Moreover, there is no widely adopted workflow for rapidly constructing collaborative, multi-agent materials design systems based on newly mined datasets.

Recent work has increasingly adopted multi-step, multimodal pipelines to mine scientific literature beyond plain text by explicitly incorporating figures, tables, and cross-modality constraints. For example, OpenChemIE provides an information-extraction toolkit for chemistry literature that integrates multimodal extraction components.[14] In electrosynthesis, MERMES demonstrates an end-to-end multimodal workflow that leverages multimodal LLMs to parse reaction diagrams and resolve cross-modality dependencies from publications.[15] In materials/reticular chemistry, Zheng *et al.* show that GPT-4V can be used to categorize and mine diverse graphical sources (*e.g.*, isotherms and diffraction patterns) at scale.[16] More broadly, MERMaid proposes sequential modules for figure/table segmentation and multimodal analysis to convert PDF-embedded chemical information into machine-actionable representations.[17] Building on these advances, we

[a] *Advanced Institute for Materials Research (WPI-AIMR), Tohoku University, Sendai 980-8577, Japan. E-mail: di.zhang.a8@tohoku.ac.jp; shin-ichi.orimo.a6@tohoku.ac.jp; li.hao.b8@tohoku.ac.jp*

[b] *Frontier Research Institute for Interdisciplinary Sciences (FRIS), Tohoku University, Sendai 980-8577, Japan*

[c] *Department of Materials Engineering, The University of Tokyo, Tokyo 113-8656, Japan*

[d] *Institute of Fluid Science, Tohoku University, Sendai, 980-8577, Japan*

[e] *Institute for Materials Research, Tohoku University, Sendai, 980-8577, Japan*

develop the Descriptive Interpretation of Visual Expression (DIVE) workflow, a multi-agent workflow designed for high-throughput extraction of quantitative, figure-centric materials data (*e.g.*, PCT/TPD/discharge curves), coupled with an embedding-based evaluation protocol and a downstream inverse-design agent for hydrogen storage materials discovery. Although conceptually simple, the DIVE pipeline achieves significant gains over current open-source and commercial models, as confirmed by rigorous manual validation and scoring. Afterward, we apply DIVE to the domain of solid-state hydrogen storage materials (HSMs)—a field critical for the future of sustainable, carbon-neutral energy.[18] Hydrogen's high gravimetric energy density and environmentally benign combustion make it an ideal candidate for large-scale energy storage;[19] yet practical deployment hinges on the development of compact, safe, and cost-effective storage technologies. Solid-state HSMs, including interstitial hydrides, complex borohydrides, ionic compounds, porous frameworks, and emergent high-entropy and superhydride phases, offer a promising path forward. Despite decades of research, however, no comprehensive, structured experimental database for hydrogen storage materials currently exists.

In this work, we systematically mine over 4000 primary publications on solid-state HSMs, spanning the period from 1972 to 2025, using the DIVE workflow and optimized prompt engineering. Compared to leading multimodal and open-source models, DIVE achieves improvements of 10% to 15% and 30%, respectively, in accuracy and data completeness. The resulting database comprises more than 30 000 entries, which we leverage to construct a materials design agent (*DigHyd*) using GPTs. This agent supports natural language interaction with the HSM database and, more importantly, incorporates a machine-learning-based verifier trained on the extracted data. By integrating LLM-driven reasoning and iterative validation, we realize a streamlined materials design workflow capable of proposing novel hydrogen storage candidates that meet user-defined criteria within minutes (SI Video 1–3). Overall, this work delivers an efficient, scalable framework for AI-driven materials research and offers a transferable methodology for rapid database construction and inverse design in diverse materials domains.

Results and discussion

Fig. 1 shows the traditional workflow for extracting materials data from the literature using LLMs (Fig. 1a), as well as the schematic of our proposed DIVE workflow (Fig. 1b). In the conventional approach, the PDF file of a materials science article is first converted into text (*e.g.*, markdown format) and images. These are then directly fed into a multimodal LLM, which outputs a structured database. In contrast, our DIVE workflow introduces a much more detailed process, particularly for extracting key material properties that are often presented in figures. For HSMs, these include pressure–composition–temperature (PCT) curves, temperature-programmed desorption (TPD) curves, and discharge curves. First, a lightweight inference model scans the article's figure captions to determine whether these key figures are present. If so, the corresponding figure, its caption, and the relevant surrounding text are input into a second multimodal LLM. By carefully designing and optimizing prompts, the LLM is instructed to extract the key points from each curve in the figure, placing the results in the correct positions (as shown in Fig. 1b, prompt design), the complete prompt template used for image-based data extraction is provided in Fig. S14 and in prompt_template.py of our DIVE codebase (https://github.com/gtex-hydrogen-storage/DIVE). The extracted text then replaces the original figure in the article. We name this approach as the descriptive interpretation of visual expression, as we essentially transform visual information into descriptive text. Finally, the modified article, now with images represented as text, is input into a third LLM for the final extraction of key data (for details on all model combinations, refer to the SI).

Equally important to the multi-agent workflow is the development of effective evaluation methods. To the best of our knowledge, there is currently no well-established method for evaluating the accuracy and completeness of data extraction from articles using LLMs. To save tokens, it is common to extract multiple entries in one call, making the JSON dictionary list format particularly suitable for outputs. However, how to efficiently and reasonably compare human-extracted and AI-extracted JSONs and assign meaningful scores remains under-explored. This is particularly challenging in materials property extraction, where extraction quality cannot be judged simply as true or false, because the magnitude of numerical differences should also be considered. To address this, as shown in Fig. 1c, we propose using an embedding model to match entries between the human and AI-extracted JSONs. After matching, the units of numerical values are standardized, and the relative errors are calculated using mathematical functions to provide nuanced scoring. We divide the final score into accuracy and completeness (each normalized to 50 points, for a total of 100). In this way, hallucinated entries are not explicitly filtered at the completeness stage; instead, hallucinations are penalized during accuracy evaluation. Specifically, each AI-extracted entry is forcibly matched to the most similar ground-truth entry using an embedding-based alignment. As a result, hallucinated or severely inaccurate entries receive low per-item scores and are systematically penalized in the final accuracy metric. We use a 10% relative-error tolerance as a pragmatic choice for figure-derived values. This level is sufficiently strict to penalize clear misreads, while remaining compatible with the current limitations of multimodal models in precise visual digitization (*e.g.*, axis tick resolution, curve overlap, and image quality). This method allows for a more scientific and rapid evaluation of LLM data extraction performance and can also serve as a reward function for reinforcement learning to further fine-tune or train LLMs. The detailed evaluation functions, as well as the code for the DIVE workflow, are available in the GitHub repository in the Data and code availability section provided with this article. A representative example comparing ground-truth annotations and AI-extracted structured data is shown in SI, Table S2, further illustrating completeness and accuracy scores. To ensure the high reliability and scientific value of HSM data, the

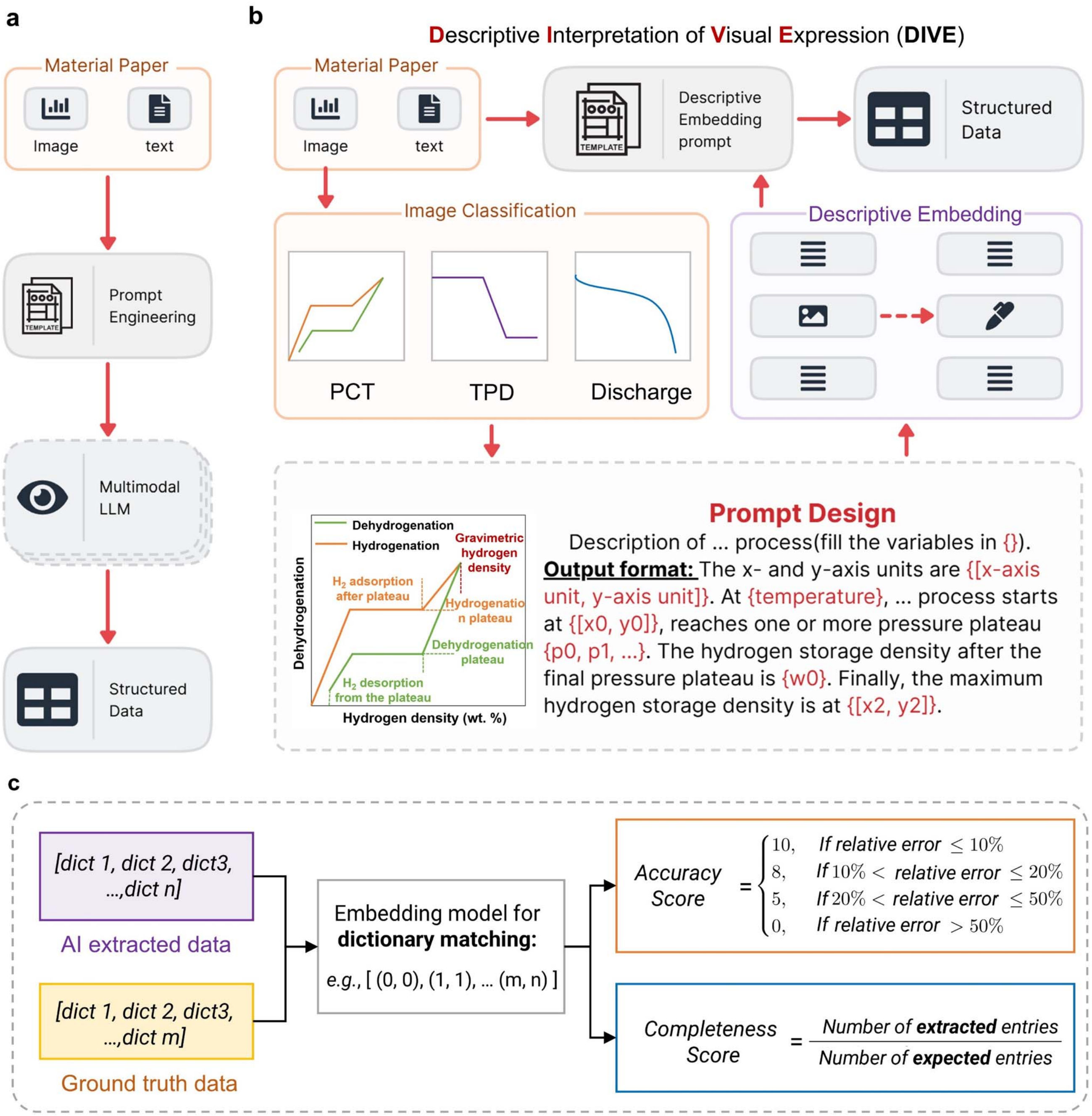

a
b
Descriptive Interpretation of Visual Expression (DIVE)
Material Paper
Image
text
Prompt Engineering
Multimodal LLM
Structured Data
TEMPLATE
Descriptive Embedding prompt
Image Classification
PCT
TPD
Discharge
Descriptive Embedding
Prompt Design
Dehydrogenation
Hydrogenation
Gravimetric hydrogen density
H2 adsorption after plateau
Hydrogenation plateau
Dehydrogenation plateau
H2 desorption from the plateau
Hydrogen density (wt. %)
Description of ... process(fill the variables in {}).
Output format: The x- and y-axis units are {[x-axis unit, y-axis unit]}. At {temperature}, ... process starts at {[x0, y0]}, reaches one or more pressure plateau {p0, p1, ...}. The hydrogen storage density after the final pressure plateau is {w0}. Finally, the maximum hydrogen storage density is at {[x2, y2]}.
c
[dict 1, dict 2, dict3, ...,dict n]
AI extracted data
[dict 1, dict 2, dict3, ...,dict m]
Ground truth data
Embedding model for dictionary matching:
e.g., [(0, 0), (1, 1), ... (m, n)]
Accuracy Score = 10, If relative error ≤ 10%; 8, If 10% < relative error ≤ 20%; 5, If 20% < relative error ≤ 50%; 0, If relative error > 50%
Completeness Score = Number of extracted entries / Number of expected entries

Fig. 1 Schematic diagram and evaluation methods of the DIVE workflow. (a) Conventional extraction pipeline based on a single multimodal LLM. (b) DIVE extraction pipeline, where descriptive prompts embed key data points and generate image replacements for structured data extraction. (c) Evaluation method for batch extraction accuracy. Both AI-extracted and manually annotated data are formatted as lists of dictionaries. A shared embedding model is used to match values across dictionaries, from which numerical values are retrieved to calculate precision and completeness scores.

DigHyd Data Checking System (curvechecking.dighyd.org; refer to the SI for details) has been developed as an efficient online platform for manual review and correction of AI-extracted data. The diversity of the test set can be found in Fig. S16.

Based on our developed DIVE workflow and the associated scoring algorithm for materials literature data extraction, we systematically evaluated several state-of-the-art commercial and open-source large language models. The score distributions for data extracted by different combinations of multimodal models and LLMs in the DIVE workflow are benchmarked against a dataset consisting of results manually curated from 100 published articles on experimental HSM reports. Fig. 2a presents the data extraction scores for the conventional direct extraction approach and under the DIVE workflow (Fig. 2b and c). Gemini-2.5-Flash,[20] currently Google's best model in terms of price-performance, achieved a total score of 77.89 when used for

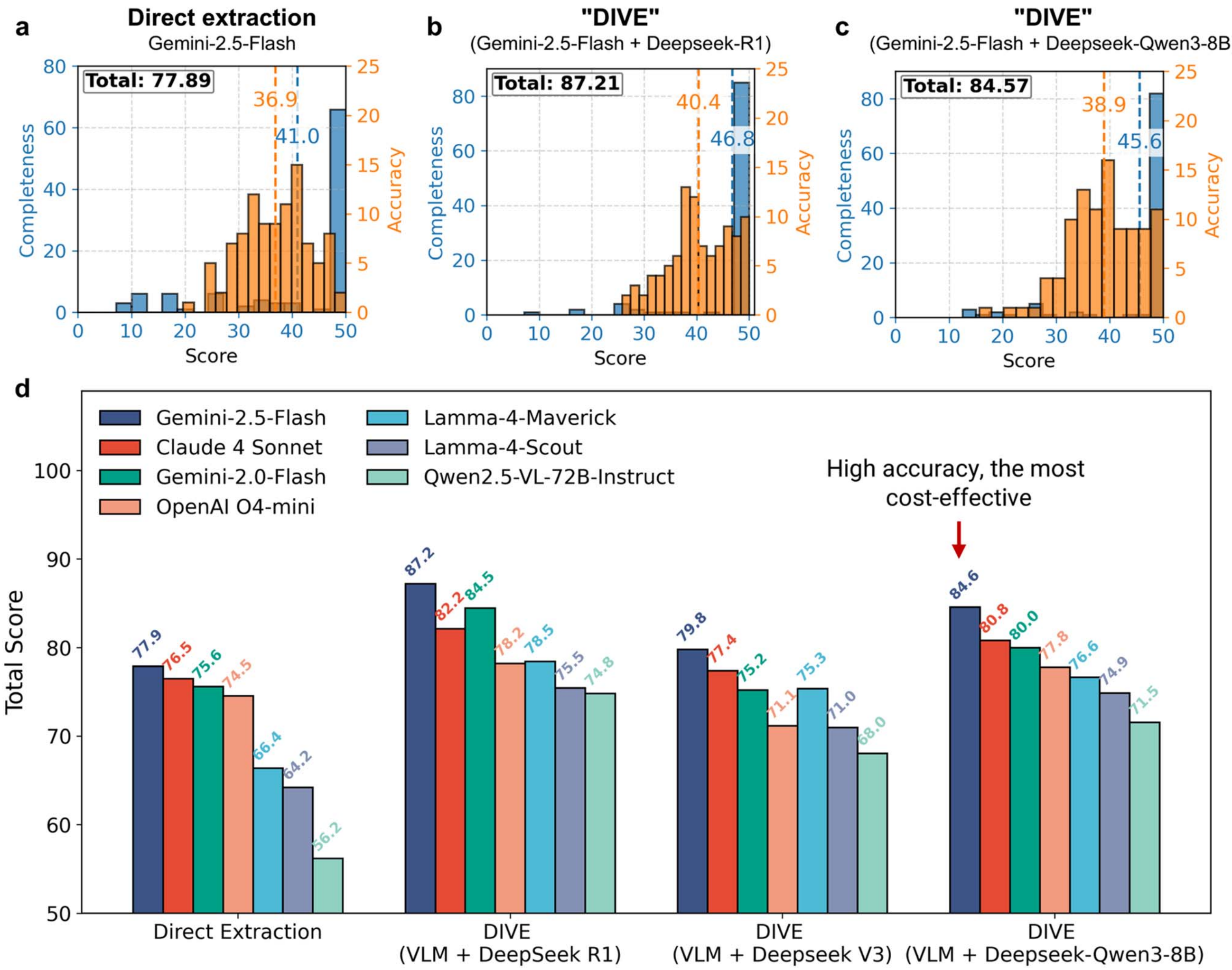


Fig. 2 Performance improvement of the DIVE data extraction workflow. (a) Conventional extraction workflow using Gemini 2.5 Flash.[20] (b) DIVE workflow integrating Gemini 2.5 Flash with DeepSeek R1. (c) DIVE workflow integrating Gemini 2.5 Flash with DeepSeek Qwen3 8B. Dotted vertical lines indicate the mean score of each corresponding score distribution. (d) Benchmark comparison across seven multimodal models, including four proprietary models (Gemini 2.5 Flash,[20] Claude 4 Sonnet, OpenAI o4 mini, and Gemini 2.0 Flash) and three open-source models (LLaMA-4-Scout, LLaMA-4-Maverick, and Qwen2.5-VL-72B-Instruct[22]). Ideally, the proposed DIVE workflow achieves a 10–15% improvement in extraction performance compared to state-of-the-art commercial models, and an over 30% improvement over leading open-source models.

direct extraction. However, when combined in a multi-stage, multi-agent DIVE workflow (Gemini-2.5-Flash[20] + DeepSeek R1 (ref. 21)), the total score increased to 87.21 (Fig. 2b), representing an improvement of nearly 12%. To further demonstrate the effectiveness of the DIVE workflow on models with even better token efficiency, we also tested DeepSeek-Qwen3-8B.[21] Despite having only 8B parameters, the model also showed about a 10% improvement compared to Gemini-2.5-Flash in the direct extraction scenario. In addition, we systematically assessed the data extraction accuracy across different combinations of mainstream commercial and open-source multimodal and text extraction models (all detailed results can be found in the SI). As shown in Fig. 2d, for the direct extraction workflow, most commercial models achieved a total score of around 75, whereas open-source models scored noticeably lower. When the multi-stage, multi-agent DIVE workflow is applied—particularly with DeepSeek R1 as the post-descriptive embedding LLM—commercial models saw typical improvements of 10–15%, and open-source models improved by 15–30%. The highest score was achieved with the combination of Gemini 2.5 Flash and DeepSeek R1. However, DeepSeek R1 is a large inference model with 685B parameters, making it relatively costly and slow. The same memory budget that supports one DeepSeek-R1-class deployment can typically support dozens of concurrent DeepSeek-Qwen3-8B instances, enabling substantially higher throughput for large-scale processing. Therefore, we further tested DeepSeek V3 and DeepSeek-Qwen3-8B as post-embedding LLMs. Surprisingly, despite its much smaller size (8B parameters), DeepSeek-Qwen3-8B achieved a total score of 84.6, second only to the Gemini 2.5 Flash + DeepSeek R1 combination, but with much faster inference speed and significantly lower computational cost.

Based on the above benchmarking, we ultimately selected the combination of Gemini 2.5 Flash and DeepSeek-Qwen3-8B for data extraction across 4053 publications. The screening strategy for selecting article DOIs is described in the SI. The processed data have been made publicly available in our Digital Hydrogen Platform (*DigHyd*: https://www.dighyd.org/). Fig. 3 provides an overview of data mining results from over 4000 hydrogen storage materials publications. As shown in Fig. 3a, aside from the years before 2010, the number of experimental publications on hydrogen storage materials has steadily increased, with 150–200 papers published annually since 2011 (except for 2021 and 2022, likely due to the global COVID-19 pandemic).

Fig. 3b shows the distribution of gravimetric hydrogen densities for different types of hydrogen storage materials. Porous carbon materials generally exhibit very low hydrogen storage capacities at room temperature. At low temperatures (*e.g.*, 77 K) and moderate pressures (*e.g.*, below 100 bar), their hydrogen uptake is typically in the range of 0–1 wt%. One of the main advantages of these materials lies in their extremely fast adsorption and desorption kinetics. Therefore, in the hydrogen storage range of 0–1 wt%, porous materials are the primary candidates.[23] The region with the highest concentration is between 1 and 2 wt%, which mainly corresponds to interstitial hydrides—the most widely studied class of hydrogen storage materials. In contrast, ionic, complex, and multi-component hydrides primarily fall in the 4–8 wt% range. By analyzing the extracted formula fields in the DIVE-generated data dictionaries, we can examine the elemental distribution in hydrogen storage materials across different gravimetric density ranges. The most frequent elements in the 0–4 wt%, 4–8 wt%, and 8–12 wt% intervals are Ni, Mg, and Li, respectively, reflecting a general shift in hydrogen storage materials from interstitial hydrides (represented by $LaNi_5$,[24,25] Ti–Mn alloys,[26] or high-entropy alloys[27]) to ionic hydrides (MgH_2) and complex hydrides ($LiBH_4$ (ref. 28) or $Mg(BH_4)_2$ (ref. 29)). Fig. 3c and d show the proportion of different types of materials in the *DigHyd* platform. Interstitial hydrides account for the largest share, but we also include a small number of superhydrides. Although superhydrides are mainly reported for superconducting applications,[30] they are emerging as a new research hotspot for hydrogen storage under ultra-high pressure conditions. Fig. 3d further illustrates the subtypes of interstitial hydrides.

After constructing the *DigHyd* database, direct data mining enables the extraction of valuable insights for materials design. Fig. 4 illustrates the top five most frequently added elements to typical hydrogen storage materials—$LaNi_5$, MgH_2, and $LiBH_4$—and the distribution of key performance metrics for materials modified with these elements. For $LaNi_5$, magnesium is the most commonly used dopant. After Mg is added, the gravimetric hydrogen density of $LaNi_5$-based materials can reach 4–6 wt% (Fig. 4b). However, the introduction of Mg also affects the hydrogen absorption and desorption pressures. In the case of MgH_2, nickel is the most frequent additive.[31] While doping MgH_2 with Ni tends to improve its hydrogen storage density (Fig. 4e), the dehydrogenation temperature of Mg–Ni systems can reach around 600 K. For $LiBH_4$-based systems, the gravimetric hydrogen density spans the widest range (0–14 wt%). Notably, introducing carbon or nitrogen can boost the hydrogen density of $LiBH_4$ materials to ~14 wt%, likely due to the catalytic effects of graphene or N-doped graphene on $LiBH_4$ (ref. 32 and 33) dehydrogenation. However, despite this high hydrogen density potential, the dehydrogenation temperature of $LiBH_4$ systems also tends to be relatively high, often requiring 700–800 K for complete hydrogen release. All the visualizations shown in Fig. 3 and 4 can be directly accessed and interacted with *via* our AI agent using natural language (SI Video 4).

Despite decades of research, most HSMs still fall short of the U.S. Department of Energy (DOE) 2030 technical targets for onboard hydrogen storage systems: >5.5 wt% system-level hydrogen capacity, >40 g H_2 per L volumetric density, operational capability between −40 to 85 °C, and cycling durability exceeding 1500 charge–discharge cycles.[34] Current benchmark materials exemplify these limitations. MgH_2, for instance, boasts a high theoretical gravimetric capacity (7.6 wt%) but requires temperatures above 300 °C for hydrogen release due to slow desorption kinetics.[35] Complex hydrides such as $LiBH_4$ and $NaAlH_4$ can achieve moderate hydrogen densities but often necessitate high temperatures, catalytic activation, or suffer from poor reversibility.[36] Porous frameworks (MOFs/COFs), while tunable and lightweight, rely primarily on weak physisorption and struggle to meet practical storage densities.[37] High-entropy alloys[38] and superhydrides, though scientifically intriguing, demand extreme synthesis or operating conditions (high pressures or cryogenic temperatures),[38,39] hindering their deployment in commercial systems.

The chemical diversity and complexity of hydrogen storage materials—ranging from AB_2, AB_3, and AB_5 interstitial hydride[40] to Mg-, Ti-, and V-based alloys, complex hydrides, and rare-earth-enriched compounds—make the search for optimal candidates challenging. Existing efforts to accelerate hydrogen storage material discovery are fragmented. Conventional computational databases primarily focus on crystalline structures and predicted thermodynamic properties, lacking integration with experimentally validated performance data. The absence of a comprehensive, machine-readable platform[41] that integrates both experimental and theoretical information has hindered the rational design and rapid screening of HSMs.

In this work, by integrating the database, machine learning models trained on this database, and LLMs, it becomes straightforward to construct materials-focused AI agents using simple instruction and schema interface functions (for more related details, refer to SI, Fig. S9–11). To initially assess the reliability of the AI agent's predictions, we did not require *DigHyd* to design entirely new materials. Instead, we focused on cases where comparable materials already exist in the database, allowing for direct validation (Fig. S12 and SI Video 1). Under these conditions, the *DigHyd* agent proposed compositions such as $Mg_2Ni_{0\cdot8}Co_{0\cdot2}$, $Mg_2Fe_{0\cdot8}Co_{0\cdot2}$, and $La_{0\cdot8}Mg_{0\cdot2}Ni_5$. Among these, $Mg_2Fe_{0\cdot8}Co_{0\cdot2}$ was predicted to exhibit a hydrogen storage capacity of 4.06 wt%. Importantly, analogous alloys already reported in the database, such as Mg_2FeH_6 and $Mg_2Fe_{1-x}Co_xH_6$, display capacities in the range of 4.5–5.5 wt%,[42,43] thereby supporting the consistency of the predictions.

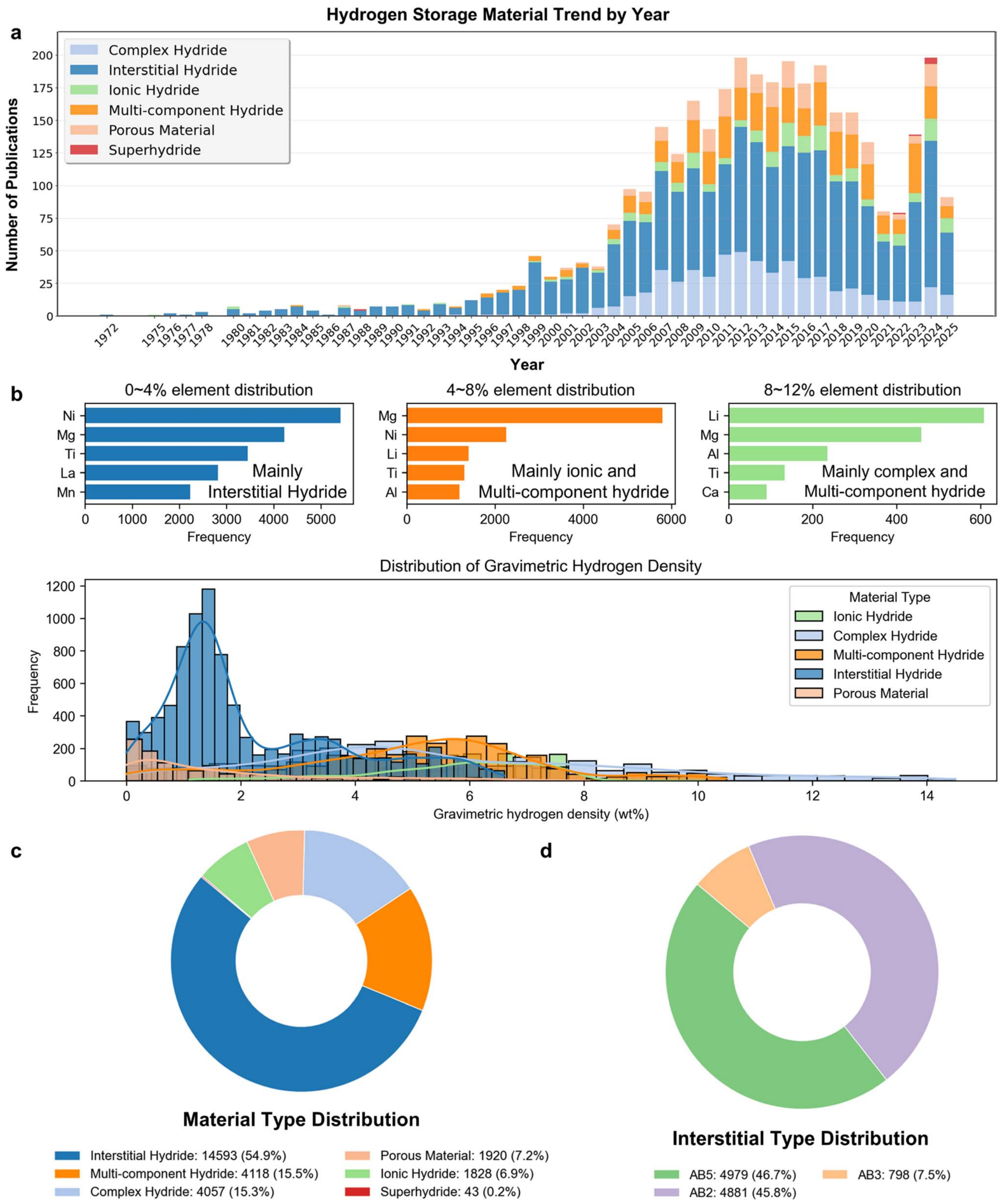


Fig. 3 Overview of data mining from over 4000 hydrogen storage materials publications. (a) Annual publication trends categorized by different types of hydrogen storage materials. (b) Distribution of 17 954 hydrogen storage capacity values, along with the elemental distribution of materials within three ranges: 0–4%, 4–8%, and 8–10%. (c) Overall distribution of hydrogen storage material types. (d) Type distribution of interstitial hydrides, classified into AB_2, AB_3, and AB_5 structures.

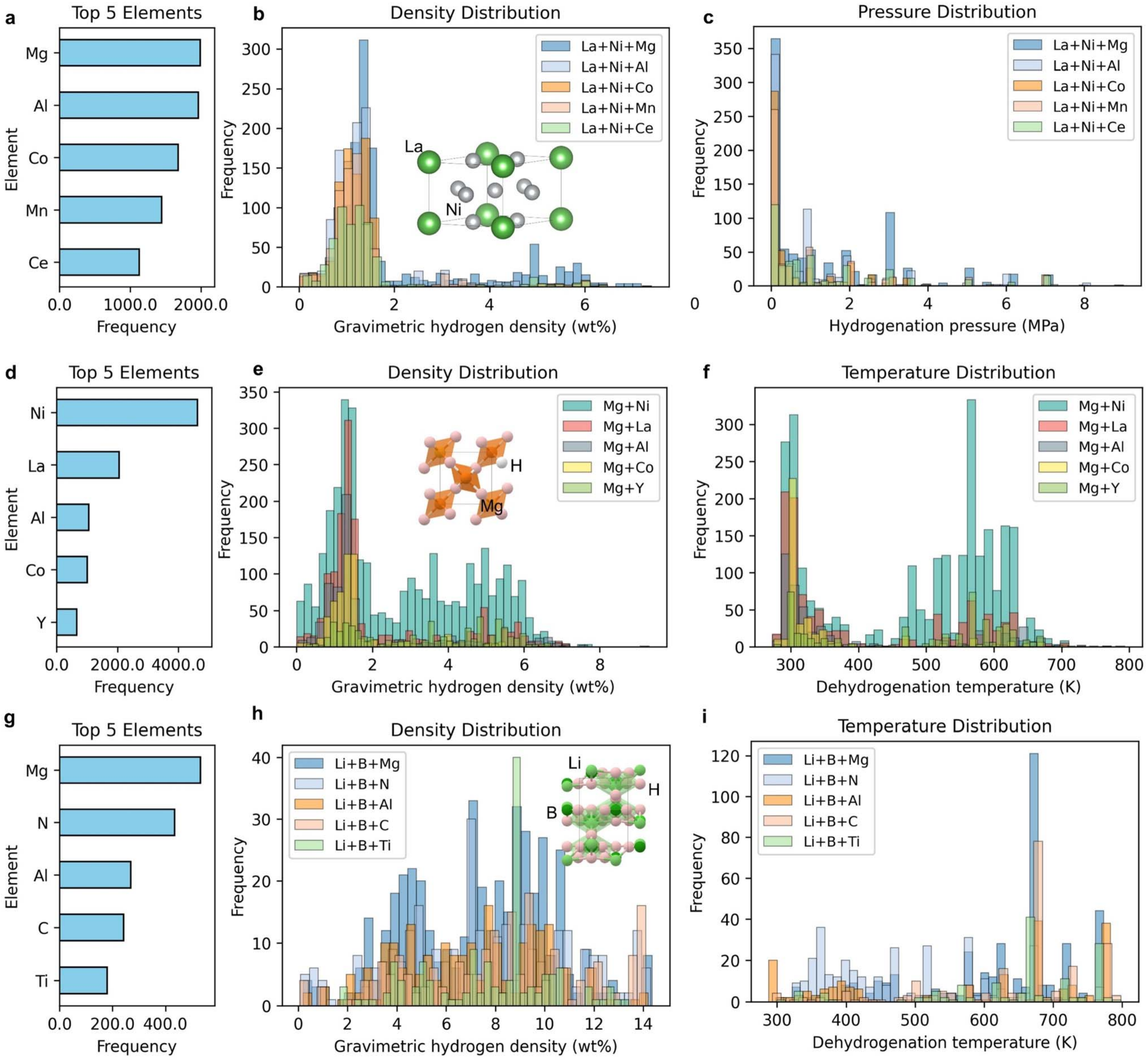
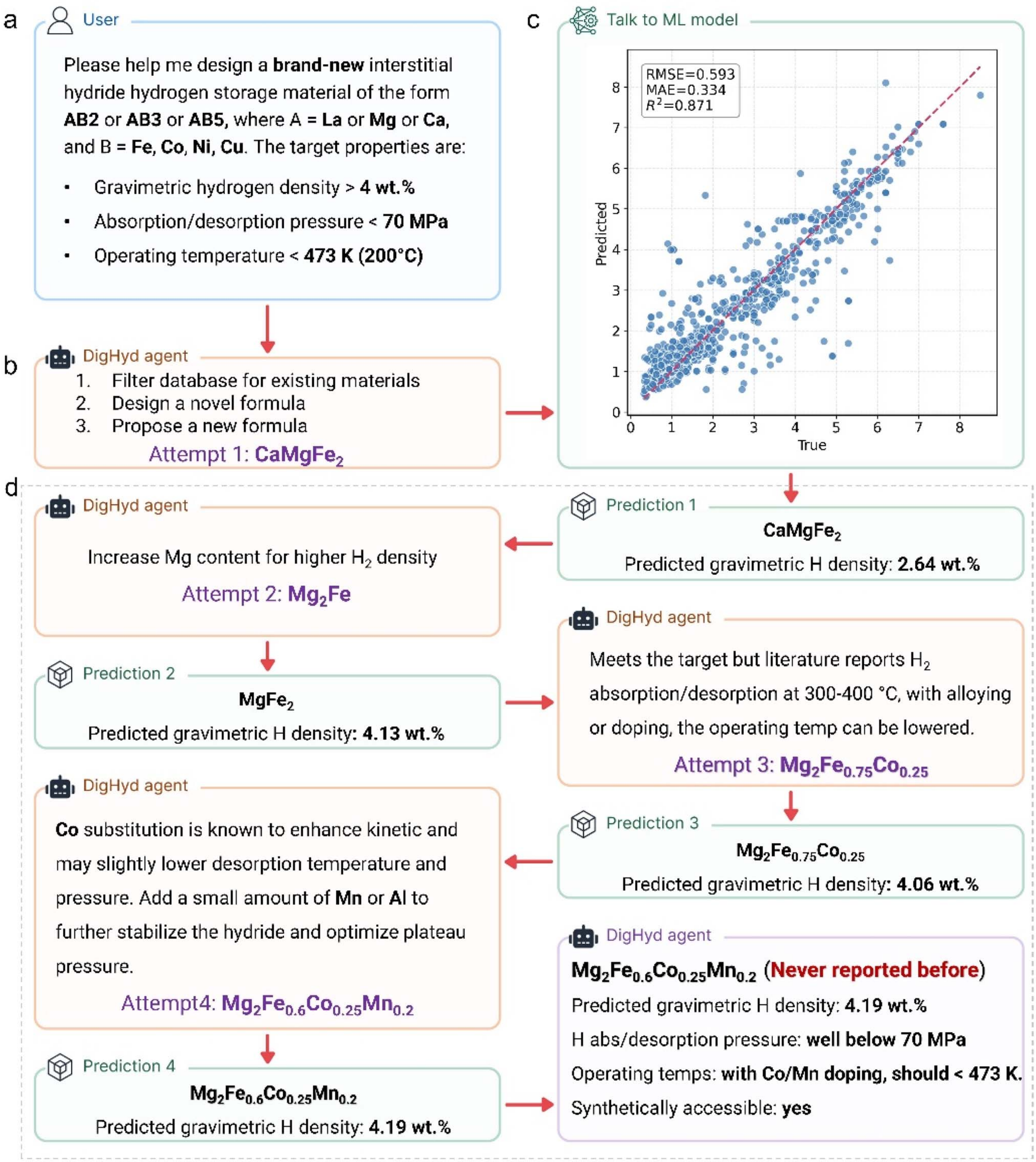


Fig. 4 Analysis of representative hydrogen storage materials. (a) Top 5 frequently added elements to $LaNi_5$, and the corresponding distributions of (b) hydrogen storage density and (c) hydrogen absorption equilibrium pressure upon element addition; (d) top 5 frequently added elements to MgH_2, and the corresponding distributions of (e) hydrogen storage density and (f) hydrogen desorption temperature; (g) top 5 frequently added elements to $LiBH_4$, and the corresponding distributions of (h) hydrogen storage density and (i) hydrogen desorption temperature.

Next, to verify that *DigHyd* can indeed design entirely new materials (Fig. 5 and SI Video 2), we applied the same prompting strategy but with explicit instructions to generate compositions never previously reported. Under these conditions, *DigHyd* demonstrated an iterative design–prediction–optimization capability, as illustrated in Fig. 5. In this workflow, researchers can guide the AI agent to propose novel materials by specifying the material class, potential elements, and target properties such as gravimetric hydrogen density, pressure, and temperature (Fig. 5a).

In the first round, leveraging the local knowledge base as well as the analytical, reasoning, and predictive capabilities of large language models, the *DigHyd* agent proposed $CaMgFe_2$ (Fig. 5b). This candidate was then evaluated using our machine learning model (see Methods: machine learning methods for model details, hyperparameters, and code), which predicts hydrogen density directly from the material formula. With an R^2 value of 0.87, the model provides a reliable first-pass screening for LLM-proposed candidates (Fig. 5c). $CaMgFe_2$ was predicted to store 2.64 wt% hydrogen (Fig. 5d). The agent subsequently suggested increasing the Mg content, resulting in Mg_2Fe with a predicted capacity of 4.13 wt%. However, literature reports indicated that this compound exhibits hydrogenation/dehydrogenation only at elevated temperatures (300–400 °C),



Please help me design a **brand-new** interstitial hydride hydrogen storage material of the form **AB2** or **AB3** or **AB5**, where A = **La** or **Mg** or **Ca**, and B = **Fe**, **Co**, **Ni**, **Cu**. The target properties are:

- Gravimetric hydrogen density > **4 wt.%**
- Absorption/desorption pressure < **70 MPa**
- Operating temperature < **473 K (200°C)**

1. Filter database for existing materials
2. Design a novel formula
3. Propose a new formula

Attempt 1: **$CaMgFe_2$**

$CaMgFe_2$

Predicted gravimetric H density: **2.64 wt.%**

Increase Mg content for higher H_2 density

Attempt 2: **Mg_2Fe**

$MgFe_2$

Predicted gravimetric H density: **4.13 wt.%**

Meets the target but literature reports H_2 absorption/desorption at 300-400 °C, with alloying or doping, the operating temp can be lowered.

Attempt 3: **$Mg_2Fe_{0.75}Co_{0.25}$**

$Mg_2Fe_{0.75}Co_{0.25}$

Predicted gravimetric H density: **4.06 wt.%**

Co substitution is known to enhance kinetic and may slightly lower desorption temperature and pressure. Add a small amount of **Mn** or **Al** to further stabilize the hydride and optimize plateau pressure.

Attempt4: **$Mg_2Fe_{0.6}Co_{0.25}Mn_{0.2}$**

$Mg_2Fe_{0.6}Co_{0.25}Mn_{0.2}$

Predicted gravimetric H density: **4.19 wt.%**

$Mg_2Fe_{0.6}Co_{0.25}Mn_{0.2}$ (Never reported before)

Predicted gravimetric H density: **4.19 wt.%**

H abs/desorption pressure: **well below 70 MPa**

Operating temps: **with Co/Mn doping, should < 473 K.**

Synthetically accessible: **yes**

Fig. 5 Workflow of AI agent-driven discovery of new hydrogen storage materials. (a) The user specifies key requirements, including material type, constituent elements, and performance targets. (b) The *DigHyd* agent proposes initial candidate compositions based on data mined from over 4000 historical publications. (c) The candidate compositions are evaluated using a pretrained machine learning model to predict their gravimetric hydrogen density. (d) The *DigHyd* agent rapidly designs, predicts, and iteratively refines candidate materials in line with researcher-defined goals within minutes. Finally, the *DigHyd* agent outputs the final material design, together with the relevant reaction conditions and an assessment of synthetic feasibility. (See SI Video 2 for the complete process and details).

failing to meet the design targets. In response, *DigHyd* refined the composition to $Mg_2Fe_{0.75}Co_{0.25}$, and further to $Mg_2Fe_{0.6}Co_{0.2}Mn_{0.2}$. The latter was predicted to achieve 4.19 wt% hydrogen storage capacity, with Mn (or alternatively Al) contributing to hydride stabilization and plateau pressure optimization. Importantly, this final composition has never been reported in the current database. Taken together, these results in Fig. 5d highlight the ability of the *DigHyd* agent to rapidly design, predict, and iteratively refine candidate materials in line with researcher-defined goals within minutes. If such AI-driven agents are directly integrated with high-throughput experimental platforms, the efficiency of materials discovery and development could be advanced to an unprecedented level.

To further increase the design difficulty, in the third case study (Fig. S13 and SI Video 3), we constrained the element space for material design (A = Mg or Ca, B =Ni). Leveraging the local knowledge base together with the analytical, reasoning, and predictive capabilities of LLM, the *DigHyd* agent proposed 8 candidate materials. Among these, one candidate exceeded the initial target of 4 wt% hydrogen capacity, while three achieved predicted performances above 3 wt%. The remaining candidates showed comparatively lower hydrogen densities. Based on these initial predictions, the *DigHyd* agent further optimized the proposed compositions by suggesting minor La and Y doping to enhance hydride phase stability and to reduce the hydrogenation/dehydrogenation temperature and pressure. The final designs, $Mg_2Ni_{2.9}La_{0.1}$ and $Mg_2NiY_{0.1}$, are derived from the Mg_2Ni system, a well-established intermetallic compound for hydrogen storage.[44] The introduction of a small amount of La or Y by partially substituting Ni is a common strategy to optimize hydrogen storage properties. The substitution ratio (3.3% for La[45] or Y[46]) is appropriate because it is sufficient to significantly influence the microstructure and hydrogen storage behavior without destroying the main phase structure. The addition of La or Y can promote grain refinement and introduce defects, which facilitate hydrogen diffusion, improve absorption/desorption kinetics, and may lower the hydrogenation/dehydrogenation temperature. Moreover, the larger atomic radii of La and Y compared to Ni lead to lattice expansion, thus reducing the activation energy for hydrogen diffusion.[46] Therefore, the proposed compositions are also rational for hydrogen storage materials, as supported by both theoretical understanding and experimental data from the literature. In fact, our database did not include this very recent paper [ref. 46] at the time of writing, which investigates the Mg–Y–Ni system. The findings presented in this work further demonstrate the reliability of the predictions made by our developed agent.

Conclusions

We developed DIVE (Descriptive Interpretation of Visual Expression), a multi-agent workflow that converts figure-embedded experimental information in scientific papers into structured, machine-readable data. By transforming key graphical elements (*e.g.*, PCT, TPD, and discharge curves) into descriptive text using schema-enforced prompts, DIVE enables efficient batch extraction. Applied to solid-state hydrogen storage materials, DIVE was used to mine 4053 publications (1972–2025) and build the *DigHyd* database with 30 435 entries. Across seven multimodal models, DIVE consistently outperforms direct extraction, with typical gains of 10–15% over commercial models and 15–30% over open-source models under the same benchmark. Building on this resource, we implemented the *DigHyd* agent, integrating natural-language querying with a machine-learning verifier to rapidly propose and refine candidate materials under user-defined constraints. Current limitations still include hallucinated fields, visual reading noise, and multi-plateau interpretation errors. Future work will focus on improving robustness to these failure modes and extending coverage to more complex figure types and long-range context, enabling more reliable literature-to-design pipelines for accelerated materials discovery.

Methods

DIVE workflow

The first step of the DIVE workflow involves converting PDF files into both text and image formats. This conversion process was accomplished using MinerU,[47] which efficiently extracts both textual content and embedded figures from scientific PDFs. All subsequent steps in the workflow were developed using the LangGraph package, enabling modular and robust pipeline construction for literature mining and data extraction. The complete set of codes, including workflow scripts, prompt engineering details, and evaluation protocols, has been made openly available in our GitHub repository (https://github.com/gtex-hydrogen-storage/DIVE) to ensure transparency and reproducibility. For the model used in our article (DeepSeek Qwen3 8B), we deployed it locally with an A6000 GPU. For other open- or closed-source models, we accessed them *via* API calls to third-party platforms or official websites—for example, service providers such as SiliconFlow (https://www.siliconflow.com/) and Groq (https://groq.com/). The exact model version strings and inference parameters (*e.g.*, temperature, maximum tokens, and retry numbers) used at each stage are summarized in SI, Table S1. Details of the document JSON structure and how surrounding textual context is constructed and passed to the models are provided in the SI (Table S3). The robustness of the DIVE workflow to multi-curve figures, overlapping curves, and low-resolution images can be found in Fig. S17. Accuracy breakdown for step 1 (caption-based figure identification), including precision, recall, and F1 scores, can be found in the SI (Table S6).

The digital hydrogen platform (*DigHyd*) database

All hydrogen storage materials data extracted *via* the DIVE workflow have been integrated into the digital hydrogen platform and are accessible through a web interface built with Streamlit (https://www.dighyd.org). As of August 2, 2025, the database currently contains 4053 literature sources and 30 435 unique entries, each corresponding to a distinct material or experimental condition. Users can interactively filter data, visualize results, and explore specific material properties or test conditions. We have also deployed the AI agent developed based on DIVE on the website. In addition, the *DigHyd* database is updated daily with newly published literature related to HSMs. The platform also provides direct access to the *DigHyd* agent and integrated machine learning regression models for data analysis and materials prediction.

Development of the *DigHyd* agent

The AI agent utilized in this study was rapidly built using OpenAI's custom GPTs and Actions functionality, allowing seamless integration with local knowledge bases and automated analysis tools. The local knowledge base is utilized *via* the OpenAI GPT "Knowledge" mechanism, which functions as an internal retrieval-augmented generation (RAG) pipeline rather than a separate, custom-built retriever. The agent's prompt instructions (Fig. S9–11), schema definitions, and action logic are also provided in the GitHub repository (https://github.com/gtex-hydrogen-storage/DIVE) for reference and reuse by the community. This infrastructure enables end-to-end question answering, data analysis, and material design based on literature-derived knowledge, supporting both interactive and automated workflows in materials research. Users who would like to use the *DigHyd* agent directly can visit https://www.dighyd.org. After registration, they may contact the corresponding authors to activate access and then launch the *DigHyd* agent *via* the public link in the sidebar.

Machine learning methods

We developed a machine learning workflow to predict material properties from chemical composition. After removing samples lacking valid target values or standard chemical formulae, each compound was parsed into the Pymatgen[48] Composition object. A total of 5357 data points were used in this study. Features were generated using the Matminer ElementProperty featurizer ("magpie" preset) and element molar fractions.[49] The XGBoost regressor was used for prediction, and model performance was evaluated by standard regression metrics. The dataset was randomly split into training and test sets with a ratio of 80% : 20%. Model training was performed using an XGBoost regressor. Hyperparameter optimization was conducted *via* GridSearchCV (with 3-fold cross-validation, scoring by negative mean squared error and parallel computation) to select the best model configuration. Model performance was evaluated using standard regression metrics. All code and scripts are available in our GitHub repository (https://github.com/gtex-hydrogen-storage/DIVE).

Author contributions

D. Z. conceived the project, coordinated the overall research, and wrote the main manuscript text. X. J. performed the primary data collection, analysis, and visualization. H. B. T. and S. H. J. contributed to theoretical calculations and assisted in data interpretation. L. Z. assisted with literature mining and data curation. R. S. contributed to methodology development. Y. H. assisted with computational workflow construction and data processing. T. S. supported experimental measurements and related analyses. K. K. contributed to data mining. S. -i. O. supervised the manuscript. H. L. conceived the project, supervised the computational and AI-driven components, guided the overall study design, and revised the manuscript. All authors discussed the results and reviewed the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

Supplementary information (SI): digital interface and codes can be accessed *via*: Digital Hydrogen Platform (*DigHyd*): https://www.dighyd.org, DigHyd Data Checking System: https://curvechecking.dighyd.org/, and Code repository: https://github.com/gtex-hydrogen-storage/DIVE. SI Video 1: *DigHyd* agent designs material $Mg_2Fe_{0.8}Co_{0.2}$. SI Video 2: *DigHyd* agent designs new material $Mg_2Fe_{0.6}Co_{0.2}Mn_{0.2}$. SI Video 3: *DigHyd* agent designs new materials $Mg_2Ni_{2.9}La_{0.1}$ and $Mg_2NiY_{0.1}$. SI Video 4: *DigHyd* agent for data analysis and visualization. SI Video 5: Main features of the Digital Hydrogen Platform (*DigHyd*). See DOI: https://doi.org/10.1039/d5sc09921h.

Acknowledgements

This work was supported by the Green Technologies of Excellence (GteX) Program, Japan (Grant No. JPMJGX23H1). We thank Mr. Kazuteru Eguchi and Mrs. Yukina Takada for their efforts in reviewing the data.

Notes and references

1 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, Scaling deep learning for materials discovery, *Nature*, 2023, **624**, 80–85.

2 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng and G. Ceder, An autonomous laboratory for the accelerated synthesis of novel materials, *Nature*, 2023, **624**, 86–91.

3 Y. Zhang, Y. Han, S. Chen, R. Yu, X. Zhao, X. Liu, K. Zeng, M. Yu, J. Tian, F. Zhu, X. Yang, Y. Jin and Y. Xu, Large language models to accelerate organic chemistry synthesis, *Nat. Mach. Intell.*, 2025, **7**, 1010–1022.

4 H. J. Kulik, Are we there yet? Adventures on a road trip through machine learning as a computational chemist, *APL Comput. Phys.*, 2025, 1.

5 X. Jiang, W. Wang, S. Tian, H. Wang, T. Lookman and Y. Su, Applications of natural language processing and large language models in materials discovery, *npj Comput. Mater.*, 2025, **11**, 79.

6 S. Miret and N. M. A. Krishnan, Enabling large language models for real-world materials discovery, *Nat. Mach. Intell.*, 2025, **7**, 991–998.

7 Y. Zheng, H. Y. Koh, J. Ju, A. T. N. Nguyen, L. T. May, G. I. Webb and S. Pan, Large language models for scientific

discovery in molecular property prediction, *Nat. Mach. Intell.*, 2025, **7**, 437–447.

8 Z. Wang, Z. Sun, H. Yin, X. Liu, J. Wang, H. Zhao, C. H. Pang, T. Wu, S. Li, Z. Yin and X.-F. Yu, Data-Driven Materials Innovation and Applications, *Adv. Mater.*, 2022, **34**, 2104113.

9 G. Zhao, L. M. Brabson, S. Chheda, J. Huang, H. Kim, K. Liu, K. Mochida, T. D. Pham, P. Prerna, G. G. Terrones, S. Yoon, L. Zoubritzky, F.-X. Coudert, M. Haranczyk, H. J. Kulik, S. M. Moosavi, D. S. Sholl, J. I. Siepmann, R. Q. Snurr and Y. G. Chung, CoRE MOF DB: A curated experimental metal-organic framework database with machine-learned properties for integrated material-process screening, *Matter*, 2025, **8**(6), 102140.

10 Y. Kang and J. Kim, ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models, *Nat. Commun.*, 2024, **15**, 4705.

11 M. P. Polak and D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, *Nat. Commun.*, 2024, **15**, 1569.

12 K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, Leveraging large language models for predictive chemistry, *Nat. Mach. Intell.*, 2024, **6**, 161–169.

13 Y. Zhang, S. Itani, K. Khanal, E. Okyere, G. Smith, K. Takahashi and J. Zang, GPTArticleExtractor: An automated workflow for magnetic material database construction, *J. Magn. Magn. Mater.*, 2024, **597**, 172001.

14 V. Fan, Y. Qian, A. Wang, A. Wang, C. W. Coley and R. Barzilay, OpenChemIE: An Information Extraction Toolkit for Chemistry Literature, *J. Chem. Inf. Model.*, 2024, **64**, 5521–5534.

15 S. X. Leong, S. Pablo-García, Z. Zhang and A. Aspuru-Guzik, Automated electrosynthesis reaction mining with multimodal large language models (MLLMs), *Chem. Sci.*, 2024, **15**, 17881–17891.

16 Z. Zheng, Z. He, O. Khattab, N. Rampal, M. A. Zaharia, C. Borgs, J. T. Chayes and O. M. Yaghi, Image and data mining in reticular chemistry powered by GPT-4V, *Digital Discovery*, 2024, **3**, 491–501.

17 S. X. Leong, S. Pablo-García, B. Wong and A. Aspuru-Guzik, MERMaid: Universal multimodal mining of chemical reactions from PDFs using vision-language models, *Matter*, 2025, **8**, 102331.

18 L. Schlapbach and A. Züttel, Hydrogen-storage materials for mobile applications, *Nature*, 2001, **414**, 353–358.

19 M. S. Dresselhaus and I. L. Thomas, Alternative energy technologies, *Nature*, 2001, **414**, 332–337.

20 G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang and E. Rosen, Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, *arXiv*, 2025, preprint arXiv:2507.06261, DOI: 10.48550/arXiv.2507.06261.

21 D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang and X. Bi, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, *arXiv*, 2025, preprint arXiv:2501.12948, DOI: 10.48550/arXiv.2501.12948.

22 S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang and J. Tang, Qwen2. 5-vl technical report, *arXiv*, 2025, preprint arXiv:2502.13923, DOI: 10.48550/arXiv.2502.13923.

23 L. Zhang, M. D. Allendorf, R. Balderas-Xicohténcatl, D. P. Broom, G. S. Fanourgakis, G. E. Froudakis, T. Gennett, K. E. Hurst, S. Ling, C. Milanese, P. A. Parilla, D. Pontiroli, M. Riccò, S. Shulda, V. Stavila, T. A. Steriotis, C. J. Webb, M. Witman and M. Hirscher, Fundamentals of hydrogen storage in nanoporous materials, *Prog. Energy*, 2022, **4**, 042013.

24 L. Lv, J. Lin, G. Yang, Z. Ma, L. Xu, X. He, X. Han and W. Liu, Hydrogen storage performance of LaNi3.95Al0.75Co0.3 alloy with different preparation methods, *Prog. Nat. Sci. Mater. Int.*, 2022, **32**, 206–214.

25 Y. Liu, D. Chabane and O. Elkedim, Optimization of LaNi5 hydrogen storage properties by the combination of mechanical alloying and element substitution, *Int. J. Hydrogen Energy*, 2024, **53**, 394–402.

26 J. Liu, L. Sun, J. Yang, D. Guo, D. Chen, L. Yang and P. Xiao, Ti–Mn hydrogen storage alloys: from properties to applications, *RSC Adv.*, 2022, **12**, 35744–35755.

27 G. Ek, M. M. Nygård, A. F. Pavan, J. Montero, P. F. Henry, M. H. Sørby, M. Witman, V. Stavila, C. Zlotea, B. C. Hauback and M. Sahlberg, Elucidating the Effects of the Composition on Hydrogen Sorption in TiVZrNbHf-Based High-Entropy Alloys, *Inorg. Chem.*, 2021, **60**, 1124–1132.

28 C. Li, P. Peng, D. W. Zhou and L. Wan, Research progress in LiBH4 for hydrogen storage: A review, *Int. J. Hydrogen Energy*, 2011, **36**, 14512–14526.

29 T. Matsunaga, F. Buchter, P. Mauron, M. Bielman, Y. Nakamori, S. Orimo, N. Ohba, K. Miwa, S. Towata and A. Züttel, Hydrogen storage properties of Mg[BH4]2, *J. Alloys Compd.*, 2008, **459**, 583–588.

30 P. Shan, L. Ma, X. Yang, M. Li, Z. Liu, J. Hou, S. Jiang, L. Zhang, L. Shi, P. Yang, C. Lin, B. Wang, J. Sun, H. Guo, Y. Ding, H. Gou, Z. Zhao and J. Cheng, Molecular Hydride Superconductor BiH4 with Tc up to 91 K at 170 GPa, *J. Am. Chem. Soc.*, 2025, **147**, 4375–4381.

31 L. Ren, Y. Li, N. Zhang, Z. Li, X. Lin, W. Zhu, C. Lu, W. Ding and J. Zou, Nanostructuring of Mg-Based Hydrogen Storage Materials: Recent Advances for Promoting Key Applications, *Nano-Micro Lett.*, 2023, **15**, 93.

32 A. A. Martínez, A. Gasnier and F. C. Gennari, From Iron to Copper: The Effect of Transition Metal Catalysts on the Hydrogen Storage Properties of Nanoconfined LiBH4 in a Graphene-Rich N-Doped Matrix, *Molecules*, 2022, **27**, 2921.

33 A. Gasnier and F. C. Gennari, Graphene entanglement in a mesoporous resorcinol–formaldehyde matrix applied to the nanoconfinement of LiBH4 for hydrogen storage, *RSC Adv.*, 2017, **7**, 27905–27912.

34 U.S. Department of Energy (DOE), Technical Targets: Onboard Hydrogen Storage for Light-Duty Vehicles (2025–2030), 2026, Available at: https://www.energy.gov/eere/

fuelcells/doe-technical-targets-onboard-hydrogen-storage-light-duty-vehicles.

35 C. Li, W. Yang, H. Liu, X. Liu, X. Xing, Z. Gao, S. Dong and H. Li, Picturing the Gap Between the Performance and US-DOE's Hydrogen Storage Target: A Data-Driven Model for MgH2 Dehydrogenation, *Angew. Chem., Int. Ed.*, 2024, **63**, e202320151.

36 S.-i. Orimo, Y. Nakamori, J. R. Eliseo, A. Züttel and C. M. Jensen, Complex Hydrides for Hydrogen Storage, *Chem. Rev.*, 2007, **107**, 4111–4132.

37 M. P. Suh, H. J. Park, T. K. Prasad and D.-W. Lim, Hydrogen Storage in Metal–Organic Frameworks, *Chem. Rev.*, 2012, **112**, 782–835.

38 R. R. Shahi, A. K. Gupta and P. Kumari, Perspectives of high entropy alloys as hydrogen storage materials, *Int. J. Hydrogen Energy*, 2023, **48**, 21412–21428.

39 Z. M. Geballe, H. Liu, A. K. Mishra, M. Ahart, M. Somayazulu, Y. Meng, M. Baldini and R. J. Hemley, Synthesis and Stability of Lanthanum Superhydrides, *Angew. Chem., Int. Ed.*, 2018, **57**, 688–692.

40 T. Sato, H. Saitoh, R. Utsumi, J. Ito, K. Obana, Y. Nakahira, D. Sheptyakov, T. Honda, H. Sagayama, S. Takagi, T. Kono, H. Yang, W. Luo, L. Lombardo, A. Züttel and S.-i. Orimo, Synthesis, Crystal Structure, and Hydrogen Storage Properties of an AB3-Based Alloy Synthesized by Disproportionation Reactions of AB2-Based Alloys, *J. Phys. Chem. C.*, 2025, **129**, 2865–2873.

41 M.-H. Van, P. Verma, C. Zhao and X. Wu, A Survey of AI for Materials Science: Foundation Models, LLM Agents, Datasets, and Tools, *arXiv*, 2025, preprint arXiv:2506.20743, DOI: 10.48550/arXiv.2506.20743.

42 D. Khan, S. Panda, Z. Ma, W. Ding and J. Zou, Formation and hydrogen storage behavior of nanostructured Mg2FeH6 in a compressed 2MgH2–Fe composite, *Int. J. Hydrogen Energy*, 2020, **45**, 21676–21686.

43 H. Bai, B. Liu, L. Kang, Y. Wang, J. Bai, S. K. Verma and Y. Xu, Reactive ball milling-induced improvement in hydrogen storage performance of Mg-Co alloys, *J. Alloys Compd.*, 2025, **1037**, 182230.

44 Y. Shang, C. Pistidda, G. Gizer, T. Klassen and M. Dornheim, Mg-based materials for hydrogen storage, *J. Magnesium Alloys*, 2021, **9**, 1837–1860.

45 A. Nobuta, F.-F. Hsieh, T. H. Shin, S. Hosokai, S. Yamamoto, N. Okinaka, T. Ishihara and T. Akiyama, Self-propagating high-temperature synthesis of La(Sr)Ga(Mg,Fe)O3−δ with planetary ball-mill treatment for solid oxide fuel cell electrolytes, *J. Alloys Compd.*, 2011, **509**, 8387–8391.

46 Y. Qi, D. Zhou, W. Sun, J. Li, Z. Cao, L. Xu, S. Guo, D. Zhao and Y. Zhang, Improving hydrogen storage characteristics of Mg–Ni-based alloys by adding Y and melt spinning, *J. Phys. Chem. Solids*, 2026, **208**, 113027.

47 B. Wang, C. Xu, X. Zhao, L. Ouyang, F. Wu, Z. Zhao, R. Xu, K. Liu, Y. Qu and F. Shang, Mineru: An open-source solution for precise document content extraction, *arXiv*, 2024, preprint arXiv:2409.18839, DOI: 10.48550/arXiv.2409.18839.

48 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

49 L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster and A. Jain, Matminer: An open source toolkit for materials data mining, *Comput. Mater. Sci.*, 2018, **152**, 60–69.